Statistica Sinica **27** (2017), 1949-1965
doi:https://doi.org/10.5705/ss.202016.0128

SOME INSIGHTS ABOUT THE SMALL BALL PROBABILITY FACTORIZATION FOR HILBERT RANDOM ELEMENTS

Enea G. Bongiorno and Aldo Goia

Università del Piemonte Orientale

Abstract: Asymptotic factorizations for the small–ball probability (SmBP) of a Hilbert–valued random element X are established and discussed. In particular, given the first d principal components (PCs) and as the radius ε of the ball tends to zero, the SmBP is asymptotically proportional to (a) the joint density of the first d PCs, (b) the volume of the d–dimensional ball with radius ε, and (c) a correction factor weighting the use of a truncated version of the process expansion. Under suitable assumptions on the spectrum of the covariance operator of X and as d diverges to infinity when ε vanishes, some simplifications occur. In particular, the SmBP factorizes asymptotically as the product of the joint density of the first d PCs and a pure volume parameter. The factorizations allow one to define a *surrogate intensity* of the SmBP that, in some cases, leads to a genuine *intensity*. To operationalize the stated results, a non–parametric estimator for the *surrogate intensity* is introduced and it is proved that the use of estimated PCs, instead of the true ones, does not affect the rate of convergence. Finally, as an illustration, simulations in controlled frameworks are provided.

Key words and phrases: Hilbert functional data, Karhunen–Loève decomposition, kernel density estimate, small ball probability.

1. Introduction

For a random element X valued in a general metric space, the measure of how it concentrates over such a space plays a central role in statistical analysis. If X is a real random vector, its joint density is, in a natural way, that measure. In practical situations, the density is helpful in defining mixture models, in detecting latent structure, in discriminant analysis, in identifying outliers, and so on. When observed data are curves, surfaces, images, objects or, briefly, *functional data* (see e.g. monographs Ferraty and Vieu (2006); Horváth and Kokoszka (2012); Ramsay and Silverman (2005), and Bongiorno et al. (2014) for recent contributions), the dimensionality of the space to which the data belong raises problems in defining

an object that plays the role of the joint density. The main problem is that, without an underlying dominant probability measure, the Radon–Nikodym derivative cannot be straightforwardly applied. To manage this, a concept of "surrogate density" can be derived from the notion of small–ball probability (SmBP in the sequel) of a random element X.

For a given point x, a semimetric Δ, and a real positive ε, consider $\varphi(x,\varepsilon) = \mathbb{P}(\Delta(X,x) < \varepsilon)$. The behaviour of $\varphi(x,\varepsilon)$ as ε vanishes (i.e. of the SmBP) provides information about the way in which X concentrates at x. From a theoretical point of view, the limiting behaviour has been developed in the small tails/deviations theory, see Li and Shao (2001); Lifshits (2012), and references therein. In functional statistics the SmBP was used to derive asymptotics in mode estimations (see, e.g. Dabo-Niang, Ferraty and Vieu (2007); Delaigle and Hall (2010); Ferraty, Kudraszow and Vieu (2012); Gasser, Hall and Presnell (1998)), as well as in non–parametric regression literature in evaluating the rate of convergence of estimators (see, e.g. Ferraty and Vieu (2006); Ferraty, Mas and Vieu (2007)). Often, the necessity to have a surrogate density available for X has involved the assumption (as done, for instance, in Ferraty, Kudraszow and Vieu (2012); Gasser, Hall and Presnell (1998)) that

$$\varphi(x,\varepsilon) = \Psi(x)\,\phi(\varepsilon) + o(\phi(\varepsilon)), \qquad \varepsilon \to 0, \tag{1.1}$$

where Ψ is the *intensity* of the SmBP that plays the role of the *surrogate density* of the random element X, whilst $\phi(\varepsilon)$ is a kind of "volume parameter". Although breaking the dependence on x and ε supplies a clear modelling advantage and the existence of $\Psi(x)$ is desirable, factorization (1.1) can be derived only in particular settings. Notable examples are the case of Gaussian processes (e.g. Li and Shao (2001); Lifshits (2012), and references therein) and the one of fractal processes for suitable semi–norms Δ (e.g. Ferraty and Vieu (2006, Chap. 13)). Hence, a crucial task is to study some asymptotic factorizations of the SmBP leading to a definition of its *intensity* or, at least a *surrogate intensity*, when it is not possible to completely isolate the dependence on x and ε. In the framework of random elements in a separable Hilbert space with Δ the induced metric, a first factorization of the SmBP that allows one to define a *surrogate intensity* was provided by Delaigle and Hall (2010). Under some technical hypothesis on the spectrum of the covariance operator of X, and assuming that principal components of X are independent with positive and sufficiently smooth marginal density functions $\{\tilde{f}_j\}$, the authors showed that $\varphi(x,\varepsilon) \sim \prod_{j\le d} \tilde{f}_j(x_j)\,\phi(\varepsilon,d)$, as $\varepsilon \to 0$, where x_j is the projection of x over the j-th principal axis, $\phi(\varepsilon,d)$ is a volumetric term,

and $d = d(\varepsilon)$ diverges to infinity as ε vanishes. From the applications point of view, the independence assumption appears quite restrictive and the spatial factor $\prod_{j\leq d} \tilde{f}_j$ results in just a *surrogate intensity* of the SmBP because of the dependence between d and ε. Moreover, one wonders if the principal component analysis is necessary to obtain the factorization.

The first part of this work proposes some more general factorizations for the SmBP in the separable Hilbert framework. The aim is to relax the hypothesis of independence, and to identify those situations which lead to a genuine *intensity*. The first result holds for any positive integer d:

$$\varphi(x,\varepsilon) \sim f_d(x_1,\ldots,x_d)V_d(\varepsilon)\mathcal{R}\left(x,\varepsilon,d\right), \qquad \text{as } \varepsilon \to 0,$$

where f_d is the joint distribution of the first d principal components, $V_d(\varepsilon)$ is the volume of a d–dimensional ball with radius ε, and $\mathcal{R}\left(x,\varepsilon,d\right) \in (0,1]$ denotes an extra factor compensating the use of $(x_1,\ldots,x_d)$ instead of x. Such factorization benefits from the fact that d is fixed but, because $\mathcal{R}$ depends on both x and ε, a genuine *intensity* cannot be defined without additional assumptions on the probability law of the process and/or on the point x at which the factorization is evaluated.

Moving further, we prove:

$$\varphi(x,\varepsilon) \sim f_d(x_1,\ldots,x_d)\phi(\varepsilon,d), \qquad \text{as } \varepsilon \to 0, \text{ and } d(\varepsilon) \to \infty,$$

where $\phi(\varepsilon,d)$ is a volume parameter that depends on the decay rate of $\{\lambda_j\}$, the eigenvalues of the covariance operator of X and f_d is the *surrogate intensity*. In particular cases, this allows one to define an *intensity*. It turns out that our factorizations can be derived for any basis but, for the second one, the principal components basis is optimal in some sense.

In the second part of the paper, to make available the *surrogate intensity* of the SmBP for statistical purposes, we propose a multivariate kernel density approach to estimating f_d. Under general conditions, we prove that, although the estimation procedure involves the estimated principal components instead of the true ones, the estimator achieves the classical non–parametric rate of convergence. To show how such an estimator performs on finite sample frameworks, we study its behaviour by means of simulated processes with known *intensities*.

The paper outline goes as follows: Section 2 introduces the framework, Section 3 considers the factorization of the SmBP when d is fixed, whereas Section 4 has d diverging to infinity as ε vanishes. Section 5 provides the statistical asymptotic theorem in estimating the joint density f_d. Section 6 illustrates some numerical examples. The proofs are in the supplementary materials.

2. Preliminaries

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 standard inner product $\langle g, h\rangle = \int_0^1 g(t)\, h(t)\, dt$ and the induced norm $\|g\|^2 = \langle g, g\rangle$. Consider a measurable map X 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 $\|\cdot\|$. Define the SmBP with $\Delta(X, x) = \|X - x\|$, $\varphi(x, \varepsilon) = \mathbb{P}(\|X - x\| < \varepsilon)$. Denote by $\mu_X = \{\mathbb{E}[X(t)], t \in [0,1]\}$, and $\Sigma[\cdot] = \mathbb{E}[\langle X - \mu_X, \cdot\rangle (X - \mu_X)]$, the mean function and covariance operator of X respectively. Consider the Karhunen–Loève expansion associated to X (see e.g. Bosq (2000)): denoting by $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$ the decreasing to zero sequence of non–negative eigenvalues and the associated orthonormal eigenfunctions of the covariance operator Σ, the random curve X admits the representation $X(t) = \mu_X(t) + \sum_{j\geq 1} \theta_j \xi_j(t)$, $0 \leq t \leq 1$, where $\theta_j = \langle X - \mu_X, \xi_j\rangle$ are the so–called principal components (PCs in the sequel) of X satisfying $\mathbb{E}[\theta_j] = 0$, $Var(\theta_j) = \lambda_j$ and $\mathbb{E}[\theta_j \theta_{j'}] = 0, j \neq j'$.

In order to achieve our aims, we need some assumptions.

(A-1) $\mu_X = 0$.

(A-2) The center of the ball $x \in \mathcal{L}^2_{[0,1]}$ is sufficiently close to the process in its high–frequency part, that is $x_j^2 \leq C_1 \lambda_j$ for any $j \geq 1$, where $x_j = \langle x, \xi_j\rangle$ for some positive constant C_1.

The latter is not a restrictive condition since it holds when x belongs to the reproducing kernel Hilbert space generated by the process X:

$$RKHS(X) = \{x \in \mathcal{L}^2_{[0,1]} : \sum_{j\geq 1} \lambda_j^{-1} \langle x, \xi_j\rangle^2 < \infty\}, \tag{2.1}$$

that is, when x is "at least smooth as the covariance function", see Berlinet and Thomas-Agnan (2004, p. 13 and p. 69). Furthermore, (A-2) is not unusual since it is equivalent to $\sup_{j\geq 1} \mathbb{E}\left[(\theta_j - x_j)^2/\lambda_j\right] < \infty$ that was used, for similar purpose by Delaigle and Hall (2010, Condition (4.1)).

(A-3) Denote by Π_d the projector onto the d–dimensional space spanned by $\{\xi_j\}_{j=1}^d$. The first d PCs, $\boldsymbol{\theta} = \Pi_d X = (\theta_1, \ldots, \theta_d)'$, admit a joint strictly positive probability density, $\boldsymbol{\vartheta} \in \mathbb{R}^d \mapsto f_d(\boldsymbol{\vartheta})$. Moreover, f_d is twice differentiable at $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d)' \in \mathbb{R}^d$, and there exists a positive constant C_2 (not depending on d) for which

$$\left|\frac{\partial^2 f_d}{\partial \vartheta_i \partial \vartheta_j}(\boldsymbol{\vartheta})\right| \leq \frac{C_2}{\sqrt{\lambda_i \lambda_j}} f_d(x_1, \ldots, x_d) \tag{2.2}$$

for any $d \in \mathbb{N}$, $i, j \leq d$ and $\boldsymbol{\vartheta} \in D^x = \{\boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j \leq d}(\vartheta_j - x_j)^2 \leq \rho^2\}$ for some $\rho \geq \varepsilon$.

From now on, with a slight abuse of notation and when it is clear from the context, $f_d(x)$ denotes $f_d(x_1, \ldots, x_d)$. It is worth noting that (A-3) is not restrictive: it includes, for instance, the case of Gaussian Hilbert–valued processes.

3. Approximations for a Given d

For a finite positive integer d, and a given point $x \in \mathcal{L}^2_{[0,1]}$, let

$$S = S(x, \varepsilon, d) = \frac{1}{\varepsilon^2} \sum_{j \geq d+1} (\theta_j - x_j)^2, \qquad \mathcal{R}(x, \varepsilon, d) = \mathbb{E}\left[(1 - S)^{d/2} \mathbb{I}_{\{S<1\}}\right], \tag{3.1}$$

and $V_d(\varepsilon) = \varepsilon^d \pi^{d/2} / \Gamma(d/2 + 1)$, the volume of the d–dimensional ball with radius ε. With X and $\varphi(x, \varepsilon)$ as above, set

$$\varphi_d(x, \varepsilon) = f_d(x) V_d(\varepsilon) \mathcal{R}(x, \varepsilon, d), \qquad \text{for } \varepsilon > 0. \tag{3.2}$$

Theorem 1. *If* (A-1), …, (A-3) *hold, then*

$$|\varphi(x, \varepsilon) - \varphi_d(x, \varepsilon)| \leq C_2 \frac{\varepsilon^2}{2\lambda_d} \varphi_d(x, \varepsilon) \qquad \text{for } \varepsilon > 0, \tag{3.3}$$

that is

$$\varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon) \mathcal{R}(x, \varepsilon, d) \qquad \text{for } \varepsilon \to 0. \tag{3.4}$$

In other words, for a fixed d and as $\varepsilon \to 0$, the SmBP $\varphi(x, \varepsilon)$ behaves as $\varphi_d(x, \varepsilon)$, the usual first order approximation of the SmBP in a d–dimensional space $f_d(x)V_d(\varepsilon)$ up to the scale factor $\mathcal{R}(x, \varepsilon, d)$. The latter, depending on x only through its high–frequency components $\{x_j\}_{j \geq d+1}$, can be interpreted as a corrective factor compensating for the use of a truncated version of the process expansion. Changing d affects all the terms in the factorization but not (3.4). Because of $\mathcal{R}(x, \varepsilon, d)$, the dependence on x and ε cannot be isolated and hence an *intensity* of the SmBP is not, in general, available.

There exist some situations in which a genuine *intensity* can be defined from the above factorization: a) $\mathcal{R}(x, \varepsilon, d)$ is independent on x; b) there exists a finite positive integer d_0 such that, for any $d \geq d_0$, $\mathcal{R}(x, \varepsilon, d) = 1$; c) for any x, as $\varepsilon \to 0$, $d(\varepsilon) \to \infty$, $\mathcal{R}(x, \varepsilon, d) \to 1$ and $\varphi(x, \varepsilon) \sim f_d(x) V_d(\varepsilon)$.

In the following, we discuss points a) and b), whereas point c) is discussed in Section 4.

D.1. $\mathcal{R}(x, \varepsilon, d)$ is independent on x. Consider, for instance, $x_j = 0$ for any $j \geq d_0 + 1$, that x belongs to the space spanned by $\{\xi_1, \ldots, \xi_{d_0}\}$. From

Theorem 1 for any $d \geq d_0$, we have $\varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon)\mathcal{R}(\varepsilon,d)$, as $\varepsilon \to 0$, where $V_d(\varepsilon)\mathcal{R}(\varepsilon,d)$ now represents a pure volumetric term while f_d is an *intensity* of the SmBP evaluated at x.

For Gaussian processes, Theorem 1 gives

$$\varphi(x,\varepsilon) \sim \exp\left\{-\frac{1}{2}\sum_{j\leq d}\frac{x_j^2}{\lambda_j}\right\}\frac{V_d(\varepsilon)\mathcal{R}(\varepsilon,d)}{\prod_{j\leq d}\sqrt{2\pi\lambda_j}} = \Psi_d(x)\mathcal{V}_d(\varepsilon), \qquad \text{as } \varepsilon \longrightarrow 0$$

where, for any $d \geq d_0$, $\Psi_d(x) = \Psi_{d_0}(x) = \exp\left\{-\sum_{j\leq d_0} x_j^2/(2\lambda_j)\right\}$ is the *intensity* of the SmBP evaluated at x. In particular, for a Wiener process on $[0,1]$, $\Psi_{d_0}(x)$ agrees with known results (see, for instance, Li and Shao (2001, Thm. 3.1) and Dereich et al. (2003, Example 5.1)). The Karhunen–Loève decomposition of a Wiener process is $W(t) = \sum_{j=1}^{\infty} Z_j \xi_j(t)$, $t \in [0,1]$, where $\{Z_j\}$ are i.i.d. as $Z \sim N(0,1)$, $\xi_j(t) = \sqrt{2}\sin\left((j-0.5)\pi t\right)/\sqrt{\lambda_j}$, $\lambda_j = (j-0.5)^{-2}\pi^{-2}$ and it is known that

$$\varphi(x,\varepsilon) \sim \exp\left\{-\frac{1}{2}\int_0^1 x'(t)^2\,dt\right\} 4\varepsilon \exp\frac{\left\{-1/(8\varepsilon^2)\right\}}{\sqrt{\pi}}, \qquad \varepsilon \to 0,$$

where $x(t)$ is sufficiently smooth. Since we are interested in the definition of an intensity, we compare the spatial parts. For any $x(t) = \sum_{j=1}^{d_0} b_j \xi_j(t)$ where $b_j \in \mathbb{R}$, straightforward computations lead to

$$\exp\left\{-\frac{1}{2}\int_0^1 x'(t)^2\,dt\right\} = \exp\left\{-\frac{1}{2}\sum_{j=1}^{d_0} b_j^2\right\} = \Psi_{d_0}(x).$$

D.2. The case $\mathcal{R}(x,\varepsilon,d) = 1$. Suppose X takes values in a d_0–dimensional subspace of the Hilbert space. Then $\lambda_j = 0$ for any $j \geq d_0 + 1$, (A-2) leads to $x_j = \theta_j = 0$, and $\mathcal{R}(x,\varepsilon,j) = 1$ for any $j \geq d_0 + 1$. Moreover, Theorem 1 can be applied only for $d \leq d_0$ because f_{d_0+1} is not strictly positive and hence (A-3) fails. Consequently $\varphi(x,\varepsilon) \sim f_{d_0}(x) V_{d_0}(\varepsilon)$, which is the usual first order approximation of the d_0–dimensional process and f_{d_0} is the *intensity* of the SmBP of the process.

D.3. Changing the basis. If $\{\xi_j\}_{j=1}^{\infty}$ is an orthonormal basis of the Hilbert space, arranged so that the sequence $Var(\langle X, \xi_j\rangle) = \lambda_j$ is in descending order, Theorem 1 still holds.

4. Approximations When d Depends on ε

We establish conditions on X that allow one to simplify (3.4), to get $\varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon)$, as $\varepsilon \to 0$. This is achieved by combining Theorem 1 and the limit behaviour of $\mathcal{R}(x,\varepsilon,d)$ to have, for any x,

$$\begin{cases} \mathcal{R}(x,\varepsilon,d) \to 1, \\ \varphi(x,\varepsilon) \sim f_d(x) V_d(\varepsilon), \end{cases} \qquad \varepsilon \to 0,\ d(\varepsilon) \to \infty. \tag{4.1}$$

Consider then the limit behaviour of $\mathcal{R}$, as ε goes to zero and d diverges to infinity.

Proposition 1. *Assume* (A-2) *and that* $\sum_{j\geq d+1} \lambda_j = o(1/d)$, *as* d *goes to infinity. One can choose* $d = d(\varepsilon)$ *so that it diverges to infinity as* ε *tends to zero and* $d\sum_{j\geq d+1} \lambda_j = o(\varepsilon^2)$. *Then, as* $\varepsilon \to 0$,

$$0 \leq 1 - \mathcal{R}(x,\varepsilon,d) \leq \frac{C_1(d+2)}{2\varepsilon^2} \sum_{j\geq d+1} \lambda_j = o(1). \tag{4.2}$$

Consider the inequality

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

that, thanks to (3.3), (4.2), and $0 < \mathcal{R} \leq 1$, leads to

$$\begin{aligned} \left| \frac{\varphi(x,\varepsilon)}{f_d(x)\, V_d(\varepsilon)} - 1 \right| &\leq C_2 \frac{\varepsilon^2}{2\lambda_d} \mathcal{R}(x,\varepsilon,d) + |\mathcal{R}(x,\varepsilon,d) - 1| \\ &\leq C_2 \frac{\varepsilon^2}{2\lambda_d} + \frac{C_1(d+2)}{2\varepsilon^2} \sum_{j\geq d+1} \lambda_j. \end{aligned} \tag{4.3}$$

Thus, the wished result holds if there exists $d = d(\varepsilon)$ such that

$$\varepsilon^2 = o(\lambda_d), \qquad \text{and} \qquad (d+2) \sum_{j\geq d+1} \lambda_j = o(\varepsilon^2). \tag{4.4}$$

To obtain (4.1) we combine conditions in (4.4) (plug the first in the second), and we get that eigenvalues must satisfy the *hyper–exponential* decay rate

$$\frac{d\sum_{j\geq d+1} \lambda_j}{\lambda_d} = o(1), \qquad \text{as } d \to \infty. \tag{4.5}$$

This rate highlights the trade–off between the approximation errors provided by Theorem 1 and Proposition 1. Moreover, it is a necessary condition to guarantee that (4.3) vanishes. One wonders if it is possible to define $d = d(\varepsilon)$ so that the errors in (4.4) vanish at the same time as ε goes to zero.

Theorem 2. *Under the conditions of Theorem* 1, *if the eigenvalues decay hyper–exponentially, it is possible to choose* $d = d(\varepsilon)$ *so that, if* $\varepsilon \to 0$, *then* $d \to \infty$ *and*

$$\varphi(x,\varepsilon) = f_d(x)\, V_d(\varepsilon) + o(f_d(x)\, V_d(\varepsilon)). \tag{4.6}$$

In what follows, we discuss assumptions and consequences of the above result.

D.4. Again about the intensity of the SmBP. Because of the relation

between d and ε, in general (4.6) does not allow one to define an *intensity* as commonly intended. Since f_d is the only term depending on x, it can be considered as a *surrogate intensity*.

Gaussian processes, or suitable generalizations, provide examples for which f_d leads to define a genuine *intensity*. At first, consider a Gaussian process X: for any $x \in \mathcal{L}^2_{[0,1]}$ and as ε goes to zero, $\varphi(x,\varepsilon) \sim \Psi_d(x)\mathcal{V}_d(\varepsilon)$; see D.1. When d tends to infinity, for any $x \in \mathcal{L}^2_{[0,1]}$, $\Psi_d(x)$ tends to $\exp\{-\sum_{j\geq 1} x_j^2/(2\lambda_j)\}$ which is the *intensity* of the small–ball probability at x. Note that it is not null if and only if x belongs to $RKHS(X)$, see (2.1).

Another situation in which an *intensity* for the SmBP can be defined, occurs when the PCs are independent each with density belonging to a subfamily of the exponential power (or generalized normal) distribution (see e.g. Box and Tiao (1973)), that is proportional to $\exp\left\{-\left(|x_j|/\sqrt{\lambda_j}\right)^q\right\}$, with $q \geq 2$. In this case, $\Psi(x) = \exp\{-1/2\sum_{j=1}^{\infty}(|x_j|/\sqrt{\lambda_j})^q\}$, for any $x \in \mathcal{L}^2_{[0,1]}$ and, it is not null if x is in $H(q) = \{x \in \mathcal{L}^2_{[0,1]} : \sum\left(|x_j|/\sqrt{\lambda_j}\right)^q < \infty\}$ that includes the $RKHS(X)$ when $q \geq 2$.

D.5. An example of hyper–exponential decay. Suppose $\lambda_j = \exp\{-\beta j^\alpha\}$ with $\beta > 0$ and $\alpha > 1$. In this case, for any real number $n \geq 1$,

$$\frac{d\sum_{j\geq d+1}\lambda_j}{\lambda_d} \leq \frac{d^n\sum_{j\geq d+1}\lambda_j}{\lambda_d} \to 0, \qquad \text{as } d \to \infty. \tag{4.7}$$

In fact, some algebra and the Bernoulli inequality give

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

Since $\exp\{-\beta\alpha d^{\alpha-1}j\} \leq (j^2 d^{n+\delta})^{-1}$ eventually (with respect to d) holds for some positive δ and for each $j \in \mathbb{N}$, (4.7) is obtained.

4.1. Changing the eigenvalues decay rate

The factorization (4.6) is obtained at the cost of the hyper–exponential eigenvalues decay (4.5). If one changes the eigenvalues decay rate, a factorization of the SmBP is still available, but the volumetric term cannot be written explicitly. We focus on the decay rates

"super–exponential": $\lambda_d^{-1}\sum_{j\geq d+1}\lambda_j = o(1)$, as $d \to \infty$, or equivalently

$$\frac{\lambda_{d+1}}{\lambda_d} \to 0, \qquad \text{as} \qquad d \to \infty. \tag{4.8}$$

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

$$\lambda_d^{-1} \sum_{j \geq d+1} \lambda_j < C, \qquad \text{for any } d \in \mathbb{N}. \tag{4.9}$$

It is possible to show that (4.5) $\Rightarrow$ (4.8) $\Rightarrow$ (4.9) but the contraries do not hold. For instance, for any $\alpha > 1$ and $\beta > 0$, $\lambda_j = \exp\{-\beta j\}$ decays exponentially but not super–exponentially, $\lambda_j = \exp\{-\beta j \ln(\ln(j))\}$ decays super–exponentially but not hyper–exponentially, while $\lambda_j = \exp\{-\beta j^\alpha\}$ decays hyper–exponentially.

Theorem 3. *Under the conditions of Theorem* 1*, as ε tends to zero, it is possible to choose $d = d(\varepsilon)$ diverging to infinity so that $\varphi(x, \varepsilon) \sim f_d(x)\, \phi(\varepsilon, d)$, where*

i) $\phi(\varepsilon, d) = \exp\{(1/2)d[\log(2\pi e\varepsilon^2) - \log(d) + o(1)]\}$ *in the super-exponential case;*

ii) $\phi(\varepsilon, d) = \exp\{1/2d[\log(2\pi e\varepsilon^2) - \log(d) + \delta(d, \alpha)]\}$ *in the exponential case, with $\lim_{\alpha\to\infty} \limsup_{s\to\infty} \delta(s, \alpha) = 0$, and α a parameter chosen so that $\lambda_d^{-1}\varepsilon^2 \leq \alpha^2$.*

In other words, $f_d(x)$ preserves the role of a *surrogate intensity* whereas $V_d(\varepsilon)$ is replaced by $\phi(\varepsilon, d)$ which depends on terms implicitly defined (namely, $o(1)$ and $\delta(s, \alpha)$). It is just the case to note that, in the exponential setting, Discussion D.4 about Gaussian and exponential power processes still holds with minor modifications.

D.6. About slower eigenvalues decay rates. This theoretical problem is partially still open. In fact, a part from the Gaussian processes and, in particular, the Wiener one (whose eigenvalues decay arithmetically but the intensity, evaluated at smooth x, can be defined as illustrated in D.1), to the best of our knowledge, there are no other attempts to provide asymptotic factorizations for the SmBP of processes whose eigenvalues decay slower than exponentially. Hence, if no information about the probability law is available, a solution is to go back to Theorem 1 to manage the dependence on x and ε in $\mathcal{R}(x, \varepsilon, d)$.

D.7. Optimal basis. Although the factorization results in Theorems 2 and 3 are stated 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 they decay sufficiently fast. In particular, using the same notations as in D.3, if the sequence $\{\lambda_j\}_{j=1}^\infty$ has an exponential decay then Theorem 3 still holds and a *surrogate intensity* can be defined. The variances obtained when one uses the PCA basis exhibit, by construction, the fastest decay: in this sense the choice of this basis can be considered optimal.

5. Estimation of the Surrogate Intensity

Theorems 1, 2 and 3 justify the use of f_d as a *surrogate intensity* for Hilbert–valued processes in statistical applications as done, for instance, within classification problems by Bongiorno and Goia (2016). We aim to make the factorization results useful for practical purposes and, in particular, to introduce an estimator of the *surrogate intensity* f_d.

Consider a sample of random curves $\{X_i, i = 1, \ldots, n\}$, i.i.d. as X. If the sequence of eigenvalues $\{\xi_j\}_{j=1}^{\infty}$ was known, one would 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 introduce the classical kernel density estimate of f_d as

$$f_{d,n}(\Pi_d x) = f_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{H_n}(\|\Pi_d(X_i - x)\|), \tag{5.1}$$

where $K_{H_n}(\mathbf{u}) = \det(H_n)^{-1/2} K(H_n^{-1/2}\mathbf{u})$, K is a kernel function, and $H_n = H_{nd}$ is a symmetric semi-definite positive $d \times d$ matrix (we drop the dependence on d). Equation (5.1) defines only a pseudo-estimate for f_d as the covariance operator Σ and the sequence $\{\xi_j\}$ are unknown. Thus, to operationalize these pseudo-estimates, we need estimates $\widehat{\theta}_i$ and $\widehat{\Pi}_d$ of θ_i and Π_d respectively. The sample versions of μ_X and Σ, are $\overline{X}_n(t) = 1/n \sum X_i(t)$, and $\widehat{\Sigma}_n[\cdot] = 1/n \sum \langle X_i - \overline{X}_n, \cdot \rangle (X_i - \overline{X}_n)$, respectively. The eigenelements $\{\widehat{\lambda}_j, \widehat{\xi}_j\}_{j=1}^{\infty}$ of $\widehat{\Sigma}_n$ provide estimates of for $\{\lambda_j, \xi_j\}_{j=1}^{\infty}$, and $\langle X_i - \overline{X}_n, \widehat{\xi}_j \rangle = \widehat{\theta}_{ji}$ estimates θ_{ji} (the asymptotic behaviour of these estimators has been widely studied; see e.g. Bosq (2000)). Plugging these estimates in (5.1), we get the kernel density estimator:

$$\widehat{f}_{d,n}\left(\widehat{\Pi}_d x\right) = \widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_{H_n}\left(\left\|\widehat{\Pi}_d(X_i - x)\right\|\right), \qquad \widehat{\Pi}_d x \in \mathbb{R}^d. \tag{5.2}$$

Since $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_n \geq 0 = \widehat{\lambda}_{n+1} = \ldots$ one could choose $d = n$ but, in practice, this is not an appropriate choice: the curse of dimensionality jeopardizes the quality of estimation. A suitable dimension $d \ll n$ has to be identified. This problem is, in practice, still open and needs developments that go beyond the scope of this paper.

We consider the problem of whether using $\widehat{f}_n$ instead of f_n has an effect on the rate of convergence of the kernel estimator. To answer this question, we study the behaviour of $\mathbb{E}[f_d(x) - \widehat{f}_n(x)]^2$ as n goes to infinity. For the sake of simplicity, we consider the special case $H_n = h_n^2 I$ where I is the identity matrix, with d fixed and independent of the observed data, and we suppose that the

following hold.

(B-1) $f_d(x)$ is positive and p times differentiable at $x \in \mathbb{R}^d$, with $p \geq 2$;

(B-2) the sequence $\{h_n\}$ satisfies: $h_n \to 0$ and $nh_n^d / \log n \to \infty$ as $n \to \infty$;

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

(B-4) there exist positive constants s and κ such that $\mathbb{E}\left[\|X - x\|^m\right] \leq m! s \kappa^{m-2}/2$ for all integers $m \geq 2$.

Assumptions (B-1), (B-2), and (B-3) are standard in the non–parametric framework, and $p \geq 2$ is required because of (A-3). Condition (B-4) holds for a wide family of processes, including the Gaussian.

First, observe that one can control the quadratic mean under study by intercalating the pseudo-estimator (5.2); thanks to the triangle inequality

$$\mathbb{E}\left[f_d(x) - \widehat{f}_n(x)\right]^2 \leq \mathbb{E}\left[f_d(x) - f_n(x)\right]^2 + \mathbb{E}\left[f_n(x) - \widehat{f}_n(x)\right]^2. \tag{5.3}$$

About the first term on the right–hand side of (5.3), it is known (see for instance Wand and Jones (1995)) that, under assumptions (B-1), ..., (B-4) and taking the optimal bandwidth

$$c_1 n^{-1/(2p+d)} \leq h_n \leq c_2 n^{-1/(2p+d)}, \tag{5.4}$$

where c_1 and c_2 are positive constants, one gets the minimax rate: $\mathbb{E}[f_d(x) - f_n(x)]^2 = O\left(n^{-2p/(2p+d)}\right)$ uniformly in $\mathbb{R}^d$. Therefore, it is enough to control the second addend on the right–hand side of (5.3).

The following theorem states that using the estimated principal components instead of the empirical ones does not affect the rate of convergence.

Theorem 4. *Assume* (B-1), ..., (B-4) *with* $p > \max\{2, 3d/2\}$, *and consider the optimal bandwidth* (5.4). *Thus* $\mathbb{E}[f_n(x) - \widehat{f}_n(x)]^2 = o\left(n^{-2p/(2p+d)}\right)$ *as* n *goes to infinity, and uniformly in* $\mathbb{R}^d$.

Formulation (5.2) requires that each random curve $X_i(t)$ is observed entirely in the continuum and without noise over $[0, 1]$. In practice, the curves are available only at design points $\{\tau_{i,1}, \ldots, \tau_{i,p_i}\}$, $\tau_{i,j} \in [0, 1]$, that are not necessarily the same for each i. Thus, some numerical approximations to compute the estimates are necessary. When each curve is observed without errors over the same fixed equispaced grid, with p sufficiently large, one can replace integrals by summations: the empirical covariance operator is approximated by a matrix

and its eigenelements are computed by standard numerical algorithms (see Rice and Silverman (1991)). This is the approach we follow in the simulations in Section 6. A more general situation occurs when observed data are discretely sampled and corrupted by noise. Suppose then that, one has observed pairs $\{(\tau_{i,j}, Y_{i,j}), i = 1, \ldots, n, j = 1, \ldots, p_i\}$, where $Y_{i,j} = X_i(\tau_{i,j}) + \varepsilon_{ij}$ and the errors ε_{ij} are i.i.d. with zero mean and finite variance. If each $p_i \geq M_n$, where M_n is a suitable sequence tending to infinity with n (we refer to this case as *dense functional data*), a presmoothing process is run before performing PCA using the sample mean and covariance computed from the smoothed curves (see, for instance, Hall, Müller and Wang (2006)). Under suitable assumptions, the estimators of eigenelements are root-n consistent and first-order equivalent to the estimators obtained if curves were directly observed (see Hall, Müller and Wang (2006, Theorem 3)).

6. Finite Sample Performances in Estimating the Surrogate Density

We illustrate the feasibility of the SmBP factorization approach by exploring how the proposed estimator works in a finite sample setting. We considered only two situations because of the difficulty in finding explicit expressions for the *intensity*. First, we focused on a finite–dimensional process for which the surrogate density is straightforwardly derived. Then, we dealt with the Wiener process. In both cases, we studied how the estimates behaved varying the sample size and d. All simulations rested on the density estimator defined in (5.2), and were performed on a suitable grid of the d–dimensional factor space: the algorithms were implemented in R, and exploited the function `kde` in the package `ks` (see Duong (2007)).

6.1. Finite dimensional setting

Consider the one-dimensional random process $X(t) = a\sqrt{2/\pi}\sin(t)$, $t \in [0, \pi]$, where a is a random variable with zero mean, unitary variance, density f_a, and cumulative distribution function F_a. Given $x(t) = b\sqrt{2/\pi}\sin(t)$ with $b \in \mathbb{R}$, for any $\varepsilon > 0$, $\varphi(x, \varepsilon) = F_a(b + \varepsilon) - F_a(b - \varepsilon)$ and, as ε goes to zero, $\varphi(x, \varepsilon) \sim 2\varepsilon f_a(b)$. This asymptotic is the same as obtained from the SmBP factorization: since the first PC is $\theta = a$ and $x_1 = b$, it holds $\varphi(x, \varepsilon) \sim f_1(x_1)\,\varepsilon\pi^{1/2}/\Gamma(1/2 + 1) = 2f_a(b)\,\varepsilon$, $\varepsilon \to 0$, with f_a being the *intensity* of the SmBP. Here, f_a was compared with its estimates $\widehat{f}_{1,n}$ from a sample of curves, for different $x(t)$, varying the nature of a and the sample size. We generated $1,000$ samples $\{X_i(t), i = 1, \ldots, n\}$, i.i.d. as $X(t)$, (with $n = 50, 100, 200, 500, 1,000$)

Table 1. Mean and standard deviation of RMSEP ($\times 100$) for Gaussian, t, and χ^2 distributions, computed over $1,000$ Monte Carlo replications varying the sample size n.

	$N(0,1)$		$t(5)/\sqrt{5/3}$		$\left(\chi^2(8)-8\right)/4$	
n	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
50	3.235	(2.681)	5.921	(2.557)	4.081	(2.842)
100	1.860	(1.444)	4.775	(1.503)	2.401	(1.619)
200	1.091	(0.824)	4.138	(0.878)	1.422	(0.887)
500	0.546	(0.355)	3.737	(0.477)	0.753	(0.443)
1,000	0.330	(0.220)	3.606	(0.327)	0.453	(0.233)

where every curve was discretized over a mesh consisting on 100 equispaced points $\{t_j = (j-1)\,\pi/99,\ j = 1,\ldots,100\}$. For each sample, we estimated the eigenfunction $\xi(t)$, the associated PC θ and its density via kernel procedure. Besides such samples, we built a set of curves $x^b(t) = b\sqrt{2/\pi}\sin(t)$ (discretized on the same grid as $X(t)$), where b is a suitable increasing sequence of real values. The estimated density $\widehat{f}_{1,n}$ was then evaluated at the points $\widehat{x}_1^b = \langle x^b(t), \widehat{\xi}(t)\rangle$ and compared with the true values $f_a(b)$ in term of relative mean square prediction error (RMSEP $= \sum_b [\widehat{f}_{1,n}\left(\widehat{x}_1^b\right) - f_a(b)]^2 \,/\, \sum_b f_a^2(b)$) over the $1,000$ replications. We also investigated for which values b the estimate of the surrogate density is better, by using the absolute percentage error (APE $= |\widehat{f}_{1,n}\left(\widehat{x}_1^b\right) - f_a(b)\,|/f_a(b)$). In the experiment we took a distributed as: i) $\mathcal{N}(0,1)$; ii) $t(5)/\sqrt{5/3}$; iii) $\left(\chi^2(8)-8\right)/4$. For b, we used sequences consisting of 160 equispaced points over the interval $[-4,4]$ for the distributions i) and ii), and $[-2,6]$ for the asymmetric distribution iii). The MSEP (multiplied by 100) obtained under the different experimental conditions are collected in Table 1. As expected, results improve as the sample size increases. This is due to the better estimates of projections $\widehat{\theta}$ and $\widehat{x}^b$ and to the better performances of the kernel estimator. On the other hand, differences due to the shape of distributions occur: long tails and asymmetries produce a deterioration in estimates. The APE (multiplied by 100) for some selected values b when $n = 200$ are reproduced in Figure 1. As one might expect, the quality of estimate worsens at the edges of the distributions, when b is rather far from zero. This fact is connected to the limitations of kernel density estimator in evaluating the tails of distributions.

6.2. Infinite dimensional setting

We dealt with an infinite–dimensional setting in order to study how the estimation of the *intensity* of the SmBP behaves according to the sample size and the dimensional parameter d. We considered a Wiener process X on $[0,1]$

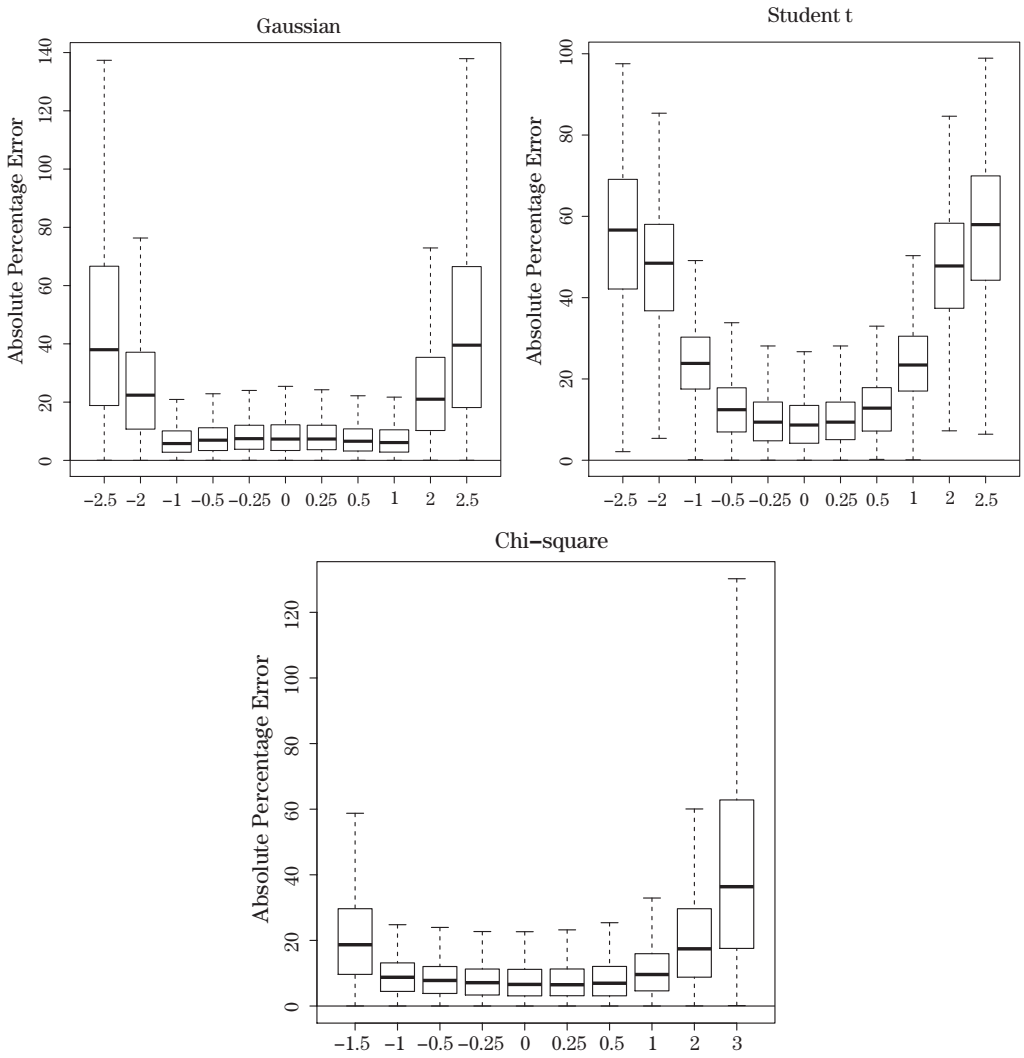

Figure 1. APE ($\times 100$) in estimating $f_a(b)$ varying b for Normal, t, and χ^2 distributions, respectively.

and the smooth function $x(t) = \sum_{j=1}^{d_0} b_j \xi_j(t)$ with, for the sake of simplicity, $d_0 = 1$,

$$x(t) = b\frac{2\sqrt{2}}{\pi}\sin\left(\frac{\pi t}{2}\right), \qquad t \in [0,1], \tag{6.1}$$

where $b \in \mathbb{R}$, so that the *intensity* is $\Psi_{d_0}(x) = \exp\left\{-b^2/2\right\}$. We generated $1,000$ samples $\{X_i(t), i = 1, \ldots, n\}$ (with $n = 50, 100, 200, 500, 1,000$), where every curve was discretized over 100 equispaced points $\mathcal{G} = \{t_j = (j-1)/99, j = 1, \ldots,$

Table 2. Mean and standard deviation (in parentheses) of RMSEP (×100) for Wiener process, computed over 1,000 Monte Carlo replications varying the sample size n and the dimension d.

n	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$
50	3.36 (2.51)	7.20 (3.73)	13.53 (7.34)	22.03 (12.05)	31.90 (15.87)	42.05 (19.16)
100	1.95 (1.20)	4.82 (2.59)	9.47 (5.54)	15.86 (8.71)	23.99 (12.27)	33.73 (15.88)
200	1.16 (0.72)	3.14 (1.60)	6.64 (3.78)	11.51 (6.30)	17.89 (9.48)	25.51 (13.10)
500	0.57 (0.33)	1.78 (0.93)	4.17 (2.36)	7.77 (4.23)	12.96 (6.88)	18.99 (9.28)
1,000	0.35 (0.19)	1.15 (0.63)	2.82 (1.64)	5.86 (3.13)	10.09 (5.43)	15.29 (7.65)

100} and 160 fixed curves $x^b(t)$ generated according to (6.1) and discretized over $\mathcal{G}$ (b was an increasing sequence of equispaced points, over the interval $[-4,4]$). For each sample, once empirical eigenfunctions $\widehat{\xi}_j(t)$ were obtained, we estimated f_d (with $d=1,\ldots,6$) and computed them at $\left(\widehat{x}_1^b,\ldots,\widehat{x}_d^b\right)'$ where $\widehat{x}_j^b = \left\langle x^b(t), \widehat{\xi}_j(t)\right\rangle$. Finally, we compared the estimated surrogate density with the true one in term of relative mean square prediction error (MSEP) over the 1,000 replications. The obtained results (multiplied by 100), varying n and d, are reported in Table 2. As a general comment, one can observe that, for each d, the MSPE reduces (both in mean and in variability) with increasing n, whereas, for each n, the MSPE increases (both in mean and in variability) with d. To perceive the relation between d and n, one has to read the table in a diagonal direction: it is possible to use large d at the cost of large samples. For instance, we got around 3% using $n=50$ and $d=1$, or $n=200$ and $d=2$, or when $n=1,000$ and $d=3$. On the other hand, results benefit from the fact that the spectrum of the process is rather concentrated. In fact, the Fraction of Explained Variance (defined as $\mathrm{FEV}(d)=\sum_{j\le d}\lambda_j/\sum_{j\ge 1}\lambda_j$) are: $\mathrm{FEV}(1)=0.811$, $\mathrm{FEV}(2)=0.901$, $\mathrm{FEV}(3)=0.933$, $\mathrm{FEV}(4)=0.950$, $\mathrm{FEV}(5)=0.960$ and $\mathrm{FEV}(6)=0.966$. Hence, good estimates for the surrogate density are already possible with $d=1$ or $d=2$, also for medium size samples.

Supplementary Materials

Proofs are collected in a supplementary document available on–line.

Acknowledgment

The authors thank the Editor, an associate editor and two anonymous referees for their constructive suggestions that led to improvement in the presentation of the paper. Special thanks go to G. González–Rodríguez and P. Vieu for fruit-

ful discussions and kind hospitality. The authors thank "Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni" (GNAMPA) of the "Istituto Nazionale di Alta Matematica" (INdAM) and CRoNoS (COST Action IC1408) for their support. The careful proofreading of an editorial assistant was appreciated.

References

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA.

Bongiorno, E. G. and Goia, A. (2016). Classification methods for hilbert data based on surrogate density. *Comput. Statist. Data Anal.* **99**, 204–222.

Bongiorno, E. G., Goia, A., Salinelli, E. and Vieu, P., eds. (2014). *Contributions in Infinite-Dimensional Statistics and Related Topics*. Società Editrice Esculapio.

Bosq, D. (2000). *Linear Processes in Function Spaces*, vol. 149 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.

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

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

Dereich, S., Fehringer, F., Matoussi, A. and Scheutzow, M. (2003). On the link between small ball probabilities and the quantization problem for Gaussian measures on Banach spaces. *J. Theoret. Probab.* **16**, 249–265.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *J. Stat. Softw.* **21**, 1–16.

Ferraty, F., Kudraszow, N. and Vieu, P. (2012). Nonparametric estimation of a surrogate density function in infinite-dimensional spaces. *J. Nonparametr. Stat.* **24**, 447–464.

Ferraty, F., Mas, A. and Vieu, P. (2007). Nonparametric regression on functional data: inference and practical aspects. *Aust. N. Z. J. Stat.* **49**, 267–286.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Series in Statistics. Springer, New York.

Gasser, T., Hall, P. and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**, 681–691.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34**, 1493–1517.

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

Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic Processes: Theory and Methods*, vol. 19 of *Handbook of Statist.* North-Holland, Amsterdam, 533–597.

Lifshits, M. A. (2012). *Lectures on Gaussian Processes*. Springer Briefs in Mathematics.

Springer, Heidelberg.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, 2nd edn.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B* **53**, 233–243.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.

Università degli Studi del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, Via Perrone 18, 28100, Novara, Italia.

E-mail: enea.bongiorno@uniupo.it

Università degli Studi del Piemonte Orientale, Dipartimento di Studi per l'Economia e l'Impresa, Via Perrone 18, 28100, Novara, Italia.

E-mail: aldo.goia@uniupo.it

(Received March 2016; accepted October 2016)

SOME INSIGHTS ABOUT THE SMALL BALL PROBABILITY FACTORIZATION FOR HILBERT RANDOM ELEMENTS

Enea G. Bongiorno and Aldo Goia

Università del Piemonte Orientale

enea.bongiorno@uniupo.it, aldo.goia@uniupo.it

Supplementary Material

This document collects the detailed proofs of the results presented in the paper "Some Insights About the Small Ball Probability Factorization for Hilbert Random Elements", by Enea G. Bongiorno and Aldo Goia.

Proof of Theorem 1

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

$$\begin{aligned}\varphi(x,\varepsilon) =& \mathbb{P}\left(S_1 + \varepsilon^2 S \le \varepsilon^2\right) = \mathbb{P}\left(S_1 \le \varepsilon^2\left(1-S\right)\right)\\ =& \mathbb{P}\left(\left\{S_1 \le \varepsilon^2\left(1-S\right)\right\} \cap \left\{S \ge 1\right\}\right) +\\ &+ \mathbb{P}\left(\left\{S_1 \le \varepsilon^2\left(1-S\right)\right\} \cap \left\{0 \le S < 1\right\}\right)\\ =& \mathbb{P}\left(\left\{S_1 \le \varepsilon^2\left(1-S\right)\right\} \cap \left\{0 \le S < 1\right\}\right)\\ =& \int_0^1 \varphi(s|x,\varepsilon,d)dG\left(s\right) \qquad \text{(S1.1)}\end{aligned}$$

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 $\{\xi_j\}_{j\leq d}$. In terms of $f_d(\cdot)$, the probability density function of $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_d)'$, it can be written as

$$\varphi(s|x, \varepsilon, d) = \int_{D^x} f_d(\boldsymbol{\vartheta})\, d\boldsymbol{\vartheta},$$

where $D = D^x = \left\{\boldsymbol{\vartheta} \in \mathbb{R}^d : \sum_{j\leq d} (\vartheta_j - x_j)^2 \leq \varepsilon^2 (1-s)\right\}$ is a 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 $\Pi x = \Pi_d x$,

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

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

$$\begin{aligned} \varphi(s|x, \varepsilon, d) =& \int_D \Bigg(f(x_1, \ldots, x_d) + \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \ldots, x_d)\rangle + \\ &+ \frac{1}{2}(\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t)(\boldsymbol{\vartheta} - \Pi x) \Bigg) d\boldsymbol{\vartheta} \\ =& f(x_1, \ldots, x_d) \int_D d\boldsymbol{\vartheta} + \int_D \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \ldots, x_d)\rangle\, d\boldsymbol{\vartheta} + \\ &+ \frac{1}{2}\int_D (\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t)(\boldsymbol{\vartheta} - \Pi x) d\boldsymbol{\vartheta} \\ =& f(x_1, \ldots, x_d) I + \\ &+ \frac{1}{2}\int_D (\boldsymbol{\vartheta} - \Pi x)' H_f(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t)(\boldsymbol{\vartheta} - \Pi x) d\boldsymbol{\vartheta} \quad \text{(S1.2)} \end{aligned}$$

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

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

and, the addend $\int_D \langle \boldsymbol{\vartheta} - \Pi x, \nabla f(x_1, \ldots, x_d)\rangle\, d\boldsymbol{\vartheta}$ is null since the integrand is a linear functional integrated over the symmetric – with respect to the center $(x_1, \ldots, x_d)$ – domain D. Thus from (S1.2), thanks to: the boundedness of second derivatives (2.3), the fact that symmetry arguments lead

to $\int_D (\vartheta_i - x_i)(\vartheta_j - x_j) d\boldsymbol{\vartheta} = 0$ for $i \neq j$ and monotonicity of eigenvalues, it follows

$$\begin{aligned}
&|\varphi(s|x,\varepsilon,d) - f(x_1,\ldots,x_d)I| = \\
&\qquad = \left| \frac{1}{2} \int_D \sum_{i\leq d} \sum_{j\leq d} (\vartheta_i - x_i)(\vartheta_j - x_j) \frac{\partial^2 f}{\partial \vartheta_i \partial \vartheta_j} \left(\Pi x + (\boldsymbol{\vartheta} - \Pi x)t\right) d\boldsymbol{\vartheta} \right| \\
&\qquad \leq \frac{1}{2} C_2 f(x_1,\ldots,x_d) \left| \sum_{i\leq d} \sum_{j\leq d} \int_D \frac{(\vartheta_i - x_i)(\vartheta_j - x_j)}{\sqrt{\lambda_i}\sqrt{\lambda_j}} d\boldsymbol{\vartheta} \right| \\
&\qquad = \frac{1}{2} C_2 f(x_1,\ldots,x_d) \int_D \sum_{j\leq d} \frac{(\vartheta_j - x_j)^2}{\lambda_j} d\boldsymbol{\vartheta} \\
&\qquad \leq \frac{C_2}{2\lambda_d} f(x_1,\ldots,x_d) \int_D \sum_{j\leq d} (\vartheta_j - x_j)^2 d\boldsymbol{\vartheta}.
\end{aligned}$$

Note that

$$\int_D \sum_{j\leq d} (\vartheta_j - x_j)^2 d\boldsymbol{\vartheta} = \int_{\|\boldsymbol{\vartheta}\|^2_{\mathbb{R}^d} \leq \varepsilon^2(1-s)} \|\boldsymbol{\vartheta}\|^2_{\mathbb{R}^d} d\boldsymbol{\vartheta}$$

whose integrand is a radial function (i.e. a map $H : \mathbb{R}^d \to \mathbb{R}$ such that $H(\boldsymbol{\vartheta}) = h(\|\boldsymbol{\vartheta}\|_{\mathbb{R}^d})$ with $h : \mathbb{R} \to \mathbb{R}$), for which the following identity applies

$$\int_{\|\boldsymbol{\vartheta}\|_{\mathbb{R}^d} \leq R} H(\boldsymbol{\vartheta}) d\boldsymbol{\vartheta} = \omega_{d-1} \int_0^R h(\rho) \rho^{d-1} d\rho,$$

where ω_{d-1} denotes the surface area of the sphere of radius 1 in $\mathbb{R}^d$. Hence

$$\int_{\|\boldsymbol{\vartheta}\|^2_{\mathbb{R}^d} \leq \varepsilon^2(1-s)} \|\boldsymbol{\vartheta}\|^2_{\mathbb{R}^d} d\boldsymbol{\vartheta} = \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_0^{\varepsilon\sqrt{1-s}} \rho^{d+1} d\rho = \frac{d}{(d+2)} I\varepsilon^2(1-s) \leq I\varepsilon^2,$$

where the latter inequality follows from the fact that $s \in [0,1)$. This leads to

$$|\varphi(s|x,\varepsilon,d) - f(x_1,\ldots,x_d)I| \leq C_2 \frac{\varepsilon^2 I}{2\lambda_d} f(x_1,\ldots,x_d). \tag{S1.3}$$

Come back to the SmBP (S1.1),

$$\varphi(x,\varepsilon) = \int_0^1 f(x_1,\ldots,x_d) I dG(s) + \int_0^1 \left(\varphi(s|x,\varepsilon,d) - f(x_1,\ldots,x_d)I\right) dG(s), \tag{S1.4}$$

and note that, thanks to (S1.3) and because d is fixed, the second addend in the right–hand side of (S1.4) is infinitesimal with respect to the first addend

$$\left|\frac{\int_0^1 \left(\varphi(s|x,\varepsilon,d) - f(x_1,\ldots,x_d)I\right) dG\left(s\right)}{\int_0^1 f(x_1,\ldots,x_d) I dG\left(s\right)}\right| \leq$$
$$\leq \left|\frac{C_2 \frac{\varepsilon^2}{2\lambda_d} f(x_1,\ldots,x_d) \int_0^1 I dG\left(s\right)}{f(x_1,\ldots,x_d) \int_0^1 I dG\left(s\right)}\right| = C_2 \frac{\varepsilon^2}{2\lambda_d}.$$

Noting that

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

we obtain

$$|\varphi(x,\varepsilon) - \varphi_d(x,\varepsilon)| \leq C_2 \frac{\varepsilon^2}{2\lambda_d} \varphi_d(x,\varepsilon) \tag{3.6}$$

where,

$$\varphi_d(x,\varepsilon) = f(x_1,\ldots,x_d) \frac{\varepsilon^d \pi^{d/2}}{\Gamma\left(d/2+1\right)} \mathbb{E}\left[\left(1-S\right)^{d/2} \mathbb{I}_{\{S\leq 1\}}\right]. \tag{3.5}$$

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

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

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

Proof Proofs of Proposition 1, and theorems 2 and 3

To prove Proposition 1 we need the following Lemma.

Lemma 1. *Assume (A-1) and (A-2). Then, it is possible to choose $d = d(\varepsilon)$ so that it diverges to infinity as ε tends to zero and*

$$\sum_{j\geq d+1} \lambda_j = o(\varepsilon^2). \tag{S1.5}$$

Moreover, 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.

Proof. A possible choice for $d = d(\varepsilon)$ satisfying (S1.5) can be, for a fixed $\delta > 0$, as follows

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

Such a minimum is well defined since eigenvalues series is convergent.
Let us prove that S converges to zero in probability. For any $k > 0$, by Markov inequality and, thanks to Assumption (A-2),

$$\begin{aligned}\mathbb{P}\left(|S| > k\right) = \mathbb{P}\left(S > k\right) &= \mathbb{P}\left(\frac{1}{\varepsilon^2}\sum_{j\geq d+1}\left(\theta_j - x_j\right)^2 > k\right) \\ &\leq \frac{\mathbb{E}\left[\frac{1}{\varepsilon^2}\sum_{j\geq d+1}\left(\theta_j - x_j\right)^2\right]}{k^2} \leq \frac{C_1}{k^2}\frac{\sum_{j\geq d+1}\lambda_j}{\varepsilon^2}. \qquad \text{(S1.6)}\end{aligned}$$

Thanks to (S1.5) we get the convergence 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 (S1.6), guarantees the almost sure convergence of S to zero (e.g. Shiryayev (1984, Theorem 10.3.1)) as ε tends to zero. Moreover, the monotone convergence theorem guarantees the L^1 convergence. ■

Proof of Proposition 1. Note that if $d(\varepsilon)$ satisfies $d\sum_{j\geq d+1}\lambda_j = o(\varepsilon^2)$, then (S1.5) and Lemma 1 hold. For a fixed $\delta > 0$, a possible choice of such $d = d(\varepsilon)$ can be

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

where the minimum is achieved thanks to the eigenvalues hyperbolic decay assumption.
At this stage, note that

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

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 (A-2),

we have (for any $d \geq 2$)

$$\begin{aligned}
0 \leq & 1 - \mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S<1\}}\right] \leq 1 - \mathbb{E}\left[\left(1-\frac{d}{2}S\right)\mathbb{I}_{\{S<1\}}\right] \\
\leq & \mathbb{P}(S \geq 1) + \mathbb{E}\left[\frac{d}{2}S\,\mathbb{I}_{\{S<1\}}\right] \leq \mathbb{E}\left[\frac{(d+2)}{2\varepsilon^2}\sum_{j\geq d+1}(\theta_j - x_j)^2\right] \\
\leq & \frac{C_1(d+2)}{2\varepsilon^2}\sum_{j\geq d+1}\lambda_j.
\end{aligned}$$

Choosing d according to $d\sum_{j\geq d+1}\lambda_j = o(\varepsilon^2)$ the thesis follows. ■

Proof of Theorem 2. Thanks to hyper–exponentiality (4.12), there exists $d_0 \in \mathbb{N}$ so that for any $d \geq d_0$

$$d\sum_{j\geq d+1}\lambda_j < \lambda_d.$$

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

$$0 \leq d\sum_{j\geq d+1}\lambda_j \leq b(d, \{\lambda_j\}_{j\geq d+1}, \delta_1) < B(d, \{\lambda_j\}_{j\leq d}, \delta_2) \leq \lambda_d, \qquad \text{(S1.7)}$$

where

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

As instance, for a given $d \geq d_0$, fix $\delta_1 \in (0,1)$ and solve (S1.7) with respect to δ_2, that is $\delta_2 \in (\min\{0, \beta(\delta_1)\}, 1)$ where $\beta(\delta_1) = 1 - (1-\delta_1)\ln\left(d\sum_{j\geq d+1}\lambda_j\right)/\ln(\lambda_d)$. As a consequence, for any $\varepsilon > 0$ and for such a choice of δ_1, δ_2, the following minimum is well–defined

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

This guarantees that the right–hand side of (4.10) vanishes as ε goes to zero. ■

To prove Theorem 3 we need the following Lemma.

Lemma 2. *Assume (A-1) and (A-2). Then, as $\varepsilon \to 0$,*

$$\mathcal{R}(x, \varepsilon, d)^{2/d} \to 1, \qquad \text{or,} \qquad \log(\mathcal{R}(x, \varepsilon, d)) = o(d). \qquad \text{(S1.8)}$$

Proof. Jensen inequality for concave functions (i.e. $\mathbb{E}[f(g)] \leq f(\mathbb{E}[g])$ if f is a concave function) guarantees that

$$\mathbb{E}\left[\left((1-S)\mathbb{I}_{\{S<1\}}\right)^{\frac{d}{2}}\right] = \mathbb{E}\left[\left((1-S)\mathbb{I}_{\{S<1\}}\right)^{\frac{d+1}{2}\frac{d}{d+1}}\right] \leq \left\{\mathbb{E}\left[\left((1-S)\mathbb{I}_{\{S<1\}}\right)^{\frac{d+1}{2}}\right]\right\}^{\frac{d}{d+1}},$$

noting that $S(x,\varepsilon,d+1) =: S_{d+1} \leq S_d := S(x,\varepsilon,d)$ and $\mathbb{I}_{\{S_d<1\}} \leq \mathbb{I}_{\{S_{d+1}<1\}}$, then

$$\mathbb{E}\left[\left((1-S_d)\mathbb{I}_{\{S_d<1\}}\right)^{\frac{d}{2}}\right] \leq \left\{\mathbb{E}\left[\left((1-S_{d+1})1_{\{S_{d+1}<1\}}\right)^{\frac{d+1}{2}}\right]\right\}^{\frac{d}{d+1}}.$$

The latter guarantees that $\mathbb{E}\left[(1-S)^{d/2}\,\mathbb{I}_{\{S<1\}}\right]^{2/d}$ is a non–decreasing monotone sequence with respect to d whose values are in $(0,1]$ and eventually bounded away from zero. ■

Proof of Theorem 3. Given results in Theorem 1, thesis holds using the same arguments as in Delaigle and Hall (2010, Proof of Theorem 4.2.): the idea is to combine together (S1.8), the Stirling expansion of the Gamma function in V_d and the (super–)exponential eigenvalues decay. ■

Proof of Theorem 4

In what follows, as in Section 5, we simplify the notations dropping the dependence on d for the density estimators f_n and $\widehat{f}_n$. Moreover, C denotes a general positive constant. The proof of Theorem 4 uses similar arguments as in Biau and Mas (2012).

Since $H_n = h_n^2 I$, it holds $K_{H_n}(u) = h_n^{-d} K(u)$. Consider

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

then the pseudo-estimator and the estimator are given by

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

and, hence,

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

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

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

then we have 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 $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned} \mathbb{E}\left[S_n(x) - \widehat{S}_n(x)\right]^2 \leq &2\mathbb{E}\left[\sum_{i=1}^{n} \left(K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right)\right) \mathbb{I}_{A_i \cap B_i}\right]^2 + \\ &+ 2\mathbb{E}\left[\left(\sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right) \mathbb{I}_{A_i \cap \overline{B}_i}\right)^2 + \right. \\ &\left. + \left(\sum_{i=1}^{n} K\left(\frac{\widehat{V}_i}{h_n}\right) \mathbb{I}_{\overline{A}_i \cap B_i}\right)^2\right]. \end{aligned} \tag{S1.9}$$

Consider now the first addend in the right–hand side of (S1.9): Assumption (B-3) and 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, lead to

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

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

$$C\mathbb{E}\left[\left\|\Pi_d - \widehat{\Pi}_d\right\|_\infty^2\right] \mathbb{E}\left[\left(\sum_{i=1}^{n} \|X_i - x\| \, \mathbb{I}_{A_i \cap B_i}\right)^2\right]. \tag{S1.10}$$

About the first factor in (S1.10), Biau and Mas (2012, Theorem 2.1 (ii)) established that

$$\mathbb{E}\left[\left\|\Pi_d - \widehat{\Pi}_d\right\|_\infty^2\right] = O\left(\frac{1}{n}\right). \tag{S1.11}$$

Consider now the second term in (S1.10). Thanks to the Chebyshev's algebraic inequality (see, for instance, Mitrinović et al. (1993, page 243)) and since $\mathbb{E}\left[\mathbb{I}_{A_i\cap B_i}\right] \leq \mathbb{E}\left[\mathbb{I}_{A_i}\right]$, for any $k \geq 1$ it holds

$$\mathbb{E}\left[\|X-x\|^k \mathbb{I}_{A_i\cap B_i}\right] \leq \mathbb{E}\left[\|X-x\|^k\right]\mathbb{E}\left[\mathbb{I}_{A_i}\right].$$

The fact that $\mathbb{E}\left[\mathbb{I}_{A_i}\right] \sim h_n^d$ and Assumption (B-4) give

$$\mathbb{E}\left[\|X-x\|^k \mathbb{I}_{A_i\cap B_i}\right] \leq C\frac{k!}{2}b^{k-2}h_n^d,$$

with $b > 0$. Hence, the Bernstein inequality (see e.g. Massart (2007)) can be applied: for any $M > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^n \|X_i-x\|\,\mathbb{I}_{A_i\cap B_i} - \mathbb{E}\left[\sum_{i=1}^n \|X_i-x\|\,\mathbb{I}_{A_i\cap B_i}\right]\right| \geq Mnh^d\right) \leq \\ \leq \exp\left(-CM^2nh^d\right).$$

This result, together with the Borel-Cantelli lemma, leads to:

$$\sum_{i=1}^n \|X_i-x\|\,\mathbb{I}_{A_i\cap B_i} \leq Cnh^d \qquad a.s.$$

and therefore,

$$\mathbb{E}\left[\left(\sum_{i=1}^n \|X_i-x\|\,\mathbb{I}_{A_i\cap B_i}\right)^2\right] \leq Cn^2h^{2d}. \tag{S1.12}$$

Finally, combining results (S1.11) and (S1.12), we obtain:

$$\frac{1}{(nh_n^d)^2}\mathbb{E}\left[\sum_{i=1}^n\left(K\left(\frac{V_i}{h_n}\right) - K\left(\frac{\widehat{V}_i}{h_n}\right)\right)\mathbb{I}_{A_i\cap B_i}\right]^2 \leq C\frac{1}{nh_n^2}. \tag{S1.13}$$

Consider now the second addend in the right–hand side of (S1.9). We only look at

$$\mathbb{E}\left[\sum_{i=1}^n K\left(\frac{V_i}{h_n}\right)\mathbb{I}_{A_i\cap\overline{B}_i}\right]^2, \tag{S1.14}$$

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

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

The latter inclusion and Assumption (B-3) allow to control (S1.14) by

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}_{A_i \cap \overline{B}_i}\right]^2 \leq & 2\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}_{\{h_n(1-\kappa_n) < V_i \leq h_n\}}\right]^2 + \\
& + 2\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{I}_{\left\{\left\|\widehat{\Pi}_d - \Pi_d\right\| \|X_i - x\| > C\kappa_n h_n\right\}}\right]^2. \qquad \text{(S1.15)}
\end{aligned}$$

About the first term in the right–hand side 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

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

Similarly, for what concerns the other addend in the right–hand side of (S1.15), we have

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

Thanks to the Markov inequality, Biau and Mas (2012, Theorem 2.1 (iii)) and Assumption (B-4), it follows

$$\mathbb{P}\left(\left\|\widehat{\Pi}_d - \Pi_d\right\| \|X - x\| > C\kappa_n h_n\right) = O\left(\frac{1}{n^{1/2} h_n \kappa_n}\right).$$

Combining the previous results we obtain:

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

If we choose $\kappa_n = \left(n^{5/2}h_n^{2d}\right)^{-1/2}$ with $n^{5/4}h_n^d \to \infty$, as $n \to \infty$, we obtain:

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} K\left(\frac{V_i}{h_n}\right)\mathbb{I}_{A_i\cap\overline{B}_i}\right)^2 + \left(\sum_{i=1}^{n} K\left(\frac{\widehat{V}_i}{h_n}\right)\mathbb{I}_{\overline{A}_i\cap B_i}\right)^2\right] \leq C\frac{1}{n^{5/4}h_n^{2d}}. \tag{S1.16}$$

In conclusion, (S1.13) and (S1.16) lead to:

$$\frac{1}{\left(nh_n^d\right)^2}\mathbb{E}\left[S_n(x) - \widehat{S}_n(x)\right]^2 = O\left(\frac{1}{nh_n^2}\right) + O\left(\frac{1}{n^{5/4}h_n^{2d}}\right).$$

Choose the optimal bandwidth (5.20) and $p > \max\{2, 3d/10\}$, then, as n goes to infinity, the first addend becomes negligible compared to the second one that turns to be $O\left(n^{-(10p-3d)/4(2p+d)}\right)$. Moreover, a direct computation shows that such bound is definitively negligible when compared to the "optimal bound" $n^{-2p/(2p+d)}$, for any $p > \max\{2, 3d/2\}$ and $d \geq 1$. This concludes the proof.

References

Biau, G., Mas, A., 2012. PCA-kernel estimation. Stat. Risk Model. 29 (1), 19–46.

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

Massart, P., 2007. Concentration inequalities and model selection. Vol. 1896 of Lecture Notes in Mathematics. Springer, Berlin, lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

Mitrinović, D. S., Pečarić, J. E., Fink, A. M., 1993. Classical and new inequalities in analysis. Vol. 61 of Mathematics and its Applications (East European Series). Kluwer Academic Publishers Group, Dordrecht.

Shiryayev, A. N., 1984. Probability. Vol. 95 of Graduate Texts in Mathematics. Springer-Verlag, New York.