

Classification methods for Hilbert data based on surrogate density

Enea G. Bongiorno¹, Aldo Goia
Università del Piemonte Orientale, Italy

Abstract

An unsupervised and a supervised classification approaches for Hilbert random curves are studied. 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 focus is on the illustration of the classification algorithms and the computational implications, with particular attention to the tuning of the 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.

Revised Personal Version of the text of the final journal article:

E. G. Bongiorno, A. Goia. *Classification methods for Hilbert data based on surrogate density*.
Computational Statistics & Data Analysis 99 (2016), pp. 204-222. (doi: 10.1016/j.csda.2016.01.019)
<http://www.sciencedirect.com/science/article/pii/S0167947316300056>

It reflects changes made in the peer review process.

Introduction

In multivariate classification problems, whether they are supervised or unsupervised, joint density, or better its estimate, plays a central role. In order to make it clear, one has just to think about the literature on 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 [Bock et al., 2013, 2014](#), [Gimelfarb et al., 2012](#)).

When one deals with data belonging to functional spaces (for a general introduction on this topic, one can refer to the monographs of [Ferraty and Vieu, 2006](#), [Horváth and Kokoszka, 2012](#) and [Ramsay and Silverman, 2005](#), and to the recent book [Bongiorno et al.](#),

¹Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Via Perrone 18, 28100, Novara, Italia. Tel: (+39) 0321 375 317 e-mail: enea.bongiorno@uniupo.it

2014), the dimensionality problem arises immediately, and, as a consequence, a probability density function generally does not exist (see [Delaigle and Hall, 2010](#)). Hence, a direct extension of density oriented classical multivariate classification approaches to functional data cannot be implemented: usually, a reduction of dimensionality based on projection over suitable finite subspaces, is made as a preliminary step to tackle the problem. This route is followed for instance by [James and Sugar \(2003\)](#), where model based clustering methods are proposed, and by [James and Hastie \(2001\)](#) and [Shin \(2008\)](#) 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 the underlying process. It is worth noting that also other dimensionality reduction approaches are possible: for instance, one can recall the techniques, which rest on the most important points, illustrated in [Delaigle et al. \(2012\)](#).

Another way to proceed, that aims at working directly on the distribution of the process, refers to the concept of surrogate (or pseudo) density. The general principle, dating back to [Gasser et al. \(1998\)](#), is to factorize the small ball probability associated with the functional data, when the radius of the ball tends to zero, as a 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 finite dimensional classification methods. Theoretical conditions that allow such a 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 [Delaigle and Hall \(2010\)](#) and then in [Bongiorno and Goia \(2015\)](#), where some assumptions are relaxed. One can observe that this approach 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 [Jacques and Preda \(2014\)](#) in defining a Gaussian mixture clustering procedure. On the other hand, a full nonparametric approach is possible as done in [Ferraty et al. \(2012\)](#) where a k -NN procedure is proposed to estimate the pseudo-density.

In this paper we consider an intermediate approach for evaluating 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, in both clustering and discriminant framework, to mixtures presenting non-spherical clusters.

The paper is organised as follows: in Section 1 we introduce the mixture model and its factorization; Section 2 is devoted to illustrating the classification algorithms (clustering and discriminant) and to discussing the parameters used; Section 3 collects the applications to simulated data and real cases; Section 4 gathers some final comments; finally, in Appendix A we sketch the proofs of the main theoretical results.

1. Theoretical framework

Let X be a random element defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, taking 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 Ferraty and Vieu, 2006 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 Ω is partitioned in G (unknown and finite) sub-sets Ω_g , and let Y be a \mathbb{N} -valued random variable defined by

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

(here \mathbb{I}_A denotes 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 for approaching 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 leads 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 tackling it directly, we want to simplify it by exploiting an approximation result sketched below (for more details see [Bongiorno and Goia, 2015](#)). 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 is made.

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 Σ , 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

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

Without loss of generality, from now on 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) 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.3) 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.4) the spectrum of Σ is rather concentrate: $\{\lambda_j\}_{j=1}^\infty$ decays to zero exponentially, that is

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j < C, \quad \text{for any } d \in \mathbb{N}. \quad (2)$$

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

$$\varphi(x, \varepsilon) \sim f_d(x_1, \dots, x_d) \phi(d, \varepsilon). \quad (3)$$

From a practical point of view, and as we will see in the sequel, the exponential decay required by Assumption (A.4) is sufficient to reach the scopes of this paper. Nevertheless, to better appreciate the interpretation of factors in (3) and, in particular, because the form of $\phi(d, \varepsilon)$ depends on them, we list below the forms of ϕ associated to the different eigenvalues decays. In particular:

- if $\{\lambda_j\}_{j=1}^\infty$ decays exponentially (2), 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 α is a parameter chosen in such a way 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

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

The fact that in the hyper-exponential case, $\phi(d, \varepsilon)$ is the volume of a d -dimensional ball with radius ε , justifies to interpret, in Equation (3), $\phi(d, \varepsilon)$ 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. Note that (5) \Rightarrow (4) \Rightarrow (2), whereas the vice versa does not hold. For instance, for any $\alpha > 1$ and $\beta > 0$, $\lambda_d = \exp\{-\beta d\}$ decays exponentially but not super-exponentially, $\lambda_d = \exp\{-\beta d \ln(\ln(d))\}$ decays super-exponentially but not hyper-exponentially while $\lambda_d = \exp\{-\beta d^\alpha\}$ decays hyper-exponentially.

Remark 3. Although the results are exposed using the Karhunen-Loève (or PCA) basis, they hold for any orthonormal basis ordered according to the decreasing values of the variances of the projections, provided that they decay sufficiently fast (see [Bongiorno and Goia, 2015](#)). Note that the variances obtained when one uses the PCA basis present, by construction, the faster decay: in this sense the choice of this basis can be considered optimal.

2. Classification

This section is devoted to defining classification procedures in a functional framework. The idea is to exploit the asymptotic factorization results provided by Proposition 1. It is divided in two subsections: the first one illustrates a clustering algorithm, whereas the second one deals with a discriminant analysis procedure.

2.1. Unsupervised classification

Consider a sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ drawn from (X, Y) defined as in Section 1, where X 's are observed while the group variables Y 's are latent. Our 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 $(X|Y = g)$ is specified then a full parametric approach applies; this has been done, for example, in Jacques and Preda (2014) where the authors used a maximum likelihood and expectation maximization approach to identify the distribution parameters of a Gaussian mixture assumed for f_d . Instead, if no information is 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:

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

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; hence, from now on, we assume that there exists a positive integer d^* such that f_d is G -modal for any $d \geq d^*$. Therefore, our scope is to identify the groups by a “locally high (surrogate) density regions” principle: the clustering algorithm computes the estimates $\hat{m}_{d,g}$ of the modes $m_{d,g}$, that is the local maxima of f_d , whose number \hat{G} estimates G ; for each g , it finds the largest connected upper-surface containing only $\hat{m}_{d,g}$, and hence it assigns group labels to each observation consistently with its proximity to these sets. The algorithm procedure is illustrated below:

Step 1. Estimate the covariance operator and the eigenelements.

Step 2. Fixed d , compute $\hat{f}_{d,n}$ (an estimation of the joint distribution density f_d).

Step 3. Look for the local maxima $\hat{m}_{d,g}$ of $\hat{f}_{d,n}$, $g = 1, \dots, \hat{G}$ over a grid.

Step 4. *Finding Prototypes*: for each g in $\{1, \dots, \hat{G}\}$, the g -th “prototypes” group is formed by those X_i whose estimated PCs $(\hat{\theta}_{1,i}, \dots, \hat{\theta}_{d,i})$ belong to the largest connected upper-surface of $\hat{f}_{d,n}$ that contains only the maximum $\hat{m}_{d,g}$. In other words, for such individual, $\hat{Y}_i = g$.

Step 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 $\xi_j(t)$. In the remaining part of this section, some theoretical and practical aspects of the algorithm will be discussed.

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, $\widehat{\Pi}_d(\cdot) = \sum_{j=1}^d \widehat{\xi}_j \langle \widehat{\xi}_j, \cdot \rangle$ is the projection operator over the subspace spanned by $\{\widehat{\xi}_1, \dots, \widehat{\xi}_d\}$, i.e. the first d eigenfunctions of $\widehat{\Sigma}_n$ (the sample version of Σ) so that the kernel argument depends on the estimated PCA semi-metric (Ferraty and Vieu, 2006, Section 8.2). It is worth noticing that in estimating f_d the use of the estimated projector $\widehat{\Pi}_d$ instead of Π_d introduces a non-standard source of noise that might modify the consistency properties and, hence, should be considered with care. In fact, 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 Π_d is known. A positive answer was provided in Bongiorno and Goia (2015); 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 a Lipschitz, bounded, integrable density function with compact support $[0, 1]$;
- (B.4) the process X is bounded.

Then, the following result holds:

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

$$c_1 n^{-\frac{1}{2p+d}} \leq h_n \leq c_2 n^{-\frac{1}{2p+d}} \quad (6)$$

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

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

uniformly in \mathbb{R}^d .

From a practical point of view, the bandwidth selection is an important task: our choice is to consider a diagonal bandwidth matrix (see [Duong and Hazelton, 2005](#) for a heuristic justification) whose non-null entries are the univariate bandwidth provided by [Silverman \(1986, p.48\)](#). Anyway, it is clear that the larger is $|H|$, the “smoother” is $\widehat{f}_{d,n}$, the 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 by applying to H a scale factor $\delta > 0$, whose optimality (in a sense to be specified) is discussed below.

2.1.2. Prototypes identification and Modes

The identification of prototypes is the core of the algorithm: it rests on the largest connected upper-level sets of $\widehat{f}_{d,n}$ related to each mode $\widehat{m}_{d,g}$. The latter plays just an instrumental part in identifying prototypes. More in details, we use the graphical visualization system of R software: X_i is assigned to the g -th prototype group if its estimated PCs belong to the g -th upper-level sets, by using the algorithm described in [Liu et al. \(2010\)](#) and available in the R-package *ptinpoly*. From a practical point of view, a problem concerns spurious modes caused by the choice of δ and sampling variability. To tackle this issue, we select only those modes $\widehat{m}_{d,g}$ for which $\widehat{f}_d(\widehat{m}_{d,g})$ is maximum over the parallelepiped $\widehat{m}_{d,g} \pm [0, rh_1] \times \dots \times [0, rh_d]$, 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.

It is worth recalling that, in the multivariate literature, alternative techniques in detecting high density regions can be found (see for instance, [Azzalini and Torelli, 2007](#), [Sager, 1979](#), [Rinaldo et al., 2012](#) and references therein).

To conclude, we provide a glimpse of the theoretical aspects about mode estimation. For a fixed d , consistency properties for estimated modes have been considered, for instance, in [Abraham et al. \(2003\)](#), [Chen et al. \(2015\)](#) and [Sager \(1979\)](#). Nevertheless, such theoretical issues are little studied in the infinite dimensional framework, since the mathematical concept of the density is still under-developed; some attempts can be found in [Gasser et al. \(1998\)](#) and, more recently, in [Bongiorno and Goia \(2015\)](#), [Dabo-Niang et al. \(2007\)](#) and [Delaigle and Hall \(2010\)](#).

2.1.3. Tuning parameters

Given a data set, different values of d , r and δ may lead to different cluster results: thus, in many situations, a criterion to choose them is opportune. In literature it is acknowledged that, for the most clustering algorithms, there is not a selection “golden rule” of the parameters. In practice, a series of trials and repetitions are performed to tune the parameters (see

Xu and Wunsch II, 2005 and references therein). In this view, automatic selection rules can be used only as support in decisional steps.

In view of Proposition 1, parameter d should be large enough to guarantee a good approximation for the SmBP (3), 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, δ) , and to provide some insights on the quality of clustering solutions, we exploit both external and internal criteria although they are not universal and effective tests for an optimal choice. 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, a combination of the two indices. Thus, accordingly to the nature of data, the idea is to look at those (r, δ) which furnish the best value for these validation criteria and, as a consequence, suggesting the number of clusters G . In the remaining part of this section, we summarise these two criteria; more details can be found in Xu and Wunsch II (2005) and references therein.

The purity index measures how close a clustering is 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 a member of cluster g belongs to class j as $p_{gj} = \pi_{gj} / \pi_g$, where π_g is the proportion of objects in cluster g , and π_{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 well-known and often achieves the best performance (see Dubes, 1993). 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 clusters obtained by choosing the couple (δ, r) , and $Tr(S_B)$ and $Tr(S_W)$ are the traces of the estimated between and within covariance matrices, respectively. The couples (δ, r) that maximize CH are selected as optimal, and the number of clusters K is consequently obtained.

It is worth noticing that purity is not affected by the geometry of the point clouds, whereas CH is. In particular, since CH definition is based on variances, it is expected that CH provides good results whenever the clusters are elliptical point clouds and not otherwise.

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 group 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 corresponds the highest *a posteriori* 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 the index g' in $\{1, \dots, G\}$ such that

$$\frac{\mathbb{P}(Y = g'|X = x)}{\mathbb{P}(Y = g|X = x)} > 1, \quad \text{for any } g = 1, \dots, G \text{ and } g \neq g'. \quad (7)$$

If a probability density of X in the g -th group $f(x|g)$ were known (with $f(x|g) > 0$), thanks to the Bayes formula, Equation (7) would simplify 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 would become:

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

It is clear that such arguments do not apply straightforwardly in 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, as ε tends to 0,

$$\frac{\mathbb{P}(Y = g \mid \|X - x\| < \varepsilon)}{\mathbb{P}(Y = g' \mid \|X - x\| < \varepsilon)} > 1, \quad \text{for any } g' \neq g. \quad (8)$$

At a glance, it is evident that it is hard to use in practice. Anyway, thanks to the Bayes formula, the ratio in (8) 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 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 Σ_g of the group g . Clearly, the asymptotic behaviour of the ratio 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 assumptions are imposed on the mixture process X , since the spectrum decay of Σ controls the one of each Σ_g (the conditional covariance operator corresponding to 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 (in decreasing order). The latter inequality ensures that the eigenvalues of B , W and Σ_g ($g = 1, \dots, G$) have a decay fast at least as the one of Σ . In other words, since the eigenvalues' decay rate is a measure of how much X is concentrated in the space, the process $(X|Y = g)$ in each sub-population must be “concentrated” at least as much as X . As a consequence, if the spectrum of Σ decays exponentially (according to (2)), 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 (8) similarly to the multivariate case, by replacing a probability density with a surrogate version: assign a new functional observation x to the g -th group for which, as ε 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 $+\infty$,

$$\gamma(x, d) = \arg \max_{g=1, \dots, G} \pi_g f_d(x|g). \quad (9)$$

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 on ε by means of d (see Proposition 1), it is not so restrictive to assume that

(A.5) $\gamma(x, d)$ is constant as d goes 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 [James and Hastie \(2001\)](#) 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 previously 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 assumes an underlying Gaussian mixture model, both approaches lead to the same classifier: the present section theoretically justifies the use of (9) in a finite dimensional subspace. However, if

the eigenvalues of Σ decay slowly, we cannot ensure that d_g is the same varying g , and the volume terms $\phi(d_g, \varepsilon)$ cannot be neglected in the classification rule. Consequently, the approach based on a pure projective method could be unfruitful.

The illustrated method can be framed by the full parametric discrimination described in [James and Hastie \(2001\)](#) and the full nonparametric one proposed by [Ferraty and Vieu \(2003\)](#), where the *a posteriori* 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) , with $d_g = d$ for each $g = 1, \dots, G$; a kernel density based estimator for (9) is given by:

$$\hat{\gamma}_n(x, d) = \arg \max_{g=1, \dots, G} \frac{\hat{\pi}_g}{n_g} \sum_{i=1}^n \mathbb{I}_{\{Y_i=g\}} K_{H_g} \left(\left\| \hat{\Pi}_{g,d} (X_i - x) \right\| \right) \quad (10)$$

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 π_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, and 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 (10). 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. Σ_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 concerns 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 δ 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 (10). 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 the results by [Devroye \(1981\)](#), Section 5.

Proposition 5. *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 the previously described methods: the first two subsections are dedicated to clustering (3.1 considers an experiment under a controlled set-up, whereas in 3.2 we apply clustering to a real dataset), and the last one (Section 3.3) is dedicated to discriminant analysis.

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 ability in detecting the underlying group structure by measuring a misclassification error, as if it was a supervised exercise. Besides, by construction the SmBP clustering provides an estimate of the number of clusters that must be studied as well, it being a source of noise. As pointed out in Section 2.1.3, both misclassification error and number of detected clusters depend on the choice of parameters: keeping this in mind, the simulation exercise is coherently calibrated.

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

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

The mixture 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:

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

with (ϑ_i) i.i.d. as a *Beta*(5,5) scaled on $[-\pi, \pi]$ and $(\epsilon_{i,l})_{l=1}^L \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$. 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 ϵ (randomness is confined in the polar angle ϑ and in the noise ϵ). In particular, limited to the first three components, we deal with two noised semi-circumferences laying on orthogonal planes, with unitary radii, 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 interlocked 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. Concerning σ , one can choose $(0, k/3)$ to avoid that groups overlap too much due to noise variability.

With such choices, we are concentrating the process along three orthogonal directions so that the PCs tend to replicate the τ 's structure. Moreover, this setting ensures that Proposition 1 applies. In fact, the eigenvalues (of Σ) decay faster or at least equally to $\{\beta_l \text{Var}(\tau_l)\}_{l=1}^L$. Due to boundedness of $\text{Var}(\tau_l)$, it inherits the same decay type of $\{\beta_l\}_{l=1}^L$ that is exponentially (2) with $C = 1/3$.

In our simulations, we consider $n_1 = n_2 = 300$, $\sigma = \sqrt{0.005}$ and $k = 0.5$. This setting leads to have $FEV(3)$ always greater than 99%, that suggests 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 prototype regions obtained with our algorithm when $\delta = 2$ and $r = 5$ (that produced $\widehat{G} = 2$).

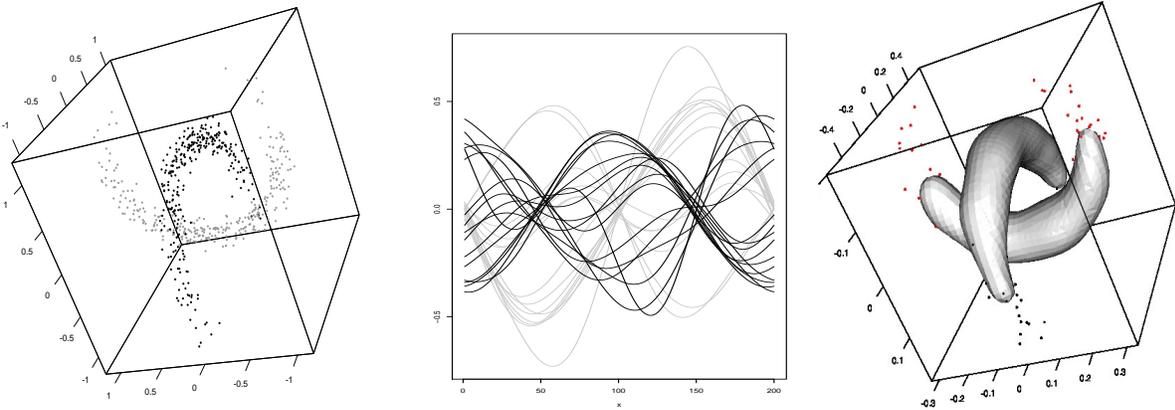


Figure 1: Left to right: simulated coefficients $(\tau_{i,l}^{(k)})_{l=1}^3$, a sample of simulated curves and upper level sets associated to the estimated modes in the factor space.

We generate 400 Monte Carlo samples according to the above setting. To each replication, we apply the SmBP clustering that returns the corresponding estimated number of clusters \widehat{G} and the misclassification error. According to FEV criterion 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 [Febrero-Bande and Oviedo de la Fuente, 2012](#)) with $G = 2$;
- (GM) the EM clustering method based on a 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 [Biernacki et al., 2006](#)).

Table 1 collects 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 \widehat{G} varying δ 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 the parameters $\delta = 1.4$ and $r = 10$ for the SmBP clustering for which the two clusters are correctly recognized at least in 90% of the 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 has to be estimated, since BIC overestimates the number of clusters.

Algorithm	Parameters		Miscl. Error		\widehat{G}		
	δ	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	0.088	0.174	2	2	2
		# 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) works when applied to a real dataset. The aim is twofold: on one hand, it shows the cognitive support that the method could bring to the studied phenomenon and, on the other hand, what kind of practical problems could occur and how to treat them.

The 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 inexpensive technique which provides an estimate of the composition of an aliment based of 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 have been widely explored for this kind of data: see for instance, [Ferraty and Vieu \(2006\)](#), [Delaigle et al. \(2012\)](#) in the supervised classification framework.

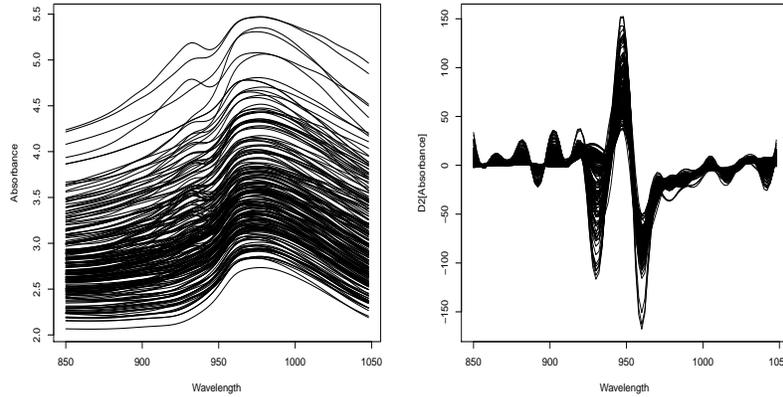


Figure 2: Tecator curves (left) and their 2-nd derivatives (right).

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 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 the same number of 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, to avoid the well-known “calibration problem” due to the presence of shifts in the curves (see [Ferraty and Vieu, 2006](#)). 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 linearly correlated: in particular, fat and water present a linear correlation coefficient equal to -0.988 , whereas the content of protein exhibits a positive correlation with fat (0.814) and a negative one with water (-0.861). This suggests using 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 kinds 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 as, that 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 find that the spectrum is rather concentrated: the first three PCs explain 98.5% of the total

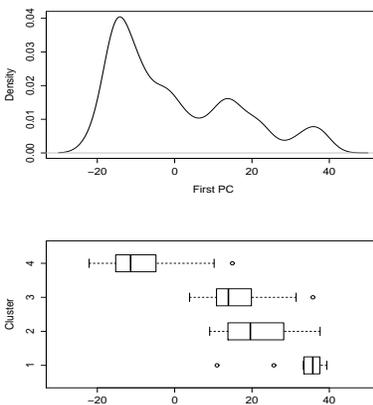


Figure 3: Density estimate of the first PC of chemical components and the same stratified according to clusters analysis carried out with $r = 5$ and $\delta = 0.2$.

variability, and this suggests to use $d = 3$ in our approximation. The selection of parameters r and δ is performed according to the maximization of CH index over a bivariate grid built with $r = 1, \dots, 7$ and δ varying from 0.2 to 1 with step 0.1. Index CH reaches its maximum for $r = 5$ and $\delta = 0.2$, to which it corresponds $k = 4$ clusters. In order to understand the appropriateness of this choice, besides the internal criterion CH we use an external criterion by computing, for each couple of parameters r, δ , the index of purity according to the three-group structure shown by the first PC which summarizes the chemical variables. It emerges that the couple (r, δ) maximizing CH provides also a high degree of purity: this fact can be appreciated by inspecting Figure 3, and it provides a heuristic support of the possibility of reproducing 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 the distribution of heat, generated in a centralized location, for entire districts through a network of insulated pipes. Due to its efficiency and to the pollution control, this system is spreading to many cities. In order to guarantee an optimal scheduling for generating heat, which allows choosing the right mix of on-line capacity, the analysis of the flows of heating demand is crucial. These flows depend mainly on two factors: an intra-daily pattern of the load demand, known as the load curve, and seasonal aspects.

To manage data from a 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, since consumers characteristics are very different according to seasons and weather conditions. In what follows, we propose an application of our clustering algorithm to data on heat consumption in Turin, a northern Italian city, where the district heating is produced through a co-generation system. The dataset has been used previously in a

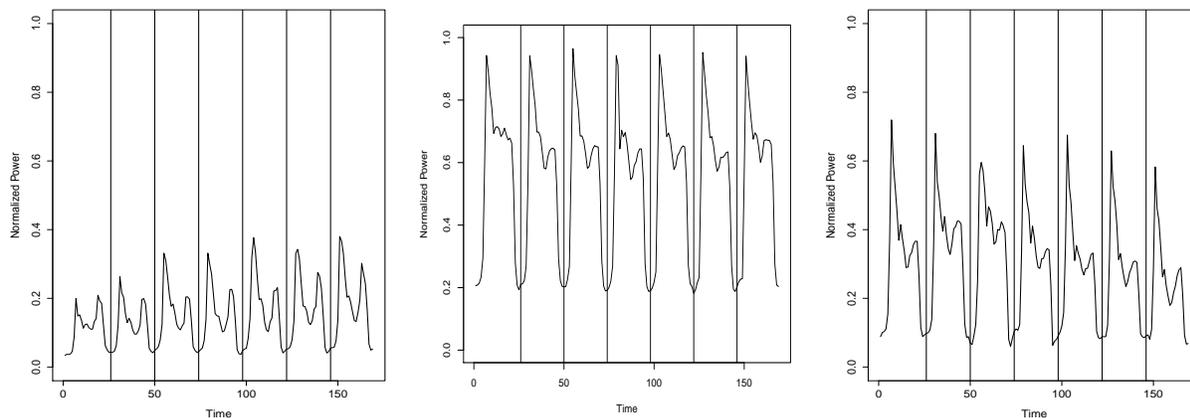


Figure 4: Demand of heat in a selected week of November 2002, January and March 2003.

forecasting context based on regression approaches (see [Goia et al., 2010](#), [Goia, 2012](#)).

The dataset consists of hourly measurements of heat consumption for residential and commercial buildings during the periods October 15 – April 20, covering the years 2001–02, 2002–03, 2003–04 and 2004–05. Due to privacy requests from the data supplier, the data have been normalized. Figure 4 displays the behaviour of the 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, as well as the seasonal evolution. Differently from electricity power demand, intra weekly differences among working days and weekends do not appear.

Taking advantage of the functional nature of the dataset, 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 of 748 curves discretized over an equispaced mesh of 24 points. Figure 5 displays the set of curves.

The first step of our procedure is to perform a functional principal components analysis. It emerges that the spectrum is rather concentrated: the first three PCs explain more than the 97% of the total variance, and thus it is sufficient to limit our analysis to $d \leq 3$. In order to provide an interpretation of the contribution of the relevant PCs, we exploit a graphical tool where we report the estimated mean curve plus and minus a suitable multiple M_j of each estimated eigenfunction: $\hat{\mu} \pm M_j \hat{\xi}_j \sqrt{\hat{\lambda}_j}$ (see e.g. [Ramsay and Silverman, 2005](#)). The results, visualized in Figure 6, show that the first eigenfunction, which does not present sign changes, 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 to be 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 to be 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.

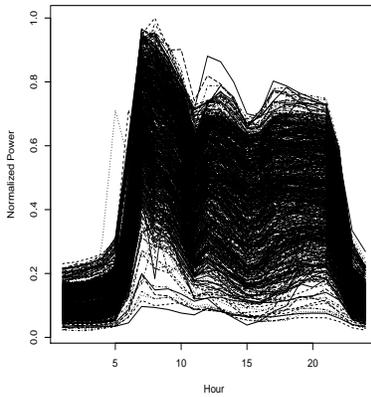


Figure 5: Normalized load curves, corresponding to daily profile.

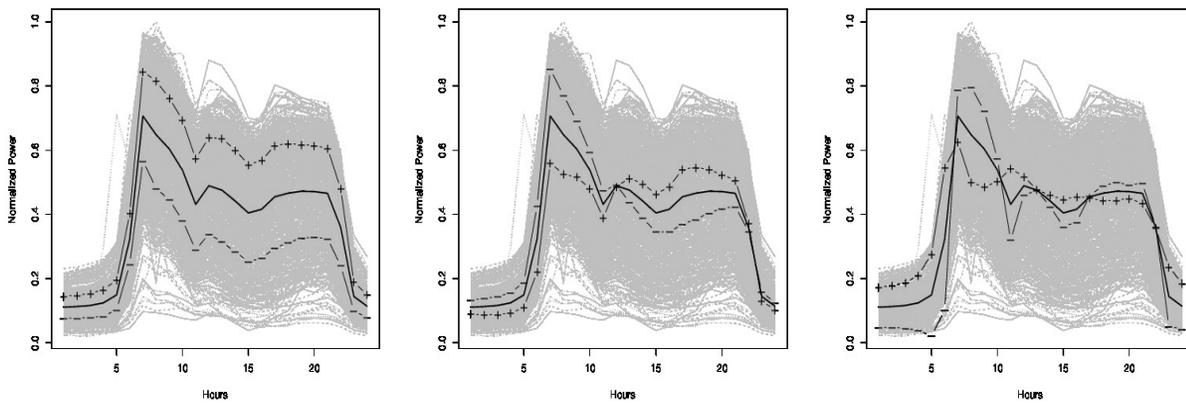


Figure 6: Contribution of the first three principal components.

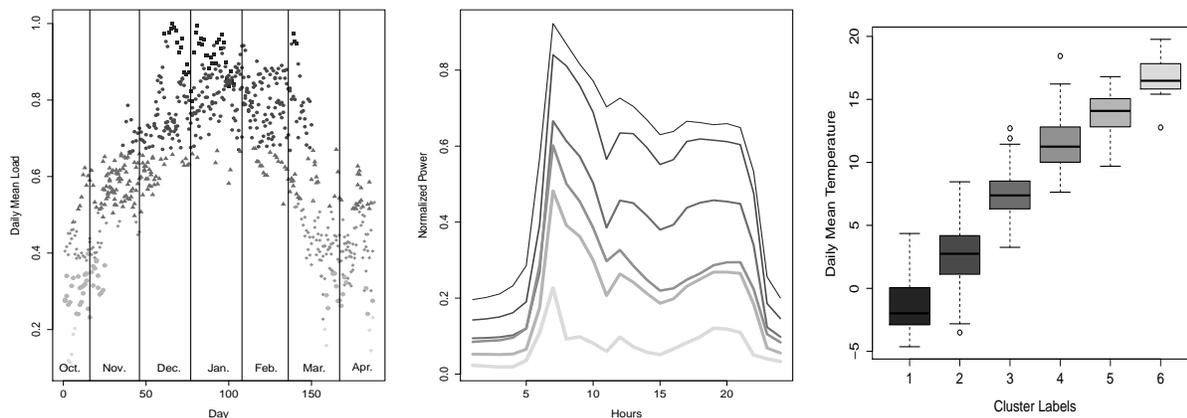


Figure 7: Calendar positioning of clusters against the daily mean load, corresponding modal curves and relationship among clusters and daily mean temperature. Along the panels, a cluster is identified by the same grey level.

In order to apply the SmbP cluster algorithm, one has to preventively select the parameters r and δ . 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, in Figure 7 we report the calendar positioning of each element of the 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. We also report the box-plots resulting after a stratification of the daily mean temperature by the cluster labels: one can note that the temperature, without presenting a multi-modal density, is one of the most important external variables that can be used in such a 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 was to set up a forecasting model, an accurate prediction of temperature would be the key to making accurate prediction in the demand of heating.

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 distinguishing 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 to the spike sorting:

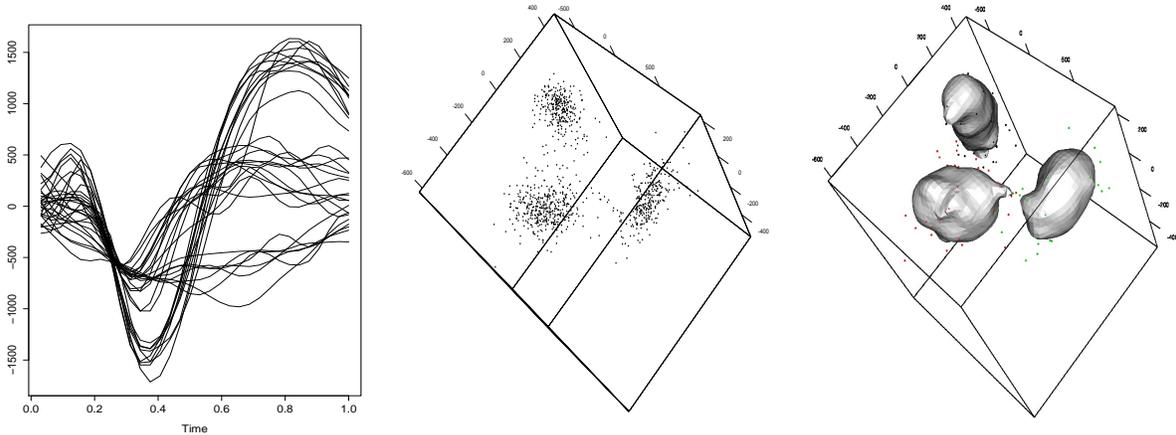


Figure 8: Neuronal Experiment (left to right): a random sample of 30 curves, 3-D Scatter plot of the first three Principal Components and the corresponding maximal level set when $r = 7$ and $\delta = 0.8$.

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 Todorova et al., 2014 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 into 1000 functional data representing the voltage of neurons versus the time, discretized over a grid of 32 equispaced time points (normalized between 0 and 1). A sample, of 30 selected randomly curves, is shown in the left panel of Figure 8. An analysis of (a larger set of) these curves can also be found in Todorova et al. (2014).

Performing the functional PCA on such a 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 the explained variance by the first three PCs is about 96.4%. Observing the 3-dimensional scatterplot of the first three PCs (see Figure 8, left panel), three clouds appear evidently.

In order to detect a good choice for the parameters r and δ , we perform our clustering 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 the obtained values leads to various admissible configurations for the couple (r, δ) , to which there always correspond $k = 3$ clusters: for instance $r = 3, \dots, 7$ combined with $\delta = 0.8, 0.9, 1$ produce the same number of clusters and the maximal CH . The middle panel of Figure 8 shows the maximal level set when one uses $r = 7$ and $\delta = 0.8$.

The result of the clustering procedure is visualized in Figure 9, where the centers of prototypes (the modal curves) and the corresponding clusters are reproduced.

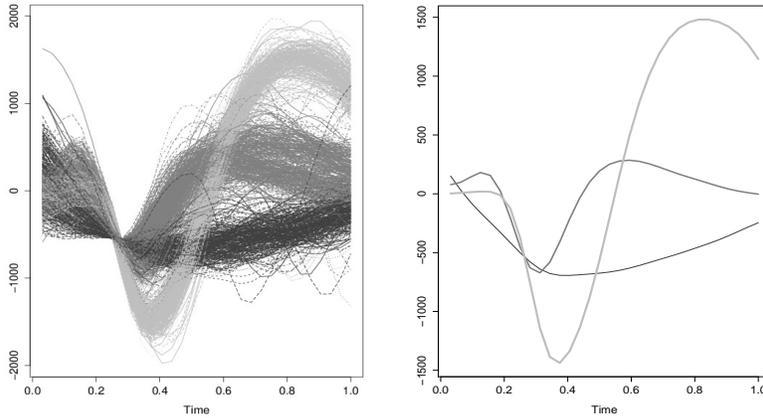


Figure 9: Neuronal Experiment: clusters and modal curves when $r = 7$ and $\delta = 0.8$.

3.3. Discriminant: simulation and real data illustration

The aim of this section is to assess the performance 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 details, for each available sample, at each iteration, $2/3$ 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.

According to the comments in Section 2.1.3, classifiers are implemented with $d = 2, \dots, 5$, where $d = 5$ is considered only to see how the results worsen due to the curse of dimensionality. The out-of-sample errors are compared with those computed by using parametric and nonparametric 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 Ferraty and Vieu, 2006). All computations are done with the software R: in particular, the competitor algorithms are taken from the package *fda.usc* (see Febrero-Bande and Oviedo de la Fuente, 2012).

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 $n_{in} = 100, 200, 300$;
- we consider both the balanced case ($\pi_1 = 1/2$), and two unbalanced cases (with $\pi_1 = 1/3$ and $\pi_1 = 1/4$);
- the vertical translation parameter k equals 0.5;

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

Table 2: Estimated mean and standard deviation (in parentheses) of misclassification error for the two-horseshoes setting. Two balanced groups: $\pi_1 = 1/2$.

- three degrees of variability are considered to model the noise around the two semi-circumferences: $\sigma = 0.05, 0.10, 0.15$, (small, medium and high variability).

After performing the discrimination exercise, one obtains the misclassification error distributions whose summary measures (mean and standard deviation) are collected in tables 2, 3 and 4. In Figure 10 such error distributions when $n_{in} = 200$, $\sigma = 0.05, 0.10, 0.15$ are reproduced.

From the tables and plots, it emerges that the SmBP classifier produces the best results, both in the balanced and unbalanced cases when $d = 3$. This is coherent with the fact that $FEV(3) \geq 99\%$ (see Section 3.1): for fixed n , increasing d further does not produce benefits; on the contrary, dimensionality causes a worsening in the classification abilities. Results with $d = 3$ 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.

3.3.2. Analysis of real datasets

In what follows we analyze the 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 Jacques and Preda (2014) in an unsupervised classification framework.

n_{in}	σ	n_{out}	$d = 2$	$d = 3$	$d = 4$	$d = 5$	GLM	k-NN	NP
100	0.05	50	0.156	0.051	0.067	0.069	0.342	0.036	0.060
	(0.045)		(0.029)	(0.038)	(0.039)	(0.067)	(0.023)	(0.039)	
	0.208		0.154	0.154	0.172	0.287	0.101	0.163	
	0.10		(0.044)	(0.048)	(0.047)	(0.045)	(0.057)	(0.043)	(0.043)
	0.15		0.241	0.212	0.225	0.252	0.366	0.175	0.221
			(0.050)	(0.048)	(0.052)	(0.061)	(0.058)	(0.048)	(0.050)
200	0.05	100	0.150	0.035	0.034	0.045	0.340	0.030	0.034
	(0.045)		(0.013)	(0.014)	(0.015)	(0.036)	(0.014)	(0.015)	
	0.211		0.133	0.153	0.175	0.342	0.129	0.146	
	0.10		(0.038)	(0.028)	(0.029)	(0.036)	(0.038)	(0.030)	(0.029)
	0.15		0.263	0.180	0.188	0.201	0.359	0.169	0.193
			(0.038)	(0.035)	(0.033)	(0.032)	(0.042)	(0.036)	(0.034)
300	0.05	150	0.137	0.036	0.039	0.051	0.327	0.039	0.043
	(0.025)		(0.010)	(0.012)	(0.015)	(0.042)	(0.015)	(0.013)	
	0.155		0.083	0.098	0.100	0.317	0.080	0.096	
	0.10		(0.028)	(0.020)	(0.021)	(0.022)	(0.031)	(0.021)	(0.021)
	0.15		0.200	0.167	0.170	0.182	0.341	0.154	0.185
			(0.031)	(0.029)	(0.028)	(0.028)	(0.038)	(0.028)	(0.031)

Table 3: Estimated mean and standard deviation (in parentheses) of misclassification error for the two-horseshoes setting. Two unbalanced groups: $\pi_1 = 1/3$.

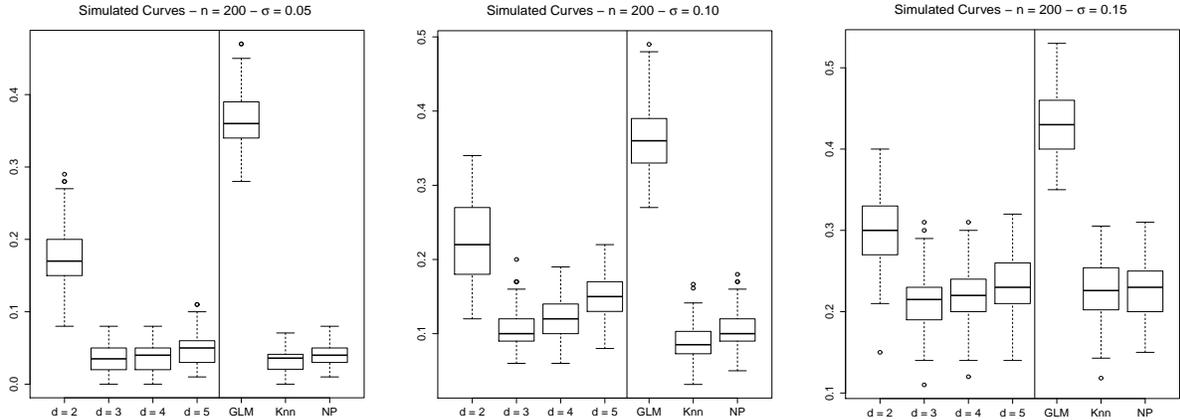


Figure 10: Distributions of misclassification errors estimated over 100 replications when $n_{in} = 200$ and $\sigma = 0.05, 0.10, 0.15$ (form left to right).

n_{in}	σ	n_{out}	$d = 2$	$d = 3$	$d = 4$	$d = 5$	GLM	k-NN	NP
100	0.05	50	0.112	0.042	0.046	0.064	0.279	0.018	0.040
			(0.048)	(0.031)	(0.031)	(0.042)	(0.070)	(0.019)	(0.035)
	0.10		0.183	0.151	0.156	0.164	0.266	0.082	0.145
			(0.044)	(0.047)	(0.047)	(0.046)	(0.054)	(0.041)	(0.048)
	0.15		0.169	0.148	0.142	0.161	0.260	0.092	0.147
			(0.042)	(0.042)	(0.047)	(0.049)	(0.056)	(0.047)	(0.040)
200	0.05	100	0.098	0.024	0.031	0.033	0.252	0.020	0.032
			(0.021)	(0.013)	(0.011)	(0.013)	(0.037)	(0.015)	(0.018)
	0.10		0.152	0.121	0.128	0.138	0.258	0.084	0.120
			(0.040)	(0.033)	(0.033)	(0.032)	(0.035)	(0.025)	(0.032)
	0.15		0.181	0.152	0.170	0.179	0.271	0.129	0.150
			(0.034)	(0.031)	(0.031)	(0.031)	(0.030)	(0.030)	(0.032)
300	0.05	150	0.119	0.013	0.021	0.033	0.271	0.018	0.020
			(0.020)	(0.009)	(0.013)	(0.016)	(0.033)	(0.008)	(0.008)
	0.10		0.153	0.127	0.131	0.128	0.265	0.107	0.139
			(0.028)	(0.024)	(0.025)	(0.025)	(0.023)	(0.024)	(0.025)
	0.15		0.174	0.169	0.181	0.194	0.256	0.156	0.181
			(0.027)	(0.028)	(0.025)	(0.027)	(0.028)	(0.024)	(0.025)

Table 4: Estimated mean and standard deviation (in parentheses) of misclassification error for the two-horseshoes setting. Two unbalanced groups: $\pi_1 = 1/4$.

The first dataset comes from the UCR Time Series Classification and Clustering website (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 [Olszewski, 2001](#) for more details).

The second dataset is the well-known Berkeley growth dataset (see [Tuddenham and Snyder, 1954](#)). 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 with a monotone smoothing method (see [Ramsay and Silverman, 2005](#)). The aim is to discriminate the curves on the basis of gender.

The third dataset, described in detail in [Lévêder et al. \(2004\)](#), 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 measurements of dough resistance 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 of medium quality and 40 of low quality. The goal of the analysis is to classify the functional dataset based on the quality of cookies.

The three datasets of curves are plotted at the top of [Figure 11](#): in the plots, the group membership of each individual (patient, child or flour) is highlighted by using different colours. The estimated conditional pseudo-densities $\hat{f}_d(\cdot|g)$, with $d = 3$, are depicted in [Figure 11](#). Finally, [Table 6](#) reports the estimated FEV in the three cases.

We apply our SmBP classification method with balanced groups with the dimension d which varies from 2 to 5. The distributions of misclassification errors for our approach and the competitors are reproduced at the bottom of [Figure 11](#). Moreover, to allow a direct comparison, [Table 5](#) 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 group structure (requiring the estimate of prior probabilities π_g) does not change the obtained results.

The results reveal how the SmBP classifier behaves with d : if, in principle, misclassification errors should reduce with d coherently with FEV and the mixture structure, in practice, larger values may amplify the noise due to bad estimation and, in the proposed examples, we find that a good compromise between approximation and dimensionality is reached with $d = 3$ (for growth curves and kneading process dataset) and $d = 4$ (for ECG dataset).

It is worth noting that our method performs rather well when compared to the other ones: despite the fact that 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: the quality of the results is explainable by observing the estimated conditional pseudo-densities, where a thin overlapping “grey zone” emerges. For the ECG dataset, the best result (for $d = 4$) is comparable with that obtained by nonparametric classifier. For what concerns the kneading process dataset, all of the proposed methods suffer from a relatively wide overlapping region of the estimated conditional pseudo-densities.

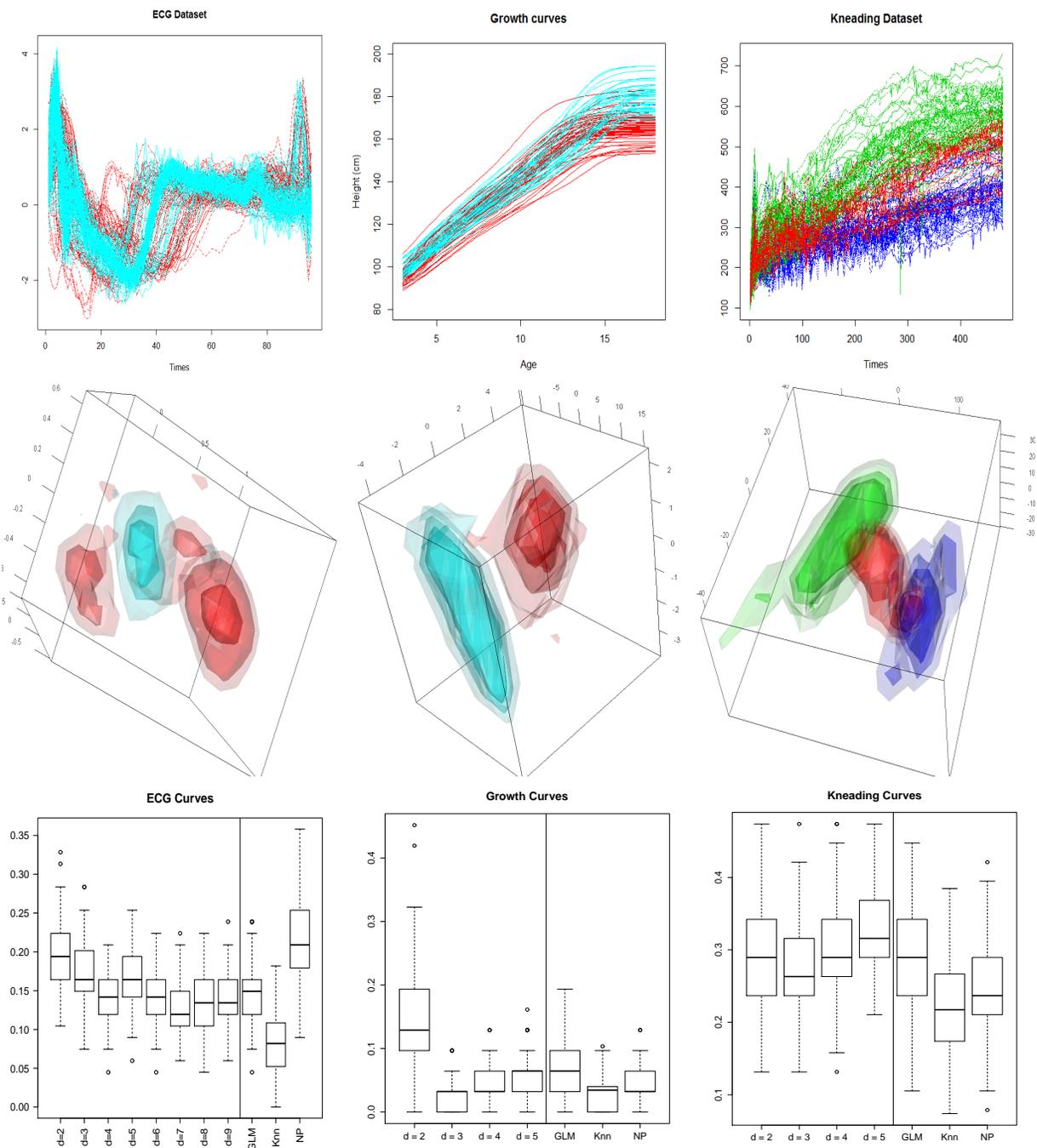


Figure 11: Top to bottom: curves of the considered dataset, conditional densities of the first three PCs scores and Out-of-sample Misclassification errors over 100 replications. Case studies (left to right): ECG (2 groups), growth curves (2 groups) and kneading process (3 groups).

	ECG		Growth Curves		Kneading Process	
SmBP	$d = 4$	0.148 (0.035)	$d = 3$	0.028 (0.024)	$d = 3$	0.273 (0.061)
GLM		0.213 (0.037)		0.075 (0.049)		0.286 (0.070)
k-NN		0.090 (0.038)		0.030 (0.027)		0.221 (0.068)
NP		0.143 (0.036)		0.041 (0.030)		0.251 (0.064)

Table 5: 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.

d	1	2	3	4	5
ECG	38.6	63.4	73.3	79.2	83.0
Growth Curves	81.7	95.9	98.7	99.5	99.8
Kneading Process	89.6	93.2	95.6	96.3	96.8

Table 6: Estimated FEV for the considered datasets.

4. Conclusions

In this paper, an unsupervised and a supervised classification method based on the concept of SmBP mixture for Hilbert-valued process have been introduced and analyzed. The novelty lies in the use of the theoretical factorization of the SmBP due to [Bongiorno and Goia \(2015\)](#) and reported in Proposition 1. Such a result introduces a surrogate-density for Hilbert-valued processes that, on the one hand, endorses a “density oriented” clustering approach for detecting the latent structure by incorporating the information on the mixture, and, on the other one, leads to define an optimal Bayes classifier in a supervised classification (discriminant) context. From a theoretical point of view, the approaches proposed here can be seen as semi-parametric: the coefficients of the Karhunen–Loève decomposition, truncated at a suitable order d , define the pseudo-density of a mixture model which is not entirely specified and is estimated nonparametrically. In this view, the detection of a reasonable dimension d which balances the trade-off between a good approximation and the curse of dimensionality is an important task. Furthermore, dealing with joint and conditional pseudo-densities leads naturally to represent them graphically, for $d \leq 3$, in order to understand the underlining mixture “structure” and to evaluate why classification errors may arise. In addition to the theoretical aspects, computational issues deserve attention as well: especially for clustering, the problem of tuning parameters is deeply considered, and known tools are implemented. The use of such standard tools has shown some shortcomings strictly related to the open long-standing problem of finding a universal optimal criteria for validating clustering procedures. The issue of efficient tuning methods requires further study, which is beyond the scope of this work.

Appendix A. Sketch of the Proofs

This Section contains a sketch of Propositions 1 and 4; for a detailed and theoretical discussion the interested reader can refer to [Bongiorno and Goia \(2015\)](#).

Appendix A.1. Sketch of proof of Proposition 1

At the beginning, fix $d \in \mathbb{N}$ and consider the quantities

$$S_1 = S_1(d, x) = \sum_{j \leq d} (\theta_j - \langle x, \xi_j \rangle)^2 \quad \text{and} \quad S = S(d, \varepsilon, x) = \frac{1}{\varepsilon^2} \sum_{j \geq d+1} (\theta_j - \langle x, \xi_j \rangle)^2.$$

These allow to rewrite the SmBP as follows

$$\varphi(\varepsilon, x) = \mathbb{P}(\|X - x\|^2 \leq \varepsilon^2) = \mathbb{P}(S_1 + \varepsilon^2 S \leq \varepsilon^2) = \int_0^1 \mathbb{P}(S_1 \leq (1-s)\varepsilon^2) dG(s),$$

where G is the cdf of S . In terms of $f_d(\cdot)$, the probability density function of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)'$, it holds

$$\mathbb{P}(S_1 \leq (1-s)\varepsilon^2) = \int_D f_d(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

with $D = \left\{ \boldsymbol{\theta} \in \mathbb{R}^d : \sum_{j \leq d} (\theta_j - x_j)^2 \leq \varepsilon^2 (1-s) \right\}$. The Taylor expansion of f_d about the point (x_1, \dots, x_d) leads to the following first order approximation

$$\varphi(\varepsilon, x) \sim f_d(x) V_d(\varepsilon) \mathbb{E} \left[(1-S)^{d/2} \mathbb{I}_{\{S \leq 1\}} \right], \quad \varepsilon \rightarrow 0.$$

According to the type of eigenvalue decay rate, it is possible to choose $d = d(\varepsilon)$ as a function of ε so that it approaches infinity when ε goes to zero and such that

$$S \rightarrow 0, \quad \text{and} \quad \mathbb{E} \left[(1-S)^{d/2} \mathbb{I}_{\{S \leq 1\}} \right] \rightarrow 1.$$

Finally, errors due to both the latter approximation and the Taylor expansion can be simultaneously controlled by exploiting the kind of eigenvalues decay rate; in particular, it turns out that the faster they decay, the smaller is the total error.

Appendix A.2. Sketch of proof of Proposition 4

Consider $f_{d,n}$, the pseudo-estimator for f_d , given by

$$f_{d,n}(\Pi_d x) = \frac{1}{n} \sum_{i=1}^n K_{H_n}(\|\Pi_d(X_i - x)\|), \quad \Pi_d x \in \mathbb{R}^d,$$

that involves the true but unknown projector operator $\Pi_d(\cdot) = \sum_{j=1}^d \xi_j \langle \xi_j, \cdot \rangle$. By the triangle inequality,

$$\mathbb{E} \left[f_d(x) - \widehat{f}_{d,n}(x) \right]^2 \leq \mathbb{E} [f_d(x) - f_{d,n}(x)]^2 + \mathbb{E} \left[f_{d,n}(x) - \widehat{f}_{d,n}(x) \right]^2. \quad (\text{A.1})$$

Regarding the first term on the right-hand side of (A.1), it is well known in the literature (see for instance Wand and Jones, 1995) that, under Assumptions (B.1)–(B.4) and taking the optimal bandwidth (6), one gets the minimax rate:

$$\mathbb{E} [f_d(x) - f_{d,n}(x)]^2 = O(n^{-2p/(2p+d)})$$

uniformly in \mathbb{R}^d . Regarding the second addend on the right-hand side of (A.1), since $H_n = h_n^2 I$, it holds

$$(nh_n^d)^2 \mathbb{E} [f_{d,n}(x) - \hat{f}_{d,n}(x)]^2 = \mathbb{E} \left[\sum_{i=1}^n \left(K \left(\frac{V_i}{h_n} \right) - K \left(\frac{\hat{V}_i}{h_n} \right) \right) \right]^2$$

where $V_i = \|\Pi_d(X_i - x)\|$, $\hat{V}_i = \|\hat{\Pi}_d(X_i - x)\|$. Consider the events $A_i = \{V_i \leq h_n\}$, $B_i = \{\hat{V}_i \leq h_n\}$; we get that

$$\begin{aligned} (nh_n^d)^2 \mathbb{E} [f_{d,n}(x) - \hat{f}_{d,n}(x)]^2 &\leq 2 \mathbb{E} \left[\sum_{i=1}^n \left(K \left(\frac{V_i}{h_n} \right) - K \left(\frac{\hat{V}_i}{h_n} \right) \right) \mathbb{I}_{A_i \cap B_i} \right]^2 \\ &\quad + 4 \mathbb{E} \left[\left(\sum_{i=1}^n K \left(\frac{V_i}{h_n} \right) \mathbb{I}_{A_i \cap \bar{B}_i} \right)^2 + \left(\sum_{i=1}^n K \left(\frac{\hat{V}_i}{h_n} \right) \mathbb{I}_{\bar{A}_i \cap B_i} \right)^2 \right]. \end{aligned}$$

Under Assumptions (B.1)–(B.4) and after some computations, it is possible to show that, for any $d \geq 1$ and as $n \rightarrow \infty$

$$\frac{1}{(nh_n^d)^2} \mathbb{E} \left[\sum_{i=1}^n \left(K \left(\frac{V_i}{h_n} \right) - K \left(\frac{\hat{V}_i}{h_n} \right) \right) \mathbb{I}_{A_i \cap B_i} \right]^2 \leq C \frac{h_n^{2(d-1)}}{n},$$

whereas

$$\frac{1}{(nh_n^d)^2} \mathbb{E} \left[\left(\sum_{i=1}^n K \left(\frac{V_i}{h_n} \right) \mathbb{I}_{A_i \cap \bar{B}_i} \right)^2 + \left(\sum_{i=1}^n K \left(\frac{\hat{V}_i}{h_n} \right) \mathbb{I}_{\bar{A}_i \cap B_i} \right)^2 \right] \leq C \left(\frac{1}{nh_n^{2(2d+1)}} \right)^{1/3},$$

and, hence,

$$\mathbb{E} [f_{d,n}(x) - \hat{f}_{d,n}(x)]^2 = O \left(\frac{h_n^{2(d-1)}}{n} \right) + O \left(\left(\frac{1}{nh_n^{2(2d+1)}} \right)^{1/3} \right). \quad (\text{A.2})$$

For any $d \geq 1$, a direct computation shows that, taking the optimal bandwidth (6) and $p > (3d + 2)/2$, the bounds in equation (A.2) are definitively negligible compared to the “optimal bound” $n^{-2p/(2p+d)}$.

Acknowledgements. The authors wish to thank an Associate Editor and two anonymous referees for their advice and suggestions that have allowed to improve the presentation of the paper. The authors even wish to thank H. H. Bock, who provided clear comments and references about density based clustering, 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. We thank F. Centrone and M. Karathanasis for their careful proofreading of the manuscript; in any case, every mistake is ascribed entirely only to the authors. 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

- Abraham, C., Biau, G., Cadre, B., 2003. Simple estimation of the mode of a multivariate density. *The Canadian Journal of Statistics. La Revue Canadienne de Statistique* 31, 23–34.
- Azzalini, A., Torelli, N., 2007. Clustering via nonparametric density estimation. *Statistics and Computing* 17, 71–80.
- Biernacki, C., Celeux, G., Govaert, G., Langrognet, F., 2006. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis* 51, 587–600.
- Bock, H.H., Ingrassia, S., Vermunt, J.K., 2013. Special issue on “Model-based clustering and classification”. *Advances in Data Analysis and Classification*. 7.
- Bock, H.H., Ingrassia, S., Vermunt, J.K., 2014. Special issue on “Model-based clustering and classification” (part 2). *Advances in Data Analysis and Classification*. 8.
- Bongiorno, E.G., Goia, A., 2015. Some insights about the small ball probability factorization for Hilbert random elements. Preprint .
- Bongiorno, E.G., Goia, A., Salinelli, E., Vieu, P. (Eds.), 2014. Contributions in infinite-dimensional statistics and related topics, Società Editrice Esculapio.
- Chen, Y.C., Genovese, C.R., Wasserman, L., 2015. Asymptotic theory for density ridges. *The Annals of Statistics* 43, 1896–1928.
- Dabo-Niang, S., Ferraty, F., Vieu, P., 2007. On the using of modal curves for radar waveforms classification. *Computational Statistics & Data Analysis* 51, 4878–4890.
- Delaigle, A., Hall, P., 2010. Defining probability density for a distribution of random functions. *The Annals of Statistics* 38, 1171–1193.
- Delaigle, A., Hall, P., Bathia, N., 2012. Componentwise classification and clustering of functional data. *Biometrika* 99, 299–313.

- Devroye, L., 1981. On the almost everywhere convergence of nonparametric regression function estimates. *The Annals of Statistics* 9, 1310–1319.
- Dubes, R.C., 1993. Cluster analysis and related issues, in: Chen, C.H., Pau, L.F., Wang, P.S.P. (Eds.), *Handbook of Pattern Recognition and Computer Vision*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, pp. 3–32.
- Duong, T., Hazelton, M.L., 2005. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics. Theory and Applications* 32, 485–506.
- Febrero-Bande, M., Oviedo de la Fuente, M., 2012. Statistical computing in functional data analysis: The R package *fda.usc*. *Journal of Statistical Software* 51, 1–28.
- Ferraty, F., Kudraszow, N., Vieu, P., 2012. Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *Journal of Nonparametric Statistics* 24, 447–464.
- Ferraty, F., Vieu, P., 2003. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis* 44, 161–173.
- Ferraty, F., Vieu, P., 2006. *Nonparametric functional data analysis*. Springer Series in Statistics, Springer, New York.
- Gasser, T., Hall, P., Presnell, B., 1998. Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* 60, 681–691.
- Gimelfarb, G., Hancock, E., Imiya, A., Kuijper, A., Kudo, M., Omachi, S., Windeatt, T., Yamada, K., 2012. *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop, SSPR&SPR 2012, Hiroshima, Japan, November 7-9, 2012. Proceedings*. Springer.
- Goia, A., 2012. A functional linear model for time series prediction with exogenous variables. *Statistics & Probability Letters* 82, 1005–1011.
- Goia, A., May, C., Fusai, G., 2010. Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting* 26, 700–711.
- Horváth, L., Kokoszka, P., 2012. *Inference for functional data with applications*. volume 200. Springer Science & Business Media.
- Jacques, J., Preda, C., 2014. Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis* 71, 92–106.
- James, G.M., Hastie, T.J., 2001. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63, 533–550.

- James, G.M., Sugar, C.A., 2003. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98, 397–408.
- Lévédér, C., Abraham, C., Cornillon, P., Matzner-Lober, E., Molinari, N., 2004. Discrimination de courbes de pétrissage. *Chimiométrie* , 37–43.
- Liu, J., Chen, Y., Maisog, J.M., Luta, G., 2010. A new point containment test algorithm based on preprocessing and determining triangles. *Computer-Aided Design* 42, 1143 – 1150.
- Olszewski, R.T., 2001. Generalized feature extraction for structural pattern recognition in time-series data. PhD Thesis.
- Ramsay, J.O., Silverman, B.W., 2005. *Functional data analysis*. Springer Series in Statistics. second ed., Springer, New York.
- Rinaldo, A., Singh, A., Nugent, R., Wasserman, L., 2012. Stability of density-based clustering. *Journal of Machine Learning Research* 13, 905–948.
- Sager, T.W., 1979. An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association* 74, 329–339.
- Shin, H., 2008. An extension of fisher’s discriminant analysis for stochastic processes. *Journal of Multivariate Analysis* 99, 1191–1216.
- Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability, Chapman & Hall, London.
- Todorova, S., Sadtler, P., Batista, A., Chase, S., Ventura, V., 2014. To sort or not to sort: the impact of spike-sorting on neural decoding performance. *Journal of neural engineering* 11, 056005.
- Tuddenham, R., Snyder, M., 1954. Physical growth of california boys and girls from birth to age 18. *California Publications on Child Development* 1, 183–364.
- Wand, M.P., Jones, M.C., 1995. Kernel smoothing. volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.
- Xu, R., Wunsch II, D., 2005. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on* 16, 645–678.