SIS | 2022

51st Scientific Meeting
of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

800 1222-2022 ANNI

Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Stata

sas

Università degli Studi del Sannio

Pearson

ISBN 9788891932310

A parsimonious approach to representing functional data

Un approccio parsimonioso per rappresentare dati funzionali

Enea G. Bongiorno and Aldo Goia

Abstract The correction term appearing in a Small ball probability factorization for functional Hilbert data is considered and some properties are presented. Such term leads to a new local dimensionality reduction method that allows a parsimonious representation of data. For the sake of illustration, this approach is applied to the Tecator dataset.

Abstract *Vengono descritte alcune proprietà del fattore correttivo che appare in una fattorizzazione della probabilità di una piccola bolla per dati funzionali in spazi di Hilbert. Questo termine correttivo porta a definire un nuovo metodo di riduzione della dimensionalità locale che permette una rappresentazione parsimoniosa dei dati. A fini illustrativi, questo approccio è applicato al dataset Tecator.*

Key words: Small ball probability factorization, local dimension, correction term

1 Introduction

One of the main problem in the functional data analysis, that is the toolkit of statistical methodologies to treat sample of curves, surfaces or other objects taking values in infinite dimensional spaces (for a review, see e.g. the monographes [3], [4] or [5]), is the representation of the data in small dimension.

To achieve the goal, a typical approach is to use a truncated version of the Karhunen–Loève decomposition: given a separable Hilbert space $\mathscr{H}$ equipped with an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\|\cdot\|$ and a functional random element X

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

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

taking values in $\mathscr{H}$ with mean μ and covariance operator Σ, one can write

$$X \approx \mu + \sum_{j=1}^{d} \xi_j v_j \qquad \mathbb{E}[\xi_j \xi_k] = \lambda_j \delta_{jk} \tag{1}$$

where $d < +\infty$, $\xi_j = \langle X, v_j \rangle$ are the so-called Principal Components (PCs) of X, (λ_j, v_j) are the eigenelements of the covariance operator of X, and $\delta_{jk} = 1$ if $j = k$, and zero otherwise. The quality of the approximation provided by (1) is often measured by the so-called fraction of explained variance (FEV), that is

$$FEV(d) = \frac{\sum_{j=1}^{d} \lambda_j}{\sum_{j=1}^{\infty} \lambda_j} 100\%$$

The proposed criterium is global: given a sample of functional data, a unique dimension d is selected for all the data. As a consequence, d could be too large for some of them and too small for other ones, thus producing inefficient or inadequate representations.

This paper aims to overcome this drawback, illustrating an approach to customize the choice of dimension for each element in a sample, through a local–based methodology. The latter exploits the properties of the correction term C_d appearing in the following Small-Ball Probability (SmBP) factorization (see [1] and [2]): given a positive integer d and a point $x \in \mathscr{H}$

$$\mathbb{P}(X \in B(x,h)) \sim f_d(x) V_d(h) C_d(x,h), \qquad h \to 0, \tag{2}$$

where $B(x,h)$ is the ball centred at x with radius h, $f_d(x)$ is the pdf of the first d PCs, V_d is the volume of the d–dimensional ball of radius h. In an intuitive way, for a fixed x, $C_d(x,h)$ provides a compensation for the use of the finite dimensional factorization $f_d V_d$. If that correction term is close to zero, it means that the selected dimension d is inadequate, because of the factorization $f_d V_d$ badly approximates the SmBP being x element of a space having dimension greater than d. On the other hands, if $C_d(x,h)$ reaches its maximum, d is a good choice to approximate x. These arguments allow to interpret $C_d(x,h)$ as a local measure of the quality of the representation of x as an element of a d-dimensional subspace of $\mathscr{H}$.

In this paper, this idea is described and applied. The outline is as follows: Section 2 illustrates the properties of the correction term that allow to interpret it as a quality index for a small-dimension representation of a functional data; in Section 3 a non-parametric estimate is introduced and an algorithm to select the dimensionality at x is described; finally, in Section 4 an application illustrates the advantages in using such an approach. More theoretical and computational details can be found in [1].

2 The correction factor in the SmBP factorization as quality index

This section collects some theoretical results that justify the use of the correction factor $C_d(x,h)$ as a measure of the quality in approximating x by means of a d-dimensional representation.

By definition, the correction term is:

$$C_d(x,h) = \mathbb{E}\left[\left(\left(1 - \frac{\|\Pi_d^{\perp}(X-x)\|^2}{h^2}\right)\mathbb{I}_{\left\{\|\Pi_d^{\perp}(X-x)\|^2 \le h^2\right\}}\right)^{d/2}\middle|\,\Pi_d x\right] \tag{3}$$

where Π_d denotes the projector onto $\mathscr{H}_d = span\{v_1,\ldots,v_d\}$, and $\Pi_d^{\perp}$ is its orthogonal projector. Note that $C_d(x,h) \in (0,1]$.

It can be proven that, varying $x \in \mathscr{H}$, the term $C_d(x,h)$ reaches its maximum over $\mathscr{H}_d$, as stated below.

Proposition 1. *Fix $h > 0$ and a strictly positive integer d, and suppose that the assumptions that guarantee the existence of the factorization (2) hold. Assume that the r.v. $((1 - \|\Pi_d^{\perp}(X-x)\|^2/h^2)\mathbb{I}_{\{\|\Pi_d^{\perp}(X-x)\|^2 \le h^2\}})^{d/2}$ is uncorrelated with $\{\Pi_d X = \Pi_d x\}$. Then, $C_d(x,h)$ admits a maximum $M_d(h)$ over $\mathscr{H}$ and it is achieved for any $x \in \mathscr{H}_d$.*

In other words, the maximum of $C_d(x,h)$ is reached for any x such that $\langle x, v_j\rangle = 0$ for any $j > d$. As a consequence, $C_d(x,h)$ helps to identify d to represent in small-dimension x: the closer $C_d(x,h)$ and $M_d(h)$ are, the more accurate the representation of x over the subspace $\mathscr{H}_d$ is, and adding further dimensions does not improve the quality of the representation.

Finally, the following characterization result can be stated:

Proposition 2. *Let X' be an independent copy of X, d a strictly positive integer and $h > 0$. Then the following statements are equivalent:*

i) $\mathbb{E}[C_d(X',h)] = 1$;
ii) $C_d(X',h) = 1$ *a.s.*;
iii) $\lambda_{d+1} = 0$;
iv) the process admits the d–dimensional representation $X = \sum_{j=1}^{d} \xi_j v_j$ *a.s.*.

3 A dimensionality selection algorithm

In order to make possible to use in practice the ideas described in the previous section, estimates of $C_d(x,h)$ and $M_d(h)$ must be defined. In this perspective, let $X_1,\ldots,X_n$ be a sample drawn from X. A Nadaraya–Watson type estimate of $C_d(x,h)$ is given by

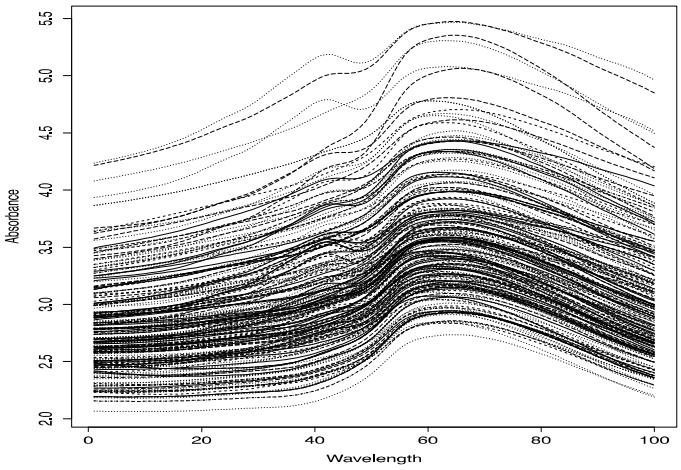

Fig. 1 The Tecator dataset.

$$\widehat{C}_{d,n}(x,h) = \sum_{i=1}^{n} \left(\left(1 - \frac{\|\widehat{\Pi}_d^{\perp}(X_i - x)\|^2}{h^2} \right) \mathbb{I}_{\left\{\|\widehat{\Pi}_d^{\perp}(X_i-x)\|^2 \le h^2\right\}} \right)^{d/2} \times \\ \times \frac{K(\|\widehat{\Pi}_d(X_i - x)\|/b)}{\sum_j K(\|\widehat{\Pi}_d(X_j - x)\|/b)}$$

where b is a bandwidth (in general depending on n), K a suitable kernel, $\widehat{\Pi}_d$ and $\widehat{\Pi}_d^{\perp}$ are the empirical estimates of the projectors Π_d and $\Pi_d^{\perp}$. For any d, an estimate of the upper bound $M_d(h)$ is provided by $\widehat{M}_{d,n}(h) = \max_i \widehat{C}_{d,n}(X_i,h)$.

At this stage, a procedure to select the local dimension can be defined. Given $\{\chi_j, j = 1,\dots,N\}$, possibly coincident with the sample, for each χ_j the local dimension $d_j^{\star}$, that should be used in (1), is selected as the smallest d for which $\widehat{C}_{n,d}(\chi_j,h)$ is close enough to $\widehat{M}_{d,n}(h)$. The proximity to this bound is quantified by considering if the relative measure $\widehat{C}_{n,d}(\chi_j,h)/\widehat{M}_{d,n}(h)$ is larger or smaller than a threshold $\alpha \in (0,1)$ suitably selected.

4 Application

To illustrate the performances of the local dimensionality selection algorithm described above, it is applied to the so–called Tecator dataset. It consists of near-infrared absorbance spectra of 215 chopped pieces of meat, discretized on 100 equally spaced wavelengths in the range $852-1050$ nm (these curves are visualized in the top panel of Figure 1).

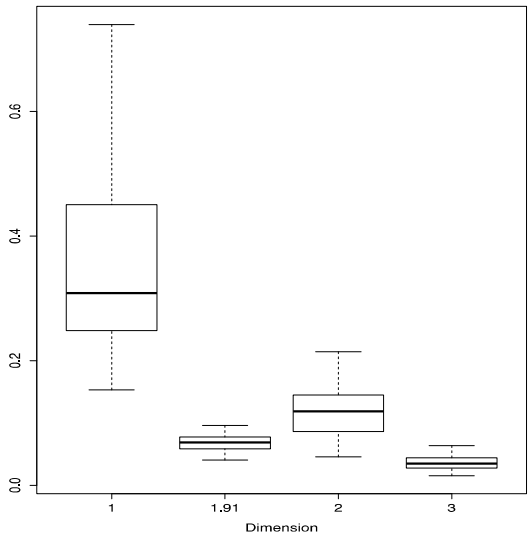

Fig. 2 Empirical distributions of the means of the ISEs, varying the dimension. The second (from the left) boxplot corresponds to the ISE computed when local dimensions are used.

The curves are rather smooth and a vertical shift appears: a good representation of them by using (1) can be obtained with the global dimension $d = 3$, that corresponds to a fraction of explained variance equal to 99.9%.

Clearly, that dimension could be too large for some of the curves in the dataset, and a parsimonious representations based on the algorithm, with a similar precision, can be adopted. In practice, the dataset is randomly split in two parts: the first one, containing 200 curves, is used to estimate the bounds $M_d(h)$ for $d = 1,\ldots,5$, whereas the remaining part $\{\chi_j, j = 1,\ldots,15\}$ is used to evaluate the local dimensions $d_j^\star$. The used kernel is the Epanechnikov one whereas the bandwidth is selected as the 10%–quantile of the distances between the curves in the training set projected by means of $\widehat{\Pi}_d$. The goodness of the approximation of χ_j by means of its $d_j^\star$-dimensional approximation $\chi_j^\star$ is measured by means of the Integrated Square Error (ISE), that is $\int_0^1 \left(\chi_j(t) - \chi_j^\star(t)\right)^2 dt$. The CV procedure is repeated 100 times: in each replication, the means of ISEs calculated by using both a global dimension d and the local one are computed, as well as the mean dimension $\overline{d}_m^\star$.

The choice of α is carried out by comparing the ISE behaviour varying α over a grid of possible values and the ISE obtained by using a global dimension. For this dataset, a good compromise is $\alpha = 0.87$ for which one gets $\overline{d}_m^\star = 1.91$ with a mean ISE equals to 0.068; if one compares this error with the one obtained when a global dimension is adopted, it is evident that the customization produces an efficient representation (see the distributions of the mean ISEs, multiplied by 100, in Figure 2).

References

1. Aubin, J.B., Bongiorno, E.G., Goia, A.: The correction term in a Small-Ball Probability factorization for random curves. Journal of Multivariate Analysis, In Press (2022)
2. Bongiorno, E.G., Goia, A.: Some insights about the Small Ball Probability factorization for Hilbert random elements. Statistica Sinica, **27**, 1949–1965 (2017)
3. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis. Springer Series in Statistics. Springer, New York (2006)
4. Horvath, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer Series in Statistics. Springer, New York (2012)
5. Ramsay, J.O., Silverman, B.W.: Functional Data Analysis, 2nd Edition. Springer Series in Statistics. Springer, New York (2005)