Chapter 7
An asymptotic factorization of the Small–Ball Probability: theory and estimates

Jean-Baptiste Aubin, Enea G. Bongiorno and Aldo Goia

Abstract This work reviews recent results on an asymptotic factorization of the Small–Ball Probability of a $\mathscr{L}^2_{[0,1]}$–valued process, as the radius of the ball tends to zero. This factorization involves a volumetric term, a pseudo–density for the probability law of the process, and a correction factor. Estimators of the latter two factors are introduced and some of their theoretical properties considered.

7.1 Introduction

Since the seminal works of [9, 14], functional data analysis continues to massively attract the attention of researchers as proven by the recent monograph [5], special issues [8, 11] and activities [4] on the topic. In this framework, the Small–Ball Probality (SmBP) theory has been playing (and still now plays) an important role. It refers to the study of the asymptotic behaviour of $\mathbb{P}(X \in B(x,\varepsilon))$ as ε vanishes, where X is a random element taking its values in some topological space and $B(x,\varepsilon)$ denotes a suitable ball in such topology. From a theoretical point of view, researchers have mainly focused on different Gaussian processes and in providing the convergence rate (refer to the small tails probability theory; see [12, 13] and references therein). In functional regression, SmBP is a technical instrument used to express the convergence rate of estimators (see [9]). Recently, in the context of $\mathscr{L}^2_{[0,1]}$–valued random elements,

Jean-Baptiste Aubin
INSA-Lyon, ICJ, 20, Rue Albert Einstein, 69621 Villeurbanne Cedex, France, e-mail: jean-baptiste.aubin@insa-lyon.fr

Enea G. Bongiorno
Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Via Perrone, 18, 28100, Novara, Italy e-mail: enea.bongiorno@uniupo.it

Aldo Goia (✉)
Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Via Perrone, 18, 28100, Novara, Italy e-mail: aldo.goia@uniupo.it

G. Aneiros et al. (eds.), *Functional Statistics and Related Fields*,
Contributions to Statistics, DOI 10.1007/978-3-319-55846-2_7

the SmBP has been used to derive a concept of surrogate density and to introduce some non–parametric estimators for it (see [3, 6]). In particular, in [3] it has been shown that, for a fixed number d 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 principal components (PCs) evaluated at the center of the ball, (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 decay rate of the eigenvalues of the covariance operator of the process, it has been shown that the correction factor in (c) tends to 1 as the number of considered dimension increases (see [3]). This fact provides a clear advantage in modelling the SmBP since justifies the use of the lonely term (a) as a surrogate density of the process.

In this work, after recalling in Section 7.2 the theoretical conditions that lead to the mentioned factorization, we illustrate in Section 7.3 how to estimate the terms (a) and (c) providing asymptotic properties. The model advantages and potential applications are discussed in the last Section 7.4.

7.2 Framework and Notations

Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space and $\mathscr{L}^2_{[0,1]}$ 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 $(\Omega,\mathscr{F})$ taking values in $(\mathscr{L}^2_{[0,1]},\mathscr{B})$, where $\mathscr{B}$ denotes the Borel sigma–algebra induced by $\|\cdot\|$. Suppose $\mathbb{E}\|X\|^2<+\infty$ and denote by

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

its mean function and covariance operator respectively. Consider the Karhunen–Loève expansion of X: denoting by $\left\{\lambda_j,\xi_j\right\}_{j=1}^{\infty}$ the decreasing to zero sequence of positive eigenvalues and their associated orthonormal eigenfunctions of Σ, it holds

$$X(t)=\mu_X(t)+\sum_{j=1}^{\infty}\theta_j\xi_j(t),\qquad 0\le t\le 1, \tag{7.1}$$

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

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

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

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

(A-2) there exists a strictly positive constant C such that $x_j^2 \leq C\lambda_j$ for any $j \geq 1$, with $x_j = \langle x, \xi_j \rangle$;

(A-3) f_d is sufficiently smooth (differentiable p times) and there exists a strictly positive constant C for which, for any $d \in \mathbb{N}$

$$\sup_{i,j \in \{1,\ldots,d\}} \sqrt{\lambda_i \lambda_j} \left| \frac{\partial^2 f_d(\boldsymbol{\vartheta})}{\partial \vartheta_i \partial \vartheta_j} \right| \leq C f_d(x_1, \ldots, x_d), \qquad \text{for any } \boldsymbol{\vartheta} \in D,$$

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

Now, consider the small ball probability of the process X defined by

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

In [3], authors have proven that, for a given $d \in \mathbb{N}$ and under assumptions (A-1),..., (A-3),

$$\varphi(x, \varepsilon) \sim f_d(x_1, \ldots, x_d) \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2+1)} C(x, \varepsilon, d), \qquad \text{as } \varepsilon \to 0, \tag{7.2}$$

where

$$C(x, \varepsilon, d) = \mathbb{E}\left[(1 - S_d)^{d/2} \, \mathbb{I}_{\{S_d \leq 1\}} \right],$$

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

and $\mathbb{I}_A$ is the indicator function of the event A. Roughly speaking, (7.2) means that, for a given positive integer d and as $\varepsilon \to 0$, the SmBP $\varphi(x, \varepsilon)$ behaves as the usual first order approximation of the SmBP in a d–dimensional space (i.e. the probability density function of the first d PCs evaluated at $(x_1, \ldots, x_d)$ times the volume of the d–dimensional ball of radius ε) up to the scale factor $C(x, \varepsilon, d)$ that balances the use of a truncated version of the process expansion (7.1).

To fully split the dependence on x and ε in factorization (7.2), the following assumption can be considered:

(A-4) The eigenvalues $\{\lambda_j\}_{j \in \mathbb{N}}$ decay hyper–exponentially, that is $d \sum_{j \geq d+1} \lambda_j = o(\lambda_d)$, as $d \to \infty$.

Under (A-1),...,(A-4), it can be proven that, as d tends to infinity and for a suitable choice of $\varepsilon = \varepsilon(d)$,

$$\begin{cases} C(x, \varepsilon, d) \to 1, \\ \varphi(x, \varepsilon) \sim f_d(x_1, \ldots, x_d) \frac{\varepsilon^d \pi^{d/2}}{\Gamma(d/2+1)}, \end{cases} \tag{7.3}$$

see [3]. A practical choice for ε is

$$\varepsilon^2(d) = \sqrt{d\lambda_d \sum_{j=d+1}^{\infty} \lambda_j}. \tag{7.4}$$

To take advantage of the above factorizations, in the next section some estimators for f_d and C are introduced and their basic properties stated. Further discussions and applications are discussed in the last section.

7.3 Estimates

Consider $(X_1,\dots,X_n)$ a sample of RCs distributed as X, $\overline{X}_n$ and $\widehat{\Sigma}_n$ the empirical versions of μ_X and Σ from which it is possible to estimate the empirical eigensystem $\{\widehat{\lambda}_j,\widehat{\xi}_j\}_{j\in\mathbb{N}}$ and $\{\widehat{\theta}_{i,j} = \langle X_i,\xi_j\rangle\}_{j\in\mathbb{N}}$ the estimated scores associated to X_i for any $i = 1,\dots,n$. It is known that such estimators are consistent; see, for instance, [5].

For what concerns the surrogate density f_d, for a fixed d, let us introduce the kernel density estimate:

$$\widehat{f}_{d,n}\left(\widehat{\Pi}_d x\right) = \widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{H_n}\left(\left\|\widehat{\Pi}_d(X_i - x)\right\|\right) \tag{7.5}$$

where $K_{H_n}(\mathbf{u}) = \det(H_n)^{-1/2} K(H_n^{-1/2}\mathbf{u})$, K is a kernel function, H_n is a symmetric semi-definite positive $d\times d$ matrix and $\widehat{\Pi}_d$ denotes the projector onto the d–dimensional space spanned by $\{\widehat{\xi}_j\}_{j=1}^{d}$. Under regularity assumptions on f_d and on the kernel K, if one takes $H_n = h_n I$ with $h_n \to 0$ and $nh_n^d/\log n \to \infty$ as $n\to\infty$, the following result has been proven in [3].

Proposition 7.1. *Take the optimal bandwidth $c_1 n^{-1/(2p+d)} \le h_n \le c_2 n^{-1/(2p+d)}$ and $p > \max\{2, 3d/2\}$. Thus*

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

as n goes to infinity and uniformly in $\mathbb{R}^d$.

Regarding the corrective factor $C(x,\varepsilon,d)$ an estimator is provided by the empirical one:

$$\widehat{C}_{n,d} = \widehat{C}_n(x,\widehat{\varepsilon},d) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \widehat{S}_i(x,\widehat{\varepsilon},d)\right)^{d/2} \mathbb{I}_{\{\widehat{S}_i(x,\widehat{\varepsilon},d)\le 1\}},$$

with $\widehat{S}_i(x,\widehat{\varepsilon},d) = \widehat{\varepsilon}^{-2}\sum_{j\ge d+1}\left(\widehat{\theta}_{i,j} - \widehat{x}_j\right)^2$, $\widehat{\theta}_{i,j} = \langle X_i,\widehat{\xi}_j\rangle$, $\widehat{x}_j = \langle x,\widehat{\xi}_j\rangle$ and where $\widehat{\varepsilon}$ is the empirical version of (7.4). Asymptotics on such estimator have been provided in [1] and collected in the following proposition.

Proposition 7.2. *As n tends to infinity, $\widehat{\varepsilon}^2$ and $\widehat{C}_{n,d}$ are consistent estimator in the $L^1[\Omega,\mathscr{F},\mathbb{P};\mathbb{R}]$ metric. Moreover, $\sqrt{n}(\widehat{C}_{n,d} - C)$ is asymptotically normal distributed.*

7.4 Conclusions

This work collects some theoretical results concerning the factorization of the SmBP. These clarify those conditions under which it is possible to separate by means of distinct factors the spatial and volumetric components. The asymptotic (7.3) provides a modelling advantage: for d large enough, it justifies the use of factorized version of the SmBP since the corrective factor $C(x,\varepsilon,d)$ will be close to 1.
On the one hand, such approximation yields f_d a surrogate density of the process whose estimation can be tackled in a non–parametric manner (see [3, 6]) or parametrically (see [10] in the Gaussian mixture case). In [2], the estimate (7.5) is the starting point to build a pseudo–density oriented clustering algorithm where clusters are identified by the largest connected upper–surfaces containing only one mode. This technique was applied to different real datasets.
On the other hand, the convergence to 1 of $C(x,\varepsilon,d)$ holds theoretically only for $d \to \infty$. From the practical point of view, when d is fixed and in order to assess the goodness of f_d as a surrogate density, it is useful to evaluate how close to 1 is this correction factor C. This qualitatively suggests the dimension d to be used in practice (see [1]).

Acknowledgements The authors are grateful to R. Ignaccolo for the fruitful discussions they had at different stages of this work. E. Bongiorno and A. Goia 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) that funded a visit by J.B. Aubin at Università del Piemonte Orientale.

References

[1] Aubin J.B., Bongiorno E.G.: Optimal local dimension for suitable Hilbert–valued processes. Preprint (2017)

[2] Bongiorno E.G., Goia A.: Classification methods for Hilbert data based on surrogate density. Comput. Statist. Data Anal. **99**, 204 – 222, (2016)

[3] Bongiorno E.G., Goia A.: Some insights about the small ball probability factorization for Hilbert random elements. Statist. Sinica (To Appear) (2017)

[4] Bongiorno E.G., Goia A., Salinelli E., Vieu P. (eds): Contributions in infinite-dimensional statistics and related topics, Società Editrice Esculapio, (2014)

[5] Bosq D.: Linear processes in function spaces, Lecture Notes in Statistics, vol 149. Springer-Verlag, New York, (2000)

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

[9] Ferraty F., Vieu P.: Nonparametric functional data analysis. Springer Series in Statistics, Springer, New York (2006)

[8] Goia A., Vieu P.: An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal. **146**, 1–6, (2016)

[5] Horváth L., Kokoszka P.: Inference for functional data with applications. Springer Series in Statistics, Springer, New York, (2012)

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

[11] Kokoszka P., Oja H., Park B., Sangalli L.: Special issue on functional data analysis. Econometrics and Statistics, **1**, 99–10, (2017)

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

[13] Lifshits M.A.: Lectures on Gaussian processes. Springer Briefs in Mathematics, Springer, Heidelberg, (2012)

[14] Ramsay J.O., Silverman B.W.: Functional data analysis, 2nd edn. Springer Series in Statistics, Springer, New York (2005)