J. R. Statist. Soc. B (2006)
68, *Part* 1, *pp.* 109–126

On properties of functional principal components analysis

Peter Hall and Mohammad Hosseini-Nasab

Australian National University, Canberra, Australia

[Received August 2004. Final revision September 2005]

Summary. Functional data analysis is intrinsically infinite dimensional; functional principal component analysis reduces dimension to a finite level, and points to the most significant components of the data. However, although this technique is often discussed, its properties are not as well understood as they might be. We show how the properties of functional principal component analysis can be elucidated through stochastic expansions and related results. Our approach quantifies the errors that arise through statistical approximation, in successive terms of orders $n^{-1/2}, n^{-1}, n^{-3/2}, \ldots$, where n denotes sample size. The expansions show how spacings among eigenvalues impact on statistical performance. The term of size $n^{-1/2}$ illustrates first-order properties and leads directly to limit theory which describes the dominant effect of spacings. Thus, for example, spacings are seen to have an immediate, first-order effect on properties of eigenfunction estimators, but only a second-order effect on eigenvalue estimators. Our results can be used to explore properties of existing methods, and also to suggest new techniques. In particular, we suggest bootstrap methods for constructing simultaneous confidence regions for an infinite number of eigenvalues, and also for individual eigenvalues and eigenvectors.

Keywords: Confidence interval; Cross-validation; Eigenfunction; Eigenvalue; Linear regression; Operator theory; Principal component analysis; Simultaneous confidence region

1. Introduction

Principal component analysis (PCA) is widely used in the study of functional data, since it allows finite dimensional analysis of a problem that is intrinsically infinite dimensional. See, for example, chapter 6 of Ramsay and Silverman (1997), and several of the examples that were treated by Ramsay and Silverman (2002). In traditional PCA the effects of truncating to a finite number of dimensions are often explored in terms of a finite dimensional parametric model, e.g. when the data are Gaussian. However, this approach is often not feasible in the case of functional data analysis, and as a result the justification for methodology there tends to be more *ad hoc*.

In this paper we develop theory that is based on stochastic expansions of eigenvalue and eigenvector estimators, providing not only a new understanding of the effects of truncating to a finite number of principal components but also pointing to new methodology. Our results, which are summarized in Section 2.2, lead directly to first-order properties of eigenvalue and eigenfunction estimators, and also to higher order theory if the expansions are taken to greater numbers of terms. In Section 2.2 we give expansions which explicitly include terms of sizes $n^{-1/2}$ and n^{-1}, where n denotes sample size, and a remainder of order $n^{-3/2}$. This work shows that eigenvalue spacings have only a second-order effect on properties of eigenvalue estimators, but

Address for correspondence: Peter Hall, Centre for Mathematics and Its Applications, Australian National University, Canberra, ACT 0200, Australia.
E-mail: halpstat@maths.anu.edu.au

 1369–7412/06/68109

a first-order effect on properties of eigenfunction estimators. Our expansions are immediately valid for any finite number of principal components, but they are also available uniformly in increasingly many components; the issue of uniformity is addressed in Appendix A.1.

The problem of determining estimator accuracy, uniformly over many components, prompts consideration of explicit uniform bounds that are obtainable via the mathematical theory of infinite dimensional operators. In Section 3 we show how stochastic bootstrapped versions of these ideas can be used to construct simultaneous confidence regions for literally all eigenvalue estimates, and for increasing numbers of eigenfunction estimates. The nature of these regions means that their accuracy of coverage errs on the side of conservatism. In the case of simultaneous confidence regions for eigenvalues the degree of conservatism is small, and the bootstrap confidence regions are attractive practical tools.

We also address the problem of bootstrap confidence regions for individual, or relatively small numbers of, eigenvalues and eigenvectors. The theory that is developed in Section 2.2 makes it possible to justify bootstrap methods of that type, in terms of asymptotic theory; see Appendix A.3. Sections 5.2 and 5.3 summarize numerical properties of such methods.

The theory in Section 2.2 also provides new insight into more conventional functional data analysis methods, including those which are used for linear regression. Indeed, the effect of eigenvalue spacings on properties of a wide range of techniques for functional data analysis is made quite transparent by the expansions in Section 2.2, and in Section 4 those results are used to explore the validity of simple accounts of the performance of functional linear regression. It is observed that those accounts are valid if eigenvalues are reasonably well separated, but not otherwise. The use of cross-validation as a tool for 'tuning' functional linear regression is suggested by this discussion and can be justified theoretically by using the expansion approach that is developed in Section 2. See Appendix A.3, and the numerical work in Section 5.4.

Early work on PCA for functional data includes that of Besse and Ramsay (1986), Ramsay and Dalzell (1991), Rice and Silverman (1991) and Silverman (1995, 1996). More recent research includes contributions to techniques for functional PCA (see for example Brumback and Rice (1998), Cardot (2000), Cardot *et al.* (2000), Girard (2000), James *et al.* (2000), Boente and Fraiman (2000) and He *et al.* (2003)), functional PCA in different metrics (e.g. Ocana *et al.* (1999)) and applications of PCA to fields as diverse as longitudinal data analysis (e.g. Yao *et al.* (2005)), time series analysis for functional data (e.g. Aguilera *et al.* (1999a, b)) and exploratory methods (e.g. Kneip and Utikal (2001)). Work of Dauxois *et al.* (1982), Bosq (1989), Besse (1992), Huang *et al.* (2002) and Mas (2002), for example, addresses the empirical basis function approximation and approximations of covariance operators, and so is directly related to our own.

2. Definition and properties of principal component expansions

2.1. Definition of expansions

In this section we summarize several known properties of principal component expansions. Let X denote a random function, or equivalently a stochastic process, that is defined in the interval $\mathcal{I} = [0, 1]$ and satisfying $\int_{\mathcal{I}} E(X^2) < \infty$. Put $\eta = E(X)$, a conventional function. The principal component expansion of $X - \eta$ may be constructed via the covariance function

$$K(u, v) = E[\{X(u) - \eta(u)\}\{X(v) - \eta(v)\}], \tag{2.1}$$

which we assume to be square integrable and interpret as the kernel of a mapping, or operator, on the space $L_2(\mathcal{I})$ of square integrable functions from $\mathcal{I}$ to the real line. To reduce the notational load we shall denote the operator by K, also; it takes $\psi \in L_2(\mathcal{I})$ to $K\psi$, where

$$(K\psi)(u) = \int_{\mathcal{I}} K(u,v)\,\psi(v)\,\mathrm{d}v.$$

Then (see for example Indritz (1963), chapter 4) we may write

$$K(u,v) = \sum_{j=1}^{\infty} \theta_j\,\psi_j(u)\,\psi_j(v), \tag{2.2}$$

where $\theta_1 \geqslant \theta_2 \geqslant \ldots \geqslant 0$ is an enumeration of the eigenvalues of K, and the corresponding orthonormal eigenfunctions are $\psi_1, \psi_2, \ldots$.

The Karhunen–Loève expansion of $X - \eta$ is given by

$$X(u) - \eta(u) = \sum_{j=1}^{\infty} \xi_j\,\psi_j(u), \tag{2.3}$$

where the random variables $\xi_1, \xi_2, \ldots$ are given by $\xi_j = \int_{\mathcal{I}} (X-\eta)\psi_j$. It follows that they are uncorrelated and have zero means, and that $\theta_j = E(\xi_j^2)$ and

$$\sum_{j \geqslant 1} \theta_j = \int_{\mathcal{I}} E(X-\eta)^2 < \infty.$$

Next we describe the empirical expansion. Suppose that we are given a set $\mathcal{X} = \{X_1, \ldots, X_n\}$ of independent random functions, all distributed as X. The standard empirical approximation to $K(u,v)$ is

$$\hat{K}(u,v) = \frac{1}{n} \sum_{i=1}^{n} \{X_i(u) - \bar{X}(u)\}\{X_i(v) - \bar{X}(v)\},$$

where $\bar{X} = n^{-1}\Sigma_i X_i$. Analogously to equation (2.2) we may write

$$\hat{K}(u,v) = \sum_{j=1}^{\infty} \hat{\theta}_j\,\hat{\psi}_j(u)\,\hat{\psi}_j(v) \tag{2.4}$$

and define a mapping $\hat{K}$ from $L_2(\mathcal{I})$ to itself by

$$(\hat{K}\psi)(u) = \int_{\mathcal{I}} \hat{K}(u,v)\,\psi(v)\,\mathrm{d}v.$$

In equation (2.4) the random variables $\hat{\theta}_1 \geqslant \hat{\theta}_2 \geqslant \ldots \geqslant 0$ are eigenvalues of the operator $\hat{K}$, and $\hat{\psi}_1, \hat{\psi}_2, \ldots$ is the corresponding sequence of eigenvectors. To overcome problems arising from the fact that ψ_j and $\hat{\psi}_j$ are defined only up to a change in sign, and to ensure that $\hat{\psi}_j$ may be viewed as an estimator of ψ_j rather than of $-\psi_j$, we shall tacitly assume, below, that the sign of $\hat{\psi}_j$ is chosen so that $\int_{\mathcal{I}} \psi_j \hat{\psi}_j \geqslant 0$.

The functions $\hat{\psi}_1, \hat{\psi}_2, \ldots$ form a complete orthonormal basis of $L_2(\mathcal{I})$. Thus, given a function $b \in L_2(\mathcal{I})$ we may write

$$\begin{aligned} X_i &= \sum_{j=1}^{\infty} \xi_{ij}\hat{\psi}_j, \\ b &= \sum_{j=1}^{\infty} b_j\hat{\psi}_j, \end{aligned} \tag{2.5}$$

where $\xi_{i1}, \xi_{i2}, \ldots$ and $b_1, b_2, \ldots$ are generalized Fourier coefficients, each of them being a random function of the data $\mathcal{X}$, and hence a random variable. In most statistical studies, e.g. many of those which were cited in Section 1, $\hat{\psi}_j$ is treated as an approximation to ψ_j.

2.2. Properties of expansions

This section summarizes statistical properties of the estimators $\hat{\theta}_j$ and $\hat{\psi}_j$. First we use stochastic expansions to evaluate the accuracy of $\hat{\psi}_j$ as an approximation to ψ_j. Assume that the eigenvalues θ_j are all distinct. The case where there is only a finite number of ties among the θ_js can be treated without much difficulty, but other settings are more awkward. Distinctness of eigenvalues implies that the operator K is strictly positive definite, i.e. each $\theta_j > 0$. We may write

$$\hat{\psi}_j = \sum_{k \geqslant 1} a_{jk} \psi_k,$$

where the generalized Fourier coefficients a_{jk} are functionals of the data $\mathcal{X}$. Then, for each $j \neq k$,

$$a_{jj} = 1 - \tfrac{1}{2} n^{-1} \sum_{l:l \neq j} (\theta_j - \theta_l)^{-2} \biggl(\int Z \psi_j \psi_l \biggr)^2 + O_p(n^{-3/2}), \tag{2.6}$$

$$\begin{aligned} a_{jk} = {} & n^{-1/2} (\theta_j - \theta_k)^{-1} \int Z \psi_j \psi_k + n^{-1} \biggl\{ (\theta_j - \theta_k)^{-1} \sum_{l:l \neq j} (\theta_j - \theta_l)^{-1} \biggl(\int Z \psi_j \psi_l \biggr) \biggl(\int Z \psi_k \psi_l \biggr) \\ & - (\theta_j - \theta_k)^{-2} \biggl(\int Z \psi_j \psi_j \biggr) \biggl(\int Z \psi_j \psi_k \biggr) \biggr\} + O_p(n^{-3/2}), \end{aligned} \tag{2.7}$$

where $Z = n^{1/2}(\hat{K} - K)$, $\int Z \psi_r \psi_s$ denotes $\int \int_{\mathcal{I}^2} Z(u, v)\, \psi_r(u)\, \psi_s(v)\, \mathrm{d}u\, \mathrm{d}v$ and, here and in equations (2.8)–(2.10) below, the infinite series converge for each fixed j. (In Appendix A.1 we shall discuss conditions under which equations (2.6)–(2.11) hold, and in particular under which the infinite series there converge.) Of course, there are analogues of equations (2.6) and (2.7) with remainders $O_p(n^{-r/2})$ for any positive integer r. We have taken $r = 3$ only for brevity and simplicity.

The expected values of the terms in $n^{-1/2}$ and $n^{-3/2}$, on the right-hand sides of equations (2.6) and (2.7), may be shown to be of orders n^{-1} and n^{-2} respectively. It then follows from equations (2.6) and (2.7) that, for example, $a_{jk} = \delta_{jk} + O_p(n^{-1/2})$ and $E(a_{jk}) = \delta_{jk} + O(n^{-1})$, where δ_{jk} is the Kronecker delta.

Results (2.6) and (2.7) point to the following expansion:

$$\begin{aligned} \hat{\psi}_j(t) - \psi_j(t) = {} & n^{-1/2} \sum_{k:k \neq j} (\theta_j - \theta_k)^{-1} \psi_k(t) \int Z \psi_j \psi_k - \tfrac{1}{2} n^{-1} \psi_j(t) \sum_{k:k \neq j} (\theta_j - \theta_k)^{-2} \biggl(\int Z \psi_j \psi_k \biggr)^2 \\ & + n^{-1} \sum_{k:k \neq j} \psi_k(t) \biggl\{ (\theta_j - \theta_k)^{-1} \sum_{l:l \neq j} (\theta_j - \theta_l)^{-1} \biggl(\int Z \psi_j \psi_l \biggr) \biggl(\int Z \psi_k \psi_l \biggr) \\ & - (\theta_j - \theta_k)^{-2} \biggl(\int Z \psi_j \psi_j \biggr) \biggl(\int Z \psi_j \psi_k \biggr) \biggr\} + O_p(n^{-3/2}). \end{aligned} \tag{2.8}$$

Analogously to equations (2.6) and (2.7), it may be shown that

$$\hat{\theta}_j - \theta_j = n^{-1/2} \int Z \psi_j \psi_j + n^{-1} \sum_{k:k \neq j} (\theta_j - \theta_k)^{-1} \biggl(\int Z \psi_j \psi_k \biggr)^2 + O_p(n^{-3/2}). \tag{2.9}$$

Similarly, the following shorter expansions may be derived:

$$\begin{aligned} n \| \hat{\psi}_j - \psi_j \|^2 & = n \sum_{k=1}^{\infty} (a_{jk} - \delta_{jk})^2 \\ & = \sum_{k:k \neq j} (\theta_j - \theta_k)^{-2} \biggl(\int Z \psi_j \psi_k \biggr)^2 + o_p(1), \end{aligned} \tag{2.10}$$

$$n^{1/2}(\hat\theta_j-\theta_j)=\int Z\psi_j\psi_j+o_p(1). \tag{2.11}$$

Results (2.10) and (2.11) lead directly to limit theorems for $\hat\psi_j$ and $\hat\theta_j$, as follows. If part (a) of condition (A.1) in Appendix A.1 holds, then the weak limit of Z is a Gaussian process, ζ say. The covariance function of ζ is

$$\begin{aligned}\operatorname{cov}\{\zeta(u_1,v_1)\,\zeta(u_2,v_2)\}&=\sum_{j_1=1}^{\infty}\ldots\sum_{j_4=1}^{\infty}E(\xi_{j_1}\ldots\xi_{j_4})\,\psi_{j_1}(u_1)\,\psi_{j_2}(v_1)\,\psi_{j_3}(u_2)\,\psi_{j_4}(v_2)\\&\quad-\sum_{j_1=1}^{\infty}\sum_{j_2=1}^{\infty}\theta_{j_1}\theta_{j_2}\,\psi_{j_1}(u_1)\,\psi_{j_1}(v_1)\,\psi_{j_2}(u_2)\,\psi_{j_2}(v_2)\\&=\sum_{j_1\ne j_2}\sum\theta_{j_1}\theta_{j_2}\{\psi_{j_1}(u_1)\,\psi_{j_2}(v_1)\,\psi_{j_1}(u_2)\,\psi_{j_2}(v_2)+\psi_{j_1}(u_1)\,\psi_{j_2}(v_1)\,\psi_{j_2}(u_2)\,\psi_{j_1}(v_2)\}\\&\quad+2\sum_{j=1}^{\infty}\theta_j^2\,\psi_j(u_1)\,\psi_j(v_1)\,\psi_j(u_2)\,\psi_j(v_2)\,,\end{aligned}$$

where the first identity is true generally and the second holds for processes X, such as Gaussian processes, where the variables ξ_j are independent, rather than merely uncorrelated, and have zero kurtosis.

Accounts of asymptotic normality of eigenvalues, eigenvectors and their projections have been given by Dauxois *et al.* (1982) and Bosq (2000). In connection with the results that were discussed above, it can be seen from equation (2.10) that $n^{1/2}\|\hat\psi_j-\psi_j\|\to U_j$ in distribution, where

$$U_j^2=\sum_{k:k\ne j}(\theta_j-\theta_k)^{-2}N_{jk}^2 \tag{2.12}$$

and the random variables $N_{jk}=\int\zeta\psi_j\psi_k$ are jointly normally distributed with zero mean. If the random function X is a Gaussian process then $N_{j1},N_{j2},\ldots$ are independent as well as normally distributed. Note that, since $\int_{\mathcal I}E(X^4)<\infty$,

$$\begin{aligned}\sum_{k=1}^{\infty}E(N_{jk}^2)&=\int_{\mathcal I}E\biggl\{\int_{\mathcal I}\zeta(u,v)\psi_j(v)\,\mathrm dv\biggr\}^2\,\mathrm du\\&\leqslant\iint_{\mathcal I^2}E(\zeta^2)\\&\leqslant E\biggl(\int_{\mathcal I}X^2\biggr)^2<\infty,\end{aligned}$$

from which it follows that the series defining U_j^2 is finite provided that the eigenvalue θ_j is not repeated.

Our next result gives explicit bounds on $\|\hat\psi_j-\psi_j\|$ in terms of spacings, and a spacings-free bound for $|\hat\theta_j-\theta_j|$. Define $\hat\Delta=(\int|\hat K-K|^2)^{1/2}$,

$$\begin{aligned}\delta_j&=\min_{1\leqslant k\leqslant j}(\theta_k-\theta_{k+1}),\\J&=\inf\{j\geqslant1:\theta_j-\theta_{j+1}\leqslant2\hat\Delta\},\\\hat\delta_j&=\min_{1\leqslant k\leqslant j}(\hat\theta_k-\hat\theta_{k+1}),\\\hat J&=\inf\{j\geqslant1:\hat\theta_j-\hat\theta_{j+1}\leqslant2\hat\Delta\}.\end{aligned} \tag{2.13}$$

The only assumptions that are needed for the theorem below are that $X_1, \ldots, X_n$ are square integrable random functions and K is a covariance such as that at equation (2.1). The theorem follows from results of Dauxois *et al.* (1982), Bhatia *et al.* (1983), Bosq (1991) and Bosq (2000), lemma 4.3.

Theorem 1.

(a) With probability 1, $\sup_{j \geqslant 1} |\hat{\theta}_j - \theta_j| \leqslant \hat{\Delta}$ and, for all $1 \leqslant j \leqslant J-1$,

$$\|\hat{\psi}_j - \psi_j\| \leqslant 2^{1/2}[1 - \{1 - 4(\hat{\Delta}/\delta_j)^2\}^{1/2}]^{1/2} \leqslant 8^{1/2}\hat{\Delta}/\delta_j.$$

(b) This result continues to hold if (J, δ_j) is replaced by $(\hat{J}, \hat{\delta}_j)$ throughout.

In Section 3 we shall use part (b) of the theorem to develop confidence statements for θ and ψ_j.

3. Quantifying the accuracy of $\hat{\theta}_j$ and $\hat{\psi}_j$

Here we suggest bootstrap methods, which are justifiable by using the expansion approach that was developed in Section 2 (see also Appendix A.3), for quantifying the accuracy of $\hat{\theta}_j$ and $\hat{\psi}_j$ as approximations to θ_j and ψ_j respectively. Draw a resample, $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$, by sampling randomly, with replacement, from the sample $\mathcal{X}$ of random functions. For this resample, compute the analogues $\hat{\theta}_j^*$ and $\hat{\psi}_j^*$ of $\hat{\theta}_j$ and $\hat{\psi}_j$. Approximate the unconditional distribution of $\hat{\theta}_j - \theta_j$ by the distribution of $\hat{\theta}_j^* - \hat{\theta}_j$ conditional on $\mathcal{X}$, and approximate the unconditional distribution of the random function $\hat{\psi}_j - \psi_j$ by the conditional distribution of $\hat{\psi}_j^* - \hat{\psi}_j$. In this way, develop confidence statements about the sizes of

(a) $\hat{\theta}_j - \theta_j$,
(b) $\sup_t |\hat{\psi}_j(t) - \psi_j(t)|$ or
(c) $\|\hat{\psi}_j - \psi_j\|$.

Using (a) we construct percentile bootstrap confidence intervals for θ_j, or for a collection of θ_js if we address several eigenvalues simultaneously; using (b) we obtain simultaneous bootstrap confidence bands for ψ_j; using (c) we obtain confidence intervals for the L_2-distance of $\hat{\psi}_j$ from ψ_j.

For example, if the critical point $\hat{z}_\alpha$ is defined by

$$P\{\sup_{t \in \mathcal{I}} |\hat{\psi}_j^*(t) - \hat{\psi}_j(t)| \leqslant \hat{z}_\alpha | \mathcal{X}\} = 1 - \alpha,$$

then the bivariate region that is given by $\{(t, u): t \in \mathcal{I} \text{ and } |\hat{\psi}(t) - u| \leqslant \hat{z}_\alpha\}$ is a nominal $(1-\alpha)$-level confidence region for graphs of ψ_j. Bootstrap versions of the asymptotic theory that was discussed in Section 2 can be used to show that these two-sided, percentile method confidence statements (for either θ_j or ψ_j) have coverage error equal to $O(n^{-1})$, and that this reduces to $O(n^{-2})$ after double-bootstrap calibration. See Appendix A.3 for discussion of theory for the bootstrap. Section 5.2 will summarize numerical properties of such regions.

However, this approach is problematic if we want the approximation to be valid for a large number of values of j. In theory, we would wish that number to diverge as n increases. Theorem 1 suggests an approach which we can use with a degree of conservatism in such cases, as follows. Suppose that we have a one-sided prediction interval for $\hat{\Delta}$, of the form $P(\hat{\Delta} \leqslant \hat{\Delta}_{\mathrm{upp}}) = 1 - \alpha_n$, say, where $\hat{\Delta}_{\mathrm{upp}}$ is computable from data and the subscript denotes 'upper bound'. Define

$$\hat{J}_{\mathrm{upp}} = \inf\{j \geqslant 1 : \hat{\theta}_j - \hat{\theta}_{j+1} \leqslant 2\hat{\Delta}_{\mathrm{upp}}\}.$$

Then, in view of theorem 1, the following is true:

$$\begin{aligned}&\text{with probability at least } 1-\alpha_n,\ \sup_{j\geqslant 1}|\hat\theta_j-\theta_j|\leqslant\hat\Delta_{\mathrm{upp}},\ \text{and, for all}\\ &1\leqslant j\leqslant\hat J_{\mathrm{upp}}-1,\ \|\hat\psi_j-\psi_j\|\leqslant 2^{1/2}[1-\{1-4(\hat\Delta_{\mathrm{upp}}/\hat\delta_j)^2\}^{1/2}]^{1/2}.\end{aligned}\tag{3.1}$$

We may readily compute $\hat\Delta_{\mathrm{upp}}$ by using bootstrap methods, as follows. Let $\hat K^*$ denote the bootstrap version of $\hat K$, computed from $\mathcal{X}^*$ rather than $\mathcal{X}$, and put $\hat\Delta^*=\|\hat K^*-\hat K\|$. Given $0<\alpha<1$, e.g. $\alpha=0.05$, take $\hat\Delta_{\mathrm{upp}}$ to be the upper α-level critical point of the distribution of $\hat\Delta^*$, given $\mathcal{X}$.

Note that $P(\hat\Delta\leqslant\hat\Delta_{\mathrm{upp}})$ converges to $1-\alpha$ as $n\to\infty$. To appreciate why, recall from Section 2 that $n^{1/2}(\hat K-K)$ converges weakly to a Gaussian process ζ. Analogously, and conditional on $\mathcal{X}$, $n^{1/2}(\hat K^*-\hat K)$ converges weakly to the same ζ. Therefore, the limit of the distribution of $n^{1/2}\Delta^*$, conditional on $\mathcal{X}$, is identical to the limit of the unconditional distribution of $n^{1/2}\hat\Delta$. It follows that $n^{1/2}\hat\Delta_{\mathrm{upp}}$ converges in probability to the upper α-level critical point of the distribution of $\int\int_{\mathcal{I}^2}\zeta^2$, and hence that $P(\hat\Delta\leqslant\hat\Delta_{\mathrm{upp}})$ converges to $1-\alpha$.

The simultaneous bootstrap confidence interval for θ_j, which is suggested by result (3.1), is indeed conservative but not especially so. Its numerical properties will be discussed in Section 5.3. The confidence band for ψ_j tends to be quite conservative, however.

4. Properties of linear regression estimators

The functional simple linear regression model is

$$Y_i=a+\int_{\mathcal{I}}bX_i+\varepsilon_i,\qquad 1\leqslant i\leqslant n,\tag{4.1}$$

where b and X_i are square integrable functions from $\mathcal{I}$ to the real line, a, Y_i and ε_i are scalars, a and b are deterministic, the pairs $(X_1,\varepsilon_1),(X_2,\varepsilon_2),\ldots$ are independent and identically distributed, the random functions X_i are independent of the errors ε_i, $\sigma^2=E(\varepsilon^2)<\infty$, $E(\varepsilon)=0$ and $\int_{\mathcal{I}}E(X^2)<\infty$, where ε and X are distributed as ε_i and X_i respectively. If X_i and b are expressed in terms of the orthonormal basis $\hat\psi_1,\hat\psi_2,\ldots$ as at equation (2.5), then model (4.1) may be written equivalently as

$$Y_i=a+\sum_{j\geqslant 1}b_j\xi_{ij};$$

this motivates expression (4.2) below.

The true value, (a^0,b^0) say, of (a,b) may be estimated by minimizing

$$\sum_{i=1}^n\left(Y_i-a-\sum_{j=1}^r b_j\xi_{ij}\right)^2\tag{4.2}$$

with respect to $a,b_1,\ldots,b_r$, and taking $b_j=0$ for $j\geqslant r+1$. See, for example, Ramsay and Silverman (1997), chapter 10. This gives

$$\begin{aligned}\hat a&=\bar Y-\sum_{j=1}^r\hat b_j\bar\xi_j,\\ \hat b_{(r)}&=(\hat b_1,\ldots,\hat b_r)^{\mathrm{T}}=\hat\Sigma_{(r)}^{-1}\hat Z_{(r)},\end{aligned}\tag{4.3}$$

where $\bar Y=n^{-1}\Sigma_i Y_i$, $\bar\xi_j=n^{-1}\Sigma_i\xi_{ij}$, $\hat\Sigma_{(r)}$ is the $r\times r$ matrix with (j,k)th component $\hat\sigma_{jk}$, $\hat Z_{(r)}=(\hat Z_1,\ldots,\hat Z_r)^{\mathrm{T}}$,

$$\begin{aligned}\hat{\sigma}_{jk} &= \frac{1}{n}\sum_{i=1}^{n}(\xi_{ij}-\bar{\xi}_j)(\xi_{ik}-\bar{\xi}_k)\\ &= \hat{\theta}_j\delta_{jk},\\ \hat{Z}_j &= \frac{1}{n}\sum_{i=1}^{n}(Y_i-\bar{Y})(\xi_{ij}-\bar{\xi}_j),\end{aligned}$$

and again δ_{jk} is the Kronecker delta. Therefore, $\hat{\Sigma}_{(r)} = \mathrm{diag}(\hat{\theta}_1,\ldots,\hat{\theta}_r)$. Hence, by equation (4.3), our estimator of b is

$$\begin{aligned}\hat{b} &= \sum_{j=1}^{r}\hat{b}_j\hat{\psi}_j\\ &= \sum_{j=1}^{r}\hat{\theta}_j^{-1}\hat{Z}_j\hat{\psi}_j.\end{aligned}$$

Elementary manipulations allow us to write the mean-squared error of $\hat{b}$, conditional on $\mathcal{X} = \{X_1,\ldots,X_n\}$, in a simple form, as follows. Assume that the true value b^0 of the function b is square integrable. Then, for each realization of $\mathcal{X}$ for which $\hat{\theta}_r > 0$,

$$\int_{\mathcal{I}} E\{(\hat{b}-b^0)^2|\mathcal{X}\} = n^{-1}(1-n^{-1})\sigma^2\sum_{j=1}^{r}\hat{\theta}_j^{-1} + \sum_{j=r+1}^{\infty}\left(\int_{\mathcal{I}} b^0\hat{\psi}_j\right)^2 \tag{4.4}$$

$$\begin{aligned}&= n^{-1}(1-n^{-1})\sigma^2\sum_{j=1}^{r}\hat{\theta}_j^{-1} + \sum_{j=r+1}^{\infty}\left(\int_{\mathcal{I}} b^0\psi_j\right)^2\\ &\quad - 2\sum_{j=1}^{r}\left(\int_{\mathcal{I}} b^0\psi_j\right)\left\{\int_{\mathcal{I}} b^0(\hat{\psi}_j-\psi_j)\right\} - \sum_{j=1}^{r}\left\{\int_{\mathcal{I}} b^0(\hat{\psi}_j-\psi_j)\right\}^2.\end{aligned} \tag{4.5}$$

Result (4.4) suggests, but of course does not prove, that

$$\int_{\mathcal{I}} E(\hat{b}-b^0)^2 \sim \frac{\sigma^2}{n}\sum_{j=1}^{r}\theta_j^{-1} + \sum_{j=r+1}^{\infty}\left(\int_{\mathcal{I}} b^0\psi_j\right)^2, \tag{4.6}$$

where '$A_n \sim B_n$' means that the ratio of the random variables A_n and B_n converges to 1 as $n\to\infty$. (For technical reasons, to guarantee that the expected value on the left-hand side in expression (4.6) is finite, we replace $\hat{b}_j$ by an arbitrary fixed constant if $|\hat{b}_j| > c_1 n^{c_2}$, where $c_1, c_2 > 0$ are also arbitrary.) If result (4.6) were correct then the first term on the right-hand side would denote the dominant contribution from the error about the mean to integrated squared error, and the second term would be the dominant contribution from the squared bias. The right-hand side of expression (4.6) is reminiscent of familiar formulae for the mean integrated squared error of orthogonal series estimators; see, for example, Kronmal and Tarter (1968).

Conditions under which

$$\left(\sum_{j=1}^{r}\hat{\theta}_j^{-1}\right)\Big/\left(\sum_{j=1}^{r}\theta_j^{-1}\right) \to 1 \tag{4.7}$$

in probability, and hence for which the first term on the right-hand side of expression (4.6) provides a valid approximation to the first term on the right-hand side of equation (4.4), can be established in many circumstances; see Appendix A.1. However, the second term on the right-hand side of expression (4.6) is not always appropriate. We know from Section 2 that properties

of expansions that are founded on the basis functions $\hat{\psi}_j$ should depend on the spacings of the eigenvalues θ_j, and particularly in that respect the behaviour of the squared bias approximation should be treated carefully. Indeed, it may be shown from equations (2.6) and (2.7) that to first order, if X is a Gaussian process,

$$\sum_{j=r+1}^{\infty} E\Big(\int_{\mathcal{I}} b^0 \hat{\psi}_j\Big)^2 = \sum_{j=r+1}^{\infty} \Big(\int_{\mathcal{I}} b^0 \psi_j\Big)^2 + n^{-1} \sum_{j=1}^{r} \theta_j \sum_{k:k\neq j} \theta_k \Big\{ \frac{\beta_k^2}{(\theta_j-\theta_k)^2} - \frac{\beta_j^2}{(\theta_j-\theta_k)} \Big\} + \text{higher order terms}, \tag{4.8}$$

where $\beta_j = \int_{\mathcal{I}} b^0 \psi_j$. A derivation of equation (4.8) is given in Appendix A.1. The effects of eigenvalue spacings are clear in the second term on the right-hand side of equation (4.8) and also make their presence felt in the higher order terms there, although they are not visible in the first term, which of course is the only term that appears on the right-hand side of expression (4.6).

If, along the sequence $\theta_1, \theta_2, \ldots$, there are from time to time very closely spaced eigenvalues, then the term in n^{-1} on the right-hand side of equation (4.8) can make a non-negligible contribution, and the approximation at result (4.6) can fail. However, in other cases result (4.6) is valid; see Appendix A.2 for discussion. Less generally, if the β_js decrease to 0 very rapidly, and in particular if only a finite number of them are non-zero, then difficulties with spacings will be minor.

In the context of functional data analysis, the predictive cross-validation criterion is given by

$$\mathrm{CV}(r) = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \hat{a}_{-i;r} - \int_{\mathcal{I}} \hat{b}_{-i;r} X_i \Big)^2. \tag{4.9}$$

Here, $(\hat{a}_{-i;r}, \hat{b}_{-i;r})$ denotes the least squares estimator of (a, b) that is obtained by confining attention to the set $\mathcal{Z}_i$, say, of all data pairs (X_j, Y_j) excluding the ith, and both $\hat{a}_{-i;r}$ and $\hat{b}_{-i;r}$ use the empirical Karhunen–Loève expansion of length r that is computed from $\mathcal{Z}_i$. We choose r to minimize $\mathrm{CV}(r)$. See Section 5 and Appendix A.3.

5. Numerical properties

5.1. Models used in simulation study

Each X_i was distributed as $X = \Sigma_{j\geqslant 1} \xi_j \psi_j$ and was defined on $\mathcal{I} = [0, 1]$, with $\psi_j(t) = 2^{1/2} \cos(j\pi t)$ and the ξ_js denoting independent variables with zero means and respective variances $\theta_j = j^{-2l}$, for $l = 1, 2, 3$. The last three cases will be referred to as models (i), (ii) and (iii) respectively. The distributions of the ξ_js were either normal $N(0, \theta_j)$ or centred exponential with the same variance. When treating the regression problem the errors ε_i were normal $N(0, 1)$ and we took $a = 0$ and

$$b(t) = \pi^2(t^2 - \tfrac{1}{3}) = \sum_j 2^{3/2}(-1)^j j^{-2} \psi_j(t).$$

For numerical calculation we truncated infinite series, defining X and b, at $j = 20$. All coverages of confidence regions were computed by averaging over 1000 simulated data sets. However, median values of the integrated squared error, which is discussed in Section 5.4, were calculated from 5000 simulated samples.

5.2. Confidence intervals and bands for θ_j and ψ_j

Coverage levels of two-sided, nominal 95% confidence intervals and bands for θ_j and ψ_j are shown in Tables 1 and 2 respectively, in the cases $j = 1, 2$, models (i)–(iii), $n = 20$, 50, 100, 200

Table 1. Coverages of two-sided bootstrap confidence bands for θ_j†

n	*Results for the following models:*					
	(i)		*(ii)*		*(iii)*	
	θ_1	θ_2	θ_1	θ_2	θ_1	θ_2
20	0.88	0.87	0.86	0.84	0.86	0.83
50	0.91	0.91	0.90	0.90	0.90	0.89
100	0.92	0.92	0.92	0.92	0.92	0.92
200	0.94	0.93	0.93	0.94	0.94	0.93

†Blocks of columns refer to cases (i), (ii) and (iii), where $\theta_j = j^{-2l}$ and $l = 1, 2, 3$. Pairs of columns in each block represent θ_j for $j = 1, 2$. Rows indicate sample sizes, $n = 20, 50, 100, 200$. For a given model, a given j and a given sample size, the value in the table gives the coverage, to two-decimal-place accuracy, of a confidence band for θ_j, when the nominal coverage was $1 - \alpha = 0.95$.

Table 2. Coverages of two-sided bootstrap confidence bands for ψ_j†

n	*Results for the following models:*					
	(i)		*(ii)*		*(iii)*	
	ψ_1	ψ_2	ψ_1	ψ_2	ψ_1	ψ_2
20	0.97	0.97	0.95	0.96	0.95	0.98
50	0.97	0.97	0.95	0.96	0.95	0.98
100	0.98	0.97	0.95	0.96	0.96	0.99
200	0.98	0.98	0.96	0.98	0.96	0.99

†All values should be interpreted as in the case of Table 1, except that the tabulated coverages are of confidence bands for ψ_j rather than of confidence intervals for θ_j. The bands are simultaneous in t, although not in j.

and for Gaussian X. These results, and more extensive results that are not given here, reveal the following features. Confidence intervals for θ_j are generally anticonservative, whereas bands for ψ_j tend to be conservative (i.e. have coverage that is greater than the nominal level). The accuracy of coverage of θ_j-intervals generally decreases, in the direction of greater anticonservatism, as j increases, although the accuracy of ψ_j-bands remains relatively stable. Whether the model is (i), (ii) or (iii) has relatively little effect.

If we alter the distribution of ξ_j to centred exponential then the accuracy of coverage of confidence intervals for θ_j declines, in the sense that the intervals become more anticonservative. However, here and in the Gaussian case, coverage correction using the double bootstrap

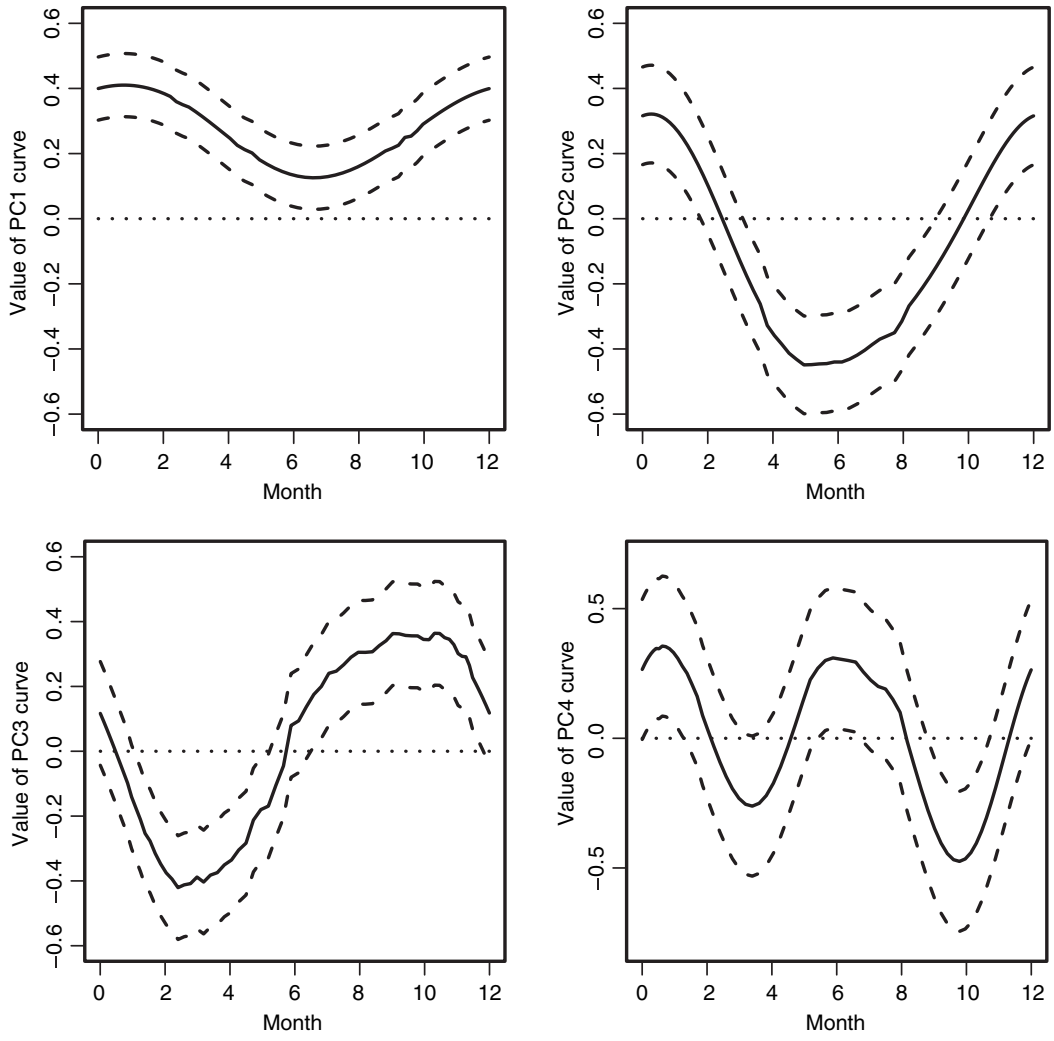

Fig. 1. Simultaneous confidence bands for the first four principal components in Ramsay's Canadian temperature data set: the central curve shows the point estimator $\hat{\psi}_j$, and the upper and lower curves show $\hat{\psi}_j \pm \delta_j$, where δ_j was constructed by double-bootstrap calibration so that a graph of the true function ψ_j lies between the two bounds with probability approximately 0.95

substantially improves the performance, usually removing about half the coverage error through making the intervals less conservative.

For example, for sample sizes $n = 50$, 100, 200, 500, and in the centred exponential case, coverages of nominal 95%-intervals for θ_1 are only 0.80, 0.86, 0.89 and 0.92 respectively, but these figures increase to 0.85, 0.90, 0.92 and 0.95 after double-bootstrap correction. In the cases of models (ii) and (iii) the respective coverages are almost identical to these.

Changing the distribution of ξ_j from normal to centred exponential has little effect on the accuracy of coverage of bands for ψ_j. The accuracy of coverage remains very good, and, for the parameter settings in Table 2, never falls below 0.94. As a result, although double-bootstrap calibration continues to remove about half the coverage error, the effect is much less striking.

Table 3. Coverages of two-sided, simultaneous bootstrap confidence bands for θ_j†

n	*Results for the following distributions:*		
	(i)	*(ii)*	*(iii)*
20	0.93	0.86	0.85
50	0.95	0.90	0.91
100	0.96	0.94	0.93
200	0.96	0.96	0.94

†Columns refer to cases (i), (ii) and (iii); rows are as for Tables 1 and 2. Values in the table are coverages of simultaneous confidence bands for $\theta_1, \theta_2, \ldots$, when the nominal coverage was $1-\alpha=0.95$.

As a practical illustration, Fig. 1 shows 95% bootstrap confidence bands, calibrated by using the double bootstrap, for the first four principal components in the case of J. O. Ramsay's Canadian weather-station temperature data set (Ramsay and Silverman (1997), page 89). The sample size is $n=35$. On the scale of Fig. 1, single-bootstrap confidence bands are almost indistinguishable from their double-bootstrap counterparts. This suggests that the accuracy of coverage of the single-bootstrap bands is already very good. Just this conclusion was reached in the simulation study that was discussed above.

We also explored the case where each X_i was observed discretely on a grid, with additive error, in particular $W_{ij}=X_i(j/20)+\delta_{ij}$, where $1\leqslant j\leqslant 20$ and δ_{ij} was normal $N(0,\sigma^2)$. We passed a conventional local linear smoother through these data, using the method of Sheather and Jones (1991) to choose the bandwidth, thereby constructing an estimator $\hat{X}_i$ of X_i; and then we applied our methods as though $\hat{X}_i$ were X_i. We took $\sigma^2=0.025$, 0.05, 0.1, and calculated the usual empirical approximations to coverage of confidence regions for θ_j and ψ_j, when $n=20,50,100$. Interestingly, the accuracy of coverage of confidence intervals for θ_j is almost always improved by discretizing and adding noise, the extent of improvement generally decreasing with increasing sample size and with decreasing error variance. The accuracy of coverage of confidence bands for ψ_j is hardly affected. (This also occurs in related contexts and so is not unexpected. It arises from the effects of smoothing, including the smoothing that is associated with adding noise.)

5.3. Simultaneous confidence intervals for θ_j

Table 3 gives coverages of simultaneous bootstrap confidence bands (which were introduced in Section 3) for θ_j when the nominal coverage level is 0.95 and X is Gaussian. Here model (i) enjoys especially good performance, with actual coverage 0.95 when $n=50$. For relatively small sample sizes, the accuracy of coverage declines under models (ii) and (iii). There is a further decline in the centred exponential case. For example, when $n=100$, and under models (i), (ii) and (iii), the coverages are respectively 0.90, 0.86 and 0.85. However, as in the contexts that were discussed earlier, double-bootstrap calibration provides substantial correction, removing about half the deficit in coverage.

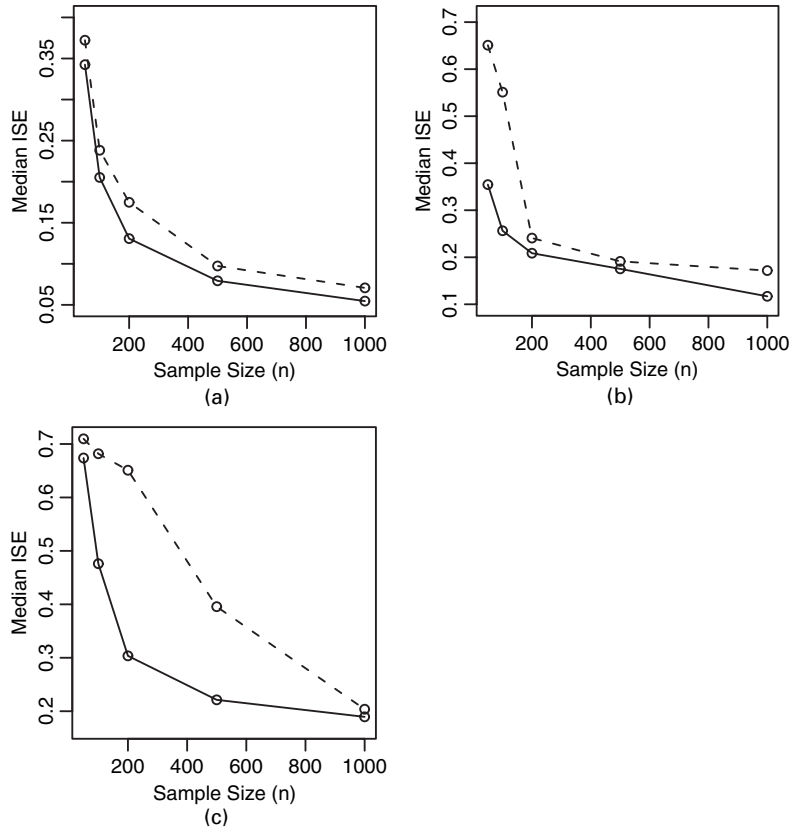

Fig. 2. Median performance of the integrated squared error, when r is chosen optimally or by cross-validation (———, median value of the integrated squared error when $r = r_0$ is chosen to minimize the median; – – –, median when r is selected by cross-validation) (the value of the sample size, n, is graphed on the horizontal axis, and the median integrated squared error is shown on the vertical axis): (a) model (i); (b) model (ii); (c) model (iii)

5.4. Cross-validation in regression

We simulated data from the regression models that were discussed in Section 5.1. For each sample we calculated the value of $\|\hat{b} - b\|^2$, representing the integrated squared error, and analysed, by simulation, the distribution of this quantity. In particular, we calculated the median, med(r) say, of the distribution and found the value r_0 of r that minimized the median. The full lines in Fig. 2 indicate med(r_0). We also computed the smallest value of r, $\hat{r}$ say, that produced a local minimum of CV(r), at equation (4.9), and calculated the median of the distribution of $\|\hat{b} - b\|^2$ when $\hat{b}$ was computed with $r = \hat{r}$. The broken lines in Fig. 2 represent the values of this median. When CV(r) suggested more than one value of r we always chose the smallest, mindful of analogous recommendations in more conventional cases (see for example Park and Marron (1990) and Hall and Marron (1991)).

It can be seen from Fig. 2 that, for model (i), the median performance of $\hat{b}$, computed by using cross-validation to choose r, lies only a little below that when r is selected optimally. The 'effective dimension' grows in passing to models (ii) and (iii), and as a result the cross-validation method has successively greater difficulty with those cases. However, we were pleasantly surprised that a technique that is ostensibly designed for prediction gave reputable performance when used

to estimate the slope function; the latter problem can be expected to require significantly more smoothing than prediction.

6. Conclusions

We have shown how to develop stochastic 'Taylor series expansions' of estimators of eigenvalues and eigenfunctions in functional PCA and have provided a theoretical account of the accuracy of those expansions. The expansions themselves, or the methods that were used to derive them, can be used as the basis for an extensive theory of properties of functional PCA, including the bootstrap. These points have been described theoretically and numerically, and illustrated in the context of regression, but they have implications beyond that setting. For example, they can be used to describe properties of functional data methods applied to classification and clustering.

Acknowledgements

We are grateful to Alan McIntosh, Hans-Georg Müller, Jane-Ling Wang and three reviewers for helpful comments.

Appendix A: Mathematical properties

A.1. Theory behind results (2.6)–(2.11), (4.7) and (4.8)

Take $\mathcal{I}$ to be the unit interval. Put $A = X - E(X)$ and $\delta_j = \min_{k \leqslant j}(\theta_k - \theta_{k+1})$, and assume that,

(a) for all $C > 0$ and some $\varepsilon > 0$,

$$\begin{aligned} &\sup_{t \in \mathcal{I}} \{E|X(t)|^C\} < \infty, \\ &\sup_{s,t \in \mathcal{I}} (E[\{|s-t|^{-\varepsilon} |X(s) - X(t)|\}^C]) < \infty, \end{aligned} \tag{A.1}$$

(b) for each integer $r \geqslant 1$, $\theta_j^{-r} E(\int_{\mathcal{I}} A\psi_j)^{2r}$ is bounded uniformly in j.

For example, assumption (A.1) holds for Gaussian processes with Hölder continuous sample paths. Define $|||\hat{K} - K|||^2 = \int (\hat{K} - K)^2$, $\|\hat{K} - K\|_{\sup}^2 = \sup_u [\int \{(\hat{K} - K)(u, v)\}^2 \, dv]$, $s_j = \sup_u |\psi_j(u)|$ and $s = \sup_u \{K(u, u)\}$. Since $s_j \leqslant (s/\theta_j)^{1/2}$ then s_j, in the bounds in theorem 2 below, may be replaced by $\theta_j^{-1/2}$, although this is generally conservative. Recall that the eigenvalues of the covariance operator K are ordered so that $\theta_1 \geqslant \theta_2 \geqslant \ldots \geqslant 0$. Let $\xi_j = \inf_{k \geqslant j}(1 - \theta_k/\theta_j)$.

Results (2.6)–(2.11) are straightforward corollaries of the following theorem, although they may be obtained separately under weaker conditions. Proofs of theorems 2 and 5, and of more general results, are obtainable from the authors.

Theorem 2. If assumption (A.1) holds, then for each j for which

$$|||\hat{K} - K||| \leqslant \tfrac{1}{2} \min(\theta_j - \theta_{j+1}, \theta_{j-1} - \theta_j), \tag{A.2}$$

the absolute values of the '$O_p(n^{-3/2})$' remainders on the right-hand sides of equations (2.8) and (2.9) are each bounded above by $n^{-3/2} U_{nj} (1 - \xi_j)^{-1/2} \delta_j^{-3} \theta_j^{-1/2} s_j$, where the random variables U_{nj} satisfy $\sup_{n,\, j \geqslant 1} \{E(U_{nj}^C)\} < \infty$ for each $C > 0$. In the case of equation (2.8), this bound is also valid uniformly in t.

We may paraphrase inequality (A.2) by saying that that condition holds for all j for which the distance of θ_j to the nearest other eigenvalue does not fall below $2|||\hat{K} - K|||$, which in turn equals $O_p(n^{-1/2})$. The bounds that are given in theorem 2 imply that the $O_p(n^{-3/2})$ remainders in equations (2.8) and (2.9) equal

$$O_p\{n^{-3/2} j^{\varepsilon} (1 - \xi_j)^{-1/2} \delta_j^{-3} \theta_j^{-1/2} s_j\}$$

uniformly in j for which inequality (A.2) holds and $1 \leqslant j \leqslant n^C$ for each $C, \varepsilon > 0$. In the case of equation (2.8) the bound is also uniform in $t \in \mathcal{I}$.

To derive equation (4.8), note that

$$\sum_{j=r+1}^{\infty}\left\{E\Big(\int_{\mathcal{I}} b^0\hat{\psi}_j\Big)^2-\Big(\int_{\mathcal{I}} b^0\psi_j\Big)^2\right\}=\sum_{j=1}^{r}\left\{\Big(\int_{\mathcal{I}} b^0\psi_j\Big)^2-E\Big(\int_{\mathcal{I}} b^0\hat{\psi}_j\Big)^2\right\}.$$

Use theorem 2 to construct an expansion of $E(\int_{\mathcal{I}} b^0\hat{\psi}_j)^2$, plus a remainder that is of the stated order uniformly in $1\leqslant j\leqslant r$. To complete the proof of equation (4.8), note that, when X is a Gaussian process,

$$E\left\{\Big(\int Z\psi_{j_1}\psi_{j_2}\Big)\Big(\int Z\psi_{j_3}\psi_{j_4}\Big)\right\}=2\theta_{j_1}^2 I_1+\theta_{j_1}\theta_{j_2}I_2+O(n^{-1}),$$

where $I_1=1$ if $j_1=j_2=j_3=j_4$, $I_2=1$ if either $j_1=j_3\neq j_2=j_4$ or $j_1=j_4\neq j_2=j_3$, and both indicators vanish in the respective contrary cases.

The complexity of the bounds in theorem 2 stems mainly from the fact that the expansions there are of relatively high order. Lower order properties, such as those immediately below, are generally simpler.

Theorem 3. If there are no ties for the eigenvalue θ_j, then

$$\sup_{j\geqslant 1}\max\{|\hat{\theta}_j-\theta_j|,\,8^{-1/2}\delta_j\|\hat{\psi}_j-\psi_j\|\}\leqslant|||\hat{K}-K|||, \tag{A.3}$$

$$\begin{aligned}\left|\hat{\theta}_j-\theta_j-\int(\hat{K}-K)\psi_j\psi_j\right|&\leqslant 2\|\hat{\psi}_j-\psi_j\|\,|||\hat{K}-K|||_{\sup}\\&\leqslant 8\delta_j^{-1}|||\hat{K}-K|||\,|||\hat{K}-K|||_{\sup}.\end{aligned} \tag{A.4}$$

The bound (A.3) follows from lemma 4.3 of Bosq (2000) and implies theorem 1 in Section 2. The first inequality in result (A.4) can be proved by using methods of operator theory, and the second follows on applying inequality (A.3) to $\|\hat{\psi}_j-\psi_j\|$.

Conditions under which result (4.7) holds can be quickly deduced from theorem 3, as follows. Assuming result (A.1), simple moment calculations give

$$E\left\{\int(\hat{K}-K)\psi_j\psi_j\right\}^2\leqslant Cn^{-1}\theta_j^2,$$

and also

$$E(|||\hat{K}-K|||\,|||\hat{K}-K|||_{\sup})=O(n^{-1}).$$

From these results and inequality (A.4) it may be proved that

$$\sum_{j=1}^{r}E|\hat{\theta}_j^{-1}-\theta_j^{-1}|=O\Big(n^{-1/2}\sum_{j=1}^{r}\theta_j^{-1}+n^{-1}\sum_{j=1}^{r}\theta_j^{-2}\delta_j^{-1}\Big),$$

from which it follows that result (4.7) holds provided that r increases sufficiently slowly, a sufficient condition for the rate being

$$\Big(\sum_{j\leqslant r}\theta_j^{-2}\delta_j^{-1}\Big)\Big/\Big(n\sum_{j\leqslant r}\theta_j^{-1}\Big)\to 0.$$

A.2. Theory related to regression

First we give a simple sufficient condition on r for the estimator $\hat{b}$ to be consistent for b^0. The condition is based on spacings of eigenvalues and is quite different from a constraint that was imposed by Cardot *et al.* (1999) in a related problem. When the spacings $\theta_j-\theta_{j+1}$ are decreasing, condition (H_3) of Cardot *et al.* (1999) assumes the form

$$(n\theta_r^2)^{-1}\Big(\sum_{j\leqslant r}\delta_j^{-1}\Big)^2\to 0,$$

which is more restrictive than condition (A.5) below. Recall that δ_j was defined at result (2.13).

Theorem 4. If $b^0 \in L_2(\mathcal{I})$ and $\int_{\mathcal{I}} E(X^4) < \infty$, where the random function X has the same distribution as the data X_i, and if $r = r(n) \to \infty$ as $n \to \infty$, in such a manner that

$$\frac{1}{n} \sum_{j=1}^{r} \delta_j^{-2} \to 0, \tag{A.5}$$

then

$$\int_{\mathcal{I}} E\{(\hat{b} - b^0)^2 | \mathcal{X}\} \to 0 \tag{A.6}$$

in probability as $n \to \infty$. In particular, if condition (A.5) holds then $\int_{\mathcal{I}} (\hat{b} - b^0)^2 \to 0$ in probability.

Next we state a result that implies approximation (4.6). Recall that $A = X - E(X)$, define $b_j^0 = \int_{\mathcal{I}} b^0 \psi_j$ and consider the following conditions:

$$\begin{aligned} &\theta_j = j^{-a} L(j) \text{ and } |b_j^0| = j^{-b} M(j), \text{ where } b > a + \tfrac{1}{2} > \tfrac{3}{2} \text{ and } L \text{ and } M \\ &\text{are slowly varying functions; } \theta_j - \theta_{j+1} \geqslant \text{constant} \times j^{-a-1}; \text{ the process } X \text{ has} \\ &\text{all moments finite; for each integer } r \geqslant 1, \theta_j^{-r} E(\textstyle\int A\psi_j)^{2r} \text{ is bounded} \\ &\text{uniformly in } j; \text{ the errors } \varepsilon_i \text{ in equation (4.1) are independent and identically} \\ &\text{distributed with all moments finite, zero mean and variance } \sigma^2; \text{ the} \\ &\text{frequency cut-off } r \text{ is in the range } 1 \leqslant r \leqslant r_0, \text{ where } r_0 = r_0(n) \text{ satisfies} \\ &r_0 = O(n^{(1-\eta)/2(a+1)}) \text{ for some } 0 < \eta < 1. \end{aligned} \tag{A.7}$$

Under these assumptions, r_0 can be chosen so that it is an order of magnitude larger than the value that minimizes the mean integrated squared error.

Theorem 5. To eliminate pathologies arising from too small values of $\hat{\theta}_j$, replace $\hat{b}_j$ by an arbitrary fixed constant if $|\hat{b}_j| > c_1 n^{c_2}$, for any given $c_1, c_2 > 0$. Then, if conditions (A.7) hold, so also does result (4.6), uniformly in $1 \leqslant r \leqslant r_0$.

Assuming that the slowly varying functions L and M in conditions (A.7) are asymptotically constant, the mean integrated squared error of $\hat{b}$ is asymptotic to $C_1 n^{-1} r^{a+1} + C_2 r^{1-2b}$, which is minimized by choosing $r = C_3 n^{1/(a+2b-1)}$, where $C_1, C_2, C_3 > 0$ are constants. The resulting mean-square convergence rate is $n^{-(2b-1)/(a+2b-1)}$. Versions of theorem 5 are available under more general conditions than conditions (A.7); the coefficients θ_j and b_j need only to be 'polynomial like', and in particular need not be regularly varying functions of j.

To derive theorem 4, note that

$$\left[\sum_{j=1}^{r} \left(\int b^0 \psi_j \right) \left\{ \int b^0 (\hat{\psi}_j - \psi_j) \right\} \right]^2 \leqslant \left\{ \int (b^0)^2 \right\}^2 \sum_{j=1}^{r} \|\hat{\psi}_j - \psi_j\|^2,$$

$$\left\{ \int b^0 (\hat{\psi}_j - \psi_j) \right\}^2 \leqslant \left\{ \int (b^0)^2 \right\} \|\hat{\psi}_j - \psi_j\|^2.$$

These results, and result (4.5), imply result (A.6), provided that

$$\frac{1}{n} \sum_{j=1}^{r} \hat{\theta}_j^{-1} \to 0 \quad \text{and} \quad \sum_{j=1}^{r} \|\hat{\psi}_j - \psi_j\|^2 \to 0 \tag{A.8}$$

in probability. Moment arguments may be used to prove that $E(\hat{\Delta}^2) = O(n^{-1})$, and hence that $\hat{\Delta} = O_p(n^{-1/2})$. If result (A.5) holds then $n^{-1/2} \delta_r^{-1} \to 0$, and so $n^{1/2} \min_{j \leqslant r} (\theta_j - \theta_{j+1}) \to \infty$. Hence, for each $C > 0$ and all sufficiently large n, $\theta_j - \theta_{j+1} > C n^{-1/2}$ uniformly in $1 \leqslant j \leqslant r$. Therefore, the probability that $\theta_j - \theta_{j+1} > 2\hat{\Delta}$ for all $1 \leqslant j \leqslant r$ converges to 1 as $n \to \infty$. Equivalently, $1 \leqslant r \leqslant J$. Result (A.3) now implies that

(a) $\max_{j \leqslant r} |\theta_j^{-1} \hat{\theta}_j - 1| \to 0$ in probability and
(b) $\Sigma_{j \leqslant r} \|\hat{\psi}_j - \psi_j\|^2 = O_p(n^{-1} \Sigma_{j \leqslant r} \delta_j^{-2})$.

In view of result (A.5), (a) entails that $\Sigma_{j \leqslant r} \hat{\theta}_j^{-1} \sim_{\mathrm{p}} \Sigma_{j \leqslant r} \theta_j^{-1} = O_p(\Sigma_{j \leqslant r} \delta_j^{-2})$. Therefore, the first part of condition (A.8) follows from result (A.5). Property (b) and result (A.5) imply the second part of condition (A.8).

A.3. Bootstrap and cross-validation

Stochastic Taylor series approximations, such as those which are addressed by theorem 2, make it relatively straightforward to develop rigorous theory for bootstrap confidence intervals and bands for eigenvalues θ_j and eigenfunctions ψ_j. In particular, since the bounds on remainders that are given by theorem 2 are explicit, we can work with the dominant terms and use the remainder bounds to show that high order terms make negligible contributions. The analogue of Cramér's continuity condition, which is needed to ensure that discretization errors do not arise, is continuity of the distributions of the principal components $\xi_j = \int_{\mathcal{I}} X\psi_j$.

If x is a function then, in many circumstances (for example, if $\theta_j \sim \text{constant} \times j^{-a}$, $|\int_{\mathcal{I}} b^0\psi_j| \sim \text{constant} \times j^{-b}$, $|\int_{\mathcal{I}} x\psi_j| \sim \text{constant} \times j^{-c}$, and b and c are sufficiently large), the predictor $\int_{\mathcal{I}} \hat{b}x$ converges to $\int_{\mathcal{I}} bx$ at rate $n^{-1/2}$, provided that the frequency cut-off r is chosen appropriately. In such cases, predictive cross-validation and related methods, such as those which are based on thresholding, can be used to choose r empirically and achieve the root n rate.

References

Aguilera, A. M., Ocana, F. A. and Valderrama, M. J. (1999a) Forecasting time series by functional PCA: discussion of several weighted approaches. *Comput. Statist.*, **14**, 443–467.

Aguilera, A. M., Ocana, F. A. and Valderrama, M. J. (1999b) Forecasting with unequally spaced data by a functional principal component approach. *Test*, **8**, 233–253.

Besse, P. (1992) PCA stability and choice of dimensionality. *Statist. Probab. Lett.*, **13**, 405–410.

Besse, P. and Ramsay J. O. (1986) Principal components-analysis of sampled functions. *Psychometrika*, **51**, 285–311.

Bhatia, R., Davis, C. and McIntosh, A. (1983) Perturbation of spectral subspaces and solution of linear operator equations. *Lin. Alg. Appl.*, **52–53**, 45–67.

Boente, G. and Fraiman, R. (2000) Kernel-based functional principal components. *Statist. Probab. Lett.*, **48**, 335–345.

Bosq, D. (1989) Propriétés des opérateurs de covariance empiriques d'un processus stationnaire hilbertien. *C. R. Acad. Sci. Par.* I, **309**, 873–875.

Bosq, D. (1991) Modelization, nonparametric estimation and prediction for continuous time processes. *NATO Adv. Sci. Inst.* C, **335**, 509–529.

Bosq, D. (2000) Linear processes in function spaces: theory and applications. *Lect. Notes Statist.*, **149**.

Brumback, B. A. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Am. Statist. Ass.*, **93**, 961–976.

Cardot, H. (2000) Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *J. Nonparam. Statist.*, **12**, 503–538.

Cardot, H., Ferraty, F. and Sarda, P. (1999) Functional linear model. *Statist. Probab. Lett.*, **45**, 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2000) Étude asymptotique d'un estimateur spline hybride pour le modèle linéaire fonctionnel. *C. R. Acad. Sci. Par.* I, **330**, 501–504.

Dauxois, J., Pousse, A. and Romain, Y. (1982) Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multiv. Anal.*, **12**, 136–154.

Girard, S. (2000) A nonlinear PCA based on manifold approximation. *Comput. Statist.*, **15**, 145–167.

Hall, P. and Marron, J. S. (1991) Local minima in cross-validation functions. *J. R. Statist. Soc.* B, **53**, 245–252.

He, G. Z., Müller, H.-G. and Wang, J.-L. (2003) Functional canonical analysis for square integrable stochastic processes. *J. Multiv. Anal.*, **85**, 54–77.

Huang, J. H. Z., Wu, C. O. and Zhou, L. (2002) Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, **89**, 111–128.

Indritz, J. (1963) *Methods in Analysis*. New York: Macmillan.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

Kneip, A. and Utikal, K. J. (2001) Inference for density families using functional principal component analysis. *J. Am. Statist. Ass.*, **96**, 519–532.

Kronmal, R. and Tarter, M. (1968) The estimation of probability densities and cumulatives by Fourier series methods. *J. Am. Statist. Ass.*, **63**, 925–952.

Mas, A. (2002) Weak convergence for the covariance operators of a Hilbertian linear process. *Stoch. Process. Applic.*, **99**, 117–135.

Ocana, F. A., Aguilera, A. M. and Valderrama, M. J. (1999) Functional principal components analysis by choice of norm. *J. Multiv. Anal.*, **71**, 262–276.

Park, B. U. and Marron, J. S. (1990) Comparison of data-driven bandwidth selectors. *J. Am. Statist. Ass.*, **85**, 66–72.

Ramsay, J. O. and Dalzell, C. J. (1991) Some tools for functional data analysis (with discussion.) *J. R. Statist. Soc.* B, **53**, 539–572.
Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. New York: Springer.
Ramsay, J. O. and Silverman, B. W. (2002) *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer.
Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc.* B, **53**, 233–243.
Sheather, S. J. and Jones, M. C. (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc.* B, **53**, 683–690.
Silverman, B. W. (1995) Incorporating parametric effects into functional principal components analysis. *J. R. Statist. Soc.* B, **57**, 673–689.
Silverman, B. W. (1996) Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, **24**, 1–24.
Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.