UNIVERSITÀ DEL PIEMONTE ORIENTALE

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

Quaderni DISEI

E. G. Bongiorno, A. Goia

A clustering method for Hilbert functional data based on the Small Ball Probability

Gennaio 2015

Quaderno n. 1/2015

A clustering method for Hilbert functional data based on the Small Ball Probability

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

January 27, 2015

Abstract

In the present work, motivated by the definition of a clustering method for functional data, the small–ball probability (SmBP) of a Hilbert valued process is considered. In particular, asymptotic factorizations for the SmBP are rigorously established exploiting the Karhunen–Loève expansion whose basis turns out to be the optimal one in controlling the approximation errors. In fact, as the radius of the ball tends to zero, the SmBP is asymptotically proportional to the joint density of an increasing number (with the radius) of principal components (PCs) evaluated at the center of the ball up to a factor depending only on the radius. As a consequence, the joint distribution of the first PCs provides a surrogate density of the process and, hence, in a very natural way, becomes the core in defining a density based unsupervised classification algorithm. To implement the latter, a non parametric estimator for such joint density is introduced and it is proved that used estimated PCs does not affect the rate of convergence. Finally, after a discussion on the proposed clustering algorithm, as an illustration, an application to a real dataset is provided.

Keywords. density based clustering; Hilbert functional data; Karhunen–Loève decomposition; kernel density estimate; small–ball probability.

Introduction

Cluster analysis, or unsupervised classification, is an exploratory tool encompassing a set of techniques whose scope is to reveal structural differences among data: the aim is to organize a collection of observations into “homogeneous” (in some sense) subsets through heuristic, or geometric as well as probabilistic approaches (some classical insights on such topics can be found in [27]).

In the multivariate context, an important class of clustering approaches is the so–called “density oriented” methods. The primitive idea dates back to a paper by Wishart [51]: clusters are identified by the “high density regions” and, in particular, by the connected components of the level set (at a given threshold c) of the joint distribution f of the data; i.e. the connected components of $\{f > c\}$ (see [27]). Along the years, this idea has been explored by various authors, as instance, in estimating the number of clusters [14] and/or in estimating the clusters themselves [15]. Such clustering approach is not very flexible and leads to an easy shortcoming: the number of connected components depends on the chosen threshold c that, consequently, may not catch all the structural differences among data. In contrast to the previous method (also known as "absolutely density clustering" [7, 8]), in order to avoid such drawbacks and inspiring to a "relative density clustering" (see e.g. [32]), it is possible to introduce a hierarchical family $\{G_\lambda\}_{\lambda\geq 0}$ of λ–level sets associated to the density function f; for each λ, the connected components of G_λ allow to defined a dendogram named *pruned tree* ([40, 41, 46, 47] and references therein). It is worth noting how such methods implicitly exploit local properties of the density and, hence, lead to identify clusters as "the *locally* high density regions". In this view, to look for the local maxima of the density (or *modes*) may allow to straightly identify the hierarchical family of clusters. Such ideas have been explored by the research stream called "mode hunting/seeking"; see [12, 18, 33] and references therein.

Clustering methods apply to wide range of situations, that may go beyond the multivariate framework. When observed data are curves, surfaces, images, objects or, briefly, *functional data* (see e.g. monographs [23] and [39], and [9] for recent contributions), the classical multivariate approaches can not always be directly used due to problems related to the dimensionality of the space to which the data belong, and hence a variety of specific clustering methods have been introduced (see e.g. the recent survey [29]). It's worth to point out that frequently, whatever the proposed method is, there exists an underlying multivariate strategy to which it is inspired.

The present work seeks to contribute to the literature of clustering for functional data, by proposing a new approach inspired to above illustrated family of "density oriented" methods. Among the other, the first problem one has to deal with is the definition of an object that plays the same role of the joint density distribution in the multivariate context. In fact, the main problem is that without an underlying dominant probability measure, the Radon–Nikodym derivative can not be used straightforward and, hence, a "density oriented" clustering approach can not be immediately extended to the functional context.

To manage this issue, in the functional statistic literature a concept of "surrogate density" is often considered. It is derived from the notion of small–ball probability of a random function X, briefly SmBP (see [23] and reference therein), defined as

$$\varphi(x,\varepsilon) = \mathbb{P}(\|X - x\| < \varepsilon), \tag{1}$$

with x in the same space where X takes its values, and $\varepsilon > 0$. Because its asymptotic behaviour as ε tends to zero can be interpreted as the intensity of concentration of the considered process, the SmBP limiting behaviour has been studied from a theoretical point of view (for instance, refer to the small tails/deviations theory [2, 34–36] that essentially focuses on weighted series of i.i.d. Gaussian random variables), often used in functional statistics to derive asymptotics in mode estimations (see [16, 24]), as well as in non parametric regression literature in evaluating the rate of convergence of estimators (see [23]). To the best of our knowledge, despite these efforts, an explicit general expression of $\varphi(x, \varepsilon)$ as ε goes to zero is still not available, and hence the problem to provide an approximation have become a subject widely discussed (see, for instance, [23] and references therein). In functional framework, one attractive way in studying the SmBP is to assume (as done, for instance, in [24, 25]) that, asymptotically, the dependence on x and ε is broken by means of two function Ψ and ϕ as follow

$$\varphi(x, \varepsilon) = \Psi(x) \phi(\varepsilon) + o(\phi(\varepsilon)), \qquad \varepsilon \to 0, \tag{2}$$

where $\phi(\varepsilon)$ is a kind of "volume parameter" which does not depend on x whilst Ψ, whose definition is strongly related to the choice of $\|\cdot\|$ and depends only on the center x, behaves as the *intensity* of the SmBP. It is worth noting that whenever the SmBP is a mixture, Ψ is a mixture too: it becomes the natural candidate in playing the role that the multivariate density has in the "local high density regions" finite dimensional clustering method and, consequently, the knowledge of its shape or characterization will be crucial.

The paper [17] goes along this direction providing a factorization analogous to (2). There the authors consider a functional Hilbert valued process X and develop the notion of intensity of SmBP for functional data in the space determined by the basis of the Karhunen–Loève decomposition (i.e. the principal components analysis of X). In particular, besides some technical hypothesis mainly concerning the eigenvalues of the covariance operator of X and the regularity of x, assuming that principal components are independent with positive and sufficiently smooth density function $\{\tilde{f}_j\}$, they showed that

$$\varphi(x, \varepsilon) \sim \prod_{j \leq d} \tilde{f}_j(x_j) \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2+1)} \exp\{o(d)\}, \qquad \varepsilon \to 0, \tag{3}$$

where $d = d(\varepsilon)$ is the number of considered terms of the decomposition diverging to infinity as $\varepsilon \to 0$. Now, even if the factorization in (3) leads to a very simple interpretation of the SmBP and overrides the curse of dimensionality in its estimation (indeed, it depends only on univariate marginal densities) allowing the implementation of a Gaussian mixture clustering procedure (see [29, 30]), the independence assumption and the use of the principal components basis turn to be quite restrictive. In fact, from a "intensity based" clustering procedure, independence drives in an over estimation of the number of modes, while, from a theoretical point of view, the use of the principal component basis (used to heuristically but not conventionally support the independence assumption) contrasts with (2) since the latter should be basis independent.

Thus, before tackling practical problems connected with the implementation of a clustering algorithm, we have faced the above drawbacks: a whole part of this paper, which constitutes a theoretical improvement and can be read independently from the rest, is devoted to propose more general factorizations for the SmBP, dropping the hypothesis of independence and being basis free. Concerning this latter task, it will turn out that the basis provided by the Kaharunen–Loève expansion (namely, the so–called functional principal component analysis or simply FPCA) is, in a sense that will be specified, the optimal one.

Such factorizations put forward, in a very natural way, the joint density distribution of the first d coefficients of the chosen basis (ordered with respect to the explained variance) to be the candidate for a “density” oriented clustering procedure in the functional framework, the motivating goal of this paper. To implement an algorithm, a precondition is obviously the estimation of such density: we propose a classical multivariate kernel density approach. Since, the estimation procedure involves the estimated coefficients instead of the true ones, one has to wonder if that could produce a deterioration of the rate of convergence for the density estimator: we prove that, under general conditions, this does not happen.

The set of obtained theoretical results constitute the foundation for our clustering method for functional data: the main idea is to find high intensity regions from the modes of the intensity mixture, by assigning each observation to a suitable “proximity domain” of a mode. To show how the approach works on real data, it has been applied to a well-known functional data set: problems connected with practical implementation are discussed.

Paper outline goes as follow: Section 1 introduces assumptions and main theoretical results concerning the SmBP factorizations. Section 2 provides the asymptotic theorem in estimating the joint density of the first d coefficients. Section 3 details the clustering notations and procedure while Section 4 presents an application to real data. Finally, in Section 5, all proofs are collected.

1 Notations and main theoretical results

For the sake of clarity, this section is divided in four parts. The first provides notations and some assumptions. The second part furnished a first factorization of the SmBP and some asymptotic results that, in the third part, drive to the main theoretical result of this paper. In this third part, the (covariance operator) eigenvalues role in balancing the error trade–off of the SmBP is emphasized. The latter part is devoted in weakening conditions on eigenvalues.

1.1 Notations and Regularity Assumptions

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{L}^2_{[0,1]}$ be the Hilbert space of square integrable real functions on $[0, 1]$ endowed with the inner product $\langle g, h\rangle = \int_0^1 g\,(t)\,h\,(t)\,dt$ and the induced norm $\|g\|^2 = \langle g, g\rangle$. A Random Curve (RC) X is a measurable map defined on $(\Omega, \mathcal{F})$ taking values in $(\mathcal{L}^2_{[0,1]}, \mathcal{B})$, where $\mathcal{B}$ denotes the Borel sigma–algebra induced by $\|\cdot\|$. Denote by

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

its mean function and covariance operator respectively. Although all results in Section 1.2 are independent on the choice of the Hilbert space basis, let us introduce a particular (and widely known) basis that turns to be optimal in some sense (see Remark 14). In this view, let us consider the Karhunen–Loève expansion associated to the RC X (see e.g. [10]): denoting by $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ the decreasing to zero sequence of non–negative eigenvalues and their associated orthonormal eigenfunctions of the covariance operator Σ, the RC X admits the following representation

$$X(t) = \mu_X(t) + \sum_{j=1}^{\infty} \theta_j \xi_j(t), \qquad 0 \leq t \leq 1, \tag{4}$$

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

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

In other words, PCs are uncorrelated real random variables (not necessarily independent) obtained projecting the process X on the eigenfunctions $\{\xi_j\}_{j=1}^{\infty}$ (that provides an orthonormal basis of the considered Hilbert space). This representation, taking advantage of the euclidean underline structure, isolates the manner in which the random function $X(\omega, t)$ depends upon t and upon ω.

In what follows and without loss of generality, suppose that $\mu_X = 0$. Moreover, denoting by Π_d the projector on the d–dimensional space spanned by $\{\xi_j\}_{j=1}^{d}$, assume that the first d PCs, namely $\boldsymbol{\theta} = \Pi_d X = (\theta_1, \ldots, \theta_d)'$, admit a sufficiently smooth joint strictly positive probability density, namely $\boldsymbol{\vartheta} \in \mathbb{R}^d \to f_d(\boldsymbol{\vartheta})$, so that the Taylor Formula about $\Pi_d x$ can be considered and the approximation error controlled assuming that

$$\sup_{i,j \in \{1,\ldots,d\}} \left| \frac{\partial^2 f_d}{\partial \vartheta_i \partial \vartheta_j}(\boldsymbol{\vartheta}) \right| \leq C(d), \qquad \text{for any } \boldsymbol{\vartheta} \in \mathbb{R}^d,$$

where the positive constant $C(d)$ can not be chosen uniformly in $d \in \mathbb{N}$. In fact, $C(d)$ increases (as a function of the eigenvalues) with d because, adding further high frequencies PCs whose variance λ_j decays to zero, the mass of probability concentrates more and more around the mean value of the process and the density function increases its values (as well as its second derivatives). As a consequence, to control such effect let us assume that f_d is strictly positive at $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d)' \in \mathbb{R}^d$, twice differentiable everywhere and such that, there exists a positive constant C_1 (not depending on d) such that

$$\sup_{d \in \mathbb{N}} \sup_{i,j \in \{1,\ldots,d\}} \sqrt{\lambda_i \lambda_j} \left| \frac{\partial^2 f_d}{\partial \vartheta_i \partial \vartheta_j}(\boldsymbol{\vartheta}) \right| / \left| f_d(\boldsymbol{\vartheta}) \right| \leq C_1, \qquad \text{for any } \boldsymbol{\vartheta} \in D, \tag{5}$$

where $D = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j \leq d} (\vartheta_j - x_j)^2 \leq \rho^2 \right\}$ for some $\rho \geq \varepsilon$. Remark 1 shows how (5) can be derived in an intuitive way considering a standardized version of the PCs.

Remark 1 *To better appreciate the meaning of* (5)*, note that it is equivalent in assuming, uniformly with respect to* $d \in \mathbb{N}$*, the boundedness of the second derivative of the density probability function, say* g_d*, of the random vector* $\mathbf{W} = (W_1, \ldots, W_d)'$ *defined as a deterministic translation of the component wise standardized version of* $\boldsymbol{\theta}$ *by*

$$W_j = \frac{1}{\sqrt{\lambda_j}} \langle X - x, \xi_j \rangle = \frac{\theta_j - \langle x, \xi_j \rangle}{\sqrt{\lambda_j}}.$$

In fact, since $\mathbf{W}$ *is a linear transformation of* $\boldsymbol{\theta}$*, its probability density function* g_d *is related to* f_d *by*

$$g_d(\mathbf{w}) = \left(\prod_{j\leq d}\sqrt{\lambda_j}\right) f_d\left(w_1\sqrt{\lambda_1}+x_1,\dots,w_d\sqrt{\lambda_d}+x_d\right),$$

where $x_j = \langle x, \xi_j\rangle$*,* $j = 1,\dots,d$ *and, hence, condition* (5) *is equivalent to the following one*

$$\sup_{d\in\mathbb{N}}\sup_{i,j\in\{1,\dots,d\}}\left|\frac{\partial^2 g_d}{\partial w_i \partial w_j}(\mathbf{w})\right| / |g_d(\mathbf{w})| \leq C_1, \qquad \text{for any } \mathbf{w}\in D', \tag{6}$$

where $D' = \left\{\mathbf{w}\in\mathbb{R}^d : \sum_{j\leq d} w_j^2\lambda_j \leq \rho^2\right\}$ *for some* $\rho \geq \varepsilon$*. Finally, it is worth to note that condition* (5)*, or* (6)*, is not a restrictive assumption since it includes, for instance, the case in which* X *is a Gaussian Hilbert valued process. In this case,* $\{\theta_j\}_{j=1}^{\infty}$ *are independent zero mean univariate Gaussian random variables each one with variance* $\{\lambda_j\}_{j=1}^{\infty}$ *and marginal densities* $\{\widetilde{f}_j\}_{j=1}^{\infty}$*. The joint probability density of the first d PCs is given by* $f_d(\vartheta_1,\dots,\vartheta_d) = \prod_{j\leq d}\widetilde{f}_j(\vartheta_j)$ *and it satisfies* (5)*.*

1.2 Approximation results

The aim of this section is twofold. The first goal is to provide an approximation theorem for the SmBP at a given point $x\in\mathcal{L}^2_{[0,1]}$ as ε goes to zero; it emphasizes, by means of a factorization, the trade–off between the first d principal components and the remainders, when d is fixed. Such approximation looks like (2) up to an extra factor. The study of its behaviour is the second task of the section: it turns out that the extra term becomes negligible provided that $d = d(\varepsilon)$ goes to infinity as ε tends to zero. Although the two results hold separately, technical problems arise when one tries to combine them in order to factorize the SmBP as in (2). These questions will be deepened in Section 1.3.
Let us introduce the first aim by:

Theorem 2 *Let* d *be a finite positive integer,* X *be a process as above,* $x \in \mathcal{L}^2_{[0,1]}$*. Consider the small ball probabilities of the process* X *to be defined as in* (1)*, that is*

$$\varphi(x,\varepsilon) = \mathbb{P}(\|X - x\| < \varepsilon), \qquad \text{for } \varepsilon > 0. \tag{1}$$

Suppose $\boldsymbol{\theta} = (\theta_1,\dots,\theta_d)'$ *admits probability density function* $f_d : \mathbb{R}^d \to \mathbb{R}$ *strictly positive at any* $\boldsymbol{\vartheta}\in\mathbb{R}^d$ *and satisfying* (5) *or, equivalently,* (6)*. Let*

$$\varphi_d(x,\varepsilon) = f(x_1,\dots,x_d)\frac{\varepsilon^d\pi^{d/2}}{\Gamma(d/2+1)}\mathbb{E}\left[(1-S)^{d/2}\mathbb{I}_{\{S\leq 1\}}\right], \qquad \text{for } \varepsilon > 0. \tag{7}$$

where

$$S = S(x,\varepsilon,d) = \frac{1}{\varepsilon^2}\sum_{j\geq d+1}(\theta_j - x_j)^2$$

and, $x_j = \langle x, \xi_j \rangle$, $j = 1, \ldots, d$. *Then*

$$|\varphi(x,\varepsilon) - \varphi_d(x,\varepsilon)| \leq \frac{C_1}{2}\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)\varepsilon^2\varphi_d(x,\varepsilon), \qquad \text{for } \varepsilon > 0 \tag{8}$$

and

$$\varphi(x,\varepsilon) = \varphi_d(x,\varepsilon) + o\left(\varphi_d(x,\varepsilon)\right), \qquad \text{for } \varepsilon \to 0. \tag{9}$$

Theorem 2 and, in particular, Equation (9) establishes that, for a fixed positive integer d and x in $\mathcal{L}^2_{[0,1]}$ as $\varepsilon \to 0$, the SmBP $\varphi(x,\varepsilon)$ behaves as $\varphi_d(x,\varepsilon)$. The latter is the usual approximation of the SmBP in a d–dimensional space (i.e. the probability density function of the first d PCs at $(x_1,\ldots,x_d)$ times the volume of the d–dimensional ball of radius ε) up to the extra term scale factor

$$\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]. \tag{10}$$

Above comments lead immediately to the second task of this section: to study the behaviour of this extra term. In fact, expected value (10) may be interpreted as a correction factor weighting the use of a truncated version of the process expansion (4). It is exactly equal to 1, whenever the norm is replaced with the PC–seminorm $\|\Pi_d[\cdot]\|$, recovering results in [23, Chapter 13]. Otherwise, its behaviour is strictly related to the real random variable $S(x,\varepsilon,d)$ that depends on x, ε (center and radius of the considered ball respectively) and d (number of considered PCs). On the one hand, whenever d and x are fixed, S diverges with ε tending to zero. On the other hand, if ε and x are fixed, the larger the number of d and the smaller the value of S. Hence, one may wonder if it is possible to balance these two effects (as instance, tying the behaviour of d to that of ε) in order to control (10) in (9). In Proposition 5, we will provide conditions for which, as $\varepsilon \to 0$, expectation (10) tends to 1 (hence, negligible in (9)); intuitively, this happens when S is sufficiently close to zero or, more precisely, when d increases "sufficiently fast" with respect to the rate of convergence (to zero) of ε.
Next efforts go in this direction providing the behaviour for both S and (10) as ε goes to zero. To do this, recall that $\{\xi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $\mathcal{L}^2_{[0,1]}$ and consider $x \in \mathcal{L}^2_{[0,1]}$ such that

$$\sup_{j\geq 1}(x_j^2/\lambda_j) < \infty, \tag{11}$$

that is, whenever x is sufficiently close to the process in its high–frequency part. The latter, is not an unusual condition since it is equivalent to $\sup_{j\geq 1}\mathbb{E}\left[(\theta_j - x_j)^2/\lambda_j\right] < \infty$ that was assumed, for example, in [17, Condition (4.1)] for similar purposes. Moreover, it holds whenever x belongs to the reproducing kernel Hilbert space generated by the process X:

$$RKHS(X) = \left\{x \in \mathcal{L}^2_{[0,1]} : \sum_{j\geq 1}\lambda_j^{-1}\langle x, \xi_j\rangle^2 < \infty\right\}; \tag{12}$$

e.g. [5, p.69] (roughly speaking, $x \in RKHS(X)$ only if it is "at least smooth as the covariance function", see [5, p.13]).

Proposition 3 *Assume* (11) *and choose $d = d(\varepsilon)$ so that it diverges to infinity as ε tends to zero and*

$$\sum_{j\geq d+1} \lambda_j = o(\varepsilon^2). \tag{13}$$

Then, as $\varepsilon \to 0$, $S(x, \varepsilon, d) \to 0$, where the convergence holds almost surely, in the L^1 norm and hence in probability.
Moreover, as $\varepsilon \to 0$,

$$\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]^{2/d} \to 1, \qquad \text{or,} \qquad \log\left(\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]\right) = o(d). \tag{14}$$

Remark 4 *Note that a possible choice for $d = d(\varepsilon)$ satisfying* (13) *can be, for a fixed $\delta > 0$, as follow*

$$d = \min\left\{k \in \mathbb{N} : \sum_{j\geq k+1} \lambda_j \leq \varepsilon^{2+\delta}\right\}, \qquad \text{for any } \varepsilon > 0.$$

Such a minimum is well defined since eigenvalues series is convergent.

Proposition 5 *Assume* (11) *and suppose an hyperbolic decay of the eigenvalues; that is*

$$\sum_{j\geq d+1} \lambda_j = o\left(1/d\right), \qquad \text{as } d \to \infty. \tag{15}$$

Choose $d = d(\varepsilon)$ so that it diverges to infinity as ε tends to zero and

$$d \sum_{j\geq d+1} \lambda_j = o(\varepsilon^2). \tag{16}$$

Then, as $\varepsilon \to 0$,

$$0 \leq 1 - \mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right] \leq \frac{C_2 d}{2\varepsilon^2} \sum_{j\geq d+1} \lambda_j = o(1). \tag{17}$$

Remark 6 *Note that a possible choice for $d = d(\varepsilon)$ satisfying* (16) *(and* (13) *as well) can be, for a fixed $\delta > 0$, as follow*

$$d = \min\left\{k \in \mathbb{N} : \quad k \sum_{j\geq k+1} \lambda_j \leq \varepsilon^{2+\delta}\right\}.$$

Such a minimum is well defined thanks to the eigenvalues hyperbolic decay assumption (15)*.*

Finally, it is worth to point out that the choice of $d(\varepsilon)$ in propositions 3 and 5 depends on the eigenvalues. Moreover, convergence of S does not require to state a particular decay rate for the eigenvalues while, on the contrary, the behaviour of (10) depends on (at least) the hyperbolic decay of the eigenvalues (15).

1.3 Errors trade–off and SmBP intensity: the eigenvalues role

The goal of this section is to establish which conditions on the process allow to simplify (9) by dropping the extra term (10) and, hence, to get

$$\varphi(x,\varepsilon) \sim f_d\left(x_1,\ldots,x_d\right)\frac{\varepsilon^d \pi^{d/2}}{\Gamma\left(d/2+1\right)}, \qquad \text{as } \epsilon \to 0.$$

Such result is reached combining Theorem 2 (claimed for a fixed d) and Proposition 5 (stated with d being a diverging function of ε): as announced at the beginning of the previous section, in what follow the arising problems are deepened. To do this, consider

$$|\varphi(x,\varepsilon) - f_d V_d(\varepsilon)| \leq |\varphi(x,\varepsilon) - \varphi_d(x,\varepsilon)| + |\varphi_d(x,\varepsilon) - f_d V_d(\varepsilon)|$$

where $f_d = f_d\left(x_1,\ldots,x_d\right)$, $V_d(\varepsilon) = \varepsilon^d \pi^{d/2}/\Gamma\left(d/2+1\right)$ (the volume of the d–dimensional ball with radius ε) and $\varphi_d(x,\varepsilon)$ is defined in (7). Then (8) and (17) lead to

$$\left|\frac{\varphi(x,\varepsilon)}{f_d V_d(\varepsilon)} - 1\right| \leq \frac{C_1}{2}\varepsilon^2 \sum_{j\leq d}\frac{1}{\lambda_j} + \frac{C_2}{2}\frac{d}{\varepsilon^2}\sum_{j\geq d+1}\lambda_j, \tag{18}$$

that furnishes the wished result whenever the right–hand side vanishes as ε goes to zero. In fact, for a suitable choice $d(\varepsilon)$, the term $d\varepsilon^{-2}\sum_{j\geq d+1}\lambda_j$ converges to zero as established in (16). Coherently with this choice of $d(\varepsilon)$, it must hold

$$\varepsilon^2 = o\left(\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)^{-1}\right), \qquad \text{as } \varepsilon \to 0. \tag{19}$$

Hence, plugging (19) into (16), the condition

$$d\left(\sum_{j\geq d+1}\lambda_j\right)\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right) = o\left(1\right), \qquad \text{as } d \to \infty, \tag{20}$$

highlights the announced trade–off between the errors approximation in (8) and in (17), which is strictly related to a suitable balance between the first d terms of the spectrum of the covariance operator and the remainders.

Remark 7 *As instance,* (20) *is satisfied whenever* $\lambda_j = \exp\{-\beta j^\alpha\}$ *with* $\beta > 0$ *and* $\alpha > 1$*. In this case, for any real number* $n \geq 2$*, it holds*

$$d\left(\sum_{j\geq d+1}\lambda_j\right)\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right) \leq \frac{d^n}{\lambda_d}\left(\sum_{j\geq d+1}\lambda_j\right) \to 0, \qquad \text{as } d \to \infty. \tag{21}$$

In fact, some algebra and the Bernoulli inequality (i.e. $(1+s)^r \geq 1+rs$ *for* $s \geq -1$ *and* $r \in \mathbb{R} \setminus (0,1)$*) give*

$$\frac{d^n}{\lambda_d}\left(\sum_{j\geq d+1}\lambda_j\right) = d^n\left(\sum_{j\geq 1}\exp\{\beta d^\alpha(1-(1+j/d)^\alpha)\}\right)$$
$$\leq d^n\left(\sum_{j\geq 1}\exp\{-\beta\alpha d^{\alpha-1}j\}\right).$$

Since $\exp\{-\beta\alpha d^{\alpha-1}j\} \leq (j^2 d^{n+\delta})^{-1}$ *holds eventually (with respect to d) for some positive* δ *and for each* $j \in \mathbb{N}$*,* (21) *is obtained.*

It is worth noting that (20) is a necessary condition (for the eigenvalues sequence) to guarantee that the right–hand side of (18) converges to zero (that is a sufficient condition so that Theorem 8 and Proposition 5 hold simultaneously). One may wonder, if (20) is a sufficient condition as well: in other words, if (20) allows to provide a suitable definition of $d(\varepsilon)$ so that (19) and (16) hold at the same time and, hence, the right–hand side of (18) vanishes. To clarify this aspect note that (20) implies the existence of $d_0 \in \mathbb{N}$ so that for any $d \geq d_0$

$$d\sum_{j\geq d+1}\lambda_j < \left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)^{-1}.$$

Moreover, there exist $\delta_1, \delta_2 \in (0,1)$ (depending on d) for which, for any $d \geq d_0$

$$0 \leq d\sum_{j\geq d+1}\lambda_j \leq b(d,\{\lambda_j\}_{j\geq d+1},\delta_1) < B(d,\{\lambda_j\}_{j\leq d},\delta_2) \leq \left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)^{-1}, \qquad (22)$$

where

$$b(d,\{\lambda_j\}_{j\geq d+1},\delta_1) = \left(d\sum_{j\geq d+1}\lambda_j\right)^{1-\delta_1}, \qquad B(d,\{\lambda_j\}_{j\leq d},\delta_2) = \left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)^{\delta_2-1}.$$

As instance, for a given $d \geq d_0$, fix $\delta_1 \in (0,1)$ and solve (22) with respect to δ_2, that is $\delta_2 \in (\min\{0,\beta(\delta_1)\},1)$ where $\beta(\delta_1) = 1+(1-\delta_1)\ln\left(d\sum_{j\geq d+1}\lambda_j\right)/\ln\left(\sum_{j\leq d}\lambda_j^{-1}\right)$. As a consequence, for any $\varepsilon > 0$ and for such a choice of δ_1, δ_2, the following minimum is well–defined

$$d(\varepsilon) = \min\left\{k \in \mathbb{N} : b(k,\{\lambda_j\}_{j\geq k+1},\delta_1) \leq \varepsilon^2 \leq B(k,\{\lambda_j\}_{j\leq k},\delta_2)\right\}. \qquad (23)$$

With this choice of $d(\varepsilon)$, we have that

$$\varepsilon^2 \leq B(d,\{\lambda_j\}_{j\leq d},\delta_2), \qquad \text{and} \qquad \varepsilon^{-2} \leq b(d,\{\lambda_j\}_{j\geq d+1},\delta_1)^{-1},$$

which guarantee that the right–hand side of (18) vanishes as ε goes to zero. The wished result is then reached and stated in the following theorem.

Theorem 8 *Consider hypothesis of Theorem 2 and assumptions* (11) *and* (20) *and choose* $d(\varepsilon)$ *as in* (23)*. Then, as* $\varepsilon \to 0$*,* $d \to \infty$ *and*

$$\varphi(x,\varepsilon) = f_d\left(x_1,\ldots,x_d\right) V_d(\varepsilon) + o(f_d V_d(\varepsilon)), \tag{24}$$

with $V_d(\varepsilon) = \varepsilon^d \pi^{d/2} / \Gamma\left(d/2+1\right)$*.*

Remark 9 *Note that Equation* (24) *is exactly the first order approximation that can be derived for the SmBP of a d–dimensional process (by considering the Taylor formula for the density of the multivariate process). Moreover, it is the same approximation provided for Hilbert valued processes in [23, Proof of Lemma 13.6] except for the fact that, there, authors define the SmBP for the semi–metric* $\left(\sum_{j=1}^d \langle x-y, e_j\rangle^2\right)^{1/2}$ *(instead of the Hilbert metric), where* d *is fixed and* $\{e_j\}_{j=1}^d$ *are elements of an orthonormal basis of the considered Hilbert space. In spite of these strong intuitive similarities, it is worth to point out that there are still differences as explained in [17, Section 4.3]. These are strictly related to the fact that* d *depends on* ε *and, hence, both are playing the role of scale factor or resolution level: the finer is the scale at which the SmBP is considered, the smaller is* ε *and the bigger is* d*.*

Despite the fact that in (24) a kind of intensity term appears explicitly, i.e. f_d, it is not the intensity of the SmBP as intended in (2), because of the relation between d and ε. Besides, it is not possible to extract from f_d the dependence on ε unless additional hypothesis on the process. Remark 10 provides a family of processes for which (24) reduces to (2).

Remark 10 *Assume that* X *is a Gaussian process. Then, under the hypothesis of Theorem 8, we have*

$$\varphi(x,\varepsilon) \sim \exp\left\{-\frac{1}{2}\sum_{j\le d}\frac{x_j^2}{\lambda_j}\right\}\frac{\varepsilon^d}{2^{d/2}\Gamma\left(d/2+1\right)\prod_{j\le d}\sqrt{\lambda_j}}.$$

The latter is asymptotically equivalent to (2) *with*

$$\Psi(x) = \exp\left\{-\frac{1}{2}\sum_{j=1}^{\infty}\frac{x_j^2}{\lambda_j}\right\}, \qquad \text{for any } x \in \mathcal{L}^2_{[0,1]}$$

and it is not null if and only if x *belongs to* $RKHS(X)$*, see* (12)*. The same arguments apply, whenever the PCs are independent each one with density belonging to a subfamily of the exponential power (or generalized normal) distribution (see e.g. [11]), that is proportional to* $\exp\left\{-\left(|x_j|/\sqrt{\lambda_j}\right)^p\right\}$*, with* $p > 0$*. In this case,* Ψ *is well-defined by*

$$\Psi(x) = \exp\left\{-\frac{1}{2}\sum_{j=1}^{\infty}\left(\frac{|x_j|}{\sqrt{\lambda_j}}\right)^p\right\}, \qquad \text{for any } x \in \mathcal{L}^2_{[0,1]}$$

and, it is not null whenever x is in

$$H(p) = \left\{ x \in \mathcal{L}^2_{[0,1]} : \sum_{j=1}^{\infty} \left(|x_j| / \sqrt{\lambda_j} \right)^p < \infty \right\}.$$

Moreover, it holds $H(q) \subseteq RKHS(X) \subseteq H(p)$ with $0 < q \leq 2 \leq p$.

1.4 Weakening the eigenvalues decay rate

Theorem 8 and, in particular, the form of the right–hand side of (24) depend strongly on the emphasized eigenvalues trade-off provided by (20). On the other hand, exploiting the asymptotic behaviour of V_d, similar results can be obtained weakening such eigenvalues decay rate. In particular, consider the following decays and their relationships:

Lemma 11 *Consider the eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ and the following decay rates*

- *"hyper–exponential":*

$$d \left(\sum_{j \geq d+1} \lambda_j \right) \left(\sum_{j \leq d} \frac{1}{\lambda_j} \right) = o(1), \qquad \text{as } d \to \infty. \tag{20}$$

- *"super–exponential":*

$$\lambda_{d+1}/\lambda_d \to 0, \qquad \text{as} \qquad d \to \infty \tag{25}$$

  or, equivalently, $\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j \to 0$ (as $d \to \infty$).

- *"exponential": there exists a positive constant C so that*

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j < C, \qquad \text{for any } d \in \mathbb{N}. \tag{26}$$

The following implications hold (20) $\Rightarrow$ (25) $\Rightarrow$ (26).

It is easy to show that vice versa does not hold: for instance, consider $\lambda_{d+1} = (\ln(d+1))^{-(d+1)}$ with $d \geq 1$ to prove that (20) $\not\Leftarrow$ (25), while Remark 12 leads to (25)$\not\Leftarrow$ (26).

Remark 12 *Suppose that $\lambda_j = \exp\{-\beta j\}$ for $j \geq 1$ and some $\beta > 0$. Then $\{\lambda_j\}_{j \geq 1}$ has exponential decay according to (26). In fact, since $\lambda_{j+1}/\lambda_j = e^{-\beta}$, we have*

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j = \sum_{j \geq d+1} \frac{\lambda_{d+1}}{\lambda_d} \cdots \frac{\lambda_j}{\lambda_{j-1}} = \sum_{j \geq d+1} e^{-\beta(j-d)} = \sum_{j \geq 1} e^{-\beta j} = C.$$

In this case, Theorem 13 applies instead of Theorem 8.

The following Theorem holds.

Theorem 13 *Consider hypothesis of Theorem 2 and assumption* (11). *Thus, as* ε *tends to zero, it is possible to choose* $d(\varepsilon)$ *diverging to infinity so that:*

- *in the super–exponential case*

$$\varphi(x,\varepsilon) \sim f_d\,(x_1,\ldots,x_d)\exp\left\{\frac{1}{2}d\left[\log(2\pi e\varepsilon^2)-\log(d)+o(1)\right]\right\}. \qquad (27)$$

- *in the exponential case*

$$\varphi(x,\varepsilon) \sim f_d\,(x_1,\ldots,x_d)\exp\left\{\frac{1}{2}d\left[\log(2\pi e\varepsilon^2)-\log(d)+\delta(d,\alpha)\right]\right\}, \qquad (28)$$

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

Proof. Given results in Theorem 2, thesis holds using same arguments as in [17, Proof of Theorem 4.2.]: the idea is to combine together (14), the Stirling expansion of the Gamma function in V_d and the (super–)exponential eigenvalues decay. ■

In other words, as d diverges to infinity, for slower than (20) eigenvalues decay arguments of Theorem 8 do not always apply and modifications of the asymptotic approximation $\varphi(x,\varepsilon) \sim f_d(x)V_d(\varepsilon)$ are required. In fact, if the decay is not too slow, such as in the (super–)exponential cases, the fast decay to zero of $V_d(\varepsilon)$ compensate the shortcoming and drive the approximation to the slightly different form $\varphi(x,\varepsilon) \sim f_d(x)\phi(\varepsilon)$ (see Theorem 13). For what concern slower rates of convergence the theoretical problem is still open, even if a pragmatic point of view may be adopted since "nonparametric methods, where the notion of a functional–data density is typically employed, have much lower performance and so are less attractive and less likely to be used" (see [17]).
Moreover, note that in the (super–)exponential setting Remark 10 still holds with straightforward modifications.

Remark 14 *As already pointed out in the introduction, the use of the principal component basis contrasts with* (2) *since the latter should be basis independent. From a theoretical point of view, results presented so far still hold whenever the Karhunen–Loève basis is replaced by an orthonormal basis of the Hilbert space, say* $\{\xi_j\}_{j=1}^{\infty}$*, for which the sequence* $\{\lambda_j\}_{j=1}^{\infty}$ *has an (hyper– or super– or) exponential decay (see Theorem 13), where* $\mathbb{E}[\theta_j^2]=\lambda_j$ *and* $\theta_j=\langle X-\mu_X,\xi_j\rangle$*; in other words, the variance along the directions* ξ_j *should decay sufficiently fast. In this view, by construction, the Karhunen–Loève expansion associated to the RC* X *provides for* $\{\lambda_j\}_{j=1}^{\infty}$ *the best decay rate and then it is, in this sense, the optimal one.*

2 Estimation of the joint distribution of the principal components

To make the factorization results (24), (27), (28) usable for practical purposes, one has to introduce an estimate of the density f_d, with $d \geq 1$ integer, from a sample of

RCs $\{X_i, i = 1, \ldots, n\}$ which we suppose i.i.d. as the RC X. From a theoretical point of view, if the sequence of eigenvalues $\{\xi_j\}_{j=1}^{\infty}$ was known, one should consider the empirical version of the vector of the first d principal components $\theta_i = (\theta_{1i}, \ldots, \theta_{di})' \in \mathbb{R}^d$, with $\theta_{ji} = \langle X_i - \mathbb{E}[X_i], \xi_j \rangle$, and then, with an abuse of notations (we drop the dependence on d in the density estimators and we use x both as an element of the Hilbert space and as its d–dimensional projection in $\mathbb{R}^d$ since its meaning is clear from the context), to introduce the classical kernel density estimate of f_d as follows:

$$f_{d,n}(\Pi_d x) = f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{H_n}(\|\Pi_d (X_i - x)\|) \tag{29}$$

where $K_{H_n}(\mathbf{u}) = \det(H_n)^{-1/2} K\left(H_n^{-1/2}\mathbf{u}\right)$, K is a kernel function and, $H_n = H_{nd}$ is a symmetric semi-definite positive $d \times d$ matrix. In practice (29) define only a pseudo-estimate for f_d: indeed, the covariance operator Σ and then the sequence $\{\xi_j\}$ are unknown. Thus, to operationalize these pseudo-estimates it is necessary to consider the estimates $\widehat{\theta}_i$ and $\widehat{\Pi}_d$ of θ_i and Π_d respectively. In this view, consider

$$\overline{X}_n(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t), \qquad \text{and} \qquad \widehat{\Sigma}_n[\cdot] = \frac{1}{n}\sum_{i=1}^{n} \langle X_i - \overline{X}_n, \cdot \rangle (X_i - \overline{X}_n)$$

being the sample versions of μ_X and Σ respectively. The eigenelements of $\widehat{\Sigma}_n$ provide an estimation for $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ of Σ, as well as $\langle X_i - \overline{X}_n, \widehat{\xi}_j \rangle = \widehat{\theta}_{ji}$ estimates θ_j (the asymptotic behaviour of these estimators has been widely studied; e.g. [10]). Thus, plugging the estimates of the principal components (or of the eigen-projectors) in (29), we get (with some abuse of notations) the kernel density estimator:

$$\widehat{f}_{d,n}\left(\widehat{\Pi}_d x\right) = \widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{H_n}\left(\left\|\widehat{\Pi}_d (X_i - x)\right\|\right), \qquad \widehat{\Pi}_d x \in \mathbb{R}^d. \tag{30}$$

If, from a computational point of view, such replacement is a natural way to manage the problem in practice, one may wonder if it can influence the rate of convergence of the kernel estimator, or, in other words, if using $\widehat{f}_n$ instead of f_n has no effect on this rate. To answer this question, we study the behaviour of $\mathbb{E}\left[f_d(x) - \widehat{f}_n(x)\right]^2$ as n goes to infinity. For the sake of simplicity, we confine the study to the special case $H_n = h_n^2 I$, and we suppose that the following assumptions (that are standard in the nonparametric framework) occurred:

(A1) the density $f_d(x)$ is positive and p times differentiable at $x \in \mathbb{R}^d$;

(A2) the sequence of windows h_n is such that:

$$h_n \to 0 \qquad \text{and} \qquad \frac{nh_n^d}{\log n} \to \infty \qquad \text{as } n \to \infty;$$

(A3) the kernel K is Lipschitz, bounded, integrable density function with compact support $[0, 1]$;

(A4) the process X is bounded, i.e. there exits a positive constant M such that:

$$\|X\| \leq M < \infty \quad \text{a.s..}$$

Observe firstly that one can control the quadratic mean under study by intercalating the pseudo-estimator (30); in fact, thanks to the triangle inequality

$$\mathbb{E}\left[f_d(x) - \widehat{f}_n(x)\right]^2 \leq \mathbb{E}\left[f_d(x) - f_n(x)\right]^2 + \mathbb{E}\left[f_n(x) - \widehat{f}_n(x)\right]^2. \tag{31}$$

About the first term on the right–hand side of (31), it is well known in the literature (see for instance [49]) that under Assumptions (A1)–(A4), and taking the optimal bandwidth

$$c_1 n^{-\frac{1}{2p+d}} \leq h_n \leq c_2 n^{-\frac{1}{2p+d}} \tag{32}$$

where c_1 and c_2 are two positive constants, one gets the minimax rate:

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

uniformly in $\mathbb{R}^d$. Therefore, it is enough to control the second addend on the right–hand side of (31).

The following proposition, whose proof can be found in Section 5, states that, assuming a suitable degree of regularity for the density f_d depending on d, the rate of convergence in quadratic mean of $\widehat{f}_n(x)$ towards $f_n(x)$ is negligible with respect to the one of $f_n(x)$ to $f_d(x)$. Thus, to use the estimated principal components instead of the empirical ones does not affect the rate of convergence.

Proposition 15 *Assume (A1)–(A4) with $p > (3d+2)/2$ and consider the optimal bandwidth* (32). *Thus, as n goes to infinity,*

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

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

3 Small–ball probability based clustering

This section is devoted in defining a clustering procedure, whose aim is to detect the presence of distinct groups in a dataset and assign group labels to the observations, in a functional framework that takes advantage of the asymptotic factorization results provided by theorems 8 and 13. In particular, the clustering procedure here implemented is based on the premise that the observations may be regarded as a sample from some underlying unknown mixture of probability measures on feature space and that groups correspond to modes of its associated SmBP intensity. The goal then is to find the modes of the intensity mixture and assign each observation to the "proximity domain" of a mode.

In this view and for the sake of simplicity, this section is divided in two parts. The first one supplies the framework and the necessary notations generalizing the classical concept of mixture in the case of the small–ball probability. The second one concerns the definition, in the functional setting, of the clustering algorithm.

3.1 Small–ball probability mixture

Consider X as in Section 1. Suppose that Ω is partitioned in G (unknown) sub-sets Ω_g and let Y be a real r.v. defined by

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

Consider the conditioned SmBP

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

and its asymptotic behaviour as given by Theorem 8 (or 13) provided it holds:

$$\varphi(x, \varepsilon | g) \sim f_d(x|g)\, \phi(\varepsilon), \qquad \text{as } \varepsilon \to 0, \qquad g = 1, \ldots, G$$

where $f_d(x|g)$ is the conditioned joint density of the first d PCs $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)'$ of X while $\phi(\varepsilon)$ is a "volume" parameter. Thanks to the total probability law it follows

$$\begin{aligned} \varphi(x, \varepsilon) =& \mathbb{P}(\|X - x\| < \varepsilon) = \sum_{g=1}^{G} \pi_g \varphi(x, \varepsilon | g) \\ \sim& \phi(\varepsilon) \sum_{g=1}^{G} \pi_g f_d(x|g), \quad \text{as } \varepsilon \to 0, \end{aligned} \tag{33}$$

that provides the small–ball mixture representation in terms of the conditional probability joint distribution of the first d PCs with mixture coefficients π_g. Combining (33) with the asymptotic behaviour of $\varphi(x, \varepsilon)$ (provided by Theorem 8 or 13), we have $f_d(x) = \sum_{g=1}^{G} \pi_g f_d(x|g)$. In other words, a mixture model for the SmBP can be seen as a mixture model for f_d with same weights π_g, $g = 1, \ldots, G$ and, therefore in such a context, arguments on f_d apply naturally on $\varphi(x, \varepsilon)$ and vice versa.

From the clustering point of view, typical requests concern the facts that each term of the mixture is unimodal and that all the modes are distinguishable from the mixture. In particular, from now on suppose that each conditioned SmBP is unimodal, in the sense that, for any $g = 1, \ldots, G$, $f_d(x|g)$ is unimodal for any d and, assume that, for d large enough, all the G modes of the mixture are distinguishable (conditions to have this are illustrated for instance in the bivariate Gaussian case in [4, 22, 28, 42]); i.e. there exists $d_0 \in \mathbb{N}$ such that, for any $d \geq d_0$, f_d has precisely G modes, say $m_{d,g}$ with $g = 1, \ldots, G$.

Remark 16 *Whenever the principal components $(\theta_j|g)$ are mutually independent, the joint distribution can be factorized. Moreover if $f_d(\cdot|g)$ are specified, the approximation belongs to a full parametric, or model based, approach. For instance, if $(X|g)$ is a Gaussian random process with mean μ_g and covariance operator Σ_g, the principal components $(\theta_j|g)$ are mutually independent with centered Gaussian distribution and variances $(\lambda_j|g)$ and consequently the SmBP mixture is the same as in [29] and [30]. In the latter, authors used a maximum likelihood and expectation maximization approach to identify the distribution parameters and hence the mixture.*

3.2 Algorithm

In view of exposed results, we can now define the following unsupervised classification algorithm:

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 $\widehat{m}_{d,g}$, $g = 1, \ldots, \widehat{G}$.
4. *Finding Prototypes*: for each g in $\{1, \ldots, \widehat{G}\}$, the g-th "prototypes" group is formed by those X_i whose estimated PCs belong to the largest connected iso–surface of $\widehat{f}_{d,n}$ that contains only the maximum $\widehat{m}_{d,g}$.
5. (Optionally) Classify the unlabelled X_i with the $\widehat{G}$ prototypes groups by means of a k–NN procedure.

The remain part of this section is devoted in exploring those algorithm issues needing some attention.

Dimension tuning As noticed in Remark 9, factorization theorems tie the values of d to those of ε displaying that the notion of density changes as scale becomes finer. This means that rather than estimate f_d for a fixed value of d (that means to approximate SmBP for fixed values of ε), one should have to consider different values of d each one providing further insights on the process itself. Anyway, from a practical point of view, attention must be paid in tuning the feasible values of d since it should be small enough to avoid the well–known "curse of dimensionality" in estimating nonparametrically f_d but it should be sufficiently large to guarantee a good Karhunen–Loève approximation. To take care of the latter, d can be chosen so that the fraction of explained variance of the correspondent first d components is greater than a fixed threshold c, in general larger than $c \geq 0.9$; i.e. FEV criterion: $\sum_{j \leq d} \lambda_j / \sum_{j \geq 1} \lambda_j \geq c \geq 0.9$. With such criterion, the (hyper– or super– or) exponential eigenvalues decay likely provides small values of d ridding the algorithm of the curse of dimensionality as well. In this view, the distribution free approach (provided by a nonparametric estimation of f_d) is more general than a full parametric one; see also Remark 16.
Moreover, it is worth noticing the case of a finite dimensional process that clearly satisfies (20) and for which literature has provided so far different test procedure strategies in determining the correct dimension; e.g. [26].

Remark 17 *The algorithm here presented take advantage of some visualization tools already developed in graphics libraries, such as contour lines that are typically available up to three dimensions. Such cases are the most interesting from the theoretical point of view, indeed, if the FEV criterion was fulfilled only for d greater than three, it might mean that the eigenvalues decay is not sufficiently fast to guarantee the applicability of the factorization theorems 8 and 13 from which the algorithm is derived. Anyway, whenever d is greater than three, the interested reader can consider to slightly modify the algorithm by adapting similar multivariate density based clustering procedure as the ones described in [31, 47] where authors mix single-linkage procedure with non* *parametric estimations of the density, or in [3, 38] that exploit Delaunay diagrams properties; in any case, all of them explicitly (if using kernel estimation) or implicitly (if using k–NN) suffer from high dimensions settings.*

Nonparametric estimation In order to estimate f_d, consider $\widehat{f}_{d,n}$ defined as in (30) that is the classical multivariate Nadaraya–Watson density estimator and, computed on a grid of the d–dimensional factor space. An important task in the nonparametric estimation regards the bandwidth selection. In fact, even assuming independence for PCs (as in [17]), the kernel density approach require a selection procedure for the bandwidth matrix (which identifies an ellipse whose axes are, in general, not parallel to main directions) since the estimated number of clusters depends on it: the larger is $|H|$, the "smoother" is $\widehat{f}_{d,n}$, the smaller is the number of modes. In [19–21] authors show that a diagonal bandwidth matrix (identifying an ellipse whose axes are along the orthogonal directions decomposing the process) is most useful when there is large probability mass oriented along the coordinate directions (as is in the case of PCs with fast eigenvalues decay rate) as well as in mixture detections. Thus, unless further considerations on the process/sample, an optimal choice is to consider a diagonal bandwidth matrix whose non–null entries are the univariate bandwidth provided by [45, p.48]. Anyway, as usual in the nonparametric density estimation framework, different choices for the bandwidth may be considered in order to catch different phenomenon scales at the chosen resolution level d.

Remark 18 *Alternatively to the nonparametric approach, a k–NN density estimation could be considered. In this case, after having paid attention to the tuning of k (with remarks similar to those above concerning the bandwidth selection), one can take advantage of the computationally simply estimator. Nevertheless, whenever the empirical PCs scores are considered, the consistency of such estimator as well as its rate of convergence should be studied and, as usual in a k–NN framework, difficulties may arise in handling the randomness of the radius of the k-th nearest neighbourhood.*

Modes and Prototypes Density estimates tend to have spurious modes caused by sampling variability. The varying of the bandwidth matrix as explained before may help in detecting them. Another way is to look for $\widehat{f}_{d,n}$ local maxima whose estimate density is greater than all other points in the grid within a fixed distance, say r (the larger is r the less likely the point is recognized as a mode). (Alternatively, in a multivariate framework, modes can be estimated as in [1].) At this point the algorithm provides $\widehat{G}$ estimated modes, namely $\widehat{m}_{d,g}$, $g = 1, \ldots, \widehat{G}$. For each g in $\{1, \ldots, \widehat{G}\}$, the g–th "prototypes" group is formed by those X_i in the sample whose estimated PCs belong to the largest connected iso–surface of $\widehat{f}_{d,n}$ that contains only the local maximum $\widehat{m}_{d,g}$.

Classify the unlabelled sample curves At this stage, even thought the main goal of clustering is reached (indeed modes and prototypes already display the structural changes among data that cluster procedure is looking for), one may be interested in classify even those curves of the sample that, at the previous stage, are not classified

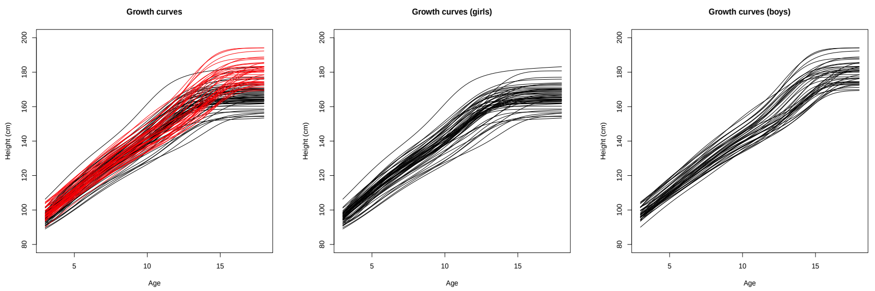

Figure 1: Smoothed Growth Curves: whole dataset (left), girls only (center), boys only (right).

as prototypes. To do this, the algorithm adopted a k–NN (with $k = 1$) procedure to label such sample curves: each unclassified curve is labelled with the same label of the nearest group.

Remark 19 *It is worth to note that once modes are computed, prototypes (i.e. those curves whose PCs belong to some largest connected iso–surface containing only one mode) are equivalently classified by means of a mode–seeking algorithm: shift data PCs according to the estimated $\widehat{f}_{d,n}$ density gradient till the shift is sufficiently small, i.e. when the shifted PCs are close to some mode (see, for instance, [8, 12, 18, 33, 50] and references therein). In view of this similarity and alternatively to the k-NN strategy to classified the unlabelled sample curves, one may shift data PCs according to the estimated $\widehat{f}_{d,n}$ density gradient till they belong to the largest connected component associated to a mode from which the unclassified curves inherits the label.*
Another classification strategy for the unlabelled curves can be derived adapting method proposed in [3, 38].

4 Real data illustration

In this section we illustrate an application of the clustering algorithm proposed in Section 3 to the well-known Berkeley growth dataset (see [48]), in order to show how the method works in practice, the cognitive support on the studied phenomenon it could bring, and what kind of practical problems could occur.

Dataset illustration The dataset 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 we use in the subsequent analysis, the original raw data are preprocessed: a monotone smoothing method was fitted to each individual set of discretized data (for more details see [39]). The final sample of curves is visualized in Figure 1.

This dataset is a benchmark in the functional data analysis framework and it has been used, for instance, in regression modelization (see for instance, [13]), supervised classification and clustering (see for instance, [43] and [30]). In this latter the aim of

j	$\widehat{\lambda}_j$	$\sum_{i\leq j}\widehat{\lambda}_i/\sum_i\widehat{\lambda}_i$
1	36.2977	81.67
2	6.3159	95.88
3	1.2440	98.67
4	0.3795	99.53
5	0.1086	99.77
6	0.0568	99.90

Table 1: Estimated first six eigenvalues $\widehat{\lambda}_j$ and correspondent explained variance.

the exercise is to retrieve the gender of subjects: sex is an hidden variable which is used a posteriori to assess clustering performances. In such literature, the problem to carry out a registration step before clustering, in order to remove amplitude and phase variabilities of curves, is debated. In fact, there are two opposing positions: in some works a registration step is recommended and even merged in the clustering algorithm (see e.g. [37] and [43]), in some others, it is believed that amplitude and phase variation of curves are typical features that allow to characterize the clusters, and then their removal could lead to a deterioration of the performances of the algorithm (see [30]). Since we agree with this second line of thought (also supported by the empirical evidences on clustering abilities), we do not perform any data registration.

Functional Principal Component Analysis The first step of our application is to perform a principal components analysis. Table 1 collects the first six estimated eigenvalues of the (estimated) covariance operator. One can note that the spectrum is rather concentrate: the first three PCs explains more than 98% of the total variance, and thus, according to Remark 17 it is sufficient limit our analysis to $d \leq 3$. About the behaviour of the estimated eigenvalues, one may wonder if their decay is fast enough to guarantee the factorization of the SmBP as provided by theorems 8 and 13 and, hence, to justify the use of the joint density of the considered scores. In fact, Figure 2 heuristically shows that eigenvalues decay is of type $\lambda_j = e^{-\beta j}$ and hence exponential as well, see Remark 12.

To conclude the first step in our analysis, we provide an interpretation of the contribution of the first three CPs: a useful graphical tool is, besides the representation of the estimated eigenfunctions, to plot the estimated mean curve plus and minus a suitable multiple of each estimated eigenfunction (see e.g. [39]). To obtain a good illustration and, at the same time, taking the weight of each PC into account, we displayed $\widehat{\mu} \pm 3/2\sqrt{\widehat{\lambda}_j}\widehat{\xi}_j$ (see Figure 3). The first eigenfunction, which does not present sign change, appears monotonic for ages up to 15 and almost constant for greater ages: it describes a "fan effect" in the first part of life and represents only a vertical shift in the remainder. As a consequence, the scores $\widehat{\theta}_1$ are highly correlated with the integral of the growth curves. The second and the third eigenfunctions are connected with the pubertal spurt, and then they appear very important in the clustering exercise.

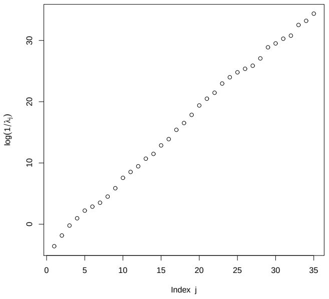

Figure 2: Empirical evidence on exponential eigenvalue decay: index j against $\log\left(1/\widehat{\lambda}_j\right)$.

Clustering In what follow, we illustrate the results of the clustering method for the case studied: the analysis is performed varying the main parameters, that is, the dimension d and the bandwidth matrix H in the kernel density estimation. About d, according to the above comments on the behaviour of the estimated spectrum, we use $d = 2$ and $d = 3$. According to Section 3, the matrix H is chosen diagonal and, to better appreciate the role of the smoothing parameters in revealing an underlying structure, we apply a shrinkage factor $\delta \in (0, 1)$. In this way, δH may be tuned to allowing us to pass from a macro–scale to a micro–scale analysis (from large to small δ).

Consider first the case $d = 2$. If one uses the bandwidth selected automatically as illustrated in Section 3, one obtains two prototypes groups to which the k–NN procedure connects the remaining points to create two clusters that overlap (almost perfectly) the groups corresponding to the boys and the girls: only 1 male is assigned to the group containing mainly female, and 4 females are assigned to the cluster of males, with a correct classification rates of 94.6% of retrieved subjects respect to the latent variable sex. When one applies a coefficient δ smaller than 1, the micro–scale analysis reveals the existence of sub-group among the girls: with $0.5 \leq \delta \leq 0.8$ all the boys are assigned to a specific cluster, while a group of 8 little girls is separated from the others. Figure 4 illustrates the factorial plane, the evolution of prototype groups obtained using the automatic selected bandwidth and the same multiplied by $\delta = 0.8$, and the corresponding modal curves.

We treat now the case $d = 3$. The evolution of the prototypes and of the corresponding modal curves, varying the coefficient δ (with $\delta = 0.7, 0.6, 0.5$ and 0.4) is illustrated in Figure 5. One can observe that when δ is relatively large the algorithm produces a first segmentation separating the boys from the girls, then, for smaller values of δ, it divides the group of boys in sub-groups. More in detail, for $\delta = 0.8$ or 0.9 only girls are completely recognized, whereas only 20 boys are separated in a second cluster. Reducing further δ, the composition of the two groups fit better with the sex of subjects: if one considers the sex as a latent variable, the correct classification rate is 90.3% when $\delta = 0.7$ and 97.9% when $\delta = 0.6$ (in this case, only 2 girls are not recognized). This latter result outperforms the ones obtained by using competitive cluster algorithms (see results collected in [30, Table 1]) and, moreover, it confirms that the

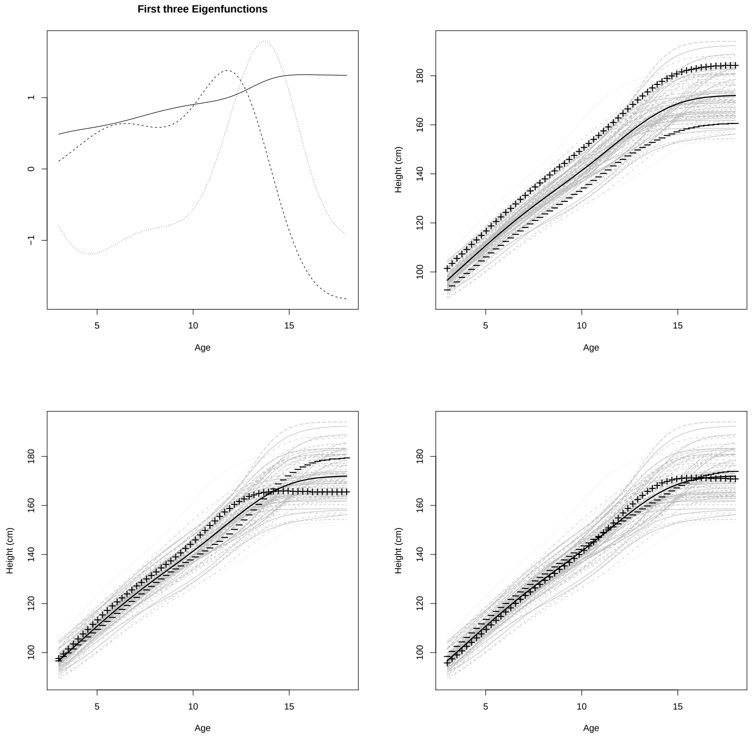

Figure 3: *Top-left*: The first three eigenfunctions. *Top-right and Bottom*: The mean growth curve perturbed by adding $(+)$ and subtracting $(-)$ a multiple $(1.5\sqrt{\lambda_j})$ of each eigenfunction $\widehat{\xi}_j$ $(j = 1, 2, 3)$.

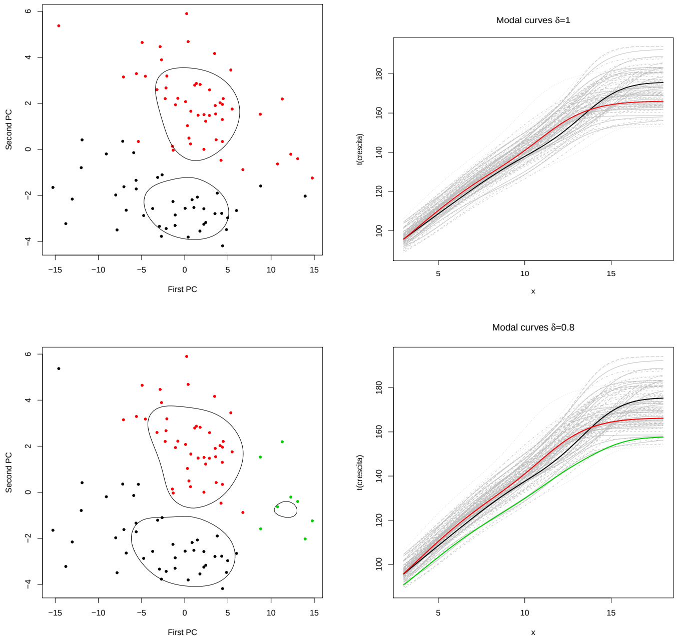

Figure 4: Prototype level-sets and modal curves for $\delta = 1, 0.8$ (from top to bottom) when $d = 2$.

registration step is not useful for the curves under analysis. Observing carefully the prototypes corresponding to the case $\delta = 0.6$, it appears that the level set that defines the clusters of boys includes de facto two more homogeneous sub-groups: this suggests to deepen the analysis by reducing further the shrinkage coefficient. Hence, passing to $\delta = 0.5$ and $\delta = 0.4$, the group of boys is segmented in three parts that one could define *small*, *normal*, and *extra–size*. Finally, repeating the procedure for very small δ ($\delta = 0.3$), the algorithm detects also the same sub-group of girls found in the case $d = 2$ when $\delta = 0.8$.

In conclusion of this Section, one can remark that if the goal of the method was to retrieve the gender of subjects, the proposed algorithm provide the best classification results, with a suitable choice of the smoothing parameters, among clustering competitors. On the other hand, if the goal was to reveal particular homogeneous structure in the data, the introduced method provide a feasible and interpretable tool that can underline some specificities within the group of boys and the one of girls.

5 Proofs

This section collects proofs of results exposed above.

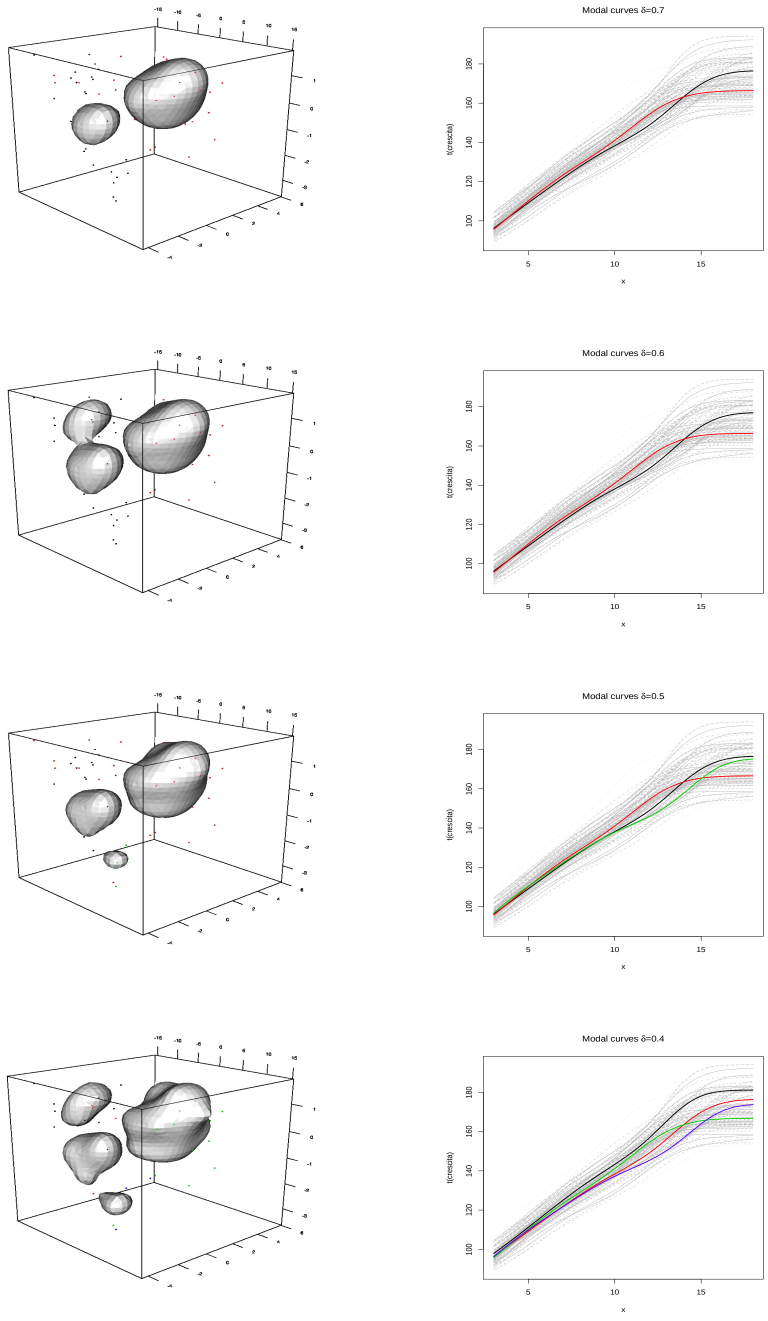

Figure 5: Prototype level-sets and modal curves form macro–scale to micro–scale analysis with $\delta = 0.7, 0.6, 0.5, 0.4$ (from top to bottom) when $d = 3$.

5.1 Proof of Theorem 2

We are interested in the asymptotic behaviour, whenever ε tends to zero, of the SmBP of the process X, that is

$$\begin{aligned}\varphi(x,\varepsilon) &= \mathbb{P}\left(\|X-x\| \le \varepsilon\right) = \mathbb{P}\left(\|X-x\|^2 \le \varepsilon^2\right)\\ &= \mathbb{P}\left(\sum_{j=1}^{+\infty}\langle X-x,\xi_j\rangle^2 \le \varepsilon^2\right) = \mathbb{P}\left(\sum_{j=1}^{+\infty}\left(\theta_j - x_j\right)^2 \le \varepsilon^2\right), \qquad \text{as } \varepsilon \to 0\end{aligned}$$

Let $S_1 = \sum_{j\le d}\left(\theta_j - x_j\right)^2$ and $S = \frac{1}{\varepsilon^2}\sum_{j\ge d+1}\left(\theta_j - x_j\right)^2$ be the truncated series and the scaled version of the remainder respectively. Thus, the SmBP is

$$\begin{aligned}\varphi(x,\varepsilon) &= \mathbb{P}\left(S_1 + \varepsilon^2 S \le \varepsilon^2\right) = \mathbb{P}\left(S_1 \le \varepsilon^2\left(1-S\right)\right)\\ &= \mathbb{E}\left[\mathbb{E}\left[S_1 \le \varepsilon^2\left(1-S\right)\Big| S\right]\right]\\ &= \mathbb{E}\left[\varphi(S|x,\varepsilon,d)\right] = \int_0^1 \varphi(s|x,\varepsilon,d)dG\left(s\right). \end{aligned} \tag{34}$$

where G is the cumulative distribution function of S. At first, for any $s \in (0,1)$, let us consider $\varphi(s|x,\varepsilon,d)$, that is the SmBP about $\Pi_d x$ of the process $\Pi_d X$ in the space spanned by $\{\xi_j\}_{j\le d}$. In terms of $f_d\left(\cdot\right)$, the probability density function of $\boldsymbol{\vartheta} = (\vartheta_1,\ldots,\vartheta_d)'$, it can be written as

$$\varphi(s|x,\varepsilon,d) = \int_D f_d\left(\boldsymbol{\vartheta}\right)d\boldsymbol{\vartheta},$$

where $D = \left\{\boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j\le d}\left(\vartheta_j - x_j\right)^2 \le \varepsilon^2\left(1-s\right)\right\}$. Note that D is an d–dimensional ball centered about $\Pi_d x = (x_1,\ldots,x_d)$ with radius $\varepsilon\sqrt{1-s}$. Now, consider the Taylor expansion of $f = f_d$ about $\Pi x = \Pi_d x$ with Lagrange remainder,

$$\begin{aligned}f\left(\boldsymbol{\vartheta}\right) =& f(x_1,\ldots,x_d) + \langle\boldsymbol{\vartheta} - \Pi x, \nabla f(x_1,\ldots,x_d)\rangle\\ &+ \frac{1}{2}(\boldsymbol{\vartheta} - \Pi x)' H_f\left(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t\right)(\boldsymbol{\vartheta} - \Pi x),\end{aligned}$$

for some $t \in (0,1)$ and with H_f denoting the Hessian matrix of f. (In general, t depends on $\boldsymbol{\vartheta} - \Pi x$, but we are not interested in the actual value of it because the boundedness of the second derivatives of f allows us to drop, in what follows, those terms depending on t). Then we can write

$$\begin{aligned}\varphi(s|x,\varepsilon,d) = &\int_D \Big(f(x_1,\dots,x_d) + \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1,\dots,x_d)\rangle \\ &+\frac{1}{2}(\boldsymbol{\vartheta} - \Pi x)' H_f \left(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t\right)(\boldsymbol{\vartheta} - \Pi x)\Big)\, d\boldsymbol{\vartheta} \\ =&f(x_1,\dots,x_d)\int_D d\boldsymbol{\vartheta} + \int_D \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1,\dots,x_d)\rangle\, d\boldsymbol{\vartheta} \\ &+ \frac{1}{2}\int_D (\boldsymbol{\vartheta} - \Pi x)' H_f \left(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t\right)(\boldsymbol{\vartheta} - \Pi x) d\boldsymbol{\vartheta} \\ =&f(x_1,\dots,x_d)I + \frac{1}{2}\int_D (\boldsymbol{\vartheta} - \Pi x)' H_f \left(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t\right)(\boldsymbol{\vartheta} - \Pi x) d\boldsymbol{\vartheta} \quad (35)\end{aligned}$$

where $I = I(s,\varepsilon,d)$ denotes the volume of the d–dimensional ball D that is

$$I = \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2+1)}(1-s)^{d/2}$$

and, the addend $\int_D \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1,\dots,x_d)\rangle\, d\boldsymbol{\vartheta}$ is null since $\langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1,\dots,x_d)\rangle$ is a linear functional integrated over the symmetric – with respect to the center $(x_1,\dots,x_d)$ – domain D. Thus from (35) it follows

$$\begin{aligned}|\varphi(s|x,\varepsilon,d) - f(x_1,\dots,x_d)I| =& \\ &= \left| \frac{1}{2}\int_D \sum_{i\le d}\sum_{j\le d}(\vartheta_i - x_i)(\vartheta_j - x_j)\frac{\partial^2 f}{\partial\vartheta_i\partial\vartheta_j}\left(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t\right) d\boldsymbol{\vartheta}\right| \\ &\le \frac{1}{2}C_1 f(x_1,\dots,x_d)\left|\sum_{i\le d}\sum_{j\le d}\int_D \frac{(\vartheta_i - x_i)(\vartheta_j - x_j)}{\sqrt{\lambda_i}\sqrt{\lambda_j}} d\boldsymbol{\vartheta}\right| \\ &= \frac{1}{2}C_1 f(x_1,\dots,x_d)\left|\sum_{j\le d}\int_D \frac{(\vartheta_j - x_j)^2}{\lambda_j} d\boldsymbol{\vartheta}\right|\end{aligned}$$

where C_1 is given by (5) and the latter equality holds because symmetry arguments lead to $\int_D(\vartheta_i - x_i)(\vartheta_j - x_j)d\boldsymbol{\vartheta} = 0$ for $i \ne j$. Furthermore, definition of D and $s \in (0,1)$ guarantee that

$$\begin{aligned}|\varphi(s|x,\varepsilon,d) - f(x_1,\dots,x_d)I| \le& \frac{1}{2}C_1 f(x_1,\dots,x_d)\left(\sum_{j\le d}\frac{1}{\lambda_j}\right)\left|\int_D \varepsilon^2(1-s)d\boldsymbol{\vartheta}\right| \\ \le& \frac{\varepsilon^2}{2}C_1 f(x_1,\dots,x_d)\left(\sum_{j\le d}\frac{1}{\lambda_j}\right) I. \quad (36)\end{aligned}$$

Come back to the SmBP (34),

$$\varphi(x,\varepsilon) = \int_0^1 f(x_1,\dots,x_d) I dG(s) + \int_0^1 \left(\varphi(s|x,\varepsilon,d) - f(x_1,\dots,x_d)I\right) dG(s), \quad (37)$$

and note that, thanks to (36) and because d is fixed, the second addend in the right–hand side of (37) is infinitesimal with respect to the first addend

$$\left|\frac{\int_0^1 \left(\varphi(s|x,\varepsilon,d) - f(x_1,\ldots,x_d)I\right) dG\,(s)}{\int_0^1 f(x_1,\ldots,x_d)IdG\,(s)}\right| \leq$$

$$\leq \left|\frac{\frac{\varepsilon^2}{2}C_1 f(x_1,\ldots,x_d)\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)\int_0^1 IdG\,(s)}{f(x_1,\ldots,x_d)\int_0^1 IdG\,(s)}\right| = \frac{C_1}{2}\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)\varepsilon^2.$$

Noting that

$$\int_0^1 I(s,\varepsilon,d)dG(s) = \frac{\varepsilon^d \pi^{d/2}}{\Gamma\left(d/2+1\right)}\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right],$$

we have

$$|\varphi(x,\varepsilon) - \varphi_d(x,\varepsilon)| \leq \frac{C_1}{2}\left(\sum_{j\leq d}\frac{1}{\lambda_j}\right)\varepsilon^2\varphi_d(x,\varepsilon)$$

where,

$$\varphi_d(x,\varepsilon) = f(x_1,\ldots,x_d)\frac{\varepsilon^d \pi^{d/2}}{\Gamma\left(d/2+1\right)}\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]. \tag{7}$$

Thus, since d is fixed, as ε tends to zero,

$$\varphi(x,\varepsilon) = \int_0^1 \varphi(s|x,\varepsilon,d)dG\,(s) = \varphi_d(x,\varepsilon) + o\left(\frac{\varphi_d(x,\varepsilon)}{f(x_1,\ldots,x_d)}\right)$$

or, equivalently, $\varphi(x,\varepsilon) \sim \varphi_d(x,\varepsilon)$ that concludes the proof.

5.2 Proofs of Proposition 3 and Proposition 5

In this section we consider results concerning asymptotics behaviour of S (Proposition 3) and $\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]$ (Proposition 5).

Proof of Proposition 3. Let us first prove that S converges to zero in probability. For any $k > 0$, by Markov inequality and, thanks to Equation (11),

$$\mathbb{P}\left(|S| > k\right) = \mathbb{P}\left(S > k\right) = \mathbb{P}\left(\frac{1}{\varepsilon^2}\sum_{j\geq d+1}(\theta_j - x_j)^2 > k\right)$$

$$\leq \frac{\mathbb{E}\left[\frac{1}{\varepsilon^2}\sum_{j\geq d+1}(\theta_j - x_j)^2\right]}{k^2} \leq \frac{C_2}{k^2}\frac{\sum_{j\geq d+1}\lambda_j}{\varepsilon^2}. \tag{38}$$

Thanks to (13) we get the converges in probability. Since $S = S(x,\varepsilon,d)$ is non–increasing when d increases,

$$\mathbb{P}\left(\sup_{j\geq d+1}|S(x,\varepsilon,j) - 0| \geq k\right) = \mathbb{P}\left(S(x,\varepsilon,d+1) \geq k\right)$$

holds for any $k > 0$ and any x. This fact, together with (38), guarantees the almost sure convergence of S to zero (e.g. [44, Theorem 10.3.1]) as ε tends to zero. Moreover, the monotone convergence theorem guarantees the L^1 convergence.
Finally, in order to prove (14), Hölder inequality guarantees that $\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right]^{2/d}$ is a non–decreasing monotone sequence with respect to d whose values are in $[0,1]$ and eventually bounded away from zero. ■
Proof of Proposition 5. At first note that

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

then, after some algebra, thanks to Bernoulli inequality (i.e. $(1+s)^r \geq 1+rs$ for $s \geq -1$ and $r \in \mathbb{R} \setminus (0,1)$), Markov inequality and Assumption (11), we have (for any $d \geq 2$)

$$\begin{aligned} 0 \leq & 1 - \mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S\leq 1\}}\right] \leq 1 - \mathbb{E}\left[\left(1-\frac{d}{2}S\right)\mathbb{I}_{\{S\leq 1\}}\right] \\ \leq & \mathbb{E}\left[\frac{d}{2}S\,\mathbb{I}_{\{S\leq 1\}}\right] \leq \mathbb{E}\left[\frac{d}{2\varepsilon^2}\sum_{j\geq d+1}(\theta_j - x_j)^2\right] \leq \frac{C_2 d}{2\varepsilon^2}\sum_{j\geq d+1}\lambda_j. \end{aligned}$$

Choosing d according to (16) the thesis follows. ■

5.3 Proof of Proposition 15

Recall that, as did in Section 2, we abuse notations dropping the dependence on d in the density estimators f and $\widehat{f}$ and, we use x both as an element of the Hilbert space and as its d–dimensional projection in $\mathbb{R}^d$ since its meaning will be clear from the context.

The proof of Proposition 15 uses similar arguments as in [6] and then some steps are only sketched. In the following C denotes a general positive constant.
Since $H_n = h_n^2 I$, it holds $K_{H_n}(u) = h_n^{-d} K(u)$. Denoting:

$$S_n(x) = \sum_{i=1}^{n} K\left(\frac{\|\Pi_d(X_i - x)\|}{h_n}\right) \qquad \text{and} \qquad \widehat{S}_n(x) = \sum_{i=1}^{n} K\left(\frac{\left\|\widehat{\Pi}_d(X_i - x)\right\|}{h_n}\right)$$

the pseudo-estimator and the estimator write respectively:

$$f_n(x) = \frac{S_n(x)}{nh_n^d} \qquad \text{and} \qquad \widehat{f}_n(x) = \frac{\widehat{S}_n(x)}{nh_n^d}$$

and hence:

$$\mathbb{E}\left[f_n(x) - \widehat{f}_n(x)\right]^2 = \frac{1}{(nh_n^d)^2}\mathbb{E}\left[S_n(x) - \widehat{S}_n(x)\right]^2.$$

Setting

$$V_i = \|\Pi_d(X_i - x)\|, \qquad \widehat{V}_i = \left\|\widehat{\Pi}_d(X_i - x)\right\|,$$

and considering the events

$$A_i = \{V_i \leq h_n\}, \qquad B_i = \left\{\widehat{V}_i \leq h_n\right\}$$

we get the decomposition:

$$\begin{aligned} S_n(x) - \widehat{S}_n(x) =& \sum_{i=1}^{n} \left[K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right)\right] \mathbb{I}_{A_i \cap B_i} + \\ &+ \sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right) \mathbb{I}_{A_i \cap \overline{B}_i} - \sum_{i=1}^{n} K\left(\frac{\widehat{V}_i}{h_n}\right) \mathbb{I}_{\overline{A}_i \cap B_i}. \end{aligned}$$

Since $(a+b)^2 \leq 2a^2 + 2b^2$, it holds:

$$\begin{aligned} \mathbb{E}\left[S_n(x) - \widehat{S}_n(x)\right]^2 \leq& 2\mathbb{E}\left[\sum_{i=1}^{n} \left(K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right)\right) \mathbb{I}_{A_i \cap B_i}\right]^2 \\ &+ 4\mathbb{E}\left[\left(\sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right) \mathbb{I}_{A_i \cap \overline{B}_i}\right)^2 + \left(\sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right) \mathbb{I}_{\overline{A}_i \cap B_i}\right)^2\right]. \end{aligned} \tag{39}$$

Consider now the first addend in the right-hand side of (39): Assumption (A3), the fact that $\left|V_i - \widehat{V}_i\right| \leq \left\|\Pi_d - \widehat{\Pi}_d\right\|_\infty \|X_i - x\|$, where $\|\cdot\|_\infty$ denotes the operator norm, and Assumption (A4) lead to:

$$\mathbb{E}\left[\sum_{i=1}^{n} \left(K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right)\right) \mathbb{I}_{A_i \cap B_i}\right]^2 \leq C\mathbb{E}\left[\left\|\Pi_d - \widehat{\Pi}_d\right\|_\infty^2 \left(\sum_{i=1}^{n} \mathbb{I}_{A_i \cap B_i}\right)^2\right]$$

Thanks to the Cauchy-Schwartz inequality we control the previous bound by:

$$\left(\mathbb{E}\left[\left\|\Pi_d - \widehat{\Pi}_d\right\|_\infty^4\right] \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{I}_{A_i \cap B_i}\right)^4\right]\right)^{1/2}$$

On the one hand, using similar arguments as in the proof of Theorem 2.1 (iii) in [6], it holds

$$\left(\mathbb{E}\left[\left\|\Pi_d - \widehat{\Pi}_d\right\|_\infty^4\right]\right)^{1/2} = O\left(\frac{1}{n}\right).$$

On the other hand, we note that the r.v. $\sum_{i=1}^{n} \mathbb{I}_{A_i \cap B_i}$ is Binomial distributed with parameters n and $P = \mathbb{P}\left(\{V_i \leq h_n\} \cap \left\{\widehat{V}_i \leq h_n\right\}\right)$. Since, when n goes to infinity, $P \sim h_n^{2d}$ and the raw fourth moment of $Bin(n, P)$ is asymptotically equivalent to $(nP)^4$, it follows

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbb{I}_{A_i \cap B_i}\right)^4\right] \sim nh_n^{8d}.$$

Finally, combining previous results, we obtain:

$$\frac{1}{\left(nh_n^d\right)^2}\mathbb{E}\left[\sum_{i=1}^{n}\left(K\left(\frac{V_i}{h_n}\right)-K\left(\frac{\widehat{V}_i}{h_n}\right)\right)\mathbb{I}_{A_i\cap B_i}\right]^2 \leq C\frac{h_n^{2(d-1)}}{n} \tag{40}$$

that goes to zero when $n\to\infty$ for any $d\geq 1$.
We work now on the second addend in the right-hand side of (39). We consider only the term:

$$\mathbb{E}\left[\sum_{i=1}^{n}K\left(\frac{V_i}{h_n}\right)\mathbb{I}_{A_i\cap\overline{B}_i}\right]^2 \tag{41}$$

because the behaviour of the other is similar. Define the sequence κ_n so that $\kappa_n\to 0$ as $n\to\infty$, the following inclusions hold:

$$\begin{aligned}
A_i\cap\overline{B}_i &= \{V_i\leq h_n\}\cap\left\{\widehat{V}_i>h_n\right\}\\
&= \left(\{h_n(1-\kappa_n)<V_i\leq h_n\}\cup\{V_i\leq h_n(1-\kappa_n)\}\right)\cap\left\{\widehat{V}_i-V_i>h_n-V_i\right\}\\
&\subseteq \{h_n(1-\kappa_n)<V_i\leq h_n\}\cup\left\{V_i\leq h_n(1-\kappa_n),\widehat{V}_i-V_i>h_n-V_i\right\}\\
&\subseteq \{h_n(1-\kappa_n)<V_i\leq h_n\}\cup\left\{\widehat{V}_i-V_i>\kappa_n h_n\right\}.
\end{aligned}$$

The latter result and Assumptions (A3) and (A4) allow to control (41) by

$$\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}_{A_i\cap\overline{B}_i}\right]^2\leq 2\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}_{\{h_n(1-\kappa_n)<V_i\leq h_n\}}\right]^2+2\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}_{\left\{\left\|\widehat{\Pi}_d-\Pi_d\right\|>C\kappa_n h_n\right\}}\right]^2.$$

About the first term in the second member of the latter, the Cauchy-Schwartz inequality gives:

$$\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}_{\{h_n(1-\kappa_n)<V_i\leq h_n\}}\right]^2\leq n^2\mathbb{P}\left(h_n(1-\kappa_n)<V\leq h_n\right).$$

Since $\mathbb{P}\left(h_n(1-\kappa_n)<V\leq h_n\right)\sim h_n^d\left(1-(1-\kappa_n)^d\right)$, performing a first order Taylor expansion of $(1-\kappa_n)^d$ in $\kappa_n=0$, we get asymptotically:

$$\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}_{\{h_n(1-\kappa_n)<V_i\leq h_n\}}\right]^2\leq Cn^2h_n^d\kappa_n.$$

Similarly, for what concerns the other addend, we have

$$\mathbb{E}\left[\sum_{i=1}^{n}\mathbb{I}_{\left\{\left\|\widehat{\Pi}_d-\Pi_d\right\|>\kappa_n h_n/M\right\}}\right]^2\leq n^2\mathbb{P}\left(\left\|\widehat{\Pi}_d-\Pi_d\right\|>C\kappa_n h_n\right)$$

with

$$\mathbb{P}\left(\left\|\widehat{\Pi}_d-\Pi_d\right\|>C\kappa_n h_n\right)=O\left(\exp\left(-nh_n^2\kappa_n^2\right)\right)$$

thanks to [6, Theorem 2.1 (i)].
Combining the previous results we obtain:

$$\frac{1}{\left(nh_n^d\right)^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right)\mathbb{I}_{A_i\cap\overline{B}_i}\right)\right]^2 = O\left(\frac{\kappa_n}{h_n^d}\right) + O\left(\frac{1}{nh_n^{2(d+1)}\kappa_n^2}\right).$$

If we choose $\kappa_n = \left(nh_n^{d+2}\right)^{-1/3}$, with $nh_n^{2(2d+1)} \to \infty$, as $n \to \infty$, we obtain:

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right)\mathbb{I}_{A_i\cap\overline{B}_i}\right)^2 + \left(\sum_{i=1}^{n} K\left(\frac{\widehat{V}_i}{h_n}\right)\mathbb{I}_{\overline{A}_i\cap B_i}\right)^2\right] \leq C\left(\frac{1}{nh_n^{2(2d+1)}}\right)^{1/3}. \tag{42}$$

In conclusion, (40) and (42) lead to

$$\frac{1}{\left(nh_n^d\right)^2}\mathbb{E}\left[S_n\left(x\right) - \widehat{S}_n\left(x\right)\right]^2 = O\left(\frac{h_n^{2(d-1)}}{n}\right) + O\left(\left(\frac{1}{nh_n^{2(2d+1)}}\right)^{1/3}\right)$$

which, when one chooses the optimal bandwidth (32), is equal to

$$O\left(n^{-(3d+2p-2)/(2p+d)}\right) + O\left(n^{-(2p-3d-2)/(3(2p+d))}\right) \tag{43}$$

provided that $3d + 2p - 2 > 0$ and $2p - 3d - 2 > 0$, that is, for each $d \geq 1$, taking $p > (3d + 2)/2$. A direct computation allows to show that (43) is definitively negligible compared to the "optimal bound" $n^{-2p/(2p+d)}$, for any $d \geq 1$.

Acknowledgements The authors are grateful to P. Vieu for discussions held while visiting University Paul Sabatier in Toulouse whose hospitality is appreciated. E. Bongiorno thanks G. González–Rodríguez for the fruitful discussions they had during a visit of the author at Universidad de Oviedo (towards which he is grateful for the kind hospitality) and, H. H. Bock that had furnished clear comments and references about density based clustering.
The authors are members of 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. Abraham, G. Biau, and B. Cadre. Simple estimation of the mode of a multivariate density. *Canad. J. Statist.*, 31(1):23–34, 2003.

[2] A. de Acosta. Small deviations in the functional central limit theorem with applications to functional laws of the iterated logarithm. *Ann. Probab.*, 11(1):78–101, 1983.

[3] A. Azzalini and N. Torelli. Clustering via nonparametric density estimation. *Stat. Comput.*, 17(1):71–80, 2007.

[4] J. Behboodian. On the modes of a mixture of two normal distributions. *Technometrics*, 12(1):pp. 131–139, 1970.

[5] A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, Boston, MA, 2004.

[6] G. Biau and A. Mas. PCA-kernel estimation. *Stat. Risk Model.*, 29(1):19–46, 2012.

[7] H. H. Bock. *Automatische Klassifikation*. Vandenhoeck & Ruprecht, Göttingen, 1974.

[8] H. H. Bock. Clustering by density estimation. In R. Tomassone, editor, *Analyse de donnes et informatique*, 1979.

[9] E. G. Bongiorno, G. Aldo, S. Ernesto, and V. Philippe, editors. *Contributions in infinite-dimensional statistics and related topics*, 2014. Società Editrice Esculapio.

[10] D. Bosq. *Linear processes in function spaces*, volume 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2000.

[11] G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont., 1973.

[12] P. Burman and W. Polonik. Multivariate mode hunting: data analytic tools with measures of significance. *J. Multivariate Anal.*, 100(6):1198–1218, 2009.

[13] K. Chen and H.-G. Müller. Conditional quantile analysis when covariates are functions, with application to growth data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(1):67–89, 2012.

[14] A. Cuevas, M. Febrero, and R. Fraiman. Estimating the number of clusters. *Canad. J. Statist.*, 28(2):367–382, 2000.

[15] A. Cuevas, M. Febrero, and R. Fraiman. Cluster analysis: a further approach based on density estimation. *Comput. Statist. Data Anal.*, 36(4):441–459, 2001.

[16] S. Dabo-Niang, F. Ferraty, and P. Vieu. On the using of modal curves for radar waveforms classification. *Comput. Statist. Data Anal.*, 51(10):4878–4890, 2007.

[17] A. Delaigle and P. Hall. Defining probability density for a distribution of random functions. *Ann. Statist.*, 38(2):1171–1193, 2010.

[18] R. Duin, A. Fred, M. Loog, and E. Pkalska. Mode seeking clustering by knn and mean shift evaluated. In G. Gimelfarb, E. Hancock, A. Imiya, A. Kuijper, M. Kudo, S. Omachi, T. Windeatt, and K. Yamada, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 7626 of *Lecture Notes in Computer Science*, pages 51–59. Springer Berlin Heidelberg, 2012.

[19] T. Duong. ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of Statistical Software*, 21(7):1–16, 10 2007.

[20] T. Duong and M. L. Hazelton. Plug-in bandwidth matrices for bivariate kernel density estimation. *J. Nonparametr. Stat.*, 15(1):17–30, 2003.

[21] T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scand. J. Statist.*, 32(3):485–506, 2005.

[22] I. Eisenberger. Genesis of bimodal distributions. *Technometrics*, 6:357–363, 1964.

[23] F. Ferraty and P. Vieu. *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York, 2006.

[24] 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.

[25] 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.

[26] P. Hall and C. Vial. Assessing the finite dimensionality of functional data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(4):689–705, 2006.

[27] J. A. Hartigan. *Clustering algorithms.* John Wiley & Sons, New York-London-Sydney, 1975.

[28] F. d. Helguero. Sui massimi delle curve dimorfiche. *Biometrika*, 3(1):pp. 84–98, 1904.

[29] J. Jacques and C. Preda. Functional data clustering: a survey. *Adv. Data Anal. Classif.*, 8(3):231–255, 2014.

[30] J. Jacques and C. Preda. Model-based clustering for multivariate functional data. *Comput. Statist. Data Anal.*, 71:92–106, 2014.

[31] B. P. Kent, A. Rinaldo, F.-C. Yeh, and T. Verstynen. Mapping topographic structure in white matter pathways with level set trees. *PLoS ONE*, 9(4):e93344, 04 2014.

[32] H.-P. Kriegel, P. Krger, J. Sander, and A. Zimek. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[33] J. Li, S. Ray, and B. G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *J. Mach. Learn. Res.*, 8:1687–1723, 2007.

[34] W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic processes: theory and methods*, volume 19 of *Handbook of Statist.*, pages 533–597. North-Holland, Amsterdam, 2001.

[35] M. A. Lifshits. On the lower tail probabilities of some random series. *Ann. Probab.*, 25(1):424–442, 1997.

[36] M. A. Lifshits. *Lectures on Gaussian processes.* Springer Briefs in Mathematics. Springer, Heidelberg, 2012.

[37] X. Liu and M. C. K. Yang. Simultaneous curve registration and clustering for functional data. *Comput. Statist. Data Anal.*, 53(4):1361–1376, 2009.

[38] G. Menardi and A. Azzalini. An advancement in clustering via nonparametric density estimation. *Stat. Comput.*, 24(5):753–767, 2014.

[39] J. O. Ramsay and B. W. Silverman. *Functional data analysis.* Springer Series in Statistics. Springer, New York, second edition, 2005.

[40] A. Rinaldo and L. Wasserman. Generalized density clustering. *Ann. Statist.*, 38 (5):2678–2722, 2010.

[41] A. Rinaldo, A. Singh, R. Nugent, and L. Wasserman. Stability of density-based clustering. *J. Mach. Learn. Res.*, 13:905–948, 2012.

[42] C. A. Robertson and J. G. Fryer. Some descriptive properties of normal mixtures. *Skand. Aktuarietidskr.*, 1969:137–146 (1970), 1969.

[43] L. M. Sangalli, P. Secchi, S. Vantini, and V. Vitelli. k-mean alignment for curve clustering. *Comput. Statist. Data Anal.*, 54(5):1219–1233, 2010.

[44] A. N. Shiryayev. *Probability*, volume 95 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1984.

[45] B. W. Silverman. *Density estimation for statistics and data analysis.* Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986.

[46] W. Stuetzle. Estimating the cluster type of a density by analyzing the minimal spanning tree of a sample. *J. Classification*, 20(1):25–47, 2003.

[47] W. Stuetzle and R. Nugent. A generalized single linkage method for estimating the cluster tree of a density. *J. Comput. Graph. Statist.*, 19(2):397–418, 2010.

[48] 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.

[49] M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability.* Chapman and Hall, Ltd., London, 1995.

[50] W.-J. Wang, Y.-X. Tan, J.-H. Jiang, J.-Z. Lu, G.-L. Shen, and R.-Q. Yu. Clustering based on kernel density estimation: nearest local maximum searching algorithm. *Chemometrics and Intelligent Laboratory Systems*, 72(1):1 – 8, 2004.

[51] D. Wishart. Mode analysis: A generalization of nearest neighbour which reduces chaining effects. In C. A.J., editor, *Numerical Taxonomy*, pages 282–311. Academic Press, 1969.

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