# Principal components for multivariate functional data

J.R. Berrendero*, A. Justel*[1] and M. Svarc**

*Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain
**Departamento de Matemática y Ciencias, Universidad de San Andrés and CONICET, Argentina

**Abstract**

A principal component method for multivariate functional data is proposed. Data can be arranged in a matrix whose elements are functions so that for each individual a vector of $p$ functions is observed. This set of $p$ curves is reduced to a small number of transformed functions, retaining as much information as possible. The criterion to measure the information loss is the integrated variance. Under mild regular conditions, it is proved that if the original functions are smooth this property is inherited by the principal components. A numerical procedure to obtain the smooth principal components is proposed and the goodness of the dimension reduction is assessed by two new measures of the proportion of explained variability. The method performs as expected in various controlled simulated data sets and provides interesting conclusions when it is applied to real data sets.

*Keywords:* Eigenvalue functions; Explained variability; Dimension reduction.
*A.M.S. subject classification. Primary:* 62H25; *secondary:* 62M86.
*Running Title:* Multivariate Functional PCA

# 1 Introduction

During the last decade the flood of data coming from internet traffic, computers, sensors and other technical devices has increased in an impressive way. Accordingly, statistical techniques aiming at handling huge amounts of information are specially relevant. For instance, sensors are currently able to measure a quantity of interest every few seconds, what entitle us to assume that our data are indeed functions, rather than vectors. As Ramsay and Dalzell (1991) highlight "some modelling problems are more natural

[1] Corresponding author: A. Justel, Departamento de Matemáticas, Universidad Autónoma de Madrid. Campus de Cantoblanco, 28049 Madrid, Spain. Email: ana.justel@uam.es

to think through in functional terms even though only finite number of observations are available". On the other hand, classical dimension reduction techniques, such as principal components, are significant in that they allow us to compress the information without much loss.

The idea behind most dimension reduction methods is to transform the original set of variables in such a way that only a few of the new transformed variables incorporates most of the information contained in the original ones. The key points are the set of transformations we are willing to consider, and the criterion to quantify the information. The most popular technique is the linear principal component analysis (PCA). In PCA, the new variables are uncorrelated linear combinations of the original ones, and the criterion is the variance. Other non-linear approaches include projection pursuit (Friedman and Tukey, 1974), independent component analysis (Hyvärinen and Oja, 2000) or principal curves (Hastie and Stuetzle, 1989, or Delicado, 2001), among others. In all these cases the dimension of the variable space is finite.

In this paper we propose a dimension reduction technique for multivariate functional data. By multivariate functional data we mean that each observation is a finite dimension vector whose elements are functions. These functions can be viewed as trajectories of stochastic processes defined on a given infinite dimensional function space. The structure of our data matrix is displayed in Figure 1. Our goal is to simplify the structure of the data set by summarizing the vector of functions for each individual (each row of the matrix) with a single function (or a very small set of functions) that retains as much information as possible from the original vector of functional observations. This goal parallels the one of the classical PCA, but here each entry of the data matrix is a function instead of a scalar. Observe that our purpose, although related, differs from that of the technique termed as *functional principal components analysis* (FPCA) in the monographs by Ramsay and Silverman (2002, 2005). In FPCA the goal is to reduce a sample of curves using feature vectors that represent in a low dimensional space the patterns of variability of the curves. Then, only one curve is observed for each individual, and the objective is to summarize it with a vector of real numbers. To deal with multivariate functional data Ramsay and Silverman (2005) propose to concatenate the functions into a single long function for each observation and then perform FPCA for the concatenated functions. The final result is again a vector of real numbers but not a function. Our procedure carries out a classical multivariate PCA for each value of the domain on which the functions are defined and consider the integrated variance as a suitable criterion to quantify the information retained by each component. In this way, the principal components we define are linear combinations of the original functions in the data set. However, given that the principal components are not unique (since the variance is not affected by a change of sign) a problem that arises is how to select the

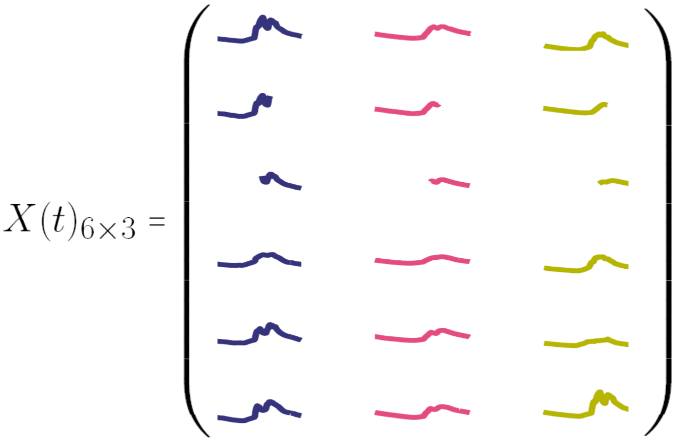


Figure 1: Data matrix for multivariate functional data.

solution that gives an appropriate degree of smoothing to the components.

Different versions and properties of FPCA have been studied by many authors. In classical multivariate analysis the principal components are linear combinations of the variables that represent the most significant modes of variation in the data. The weights of these linear combinations are obtained by solving an optimization problem that can be expressed in terms of the eigenvalues and eigenvectors of the covariance matrix with constraints involving the euclidean norm of the vector of weights. The natural extension to functional data is to replace the euclidean norm by the $L^2$-norm and the covariance matrix by the covariance function of the process generating the data. This general approach is described by Ramsey and Silverman (2005) and Ferraty and Vieu (2006). Nonetheless, the existing methods differ in the different choices of the functional space and in the way they handle the smoothing of the principal components. Rice and Silverman (1991) propose to project the data onto a finite-dimensional basis and then perform a standard multivariate analysis with a roughness penalty on the weight functions to achieve smoothness. Silverman (1996) incorporates smoothing by replacing the usual $L^2$-orthonormality constraint by orthonormality with respect to an inner product that takes into account the roughness of the functions. Boente and Fraiman (2001) propose a kernel-based estimate of the principal components. James *et al.* (2000) and Müller (2005) consider the problem of FPCA in the case of sparse data. More recently, Manté *et al.* (2007) apply PCA to densities relative to some fixed measure, Van der Linde (2008) approaches the problem from a Bayesian point of view, Park *et al.* (2009) study the problem of structural components which is related to FPCA and can provide a more helpful interpretation on certain data sets, and Delicado (2011) compare several dimensionality reduction methods for data that are density functions.

FPCA not only provides an enlightening way to look at data but also can be useful in practice. There are many examples that exhibit the usefulness of these techniques for finding interesting structures in real data sets. Among others, Ramsay and Silverman (2002) analyze data from several fields such as meteorology or growth. Wang *et al.* (2008) study data sets from online auction prices, and Gonzalez-Manteiga and Vieu (2007) study geophysical observations. Locantore *et al.* (1999) describe a robust procedure which is suitable for handling more complex data sets such as images.

Multivariate functional data have also been previously considered in multi-functional regression problems. In situations where many covariates are observed, Aneiros and Vieu (2006) introduce a semi-functional partial linear regression model, Shi *et al.* (2007) consider Gaussian process regression models, Ferraty and Vieu (2009) introduce models with an additive structure, and Shin (2009) proposes partial functional linear models. Some other papers on FPCA and problems concerning multivariate functional data have been included in recently published monographs and special issues of journals (see González-Manteiga and Vieu, 2007, Valderrama, 2007, Ferraty, 2010 and Ferraty and Romain, 2011)

Since our method applies to vectors of curves, some potential applications include the definition of rankings for such vectors, detection of clusters for multivariate functional data, or a way to deal with sensitivity to high dimensionality and multicolinearity in multi-functional regression models.

The remainder of the paper is organized as follows. In section 2 we present the dimension reduction method for multivariate functional data and provide some theoretical results about the conditions under which the principal components inherit the smooth behavior of the original functions. In section 3 we introduce a criteria to select an appropriate solution so that the resulting components are smooth and easier to interpret. We also introduce two different measures of the proportion of variability explained by each component. Section 4 is devoted to analyze the procedure performance on simulated and real data sets. Section 5 includes some final remarks. The proofs of the results are given in the appendix.

# 2 Multivariate functional principal components

## 2.1 Definitions and basic properties

We observe a vector of $p$ functions in a set of $n$ individuals, that is, the data that we consider can be arranged in an $n \times p$ matrix $M$ whose $(i, j)$ entry (denoted by $x_{ij}$) is the function $j$ corresponding to the individual $i$. We assume that all the functions are defined on the same compact real interval $[c, d]$ and take values in $I\!\!R$. Our goal is to simplify the structure of the data set by summarizing the vector of functions for each individual with a single function (or a very small set of functions) that retains as much information as possible from the original vector of functional observations.

Each row of the data matrix $M$ can be seen as a realization of a $p$–dimensional stochastic process $X := (X_1, \ldots, X_p)'$ defined on a probability space $(\Omega, \mathcal{F}, P)$. Without loss of generality we assume that, for each $t \in [c, d]$, the random vector $X(t) := (X_1(t), \ldots, X_p(t))'$ has a mean vector $\mu(t) = 0$. Whenever $\mu(t) \neq 0$, we work with $\tilde{X}(t) := X(t) - \mu(t)$ instead of $X(t)$. We also assume that $X(t)$ has a positive-definite covariance matrix $\Sigma(t) = X(t)X(t)'$.

We seek for a linear function of the components of $X$ that accounts for most of the information contained in $X$. Notice that, for any given function $a : [c, d] \to I\!\!R^p$ and all $t \in [c, d]$, it holds $\mathrm{Var}[a(t)'X(t)] = a(t)'\Sigma(t)a(t)$. The criterion we consider to measure the information is the integrated variance so that the weighting function $a_1 : [c, d] \to I\!\!R^p$ is defined as the function maximizing

$$\int_c^d a(t)'\Sigma(t)a(t)dt, \tag{2.1}$$

subject to $\|a(t)\| = 1$, for all $t \in [c, d]$, where $\| \cdot \|$ denotes the usual euclidean norm. The restriction on the norm of $a$ is needed to reach a unique solution for each $t$, except for the sign. We say that $Z_1(t) = a_1(t)'X(t)$ is the first principal component of $X$.

We proceed further by defining the rest of principal components

$$Z_r(t) = a_r(t)'X(t), \quad r = 2, \ldots, p. \tag{2.2}$$

Now, the weighting functions $a_r$ maximize (2.1) subject to $\|a(t)\| = 1$ and $a(t)'a_\ell(t) = 0$, for all $t \in [c, d]$ and $\ell = 1, \ldots, r-1$.

The following proposition collects some basic facts about the principal components as defined above.

**Proposition 1.** *For each $t \in [c, d]$, let $\lambda_1(t) > \cdots > \lambda_p(t) > 0$ be the eigenvalues of $\Sigma(t)$. Then:*
*(a) For all $t \in [c, d]$ and $r = 1, 2, \ldots, p$, $a_r(t)$ is a unit norm eigenvector corresponding to $\lambda_r(t)$.*
*(b) $\langle Z_r, Z_s \rangle = 0$, for $r \neq s$, where $\langle Z_r, Z_s \rangle := E\left[\int_c^d Z_r(t)Z_s(t)dt\right]$ is the inner product of $Z_r$ and $Z_s$ with respect to the product measure $dt \times dP$.*
*(c) $\|Z_r\|^2 = \int_c^d \lambda_r(t)dt$, for $r = 1, 2, \ldots, p$, where $\|Z_r\| = \left(E\left[\int_c^d Z_r(t)^2dt\right]\right)^{1/2}$ is the norm associated to the inner product defined in (b).*

This proposition is a direct consequence of the fact that variance is nonnegative. The solution to the sequence of optimization problems described above is reduced to find the principal components of $\Sigma(t)$ for each $t$. As a consequence, the weighting vectors $a_r(t)$ are the solutions to the classical principal component analysis in $I\!R^p$, that is, they are unit eigenvectors of $\Sigma(t)$.

If, for a given value of $t$, the vector $a(t)$ maximizes (2.1) subject to the appropriate length and orthogonality restrictions, then so does the vector $-a(t)$. We have therefore two different choices for each $t$ and, as a consequence, the number of possible weighting functions $a(t)$ defining each principal component is non finite. If the trajectories we observe are smooth, it would be reassuring that the corresponding principal components are also smooth. The following result gives conditions under which it is possible to choose a smooth $a(t)$ so that the principal components preserve the good behavior of the trajectories. It also gives formulas for the variation rate of both the eigenvalues and the eigenvectors of $\Sigma(t)$.

**Proposition 2.** *Let $t^* \in [c, d]$ be such that all the eigenvalues of $\Sigma(t^*)$ have multiplicity 1, and that the entries of $\Sigma(t)$ are differentiable functions at $t = t^*$. Let $\dot{\Sigma}(t^*)$ be the corresponding matrix of derivatives. Then, for $r = 1, \ldots, p$, it holds:*
*(a) The function $\lambda_r(t)$ is differentiable at $t^*$ with derivative $\dot{\lambda}(t^*)$ given by*

$$\dot{\lambda}(t^*) = a_r(t^*)'\dot{\Sigma}(t^*)a_r(t^*),$$

*where $a_r(t^*)$ is a unit eigenvector corresponding to $\lambda_r(t^*)$.*
*(b) It is possible to choose $a_r(t)$ so that it is differentiable at $t = t^*$. For this choice, the vector of derivatives $\dot{a}(t^*)$ is given by*

$$\dot{a}(t^*) = -\left[\sum_{\ell \neq r}[\lambda_\ell(t^*) - \lambda_r(t^*)]^{-1} a_\ell(t^*)a_\ell(t^*)'\right]\dot{\Sigma}(t^*)a_r(t^*). \tag{2.3}$$

Formula (2.3) implies that the behavior of $a_r(t)$ can be rather unstable when there exists $\ell \neq r$ such that $\lambda_r(t^*) \approx \lambda_\ell(t^*)$. Note that Proposition 2 is relevant even in the case when $\Sigma(t)$ does not depend on $t$ or it has eigenvalues with multiplicity greater than 1. The reason is that in practice we do not analyze $\Sigma(t)$ but an estimate $\hat{\Sigma}(t)$ that always depends on $t$ and has simple eigenvalues with probability 1. For instance, consider the case when the curves are equicorrelated with constant correlation $\rho$ and variance 1 so that $\Sigma(t) = \Sigma = (1-\rho)I_p + \rho 1_p 1_p'$ does not depend on $t$. Here, $I_p$ is the $p$–dimensional identity matrix and $1_p$ is a $p$–dimensional vector of ones. Obviously, both $\dot{\lambda}_r(t)$ and $\dot{a}_r(t)$ vanish for all $r$. However, as $\rho \to 0$, the difference $\lambda_2 - \lambda_1$ between the two largest eigenvalues of $\Sigma$ decreases so that the estimated difference $\hat{\lambda}_2(t) - \hat{\lambda}_1(t)$ is also expected to decrease for all $t \in [c, d]$. As a consequence, when $\rho$ is small, $\hat{a}_1(t)$ will tend to have an unstable behavior. Of course, the opposite will happen for large values of $\rho$.

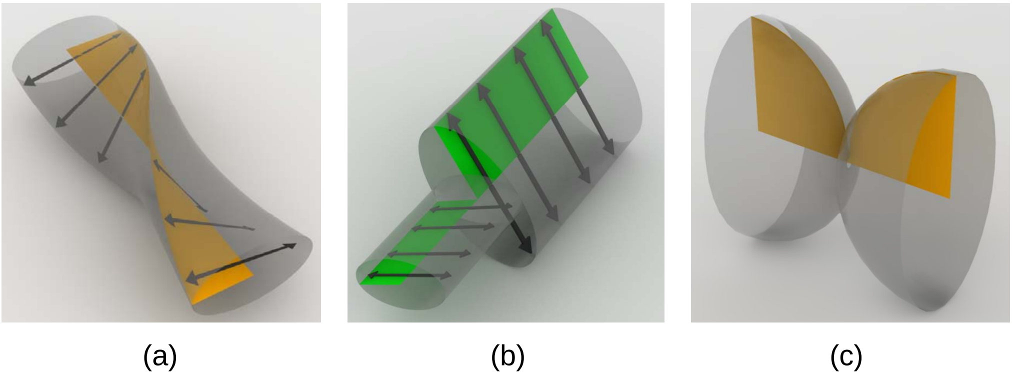


Figure 2: Curves inside these figures have covariance matrices with: (a) All the eigenvalues with multiplicity 1 and entries differentiable at all $t$; (b) All the eigenvalues with multiplicity 1 and discontinuous entries at a given $t$; (c) Eigenvalues with multiplicity 2 and continuous but non-differentiable entries at one $t$.

In some situations it may be contrived to assume that $\Sigma(t)$ is differentiable whereas the assumption of continuity is more natural. Under the assumption of absolute continuity of $\Sigma(t)$ it is also possible to show the continuity of $\lambda_r(t)$ and the existence of a continuous version of $a_r(t)$ [see e.g. Acker (1974)] so that principal components also inherit in this case the regularity of the covariances and the trajectories. Figure 2 exhibits the typical shapes of bidimensional data sets whose covariance matrices have entries with different degrees of regularity.

## 2.2 Empirical computation of the principal components

In practice, we must approximate the population principal components defined in section 2.1 using the matrix of sample functions. Following Proposition 1, the computation of the components at a particular $t \in [c, d]$ requires to carry out a $p$ dimensional principal component analysis of $\Sigma(t)$. Since $\Sigma(t)$ is unknown, it is natural to replace it by $\hat{\Sigma}(t)$, the sample covariance matrix corresponding to the $n \times p$ data matrix whose entries are $x_{ij}(t)$. Of course, for computing $\hat{\Sigma}(t)$ we have to evaluate at $t$ all the functions in the sample.

More precisely, if $\hat{\lambda}_1(t) < \cdots < \hat{\lambda}_p(t)$ are the eigenvalues of $\hat{\Sigma}(t)$, then the estimated weight function of the $r$-th component is given by $\hat{a}_r(t)$, a unit eigenvector corresponding to $\hat{\lambda}_r(t)$. Finally, for $r = 1, \ldots, p$, and $i = 1, \ldots, n$, the value of the $r$-th component for the observation $i$ is given by

$$z_{ir}(t) = \hat{a}_r(t)' x_i(t),$$

where $x_i(t) := (x_{i1}(t), \ldots, x_{ip}(t))'$.

In general, we are not only interested in estimating the principal components at a single value $t$, but over the whole interval $[c, d]$. In practice, we consider a grid of $N$ points $c = t_0 < t_1 < \cdots < t_N = d$, for large $N$, and assume that all the functions in the sample can be evaluated over such a grid. A preliminary smoothing of the data may be needed for this assumption to hold. For instance, a cubic spline smoothing can be applied (see Schumaker, 2007). The weights and values of the components for each point of the grid can then be obtained following the same procedure.

# 3 Practical details

In this section we address some practical issues that arise when one tries to implement the general procedure we have described in section 2. First, we have seen that the weights at $t$ are unique only up to a change of sing. This means that there exist infinite weighting functions for each component. In subsection 3.1 we propose a criterion for choosing the sign so that the resulting components are smooth and easier to interpret. On the other hand, once the principal components have been computed it is helpful to have an idea of the information contributed by each component compared with the information provided by the whole set of variables. In subsection 4 we discuss two different measures of the proportion of variability explained by each component. These measures can in turn be used to select the number of components we must retain.

## 3.1 Selecting the sign of the components

It seems desirable that the weighting functions $\hat{a}_r(t)$ defining the principal components do not change abruptly with $t$, since a steady behavior of the weights makes it easier the interpretation of the corresponding components. Regarding this question, an appropriate choice of the sign of $\hat{a}_r(t)$ is crucial, which otherwise does not matter in terms of explained variability. When the sample covariance matrix $\hat{\Sigma}(t)$ varies smoothly with $t$, applying Proposition 2 ensures the existence of smooth weighting functions. In this subsection we propose a criterion that aims to find such functions. The basic idea is to select the sign at $t$ which is closer on average to the signs already determined for values in a neighborhood of $t$.

Suppose first that we want to select the sign of the weighting function $\hat{a}_r(t^*)$ corresponding to the $r$-th component for a given $t^* \in [c, d]$, and that we have already selected the sign of $\hat{a}_r(t)$, for $t < t^*$. Let $u$ be a unit eigenvector of $\hat{\Sigma}(t^*)$ corresponding to the eigenvalue $\hat{\lambda}_r(t^*)$. We have to choose between the two options $\hat{a}_r(t^*) = u$ or $\hat{a}_r(t^*) = -u$. As mentioned above, the choice depends on the signs already chosen in a neighborhood of $t^*$. We propose to use an appropriate bandwidth $w$ to control the size of the neighborhood. As $w$ increases, the neighborhood is larger, that is, we take into account more previous choices. Once we have determined the neighborhood, we average over it the differences between the already chosen directions and the directions corresponding to $u$ and $-u$, and select the sign for which the average is smaller. More precisely, fix $w > 0$ and define $\tilde{t} = \max\{0, t^* - w\}$. We propose the following criterion depending on the bandwidth $w$: choose $\hat{a}_r(t^*) = u$ whenever

$$\int_{\tilde{t}}^{t^*} \|\hat{a}_r(t) - u\| \, dt \leq \int_{\tilde{t}}^{t^*} \|\hat{a}_r(t) + u\| \, dt, \tag{3.4}$$

and, otherwise, choose $\hat{a}_r(t^*) = -u$.

As we have mentioned in section 2, we will often consider a grid of $N$ points $c = t_0 < t_1 < \cdots < t_N = d$ over which the principal components are computed. In this case we must decide the value of $2^{N+1}$ signs. In this setting, the criterion given by (3.4) is equivalent to choose an arbitrary sign for $\hat{a}_r(t_0)$ and then, for $k = 1, \ldots, N$, select $\hat{a}_r(t_k) = u$ when

$$\sum_{j=\ell}^{k-1} \|\hat{a}_r(t_k) - u\| \leq \sum_{j=\ell}^{k-1} \|\hat{a}_r(t_k) + u\|,$$

where $\ell = \min\{j < k : t_j \geq t_k - w\}$ (with $\ell = 0$ when $t_{k-1} < t_k - w$). Furthermore, if the points in the grid are equispaced, fix a window $w$ amounts to fix a certain number $h = k - \ell$ of lags and select the sign of $\hat{a}_r(t_k)$ which is closer on average to the signs of its neighbors $\hat{a}_r(t_{k-h}), \ldots, \hat{a}_r(t_{k-1})$.

Later on, we will give some advice about which are the appropriate values of the bandwidth $w$ (or equivalently the number of lags, $h$). Our experiments show (see section 5.1) that the method works satisfactorily for a wide range of values of $w$ in that it gives stable weights under different settings. Still, performance depends on the dimension $p$ and the sample size $n$.

# 4 Proportion of variability explained by the components

Let $Z_r(t)$ be the $r$-th principal component as defined in (2.2). It is straightforward [see Proposition 1 (c)] to prove that $\mathrm{Var}[Z_r(t)] = \lambda_r(t)$, for each $t \in [c, d]$. We define the total variance at a given point $t \in [c, d]$ as $v(t) = \sum_{j=1}^{p} \lambda_j(t)$. In principle, there are two possible approaches to measure the proportion of variability accounted for by the $r$-th component. For the first approach we consider the local fraction of variability explained at $t$ by the component, $\lambda_r(t)/v(t)$, and then compute the average of all the local fractions. These considerations lead to define

$$\pi_{1r} = \frac{1}{d-c}\int_c^d \frac{\lambda_r(t)}{v(t)}dt. \tag{4.5}$$

Alternatively, we can also integrate out the variance of the component at $t$ and compare the result with the integral of the total variance. This approach yields the measure

$$\pi_{2r} = \frac{\int_c^d \lambda_r(t)dt}{\int_c^d v(t)dt}.$$

Both measures fulfill the natural requirement $\sum_{r=1}^p \pi_{1r} = \sum_{r=1}^p \pi_{2r} = 1$ but are different in general. Observe that the second measure can be rewritten as

$$\pi_{2r} = \int_c^d \frac{\lambda_r(t)}{v(t)}\omega(t)dt, \tag{4.6}$$

where

$$\omega(t) = \frac{v(t)}{\int_c^d v(s)ds}.$$

Comparing (4.5) and (4.6) we see that $\pi_{2r}$ can be understood as a weighted version of $\pi_{1r}$ where the most influential values of $t$ are those such that $v(t)$ is large with respect to the integrated total variance $\int_c^d v(t)dt$. As a consequence, when $\omega(t) = 1/(d-c)$, that is, when $v(t)$ does not depend on $t$, both measures coincide. Furthermore, when the component explains the same fraction of variability for all $t$ so that $\lambda_r(t)/v(t) = \kappa_r$ does not depend on $t$, it is also straightforward to check that $\pi_{1r} = \pi_{2r} = \kappa_r$.

To understand better the relative behavior of the two proposed measures of explained variability, consider the following simple example: given a positive constant $\bar{v} > 0$ and four independent standard normal random variables $Z_{11}, Z_{12}, Z_{21}, Z_{22}$, define $X(t) = (\epsilon_1(t), \epsilon_2(t))'$, where

$$\epsilon_1(t) = \begin{cases} \sqrt{0.5}\, Z_{11}, & t \in [0, 0.5) \\ \sqrt{0.9\bar{v}}\, Z_{12}, & t \in [0.5, 1] \end{cases},$$

and

$$\epsilon_2(t) = \begin{cases} \sqrt{0.5}\, Z_{21}, & t \in [0, 0.5) \\ \sqrt{0.1\bar{v}}\, Z_{22}, & t \in [0.5, 1] \end{cases}.$$

When $t \in [0, 0.5)$ the first principal component always explains 50% of the total variability $v(t) = 1$. Moreover, when $t \in [0.5, 1]$, the first principal component explains 90% of the total variability, which in this case is $v(t) = \bar{v}$. Therefore, neither the proportion $\lambda_1(t)/v(t)$ nor $\pi_{11}$ depend on $\bar{v}$. In fact, $\pi_{11} = 0.7$ for any $\bar{v}$. However, as $\bar{v}$ increases, the behavior of the functions in $[0.5, 1]$ has more weight in the final value of $\pi_{21}$. Hence, according to $\pi_{21}$ a large value of $\bar{v}$ implies that the first principal component explains a large amount of the total variability. Indeed, $\pi_{21} = (0.5 + 0.9\bar{v})(1 + \bar{v})^{-1}$. Thus, depending on $\bar{v}$ we get $\pi_{11} < \pi_{21}$ (when $\bar{v} > 1$) or $\pi_{11} > \pi_{21}$ (when $\bar{v} < 1$). We also have that $\pi_{21} \to 0.5$ as $\bar{v} \to 0$, and $\pi_{21} \to 0.9$ as $\bar{v} \to \infty$.

The considerations above allow us to identify situations for which both measures yield similar or different results. Depending on the particular problem at hand it may be of interest to give more relevance to areas of the interval $[c, d]$ for which the variance is larger. In these cases $\pi_2$ could be more suitable. For other problems, we could also use an ad-hoc weighting function $\omega(t)$ and use the corresponding measure derived from (4.6).

In practice, we will estimate both $\pi_{1r}$ and $\pi_{2r}$ from a sample of functions evaluated over a grid. In this case we replace the true eigenvalues by their natural estimators, and the integrals by sums over the points of the grid. Therefore, equations (4.5) and (4.6) lead to the following estimators:

$$\hat{\pi}_{1r} = \frac{1}{N+1} \sum_{k=1}^{N} \frac{\hat{\lambda}_r(t_k)}{\hat{v}(t_k)}$$

and

$$\hat{\pi}_{2r} = \frac{\sum_{k=1}^{N} \hat{\lambda}_r(t_k)}{\sum_{k=1}^{N} \hat{v}(t_k)},$$

where $\hat{v}(t) = \sum_{j=1}^{p} \hat{\lambda}_j(t)$.

When both the sample size, $n$, and the size of the grid, $N$, go to infinity, the estimates $\hat{\pi}_{1r}$ and $\hat{\pi}_{2r}$ converge to the true values $\pi_{1r}$ and $\pi_{2r}$ with probability 1. A sketch of the proof of this fact for $\pi_{1r}$ is as follows (a similar argument works for $\pi_{2r}$): define

$$\tilde{\pi}_{1r} = \frac{1}{N+1} \sum_{k=1}^{N} \frac{\lambda_r(t_k)}{v(t_k)}.$$

and observe that $|\hat{\pi}_{1r} - \pi_{1r}| \leq |\hat{\pi}_{1r} - \tilde{\pi}_{ir}| + |\tilde{\pi}_{ir} - \pi_{1r}|$. Since $\hat{\lambda}_r(t)$ and $\hat{v}(t)$ are consistent estimates for $\lambda_r(t)$ and $v(t)$, it holds $|\hat{\pi}_{ir} - \tilde{\pi}_{ir}| \to 0$ a.s., as $n \to \infty$. Moreover, under mild regularity assumptions on $\Sigma(t)$ (consider, for instance, the conditions of Proposition 2) it is possible to guarantee that $\lambda_r(t)/v(t)$ is Riemann-integrable so that $|\tilde{\pi}_{ir} - \pi_{1r}| \to 0$, as $N \to \infty$.

# 5 Examples

## 5.1 Simulated toy data

We first show the performance of the multivariate functional principal component method in a simple example with only two sinusoidal functions. We analyze the connection between the conditions in Proposition 2 and the smoothness properties of the principal components and weight functions, as well as the sensitivity of the sing selection method introduced in section 3.1 to the bandwidth choice.

We generate 100 pairs of functions from

$$X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} = \begin{pmatrix} \sin(t) + 0.5\epsilon_1(t) \\ 3\sin(t) + 0.5\epsilon_2(t) \end{pmatrix} \qquad t \in [0, 2\pi], \tag{5.7}$$

where the error distributions are

$$\begin{pmatrix} \epsilon_1(t) \\ \epsilon_2(t) \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \qquad t \in [0, 2\pi], \tag{5.8}$$

with $\epsilon_i(t)$ independent of $\epsilon_i(s)$ for $t \neq s$ and $i = 1, 2$.

Figures 3 (a) and (c) show the simulated functions for the two extreme cases, almost uncorrelated data, $\rho = 0.1$, and high correlated data, $\rho = 0.9$. Apparently there are no differences between both figures since they do not show the pairwise correlations among functions.

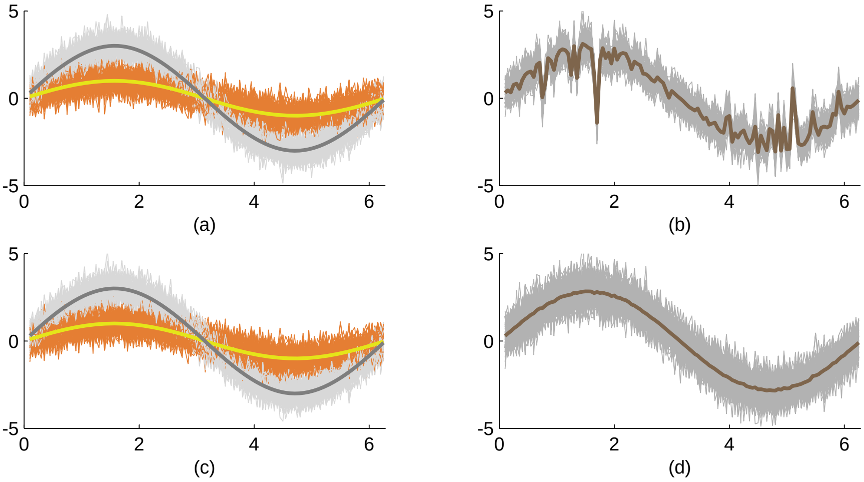


Figure 3: (a) and (c) are the data generated from (5.7) and (5.8), with $\rho = 0.1$ and $\rho = 0.9$, respectively. Dark lines are the generated $X_1(t)$ functions and light grey lines are the $X_2(t)$; (b) and (d) are the corresponding first principal component functions. The thick lines are the projections of the non-pertubed functions sin(t) and 3sin(t) considering the weight function given by the first principal component.

The first principal components for the data on (a) and (b) are displayed in Figures 3 (b) and (d), respectively. In both cases we use a thin grid of equispaced points and the sign is decided with the information given by the $h = 8$ previous neighbors, that is equivalent to a bandwidth $w = 0.4$. The thick lines are the principal components for the non–perturbed functions $\sin(t)$ and $3\sin(t)$. The first principal component functions for

$\rho = 0.9$ are steadier than for $\rho = 0.1$, even though $\Sigma(t)$ is a constant matrix in both cases. Looking at the principal component weight functions in Figure 4, we observe essentially the same results in the two cases. As expected, the first principal components are the sum of the two functions, whereas the second ones are the difference. The pictures in Figure 4 (a) and (d), and (b) and (e) seem different, but this is a matter of data variability and the particular choices of $\rho$ values. For $\rho = 0.1$ we do not succeed in finding smooth principal component trajectories as shown in Figure 3 (b), even for the case of the non-perturbed function that is clearly smooth in Figure 3 (a). This reflects the fact that, in practice, matrix $\hat{\Sigma}(t)$ is not smooth when correlation is small, as it is required in Proposition 2. Some of the marked peaks and valleys in Figure 3 (b) correspond to instants at which the sample estimation of $\rho$ is approximately zero. As we observe in Figure 5 (a), the estimated first principal component directions randomly shift for consecutive cross-sectional data ($t = 1.5, 1.55, \ldots, 1.75$). However, the sequence of estimated directions is stable when correlation increases, as in Figure 5 (b) for $\rho = 0.9$.

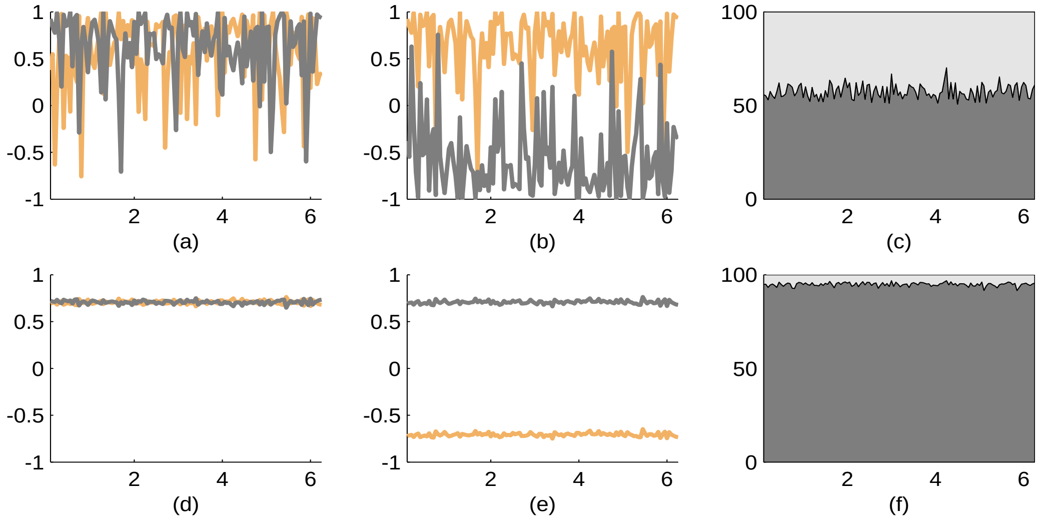
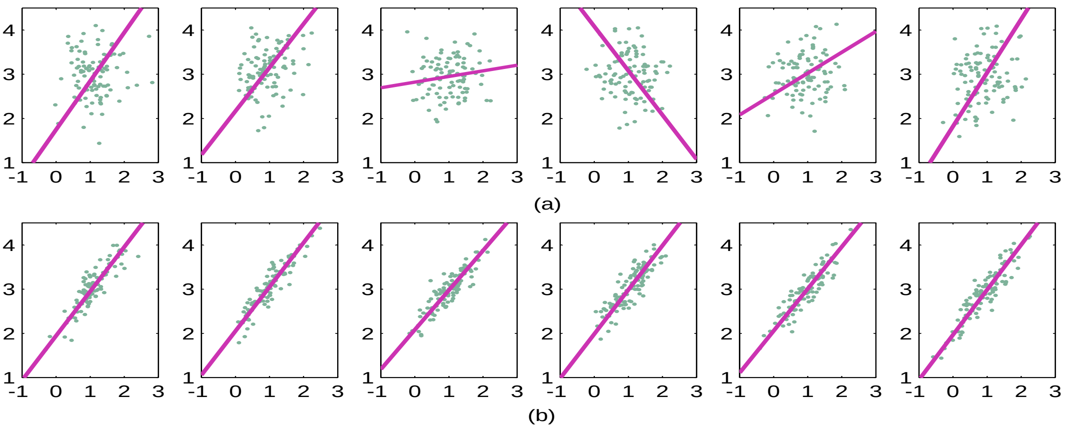


Figure 4: For $\rho = 0.1$, (a) first principal component weight functions, (b) second principal component weight functions, (c) local percentage of explained variability by the first (dark area) and second (light area) principal components. For $\rho = 0.9$, the corresponding plots are (d), (e) and (f).

The local percentages of explained variability by the principal components, $\hat{\lambda}_1(t)/\hat{v}(t)$ and $\hat{\lambda}_2(t)/\hat{v}(t)$, are showed in Figure 4 (c) and (f) for $\rho = 0.1$ and $\rho = 0.9$, respectively. The first principal component explained variability (dark area) is much higher and has less variability for $\rho = 0.9$ since in this case the first principal component contains almost the same information than the two original functions. The mean explained variability percentage by the first principal component ($\hat{\pi}_{11}$) ranges from 57.45% for $\rho = 0.1$ to 94.87% for $\rho = 0.9$. While $\hat{\pi}_{21}$ ranges from 57.18% to 94.94%. Notice that the theoretical values are $\pi_{11} = \pi_{21} = 0.55$, for $\rho = 0.1$, and $\pi_{11} = \pi_{21} = 0.95$, for


(a)

(b)

Figure 5: Scatters plots of the cross-sectional data at $t = 1.5, 1.55, \ldots, 1.75$, (a) for $\rho = 0.1$ and (b) for $\rho = 0.9$. The lines show the estimated first principal component directions.

$\rho = 0.9$.

We now increase the dimension $p$ of the multivariate functional data set in order to analyze the sensitivity of the results to the bandwidth choice (or number of lags $h$ in equispaced grids) for selecting the signs on the weight functions $\hat{a}_r(t)$. For dimensions $p = 2$, 4 ,6 ,10 and 30, we generate 1,000 multivariate functional data sets of size $n = 100$ from the vector of functions $X(t) = (X_1(t), \ldots, X_p(t))$ defined on $t \in [0, 2\pi]$, where $X_1(t) = \sin(t) + 0.5\epsilon_1(t)$ and $X_i(t) = 3(i-1)\sin(t) + 0.5\epsilon_i(t)$ for $i = 2, \ldots, p$. The distribution of the error vector $(\epsilon_1(t), \ldots, \epsilon_p(t))$ is multivariate $\mathcal{N}_p(0, S(t))$, where

$$S(t) = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \ldots & \rho & 1 \end{pmatrix}, \quad t \in [0, 2\pi],$$

and $\epsilon_i(t)$ independent of $\epsilon_i(s)$ for $t \neq s$ and $i = 1, \ldots, p$.

For any value of $\rho$, the eigenvector function $a_1(t)$ associated to the function of maximum eigenvalues of the covariance matrices $\Sigma(t)$ takes the values $(1/\sqrt{p}, \ldots, 1/\sqrt{p})$ or $(-1/\sqrt{p}, \ldots, -1/\sqrt{p})$ for all $t$. This means that if the covariance matrix is well estimated, all the coordinates must have the same sign, for all $t$. In addition, if we want $\hat{a}_1(t)$ to be a smooth function, then the sign should also be kept constant for all $t$. For the rest of eigenvalues and for all $t$, the corresponding eigenvectors have at least one coordinate with different sign.

For different values of $\rho$ and any of the 1,000 data sets, we compute the proportion of times, along the $[0, 2\pi]$ grid, that the first principal component is estimated correctly, in the sense that all the coordinates have the same sign. When $\rho > 0.3$ and $h = 1$ no mistakes have been observed. This means that for a moderate value of correlation it is enough to consider only one step backwards to choose the sign, at least for $n = 100$ and $p$ dimension from 2 to 30. We start to observe some errors just in the case of low correlation. For $\rho = 0.1$, we report on Table 1 the mean for the 1,000 data sets of the percentage of times $t$ where $\hat{a}_1(t)$ includes at least one coordinate with different sign to the others. Considering the rest of the cases in which the estimation is correct, we compute for each data set the relative error rate, that is, the proportion of times that all the coordinate signs shift at the same time and, therefore, the function $\hat{a}_1(t)$ is not smooth on these points. In Figure 6 we represent for different dimensions the mean relative error rate curves against the number of lags consider to select the sign. In a conservative strategy we recommend to consider $h = 5$ (a bandwidth of size 0.25), but in general, with 1 or 2 lags we will reach very satisfactory results. Notice that the models with low correlation are not of practical interest in the context of dimension reduction.

| $p$ | 2 | 4 | 6 | 10 | 30 |
|---|---|---|---|---|---|
| Error mean | 16% | 21% | 19% | 14% | 8% |

Table 1: Mean of the percentages of times where the first principal component includes at least one coordinate with different sign to the rest (1,000 data sets of size $n = 100$, with $\rho = 0.1$)

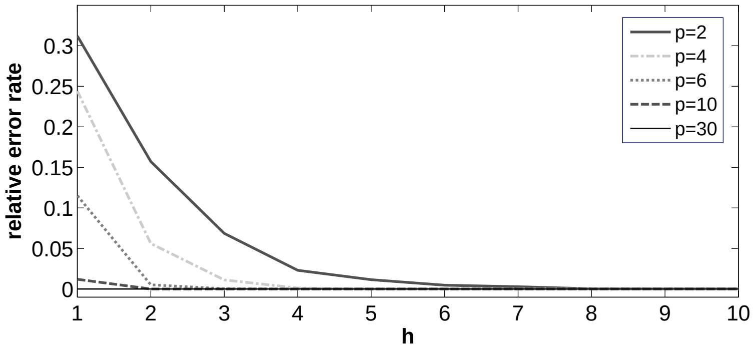


Figure 6: Mean of the relative error rate against the number of lags consider to select the sign (1,000 data sets of size $n = 100$, with $\rho = 0.1$)

Finally, we introduce a variation on (5.7) and (5.8) in order to generate smooth functions. Instead of generating different pair of errors at each $t \in [0, 2\pi]$, we consider constant error functions $\epsilon_1(t) = \epsilon_1$ and $\epsilon_2(t) = \epsilon_2$. We generate 100 pairs of functions from

$$X(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \end{pmatrix} = \begin{pmatrix} k_1 \sin(t) + 0.5\epsilon_1 \\ 3k_2 \sin(t) + 0.5\epsilon_2 \end{pmatrix} \qquad t \in [0, 2\pi], \tag{5.9}$$

where $k_1$ and $k_2$ are independent random variables with uniform distribution on the interval $(0, 2)$, and the error distribution is

$$\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right). \tag{5.10}$$

Figure 7 (a) shows the simulated multivariate functional data set. To compute the principal component functions we use a thin grid of equidistant points and, based on the previous results, we select the sign with only $h = 2$ lags, that is equivalent to a bandwidth $w = 0.1$. The first principal component functions are showed in Figure 7 (b).

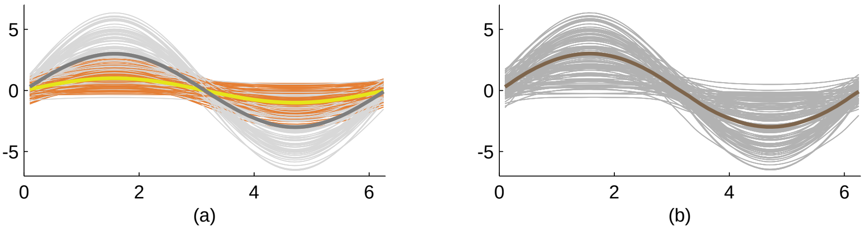


Figure 7: (a) Data generated from (5.9) and (5.10), the dark lines are the generated functions $X_1(t)$ and the light grey lines are the $X_2(t)$; and (b) first principal component functions. The thick lines are the projections of the non-pertubed functions sin(t) and 3sin(t) considering the weight function given by the first principal component.

Now, the entries of the covariance matrix $\Sigma(t)$ are non constant continuous functions. At $t = 0$, $\pi$, $2\pi$, the two variance functions reach the minimum value 0.25, and the correlation function the maximum value 0.9. The immediate consequence is that the principal component weight functions are non constant functions and inherit the same behavior in terms of continuity and differentiability as can be seen in Figures 8 (a) and (b). The shape of the weight functions, which in turn is determined by the sign choice, guarantees the smoothness of the principal component functions in Figure 7 (b).

The local percentage of variability explained by the first principal component, $\hat{\lambda}_1(t)/v(t)$, also changes with $t$, and is higher around $t = 0$, $\pi$, $2\pi$ as we observe in Figure 8(c).

## 5.2 Brownian motion simulated data

We consider a new simulated multivariate functional data set of size $n = 50$ and dimension four. The functions are generated according to the following distribution,

$$(X_1, X_2, X_3, X_4)' = A \times (B_1, B_2, B_3, B_4)',$$

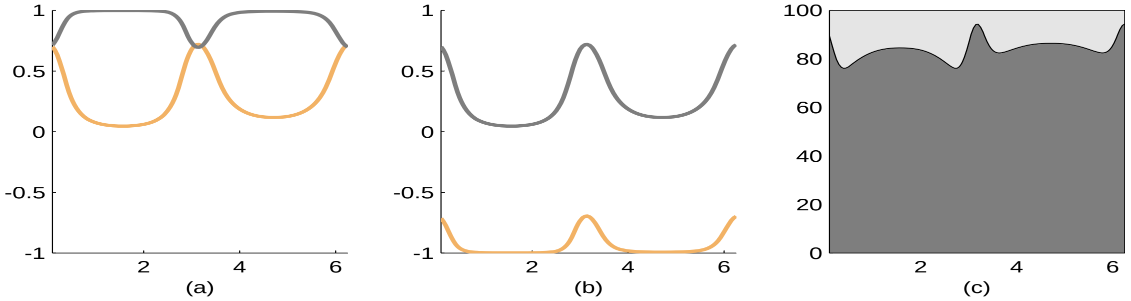


Figure 8: (a) First principal component weight functions, (b) second principal component weight functions, and (c) local percentage of explained variability by the first (dark area) and second (light area) principal components. Dark lines are the weight functions for $X_1(t)$ and the light grey lines are for $X_2(t)$.

where the $B_i's$ are four independent Brownian motions and $A$ is the Toeplitz matrix given by

$$A = \begin{pmatrix} 1 & 0.7 & 0.35 & 0.17 \\ 0.7 & 1 & 0.7 & 0.35 \\ 0.35 & 0.7 & 1 & 0.7 \\ 0.17 & 0.35 & 0.7 & 1 \end{pmatrix}.$$

We generate the data at the grid points $t = 0, 1, \ldots, 100$, and do not consider the values for $t = 0$. The window size is $h = 8$ back-steps.

With this data set we intend to approximate a real data problem where usually the first principal component is a weighted mean of all the variables. The weights depend on the covariance structure. In this case $\Sigma(t)$ is given by

$$\Sigma(t) = tAA' = t \begin{pmatrix} 1.64 & 1.7 & 1.31 & 0.83 \\ 1.7 & 2.1 & 1.89 & 1.31 \\ 1.31 & 1.89 & 2.1 & 1.7 \\ 0.83 & 1.31 & 1.7 & 1.64 \end{pmatrix}.$$

This covariance structure provides weights that are supposed to be constant. In Figure 9, we observe some variability on the estimated weight functions for the first principal component due to the sample estimation of $\Sigma(t)$. As expected, the higher weights correspond to the variables $X_2$ and $X_3$. Figure 10 (a) and (b) exhibit two randomly selected data from this multivariate functional data set. The generated functions are represented on the top of the two figures and the corresponding first principal component functions are represented on the bottom. The first principal components capture adequately the very different evolution of the functions for the two multivariate functional data. The mean explained variability percentage by the first principal component is $\hat{\pi}_{11} = 83.99\%$ and $\hat{\pi}_{21}$ is 82.84%. The theoretical values are $\pi_{11} = \pi_{21} = 0.8467$.

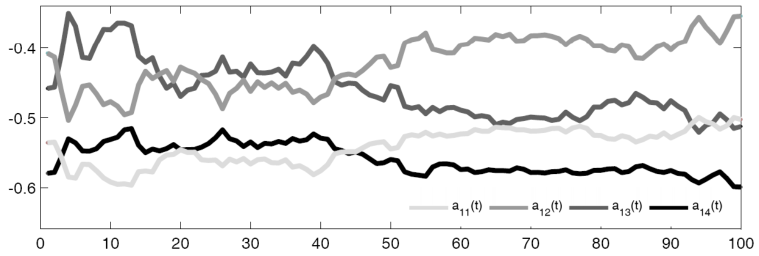


Figure 9: Weight functions for the first principal component of the four Brownian motion simulated processes.

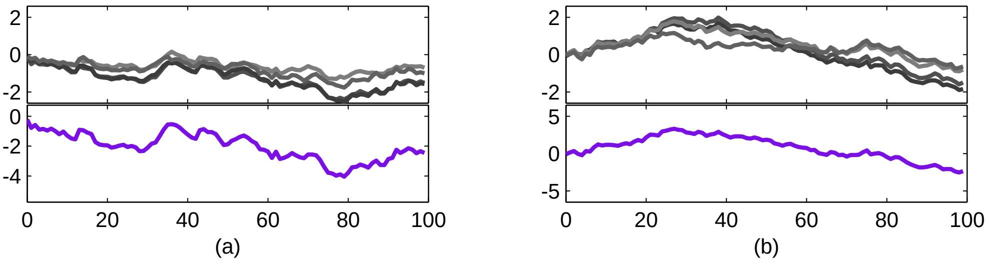


Figure 10: (a) and (b) are two randomly selected data from the four Brownian motion simulated data set. On the top, the four generated functions. On the bottom the corresponding first principal component functions.

## 5.3 Real data: Temperature summary on road experiments

Increased travel demand and the associated increased congestion will exacerbate the problem of reducing delays caused by pavement maintenance and repair. On the design of new pavement materials, Spanish civil engineers carry out complex experiments to study in a controlled environment the expected useful life time of the pavement, as well as the possible fatigue causes. By simulating the traffic load, they analyze the pavement response and performance under accelerated accumulation of damage in a compressed time period (see Coetzee *et al.* 2000). Throughout the experiment, two heavy vehicle simulators go around an oval pavement test track until the pavement wears down. The full-scale and accelerated pavement testing facility is located 18 km north of Madrid (Spain) at the Spanish Center for Road Transport Studies (CEDEX). There are several sensors at certain points on the test track that measure more than 100 parameters. We focus the attention on the temperature registers, measured at 3, 9 and 12 cm under surface. It is well known that temperature affects the pavement wear, but including the three values in any explicative model leads us to the problem of multicolinearity. We propose to summarize the three temperatures with the multivariate functional principal component method proposed in section 2.

The data consists of the three daily temperature functions registered during 21 days of November in 2007 (Madrid fall season), $x_3(t)$, $x_9(t)$ and $x_{12}(t)$, where $t$ cover a complete day (see Figure 11). The temperatures are registered when the truck pass over the sensors at almost equidistant time intervals, six minutes. The truck's speed is constant, but the huge amount of data that should be registered at the same time frequently collapses the data logging system. The sensor malfunctions are not easy to repair immediately and adjustments may require some days. The immediate consequence is that there is a large amount of missing data. We use a grid of 240 equidistant points, but excluding 48 instants in which the number of observed days is less than four.

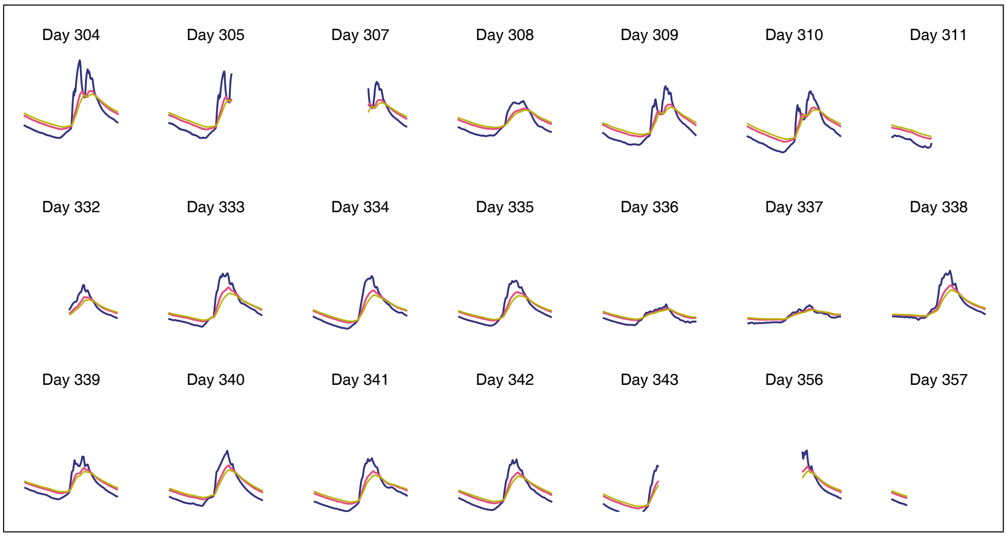


Figure 11: Daily temperatures registered by the 3 cm (dark/blue lines), 9 cm (grey/red lines) and 12 cm (light/green lines) deep sensors located at CEDEX pavement test track (Madrid, Spain) in 2007 (days indexed by Julian calendar).

The minimum temperature is 0.4°C, registered in the early morning, and the maximum is 29°C, registered by the 3 cm deep sensor in the afternoon. The three temperatures are similar at night and homogeneous along the observed period. However during the day, the different meteorological conditions produce the variability that we observe on the 3 cm deep temperature in Figure 12(a). The unusual peaks that appear some days in the afternoon (clear sky days) are caused by a column shadow projected on the pavement at the sensor location. Sensors at 9 and 12 cm deep are less affected by surface temperatures and present more homogeneous behavior. The multivariate functional principal components displayed in Figure 12 reflect this pattern. The first principal component is a function that average the three temperature curves with the vector of weight functions $a_1(t) = (a_{1,3cm}(t), a_{1,9cm}(t), a_{1,12cm}(t))$ showed in Figure 12(b). The first principal component function is the mean of the three temperatures at night. During the day the weight function $a_{1,3cm}(t)$ is higher that the other two, except at noon and the shadow time interval. The first principal component can be considered

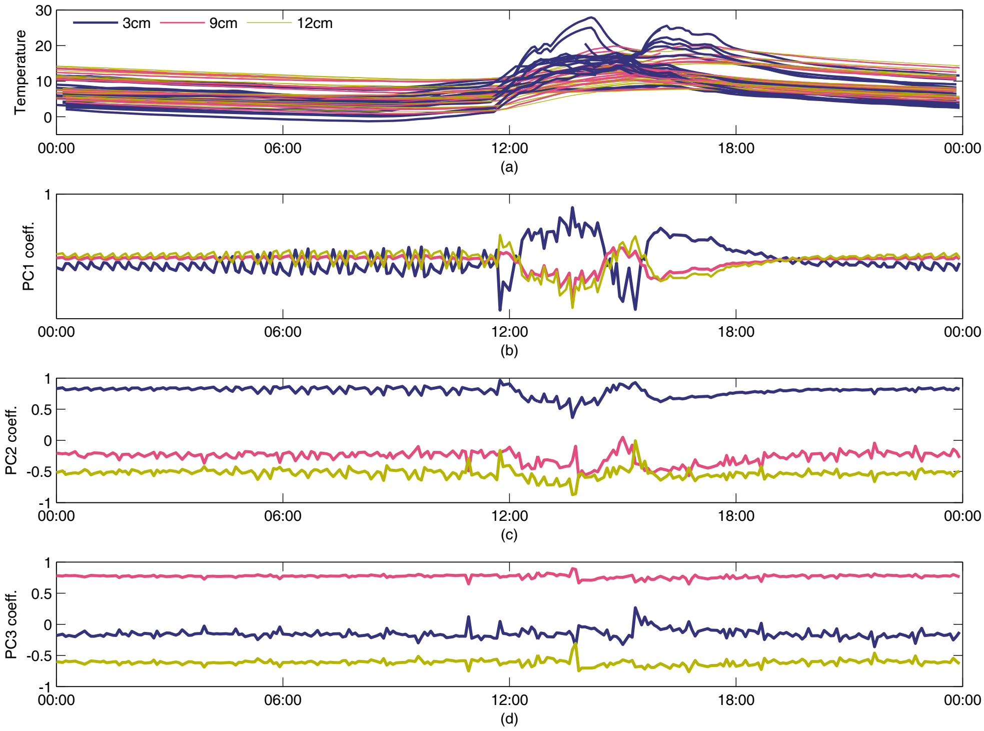


Figure 12: (a) Daily temperatures registered by the 3 cm, 9 cm and 12 cm deep sensors located at CEDEX pavement test track (Madrid, Spain) in 2007, (b)-(d) vectors of weight functions $a_1(t)$, $a_2(t)$ and $a_3(t)$ for the first, second and third principal component functions, respectively.

a good summary of the three temperatures since the mean of the local proportion of explained variability $\hat{\pi}_{11}$ is 98,65% and the integrated proportion of explained variability $\hat{\pi}_{21}$ is 99,10%. Therefore the second and third principal components displayed in Figure 12(c) and (d) are of much less importance. In Figure 13 we show the first principal component trajectories that could be used to substitute the three daily temperatures in any explicative model. Comparing with the mean functions displayed in the same figure, the principal component functions take into account the relative importance of each temperature along the day.

To complete the analysis, we analyze the impact in the principal component analysis of a previous smoothing of the curves with cubic smoothing splines and the use of different sizes of the grid. Suitable smoothing parameters range between 0.7 and 1 (the latter case is the case of the interpolating cubic spline), otherwise the functions are oversmoothed and none of the results that we present are significantly affected by this parameter. For long periods of time without recorded temperatures we use the information provided by the rest of the curve (this happen in 7 of the 21 cases) to carry out the smoothing. For every curve the knots where the points were the information was recorded. As the smoothed functions must be evaluated on a fine grid, we prove

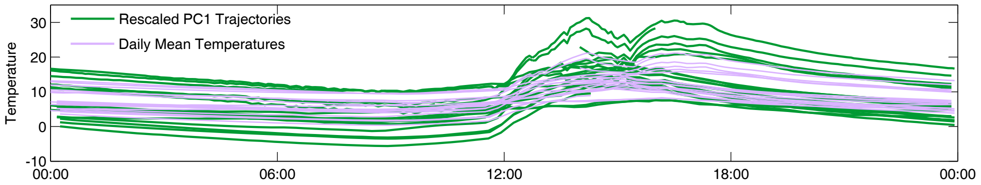


Figure 13: Rescaled first principal component trajectories and daily mean temperatures for the data set Daily temperatures registered by the 3 cm, 9 cm and 12 cm deep sensors located at CEDEX pavement test track (Madrid, Spain)

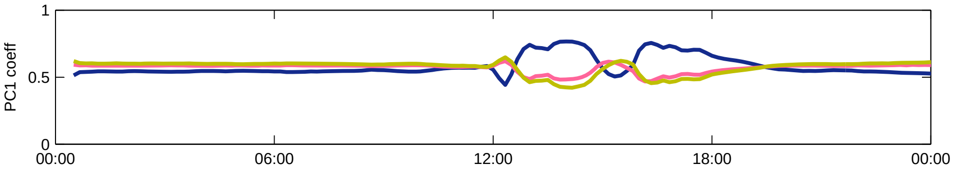


Figure 14: Vector of weight functions $a_1(t) = (a_{1,3cm}(t), a_{1,9cm}(t), a_{1,12cm}(t))$ for the first principal component, when the functional data are previously smoothed.

the sensitivity of our method to the grid size considering different options: every 1, 2, 5, 10 and 20 minutes. In all the cases, we obtain very similar weight functions for the principal components as those obtained with the raw data. Figure 14 shows the weights functions $a_1(t) = (a_{1,3cm}(t), a_{1,9cm}(t), a_{1,12cm}(t))$ for the first principal component with a grid consisting of a point every 1 minute. The weight functions inherit the smoothness from the original curves. The percentage of explained variability by the first principal component $\hat{\pi}_{11}$ ranges from 97.84% to 98.13%.

## 5.4 Real data: The Gait Data Set

The gait data set was introduced by Ramsay and Silverman (2005) to illustrate the multivariate PCA. The data functions are the simultaneous variation of the hip and knee angles for 39 children at 20 equally spaced time points. Figure 15 (a) shows that the first principal component calculated with our MFPCA method is a weighted mean of the two angles and the weight functions give more importance to hip or knee in different periods of the gait cycle. The percentages of explained variability by the first principal component are $\hat{\pi}_{11} = 75.94\%$ and $\hat{\pi}_{21} = 76.16\%$.

The comparison of both methods is not easy since the final result of the Ramsay and Silverman (2005) extension of PCA to multivariate functional data is a projection in a finite dimensional space. An ad-hoc use of the weights of these principal components has been used by Aneiros-Pérez *et al.* (2004) and Sangalli *et al.* (2010) to obtain functions that can be interpreted as principal component functions. The weights on each instant $t$

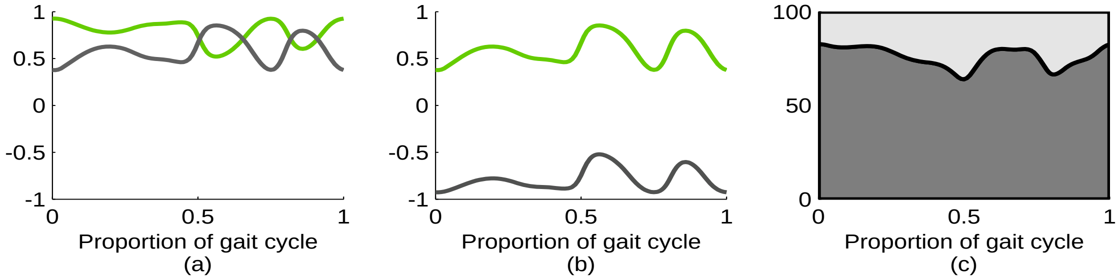


Figure 15: (a) First principal component weight functions, (b) second principal component weight functions, and (c) local percentage of explained variability by the first (dark area) and second (light area) principal components. Light/green lines are the weight functions for hips and the dark/grey lines are for knees.

are the corresponding to the non concatenated functions. When this approach is applied to the gait data set the results are now comparable with the ones from our MFPCA. This approach fails in the attempt of dimension reduction since the departure space is two–dimensional and four principal components are necessary to explain 88% of the total variability.

# 6 Conclusions and Final Remarks

A principal component method is suggested for dimension reduction applied to multivariate functional data (MFPCA). The problem of computing principal components when we observe a vector of $p$ functions in a set of $n$ individuals was considered previously in the monograph by Ramsay and Silverman (2005). They proposed the extension of FPCA to deal with multivariate functional data. In their approach, the functions corresponding to each observation are concatenated into a single long function (even in cases in which the functions correspond to the observation of different phenomenons) and then a FPCA is carried out using the concatenated functions. The final result is a projection in a finite dimensional space. Our method provides a reduced set of functions that accounts for most of the total variation. Accordingly, the maximum number of different eigenvalue-eigenfunction pairs in our method is the number of curves in each observation, $p$, whereas if we adapt FPCA for multivariate observations, this number is infinity. ~~Although useful and sensible, the approach of Ramsay and Silverman (2005) may have some undesirable effects depending on the final use we give to the results. For instance, if the variability over an interval of the domain is substantially greater than in others, the first principal component only takes into account this interval and neglects its complementary. Then we may not be able to summarize the curves in some parts of their domain without considering an intractable number of principal components.~~

~~Moreover, as was pointed out with the gait data set, even using the ad-hoc approach to derive principal components functions from the concatenation proposed in Ramsay and Silverman (2005), the dimension reduction is not always successful and sometimes to summarize we need a number of functions larger than the number of original curves.~~

The basic steps to be carried out for the principal component methods are the same, whether the data are multivariate, functional or multivariate functional. We transform the functional eigenanalysis problem in an equivalent matrix eigenanalysis problem. Differences arise both in the complexity of the correlation structure we are willing to consider, and also in the relevance given to the smoothness of the solutions. If the observed trajectories $X(t)$ are smooth, it would be reassuring that the corresponding principal components are also smooth. Considering that the number of possible weighting functions $a_r(t)$ defining each principal component is infinite, we have chosen the sign of $a_r(t)$ in a way that give easily interpretable results. Still, interpreting the components in MFPCA is not always straightforward as it happens in other PCA problems. We have considered some techniques that may aid in the interpretation of the results.

In high-dimensional problems, finding the principal components could be a computationally expensive problem. The most popular methods for calculating eigenvalues and eigenvectors in high dimension are iterative (see Golub *et al.*, 1996). Our method can be easily adapted to these iterative computations since it would be possible to use the results at $t$ as the initial values for $t+1$. Thus the computing time would be drastically reduced.

Dimension reduction is potentially helpful in a regression setup in which there is a scalar response variable $Y$ and a large number $p$ of functional regressors. The method described in this paper could be used to fit a model with only one functional regressor (namely, the first principal component) instead of $p$. Nevertheless, partial least squares (PLS), being a dimension reduction technique which takes into account the relationship between the response and the predictor variables, is an appealing alternative in this context. The first PLS algorithm was introduced by Wold (1966) and, more recently, several authors have proposed versions of PLS suitable for regression or classification models with functional predictors (see Preda and Saporta (2005) or Escabias *et al.* (2007) and references therein). The basic ideas introduced in this paper could also be applied to define multivariate functional partial least squares (MFPLS). The first PLS component would be given by a linear combination of the $p$ regressors, $T_\alpha(t) = \alpha'(t)X(t)$, such that the weight function $\alpha : [c,d] \to I\!R^p$ maximizes

$$\int_c^d [\alpha'(t)\Sigma(t)\alpha(t)]\,\mathrm{Corr}^2[Y, T_\alpha(t)]dt,$$

subject to $\|\alpha(t)\| = 1$, for each $t \in [c,d]$. If we compare the PLS criterion above with the PCA criterion proposed in equation (2.1), we see that PLS gives more weight to

$\mathrm{Var}[T_\alpha(t)] = \alpha'(t)\Sigma(t)\alpha(t)$ for those values of $t$ such that there is a high correlation between $T_\alpha(t)$ and the response variable Y. The obtention of more PLS components would require to solve a similar problem with additional orthogonality constraints [see equation (3.64) in Hastie *et al.* (2009)]. Practical implementation of the method would require to make a careful choice among the alternative algorithms for computing PLS components proposed in the literature. Criteria for choosing a unique weighting function $\alpha(t)$, similar to those proposed in this paper, would also be needed in this case.

# Acknowledgements

We are grateful to Julio Rodríguez for the fruitful discussions we have had on the ideas developed in this article, and Javier Pérez for providing us with data from the accelerated pavement testing facility at the Spanish Center for Road Transport Studies (CEDEX), that motivated this article. José Ramón Berrendero and Ana Justel have been partially supported by the Spanish Ministry of Science and Innovation, grant MTM2007-66632. Ana Justel has been also partially supported by Comunidad Autónoma de Madrid, grant CCG07-UAM/ESP-1761.

# Appendix

**Proof of Proposition 2**: (a)Define the function $G(t, \lambda) := \det(\Sigma(t) - \lambda I_p)$, where $\det(A)$ stands for the determinant of a matrix $A$. Any eigenvalue $\lambda_r(t)$ of $\Sigma(t)$ satisfies $G(t, \lambda_r(t)) = 0$. By the Implicit Function Theorem, $\lambda_r(t)$ is differentiable at $t^*$ if $G(t, \lambda)$ is differentiable at $(t^*, \lambda(t^*))$ and $G_\lambda(t^*, \lambda_r(t^*)) \neq 0$ (subscripts are used for partial derivatives). Moreover, in this case we have:

$$\dot{\lambda}_r(t^*) = -\frac{G_t(t^*, \lambda_r(t^*))}{G_\lambda(t^*, \lambda_r(t^*))}. \tag{6.11}$$

From the main result and equation (1) in Golberg (1972), under our assumptions it holds that $G(t, \lambda)$ is differentiable. Moreover, its partial derivatives are given by

$$G_t(t^*, \lambda_r(t^*)) = \mathrm{tr}[\mathrm{adj}(\Sigma(t^*) - \lambda_r(t^*) I_p)\dot{\Sigma}(t^*)]$$

and

$$G_\lambda(t^*, \lambda_r(t^*)) = -\mathrm{tr}[\mathrm{adj}(\Sigma(t^*) - \lambda_r(t^*) I_p)],$$

where $\mathrm{tr}(A)$ and $\mathrm{adj}(A)$ are the trace and the adjoint of $A$ respectively. It is not difficult to prove that

$$\mathrm{tr}[\mathrm{adj}(\Sigma(t^*) - \lambda_r(t^*) I_p)] = \prod_{\ell \neq r} [\lambda_\ell(t^*) - \lambda_r(t^*)] \neq 0,$$

since we assume the eigenvalues have multiplicity 1. Therefore, $G_\lambda(t^*, \lambda_r(t^*)) \neq 0$ and $\lambda_r(t)$ is differentiable at $t^*$. On the other hand, it can also be shown that

$$\operatorname{tr}[\operatorname{adj}(\Sigma(t^*) - \lambda_r(t^*)I_p)\dot{\Sigma}(t^*)] = \prod_{\ell \neq r}[\lambda_\ell(t^*) - \lambda_r(t^*)]a_r(t^*)'\dot{\Sigma}(t^*)a_r(t^*).$$

Then, from (6.11) and the last four displayed equations, we deduce $\dot{\lambda}_r(t^*) = a_r(t^*)'\dot{\Sigma}(t^*)a_r(t^*)$.

(b) From Theorem 8 in Lax (1997), p. 102, it is possible to choose $a_r(t)$ so that it is differentiable. We are going to show that the formula of the derivatives is given by (2.3). First, we differentiate the equation $\Sigma(t^*)a_r(t^*) = \lambda_r(t^*)a_r(t^*)$ and get

$$\dot{\Sigma}(t^*)a_r(t^*) + \Sigma(t^*)\dot{a}_r(t^*) = \dot{\lambda}_r(t^*)a_r(t^*) + \lambda_r(t^*)\dot{a}_r(t^*).$$

Rearranging terms:

$$[\Sigma(t^*) - \lambda_r(t^*)I_p]\dot{a}_r(t^*) = \dot{\lambda}_r(t^*)a_r(t^*) - \dot{\Sigma}(t^*)a_r(t^*). \tag{6.12}$$

Notice that

$$\Sigma(t^*) - \lambda_r(t^*)I_p = \sum_{\ell \neq r}[\lambda_\ell(t^*) - \lambda_r(t^*)]a_\ell(t^*)a_\ell(t^*)',$$

and define $M := \sum_{\ell \neq r}[\lambda_\ell(t^*) - \lambda_r(t^*)]^{-1}a_\ell(t^*)a_\ell(t^*)'$. Observe that

$$M[\Sigma(t^*) - \lambda_r(t^*)I_p]\dot{a}_r(t^*) = \sum_{\ell \neq r} a_\ell(t^*)a_\ell(t^*)'\dot{a}_r(t^*) = \dot{a}_r(t^*). \tag{6.13}$$

In the last equality we are using that $\|a(t^*)\| = 1$ implies that $\dot{a}(t^*)$ is orthogonal to $a(t^*)$ and therefore $\dot{a}(t^*)$ belongs to the subspace spanned by $\{a_\ell(t^*) : \ell \neq r\}$. From (6.13), if we pre-multiply both terms of (6.12) by $M$,

$$\begin{aligned}
\dot{a}_r(t^*) &= \dot{\lambda}_r(t^*)Ma_r(t^*) - M\dot{\Sigma}(t^*)a_r(t^*) \\
&= -\left[\sum_{\ell \neq r}[\lambda_\ell(t^*) - \lambda_r(t^*)]^{-1}a_\ell(t^*)a_\ell(t^*)'\right]\dot{\Sigma}(t^*)a_r(t^*),
\end{aligned}$$

since $Ma_r(t^*) = 0$ because $a_\ell(t^*)'a_r(t^*) = 0$ for $l \neq r$.

# References

Acker, A.F., 1974. Absolute continuity of eigenvectors of time-varying operators. Proceedings of the American Mathematical Society 42, 198–201.

Aneiros Pérez, G., Cardot, H., Estévez Pérez, G., Vieu, P., 2006. Maximum ozone concentration forecasting by functional nonparametric approaches. Environmetrics 15, 675–685.

Aneiros Pérez, G., Vieu, P., 2006. Semi-functional partial linear regression. Statist. Probab. Lett. 11, 1102–1110.

Boente, G., Fraiman, R. 2000., Kernel-based functional principal components. Statist. Probab. Lett. 48, 335–345.

Coetzee, N.F., Nokes, W., Monismith, C., Metcalf, J., Mahoney, J., 2000. Full-scale/accelerated pavement testing: current status and future directions. Millennium Paper Series.

Delicado, P., 2001. Another look at principal curves and surfaces. Journal of Multivariate Analysis 77, 84–116.

Delicado, P., 2011. Dimensionality reduction when data are density functions. Comput. Statist. Data Anal. 55, 401–420.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2007. Functional PLS logit regression model. Comput. Statist. Data Anal. 51, 4891–4902.

Ferraty, F. (2010). Special Issue : Statistical Methods and Problems in Infinite-dimensional Spaces (Ed.). J. Mult. Anal., 101, 305-490.

Ferraty, F., Romain, Y., 2011. The Oxford Handbook of Functional Data Analysis. Oxford University Press.

Ferraty, F., Vieu, P., 2006. Non Parametric Functional Data Analysis. Theory and Practice. Springer.

Ferraty, F., Vieu, P., 2009. Additive prediction and boosting for functional data. Comput. Statist. Data Anal. 53, 1400–1413.

Friedman, J.H., Tukey, J.W., 1974. A Projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers C-23, 881–890.

Golberg, M.A., 1972. The derivative of a determinant. American Mathematical Monthly 79, 1124–1126.

Golub, G.H., Van Loan, C.F., 1996. Matrix Computations, 3er ed. Johns Hopkins University Press, Baltimore.

González-Manteiga, W., Vieu, P., 2007. Introduction to the special issue on statistics for functional data. Comput. Statist. Data Anal. 51, 4788–4792.

Hastie, T., Stuetzle, W., 1989. Principal curves. J. Amer. Statist. Assoc. 84, 502–516.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd ed. Springer Series in Statistics. Springer, New York.

Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Networks 13, 411–430.

James, G., Hastie, T., Sugar, C., 2000. A Principal component models for sparse functional data. Biometrika 87, 587–602.

Lax, P.D., 1997. Linear Algebra. Wiley, New York.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L., 1999. Robust principal component analysis for functional data. Test 8, 1–73.

Manté, C., Yao, A.-F., Degiovanni, C., 2007. Principal component analysis of measures, with special emphasis on grain-size curves. Comput. Statist. Data Anal. 51, 4969–4983.

Müller, H.G., 2005. Functional modelling and classification of longitudinal data. Scandinavian Journal of Statistics 23, 223–240.

Park, J., Gasser, T., Rousson, V., 2009. Structural components in functional data. Comput. Statist. Data Anal. 53, 3452–3465.

Preda, C., Saporta, G., 2005. PLS regression on a stochastic process. Comput. Statist. Data Anal. 48, 149–158.

Ramsay, J., Dalzell, C.J., 1991. Some tools for functional data analysis. Journal of the Royal Statistics Society B 53, 539–572.

Ramsay, J., Silverman, B.W., 2002. Applied Functional Data Analysis: Methods and Case Studies. Springer, New York.

Ramsay, J., Silverman, B.W., 2005. Functional Data Analysis, 2nd ed. Springer, New York.

Rice, J., Silverman, B.W., 1991. Estimating the mean and the covariance structure nonparametrically when the data are curves. Journal of the Royal Statistics Society B 53, 233–243.

Sangalli, L.M., Secchi, P., Vantini, S., Vitelli, V., 2010. Joint clustering and alignment of functional data: an application to vascular geometries. MOX Report No. 09/2010, (`http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/09-2010.pdf`).

Schumaker, L., 2007. Spline Functions: Basic Theory, 3rd ed. Cambridge Mathematical Library. Cambridge University Press, Cambridge.

Shi, J., Wang, B., Murray-Smith, R., Titterington, D., 2007. Gaussian process functional regresson modelling for batch data. Biometrics 63, 714–723.

Shin, H., 2009. Partial functional linear regression. Journal of Statistical Planning and Inference 139, 3405–3418.

Silverman, B.W., 1996. Smoothed functional principal components analysis by the choice of norm. Annals of Statistics 24, 1–24.

Valderrama, M., 2007. Introduction to the special issue on modelling functional data in practice. Comput. Statist. 22, 331–334.

van der Linde, A., 2008. Variational Bayesian functional PCA. Comput. Statist. Data Anal. 53, 517–533.

Wang, S., Jank, W., Shmueli, G., 2008. Explaining and forecasting online auction prices and their dynamics using functional data analysis. Journal of Business and Economic Statistics 26, 144–160.

Wold, H., 1966. Estimation of principal components and related models by iterative least squares. In Multivariate Analysis, ed. Krishnaiah, P.R., 391–420. Academic Press. New York.