Computational Statistics and Data Analysis 58 (2013) 98–113

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Description length and dimensionality reduction in functional data analysis

D.S. Poskitt *, Arivalzahan Sengarapillai [1]

Department of Econometrics and Business Statistics, Monash University, Clayton, Victoria 3800, Australia

ARTICLE INFO

Article history:
Received 18 March 2010
Received in revised form 21 March 2011
Accepted 25 March 2011
Available online 5 April 2011

Keywords:
Bootstrap
Consistency
Dimension determination
Karhunen–Loève expansion
Signal-to-noise ratio
Variance decomposition

ABSTRACT

The use of description length principles to select an appropriate number of basis functions for functional data is investigated. A flexible definition of the dimension of a random function that is constructed directly from the Karhunen–Loève expansion of the observed process or data generating mechanism is provided. The results obtained show that although the classical, principle component variance decomposition technique will behave in a coherent manner, in general, the dimension chosen by this technique will not be consistent in the conventional sense. Two description length criteria are described. Both of these criteria are proved to be consistent and it is shown that in low noise settings they will identify the true finite dimension of a signal that is embedded in noise. Two examples, one from mass spectroscopy and the other from climatology, are used to illustrate the basic ideas. The application of different forms of the bootstrap for functional data is also explored and used to demonstrate the workings of the theoretical results.

1. Introduction

In the analysis of functional data, wherein each observation is a curve or image, it is commonly supposed that random curves or functions are sampled from a stochastic process X in $\mathcal{L}^2_{[0,\tau]}$. Here, $\mathcal{L}^2_{[0,\tau]}$ is the Hilbert space of square integrable functions on the interval $[0, \tau]$, with inner product $\langle f, g\rangle = \int_0^\tau f(t)g(t)\mathrm{d}t$ for any two functions $f, g \in \mathcal{L}^2_{[0,\tau]}$ and induced squared norm $\|\cdot\|^2 = \langle\cdot,\cdot\rangle$. A Karhunen–Loève expansion of X is also assumed to exist such that

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j\, \rho_j(t), \tag{1}$$

where the mean function $\mu(t) = E[X(t)]$ and the basis functions $\rho_j(t)$ are the orthonormal eigenfunctions of the covariance kernel $\Gamma(s, t) = \mathrm{Cov}[X(s), X(t)]$. The eigenvalues corresponding to $\rho_j(t)$ are listed in decreasing order, so that, without loss of generality, $\lambda_1 > \lambda_2 > \cdots$, where $\int_0^\tau \Gamma(s, t)\rho_j(t)\mathrm{d}t = \lambda_j\rho_j(s)$ and

$$\Gamma(t, s) = \sum_{j=1}^{\infty} \lambda_j\rho_j(t)\rho_j(s). \tag{2}$$

The coefficients ξ_j are given by the projection of $X - \mu$ in the direction of the jth eigenfunction ρ_j, *i.e.* $\xi_j = \langle X - \mu, \rho_j\rangle$. ξ_j constitute an uncorrelated sequence of random variables with zero mean and variance λ_j, and since the process X lies in $\mathcal{L}^2_{[0,\tau]}$ we have $\sum_{j=1}^{\infty} \lambda_j < \infty$. The monographs by Ramsay and Silverman (2002, 2005) present original expositions of

* Corresponding author. Tel.: +61 3 9905 9378; fax: +61 3 9905 5474.
E-mail addresses: Don.Poskitt@Buseco.monash.edu.au, dp526@york.ac.uk (D.S. Poskitt).

[1] Currently Department of Mathematics and Statistics, University of Jaffna, Sri Lanka.

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.03.018

various aspects of functional data analysis; see also Ferraty and Vieu (2006). Several more recent developments appear in the *Statistics for Functional Data* special issue of *Computational Statistics and Data Analysis* (2007, Volume 51, Issue 10), edited by González Manteiga and Vieu.

Although the series expansions in Eqs. (1) and (2) are infinite dimensional, it is often found that a given functional data set can effectively be spanned by $k \ll \infty$ basis functions. Truncating the expansions after k terms and expressing the functions in terms of a low dimensional, finite basis offers considerable practical advantages, not least because it allows various techniques of multivariate statistical analysis to be applied with little or no adaptation. Asymptotic analyses of random samples of X are commonly predicated on the assumption that the truncation point $k \to \infty$ as the number of sampled curves, n, increases; see, *inter alia*, Yao et al. (2005) and Hall et al. (2006). In practical applications, however, k is always finite and must be chosen by reference to the data. The choice of k is the main focus of this paper.

Various approaches for selecting k can be contemplated (Ramsay and Silverman, 2005, Section 4.5), but it is an open question as to how conventional dimension reduction methods can be adapted to the infinite dimensional setting of functional data (Ferraty and Vieu, 2006, Section 6.4). To the current authors' knowledge there is little in the current literature that explicitly investigates the theoretical properties of dimension reduction techniques within a functional framework. A notable exception is the work by Hall and Vial (2006), that builds upon the theoretical results presented in Hall and Hosseini-Nasab (2006). Hall and Vial assume a signal-plus-noise model for the observed process and consider determining k by examining the null hypothesis that the signal has fewer than k dimensions. They show that for such a model the noise will be confounded with the signal, and suggest that the intrinsic impossibility of estimating the full extent of the noise that results from this confounding means that conventional hypothesis testing techniques will not be effective. They therefore use the bootstrap to construct a lower bound for the un-confounded part of the noise variance and conclude that the assumed number of dimensions, k, is too small if the lower bound seems to be too large.

More direct methods are analyzed here, namely, the classical variance decomposition technique and choosing k using selection criteria. Yao et al. (2005, Section 2.5) proposed using a functional version of Akaike's information criterion to select k, justified via an appeal to a pseudo-Gaussian likelihood argument and the results of Shibata (1981). Here we consider the criteria constructed using optimal encoding, description length principles. This conceptual framework, which is reviewed in Hansen and Yu (2001), see also Rissanen (2007) and Grünwald (2007), provides a well established rationale that is directly applicable to the current functional data setting. We show below that it leads to techniques that circumvent confounding issues, and we develop the theoretical properties of the techniques within a functional data framework.

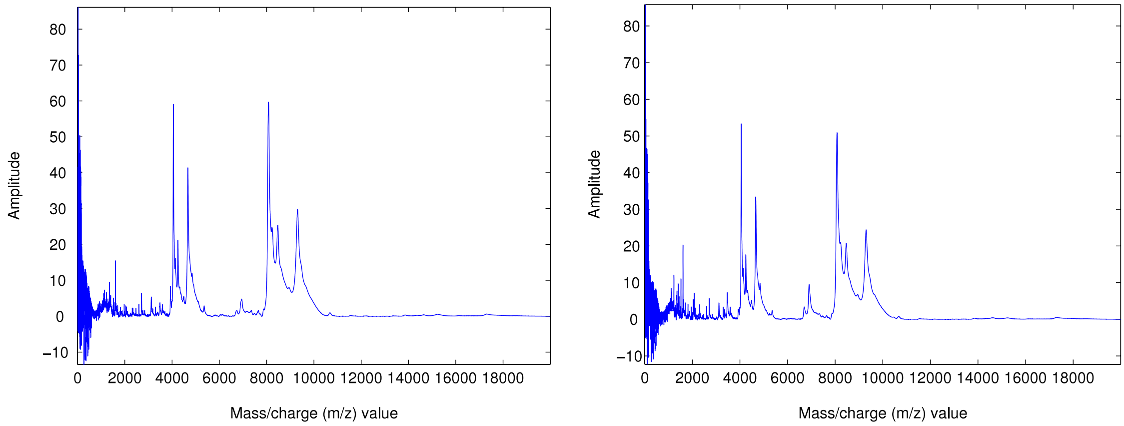
The paper proceeds as follows. The next section considers aspects of the basic structure of functional data, and introduces two examples that are used to illustrate basic ideas, the first taken from mass spectroscopy and the second from climatology. Section 3 develops some preliminary limit results under relatively weak regularity conditions. As part of the overall analysis, Section 4 provides a flexible definition of the dimension of X that depends on a signal-plus-noise decomposition derived from the Karhunen–Loève expansion of the function. By couching the concept of dimensionality directly in terms of the actually observed process the definition obviates the need to explicitly posit the existence of separate signal and noise processes, although data generating mechanisms that consist of a signal embedded in noise are encompassed as a special case and the definition coincides with the finite dimension of the signal in low noise settings. Section 5 discusses the classical variance decomposition technique. It is shown that statistics computed using this technique converge to their population counterparts, but, nevertheless, the dimension chosen by this method will not be consistent in the conventional sense. Section 6 examines two description length criteria for determining the dimension of functional data and proves that the criteria behave in a coherent manner asymptotically and that in low noise settings they will produce consistent estimates of the true finite dimension of the underlying signal. In Section 7, the data sets presented in Section 2 are used to illustrate the practical impact of the different methods considered. Section 8 examines the application and efficacy of different varieties of non-parametric and semi-parametric bootstrap, and using various versions of these demonstrate the working of the theoretical results. The proofs are assembled in the Appendix.

2. Basic data structures

Although the function X is defined on the interval $[0, \tau]$ it is seldom observed there, instead it is observed on a discrete subset of points. Here we will presume that each curve is observed on a grid of T points t_u, $u = 1, \ldots, T$, with $0 \le t_1 < t_2 < \cdots < t_T \le \tau$. Thus the raw data in a set $\mathcal{X} = \{X_1, \ldots, X_n\}$ of n observations on X will consist of an $n \times T$ data matrix $\mathbb{X} = [X_{su}]$ where

$$X_{su} = \mu(t_u) + \sum_{j=1}^{\infty} \upsilon_{s,j}\, \omega_j\, \rho_j(t_u), \tag{3}$$

for $s = 1, \ldots, n$ and $u = 1, \ldots, T$. In the expansion in (3), the $\upsilon_{s,j}$, $s = 1, 2, \ldots, n$, denote n realizations of $\upsilon_j = \xi_j/\omega_j$ where $\omega_j = \sqrt{\lambda_j}$, $j = 1, 2, \ldots,$ and ξ_j and λ_j are the random coefficients and variances that appear in Eqs. (1) and (2), respectively. In order to avoid excessively cumbersome notation, however, in what follows we will adopt the commonly employed convention of using the same symbolism for realizations of a stochastic process as for the process itself, and we will not distinguish between random variables and values of the variable. The required meaning to be attached to expressions that use this notational convention should be apparent from the context.



(a) Average proteomic spectrum of $n_1 = 100$ ovarian cancer patients.

(b) Average proteomic spectrum of $n_2 = 116$ healthy controls.

Fig. 1. Mass chromatograms from ovarian cancer data.

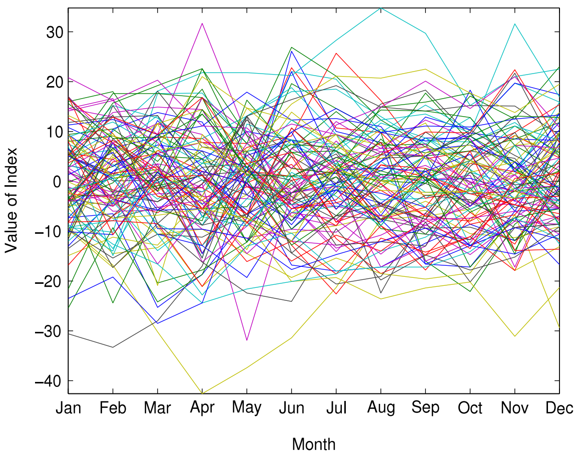


Fig. 2. Annual repeated measures on southern oscillation index: observed annual cycles in the period 1900–2004.

Example 1. The plots in Fig. 1 present mass chromatograms averaged across $n_1 = 100$ ovarian cancer patients, and $n_2 = 116$ healthy controls (including 16 individuals with benign tumors). The measurements were collected from a surface-enhanced laser desorption–ionization system (see Thiele, 2003; Banks, 2003). Each spectrum gives the relative amplitude measured at 15,154 mass–charge (μz) values on the interval [0, 18000]. Thus we have an overall sample of $n_1 + n_2 = n = 216$ curves, where each curve is the proteomic spectrum of an individual patient observed on a grid of $T = 15{,}154$ points. The ovarian cancer (OC) data was downloaded from http://clinicalproteomics.steem.com.

Despite being similar in appearance overall, the spectral profiles in Fig. 1(a) and (b) exhibit different features, witness the peaks at about $4000\mu z$ and $7000\mu z$ for example. The question of scientific interest here is whether or not differences in individual mass chromatograms can be reliably used to discriminate cancer patients from healthy controls and thereby construct a simple screening device.

Example 2. Fig. 2 presents monthly observations on the Southern Oscillation Index (SOI) for the period 1900–2004 inclusive, plotted as a sequence of repeated measures on the annual cycles. The data, constructed by the Australian Meteorological Office, can be downloaded from http://www.environment.gov.au. Treating each year as a single observation on a random function representing the annual cycle gives us a sample of $n = 105$ data points where each function is observed on a grid of $T = 12$ regularly spaced intervals. The division into yearly observations on annual cycles is natural because year-to-year variations in the SOI are thought to be very influential in determining annual weather patterns in the southern hemisphere — El Niño (drought) years and La Niña (precipitate) years. The idea of forming a functional time series from a univariate time series has been considered by several authors, including Besse et al. (2000), Ramsay and Ramsey (2002) and Ferraty and Vieu (2006, Chapter 12), where the merits of such an approach are discussed.

No obvious patterns emerge from simple visual inspection of the observed annual cycles. The values appear to fluctuate more or less randomly around zero, although there is evidence of some extreme negative values in the summer and autumn months, and extreme positive values in the winter and spring months. We would like to be able to determine if the apparently erratic behavior seen in Fig. 2 disguises more systematic patterns.

Let the observed mean of the data in $\mathbb{X}$ be $\bar{\mathbb{X}} = \mathbf{s}'\mathbb{X}/n$ where $\mathbf{s} = (1, \ldots, 1)'$ and set $\mathbf{C} = (\mathbf{I}_n - \mathbf{s}\mathbf{s}'/n)$, the centering matrix. Then the mean centered data matrix is given by $(\mathbb{X} - \mathbf{s}\bar{\mathbb{X}}) = \mathbf{C}\mathbb{X} = \mathbf{X}$, say. A standard approach for estimating the covariance kernel is to take

$$\bar{\Gamma}(t_u, t_v) = \frac{1}{n}\sum_{s=1}^{n}\{X_s(t_u) - \bar{X}(t_u)\}\{X_s(t_v) - \bar{X}(t_v)\}$$

as an estimator of $\Gamma(t_u, t_v)$ where $\bar{X}(t_u) = n^{-1}\sum_{s=1}^{n} X_s(t_u)$. Setting $\mathbf{G} = (\mathbb{X} - \mathbf{s}\bar{\mathbb{X}})'(\mathbb{X} - \mathbf{s}\bar{\mathbb{X}})/n = (\mathbf{X}'\mathbf{X})/n$ we have $\mathbf{G} = [\bar{\Gamma}(t_u, t_v)]$ for $u, v = 1, \ldots, T$.

Now let the singular value decomposition of $\mathbf{X}$ be denoted by

$$\mathbf{X} = n^{1/2}\mathbf{U}\mathbf{L}\mathbf{R}' \tag{4}$$

where $\mathbf{U}'\mathbf{U} = \mathbf{R}'\mathbf{R} = \mathbf{I}_m$, $m = rank(\mathbf{X}) = \min\{n-1, T\}$, and the diagonal matrix $\mathbf{L} = \mathrm{diag}(\sqrt{l_1}, \sqrt{l_2}, \ldots, \sqrt{l_m})$ where $l_1, \ldots, l_m$ lists the positive eigenvalues of $\mathbf{G}$ in descending order. The columns $\mathbf{u}_{\cdot 1}, \ldots, \mathbf{u}_{\cdot m}$ of $\mathbf{U}$ are the normalized eigenvectors of $\mathbf{X}\mathbf{X}'/n$ and the columns $\mathbf{r}_{\cdot 1}, \ldots, \mathbf{r}_{\cdot m}$ of $\mathbf{R}$ are the normalized eigenvectors of $\mathbf{G}$. The expansion in (4) provides an empirical counterpart to the Karhunen–Loève expansion in (3) in that a curve in $\mathbb{X}$ can be written as

$$X_s(t_u) = \bar{X}(t_u) + \sqrt{n}\sum_{j=1}^{m} u_{sj}\, w_j\, r_j(t_u), \tag{5}$$

where $w_j = \sqrt{\tau l_j/T}$ and $(\sqrt{\tau/T})r_j(t_u) = r_{uj}$, the uth element of $\mathbf{r}_{\cdot j} = (r_{1j}, \ldots, r_{Tj})'$. In addition, we have $\bar{\Gamma}(t_u, t_v) = (\tau/T)\sum_{j=1}^{m} l_j r_j(t_u) r_j(t_v)$, which in turn mimics the spectral decomposition of the covariance in (2). The pairs $(\tau l_j/T, r_j(t_u))$, which are of course the basic statistics of functional principle component analysis (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006), will be used to estimate the eigenvalue, eigenfunction pairs $(\lambda_j, \rho_j(t_u))$, $j = 1, \ldots, m$, and will form the fundamental building blocks of our subsequent practical methodology.

3. Some preliminary results

As we do not want to explicitly postulate the existence of separate signal and noise components that make up the data generating mechanism, our basic assumptions are presented in terms of the observed process X itself.

Assumption 1. The observed process $X \in \mathcal{L}^2_{[0,\tau]}$, has a Karhunen–Loève expansion as in (1), and covariance kernel $\Gamma(t, s)$ with spectral decomposition as in (2), where the eigenvalues $\lambda_j > 0, j = 1, \ldots$ are distinct.

Assumption 1 recognizes that functional data is obtained by observing realizations of a stochastic process that is inherently infinite dimensional. It is, in some ways, the functional analogue of the Wold representation employed in classical time series analysis.

Assumption 2. Let $\mathcal{X} = \{X_1, \ldots, X_n\}$ denote a sample of n observations on a process X where each curve is observed on a grid of T points t_u with $0 \le t_1 < t_2 < \cdots < t_T \le \tau$. Then for all $s = 1, \ldots, n$

$$E[X_s(t_u)] = \mu(t_u) \tag{6}$$

and

$$E[\{X_s(t_u) - \mu(t_u)\}\{X_s(t_v) - \mu(t_v)\}] = \Gamma(t_u, t_v). \tag{7}$$

Furthermore, for all $u = 1, \ldots, T$,

$$\limsup_{n\to\infty}\left(\max_{s=1,\ldots,n}\sum_{r=0}^{n-s}\mathrm{Cov}[X_s(t_u)X_{s+r}(t_u)]\right) < C_1 < \infty, \tag{8}$$

and setting $\Xi_s(uv) = \{X_s(t_u) - \mu(t_u)\}\{X_s(t_v) - \mu(t_v)\}$, then for all $u, v = 1, \ldots, T$,

$$\limsup_{n\to\infty}\left(\max_{s=1,\ldots,n}\sum_{r=0}^{n-s}\mathrm{Cov}[\Xi_s(uv)\Xi_{s+r}(uv)]\right) < C_2 < \infty. \tag{9}$$

The first part of Assumption 2 amounts to supposing that the observations behave like the realization of a weakly stationary functional process with a common mean function and covariance kernel. The second part places bounds on the auto-covariance of function values observed at different points along the abscissa, and provides sufficient conditions to ensure that $\bar{\mathbb{X}}$ and $\mathbf{G}$, the sample mean and covariance, will converge to their population counterparts $\boldsymbol{\mu} = (\mu(t_1), \ldots, \mu(t_T))$ and $\boldsymbol{\Gamma} = [\Gamma(t_u, t_v)]$, $u, v = 1, \ldots, T$, respectively.

Lemma 1. *Let* $\mathbf{d} = \bar{\mathbb{X}} - \boldsymbol{\mu}$ *and set* $\mathbf{D} = [D_{uv}]$ *where*

$$D_{uv} = \frac{1}{n}\sum_{s=1}^{n}[\{X_s(t_u) - \mu(t_u)\}\{X_s(t_v) - \mu(t_v)\}] - \Gamma(t_u, t_v) \quad u, v = 1, \ldots, T.$$

Then under Assumptions 1 and 2, the inequalities $\limsup_{n\to\infty} nE[\|\mathbf{d}\|^2] < 2C_1T$ *and* $\limsup_{n\to\infty} nE[\|\mathbf{D}\|^2] < 2C_2T^2$ *obtain for all* T.

It follows directly from Lemma 1, via Markov's inequality, that $\|\mathbf{d}\|^2 = o_p(T/n^{1-\beta})$ and $\|\mathbf{D}\|^2 = o_p(T^2/n^{1-\beta})$ for any $\beta, 0 < \beta \le 1$. From the expression $\mathbf{G} - \boldsymbol{\Gamma} = \mathbf{D} - \mathbf{d}'\mathbf{d}$ we can also deduce that $\|\mathbf{G} - \boldsymbol{\Gamma}\|^2 \le (\|\mathbf{D}\| + \|\mathbf{d}\|^2)^2$, and hence we can conclude that $\|\mathbf{G} - \boldsymbol{\Gamma}\|^2 = o_p(T^2/n^{1-\beta})$ and $\mathrm{plim}(\|T^{-1}(\mathbf{G} - \boldsymbol{\Gamma})\|^2) = 0$. These properties lead us to the following result.

Lemma 2. *Suppose that Assumptions 1 and 2 hold. Then for any* $\beta \in (0, 1]$

$$\sum_{j=1}^{m}(l_j - \ell_j)^2 = o_p(T^2/n^{1-\beta}) \tag{10}$$

as $n \to \infty$ *where* $\ell_j, j = 1, \ldots, T$, *denote the eigenvalues of* $\boldsymbol{\Gamma}$. *Moreover, if* $T \to \infty$ *then*

$$\max_{1\le j\le m}\left|\frac{l_j}{T} - \frac{\lambda_j}{\tau}\right| = o_p(1/n^{(1-\beta)/2}) + o(1) \tag{11}$$

and

$$\left|\sum_{j=1}^{m}\frac{l_j}{T} - \sum_{j=1}^{\infty}\frac{\lambda_j}{\tau}\right| = o_p(m/n^{(1-\beta)/2}) + o(m). \tag{12}$$

Interestingly enough, although our assumptions are sufficiently general to apply to time series type data, as in Example 2, the convergence rate for the eigenvalues of $o_p(n^{-(1-\beta)/2})$ given in Lemma 2 compares favorably with the $O_p(n^{-1/2})$ rate obtained by Hall and Hosseini-Nasab (2006) under simple random sampling.

By Lemma 1, $\bar{X}(t_u)$ is a consistent estimate of μ and by Lemma 2 the l_j/T provide consistent estimates of λ_j/τ. From the following lemma, we also know that the $r_j(t)$ estimate the basis functions $\rho_j(t)$ consistently.

Lemma 3. *Assume that Assumptions 1 and 2 hold and that* $(n, T) \to (\infty, \infty)$. *Let* $r_j(t)$ *be interpolating cubic smoothing splines that pass through the knots* $(t_u, \sqrt{T}r_{uj}/\sqrt{\tau})$, $u = 1, \ldots, T$, *respectively, where* $\mathbf{r}_{\cdot j} = (r_{1j}, \ldots, r_{Tj})'$ *is the jth eigenvector of* $\mathbf{G}$, $j = 1, \ldots, m$. *Then for any* $\beta \in (0, 1]$ *we have* $\|r_j - \rho_j\|^2 = o_p(1/n^{(1-\beta)/2}) + o(1)$ *and consequently* $r_j(t)$ *converges in probability to* $\rho_j(t)$ *in* $\mathcal{L}^2_{[0,\tau]}$.

It is not essential to assume uniform or balanced sampling. Alternative sampling schemes can be accommodated by considering kernel, or local smoothing, type estimates of $\Gamma(t, s)$ as described, for example, in Diggle and Verbyla (1998) and Yao et al. (2005). Similar consistency properties to those presented in Lemmas 1–3 will still hold, although the convergence rates may be different from those seen here, depending on the smoothing technique adopted. For some indication of the type of derivations and theoretical arguments that might need to be applied in such cases, see Boente and Fraiman (2000) and Yao et al. (2005, Section 3).

4. Signal +noise representations and dimension

Suppose that the Karhunen–Loève expansion of X is truncated after k terms. Then the finite expansion $S_k(t) = \mu(t) + \sum_{j=1}^{k}\xi_j\rho_j(t)$ can be used to approximate X. We can think of $S_k(t)$ as the signal component and the remainder, $N_k(t) = \sum_{j=k+1}^{\infty}\xi_j\rho_j(t)$, can be thought of as noise. Thus,

$$X(t) = S_k(t) + N_k(t) \tag{13}$$

yields an orthogonal decomposition of X in $\mathcal{L}^2_{[0,\tau]}$ that is optimal for a given k, in the sense that $S_k(t)$ provides the minimum mean squared error approximation to X. Moreover, $N_k(t)$ converges to zero as k increases. The decomposition in (13) holds true for all $k \in \mathbb{N} = \{1, 2, 3, \ldots\}$, however, and it follows that this decomposition cannot be used by and of itself to define the dimensionality of the process.

A solution of this problem is to introduce a measure of the relative magnitude of $S_k(t)$ and $N_k(t)$. Recall that the expression in (2) gives the spectral decomposition of $\Gamma(t, s)$, and the equality $\int_0^\tau \Gamma(t, t)dt = \sum_{j=1}^{\infty}\lambda_j$ is interpreted statistically as indicating the contribution of each term in (1) to the overall variance of X in $[0, \tau]$. If we set

$$\pi_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{\infty} \lambda_j}, \tag{14}$$

then $SNR(k) = \pi_k/(1-\pi_k)$ is therefore the natural measure of the signal-to-noise ratio of the decomposition in Eq. (13). Given α, where $0 \le \alpha < 1$, let $SNR_\alpha = \{k \in \mathbb{N} : SNR(k) \ge \alpha/(1-\alpha)\}$.

Definition 1. Let $k_\alpha \in SNR_\alpha$ be such that $k \ge k_\alpha$ for all $k \in SNR_\alpha$. Then X is said to be a process of dimension k_α at signal-to-noise ratio (SNR) level $\alpha/(1-\alpha)$.

It is important to recognize that in Definition 1 the dimension of X is determined *after* the signal-to-noise ratio has been assigned, the designated dimension being the smallest value of k such that SNR equals or exceeds the specified lower bound.

Suppose, following Hall and Vial (2006), that the observations are made up of realizations on an actual process of interest, $Y(t)$, that is in truth finite dimensional, to which a zero mean noise process, $\delta Z(t)$, representing experimental error, measurement error and so on, has been added. Thus

$$X(t) = Y(t) + \delta Z(t) \tag{15}$$

where $Y(t) = \mu(t) + \sum_{j=1}^{\kappa} \nu_j \varphi_j(t)$, $Z(t) = \sum_{j=1}^{\infty} \zeta_j\, \psi_j(t)$, and δ is a positive constant. The difficulty here is that (15) is observationally equivalent to

$$\begin{aligned} X(t) &= \mu(t) + \sum_{j=1}^{\kappa} \left(\nu_j + \delta \sum_{i=1}^{\infty} \zeta_i \beta_{ij} \right) \varphi_j(t) + \delta \sum_{j=\kappa+1}^{\infty} \left(\sum_{i=1}^{\infty} \zeta_i \beta_{ij} \right) \varphi_j(t) \\ &= Y'(t) + \delta Z'(t) \quad \text{say,} \end{aligned} \tag{16}$$

where the sequence $\varphi_1(t), \varphi_2(t), \ldots, \varphi_\kappa(t), \ldots$ is a complete orthonormal extension of $\varphi_j(t)$, $j = 1, \ldots, \kappa$, and $\beta_{ij} = \langle \psi_i, \varphi_j \rangle$ for $i, j = 1, 2, \ldots$, (see Hall and Vial, 2006, Section 2.1 and 2.2). It can be seen from (16) that the lower dimensional components of the noise $\delta Z(t)$ are confounded with those of the signal $Y(t)$, and that $Y'(t)$ and $\delta Z'(t)$ are orthogonal, so the original noise or error component cannot be identified.

In their analysis, Hall and Vial (2006) argue for a consideration of the low noise case, wherein the scale parameter $\delta \to 0$. They show that in this case $\delta^2 \sum_{j=\kappa+1}^{\infty} E[\zeta_j^2]$ – "the greatest knowable lower bound to all possible values of noise variance" – is identifiable and they use this as the benchmark for assessing noise levels. In empirical situations, however, the amount of noise need not be small and the representations in (15) and (16) are equivalent for all values of δ. Indeed, adopting a parallel development to that leading to (16), we also have

$$\begin{aligned} X(t) &= \mu(t) + \sum_{j=1}^{\kappa} c_j\, \psi_j(t) + \sum_{j=\kappa+1}^{\infty} c_j\, \psi_j(t) \\ &= S'_\kappa(t) + N'_\kappa(t), \end{aligned} \tag{17}$$

wherein $c_j = \sum_{l=1}^{\kappa} \beta_{jl} \nu_l + \delta \zeta_j$ for $j = 1, 2, \ldots$. In (17) the signal $Y(t)$ has been confounded with the noise $\delta Z(t)$ and the resulting decomposition is clearly observationally equivalent to (13) with $k = \kappa$. For moderate to large values of δ the question of what constitutes the dimension of the realized process X therefore remains moot without recourse to Definition 1.

From Definition 1 it is clear that k_α will depend on both the assigned level of resolution, as determined by α, and the structure of X. For the process in (15), for example, $SNR(k)$ can exceed $\alpha/(1-\alpha)$ for $k < \kappa$ if α is small, but need not do so if α is large. As $\delta \to 0$, however, $SNR(k)$ will exceed any value $\alpha/(1-\alpha) < \infty$ for all $k \ge \kappa$, and X will be deemed to be a process of dimension κ at SNR level $\alpha/(1-\alpha)$ for all $\alpha > \pi_{\kappa-1}$. To verify this let η_j, $j = 1, \ldots, \kappa$ and θ_j, $j = 1, 2, \ldots$, equal the eigenvalues of $\Gamma_Y(t,s) = \mathrm{Cov}[Y(t), Y(s)]$ and $\Gamma_Z(t,s) = \mathrm{Cov}[Z(t), Z(s)]$, respectively. Clearly $X(t)$ converges to $Y(t)$ in $\mathcal{L}^2_{[0,\tau]}$ as $\delta \to 0$ and since $Y(t)$ lies in the space spanned by $\varphi_j(1), \ldots, \varphi_j(\kappa)$,

$$\begin{aligned} \lambda_j &= \int_0^\tau \int_0^\tau \rho_j(t) \Gamma(t,s) \rho_j(s) \mathrm{d}t\mathrm{d}s \\ &= \int_0^\tau \int_0^\tau \varphi_j(t)(\Gamma_Y(t,s) + \delta^2 \Gamma_Z(t,s)) \varphi_j(s) \mathrm{d}t\mathrm{d}s + R_{\delta,j} \\ &= \begin{cases} \eta_j + \delta^2 \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2 + R_{\delta,j}, & j \le \kappa; \\ \delta^2 \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2 + R_{\delta,j}, & j > \kappa \end{cases} \end{aligned} \tag{18}$$

where $R_{\delta,j} = \int_0^\tau \int_0^\tau \Gamma(t,s)(\rho_j(t)\rho_j(s) - \varphi_j(t)\varphi_j(s))\mathrm{d}t\mathrm{d}s$. Now

$$\begin{aligned}|R_{\delta,j}| &\le \left\{\lambda_j + \left(\int_0^\tau \left[\int_0^\tau \varphi_j(t)\Gamma(t,s)\mathrm{d}t\right]^2 \mathrm{d}s\right)^{1/2}\right\} \|\rho_j - \varphi_j\| \\ &\le \left\{\lambda_j + \eta_j + 2\delta^2 \sum_{i=1}^{\infty} \theta_i\right\} \|\rho_j - \varphi_j\|\end{aligned}$$

and it can be shown, see the approximation lemmas in Hall et al. (2006, Lemma 1& 2, p. 1508), that $\|\rho_j - \varphi_j\|^2 = O(\delta^4)$. Note in addition that $\sum_{j=1}^{\infty} R_{\delta,j}$ is identically zero because $\sum_{j=1}^{\infty}\int_0^\tau\int_0^\tau \Gamma(t,s)\varphi_j(t)\varphi_j(s)\mathrm{d}t\mathrm{d}s$ equals $\sum_{j=1}^{\infty}\int_0^\tau\int_0^\tau \Gamma(t,s)\rho_j(t)\rho_j(s)\mathrm{d}t\mathrm{d}s = \sum_{j=1}^{\infty}\lambda_j$, the trace of the covariance kernel. Thus we find that

$$SNR(k) = \frac{\pi_k}{1-\pi_k} = \frac{\sum_{j=1}^{\kappa} \eta_j + O(\delta^2)}{O(\delta^2)}$$

for any $k \ge \kappa$ and hence $SNR(k)$ will exceed $\alpha/(1-\alpha)$ for any $\alpha \ge \pi_{\kappa-1}$ as $\delta \to 0$.

If $X(t) = Y(t) = \mu(t) + \sum_{j=1}^{\kappa} \nu_j\varphi_j(t)$ then it is straightforward to show that $k \in SNR_\alpha$ for all $k \ge \kappa$, and hence that $k_\alpha = \kappa$, for all $\alpha \in (\pi_{\kappa-1}, 1)$, as it should.

5. Variance decomposition

A commonly employed, classical approach for determining the number of sample principle components to retain in a description of an observed variance–covariance matrix is that based upon an examination of the proportion of variance explained. Thus, suppose that we are interested in accounting for $\alpha 100\%$ of the total variation in $\mathbb{X}$ where $0 < \alpha \le 1$. Then the variance decomposition method selects $\hat{k}_\alpha$ principle components where $\hat{k}_\alpha$ is the smallest value of k such that

$$\widehat{\pi}_k = \frac{\sum_{j=1}^{k} l_j}{\sum_{j=1}^{m} l_j} \tag{19}$$

equals or exceeds α. For the null model $\widehat{\pi}_0 \equiv 0$ and for the saturated model $\widehat{\pi}_m = 1$. For a detailed description of this and other methods see (Jolliffe, 2002, Chapter 6).

This approach is frequently adopted in the analysis of functional data (see, *inter alia*, Chiou and Li, 2007, Section 2.2.1) and in practice the value of $\hat{k}_\alpha$ is often chosen by reference to a graph of $\widehat{\pi}_k$ against k, similar to a 'scree plot'. Such a graph is monotonically non-decreasing in k with $\widehat{\pi}_k < \alpha$ for $k < \hat{k}_\alpha$ and $\widehat{\pi}_k \ge \alpha$ for $k \ge \hat{k}_\alpha$, and a popular rule-of-thumb is to look for the value of k that accounts for at least 75%–80% of the total variation.

Noting that π_k equals the proportion of variation in X attributable to the signal $S_k(t)$, and that $1-\pi_k$ equals that associated with the noise $N_k(t)$, we can see that the variance decomposition method is closely aligned with Definition 1. In particular, $\widehat{\pi}_k$ can obviously serve as an estimator of π_k.

Lemma 4. *Let π_k be defined as in (14) and $\widehat{\pi}_k$ as in (19), and suppose that Assumptions 1 and 2 hold. Then* $\max_{1\le k\le m}|\widehat{\pi}_k - \pi_k| = o_p(1/n^{(1-\beta)/2}) + o(1)$.

Similarly, the ratio $\widehat{SNR}(k) = \widehat{\pi}_k/(1-\widehat{\pi}_k)$ is the empirical counterpart to $SNR(k)$, and the above rule-of-thumb amounts to selecting k so as to obtain an observed signal-to-noise ratio $\widehat{SNR}(k) \ge \alpha/(1-\alpha)$ with the value of α *pre-assigned* by the practitioner to a value in excess of 0.75.

Theorem 1. *Assume that the conditions of Lemma 4 hold. Then $\widehat{SNR}(k)$ converges in probability to $SNR(k)$ for $k = 1, \ldots, m$ as $(n, T) \to (\infty, \infty)$. Furthermore, if X is a process of dimension k_α at SNR level $\alpha/(1-\alpha)$ then $|\hat{k}_\alpha - k_\alpha| = o_p(1)$.*

Theorem 1 indicates that the variance decomposition method behaves in a coherent way, in that the underlying statistics converge to their population counterparts. Implementing this technique as a means of selecting a dimension suitable for practical application requires the user to specify a value for α, however, and such a choice is *ad hoc*. Consider the signal-plus-noise process in (15). If $k < \kappa$ then, by Lemma 4,

$$\begin{aligned}\widehat{SNR}(k) &= \frac{\pi_k}{1-\pi_k} + o_p(1/n^{(1-\beta)/2}) + o(1) \\ &= \frac{\sum_{j=1}^{k}\eta_j + O(\delta^2)}{\sum_{j=k+1}^{\kappa}\eta_j + O(\delta^2)} + o_p(1/n^{(1-\beta)/2}) + o(1).\end{aligned}$$

Hence $\widehat{SNR}(k)$ converges to a value that will exceed $\alpha/(1-\alpha)$ if $\alpha < \pi_k$, but will remain bounded as $\delta \to 0$. It follows from Theorem 1 that $\text{plim}(\hat{k}_\alpha) = k_\alpha < \kappa$ for any $\alpha \le \pi_{\kappa-1}$. When $k \ge \kappa$, however, Lemma 4 indicates that

$$\widehat{SNR}(k) = \frac{\sum_{j=1}^{\kappa} \eta_j + O(\delta^2)}{O(\delta^2)} + o_p(1/n^{(1-\beta)/2}) + o(1),$$

implying that $k \in \widehat{SNR}_\alpha$ as $\delta \to 0$ and hence that $\hat{k}_\alpha = \kappa$ for all $\alpha \in (\pi_{\kappa-1}, 1)$ as $n \to \infty$. This indicates that the variance decomposition method does not yield a consistent estimate of κ in the conventional sense. We might attempt to retrieve the situation by setting $\alpha = \alpha(n)$ where $\alpha(n) \to 1$ as $n \to \infty$, but our current results provide no guide to a suitable choice.

6. Description length

Optimal encoding, description length principles lead to data generated rules for selecting k that will produce a finite dimensional representation of X that is as close an approximation as is possible, and uses the smallest number of parameters necessary, whilst adequately representing the structure and information contained in the data. Competing specifications are compared on the basis of their complexity, which is measured by reference to a criterion function. In the notation of this paper, one such criterion function is

$$CL_2(k) = \frac{n}{2}\log(V(k)) + \frac{k}{2}\log(n), \tag{20}$$

where the mean squared difference

$$V(k) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{u=1}^{T}\left(X_i(t_u) - \bar{X}(t_u) - \sqrt{n}\sum_{j=1}^{k} u_{ij}\, w_j\, r_j(t_u)\right)^2. \tag{21}$$

The function $CL_2(k)$ may be viewed as a two stage coding scheme, or code length, in which the first part represents the cost of the data compression and the second measures the code length used to encode the data when using k basis functions. The criterion $CL_2(k)$ achieves the stated goals since: (i) $TV(k) = \|\mathbf{G} - \widehat{\mathbf{G}}_k\|^2$ where $\widehat{\mathbf{G}}_k = \sum_{j=1}^{k} l_j \mathbf{r}_{\cdot j}\mathbf{r}_{\cdot j}'$ and if $\overline{\mathbf{G}}_k$ is a matrix of rank k used to approximate $\mathbf{G}$, then $\|\mathbf{G} - \overline{\mathbf{G}}_k\|^2$ is minimized at $\overline{\mathbf{G}}_k = \widehat{\mathbf{G}}_k$, and (ii) $CL_2(k)$ will exhibit a preference for smaller values of k, other things being equal. Note that if $V(k)$ in (21) is thought of as being the residual mean square from a multivariate regression, then $CL_2(k)$ can be seen to be analogous to BIC, after Schwarz (1978); but as pointed out in Grünwald (2007, Section 17.3) the connections between description length and information criteria are in general rather more subtle, and so we will continue to use $CL_2(k)$ to denote this criterion function.

To relate $CL_2(k)$ to the signal-plus-noise decomposition of X in (13), we can expand (21) and substitute into (20) to give $CL_2(k) = nDL_2(k)/2 + C_n$ where

$$DL_2(k) = \log(1 - \widehat{\pi}_k) + k\frac{\log(n)}{n} \tag{22}$$

and $C_n = \log(\frac{1}{nT}\sum_{i=1}^{n}\sum_{u=1}^{T}(X_i(t_u) - \bar{X}(t_u))^2)$ is a constant independent of k. The function $DL_2(k)$ gives us a description length per data point. The selected dimension, the minimum description length, is then given by $\widetilde{k}_2 = \arg\min_{0\le k<m} DL_2(k)$.

Description length criteria are not unique and an alternative criterion proposed by Rissanen (2000) for signal denoising is the, so-called, normalized minimum description length. In the current context this criterion gives rise to a consideration of

$$DL_N(k) = \log(1 - \widehat{\pi}_k) + \frac{k}{n}\log\left(\frac{\widehat{\pi}_k}{1 - \widehat{\pi}_k}\left\{\frac{nT - \nu(k)}{\nu(k)}\right\}\right) + \frac{1}{n}\log(\nu(k)(nT - \nu(k))), \tag{23}$$

wherein $\nu(k) = k(m+1) - \frac{1}{2}k(k+1)$ denotes the degrees of freedom in the kth singular value representation of the nT effective observations in $\mathbf{X}$. As above, the associated minimum description length, $\widetilde{k}_N$, is given by the value of $k \in \{0, \ldots, m-1\}$ that minimizes $DL_N(k)$. For a discussion of other encoding, description length schemes, see Hansen and Yu (2001), Rissanen (2007) and Grünwald (2007).

It is known that Schwarz criterion will produce consistent order selection under appropriate regularity conditions, including the assumption that the true data generating mechanism belongs to a finite union of parametric models. This raises the question of how, in the guise of $DL_2(k)$, it will behave under the current scenario. We therefore seek to characterize the properties of $DL_2(k)$, and $DL_N(k)$, when in truth X admits a Karhunen–Loève expansion as in (1) that cannot be a finitely parameterized, and the true structure of the process is unknown.

Towards this end, let us suppose that an oracle has told us the values of λ_j $j = 1, 2, \ldots$. Set $\overline{DL}_2(k) = \log(1 - \pi_k) + k\log(n)/n$ and let $\bar{k}_2$ denote the value of $k \in \{0, \ldots, m-1\}$ that minimizes $\overline{DL}_2(k)$. Similarly, let $\overline{DL}_N(k)$ denote the value obtained by replacing $\widehat{\pi}_k$ by π_k in $DL_N(k)$ and set $\bar{k}_N = \arg\min_{0\le k<m}\overline{DL}_N(k)$. Then for $a \in \{2, N\}$, the oracle will proclaim X to be a process of dimension $\bar{k}_a$ at SNR level $\alpha/(1-\alpha)$ where $\alpha = \pi_{\bar{k}_a}$.

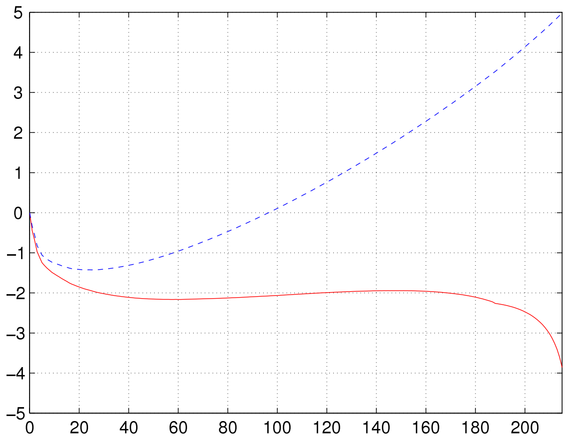


Fig. 3. Plots of $DL_N(k)$ (blue dotted line) and $DL_2(k)$ (red solid line) when computed from the OC data.

Theorem 2. *Suppose that Assumptions 1 and 2 hold. Then for both $a \in \{2, N\}$ we have $|\overline{DL}_a(k) - DL_a(k)| = o_p(1/n^{(1-\beta)/2}) + o(1)$ as $(n, T) \to \infty$. Furthermore, $\lim_{(n,T)\to(\infty,\infty)} Prob[\widetilde{k}_a = \bar{k}_a] = 1$.*

Theorem 2 indicates that for large values of n and T both $DL_2(k)$ and $DL_N(k)$ are likely to be close to the values that would be obtained by the oracle. In particular, a corollary of Theorems 1 and 2 is that for both $a \in \{2, N\}$ $\text{plim}|\widehat{SNR}(\widetilde{k}_a) - SNR(\bar{k}_a)| = 0$. Thus the criterion functions behave in a coherent manner and for (n, T) sufficiently large the practitioner will know the dimension of X that the oracle would have proclaimed and the value of SNR at which that proclamation would have been made.

We have already seen that the signal-plus-noise process X in (15) has dimension κ at SNR level $\alpha/(1-\alpha)$ for all $\alpha > \pi_{\kappa-1}$ as $\delta \to 0$. In order to relate this to the values of k selected by the description length criteria let us introduce an additional assumption.

Assumption 3. There exist constants C, $0 < C < \infty$, and d, $0 < d < 2$, such that $\inf_i \beta_{ij}^2 \geq Cg^j$, for all $j = 1, \ldots,$ where $g^2 = 1 - \delta^{2-d}$.

The generalized Fourier coefficients β_{ij} lie in the unit interval [0, 1] because $\sum_{j=1}^{\infty} \beta_{ij}^2 = 1$ and Assumption 3 bounds the coefficients away from zero. This ensures that the contribution of the noise to the overall variation of X on $[0, \tau]$ cannot be null, and that the components of $Z(t)$ that are orthogonal to $Y(t)$ cannot be identically zero.

Theorem 3. *Suppose that $X(t) = Y(t) + \delta Z(t)$, as in (15). Suppose also that $Y(t)$ and $Z(s)$ are uncorrelated for all t and s, and that Assumptions 1 and 2 hold. Then the probability that the event $\widetilde{k}_a \geq \kappa$ obtains converges to 1 as $(n, T) \to \infty$ for both $a \in \{2, N\}$. Furthermore, if Assumption 3 holds and $\delta \to 0$ such that $\delta^{2-d} nT/(n+T) \to 0$, then $\lim_{(n,T)\to(\infty,\infty)} Prob[|\widetilde{k}_a - \kappa| > \delta] = 0$ and X will be deemed to be a process of dimension κ at level SNR_α for all $\alpha > \pi_{\kappa-1}$.*

In functional data analysis it is usually preferable to use as small a number of components as possible ($k \ll \infty$) and, consequently, in practice an upper bound for k is often chosen that is much smaller than both n and T. In the previous analysis it has been assumed that the upper bound, k_{max} say, coincides with m, the rank of the centered data matrix, but other choices of k_{max} based on n and T with $k_{max} < m$ are compatible with the results presented here provided $k_{max} \to \infty$ as $m \to \infty$.

7. Illustrations

The asymptotic results presented above require that $(n, T) \to (\infty, \infty)$, but they do not impose any further restrictions on the orders of magnitude of n and T. Thus they can be thought of as being applicable to both of the examples presented previously even though the relative sizes of n and T in the two cases are very different.

Example 1. Fig. 3 plots the values of $DL_2(k)$ and $DL_N(k)$ for k in the range $0 \leq k < m$ when computed from the OC data. The figure clearly illustrates that the components of $DL_N(k)$ can counter-balance each other in such a way that the criterion has a well defined minimum at a relatively small value of k. Thus we find that $\widetilde{k}_N = 25$.

The behavior of $DL_2(k)$ merits elaboration. Starting at the origin, as k increases $DL_2(k)$ exhibits two turning points, a local minimum at $k = 57$ and a local maximum at $k = 147$, before finally reaching a global minimum at the saturation boundary. Thus, as $k \to m$ and $\widehat{\pi}_k \to 1$ the increase in $k\log(n)/n$ is no longer large enough to counteract the decrease in $\log(1 - \widehat{\pi}_k)$. This presents a problem if we continue to search for the global minimum since $\widehat{\pi}_k \geq 0.995$ for $k \geq 135$, suggesting that any variation due to bases with an index greater than 135 is very small and should be attributed to the noise

Table 1
Selected dimensions for OC data.

Criterion	k	$\widehat{\pi}_k$	$\widehat{SNR}(k)$
$\hat{k}_{0.75}$	6	0.7697	3.3424
$\hat{k}_{0.99}$	103	0.9901	100.0672
$\widetilde{k}_2$	57	0.9722	34.9152
$\widetilde{k}_N$	25	0.9235	12.0746

Table 2
Classification results for OC data.

Criterion	k	Overall error (%)	Sensitivity (%)	Specificity (%)
$\hat{k}_{0.75}$	6	19.44	83	74.85
$\hat{k}_{0.99}$	103	14.81	73	95.69
$\widetilde{k}_2$	57	10.65	83	94.83
$\widetilde{k}_N$	25	14.81	79	90.52

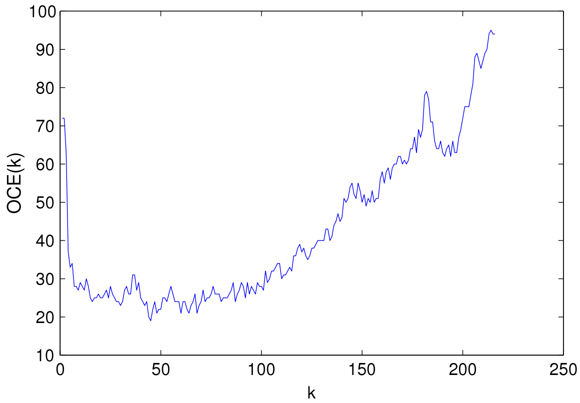


Fig. 4. Jackknife classification errors for OC data.

rather than the signal component. In this case, a straightforward solution is to restrict the search to $k \in \{0, \dots, \hat{k}_\alpha - 1\}$ for some $\alpha > 0.99$, say. Since, by Theorem 1, $\hat{k}_\alpha$ converges to k_α use of this rule implies that the smallest $(1-\alpha)100\%$ of the variation in X is being assigned to the noise component. Note that the common practice, mentioned at the end of Section 6, of choosing $k_{\max}$ on the basis of the sample size and the grid points such that $k_{\max} < m$ is equivalent to restricting the search to $k \in \{0, \dots, k_{\alpha_{\max}} - 1\}$ where $k_{\alpha_{\max}} = k_{\max}$ for some $\alpha_{\max} < 1$, and where $\alpha_{\max} \to 1$ as k_{max} increases. Using this device results in the criterion selecting $\widetilde{k}_2 = 57$.

We also evaluated $\hat{k}_\alpha$ using values of α that bound those recommended in Chiou and Li (2007, Section 2.2.1), namely $\hat{k}_{0.75} = 6$ and $\hat{k}_{0.99} = 103$. These values clearly indicate the sensitivity of $\hat{k}_\alpha$ to the assigned level of resolution. A range of 98 possible values for k is too broad to be of any help in deciding what dimension to actually use in practice, but $\hat{k}_{0.75}$ and $\hat{k}_{0.99}$ do provide useful points of comparison.

The selected dimensions are reproduced in Table 1, together with their associated estimates $\widehat{\pi}_k$ and $\widehat{SNR}(k)$. Table 2 presents the results obtained when different dimensions are used in conjunction with the non-parametric functional classification procedure introduced in Hall et al. (2001) to discriminate cancer patients from healthy controls; the Hall et al. (2001) method projects the data into the space of the first k components and discrimination then takes place in this space using non-parametric density estimates to evaluate the likelihood of different types. The overall error rate, and the sensitivity and specificity, were calculated using the *jackknife* or the *leave-one-out* method. The relative merits of the different values of k seen in Table 1 are not directly mirrored in the measures given in Table 2; in particular, the miss-allocation rates are not monotonic in k. The contrast between Tables 1 and 2 reflects that changes in mean squared approximation error are indexed by changes in k, but changes in k are neither explicitly linked to, nor necessarily indicative of, changes in classification rates. This later feature is clearly seen in Fig. 4. Fig. 4 plots the total number of jackknife classification errors as a function of k ($\mathrm{OCE}(k)$) and demonstrates that the lack of monotonicity in k seen in Table 2 is both a local and a global phenomenon.

The use of jackknife cross-validation to determine the number of components that minimizes the number of classification errors leads to the selection of $k = k_{CV} = 45$. (See Grünwald, 2007, Section 17.6 for a discussion of the links between cross-validation and description length criteria.) This gives an overall error rate of 9.26%, a sensitivity of 83% and a specificity of 97.4%. By construction AIC will select a value of $k \geq \widetilde{k}_2 = 57 > k_{CV} = 45 > \widetilde{k}_N = 25$, indicating that description length criteria can select dimensions much smaller than any of those chosen by popular benchmark criteria. Such outcomes

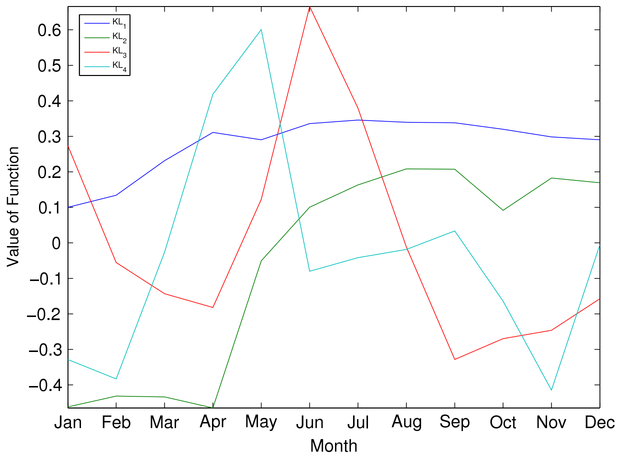


Fig. 5. Dominant basis functions for SOI data.

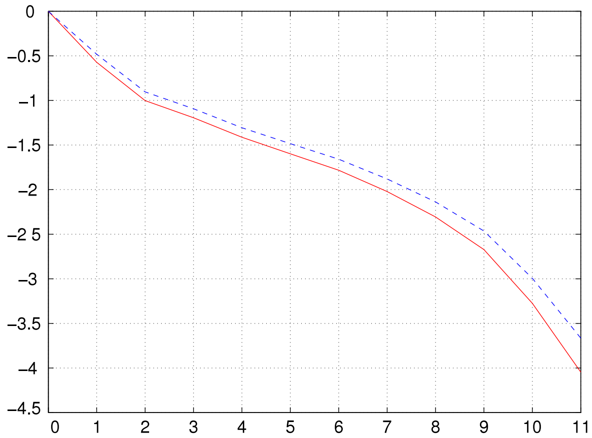


Fig. 6. Plots of $DL_N(k)$ (blue dotted line) and $DL_2(k)$ (red solid line) when computed from the SOI data.

intimate that for the OC data the best choice of k (and basis functions) for reliable discrimination remains uncertain. Nevertheless, it is apparent that even functional data that is observed on a grid of several thousands of points can be reduced to as few as fifty, or even twenty, or so dimensions whilst maintaining very creditable performance.

Example 2. Upon examination of $\widehat{\pi}_k$ for the SOI data we find that the first four basis functions account for 79.61% of the observed annual variation, suggesting that the variance decomposition method used in conjunction with the commonly employed rule-of-thumb would select $k = 4$. The first four basis functions are plotted in Fig. 5. Although it is not difficult to imagine different combinations of these basis functions giving rise to the different curves seen in Fig. 2; from Fig. 6, which plots the values of $DL_2(k)$ and $DL_N(k)$ for k in the range $0 \leq k < m$, we can see that both criteria will select the most profligate model available.

These outcomes suggest that the behavior of SOI observed in Fig. 2 cannot be attributed to variation about more dominant, common annual cycles. Rather, the oscillations and extremes are due to aberrant values of the SOI being generated throughout particular years, suggesting that predicting the so-called "g-phases", as discussed in Stone et al. (2000), could be a useful tool in forecasting future El Niño/La Niña effects and their associated weather patterns.

It should be emphasized that the investigations presented in Examples 1 and 2 are illustrative, they are conducted in the spirit of exploratory data analysis and are not meant to be definitive. Clearly, if classification is the ultimate objective with the OC data then a more classification orientated functional approach may be more appropriate, *c.f.* Ferraty and Vieu (2003), Chiou and Li (2007) and Li and Yu (2008). If the aim of the analysis of the SOI data is the prediction of future weather patterns, considerations of functional linear regression, *à la* Hall and Hosseini-Nasab (2006) and Ferraty and Vieu (2009), will be relevant. For both functional discrimination and functional regression the choice of basis functions (functional principal components) can be critical and a more goal orientated approach to criterion construction and evaluation than that considered here is likely to be necessary. One such criterion can be obtained by replacing $\widehat{\pi}_k$ in (22) and (23) by the coefficient of determination from a functional regression, or equivalently a logit/probit functional regression in the case of functional discrimination, along the lines of the description length regression criteria discussed in Rissanen (2007, Section 9.3) and Grünwald (2007, Section 14.5). A thorough examination of such possibilities would take us too far afield here however and must be left for future research.

In both examples, the observations are made on a uniform grid. For the OC data this does not present a problem since the mass-spectrometry readings are taken at $T = 15154$ points on the μz axis. For the SOI data, however, the number of grid points is only 12. Such a small number obviously flies in the face of the asymptotic requirement that $T \to \infty$. More significantly, when T is small $(\tau l_j/T, r_j(t_u))$ may not yield a reliable estimate of $(\lambda_j, \rho_j(t_u))$ and it may be preferable to use an estimation method designed for sparse functional data such as the one proposed by Yao et al. (2005).

In spite of such qualifications, the examples amply demonstrate the potential usefulness of the application of description length principles in the context of functional data analysis.

8. The bootstrap

Given the raw data $\mathcal{X} = \{X_1, \ldots, X_n\}$ of n observations on X, an obvious way to get some idea of the sampling variability of a statistic of interest is to re-sample from $\mathcal{X}$ and construct a bootstrap replication $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$. By repeatedly generating different bootstrap replications an approximation to the statistic's distribution can be constructed. This is precisely the technique employed in Hall and Vial (2006). Here we wish to investigate the application and efficacy of different forms of the bootstrap.

To begin, observe that the bootstrap replications are obtained by re-sampling from the rows of

$$\mathbb{X} = \mathbf{s}\bar{\mathbb{X}} + n^{1/2}\mathbf{ULR}', \tag{24}$$

wherein the right-hand side is the matrix-vector equivalent of (5). Writing $\mathbb{X}^*$ for a bootstrap data matrix, we have

$$\mathbb{X}^* = \mathbf{S}\mathbb{X} = \mathbf{Ss}\bar{\mathbb{X}} + n^{1/2}\mathbf{SULR}' = \mathbf{s}\bar{\mathbb{X}} + n^{1/2}\mathbf{U}^*\mathbf{LR}', \quad \text{say}, \tag{25}$$

where $\mathbf{S}$ represents a randomly chosen $n \times n$ selection matrix. From (25) we can see that the bootstrap replications of the process can be generated in the following manner: Bootstrap Step

B1. Hold the mean $\bar{X}(t_u)$, the eigenvalues $l_j, j = 1, \ldots, m$, and the basis functions $r_j(t_u), j = 1, \ldots, m$, fixed at their realized values.
B2. For $i = 1, \ldots, n$ generate bootstrap replications $u_{sj}^*, j = 1, \ldots, m$, by taking independent and identically distributed (i.i.d.) random draws from $u_{sj}, j = 1, \ldots, m$.
B3. Construct the functional data re-sample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ where, for $s = 1, \ldots, n$, the realization $X_s(t_u)^*, u = 1, \ldots, T$, is constructed as in (5) by replacing $u_{sj}, j = 1, \ldots, m$ by $u_{sj}^*, j = 1, \ldots, m$.

The consistency properties presented in Lemmas 1–3 indicate that for (n, T) reasonably large the empirical expansion in (5) will provide a close approximation to the theoretical expansion in (1), with the construction of $\mathbb{X}^*$ via (25) mimicking the decomposition of $\mathbb{X}$ in (24). An advantage of the representation in (25) is that it suggests how the raw bootstrap can be readily adapted and modified in order to meet different purposes and allow for different scenarios.

The following adaptation, for example, indicates how we can simulate different realizations of a process whose stochastic structure approximates that of the process giving rise to the original data in $\mathcal{X}$. First recall that the columns of $\mathbf{U}$ are the normalized eigenvectors of $(\mathbb{X} - \mathbf{s}\bar{\mathbb{X}})(\mathbb{X} - \mathbf{s}\bar{\mathbb{X}})'/n$, so $\mathbf{U}'\mathbf{U} = \mathbf{I}_m$ and $\mathbf{U}$ is a point on the Stiefel manifold $\mathbb{V}_{m,n}$, the space of m orthonormal vectors in $\mathbb{R}^n$. Simulated replications of the process are now generated as follow: Simulation Step

S1. As in B1. above;
S2. Generate new realizations $u_{sj}^*, s = 1, \ldots, n, j = 1, \ldots, m$, by taking i.i.d. random draws from a distribution supported on the Stiefel manifold $\mathbb{V}_{m,n}$;
S3. As in B3. above.

The Karhunen–Loève expansion tells us that the random variation observed in $\mathcal{X}$ emanates from fluctuations in the principle component scores, or equivalently, the random coefficients $\upsilon_j, j = 1, 2, \ldots$. These coefficients constitute an uncorrelated sequence of random variables, each with zero mean and unit variance, and the $u_{sj}, j = 1, 2, \ldots, m$, may be viewed as representing a realization of n values of the υ_j. Hence the assignment made in step one, the random sampling to produce $u_{sj}^*, s = 1, \ldots, n, j = 1, 2, \ldots, m$, in the second step, and the reconstruction used in the third step.

In order to illustrate these ideas, Fig. 7 plots the values of $l_k, k = 1, \ldots, m$, evaluated from the SOI data, together with the 2.5%, 50.0% and 97.5% percentile values of l_k^* computed from 25,000 bootstrap replications B1–B3. The fact that the median values of the l_k^* are virtually indistinguishable from the l_k clearly reflects the operation of Lemma 2.

The distribution of $l_k^*, k = 1, \ldots, m$, was also calculated from 25,000 simulated replications S1–S3 wherein step S2 new realizations $u_{sj}^*, s = 1, \ldots, n, j = 1, \ldots, m$, were generated by taking i.i.d. random draws from the Bingham–von Mises–Fisher family of distributions (Hoff, 2009). Of particular interest from our current perspective is the fact that when the concentration parameter is zero the von Mises–Fisher distribution collapses to the uniform distribution and as the concentration parameter increases the distribution can be well approximated by a standard normal distribution. When

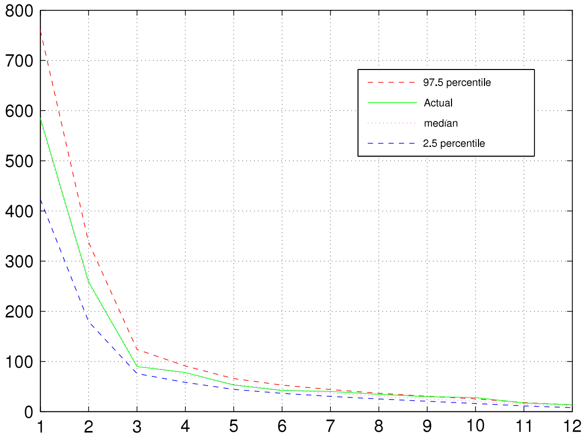


Fig. 7. Median value, and 2.5% and 97.5% percentiles, of l_k^* computed from 25,000 bootstrap replications derived from the SOI data.

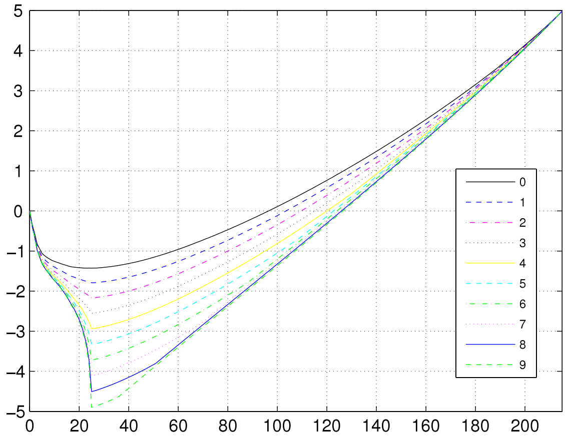


Fig. 8. Plots of $DL_N(k)$ computed from 25,000 bootstrap replications of $\mathbb{X}_g^*$ derived from the OC data.

u_{sj}^*, $s = 1, \ldots, n, j = 1, \ldots, m$, were generated as independent standard normal variables any differences in the bootstrap and simulated distributions were not statistically significant, according to a Kolmogorov–Smirnov test, at any conventional significance level. This result lends incidental support to out previous findings concerning the erratic behavior of the SOI.

Now consider replacing $\mathbb{X}^*$ in (25) by the modified version $\mathbf{s}\bar{\mathbb{X}} + n^{1/2}\mathbf{U}^*\mathbf{L}^*\mathbf{R}'$ where $\mathbf{L}^* = \text{diag}(\sqrt{l_1}, \ldots, \sqrt{l_\kappa}, \sqrt{l_{\kappa+1}^*}, \ldots, \sqrt{l_m^*})$ and $l_j^* = (\delta^*)^2 l_j, j = \kappa + 1, \ldots, m$, δ^* small. This simple modification is designed to mirror the signal-plus-noise structure in (15), when expressed as in (17). Fig. 8 presents plots of the average value of $DL_N(k)$ evaluated from ten modified data sets $\mathbb{X}_g^*$ based upon the OC data. For $g = 1, \ldots, 10$ each $\mathbb{X}_g^*$ was obtained by replacing $\mathbf{L}$ by $\mathbf{L}_g^* = \text{diag}(\sqrt{l_1}, \ldots, \sqrt{l_{25}}, \delta_g^*\sqrt{l_{26}}, \ldots, \delta_g^*\sqrt{l_{216}})$ where $\delta_g^* = (0.8)^{g-1}$. The behavior predicted in Theorem 3 is clearly apparent in the appearance of the sharply defined minimum in $DL_N(k)$ at $k = \kappa = 25$ as δ_g^* decreases.

Acknowledgments

The authors are grateful to an Associate Editor for drawing their attention to recent articles on functional data analysis that have appeared in CSDA and to two referees for helpful and constructive criticism on a preliminary version of this paper.

Appendix A. Proofs

Proof of Lemma 1. First observe that $E[\|\mathbf{d}\|^2] = \sum_{u=1}^T E[d_u^2]$ where, by definition, $d_u = n^{-1}\sum_{s=1}^n \{X_s(t_u) - \mu(t_u)\}$. Now, by (6) and (7) of Assumption 2, $E[d_u^2] = n^{-2}Var[\sum_{s=1}^n \{X_s(t_u) - \mu(t_u)\}]$, which is bounded above by

$$2n^{-2}\sum_{s=1}^{n}\sum_{r=0}^{n-s}\text{Cov}[X_s(t_u)X_{s+r}(t_u)] \leq 2n^{-1}\max_{s=1,\ldots,n}\sum_{r=0}^{n-s}\text{Cov}[X_s(t_u)X_{s+r}(t_u)].$$

It follows directly from (8) of Assumption 2 that $\limsup_{n\to\infty} n\sum_{u=1}^T E[d_u^2] < 2C_1T$.

Similarly, to establish the second part of the lemma, we have $E[\|\mathbf{D}\|^2] = \sum_{u=1}^{T}\sum_{v=1}^{T} E[D_{uv}^2]$ where, by (6) and (7) of Assumption 2,

$$E[D_{uv}^2] = E\left[\left\{\frac{1}{n}\sum_{s=1}^{n}[\{X_s(t_u)-\mu(t_u)\}\{X_s(t_v)-\mu(t_v)\}] - \Gamma(t_u,t_v)\right\}^2\right]$$
$$= \frac{1}{n^2} Var\left[\sum_{s=1}^{n}\{X_s(t_u)-\mu(t_u)\}\{X_s(t_v)-\mu(t_v)\}\right]. \quad (26)$$

Proceeding as previously, the right-hand side of Eq. (26) can be bounded by $2n^{-1}\max_{s=1,\ldots,n}\sum_{r=0}^{n-s}\mathrm{Cov}[\Xi_s(uv)\Xi_{s+r}(uv)]$, and from (9) of Assumption 2 it can be deduced that $\limsup_{n\to\infty} n\sum_{u=1}^{T}\sum_{v=1}^{T}E[D_{uv}^2] < 2C_2T^2$, giving the desired result. □

Proof of Lemma 2. Using the inequality $\sum_{j=1}^{m} l_j\ell_j \geq \mathrm{tr}(\mathbf{G}\boldsymbol{\Gamma}')$ (Anderson and Das-Gupta, 1963) we find that $\sum_{j=1}^{m}(l_j-\ell_j)^2 \leq \|\mathbf{G}-\boldsymbol{\Gamma}\|^2$, and $\|\mathbf{G}-\boldsymbol{\Gamma}\|^2$ is $o_p(T^2/n^{(1-\beta)})$ as $n\to\infty$ by Lemma 1, establishing (10).

From (10) we can readily deduce that $\max_{1\leq j\leq m}(l_j-\ell_j)^2 = o_p(T^2/n^{(1-\beta)})$, which implies that the first term on the right-hand side of the inequality

$$\max_{1\leq j\leq m}\left|\frac{l_j}{T}-\frac{\lambda_j}{\tau}\right| \leq \max_{1\leq j\leq m}\left|\frac{l_j}{T}-\frac{\ell_j}{T}\right| + \max_{1\leq j\leq m}\left|\frac{\ell_j}{T}-\frac{\lambda_j}{\tau}\right| \quad (27)$$

is $o_p(1/n^{(1-\beta)/2})$. Using arguments that parallel those employed in Hall and Hosseini-Nasab (2006, Theorem 1) it can also be shown that the second term on the right-hand side of (27) is $o(1)$. This then establishes (11).

In order to verify (12) of Lemma 2 first observe that $|\sum_{j=1}^{m} l_j/T - \sum_{j=1}^{\infty}\lambda_j/\tau|$ is bounded above by $\sum_{j=1}^{m}|l_j/T - \lambda_j/\tau| + \sum_{j=m+1}^{\infty}\lambda_j/\tau$. Now, from (11) we have $\max_{1\leq j\leq m}|l_j/T-\lambda_j/\tau| = o_p(1/n^{(1-\beta)/2}) + o(1)$, and since $\sum_{j=1}^{\infty}\lambda_j$ is a convergent series $\sum_{j=m+1}^{\infty}\lambda_j \to 0$ as $m\to\infty$. It follows therefore that $\sum_{j=1}^{m}|l_j/T-\lambda_j/\tau| \leq m\max_{1\leq j\leq m}|l_j/T-\lambda_j/\tau| = o_p(m/n^{(1-\beta)/2}) + o(m)$ and the proof of the lemma is complete. □

Proof of Lemma 4. Using (11) and (12) of Lemma 2, we obtain for each k, $1\leq k\leq m$,

$$\widehat{\pi}_k = \frac{(mT)^{-1}\sum_{j=1}^{k} l_j}{(mT)^{-1}\sum_{j=1}^{m} l_j}$$
$$= \frac{(m\tau)^{-1}\sum_{j=1}^{k}\lambda_j + o_p(k/mn^{(1-\beta)/2}) + o(k/m)}{(m\tau)^{-1}\sum_{j=1}^{\infty}\lambda_j + o_p(1/n^{(1-\beta)/2}) + o(1)}$$
$$= \frac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{\infty}\lambda_j} + \left(\frac{\sum_{j=k+1}^{\infty}\lambda_j}{\sum_{j=1}^{\infty}\lambda_j} + 1\right)(o_p(1/n^{(1-\beta)/2}) + o(1)). \quad (28)$$

From (28) we can conclude that $|\widehat{\pi}_k - \pi_k| \leq 2(o_p(1/n^{(1-\beta)/2}) + o(1))$, and hence that $|\widehat{\pi}_k - \pi_k| = o_p(1/n^{(1-\beta)/2}) + o(1)$, as required. □

Proof of Theorem 1. That $\widehat{SNR}(k)$ converges to $SNR(k)$ follows from Lemma 4 by Slutsky's Theorem. Now presume, for a given $\alpha\in[0,1)$, that $\hat{k}_\alpha > k_\alpha$ for all $n > n'$ and $T > T'$. This implies that $\widehat{SNR}(k_\alpha) \leq SNR(\hat{k}_\alpha) + \epsilon$, $\epsilon > 0$, as $(n,T)\to(\infty,\infty)$. Similarly, presuming that $\hat{k}_\alpha < k_\alpha$ for all n and T sufficiently large implies that $\widehat{SNR}(k_\alpha) \geq SNR(\hat{k}_\alpha) - \epsilon$. Thus $|\widehat{SNR}(k_\alpha) - SNR(\hat{k}_\alpha)| \to 0$ as $(n,T)\to(\infty,\infty)$ and hence $\hat{k}_\alpha = k_\alpha$ for n and T sufficiently large. □

Proof of Theorem 2. We know from Lemma 4 that $|\widehat{\pi}_k - \pi_k| = o_p(1/n^{(1-\beta)/2}) + o(1)$. From the expression $\log((1-\widehat{\pi}_k)/(1-\pi_k)) = \log(1 + (\pi_k - \widehat{\pi}_k)/(1-\pi_k))$ and the McLaurin expansion $\log(1+x) = \sum_{r\geq 1}(-)^{r-1}x^r/r$ it now follows that $|\overline{DL}_2(k) - DL_2(k)| = o_p(1/n^{(1-\beta)/2}) + o(1)$, as stated. Similarly,

$$\log\left(\frac{1-\widehat{\pi}_k}{1-\pi_k}\right)\left(\frac{\pi_k}{\widehat{\pi}_k}\right) = \log\left(1 + \frac{\pi_k - \widehat{\pi}_k}{1-\pi_k}\right) - \log\left(1 + \frac{\widehat{\pi}_k - \pi_k}{\pi_k}\right)$$

and using the McLaurin expansion of $\log(1+x)$ once again we can conclude that $|\overline{DL}_N(k) - DL_N(k)| = o_p(1/n^{(1-\beta)/2}) + o(1)$.

Now presume that $\widetilde{k}_a \neq \bar{k}_a$ for $a \in \{2, N\}$. Then we have

$$DL_a(\widetilde{k}_a) - DL_a(\bar{k}_a) = (DL_a(\widetilde{k}_a) - \overline{DL}_a(\widetilde{k}_a)) + (\overline{DL}_a(\widetilde{k}_a) - \overline{DL}_a(\bar{k}_a)). \tag{29}$$

By definition of $\widetilde{k}_a$ and $\bar{k}_a$ as the minimizing values of $DL_a(k)$ and $\overline{DL}_a(k)$, respectively, the limit-supremum of the left-hand side of (29) is zero and, given that the first term on the right-hand side converges to zero, the limit-infimum of the right-hand side is positive. Thus we have the desired result *reductio ad absurdum*. □

Proof of Theorem 3. First consider $\overline{DL}_N(k)$. Straightforward if somewhat tedious manipulations indicate that we can expand $\overline{DL}_N(k) - \overline{DL}_N(k+1)$ and express it as the product of $(n+T)/nT$ times

$$\frac{nT}{(n+T)} \log\left(1 + \frac{\lambda_{k+1}}{\sum_{j=k+2}^{\infty} \lambda_j}\right) - \log\left(\frac{nT - (k+1)(n+T)}{n+T}\right) \tag{30}$$

$$- \log\left(1 + \frac{\lambda_{k+1}}{\sum_{j=1}^{k} \lambda_j}\right) + (k+1)\log\left(\frac{nT - k(n+T)}{nT - (k+1)(n+T)}\right) \tag{31}$$

$$- \log\left(1 + \frac{\sum_{j=1}^{k+1} \lambda_j}{\sum_{j=k+2}^{\infty} \lambda_j}\right) + k\log\left(\frac{k+1}{k}\right) + \log(k). \tag{32}$$

When $k < \kappa$ we find from (18) that $\log(1 + \lambda_{k+1}/\sum_{j=k+2}^{\infty} \lambda_j)$ equals

$$\begin{cases} \log\left(1 + \dfrac{\eta_{k+1} + \delta^2 \sum_{i=1}^{\infty} \theta_i \beta_{i(k+1)}^2}{\sum_{j=k+2}^{\kappa} \eta_j + \delta^2 \sum_{j=k+2}^{\infty} \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2} + O(\delta^2)\right), & \text{when } k \leq \kappa - 2; \\[2ex] \log\left(1 + \dfrac{\eta_\kappa + \delta^2 \sum_{i=1}^{\infty} \theta_i \beta_{i\kappa}^2}{\delta^2 \sum_{j=\kappa+1}^{\infty} \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2} + O(\delta^2)\right), & \text{when } k = \kappa - 1. \end{cases}$$

Both of these expressions are positive as $\delta \to 0$, implying that the first term in (30) will dominate all others in the expansion of $\overline{DL}_N(k) - \overline{DL}_N(k+1)$ and thus that $\overline{DL}_N(k) - \overline{DL}_N(k+1)$ will be positive as $(n, T) \to (\infty, \infty)$.

Now suppose that $k \geq \kappa$ and that Assumption 3 obtains. Expression (18) implies that

$$\log\left(1 + \frac{\lambda_{k+1}}{\sum_{j=k+2}^{\infty} \lambda_j}\right) = \log\left(1 + \frac{\sum_{i=1}^{\infty} \theta_i \beta_{i(k+1)}^2}{\sum_{j=k+2}^{\infty} \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2} + O(\delta^2)\right).$$

By Assumption 3 $\sum_{j=k+2}^{\infty} \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2 > \delta^{d-2}(C \sum_{i=1}^{\infty} \theta_i - \delta^{2-d} \sum_{j=1}^{k+1} \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2) > \delta^{d-2}(C - \delta^{2-d}) \sum_{i=1}^{\infty} \theta_i = \delta^{d-2} C'$, say. We can therefore conclude that the limit-supremum of the two terms in (30) will not exceed a figure that is of magnitude $(nT/(n+T))\, O(\delta^{2-d}) - \log(nT/(n+T))$. We also find that

$$\log\left(1 + \frac{\lambda_{k+1}}{\sum_{j=1}^{k} \lambda_j}\right) = \log\left(1 + \frac{\delta^2 \sum_{i=1}^{\infty} \theta_i \beta_{i(k+1)}^2}{\sum_{j=1}^{\kappa} \eta_j + \delta^2 \sum_{j=1}^{k} \sum_{i=1}^{\infty} \theta_i \beta_{ij}^2} + O(\delta^2)\right),$$

and the two terms in (31) will be of order $-O(\delta^2) + (k+1)o(1)$, and that

$$\log\left(1+\frac{\sum_{j=1}^{k+1}\lambda_j}{\sum_{j=k+2}^{\infty}\lambda_j}\right) = \log\left(1+\frac{\sum_{j=1}^{\kappa}\eta_j+\delta^2\sum_{j=1}^{k+1}\sum_{i=1}^{\infty}\theta_i\beta_{ij}^2}{\delta^2\sum_{j=k+2}^{\infty}\sum_{i=1}^{\infty}\theta_i\beta_{ij}^2}+O(\delta^2)\right),$$

and the two terms in (32) will be of order $-O(\delta^{-2}) + O(k)$. Adding the six terms in (30)–(32) together now leads to the conclusion that when $k \geq \kappa$, $\overline{DL}_N(k) - \overline{DL}_N(k+1)$ will be negative as $\delta \to 0$, provided that $\delta^{2-d}nT/(n+T) \to 0$ as $(n, T) \to (\infty, \infty)$.

Parallel but less complicated arguments also show that: (i) $\overline{DL}_2(k)$ is monotonically decreasing in k for $k < \kappa$, and (ii) $\overline{DL}_2(k)$ is monotonically increasing in k for $k \geq \kappa$ when Assumption 3 holds and $\delta^{2-d}nT/(n+T) \to 0$.

Thus, for both $a \in \{2, N\}$ we can conclude that $\bar{k}_a \geq \kappa$, and that $\bar{k}_a = \kappa$ when Assumption 3 holds and $\delta^{2-d}nT/(n+T) \to 0$. The properties stated in the theorem now follow directly from Theorem 2. □

Proof of Lemma 3. Set $\bar{\Gamma}(t, s) = (\tau/T)\sum_{j=1}^{m} l_j r_j(t) r_j(s)$. Using Lemma 4.3 of Bosq (2000) we can deduce, as in the proof of their Theorem 1 by Hall and Hosseini-Nasab (2006), that $\|r_j - \rho_j\|^2 \leq 3\Delta/d^2$ where $d = \min_{1\leq j\leq m}(\lambda_j - \lambda_{j+1})$ and $\Delta = \int_0^\tau\int_0^\tau |\bar{\Gamma}(t, s) - \Gamma(t, s)|^2 \mathrm{d}t\mathrm{d}s$. Replacing the integrals by their approximating sums we have $\Delta \leq (\tau^2/T^2)\sum_{u=1}^{T}\sum_{v=1}^{T} |\bar{\Gamma}(t_u, t_v) - \Gamma(t_u, t_v)|^2 + o(1) = (\tau^2/T^2)\|\mathbf{G} - \boldsymbol{\Gamma}\|^2 + o(1)$. By Lemma 1 $\|\mathbf{G} - \boldsymbol{\Gamma}\|^2 = o_p(T^2/n^{1-\beta})$, and the proof is thereby completed. □

Appendix B. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.03.018.

References

Anderson, T.W., Das-Gupta, S., 1963. Some inequalities on characteristic roots of matrices. Biometrika 50, 522–524.
Banks, D., 2003. Proteomics: a frontier between genomics and metabolomics. Chance 16, 6–7.
Besse, P.C., Cardot, H., Stephenson, D.B., 2000. Autoregressive forecasting of some functional climatic variations. Scandinavian Journal of Statistics 27, 673–687.
Boente, G., Fraiman, R., 2000. Kernel-based functional principal components. Statistics and Probability Letters 48, 335–345.
Bosq, D., 2000. Linear Processes in Function Spaces: Theory and Applications. In: Lecture Notes in Statistics, Springer, New York.
Chiou, J.M., Li, P.L., 2007. Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society B69, 679–699.
Diggle, P., Verbyla, A., 1998. Nonparametric estimation of covariance structure in longitudinal data. Biometrics 54, 401–415.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.
Ferraty, F., Vieu, P., 2003. Curve discrimination: a nonparametric functional approach. Computational Statistics and Data Analysis 44, 161–173.
Ferraty, F., Vieu, P., 2009. Additive prediction and boosting for functional data. Computational Statistics and Data Analysis 53, 1400–1413.
Grünwald, P.D., 2007. The Minimum Description Length Principle. The MIT Press, Cambridge.
Hall, P., Hosseini-Nasab, M., 2006. On properties of functional principal components analysis. Journal of the Royal Statistical Society B68, 109–126.
Hall, P., Muller, H.G., Wang, J.L., 2006. Properties of principal component methods for functional and longitudinal data analysis. The Annals of Statistics 34, 1493–1517.
Hall, P., Poskitt, D.S., Presnell, B., 2001. A functional data-analytic approach to signal discrimination. Technometrics 43, 1–9.
Hall, P., Vial, C., 2006. Assessing the finite dimensionality of functional data. Journal of the Royal Statistical Society B68, 689–705.
Hansen, M., Yu, B., 2001. Model selection and the principle of minimum description length. Journal of the American Statistical Association 96, 746–774.
Hoff, P., 2009. Simulation of the matrix Bingham–von Mises–Fisher distribution, with applications to multivariate and relational data. Journal of Computational and Graphical Statistics 18, 438–456.
Jolliffe, I.T., 2002. Principal Component Analysis. Springer, New York.
Li, B., Yu, Q., 2008. Classification of functional data: a segmentation approach. Computational Statistics and Data Analysis 52, 4790–4800.
Ramsay, J.O., Ramsey, J.B., 2002. Functional data analysis of the dynamics of the monthly index of nondurable goods production. Journal of Econometrics 107, 327–344.
Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis Methods and Case Studies. Springer, New York.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, 2nd edition. Springer, New York.
Rissanen, J., 2007. Information and Complexity in Statistical Modeling. Springer, New York.
Rissanen, J., 2000. MDL denoising. IEEE Transactions on Information Theory IT-46, 2537–2543.
Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.
Shibata, R., 1981. An optimal selection of regression variables. Biometrika 68, 45–54.
Stone, R., Smith, I., McIntosh, P., 2000. Statistical Methods for Deriving Seasonal Climate Forecasts From GCM's. In: Applications of Seasonal Climate Forecasting in Agricultural and Natural Ecosystems, Kluwer Acedemic, pp. 135–147 (Chapter 10).
Thiele, H., 2003. Proteomics: mass spectrometry and bioinformatics in proteomics. Chance 16, 29–36.
Yao, F., Muller, H.G., Wang, J.L., 2005. Functional data analysis for sparse longitudinal data. Journal of American Statistical Association 100, 577–590.