



UNIVERSITÀ DEL PIEMONTE ORIENTALE

DIPARTIMENTO DI STUDI  
PER L'ECONOMIA E L'IMPRESA  
SETTORE AMMINISTRAZIONE

## Quaderni DISEI

**Enea Giuseppe Bongiorno, Aldo Goia**

***Classification methods for Hilbert data  
based on surrogate density***

*Giugno 2015*

**Quaderno n. 2/2015**

# Classification methods for Hilbert data based on surrogate density

Enea G. Bongiorno, Aldo Goia  
Università del Piemonte Orientale,  
enea.bongiorno@ aldo.goia@uniupo.it

June 4, 2015

## Abstract

We study an unsupervised and a supervised classification approaches for Hilbert random curves. Both rest on the use of a surrogate of the probability density which is defined, in a distribution-free mixture context, from an asymptotic factorization of the small-ball probability. That surrogate density is estimated by a kernel approach from the principal components of the data. The paper focuses on the illustration of the classification algorithms and the computational implications, with particular attention to the tuning of parameters involved. Some asymptotic results are sketched. Applications on simulated and real datasets show how the proposed methods work.

**Keywords.** density based clustering; discriminant Bayes rule; Hilbert data; small-ball probability mixture; functional principal component; kernel density estimate.

## Introduction

In multivariate classification problems, whether they are supervised or unsupervised, the role of joint density, or better, of its estimate, plays a central role. In order to make it clear, one has just to think to the literature about the model based clustering approaches, and recall that all the discriminant methods resting on the so-called Bayes rule require an estimation of the joint density in each group (for recent developments, see for instance [14]).

When one deals with data belonging to functional spaces (for a general introduction on this topic, one can refer to monographs [11], [17] and [24], and the recent book [3]), the dimensionality problem arises immediately, and, as a consequence, a probability density function generally, does not exist (see [4]). Hence, a direct extension of density oriented classical multivariate classification approaches to functional data can not be implemented: usually a reduction of dimensionality, based on projection over suitable finite subspaces, is a preliminary step put in place to tackle the problem. This route is followed for instance by [20], where model based clustering methods are proposed, and by [19] and [25] in defining

a discriminant approach: the general idea is to put a suitable density mixture model over the coefficients of the representation of functional data in the finite subspace, admitting that such model may summarize the distribution of underlying process. It is worthy to note that also other reduction dimensionality approaches are possible: for instance, one can recall the techniques, which rest on the most important points, that are illustrated in [5].

Another way to proceed, that aims to work directly on the distribution of the process, refers to the concept of surrogate (or pseudo) density. The general principle, dates back to [13], is to factorize the small ball probability associated with the functional data, when the radius of the ball tends to zero, as product of two terms: an “intensity term” which depends only on the center of the ball, and a kind of “volume parameter” which depends only on the radius. Since the first term reflects the latent structure of the distribution of the underlying process, it represents an ideal candidate to play the role that the multivariate density has in the finite dimensional classification methods. Theoretical conditions that allow such factorization when one considers Hilbert functional data in the space determined by the basis of the Karhunen–Loève decomposition (i.e. by the eigenfunctions of the principal components analysis), and the structure of the pseudo-density (linked to the principal components, i.e. the coefficients of the decomposition), are discussed in [4] and then in [2], where some assumptions are relaxed. One can observe that such idea allows to see the projective approaches in a more general theoretical context.

To put into effect the above factorization and take advantage of the pseudo-density, different ways are possible. A first one is to specify a suitable density model mixture for the principal components: this full parametric approach is followed by [18] in defining a Gaussian mixture clustering procedure. On the other hand, a full nonparametric approach is possible as done in [12] where a k-NN procedure is proposed to estimate the pseudo-density.

In this paper we consider an intermediate approach to evaluate the pseudo-density: after computing the first  $d$  principal components of the functional data, we obtain an estimate of their joint density  $f_d$  via the classical Parzen–Rosenblatt kernel method. We can see this approach as semi-parametric: if on one hand, we use coefficients of the Karhunen–Loève decomposition in defining the pseudo-density (it is not estimated directly), on the other hand, the mixture model is not specified.

The goal of this paper is to clarify how to use the proposed method to tackle classification problems for Hilbert data: we illustrate both a pseudo-density oriented clustering algorithm resting on the definition of clusters as a high intensity regions from the modes of  $f_d$ , and a classifier based on the Bayes rule in the discriminant context, under suitable hypothesis on the distributions of the mixture components. After introducing the general mixture model and the theoretical motivations, the algorithms are illustrated in details, focusing on computational aspects and on the tuning of different parameters involved (as, for instance, the number of considered principal components, the scale of the bandwidth matrix in estimating the joint density  $f_d$  and a “mode cap” parameter whose purpose is to prevent the springing of too much spurious modes). The study is completed with an analysis of real and synthetic datasets: a special attention is paid, both in the clustering and in the discriminant framework, to mixtures presenting non-spherical clusters.

This paper is organised as follows. In Section 1 we introduce the mixture model and its factorization, Section 2 is devoted to illustrate the classification algorithms (clustering and discriminant) and to discuss the parameters used. Finally, Section 3 collects the applications to simulated data and real cases.

## 1 Theoretical framework

Let  $X$  be a random element defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and taking its values in the Hilbert space  $\mathcal{L}_{[0,1]}^2$  endowed with the usual inner product  $\langle \cdot, \cdot \rangle$  and associated norm  $\| \cdot \|$ . Denote by

$$\mu_X = \{ \mathbb{E}[X(t)], t \in [0, 1] \}, \quad \text{and} \quad \Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot \rangle (X - \mu_X)]$$

the mean function and covariance operator of  $X$  respectively. A measure of concentration of  $X$  is given by the small-ball probability, briefly SmBP (see [11] and reference therein), defined as

$$\varphi(x, \varepsilon) = \mathbb{P}(\|X - x\| < \varepsilon), \quad x \in \mathcal{L}_{[0,1]}^2, \quad \varepsilon > 0.$$

Suppose that  $\Omega$  is partitioned in  $G$  (unknown and finite) sub-sets  $\Omega_g$  and let  $Y$  be a  $\mathbb{N}$ -valued random variable defined by

$$Y(\omega) = \sum_{g=1}^G g \mathbb{I}_{\Omega_g}(\omega), \quad \mathbb{P}(Y = g) = \pi_g > 0, \quad \sum_{g=1}^G \pi_g = 1,$$

with  $\mathbb{I}_A$  the indicator of  $A$  and consider the conditioned SmBP

$$\varphi(x, \varepsilon|g) = \mathbb{P}(\|X - x\| < \varepsilon \mid Y = g), \quad g = 1, \dots, G,$$

that leads to the mixture

$$\varphi(x, \varepsilon) = \sum_{g=1}^G \pi_g \varphi(x, \varepsilon|g), \quad x \in \mathcal{L}_{[0,1]}^2, \quad \varepsilon > 0. \quad (1)$$

The latter expression is the starting point to approach model-based classification problems: when  $Y$  is a latent variable we deal with an unsupervised classification problem focused on the left-hand side of (1), see Section 2.1; whereas when  $Y$  is observed the model drives to the construction of a Bayesian classifier whose starting point is the right-hand side of (1), see Section 2.2. In both cases, instead of tackle directly (1), we want to simplify it exploiting an approximation result sketched below (for more details see [2]). For the sake of simplicity it is presented with respect to the process  $X$ ; however the same arguments can be applied to  $(X|Y = g)$  with  $g = 1, \dots, G$ , provided a suitable change of notation.

Consider the Karhunen-Loève expansion of  $X$ : denoting by  $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$  the decreasing to zero sequence of non-negative eigenvalues and their associated orthonormal eigenfunctions of  $\Sigma$ , it holds

$$X(t) = \mu_X(t) + \sum_{j=1}^{\infty} \theta_j \xi_j(t), \quad 0 \leq t \leq 1,$$

where  $\theta_j = \langle X - \mu_X, \xi_j \rangle$  are the so-called principal components (PCs) of  $X$  satisfying

$$\mathbb{E}[\theta_j] = 0, \quad \text{Var}(\theta_j) = \lambda_j, \quad \mathbb{E}[\theta_j \theta_{j'}] = 0, \quad j \neq j'.$$

From now on, without loss of generality suppose that  $\mu_X = 0$ . Moreover, assume:

(A.1) the first  $d$  PCs  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$  admit a strictly positive and sufficiently smooth joint probability density  $f_d$ ;

(A.2) there exists a positive constant  $C$  (not depending on  $d$ ) for which

$$\sup_{d \in \mathbb{N}} \sup_{i, j \in \{1, \dots, d\}} \frac{\sqrt{\lambda_i \lambda_j}}{|f_d(\boldsymbol{\vartheta})|} \left| \frac{\partial^2 f_d(\boldsymbol{\vartheta})}{\partial \vartheta_i \partial \vartheta_j} \right| \leq C, \quad \text{for any } \boldsymbol{\vartheta} \in D,$$

where  $D = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j \leq d} (\vartheta_j - x_j)^2 \leq \rho^2 \right\}$  for some  $\rho \geq \varepsilon$ ;

(A.3)  $x$  is an element of  $\mathcal{L}_{[0,1]}^2$  such that  $\sup\{x_j^2/\lambda_j : j \geq 1\} < \infty$ , with  $x_j = \langle x, \xi_j \rangle$ ;

(A.4) the spectrum of  $\Sigma$  is rather concentrate:  $\{\lambda_j\}_{j=1}^\infty$  decays to zero at least exponentially.

**Proposition 1** *Under (A.1)–(A.4), as  $\varepsilon$  tends to zero, it is possible to choose  $d(\varepsilon)$  diverging to infinity so that:*

$$\varphi(x, \varepsilon) \sim f_d(x_1, \dots, x_d) \phi(d, \varepsilon). \quad (2)$$

The form of  $\phi(d, \varepsilon)$  depends on the eigenvalues decay; in particular:

- if  $\{\lambda_j\}_{j=1}^\infty$  decays exponentially

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j < C, \quad \text{for any } d \in \mathbb{N} \quad (3)$$

then

$$\phi(d, \varepsilon) = \exp \left\{ \frac{1}{2} d [\log(2\pi e \varepsilon^2) - \log(d) + \delta(d, \alpha)] \right\},$$

where  $\delta(\cdot, \cdot)$  is such that  $\lim_{\alpha \rightarrow \infty} \limsup_{s \rightarrow \infty} \delta(s, \alpha) = 0$  and  $\alpha$  is a parameter chosen so that  $\lambda_d^{-1} \varepsilon^2 \leq \alpha^2$ ;

- if  $\{\lambda_j\}_{j=1}^\infty$  decays super-exponentially

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j \rightarrow 0, \quad \text{as } d \rightarrow \infty \quad (4)$$

or, equivalently,  $\lambda_{d+1}/\lambda_d \rightarrow 0$  (as  $d \rightarrow \infty$ ), then

$$\phi(d, \varepsilon) = \exp \left\{ \frac{1}{2} d [\log(2\pi e \varepsilon^2) - \log(d) + \delta(d)] \right\},$$

where  $\delta(d) = o(1)$  as  $d \rightarrow \infty$ ;

- if  $\{\lambda_j\}_{j=1}^\infty$  decays hyper-exponentially

$$d \left( \sum_{j \geq d+1} \lambda_j \right) \left( \sum_{j \leq d} \frac{1}{\lambda_j} \right) = o(1), \quad \text{as } d \rightarrow \infty \quad (5)$$

then

$$\phi(d, \varepsilon) = \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2 + 1)}.$$

Note that (5)  $\Rightarrow$  (4)  $\Rightarrow$  (3); moreover, in the hyper-exponential case,  $\phi(d, \varepsilon)$  is the volume of a  $d$ -dimensional ball with radius  $\varepsilon$ . This justifies to interpret, in Equation (2),  $\phi(d, \varepsilon)$  as a  $d$ -dimensional volume parameter whilst  $f_d$ , being the only factor depending on  $x \in \mathcal{L}_{[0,1]}^2$ , as a surrogate of the density of the Hilbert process.

**Remark 2** *Although results are exposed in the Karhunen–Loève (or PCA) basis, they can be carried out with any orthonormal basis ordered according to the decreasing values of variances of the projections, provided that they decay sufficiently fast (see [2]). In this view, PCA basis is the optimal one since by construction presents the faster decays of such variances.*

## 2 Classification

This section is devoted in defining classification procedures in a functional framework that take advantage of the asymptotic factorization results provided by Proposition 1. It is divided in two subsections each one: the first one illustrates a clustering algorithm whereas the second one deal on a discriminant analysis procedure.

### 2.1 Unsupervised classification

Consider the sample  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$  drawn from  $(X, Y)$  defined as above, with  $X$ 's are observed while the group variables  $Y$ 's are latent. The aim is to determine the range of  $Y$  (i.e.  $G$ ) and, for each observed  $X_i$  the membership group (that is the value of  $Y_i$ ). If the distribution of  $(X|Y = g)$  was specified then a full parametric approach may apply; this has been done, for example, in [18] where authors used a maximum likelihood and expectation maximization approach to identify the distribution parameters of a Gaussian mixture assumed for  $f_d$ . On the other hand, if no information are available, a distribution free model could be used. In this latter view, consider the SmBP mixture (1) and apply Proposition 1 to its left-hand side to obtain:

$$\sum_{g=1}^G \pi_g \varphi(x, \varepsilon | g) = \varphi(x, \varepsilon) \sim f_d(x_1, \dots, x_d) \phi(d, \varepsilon), \quad x \in \mathcal{L}_{[0,1]}^2, \quad \varepsilon \rightarrow 0.$$

Such expression highlights how the surrogate density  $f_d$  carries the information on the mixture and, at the same time, endorses a “density oriented” clustering approach on  $f_d$  as a

fruitful tool in detecting the latent structure. Our proposal is to identify the groups by a “locally high (surrogate) density regions” principle: the clustering algorithm computes the estimates  $\widehat{m}_{d,g}$  of the modes  $m_{d,g}$  of  $f_d$  (whose number estimates  $G$ ), for each  $g$ , it finds the largest connected upper–surface containing only  $\widehat{m}_{d,g}$ , and hence, it assigns group labels to each observation consistently with its proximity to these sets (cf. [2] and references therein). The algorithm procedure is illustrated below:

1. Obtain an estimate of the covariance operator and of eigenelements.
2. Fix  $d$ , compute  $\widehat{f}_{d,n}$  (an estimation of the joint distribution density  $f_d$ ).
3. Look for its local maxima  $\widehat{m}_{d,g}$ ,  $g = 1, \dots, \widehat{G}$ .
4. *Finding Prototypes*: for each  $g$  in  $\{1, \dots, \widehat{G}\}$ , the  $g$ -th “prototypes” group is formed by those  $X_i$  whose estimated PCs  $(\widehat{\theta}_{1,i}, \dots, \widehat{\theta}_{d,i})$  belong to the largest connected upper–surface of  $\widehat{f}_{d,n}$  that contains only the maximum  $\widehat{m}_{d,g}$ . In other words, for such individual,  $\widehat{Y}_i = g$ .
5. Assign each unlabelled  $X_i$  to a group by means of a  $k$ -NN procedure (with  $k = 1$ ).

As a by–product of such algorithm, it is possible to define a center for each cluster through the “ $d$ -dimensional modal curves” built from  $\widehat{m}_{d,g}$  and defined as follows:

$$\widehat{X}_g^m(t) = \sum_{j=1}^d \widehat{m}_{d,g}^{(j)} \widehat{\xi}_j(t)$$

where  $\widehat{m}_{d,g}^{(j)}$  is the  $j$ -th term of  $\widehat{m}_{d,g}$  and  $\widehat{\xi}_j(t)$  are the empirical versions of  $\xi_j(t)$ . The remain part of this section discusses some theoretical and practical aspects on the algorithm.

### 2.1.1 Surrogate density estimation

In order to estimate  $f_d$ , we consider the classical multivariate kernel density estimator:

$$\widehat{f}_{d,n}(\widehat{\Pi}_d x) = \frac{1}{n} \sum_{i=1}^n K_H \left( \left\| \widehat{\Pi}_d (X_i - x) \right\| \right), \quad \widehat{\Pi}_d x \in \mathbb{R}^d,$$

where  $K_H(\mathbf{u}) = \det(H)^{-1/2} K(H^{-1/2}\mathbf{u})$ ,  $K$  is a kernel function,  $H$  is a symmetric semi-definite positive  $d \times d$  matrix and,  $\widehat{\Pi}_d$  is the projection operator over the subspace spanned by the first  $d$  eigenfunctions of  $\widehat{\Sigma}_n$ , the sample version of  $\Sigma$ , so that kernel argument depends on the PCA semi–metric (see [11, Section 8.2]).

From a theoretical point of view, one may wonder if such estimator is consistent for  $f_d$  and, when this is the case, if it attains the same rate of convergence that holds when  $\Pi_d$  is known. A positive answer was provided in [2]; in particular, consider the special case  $H_n = h_n^2 I$ , and suppose that:

(B.1) the density  $f_d(x)$  is positive and  $p$  times differentiable at  $x \in \mathbb{R}^d$ ;

(B.2)  $h_n$  is such that  $h_n \rightarrow 0$  and  $nh_n^d/\log n \rightarrow \infty$  as  $n \rightarrow \infty$ ;

(B.3) the kernel  $K$  is Lipschitz, bounded, integrable density function with compact support  $[0, 1]$ ;

(B.4) the process  $X$  is bounded.

Then, the following result holds:

**Proposition 3** *Assume (B.1)–(B.4) with  $p > (3d + 2)/2$  and consider the optimal bandwidth*

$$c_1 n^{-\frac{1}{2p+d}} \leq h_n \leq c_2 n^{-\frac{1}{2p+d}}$$

where  $c_1$  and  $c_2$  are two positive constants. Thus, as  $n$  goes to infinity,

$$\mathbb{E} \left[ f_d(x) - \hat{f}_n(x) \right]^2 = O \left( n^{-2p/(2p+d)} \right),$$

uniformly in  $\mathbb{R}^d$ .

From a practical point of view, an important task is the bandwidth selection: our choice is to consider a diagonal bandwidth matrix (see [8] for a heuristic justification) whose non-null entries are the univariate bandwidth provided by [26, p.48]. Anyway, it is clear that larger is  $|H|$ , “smoother” is  $\hat{f}_{d,n}$ , smaller is the number of modes and hence the number of groups. In other words, different choices for the bandwidth may be considered in order to catch different phenomenon scales; this is done applying to  $H$  a scale factor  $\delta > 0$  whose optimality (in a sense to be specified) is discussed below.

### 2.1.2 Modes and Prototypes

The construction of prototypes rests on estimated modes. A problem concerns spurious modes caused by the choice of  $\delta$  and sampling variability. To tackle this issue we select only those modes  $\hat{m}_{d,g}$  for which  $\hat{f}_d(\hat{m}_{d,g})$  is maximum over the parallelepiped  $\hat{m}_{d,g} \pm [0, rh_1] \times \dots \times [0, rh_d]$ , where  $h_j$  is the resolution along the  $j$ -th direction of the grid used to estimate the density while  $r$  is a positive integer, playing the counterpart of a tolerance coefficient.

For what concerns the identification of the largest connected upper-level sets of  $\hat{f}_{d,n}$  related to each mode, we use the graphics visualization system of R software. Hence,  $X_i$  is assigned to  $g$ -th prototype group if its estimated PCs belong to the  $g$ -th upper-level sets by using the algorithm described in [22] and available in the R-package *ptinpoly*.

### 2.1.3 Tuning parameters

Given a data set, different values of  $d$ ,  $r$  and  $\delta$  may lead to different cluster results: thus, a criterion to chose them is necessary.

In view of Proposition 1, parameter  $d$  should be large enough to guarantee a good approximation for the SmBP (2), but small enough to avoid the well-known “curse of dimensionality” in estimating non-parametrically  $f_d$ . A good compromise is to choose  $d$  so that the Fraction of Explained Variance,  $FEV(d) = \sum_{j \leq d} \lambda_j / \sum_{j \geq 1} \lambda_j$ , is larger than a suitable constant. In practice,  $FEV$  is estimated from eigenvalues of  $\widehat{\Sigma}$ .

For what concerns a choice for  $(r, \delta)$  and to provide some insights on the quality of clustering solutions, we consider both external and internal criteria; in particular, we implement: the purity index (based on some prespecified structure, which is the reflection of prior information on the data), the “Caliński and Harabasz”, or briefly “CH”, index (that does not depend on external information) and, when feasible (to not dependent on a single rule) a combination of the two indices. Thus, accordingly to the nature of data, the idea is to look at those  $(r, \delta)$  which furnish the best value for these validation criteria. As a consequence and because of the algorithm definition, such a choice for  $(r, \delta)$  will automatically lead to the optimal number of clusters  $G$ . In the remaining part of this section, we summarise these two criteria; more details can be found in [29] and references therein.

Purity index measures how close is a clustering to an available pre-specified class structure and, more precisely, the extent to which each cluster consists of objects from a single class. In particular, for each cluster, consider the class distribution of the data; i.e. for class  $j$  compute  $p_{gj}$ , the frequency that a member of cluster  $g$  belongs to class  $j$  as  $p_{gj} = \pi_{gj} / \pi_g$ , where  $\pi_g$  is the proportion of objects in cluster  $g$  and  $\pi_{gj}$  is the proportion of objects of class  $j$  in cluster  $g$ . Hence, for each cluster  $g$ , purity is calculated as

$$p_g = p_g(r, \delta) = \max\{p_{gj} : j = 1, \dots, L\},$$

where  $L$  is the number of pre-specified classes; whilst the total purity is the sum of the cluster purities weighted by the size of each cluster  $p = \sum_{g=1}^G p_g \pi_g$ . Clearly,  $p$  ranges in  $[0, 1]$  with  $p = 0$  meaning maximum separation and  $p = 1$  maximum cohesion.

Among the internal validation indices, the  $CH$  index is known to achieve the best performance (see [7]). It is defined as

$$CH = CH(\delta, r) = \begin{cases} \frac{Tr(S_B)}{K-1} / \frac{Tr(S_W)}{n-K}, & K > 1, \\ 0, & K = 1. \end{cases}$$

where  $N$  is the sample size,  $K$  is the number of cluster obtained choosing the couple  $(\delta, r)$  and  $Tr(S_B)$  and  $Tr(S_W)$  are the traces of the estimated between and within covariance matrices, respectively. The couples  $(\delta, r)$  that maximize  $CH$  are selected as optimal and the number of clusters  $K$  is consequently obtained.

## 2.2 Supervised classification

In discriminant analysis, differently from clustering, the presence of  $G$  distinct groups is established and modelled by the observed variable  $Y$ : the aim is to label each new incoming observation according to this known groups structure. To do this, a typical approach is to

use a *Bayes classification rule*: given an observation  $x$ , one assigns it to the class  $\gamma(x) \in \{1, \dots, G\}$  to which correspond the highest posterior probability  $\mathbb{P}(Y = \gamma(x)|X = x)$ :

$$\gamma(x) = \arg \max_{g=1, \dots, G} \mathbb{P}(Y = g|X = x).$$

Equivalently,  $\gamma(x)$  is that index  $g'$  in  $\{1, \dots, G\}$  such that

$$\frac{\mathbb{P}(Y = g'|X = x)}{\mathbb{P}(Y = g|X = x)} > 1, \quad \text{for any } g = 1, \dots, G \text{ and } g \neq g'. \quad (6)$$

If a probability density of  $X$  in the  $g$ -th group  $f(x|g)$  were known (with  $f(x|g) > 0$ ), thanks to Bayes formula, Equation (6) simplifies as follows:

$$\frac{\pi_{g'} f(x|g')}{\pi_g f(x|g)} > 1, \quad \text{for any } g \neq g',$$

and, consequently, the classification rule becomes:

$$\gamma(x) = \arg \max_{g=1, \dots, G} \pi_g f(x|g).$$

It is clear that such arguments do not apply straightforwardly in a functional settings without further assumptions on the probability measures. A possible way to tackle the problem is to consider the following classification rule: assign a new functional observation  $x$  to the  $g$ -th group for which eventually, as  $\varepsilon$  tends to 0,

$$\frac{\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon)}{\mathbb{P}(Y = g' \mid \|X - x\| < \varepsilon)} > 1, \quad \text{for any } g' \neq g. \quad (7)$$

At a glance, it appears hard to use in practice. Anyway, thanks to Bayes formula, ratio in (7) becomes

$$\frac{\pi_g \varphi(x, \varepsilon|g)}{\pi_{g'} \varphi(x, \varepsilon|g')}.$$

Whenever assumptions of Proposition 1 hold for each  $(X|Y = g)$  with  $g = 1, \dots, G$ , then the above ratio further reduces to

$$\frac{\pi_g f_{d_g}(x|g) \phi(d_g, \varepsilon)}{\pi_{g'} f_{d_{g'}}(x|g') \phi(d_{g'}, \varepsilon)}, \quad \text{as } \varepsilon \rightarrow 0,$$

where  $f_{d_g}(x|g)$  is the joint density of the first  $d_g$  PCs computed using the (conditional) covariance operator  $\Sigma_g$  of the group  $g$ . Clearly, the asymptotic behaviour of the ratio in the right hand part is related to the trade-off between the volume parameters  $\phi(\cdot, \varepsilon)$  and the probability densities evaluated at possibly different dimensions  $d_g$  and  $d_{g'}$ .

The classification rule may be further simplified if additional assumption may be imposed on the mixture process  $X$ , since the spectrum decay of  $\Sigma$  controls those of each  $\Sigma_g$  (the conditional covariance operator corresponding for the  $g$ -th group), the starting point to

build  $f_{d_g}$ . In particular, consider the variance decomposition  $\Sigma = B + W$ , where  $B$  and  $W = \sum_{g \in G} \pi_g \Sigma_g$  represent the between and within covariance operator respectively. A straight application of the Courant–Fischer–Weyl min–max principle for linear operators leads to

$$\lambda_k(M_1) \leq \lambda_k(M_1 + M_2),$$

where  $\{\lambda_j(M_i)\}$  denote the eigenvalues of the linear operator  $M_i$  (decreasingly ordered). The latter inequality ensures that eigenvalues of  $B$ ,  $W$  and  $\Sigma_g$  ( $g = 1, \dots, G$ ) have a decay at least fast as the one of  $\Sigma$ . In other words, since eigenvalues decay is a measure of how much  $X$  is concentrate in the space, the process  $(X|Y = g)$  in each sub–population must be “concentrate” at least as  $X$ . As a consequence, if the spectrum of  $\Sigma$  decays exponentially (according to (3)), then  $d_g$  can be chosen equal to  $d$  for any  $g = 1, \dots, G$ ,  $\phi(d, \varepsilon)$  simplifies and one can write the classification rule (7) similarly to the multivariate case, replacing a probability density with a surrogate version: assign a new functional observation  $x$  to the  $g$ –th group for which eventually, as  $\varepsilon$  tends to 0,

$$\frac{\pi_g f_d(x|g)}{\pi_{g'} f_d(x|g')} > 1, \quad \text{for any } g' \in \{1, \dots, G\}, g' \neq g,$$

or, equivalently, as  $d$  tends to  $+\infty$ ,

$$\gamma(x, d) = \arg \max_{g=1, \dots, G} \pi_g f_d(x|g). \quad (8)$$

Operatively, if one could specify the conditional densities  $f_d(x|g)$ , a full parametric approach would be possible. Although  $\gamma(x, d)$  still depends of  $\varepsilon$  by means of  $d$  (see Proposition 1), it is not so restrictive to assume that

- (A.5)  $\gamma(x, d)$  is eventually constant as  $d$  explodes to infinity; i.e. there exists a positive integer  $d^*$  such that  $\gamma(x, d) = \gamma(x, d^*)$  for any  $d \geq d^*$ .

This assumption holds, at least, in the case of a finite dimensional process.

At this point, a comparison of our approach with the one introduced in [19] is interesting and one can trace some parallelism. Indeed, in both approaches, a dimensionality reduction step, based on projection onto a finite vector subspace generated by a priori chosen basis, is implemented and so the classification rule involves the conditional joint densities of the projection coefficients. Moreover, if  $d_g = d$  for all  $g$ , and one can assume an underlying Gaussian mixture model, both approaches lead to the same classifier: the present section justifies theoretically the use of (8) in a finite dimensional subspace. However, if eigenvalues of  $\Sigma$  decay slowly, we can not ensure that  $d_g$  is the same varying  $g$  and the volume terms  $\phi(d_g, \varepsilon)$  can not be neglected in the classification rule: the approach based on a pure projective method could be unfruitful.

The illustrated method can be framed between the full parametric discrimination described in [19] and the full nonparametric one proposed by [10], where the posterior probability is estimated directly by a kernel regression approach.

### 2.2.1 Estimate classifier

Once  $d$  is chosen, since we want to work in a distribution free context, densities  $f_d(x|g)$  have to be estimated. Consider a sample  $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$  from  $(X, Y)$  and with  $d_g = d$  for each  $g = 1, \dots, G$ ; a kernel density based estimator for (8) is given by:

$$\hat{\gamma}_n(x, d) = \arg \max_{g=1, \dots, G} \frac{\hat{\pi}_g}{n_g} \sum_{i=1}^n \mathbb{I}_{\{Y_i=g\}} K_{H_g} \left( \left\| \hat{\Pi}_{g,d}(X_i - x) \right\| \right) \quad (9)$$

where  $n_g = \sum_{i=1}^n \mathbb{I}_{\{Y_i=g\}}$  is the number of observations coming from group  $g$ ,  $\hat{\pi}_g = n_g/n$  estimates the mixture coefficient  $\pi_g$ ,  $K_{H_g}(u) = \det(H_g)^{-1/2} K(H_g^{-1/2}u)$ , with  $K$  a kernel function, the bandwidth matrix  $H_g$  is  $d \times d$  and symmetric semi-definite positive, finally  $\hat{\Pi}_{g,d}$  is the projection operator over the subspace spanned by the first  $d$  eigenfunctions of the sample covariance operator  $\hat{\Sigma}_g$  for the group  $g$ .

Note that some simplifications may occur in (9). For instance, if the groups are balanced, that is  $\pi_g = 1/G$  for each  $g$ , then one can drop  $\hat{\pi}_g$  and  $n_g$ . Another example concerns the homoscedasticity case (i.e.  $\Sigma_g$  is the same for each  $g$ ), where  $\hat{\Pi}_{g,d} = \hat{\Pi}_d$  is the projector over the space spanned by the first  $d$  eigenfunctions of the within (or pooled) covariance operator  $\hat{W} = \sum_{g=1}^G \hat{\pi}_g \hat{\Sigma}_g$ .

For what concern the choice of  $d$  and  $H_g$  one can refer to the discussion in Section 2.1.3. It is scarcely necessary to observe that, in the discriminant context, coefficients  $r$  and  $\delta$  are not necessary, because we have a specific bandwidth matrix for each group and the main goal is not the estimation of a mode.

To complete the analysis, we study the asymptotic properties of the classifier  $\hat{\gamma}_n$  defined in (9). In particular, we consider the Bayes probability of error

$$L^* = \min_{\gamma} \mathbb{P}(\gamma(X) \neq Y)$$

and the conditional probability of error

$$L_n = \mathbb{P}(\hat{\gamma}_n(X, d) \neq Y \mid \{(X_1, Y_1), \dots, (X_n, Y_n)\})$$

and we study how  $L_n$  behaves when  $n$  tends to infinity. In particular, convergence of  $L_n$  to the Bayes error probability  $L^*$  is stated in the following proposition, which is a direct consequence of results in [6, Section 5].

**Proposition 4** *Take  $H_g = h_g I$ . Under assumptions (A.1)–(A.5),  $L_n$  converges to  $L^*$  in probability as  $n$  tends to  $\infty$ .*

## 3 Applications to synthetic and real data

This section concerns the simulations and applications of described methods: the first two subsections are dedicated to clustering (3.1 considers an experiment under controlled set-up whereas in 3.2 we apply clustering to real dataset), the last one (Section 3.3) is dedicated to discriminant.

### 3.1 Clustering: simulation examples and comparison with competitors

In the following a simulation study provides a quantitative comparison of the presented algorithm versus competitors. Although the methods are unsupervised, we evaluate their goodness in detecting the underlying groups structure by measuring a misclassification error like if it was a supervise exercise. Besides, by construction the SmBP clustering provides an estimate of the number of clusters that must be studied as well, being itself a source of noise. As pointed out in Section 2.1.3, both misclassification error and the number of detected clusters depend on the choice of parameters: keeping in mind this fact, the simulation exercise is coherently calibrate.

In order to generate the dataset, we use the functional basis expansion:

$$X_i^{(g)}(t) = \sum_{l=0}^L \sqrt{\beta_l} \tau_{i,l}^{(g)} \psi_l(t), \quad t \in [0, 1], \quad i = 1, \dots, n_g \quad \text{and} \quad g = 1, \dots, G,$$

where  $\beta_l = 0.7 \times 3^{-l}$  ( $l = 1, \dots, L = 150$ ) and  $\psi_l(t)$  is the  $l$ -th element of the Fourier basis

$$\psi_l(t) = \begin{cases} \sqrt{2} \sin(2\pi mt - \pi), & l = 2m - 1; \\ \sqrt{2} \cos(2\pi mt - \pi), & l = 2m. \end{cases}$$

For what concerns the mixture, it is controlled by means of  $\tau^{(g)}$ 's. Here we deal with  $G = 2$  and, to avoid spherical shaped groups, uncorrelated but dependent coefficients  $(\tau_{i,l}^{(g)})_{l=1}^L$  are generated as follow:

$$\begin{cases} \tau_{i,1}^{(g)} &= \sin(\vartheta_i) \cos(\frac{\pi}{2} \mathbb{I}_{\{g=2\}}) + \sigma \epsilon_{i,1} \\ \tau_{i,2}^{(g)} &= \sin(\vartheta_i) \sin(\frac{\pi}{2} \mathbb{I}_{\{g=2\}}) + \sigma \epsilon_{i,2} \\ \tau_{i,3}^{(g)} &= \cos(\vartheta_i) + (-k)^g + \sigma \epsilon_{i,3} \\ \tau_{i,l}^{(g)} &= \sqrt{0.1} \epsilon_{i,l}, \end{cases} \quad 4 \leq l \leq L$$

with  $(\vartheta_i)$  i.i.d. as a Beta(5,5) scaled on  $[-\pi, \pi]$  and  $(\epsilon_{i,l})_{l=1}^L \stackrel{i.i.d.}{\sim} N(0, 1)$ . In other words,  $(\tau_{i,l}^{(g)})_{l=1}^3$  are the Cartesian coordinates of the spherical ones  $(1, \theta_i^{(g)}, \frac{\pi}{2} \mathbb{I}_{\{g=2\}})_{l=1}^3$  plus a vertical translation  $(-k)^g$  and a Gaussian noise  $\epsilon$  (randomness is confined in the polar angle  $\vartheta$  and in the noise  $\epsilon$ ). In particular, limited to the first three components, we deal with two noised semi-circumferences laying on orthogonal planes, with unitary radius, whose centers are  $(0, 0, \pm k)$  and chosen so that the clouds of points of  $(\tau_{i,l}^{(g)})_{l=1}^3$  look like two bounded up horseshoes. A reasonable range for  $k$  is  $(0, 1)$ : outside this range, the two un-noised groups can be separated by means of a plane, a structure easily identifiable. For what concerns  $\sigma$ , one can choose  $(0, k/3)$  to avoid that, due to noise variability, groups overlap too much.

With such choices, we are concentrating the process along three orthogonal directions so that the PCs tend to replicate the  $\tau$ 's structure. Moreover, this setting ensures that Proposition 1 applies. In fact, eigenvalues (of  $\Sigma$ ) decay faster or at least equally to  $\{\beta_l \text{Var}(\tau)\}_{l=1}^L$ . Due to

boundedness of  $Var(\tau_l)$ , it inherits the same decay of  $\{\beta_l\}_{l=1}^L$  that is exponentially (3) with  $C = 1/3$ .

In our simulations, we consider  $n_1 = n_2 = 300$ ,  $\sigma = \sqrt{0.005}$  and  $k = 0.5$  to which corresponds  $FEV(3)$  always greater than 99% that suggest us to fix  $d = 3$ . Curves are generated over a grid of 100 equispaced points on  $[0, 1]$ . For the sake of illustration, Figure 1 depicts the scatter plot of an observed set of  $(\tau_{i,l})_{l=1}^3$ , a selection of the corresponding curves and the prototypes regions obtained with our algorithm when  $\delta = 2$  and  $r = 5$  (that produced  $\hat{G} = 2$ ).

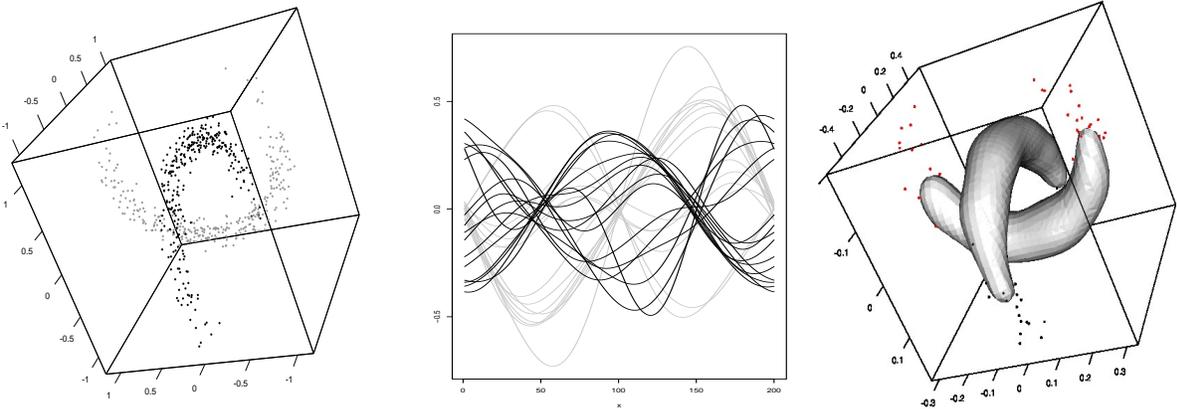


Figure 1: (Left) simulated coefficients  $(\tau_{i,l}^{(k)})_{l=1}^3$ , (middle) a sample of simulated curves and (right) upper level sets associated to the estimated modes on the factor space.

We generate 400 Monte Carlo sample according to the above setting. To each replication we apply the SmBP clustering that returns the corresponding estimated number of clusters  $\hat{G}$  and the misclassification error. According to  $FEV$  criterium we set  $d = 3$  and we explore the behaviour of the algorithm when  $\delta = 0.6, 1, 1.4$  and  $r = 1, 5, 10$ . The following competitors are considered:

- (KM) the functional  $k$ -means clustering (see [9]) with  $G = 2$ ;
- (GM) the EM clustering method based on mixture model of Gaussian components applied to the first three PCs. We consider both  $G = 2$  and  $G$  estimated by means of a Bayesian Information Criterion (BIC). The algorithm is coded in package *Rmixmod* (see [1]).

Table 1 summaries the main results. For SmBP clustering, we report the mean and the standard deviation of misclassification errors and the 0.5, 0.75, 0.9 order quantiles of  $\hat{G}$  varying  $\delta$  and  $r$ ; the same results are provided for (GM) combined with BIC, and only the misclassification error whenever  $G$  is fixed. Such results show that there exists an optimal configuration of parameters  $\delta = 1.4$  and  $r = 10$  for the SmBP clustering for which the two clusters are correctly recognized at least in the 90% of cases, with an average misclassification

error equal to 8.8%. It can be noted that the parametric method (GM) produces good results whenever  $G$  is fixed but gets worse when  $G$  have to be estimated, since BIC overestimates the number of clusters.

| Algorithm | Parameters           |     | Miscl. Error |              | $\widehat{G}$ |            |           |
|-----------|----------------------|-----|--------------|--------------|---------------|------------|-----------|
|           | $\delta$             | $r$ | Mean         | St. Dev.     | $q_{0.5}$     | $q_{0.75}$ | $q_{0.9}$ |
| SmBP      | 0.6                  | 1   | 0.676        | 0.060        | 11            | 13         | 14        |
|           |                      | 5   | 0.480        | 0.112        | 6             | 7          | 8         |
|           |                      | 10  | 0.126        | 0.140        | 2             | 3          | 3         |
|           | 1                    | 1   | 0.563        | 0.088        | 7             | 8          | 9         |
|           |                      | 5   | 0.372        | 0.141        | 4             | 5          | 6         |
|           |                      | 10  | 0.081        | 0.144        | 2             | 2          | 3         |
|           | 1.4                  | 1   | 0.463        | 0.107        | 5             | 6          | 7         |
|           |                      | 5   | 0.299        | 0.146        | 4             | 4          | 5         |
|           |                      | 10  | <b>0.088</b> | <b>0.174</b> | <b>2</b>      | <b>2</b>   | <b>2</b>  |
|           | # clusters $G$       |     |              |              |               |            |           |
| KM        | 2                    |     | 0.377        | 0.068        | —             | —          | —         |
| GM        | 2                    |     | 0.153        | 0.106        | —             | —          | —         |
|           | <i>BIC</i> selection |     | 0.666        | 0.034        | 9             | 10         | 11        |

Table 1: Misclassification errors of SmBP clustering versus competitors and (when available) quantiles of the estimated number of clusters.

## 3.2 Clustering: real data illustration

We illustrate how our clustering technique (from now on, SmBP clustering) work when applied to real dataset. The aim is twofold: from one hand to show the cognitive support that the method could bring on the studied phenomenon, and, on the other hand, what kind of practical problems could occur and how to treat them.

Presentation goes through three datasets belonging to different domains: spectrometric analysis, energy consumption and neuroscience.

### 3.2.1 Spectrometric curves

Spectroscopic analysis is a fast, non-destructive and not-expensive techniques which provides an estimate of the composition of an aliment on the based on the absorption of light emitted with different wavelengths by a spectrometer. Since the measure of absorption is a function of the wavelength, it represents a typical functional data. In the last two decades, various functional techniques has been widely explored for this kind of data: see for instance, [5, 11] in the supervised classification framework.

In the following, we illustrate an application of the SmBP clustering method to the well known Tecator dataset (available at <http://lib.stat.cmu.edu/datasets/tecator>). It

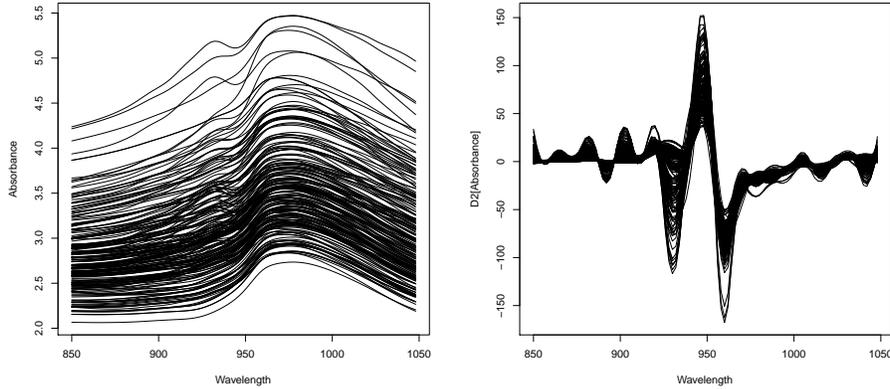


Figure 2: Tecator curves (left) and their 2-nd derivatives (right).

consists of 215 spectra in the near infra-red (NIR) wavelength range from 852 to 1,050 nm, discretized on a grid of 100 equispaced points, corresponding to as many finely chopped pork samples. Fat, protein, and water content, obtained by a traditional chemical analysis, is available for each sample. As conventionally done, in our study we consider the second derivatives of spectrometric curves instead of the original ones and that to avoid the well-known “calibration problem”, due to the presence of shifts in the curves (see [11]). Original spectrometric data and their second derivatives are visualized in Figure 2.

Since spectrometric data should represent a way to determine the chemical composition of the meat, the structure of the distribution of chemical components should be reproduced by the one of spectrometric curves. In this view, we first study the available chemical measures. The correlation analysis shows that the three components are highly linear correlated: in particular fat and water present a linear correlation coefficient equals to  $-0.988$ , whereas the content of protein exhibits a positive correlation with fat ( $0.814$ ) and negative with water ( $-0.861$ ). This suggests to use PCA in order to summarize the chemical composition: in that way, the first PC explains the 98.5% of the total variability. Observing the kernel estimate of the density of this PC (see the upper panel in Figure 3), the poly-modal distribution suggests that the sample is a mixture of three kind of meats: the three groups are detected by considering the largest upper level sets containing the modes of the estimated density that, in the one dimensional case, reduces to look for the local minima whose abscissa identify class boundaries. Hence, it is expected that the spectrometric curve distribution presents a three modal structure as well that, by construction, should be detected by the clustering algorithm defined in Section 2.1.

After running a functional PCA on the second derivatives of spectrometric curves, we get that the spectrum is rather concentrate: the first three PCs explain 98.5% of the total variability, and this suggest to use  $d = 3$  in our approximation. The selection of parameters  $r$  and  $\delta$  is performed according to the maximization of  $CH$  index over a bivariate grid built with  $r = 1, \dots, 7$  and  $\delta$  varying from 0.2 to 1 with step 0.1. Index  $CH$  reaches its maximum

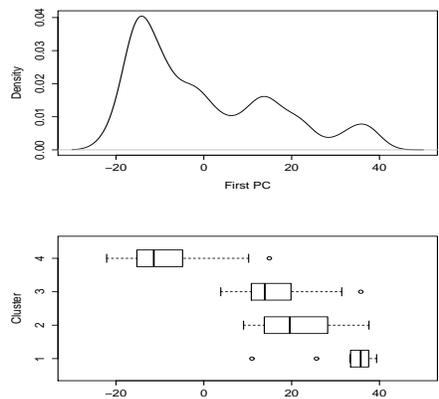


Figure 3: Density estimate of the first PC of chemical components and the same stratified according to clusters analysis carried out with  $r = 5$  and  $\delta = 0.2$ .

for  $r = 5$  and  $\delta = 0.2$  to which correspond  $k = 4$  clusters. In order to understand the appropriateness of this choice, beside the internal criterion  $CH$  we use and external criterion by computing, for each couple of parameters  $r, \delta$ , the index of purity according to the three-group structure shown by the first PC which summarize the chemical variables. It emerges that the couple  $(r, \delta)$  maximizing  $CH$  provides also an high degree of purity: this fact can be appreciated inspecting Figure 3 and provides an heuristic support about the possibility to reproduce the main features of the distribution of the chemical measures from the one of the spectrometric curves.

### 3.2.2 District heating load-curves

A district-heating system (or “teleheating”) allows to distribute heat, generated in a centralized location, for entire districts through a network of insulated pipes. Due to his efficiency and the pollution control, this system is spreading in many cities. To guarantee an optimal scheduling in generating heat, that allows to choose the right mix of on-line capacity, the analysis of flows of demand of heating is crucial. These flows depend mainly on two factors: an intra-daily pattern of the load demand, known as load curve, and on seasonal aspects.

Since consumer characteristics are very different according to seasons and weather conditions, to manage data from district-heating system, also in a forecasting perspective, it is useful to stratify the set of load curves into a few homogeneous groups exhibiting similar demand patterns. In what follows we propose an application of our clustering algorithm to data on heat consumption in Turin, a North-West Italian centre, where the district heating is produced through a co-generation system. The dataset has been used previously in a forecasting context based on regression approaches (see [15, 16]).

The dataset consists of hourly measurements of the heat consumption for residential and commercial buildings during the periods October, 15–April, 20, covering the years 2001–02,

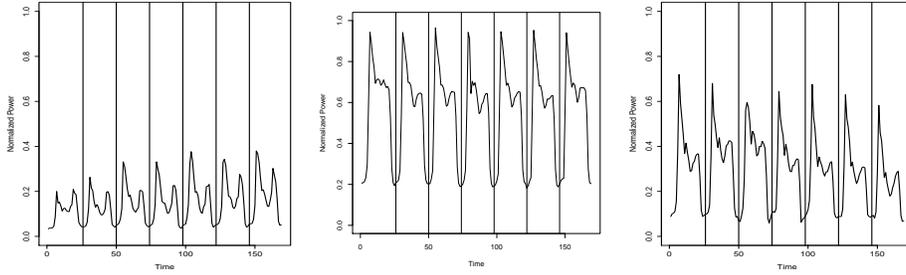


Figure 4: Demand of heat in a selected week of November 2002, January and March 2003.

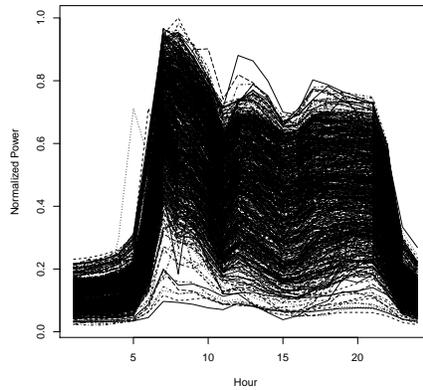


Figure 5: Normalized load curves, corresponding to daily profile.

2002–03, 2003–04 and 2004–05. Due to privacy requests from the data supplier, the data have been normalized. Figure 4 displays behaviour of heating demand in three selected weeks in autumn, winter and spring: it is possible to distinguish the intra-daily pattern, due to an inertia in the demand reflecting the aggregate behaviour of consumers, and also appreciate the seasonal evolution. Differently from electricity power demand, intra weekly differences among working days and weekend do not appear.

Taking advantage of the functional nature of the dataset, in a natural way, we split the series for each period into 187 functional observations, each one coincident with a specific daily load curve. Finally we dispose of a functional dataset consisting in 748 curves discretized over a equispaced mesh of 24 points. Figure 5 displays such set of curves.

The first step of our procedure is to perform a functional principal components analysis. It emerges that the spectrum is rather concentrate: the first three PCs explains more than 97% of the total variance, and thus it is sufficient limit our analysis to  $d \leq 3$ . In order to provide an an interpretation of the contribution of the relevant CPs, we exploit a graphical tool where we report the estimated mean curve plus and minus a suitable multiple  $M_j$

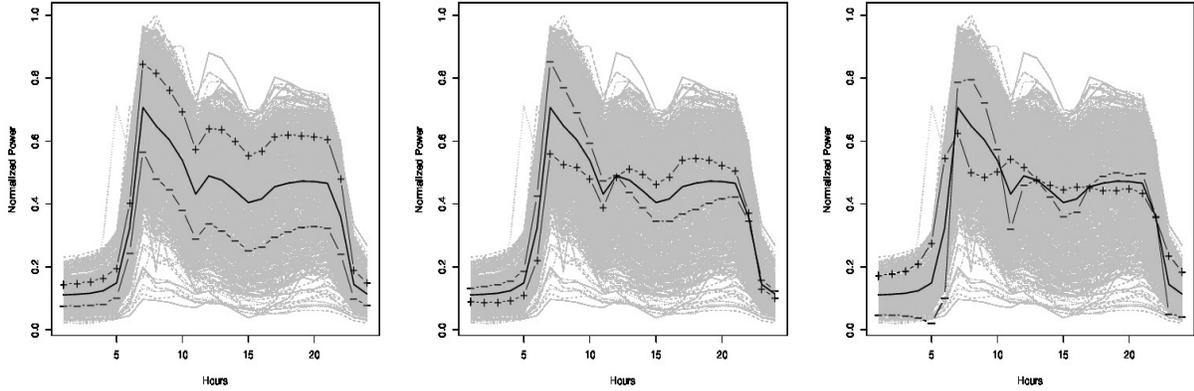


Figure 6: Contribution of the first three principal components.

of each estimated eigenfunction:  $\hat{\mu} \pm M_j \hat{\xi}_j \sqrt{\hat{\lambda}_j}$  (see e.g. [24]). The results, visualized in Figure 6, show that the first eigenfunction, which does not present sign change, describes a vertical shift effect, due to weather conditions in seasons, whereas the second eigenfunction highlights differences among demand in the morning and in the remaining part of the day: it seems related to the heat retention ability of buildings (a greater heating in the morning produces less need in the afternoon). Finally the third eigenfunction seems connected and counter-posed to the three peaks of demand that appear systematically during the day in the morning, in the afternoon and in the evening.

In order to apply the SmbP cluster algorithm one has to select preventively parameter  $r$  and  $\delta$ . Using the same grid as in Section 3.2.1, the  $CH$  criterion suggests  $r = 3$  and  $\delta = 0.5$ , a choice that leads to  $k = 6$  clusters. These clusters reflect the differences in level and behaviour of daily demand of heating in the different seasons: high levels of demand in winter with a strong peak in the morning, moderate level in autumn and spring with load curves presenting three peaks (in the morning, in the afternoon and in the evening). To better understand the effect of clustering, we report in Figure 7 the calendar positioning of each element of clusters (each point represents a specific load curves, synthesized by its daily average) and alongside the modal curves, plotted using the same level of grey. Moreover we report in Figure 7 also the box-plots resulting after a stratification of the daily mean temperature by the cluster labels: one can note as the temperature, without presenting a multi-modal density, is one of the most important external variable that can be used in such clustering exercise. Matching the results, we recognize the typical patterns for winter and mid-seasons, distinguish freezing, cold and mild days. Finally, it emerges that if one would set up a forecasting model, an accurate prevision of temperature will be the key to obtain good performances in predicting demand of heating.

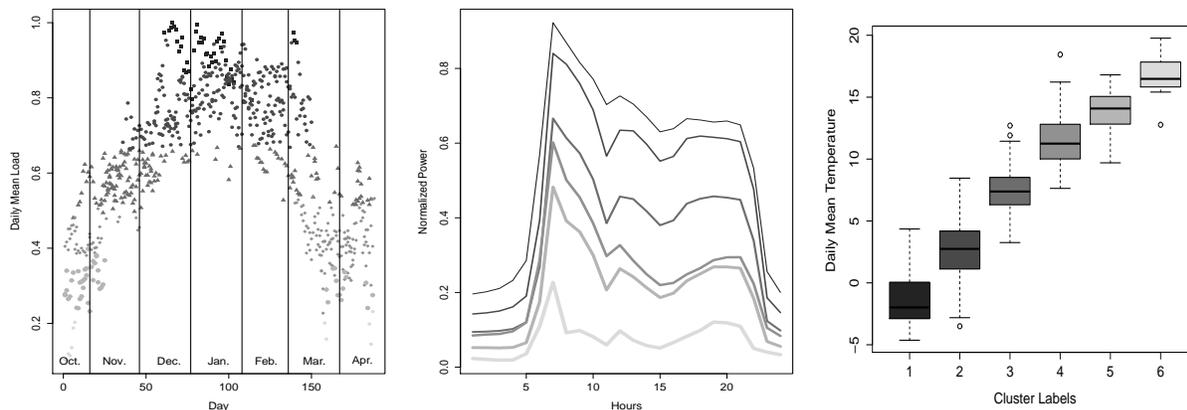


Figure 7: Calendar positioning of clusters against the daily mean load, corresponding modal curves and relationship among clusters and daily mean temperature. Along the panels, a cluster is identified by the same grey level.

### 3.2.3 Neuronal experiment

The analysis of neuronal spiking activity, recorded under various behavioural conditions, is a central tool in neuroscience: data acquired from multiple neurons are essential to explain neural information processing. The problem is that contributions of multiple cells must be disentangled from the background noise and from each other in order to analyze the activity of individual neurons. The procedure that allows to distinguish the activity of one or more neurons from a noisy time series is known as spike sorting.

In this section we show how the SmBP clustering can contribute in spike sorting perspective: each detected cluster can be thought to correspond to the activity of a single neuron. The dataset comes from a behavioural experiment performed at the Andrew Schwartz motorlab (<http://motorlab.neurobio.pitt.edu/index.php>) on a macaque monkey performing a center-out and out-center target reaching task with 26 targets in a virtual 3D environment (see [27] for a detailed description of the experiment considered). The neural activity recorded consists of all the action potentials detected above a channel-specific threshold on a 96-channel Utah array implanted in the primary motor cortex. The data set is split in 1000 functional data representing the voltage of neurons versus the times, discretized over a grid of 32 equispaced time points (normalized between 0 and 1). A sample, selected randomly, of 30 of such curves is shown in Figure 8. An analysis of (a larger set of) these curves can also be found in [27].

Performing the functional PCA on such set of curves we observe that also in this case the spectrum is concentrated: the first PC explains 78.6% of the total variability, and explained variance by the first three PCs is about 96.4%. Observing the 3-dimensional scatter plot of the first three PCs (see Figure 9, left panel), three clouds appear evidently.

In order to detect a good choice for parameters  $r$  and  $\delta$ , we perform our clustering

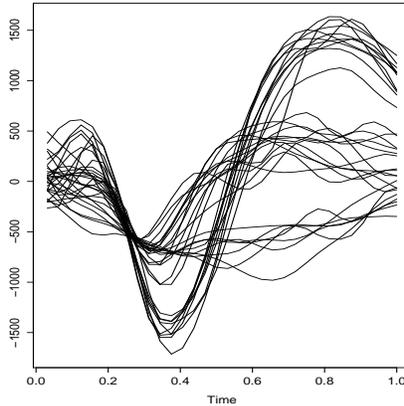


Figure 8: A random sample of 30 curves.

algorithm with  $d = 3$  over a grid built from  $r = 2, \dots, 7$  (with step equal to 1) and  $\delta = 0.2, \dots, 1$  (with step 0.1) and compute the  $CH$  indexes. The analysis of obtained values leads to various admissible configuration for the couple  $(r, \delta)$ , to which always correspond  $k = 3$  clusters: for instance  $r = 3, \dots, 7$  combined with  $\delta = 0.8, 0.9, 1$  produce the same number of cluster and the maximal  $CH$ . The right panel of Figure 9 shows the maximal level set when one uses  $r = 7$  and  $\delta = 0.8$ .

Result of clustering procedure is visualized in Figure 10 where the centers of prototypes (the modal curves) and the corresponding clusters are reproduced.

### 3.3 Discriminant: simulation and real data illustration

The aim of this section is to assess the performances of the supervised classification algorithm illustrated in Section 2.2 (briefly, SmBP classifier) by the analysis of both simulated and real datasets. The experiments consist in computing the (empirical) distribution out-of-sample misclassification error by a two-fold cross-validation procedure repeated 100 times: more in detail, for each available sample, at each iteration, the 2/3 of the data are used in evaluating the classifier, and the misclassification error is estimated on the remaining part. The estimation of the density in each group is performed by a multivariate kernel density estimator with  $H$  diagonal, selected according to Section 2.1.1.

The out-of-sample errors are compared with those computed using parametric and non-parametric competitors: the GLM classifier based on the coefficients of a basis representation for functional data, nonparametric discrimination using kernel (NP in the following) and k-NN estimation, based on the classic  $L^2$  metric (see [11]). All computations are done with the software R: in particular, the competitor algorithms are taken from the package *fda.usc* (see [9]).

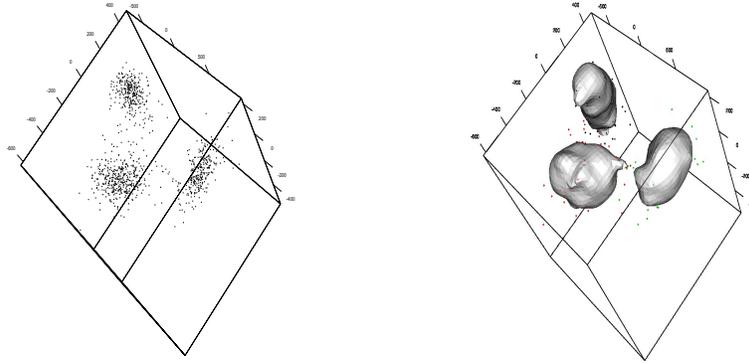


Figure 9: 3-D Scatter plot of the first three Principal Components and the corresponding maximal level set when  $r = 7$  and  $\delta = 0.8$ .

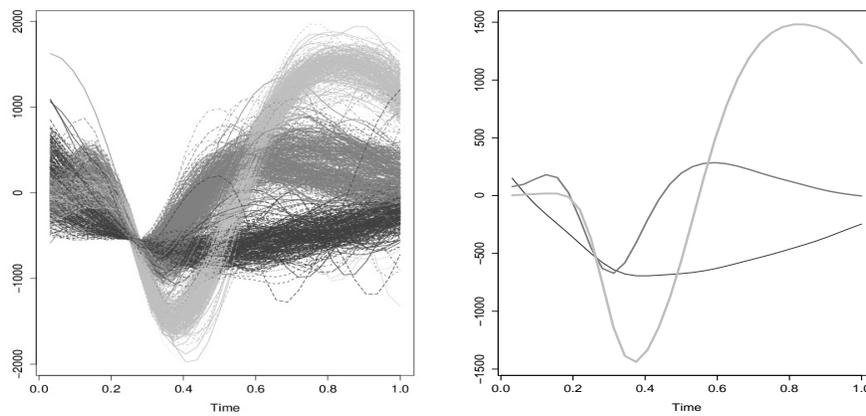


Figure 10: Clusters of curves and modal curves.

| $n_{in}$ | $\sigma$ | $n_{out}$ | $d = 2$          | $d = 3$          | $d = 4$          | $d = 5$          | GLM              | k-NN             | NP               |
|----------|----------|-----------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| 100      | 0.05     | 50        | 0.155<br>(0.039) | 0.026<br>(0.015) | 0.039<br>(0.024) | 0.058<br>(0.035) | 0.361<br>(0.060) | 0.024<br>(0.020) | 0.042<br>(0.023) |
|          | 0.10     |           | 0.228<br>(0.052) | 0.095<br>(0.032) | 0.111<br>(0.040) | 0.134<br>(0.044) | 0.345<br>(0.054) | 0.104<br>(0.041) | 0.130<br>(0.041) |
|          | 0.15     |           | 0.257<br>(0.061) | 0.188<br>(0.045) | 0.195<br>(0.050) | 0.221<br>(0.051) | 0.397<br>(0.064) | 0.159<br>(0.049) | 0.200<br>(0.048) |
| 200      | 0.05     | 100       | 0.185<br>(0.042) | 0.033<br>(0.017) | 0.042<br>(0.017) | 0.057<br>(0.019) | 0.307<br>(0.031) | 0.041<br>(0.018) | 0.047<br>(0.022) |
|          | 0.10     |           | 0.275<br>(0.053) | 0.150<br>(0.049) | 0.154<br>(0.027) | 0.167<br>(0.028) | 0.335<br>(0.029) | 0.136<br>(0.038) | 0.148<br>(0.031) |
|          | 0.15     |           | 0.255<br>(0.039) | 0.228<br>(0.033) | 0.234<br>(0.031) | 0.265<br>(0.035) | 0.440<br>(0.047) | 0.232<br>(0.040) | 0.250<br>(0.036) |
| 300      | 0.05     | 150       | 0.157<br>(0.029) | 0.020<br>(0.007) | 0.033<br>(0.011) | 0.036<br>(0.010) | 0.356<br>(0.033) | 0.030<br>(0.010) | 0.035<br>(0.011) |
|          | 0.10     |           | 0.217<br>(0.033) | 0.100<br>(0.024) | 0.117<br>(0.023) | 0.128<br>(0.026) | 0.383<br>(0.036) | 0.093<br>(0.026) | 0.108<br>(0.026) |
|          | 0.15     |           | 0.234<br>(0.030) | 0.171<br>(0.024) | 0.183<br>(0.026) | 0.192<br>(0.023) | 0.411<br>(0.030) | 0.162<br>(0.026) | 0.178<br>(0.027) |

Table 2: Estimated mean and standard deviation (in parentheses) of misclassification error for the two-horseshoes setting.

### 3.3.1 Analysis of simulated data

The dataset are generated according to the two “horseshoes” model described in Section 3.1 with the following settings:

- Sample size  $n = 150, 300, 450$  with training-sets of size  $n_{in} = 100, 200, 300$ ;
- The groups are balanced:  $\pi_g = 0.5$  for  $g = 1, 2$ ;
- The vertical translation parameter  $k$  equals 0.5;
- Three degree of variability are considered to model the noise around the two semi-circumference:  $\sigma = 0.05, 0.10, 0.15$ , (small, medium and high variability).

Performing the discrimination exercise (we use, for our classifier, a number  $d$  of PCs varying from 2 to 5), one obtains the misclassification error distributions whose summary measures (mean and standard deviation) are collected in Table 2. In Figure 11 are reproduced such error distribution when  $n_{in} = 200$ ,  $\sigma = 0.05, 0.10, 0.15$ . It emerges that when  $d = 3$  the SmBP classifier produces the best results: they are comparable with the ones of the k-NN and the nonparametric approach. As expected, due to the non-spherical nature of data, the GLM approach produces the worst results.

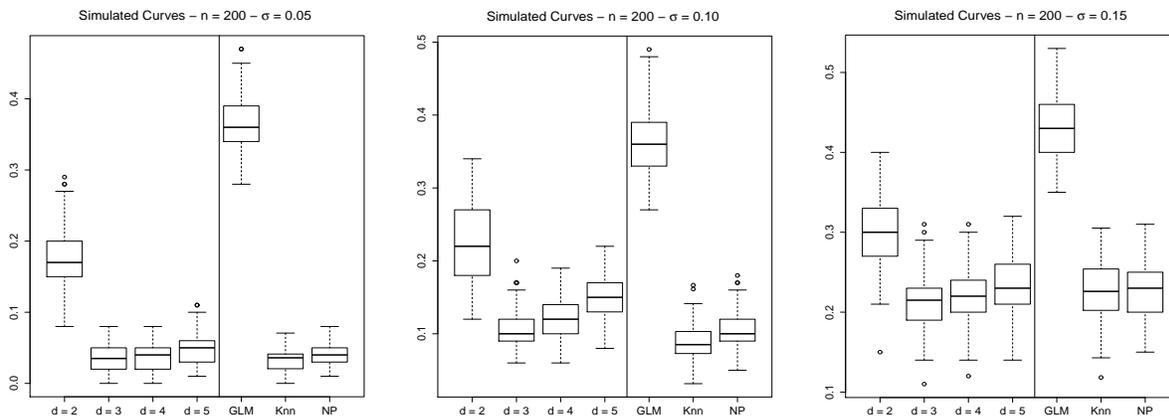


Figure 11: Distributions of misclassification errors estimated over 100 replications when  $n = 200$  and  $\sigma = 0.05, 0.10, 0.15$  (form left to right).

### 3.3.2 Analysis of real datasets

In what follows we analyze performances of our SmBP classifier on three real well-known datasets belonging to three very different research domains: electrocardiography, growth curves and quality control. The same datasets have been used previously in [18] in a unsupervised classification framework.

The first dataset comes from the UCR Time Series Classification and Clustering website ([http://www.cs.ucr.edu/~eamonn/time\\_series\\_data/](http://www.cs.ucr.edu/~eamonn/time_series_data/)). It consists of 200 electrocardiography (ECG) curves observed at 96 discretization points and related to 2 groups of patients (see [23] for more details).

The second dataset is the well-known Berkeley growth dataset (see [28]). It contains stature measurements for 54 girls and 39 boys, aged from 1 to 18 years, and observed in 31 (not equispaced) discretization points. To obtain the growth curves, the original raw data are preprocessed by fitting each individual set of discretized data by a monotone smoothing method (see [24]). The aim is to discriminate the curves on the base of the gender.

The third dataset, described in detail in [21], comes from Danone Vitapole Paris Research Center. The aim is to detect the quality of produced cookies in relationship with the flour kneading process. Each curve in the dataset collects the measures of dough resistance of flour during the kneading process at 241 equispaced instants of time in the interval  $[0, 480]$  seconds. Overall, 115 flours are analyzed: 50 of them have produced cookies of good quality, 25 produced medium quality and 40 low quality. The goal of the analysis is to classify the functional dataset on the base of quality of cookies.

The three datasets of curves are plotted in Figure 12: in the plots the group membership of each individual (patient, child or flour) is highlighted by using a different grey scale.

We apply our SmPB classification method with balanced group varying the dimension  $d$ : for the ECG dataset we use  $d = 2, 3, \dots, 9$  whereas for the other two cases,  $d = 2, \dots, 5$ . The distributions of misclassification errors for our approach and the competitors are reproduced

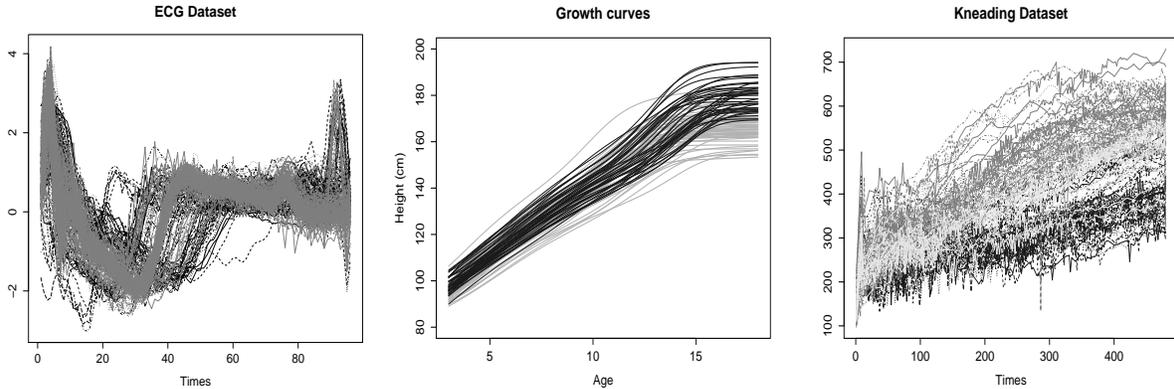


Figure 12: Curves of considered datasets: ECG (2 groups), growth curves (2 groups) and kneading process (3 groups).

| $\delta = 1$ | ECG                   | Growth Curves         | Kneading Process      |
|--------------|-----------------------|-----------------------|-----------------------|
| SmBP         | $d = 7$ 0.128 (0.036) | $d = 3$ 0.028 (0.024) | $d = 3$ 0.273 (0.061) |
| GLM          | 0.212 (0.050)         | 0.075 (0.049)         | 0.286 (0.070)         |
| k-NN         | 0.082 (0.036)         | 0.030 (0.027)         | 0.221 (0.068)         |
| NP           | 0.143 (0.039)         | 0.041 (0.030)         | 0.251 (0.064)         |

Table 3: Estimated mean and standard deviation (in parentheses) of misclassification error distribution for the three real datasets. For the SmBP approach the dimension  $d$  at which we obtain the best results is reported.

in Figure 13. Moreover, to allow a direct comparison, Table 3 collects the estimated mean and standard deviation of misclassification error distributions for the three real datasets: for the SmBP approach, we report the best results obtained and the corresponding dimension  $d$ . Unbalanced groups structure (requiring the estimate of prior probabilities  $\pi_g$ ) does not change obtained results. It appears how the SmBP classifier is sensitive to  $d$ : there exist an optimal choice for  $d$  even if larger values may amplify noise due to bad estimation of  $f_d$ . It is worthy to note that our method performs rather well in comparison with to the other ones: despite the k-NN approach tends to produce good results uniformly in all cases, our method is always comparable, with closed results. More in detail, in the growth curves case with  $d = 3$  SmBP classifier is equivalent to k-NN and better than the nonparametric approach. For the ECG dataset the best result is when  $d = 7$ ; it is comparable with the results obtained the nonparametric classifier. About the kneading process dataset, it appears a general difficulty in produce good classification results and all the proposed method suffer of this common problem.

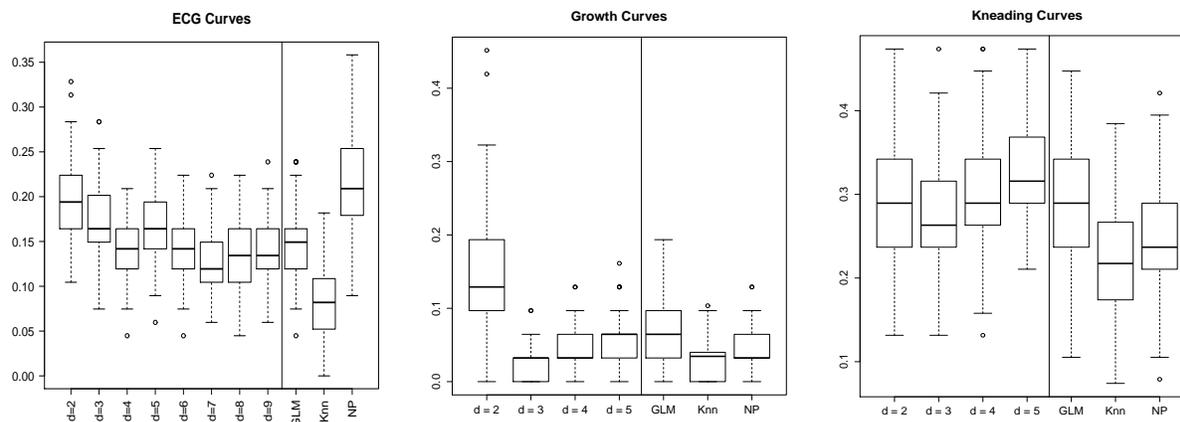


Figure 13: Out-of-sample Misclassification errors over 100 replications. Case studies (left to right): ECG, Growth and Kneading curves.

**Acknowledgements** The authors thank E. Keogh for providing the ECG dataset, C. Preda for providing the Danone Vitapole dataset and A. Schwartz, V. Ventura and S. Todorova for providing the neuronal experiment dataset.

The authors were partially funded by the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

## References

- [1] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, 51(2):587–600, 2006.
- [2] E. G. Bongiorno and G. Aldo. A clustering method for hilbert functional data based on the small ball probability. 2015.
- [3] E. G. Bongiorno, G. Aldo, S. Ernesto, and V. Philippe, editors. *Contributions in infinite-dimensional statistics and related topics*, 2014. Società Editrice Esculapio.
- [4] A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *Ann. Statist.*, 38(2):1171–1193, 2010.
- [5] A. Delaigle, P. Hall, and N. Bathia. Componentwise classification and clustering of functional data. *Biometrika*, page ass003, 2012.
- [6] L. Devroye et al. On the almost everywhere convergence of nonparametric regression function estimates. *The Annals of Statistics*, 9(6):1310–1319, 1981.

- [7] R. C. Dubes. Cluster analysis and related issues. In C. H. Chen, L. F. Pau, and P. S. P. Wang, editors, *Handbook of Pattern Recognition and Computer Vision*, pages 3–32. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 1993.
- [8] T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Statist.*, 32(3):485–506, 2005.
- [9] M. Febrero-Bande and M. Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012.
- [10] F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, 44(1):161–173, 2003.
- [11] F. Ferraty and P. Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006.
- [12] F. Ferraty, N. Kudraszow, and P. Vieu. Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *J. Nonparametr. Stat.*, 24(2):447–464, 2012.
- [13] T. Gasser, P. Hall, and B. Presnell. Nonparametric estimation of the mode of a distribution of random curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(4):681–691, 1998.
- [14] G. Gimelfarb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, T. Windeatt, and K. Yamada. *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, SSPR&SPR 2012, Hiroshima, Japan, November 7-9, 2012. Proceedings*. Springer, 2012.
- [15] A. Goia. A functional linear model for time series prediction with exogenous variables. *Statistics & Probability Letters*, 82(5):1005–1011, 2012.
- [16] A. Goia, C. May, and G. Fusai. Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting*, 26(4):700–711, 2010.
- [17] L. Horváth and P. Kokoszka. *Inference for functional data with applications*, volume 200. Springer Science & Business Media, 2012.
- [18] J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.*, 71:92–106, 2014.
- [19] G. M. James and T. J. Hastie. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):533–550, 2001.
- [20] G. M. James and C. A. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

- [21] C. Lévêder, C. Abraham, P. Cornillon, E. Matzner-Lober, and N. Molinari. Discrimination de courbes de pétrissage. *Chimiométrie 2004*, pages 37–43, 2004.
- [22] J. Liu, Y. Chen, J. M. Maisog, and G. Luta. A new point containment test algorithm based on preprocessing and determining triangles. *Computer-Aided Design*, 42(12):1143 – 1150, 2010.
- [23] R. T. Olszewski. Generalized feature extraction for structural pattern recognition in time-series data. PhD Thesis, 2001.
- [24] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.
- [25] H. Shin. An extension of fisher’s discriminant analysis for stochastic processes. *Journal of Multivariate Analysis*, 99(6):1191–1216, 2008.
- [26] B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.
- [27] S. Todorova, P. Sadtler, A. Batista, S. Chase, and V. Ventura. To sort or not to sort: the impact of spike-sorting on neural decoding performance. *Journal of neural engineering*, 11(5):056005, 2014.
- [28] R. Tuddenham and M. Snyder. Physical growth of california boys and girls from birth to age 18. *California Publications on Child Development*, 1:183–364, 1954.
- [29] R. Xu and D. Wunsch II. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

Stampato in proprio presso l'Ufficio Risorse del Dipartimento di Studi per  
l'Economia e l'Impresa, Università del Piemonte Orientale

**Dipartimento di Studi per l'Economia e l'Impresa**

Via E. Perrone 18 – 28100 Novara

Tel. 39 (0) 321- 375531 Fax 39 (0) 321 - 375535

e-mail: [segreteria.amministrativa@eco.unipmn.it](mailto:segreteria.amministrativa@eco.unipmn.it)