Selecting the Number of Principal Components in Functional Data

Yehua LI, Naisyin WANG, and Raymond J. CARROLL

Functional principal component analysis (FPCA) has become the most widely used dimension reduction tool for functional data analysis. We consider functional data measured at random, subject-specific time points, contaminated with measurement error, allowing for both sparse and dense functional data, and propose novel information criteria to select the number of principal component in such data. We propose a Bayesian information criterion based on marginal modeling that can consistently select the number of principal components for both sparse and dense functional data. For dense functional data, we also develop an Akaike information criterion based on the expected Kullback–Leibler information under a Gaussian assumption. In connecting with the time series literature, we also consider a class of information criteria proposed for factor analysis of multivariate time series and show that they are still consistent for dense functional data, if a prescribed undersmoothing scheme is undertaken in the FPCA algorithm. We perform intensive simulation studies and show that the proposed information criteria vastly outperform existing methods for this type of data. Surprisingly, our empirical evidence shows that our information criteria proposed for dense functional data also perform well for sparse functional data. An empirical example using colon carcinogenesis data is also provided to illustrate the results. Supplementary materials for this article are available online.

KEY WORDS: Akaike information criterion; Bayesian information criterion; Functional data analysis; Kernel smoothing; Model selection.

1. INTRODUCTION

Advances in technology have made functional data (Ramsay and Silverman 2005) increasingly available in many scientific fields, such as many longitudinal data in medical, biological research, electroencephalography, and functional magnetic resonance imaging data. There is tremendous research interest in functional data analysis (FDA) for the past decade. Among the newly developed methodology, functional principal component analysis (FPCA) has become the most widely used dimension reduction tool for FDA. There is some existing work on selecting the number of functional principal components, but to the best of our knowledge, none of them were rigorously studied either theoretically or empirically. In this article, we consider functional data that are observed at random, subject-specific observation times, allowing for both sparse and dense functional data. We propose novel information criteria to select the number of principal components, and investigate their theoretical and empirical performance.

There are two main streams of methods for FPCA: kernel-based FPCA methods including Yao, Müller, and Wang (2005a) and Hall, Müller, and Wang (2006) and spline-based methods including Rice and Silverman (1991), James and Hastie (2001), and Zhou, Huang, and Carroll (2008). Some applications of FPCA include functional generalized linear models (Müller and Studtmüller 2005; Yao, Müller and Wang 2005b; Cai and Hall 2006; Li, Wang, and Carroll 2010) and functional sliced inverse regression (Li and Hsing 2010a).

At this point, the kernel-based FPCA methods are better understood in terms of theoretical properties. This is due to the work of Hall and Hosseini-Nasab (2006), who proved various asymptotic expansions of the estimated eigenvalues and eigenfunction for dense functional data, and by Hall, Müller, and Wang (2006) who provided the optimal convergence rate of FPCA in sparse functional data. An important result of Hall, Müller, and Wang (2006) was that, although FPCA is applied to the covariance function estimated by a two-dimensional smoother, when the bandwidths were properly tuned, estimating the eigenvalues is a semiparametric problem and enjoys a root n convergence rate, and estimating the eigenfunctions is a nonparametric problem with the convergence rate of a one-dimensional smoother.

In the work on FDA mentioned above, functional data were classified as (a) dense functional data where the curves are densely sampled so that passing a smoother on each curve can effectively recover the true sample curves (Hall, Müller, and Wang 2006) and (b) sparse functional data where the number of observations per curve is bounded by a finite number and pooling all subjects together is required to obtain consistent estimates of the principal components (Yao, Müller, and Wang 2005a; Hall, Müller, and Wang 2006). There has been a gap in methodologies for dealing with these two types of data. Hall, Müller, and Wang (2006) showed that when the number of observations per curve diverges to ∞ with a rate of at least $n^{1/4}$, the presmoothing approach is justifiable and the errors in smoothing each individual curve are asymptotically negligible. However, in reality it is hard to decide when the observations are dense enough. In some longitudinal studies, it is possible that we have dense observations on some subjects and sparse observations on the others. In view of these difficulties, Li and

Yehua Li is Associate Professor, Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011 (E-mail: yehuali@iastate.edu). Naisyin Wang is Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 (E-mail: nwangaa@umich.edu). Raymond J. Carroll is Distinguished Professor, Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143 (E-mail: carroll@stat.tamu.edu). Li's research was supported by the National Science Foundation (DMS-1105634, DMS-1317118). Wang's research was supported by a grant from the National Cancer Institute (CA74552). Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030) and by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank the associate editor and two anonymous referees for their constructive comments that led to significant improvements in the article.

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.788980

Hsing (2010b) studied all types of functional data in a unified framework, and derived a strong uniform convergence rate for FPCA, where the number of observations per curve can be of any rate relative to the sample size.

A common finding in the aforementioned work is that higher-order principal components are much harder to estimate and harder to interpret. Because seeking sparse representation of the data is at the core of modern statistics, it is reasonable in many situations to model the high-order principal components as noise. Therefore, selecting the number of principal components is an important model selection problem in almost all practical contexts of FDA. Yao, Müller, and Wang (2005a) proposed an Akaike information criterion (AIC) criterion for selecting the number of principal components in sparse functional data. However, so far there is no theoretical justification for this approach, and whether this criterion also works for dense functional data or the types of data in the gray zone between sparse and dense functional data remains unknown. Hall and Vial (2006) included theoretical discussion about the difficulty of selecting the number of principal components using a hypothesis testing approach. The bootstrap approach proposed by Hall and Vial provides a confidence lower bound $\widehat{v}_q$ for the "unconfounded noise variance," and can provide some guidance in selecting the number of principal components. However, their approach is not a real model selection criterion, and one needs to watch the decreasing trend of $\widehat{v}_q$ and decide the cut point subjectively. The minimum description length (MDL) method by Poskitt and Sengarapillai (2013) is similar to Yao's AIC in that each principal component is counted as one parameter, although of course the criteria are numerically different. We emphasize that, in reality, each principal component consists of one variance parameter and one nonparametric function. A main point of our article is to justify how much penalty is needed in a model selection criterion, when selecting the number of nonparametric components in the data.

We approach this problem from three directions, with all approaches built upon the foundation of information criteria. In the marginal modeling approach, we focus on the decay rate of the estimated eigenvalues and develop a Bayesian information criterion (BIC)-based selection method. The advantages of this approach include that it only uses existing outcomes from FPCA, namely, the estimated eigenvalues and the residual variance, and that it is consistent for all types of functional data. As an alternative, we find that, with some additional assumptions, a modified AIC based on conditional likelihood could produce superior numerical outcomes. A referee pointed out to us that when the data are observed densely on a regular grid, where no kernel smoothing is necessary, there is some existing work in the econometrics literature based on a factor analysis model (Bai and Ng 2002) to select the number of principal components. We study this class of information criteria in our setting and find out that they are still consistent if a specific undersmoothing scheme is carried out in the FPCA method. In addition, we also provide some discussion for the case that the true number of principal components diverges to infinity.

The remainder of the article is organized as follows. In Section 2, we describe the data structure and the FPCA algorithm. In Sections 3.1 and 3.2, we propose and study the new marginal BIC and conditional AIC criteria, and we investigate the information criteria by Bai and Ng in Section 3.3. The proposed information criteria are tested by simulation studies in Section 4, and applied to an empirical example in Section 5. Some concluding remarks are given in Section 6, where we also provide discussion for the case that the true number of principal components diverges. All proofs are provided in the supplementary material.

2. FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

2.1 Data Structure and Model Assumptions

Let $X(t)$ be the functional data defined on a fixed interval $\mathcal{T} = [a, b]$, with mean function $\mu(t)$ and covariance function $R(s, t) = \mathrm{cov}\{X(s), X(t)\}$. Suppose the covariance function has the eigen-decomposition $R(s, t) = \sum_{j=1}^{\infty} \omega_j \psi_j(s)\psi_j(t)$, where the ω_j are the nonnegative eigenvalues of $R(\cdot, \cdot)$, which, without loss of generality, satisfy $\omega_1 \geq \omega_2 \geq \cdots > 0$, and the ψ_j are the corresponding eigenfunctions.

Although, in theory, the spectral decomposition of the covariance function consists of infinite number of terms, to motivate practically useful information criteria, it is sensible to assume that there is a finite-dimensional true model. Due to the nature of spectral decomposition, the higher-order terms are less reliably assessed and their estimates tend to have high variation. Consequently, even though one could assume that there are an infinite number of components, unless the data size is very large, sensible variable selection criteria will still select a relatively small number of components—the first several that can be reasonably assessed. This phenomenon is reflected by the numerical outcomes reported in Table S.7 of the supplementary material, in which a much-improved performance of BIC is observed when the sample size increases to 2000. The performance of BIC is mostly determined by the accuracy of detecting nonzero eigenvalues and that this detection can be difficult for higher-order terms. For the rest of the article, except for Section 6.2, we assume that the spectral decomposition of R ends at a finite p terms, that is, $\omega_j = 0$ for $j > p$. Then the Karhunen–Loève expansion of $X(t)$ is

$$X(t) - \mu(t) = \sum_{j=1}^{p} \xi_j \psi_j(t), \tag{1}$$

where $\xi_j = \int \psi_j(t)\{X(t) - \mu(t)\}dt$ has mean zero, with $\mathrm{cov}(\xi_j, \xi_{j'}) = I(j = j')\omega_j$. Let p_0 be the true value of p.

Suppose we sample from n independent sample trajectories, $X_i(\cdot)$, $i = 1, \ldots, n$. It often happens that the observations contain additional random errors and instead we observe

$$W_{ij} = X_i(t_{ij}) + U_{ij}, \quad j = 1, \ldots, m_i, \tag{2}$$

where U_{ij} are independent zero-mean errors, with $\mathrm{var}\{U_i(t)\} = \sigma_u^2$, and the U_{ij} are also independent of $X_i(\cdot)$. Here, the (t_{ij}) are random, subject-specific observation times. Suppose t_{ij} has a continuous density $f_1(t)$ with support $\mathcal{T}$. We adopt the framework in Li and Hsing (2010b) so that m_i can be of any rate relative to n. The only assumption on m_i is that all $m_i \geq 2$, so that we can estimate the within-curve covariance matrix. In other words, we allow m_i to be bounded by a finite number as in sparse functional data, or diverging to ∞ as in dense functional data.

2.2 Functional Principal Component Analysis

The functions $\mu(\cdot)$ and $R(\cdot,\cdot)$ can be estimated by local polynomial regression, and then $\psi_k(\cdot)$, ω_k, and σ_u^2 can be estimated using the FPCA method proposed in Yao, Müller, and Wang (2005a) and Hall, Müller, and Wang (2006). We now briefly describe the method. We first estimate $\mu(\cdot)$ by a local linear regression, $\widehat{\mu}(t) = \widehat{a}_0$, where $(\widehat{a}_0, \widehat{a}_1) = \mathrm{argmin}_{a_0,a_1}$ $n^{-1}\sum_{i=1}^n m_i^{-1}\sum_{j=1}^{m_i}\{W_{ij} - a_0 - a_1(t_{ij} - t)\}^2 K\{(t_{ij} - t)/h_\mu\}$, $K(\cdot)$ is a symmetric density function, and h_μ is the bandwidth for estimating μ. Define $C_{XX}(s,t) = \mathrm{E}\{X(s)X(t)\}$ and $M_i = (m_i - 1)m_i$. We denote the bandwidth for estimating $C_{XX}(\cdot,\cdot)$ by h_C and let $\widehat{C}_{XX}(s,t) = \widehat{b}_0$, where $(\widehat{b}_0, \widehat{b}_1, \widehat{b}_2)$ minimizes

$$n^{-1}\sum_{i=1}^n M_i^{-1}\sum_{j=1}^{m_i}\sum_{k\neq j}\{W_{ij}W_{ik} - b_0 - b_1(t_{ij} - s) - b_2(t_{ik} - t)\}^2 \times K\left(\frac{t_{ij} - s}{h_C}\right)K\left(\frac{t_{ik} - t}{h_C}\right).$$

Then $\widehat{R}(s,t) = \widehat{C}_{XX}(s,t) - \widehat{\mu}(s)\widehat{\mu}(t)$. In addition, (ω_k) and $\{\psi_k(\cdot)\}$ can be estimated from an eigenvalue decomposition of $\widehat{R}(\cdot,\cdot)$ by discretization of the smoothed covariance function, see Rice and Silverman (1991) and Capra and Müller (1997). Let $\sigma_w^2(t) = \mathrm{var}\{W(t)\} = R(t,t) + \sigma_u^2$ and $\widehat{\sigma}_w^2(t) = \widehat{c}_0 - \widehat{\mu}^2(t)$, where, with a given bandwidth, h_σ, $(\widehat{c}_0, \widehat{c}_1)$ minimizes $n^{-1}\sum_{i=1}^n m_i^{-1}\sum_{j=1}^{m_i}\{W_{ij}^2 - c_0 - c_1(t_{ij} - t)\}^2 K\{(t_{ij} - t)/h_\sigma\}$. One possible estimator if σ_u^2 is

$$\widetilde{\sigma}_{u,\mathrm{I}}^2 = (b-a)^{-1}\int_a^b\left\{\widehat{\sigma}_w^2(t) - \widehat{R}(t,t)\right\}dt. \qquad (3)$$

Define $\widehat{\omega}_k$ and $\widehat{\psi}_k(\cdot)$ to be the kth eigenvalue and eigenfunction of $\widehat{R}(s,t)$, respectively. Rates of convergence results for $\widehat{\mu}(\cdot)$, $\widehat{R}(\cdot)$, $\widehat{\sigma}_w^2(\cdot)$, $\widetilde{\sigma}_{u,\mathrm{I}}^2$, and $\widehat{\psi}_k(\cdot)$ are described in the supplementary material, Section S.1.

3. METHODOLOGY

3.1 Marginal Bayesian Information Criterion

In a traditional regression setting with sample size n, parameter size p, and normally distributed errors of mean zero and variance σ_u^2, BIC is commonly defined as

$$\log(\sigma^2) + p\log(n)/n.$$

Considering the model Equations (1) and (2), linking the current setup for each subject, and then marginalizing over all subjects, we consider a generalized BIC criterion of the structure of

$$\log\left(\widehat{\sigma}_u^2\right) + \mathcal{P}_n(p), \qquad (4)$$

where $\widehat{\sigma}_u^2$ is an estimate of σ_u^2 by marginally pooling error information from all subjects and $\mathcal{P}_n(p)$ is a penalty term. Even though the concept behind our criterion has been motivated by the traditional BIC in regression setting, there are some marked differences. For example, the ξ_j in model (1) are random. As a result, marginally, there are not np parameters. Further, unlike the traditional regression problems, we do not need to estimate/predict ξ_j. Consequently, the number of parameters in a marginal analysis is not determined by the degrees of freedom of these unknown ξ_j. Inspired by standard BIC, we let the penalty be of the form $\mathcal{P}_n(p) = C_{n,p}p$ and then determine the rate of $C_{n,p}$.

Let $\widehat{\sigma}_{u,[p]}^2$ be the estimator of $\widehat{\sigma}_u^2$ based on the residuals after taking into account of the first p principal components. Define

$$R_{[p]}(s,t) = \sum_{j=1}^p \omega_j\psi_j(s)\psi_j(t),$$
$$\widehat{R}_{[p]}(s,t) = \sum_{j=1}^p \widehat{\omega}_j\widehat{\psi}_j(s)\widehat{\psi}_j(t).$$

If p is the true number of principal components, then $R_{[p]}(s,t) = R(s,t)$. Since $\int_a^b \widehat{\psi}_k^2(t)dt = 1$ for all k, we can estimate σ_u^2 by

$$\widehat{\sigma}_{[p],\mathrm{marg}}^2 = \frac{1}{b-a}\int\left\{\widehat{\sigma}_w^2(t) - \widehat{R}_{[p]}(t,t)\right\}dt = \frac{1}{b-a}\int\widehat{\sigma}_w^2(t)dt - \frac{1}{b-a}\sum_{k=1}^p\widehat{\omega}_k. \qquad (5)$$

Replacing $\widehat{\sigma}_u^2$ by $\widehat{\sigma}_{[p],\mathrm{marg}}^2$ in (4), the new BIC criterion is given by

$$\mathrm{BIC}(p) = \log\left(\widehat{\sigma}_{[p],\mathrm{marg}}^2\right) + \mathcal{P}_n(p). \qquad (6)$$

That is, instead of estimating $\widehat{\sigma}_{u,[p]}^2$ from the estimated residuals, we will estimate it from a "marginal" approach by pooling all subjects together. This way, we avoid estimating the principal component scores and dealing with the estimation errors in them.

Denote $\|\cdot\|$ as the L^2 functional norm, and define $\gamma_{nk} = (n^{-1}\sum_{i=1}^n m_i^{-k})^{-1}$, which is the kth harmonic mean of the m_i's. When $m_i = m$ for all i, we have that $\gamma_{n1} = m$ and $\gamma_{n2} = m^2$. For any bandwidth h, define

$$\delta_{n1}(h) = [\{1 + (h\gamma_{n1})^{-1}\}/n]^{1/2},$$
$$\delta_{n2}(h) = [\{1 + (h\gamma_{n1})^{-1} + (h^2\gamma_{n2})^{-1}\}/n]^{1/2}.$$

We make the following assumptions.

- (C.1) The observations time $t_{ij} \sim f_1(t)$, $(t_{ij}, t_{ij'}) \sim f_2(t_1, t_2)$, where f_1 and f_2 are continuous density functions with bounds $0 < m_T \leq f_1(t_1), f_2(t_1, t_2) \leq M_T < \infty$ for all $t_1, t_2 \in \mathcal{T}$. Both f_1 and f_2 are differentiable with bounded (partial) derivatives.
- (C.2) The kernel function $K(\cdot)$ is a symmetric probability density function on $[-1, 1]$, and is of bounded variation on $[-1, 1]$. Denote $\nu_2 = \int_{-1}^1 t^2K(t)dt$.
- (C.3) $\mu(\cdot)$ is twice differentiable and its second derivative is bounded on $[a, b]$.
- (C.4) All second-order partial derivatives of $R(s,t)$ exist and are bounded on $[a, b]^2$.
- (C.5) There exists $C > 4$ such that $\mathrm{E}(|U_{ij}|^C) + \mathrm{E}\{\sup_{t\in[a,b]}|X(t)|^C\} < \infty$.
- (C.6) $h_\mu, h_C, h_\sigma, \delta_{n1}(h_\mu), \delta_{n2}(h_C), \delta_{n1}(h_\sigma) \to 0$ as $n \to \infty$.
- (C.7) We have $\omega_1 > \omega_2 > \cdots > \omega_{p_0} > 0$ and $\omega_k = 0$ for all $k > p_0$.

Let $\widehat{p}$ be the minimizer of BIC(p). The following theorem gives a sufficient condition for $\widehat{p}$ to be consistent to p_0.

Theorem 1. Make assumptions (C.1)–(C.7). Recall that $\mathcal{P}_n(p)$ is the penalty defined in (6), and define $\delta_n^* = h_\mu^2 + \delta_{n1}(h_\mu) + h_C^2 + \delta_{n2}(h_C)$. Suppose the following conditions hold

(i) for any $p < p_0$, $\mathrm{pr}[\mathrm{limsup}_{n\to\infty}\{\mathcal{P}_n(p_0) - \mathcal{P}_n(p)\} \le 0] = 1$;

(ii) for any $p > p_0$, $\mathrm{pr}[\mathcal{P}_n(p) > \mathcal{P}_n(p_0), \limsup_{n\to\infty} \delta_n^*/\{\mathcal{P}_n(p) - \mathcal{P}_n(p_0)\} = 0] = 1$.

Then $\lim_{n\to\infty} \mathrm{pr}(\widehat{p} = p_0) = 1$.

By Theorem 1, there is a large range of penalties that can result in a consistent BIC criterion. For example, let $N = \sum_i m_i$ and recall that the penalty term $\mathcal{P}_n(p) = C_{n,p}p$. If we let $C_{n,p} \sim \log(N)\delta_n^*$, it is easy to verify that the conditions in Theorem 1 are satisfied.

We now derive a databased version of $\mathcal{P}_n(p)$ that satisfies conditions (i) and (ii). By Lemma S.1.1 in the supplementary material, δ_n^* is actually the L^2 convergence rate of $\widehat{R}(\cdot,\cdot)$, which by Lemma S.1.3 in the supplementary material is also the bound for the null eigenvalues, $\{\widehat{\omega}_k; k > p_0\}$. In reality, $\|\widehat{R} - R\|$ not only depends on δ_n^* but also on unknown constants depending on the true function $R(\cdot,\cdot)$ and the distribution of W. To make the information criterion data-adaptive, we propose the following penalty:

$$\mathcal{P}_{n,\mathrm{adapt}}(p) = \log(N)p\|\widehat{R} - \widehat{R}_{[p]}\|/\widetilde{\sigma}_{u,\mathrm{I}}^2. \tag{7}$$

Justification for (7) is given in the supplementary material, Section S.2.

3.2 Akaike Information Criterion Based on Conditional Likelihood

The marginal BIC criterion can be computed by using outcomes from FPCA directly and it is consistent. However, its performances heavily rely on the precision in estimating ω_j, particularly when j is near the true number of principle components, p_0. It is known that the estimation of ω_j can deteriorate when j increases. In this subsection, we propose an alternative approach that, by having some additional conditions, allows us to take advantage of the use of likelihood. We consider the principal component scores as random effects, and proposed a new AIC criterion based on the conditional likelihood and estimated principal component scores. Such an approach is referred as conditional AIC in linear mixed models, see Claeskens and Hjort (2008). In an alternative context, Hurvich, Simonoff, and Tsai (1998) proposed an AIC criterion for choosing the smoothing parameters in nonparametric smoothing. The FPCA method is to project the discrete longitudinal trajectories on some nonparametric functions (i.e., the eigenfunctions), and can thus be considered as simultaneously smoothing n curves. The AIC in the FPCA context is connected to that for the nonparametric smoothing problem, but the way of counting the effective number of parameters in the model will be different. Therefore, the penalty in our AIC will also be very different from that of the nonparametric smoothing problem.

Define $\boldsymbol{W}_i = (W_{i1}, \dots, W_{i,m_i})^{\mathrm{T}}$, $\boldsymbol{\mu}_i = \{\mu(t_{i1}), \dots, \mu(t_{i,m_i})\}^{\mathrm{T}}$, and $\boldsymbol{\psi}_{ik} = \{\psi_k(t_{i1}), \dots, \psi_k(t_{i,m_i})\}^{\mathrm{T}}$. Under the assumption that there are p nonzero eigenvalues, denote $X_{i,[p]}(t) = \mu(t) + \sum_{j=1}^p \xi_{ip}\psi_j(t)$ and $\boldsymbol{X}_{i,[p]} = \{X_{i,[p]}(t_{i1}), \dots, X_{i,[p]}(t_{i,m_i})\}^{\mathrm{T}} = \boldsymbol{\mu}_i + \Psi_{i,[p]}\boldsymbol{\xi}_{i,[p]}$, where $\Psi_{i,[p]} = (\boldsymbol{\psi}_{i1}, \dots, \boldsymbol{\psi}_{ip})$ and $\boldsymbol{\xi}_{i,[p]} = (\xi_{i1}, \dots, \xi_{ip})^{\mathrm{T}}$. Under a Gaussian assumption, the conditional log-likelihood of the observed data $\{\boldsymbol{W}_i\}$ given the principal component scores is

$$\begin{aligned}
&\mathcal{L}_{n,\mathrm{cond}}\left(p, \boldsymbol{X}_{[p]}, \sigma_u^2\right)\\
&\quad= \sum_{i=1}^n \left\{-\,(m_i/2)\log\left(2\pi\sigma_u^2\right) - \left(2\sigma_u^2\right)^{-1}\|\boldsymbol{W}_i - \boldsymbol{X}_{i,[p]}\|^2\right\}\\
&\quad= -(N/2)\log\left(2\pi\sigma_u^2\right) - \left(2\sigma_u^2\right)^{-1}\\
&\qquad\times \sum_{i=1}^n \|\boldsymbol{W}_i - \boldsymbol{\mu}_i - \Psi_{i,[p]}\boldsymbol{\xi}_{i,[p]}\|^2,
\end{aligned} \tag{8}$$

where $N = \sum_i m_i$ and $\boldsymbol{X}_{[p]} = (\boldsymbol{X}_{1,[p]}^{\mathrm{T}}, \dots, \boldsymbol{X}_{n,[p]}^{\mathrm{T}})^{\mathrm{T}}$.

Following the method proposed by Yao, Müller, and Wang (2005a), we estimate the trajectories by

$$\widehat{X}_{i,[p]}(t) = \widehat{\mu}(t) + \sum_{j=1}^p \widehat{\xi}_{ij}\widehat{\psi}_j(t), \tag{9}$$

where $\widehat{\mu}(\cdot)$ and $\widehat{\psi}_j(\cdot)$ are the estimators described in Section 2. The estimated principal component scores, $\widehat{\xi}_{ij}$, are given by the principal component analysis through the conditional expectation (PACE) estimator by Yao, Müller, and Wang (2005a). Under the Gaussian model, the best linear unbiased predictor (BLUP) for $\boldsymbol{\xi}_{i,[p]}$ is $\widetilde{\boldsymbol{\xi}}_{i,[p]} = \Lambda_{[p]}\Psi_{i,[p]}^{\mathrm{T}}\Sigma_{i,[p]}^{-1}(\boldsymbol{W}_i - \boldsymbol{\mu}_i)$, where $\Lambda_{[p]} = \mathrm{diag}(\omega_1, \dots, \omega_p)$, $\Sigma_{i,[p]} = \Omega_{i,[p]} + \sigma_u^2 I_{m_i}$, and $\Omega_{i,[p]} = \Psi_{i,[p]}\Lambda_{[p]}\Psi_{i,[p]}^{\mathrm{T}}$. To estimate $\widetilde{\boldsymbol{\xi}}_{i,[p]}$, the PACE estimator requires a pilot estimator of σ_u^2, for which we can use the integral estimator $\widetilde{\sigma}_{u,\mathrm{I}}^2$ defined in (3). The PACE estimator is given by

$$\widehat{\boldsymbol{\xi}}_{i,[p]} = \widehat{\Lambda}_{[p]}\widehat{\Psi}_{i,[p]}^{\mathrm{T}}\widehat{\Sigma}_{i,[p]}^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i), \tag{10}$$

where $\widehat{\boldsymbol{\mu}}_i$, $\widehat{\Lambda}_{[p]}$, and $\widehat{\Psi}_{i,[p]}$ are the estimates using the FPCA method described in Section 2, and $\widehat{\Sigma}_{i,[p]} = \widehat{\Psi}_{i,[p]}\widehat{\Lambda}_{[p]}\widehat{\Psi}_{i,[p]}^{\mathrm{T}} + \widetilde{\sigma}_{u,\mathrm{I}}^2 I$.

To choose p, Yao, Müller, and Wang (2005a) proposed the pseudo-AIC

$$\mathrm{AIC}_{\mathrm{Yao}}(p) = \mathcal{L}_{n,\mathrm{cond}}\left(p, \widehat{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}_{u,\mathrm{I}}^2\right) + p, \tag{11}$$

where $\widehat{\boldsymbol{X}}_{[p]}$ is the estimated value of $\boldsymbol{X}_{[p]}$ by interpolating the estimated trajectories defined in (9) on the subject-specific times. By adding a penalty p to the estimated conditional likelihood, Yao et al. essentially counted each principal component as one parameter.

To motivate our own AIC criterion, we consider dense functional data satisfying

$$m_i \asymp m \to \infty \text{ for all } i, \quad \sup_i |m_i - m|/m \to 0. \tag{12}$$

We follow the spirit of the derivation of Hurvich and Tsai (1989), and define the Kullback–Leibler information to be

$$\Delta(p, \widetilde{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}^2) = \mathrm{E}_F\{-2\mathcal{L}_{n,\mathrm{cond}}(p, \widetilde{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}^2)\}, \tag{13}$$

for any fixed $\widetilde{\boldsymbol{X}}_{[p]}$ and $\widetilde{\sigma}^2$, where F is the true normal distribution given the true curves $\{X_i(\cdot), i = 1, \dots, n\}$. Using similar derivations as in Hurvich and Tsai (1989), for any fixed parameters $\widetilde{\boldsymbol{X}}_{[p]} = \{\widetilde{\boldsymbol{X}}_{i,[p]} = \widetilde{\boldsymbol{\mu}}_i + \widetilde{\Psi}_{i,[p]}\widetilde{\boldsymbol{\xi}}_{i,[p]}\}_{i=1}^n$ and $\widetilde{\sigma}^2$, we have

$$\begin{aligned}
&\Delta(p, \widetilde{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}^2)\\
&\quad = N\log(2\pi\widetilde{\sigma}^2) + \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^{n} \mathrm{E}_F\|\boldsymbol{U}_i + \boldsymbol{X}_i - \widetilde{\boldsymbol{X}}_{i,[p]}\|\\
&\quad = N\log(2\pi\widetilde{\sigma}^2) + N\frac{\sigma_u^2}{\widetilde{\sigma}^2} + \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^{n}\|(\boldsymbol{\mu}_i - \widetilde{\boldsymbol{\mu}}_i)\\
&\qquad + \Psi_{i,[p_0]}\boldsymbol{\xi}_{i,[p_0]} - \widetilde{\Psi}_{i,[p]}\widetilde{\boldsymbol{\xi}}_{i,[p]}\|^2. \qquad (14)
\end{aligned}$$

By substituting in the FPCA and PACE estimators, the estimated variance under the model with p principal components is given by

$$\begin{aligned}
\widehat{\sigma}^2_{[p]} &= N^{-1}\sum_{i=1}^{n}\|\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i - \widehat{\Psi}_{i,[p]}\widehat{\boldsymbol{\xi}}_{i,[p]}\|^2\\
&= N^{-1}\sum_{i=1}^{n}\left\|\left(I - \widehat{\Omega}_{i,[p]}\widehat{\Sigma}^{-1}_{i,[p]}\right)(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\right\|^2\\
&= N^{-1}\sum_{i=1}^{n}\left\|\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Sigma}^{-1}_{i,[p]}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\right\|^2.
\end{aligned}$$

Then the Kullback–Leibler information for these estimators is

$$\Delta\left(p, \widehat{X}_{[p]}, \widehat{\sigma}^2_{[p]}\right) = N\log\left(\widehat{\sigma}^2_{[p]}\right) + \mathcal{A}_n(p), \qquad (15)$$

where $\mathcal{A}_n(p) = N\sigma_u^2/\widehat{\sigma}^2_{[p]} + \widehat{\sigma}^{-2}_{[p]}\sum_{i=1}^n\|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i + \Psi_{i,[p_0]}\boldsymbol{\xi}_{i,[p_0]} - \widehat{\Psi}_{i,[p]}\widehat{\boldsymbol{\xi}}_{i,[p]}\|^2$.

To derive the new AIC criterion, we need the following theoretical results to evaluate the expected Kullback–Leibler information. As discussed in Hurvich, Simonoff, and Tsai (1998, p. 275), in derivation of AIC, one needs to assume that the true model is included in the family of candidate models, and any model bias is ignored. For example, Hurvich, Simonoff, and Tsai (1998) ignored the smoothing bias when developing AIC for nonparametric regressions. Following the same argument, we will ignore all the biases in $\widehat{\mu}(\cdot)$ and $\widehat{\psi}_k(\cdot)$, and only take into account the variation in the estimators.

Proposition 1. Under assumptions (C.1)–(C.7), condition (12) and the additional assumption that $n(h_\mu + h_C) \to \infty$, $\widehat{\sigma}^2_{[p_0]}/\sigma_u^2 = N^{-1}\sum_{i=1}^n\sum_{j=p_0+1}^{m_i}\mathcal{X}_{ij} + \mathcal{R}_n$, where the $\mathcal{X}_{ij}$ are independent χ_1^2 random variables and $\mathcal{R}_n = O_p\{\delta_{n1}^2(h_\mu) + \delta_{n1}^2(h_C)\} + o_p(nN^{-1})$. As a result, $\widehat{\sigma}_{[p_0]} \to \sigma_u^2$ in probability as $n \to \infty$

The next proposition gives the asymptotic expansion for $\mathrm{E}\{\mathcal{A}_n(p_0)\}$.

Proposition 2. Under the same conditions as in Proposition 1, $\mathrm{E}\{\mathcal{A}_n(p_0)\} = N + 2np_0 + o(n)$.

Thus, the expected Kullback–Leibler information is $\mathrm{E}_F\{\Delta(p_0, \widehat{X}_{[p_0]}, \widehat{\sigma}^2_{[p_0]})\} = \mathrm{E}_F\{N\log(\widetilde{\sigma}^2_{[p_0]})\} + N + 2np_0 + o(n)$. This justifies defining AIC as

$$\mathrm{AIC}(p) = N\log\left(\widehat{\sigma}^2_{[p]}\right) + N + 2np. \qquad (16)$$

When $m_i \to \infty$ and p is fixed, an intuitive interpretation for the proposed AIC in (16) is to consider FPCA as a linear regression on the observed data $\boldsymbol{W}_i - \boldsymbol{\mu}_i$ against covariates $(\boldsymbol{\psi}_{i1}, \ldots, \boldsymbol{\psi}_{ip})$ for subject i, and consider the principal component scores as the subject-specific coefficients. By pooling n independent curves together and by adding up the individual AIC, we have a total of np regression parameters, and the AIC in (16) coincides with that of a simple linear regression. The biggest difference between our AIC and that of Yao et al. in (11) is the way we count the number of parameters in the model.

3.3 Consistent Information Criteria

As pointed out by a referee, FPCA is closely related to factor models in econometrics, where there are some existing information criteria to choose the number of factors consistently (Bai and Ng 2002). We stress that the data considered in the econometrics literature are multivariate time series data observed on regular time points, while we consider irregularly spaced functional data. The estimator and criteria proposed by Bai and Ng were based on matrix projections, while our FPCA method relies heavily on kernel smoothing and operator theory. As a result, deriving consistent model selection criteria for our problem is technically much more involved.

Inspired by Bai and Ng (2002), we consider two classes of information criteria:

$$\mathrm{PC}(p) = \widehat{\sigma}^2_{[p]} + pg_n, \qquad (17)$$

$$\mathrm{IC}(p) = \log\left(\widehat{\sigma}^2_{[p]}\right) + pg_n, \qquad (18)$$

where $\widehat{\sigma}^2_{[p]}$ is the error variance estimator used in our AIC (15) and g_n is a penalty. The estimator $\widehat{\sigma}^2_{[p]}$ in Bai and Ng (2002) was a mean squared error based on a simple regression, while our estimator is based on the PACE method involving kernel smoothing and BLUP.

For any $p \le p_0$, denote $\boldsymbol{\psi}_{[p]}(t) = (\psi_1, \ldots, \psi_p)^{\mathrm{T}}(t)$, $\boldsymbol{\psi}_{[p+1:p_0]} = (\psi_{p+1}, \ldots, \psi_{p_0})^{\mathrm{T}}(t)$, and define the inner product matrices $\mathcal{J}_{1,p} = \int\boldsymbol{\psi}_{[p]}(t)\boldsymbol{\psi}^{\mathrm{T}}_{[p]}(t)f_1(t)dt$, $\mathcal{J}_{2,p} = \int\boldsymbol{\psi}_{[p+1:p_0]}(t)\,\boldsymbol{\psi}^{\mathrm{T}}_{[p+1:p_0]}(t)\ f_1(t)dt$, and $\mathcal{J}_{12,p} = \int\boldsymbol{\psi}_{[p]}(t)\boldsymbol{\psi}^{\mathrm{T}}_{[p+1:p_0]}(t)f_1(t)dt$. Put $\Lambda_{[p+1:p_0]} = \mathrm{diag}(\omega_{p+1}, \ldots, \omega_{p_0})$, and

$$\tau_p = \mathrm{tr}\left\{\left(\mathcal{J}_{2,p} - \mathcal{J}^{\mathrm{T}}_{12,p}\mathcal{J}^{-1}_{1,p}\mathcal{J}_{12,p}\right)\Lambda_{[p+1:p_0]}\right\}. \qquad (19)$$

Theorem 2. Suppose τ_p defined in (19) exists and is positive for all $0 \le p < p_0$. Let $\widehat{p}$ be the minimizer of the information criteria defined in (17) or (18) among $0 \le p \le p_{\max}$ with $p_{\max} > p_0$ being a fixed search limit, and define $\varrho_n = h_\mu^2 + h_C^2 + h_\sigma^2 + \delta_{n1}(h_\mu) + \delta_{n2}(h_C) + \delta_{n1}^2(h_\sigma)$. Under the assumptions (C.1)–(C.7) and condition (12), $\lim_{n\to\infty} pr(\widehat{p} = p_0) = 1$ if the penalty function g_n satisfies (i) $g_n \xrightarrow{p} 0$ and (ii) $g_n/(n/N + \varrho_n^2) \xrightarrow{p} \infty$.

In the factor analysis context, the penalty term in the information criteria proposed by Bai and Ng (2002) converges to 0 with a rate slower than C_n^{-2}, where $C_n = \min(m^{1/2}, n^{1/2})$ translating to our notation. Their rate shows a sense of symmetry in the roles of m and n. Indeed, when the curves are observed on a regular grid, the data can be arranged into an $n \times m$ matrix $\boldsymbol{W}$, the factor analysis can be carried out by a singular value decomposition of $\boldsymbol{W}$, and hence the roles of m and n are symmetric. For the random design that we consider, we apply nonparametric smoothing along t, not among the subjects. Therefore, m and n play different roles in our rate. Not only does the smoothing make our derivation much more involved, but also the fact that the within-subject covariance matrices are defined on subject-specific time points poses many theoretical challenges. Our proof uses many techniques from perturbation theory of random operators and matrices.

The following corollary shows that when the bandwidths are chosen properly, penalties similar to those in Bai and Ng (2002) can still lead to consistent information criteria.

Corollary 1. Suppose all conditions in Theorem 2 hold, and $h_\mu \asymp \max(n,m)^{-c_1}$, $h_C \asymp \max(n,m)^{-c_2}$, $h_\sigma \asymp \max(n,m)^{-c_3}$, where $1/4 \le c_1, c_2 \le 1$, $1/4 \le c_3 \le 3/2$. Then $\widehat{p}$ that minimizes PC(p) or IC(p) is consistent if (i) $g_n \xrightarrow{p} 0$ and (ii) $C_n^2 g_n \xrightarrow{p} \infty$, where $C_n = \min(n^{1/2}, m^{1/2})$ as defined in Bai and Ng (2002).

Bai and Ng (2002) proposed the following information criteria that satisfy the conditions in Corollary 1,

$$\begin{aligned}
\mathrm{PC}_{p1}(p) &= \widehat{\sigma}^2_{[p]} + p\widehat{\sigma}^2_{\mathrm{pilot}}\left(\frac{n+m}{nm}\right)\log\left(\frac{nm}{n+m}\right),\\
\mathrm{PC}_{p2}(p) &= \widehat{\sigma}^2_{[p]} + p\widehat{\sigma}^2_{\mathrm{pilot}}\left(\frac{n+m}{nm}\right)\log\left(C_n^2\right),\\
\mathrm{PC}_{p3}(p) &= \widehat{\sigma}^2_{[p]} + p\widehat{\sigma}^2_{\mathrm{pilot}}\left\{\frac{\log\left(C_n^2\right)}{C_n^2}\right\},\\
\mathrm{IC}_{p1}(p) &= \log\left(\widehat{\sigma}^2_{[p]}\right) + p\left(\frac{n+m}{nm}\right)\log\left(\frac{nm}{n+m}\right),\\
\mathrm{IC}_{p2}(p) &= \log\left(\widehat{\sigma}^2_{[p]}\right) + p\left(\frac{n+m}{nm}\right)\log\left(C_n^2\right),\\
\mathrm{IC}_{p3}(p) &= \log\left(\widehat{\sigma}^2_{[p]}\right) + p\left\{\frac{\log\left(C_n^2\right)}{C_n^2}\right\}, \qquad (20)
\end{aligned}$$

where $\widehat{\sigma}^2_{\mathrm{pilot}}$ is a pilot estimator for σ_u^2. In our setting, we can use $\widetilde{\sigma}^2_{u,\mathrm{I}}$ defined at (3) in place of $\widehat{\sigma}^2_{\mathrm{pilot}}$, and replace m by either the arithmetic or the harmonic mean of m_i's. Under the undersmoothing choices of bandwidths described in Corollary 1, all information criteria in (20) are consistent. One can easily see the similarity between the IC$_p$ criteria and the AIC proposed in (16). In general, the IC$_p$ criteria impose greater penalties to overfitting than AIC. By comparing AIC with the conditions in Theorem 2 and other consistent criteria we developed, we can see that the penalty term in AIC is a little bit small and that explains the nonvanishing chance of overfitting witnessed in our simulation studies, see Section 4.

4. SIMULATION STUDIES

4.1 Empirical Performance of the Proposed Criteria

To illustrate the finite sample performance of the proposed methods, we performed various simulation studies. Let $\mathcal{T} = [0, 1]$, and suppose that the data are generated from the models (1) and (2). Let the observation time points $T_{ij} \sim$ Uniform [0, 1], $m_i = m$ for all i and $U_{ij} \sim \mathrm{Normal}(0, \sigma_u^2)$.

We consider the following five scenarios.

Scenario I. Here, the true mean function is $\mu(t) = 5(t - 0.6)^2$, the number of principal components is $p_0 = 3$, the true eigenvalues are $(\omega_1, \omega_2, \omega_3) = (0.6, 0.3, 0.1)$, the variance of the error is $\sigma_u^2 = 0.2$, and the eigenfunctions are $\psi_1(t) = 1$, $\psi_2(t) = \sqrt{2}\sin(2\pi t)$, $\psi_3(t) = \sqrt{2}\cos(2\pi t)$. The principal component scores are generated from independent normal distributions, that is, $\xi_{ij} \sim \mathrm{Normal}(0, \omega_j)$. Here $\omega_3 < \sigma_u^2$.

Scenario II. The data are generated in the same way as in Scenario I, except that we replace the third eigenfunction by a rougher function $\psi_3'(t) = \sqrt{2}\cos(4\pi t)$ so that the covariance function is less smooth, and we let the principal component scores follow a skewed Gaussian mixture model. Specifically, ξ_{ij} has $1/3$ probability of following a Normal($2\sqrt{\omega_j/3}, \omega_j/3$) distribution, and $2/3$ probability of following Normal($-\sqrt{\omega_j/3}, \omega_j$), for $j = 1, 2, 3$.

Scenario III. Set $\mu(t) = 12.5(t - 0.5)^2 - 1.25$, $\phi_1(t) = 1$, $\phi_2(t) = \sqrt{2}\cos(2\pi t)$, $\phi_3(t) = \sqrt{2}\sin(4\pi t)$, and $(\omega_1, \omega_2, \omega_3, \sigma^2) = (4.0, 2.0, 1.0, 0.5)$. The principal component scores are generated from a Gaussian distribution. Here $\omega_3 > \sigma_u^2$.

Scenario IV. The mean function, eigenvalues, eigenfunction, and noise level are set to be the same as in Scenario III, but the ξ_{ij}'s are generated from a Gaussian mixture model similar to that in Scenario II.

Scenario V. In this simulation, we set $p_0 = 6$, the true eigenvalues are (4.0, 3.5, 3.0, 2.5, 2.0, 1.5) and $\sigma_u^2 = 0.5$. We assume that the principal component scores are normal random variables and let the eigenfunctions be

$$\begin{aligned}
&\psi_1(t) = 1; \quad \psi_{2k}(t) = \sqrt{2}\sin(2k\pi t), \quad \text{for } k = 1, 2, 3;\\
&\psi_{2k+1}(t) = \sqrt{2}\cos(2k\pi t), \quad \text{for } k = 1, 2.
\end{aligned}$$

In each simulation, we generated $n = 200$ trajectories from the models above, and compared the cases with $m = 5$, 10, and 50. The cases $m = 5$ and $m = 50$ may be viewed as representing sparse and dense functional data, respectively, whereas $m = 10$ represents scenarios between the two extremes. For each m, we apply the FPCA procedure to estimate $\{\mu(\cdot), R(\cdot,\cdot), \omega_k, \psi_k(\cdot), \sigma_w^2(t)\}$, then use the proposed information criteria to choose p. The simulation was then repeated 200 times for each scenario.

The performance of the estimators depends on the choice of bandwidths for $\mu(t)$, $C(s, t)$, and $\sigma_w^2(t)$, and the optimal bandwidths vary with n and m. We picked the bandwidths that are slightly smaller than those minimizing the integrated mean squared error of the corresponding functions, since undersmoothing in FPCA was also advocated by Hall, Müller, and Wang (2006) and Li and Hsing (2010b).

We consider Yao's AIC, MDL by Poskitt and Sengarapillai (2013), the proposed BIC and AIC in (6) and(16), and the criteria by Bai and Ng in (20). Yao's AIC is calculated using the publicly available PACE package (*http://anson.ucdavis.edu/mueller/data/pace.html*), where all bandwidths are data driven and selected by generalized cross-validation (GCV). The empirical distribution of $\widehat{p}$ under Scenarios I–IV is summarized in Tables 1–3. Since the true number of principal components p_0 is different in Scenario V, the distribution of $\widehat{p}$ is summarized in a separate Table 4.

The proposed BIC method is based on the convergence rate results on the eigenvalues, and does not rely much on the distributional assumptions for X and U. From Tables 1–3, we see that BIC picks the correct number of principal components with high percentage in almost all scenarios, except for the cases where the data are sparse, that is, $m = 5$. This phenomena is as expected, because it is harder to pick up the correct number of signals from sparse and noisy data.

Table 1. When $m = 5$, displayed are the distributions of the number of selected principal components $\widehat{p}$ for all methods and across Scenarios I–IV. The true number of principal components is 3

Scenario	Method	$\widehat{p} \le 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \ge 5$
I	AIC$_{\text{PACE}}$	0.000	0.008	0.000	0.121	0.870
	AIC	0.000	0.405	0.580	0.010	0.005
	BIC	0.155	0.335	0.380	0.115	0.015
	PC$_{p1}$	0.005	0.565	0.410	0.010	0.010
	IC$_{p1}$	0.000	0.215	0.735	0.045	0.005
II	AIC$_{\text{PACE}}$	0.000	0.000	0.005	0.125	0.870
	AIC	0.000	0.205	0.630	0.155	0.010
	BIC	0.230	0.395	0.245	0.110	0.020
	PC$_{p1}$	0.000	0.000	0.375	0.440	0.185
	IC$_{p1}$	0.000	0.140	0.605	0.210	0.045
III	AIC$_{\text{PACE}}$	0.000	0.025	0.005	0.130	0.840
	AIC	0.000	0.035	0.720	0.170	0.075
	BIC	0.335	0.260	0.325	0.080	0.000
	PC$_{p1}$	0.000	0.220	0.640	0.075	0.065
	IC$_{p1}$	0.000	0.005	0.590	0.280	0.125
IV	AIC$_{\text{PACE}}$	0.000	0.015	0.015	0.145	0.825
	AIC	0.000	0.020	0.710	0.185	0.085
	BIC	0.315	0.180	0.410	0.070	0.025
	PC$_{p1}$	0.000	0.160	0.640	0.095	0.105
	IC$_{p1}$	0.000	0.015	0.560	0.260	0.165

Table 3. For $m = 50$, displayed are the distributions of the number of selected principal components $\widehat{p}$ for all methods and across Scenarios I–IV. The true number of principal components is 3

Scenario	Method	$\widehat{p} = 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \ge 5$
I	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.830	0.150	0.020
	PC$_{p1}$	0.000	0.000	1.000	0.000	0.000
	IC$_{p1}$	0.000	0.000	1.000	0.000	0.000
II	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.630	0.320	0.050
	BIC	0.000	0.000	0.795	0.185	0.020
	PC$_{p1}$	0.000	0.000	0.955	0.045	0.000
	IC$_{p1}$	0.000	0.000	0.945	0.055	0.000
III	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.775	0.200	0.025
	PC$_{p1}$	0.000	0.000	1.000	0.000	0.000
	IC$_{p1}$	0.000	0.000	1.000	0.000	0.000
IV	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.945	0.055	0.000
	BIC	0.000	0.000	0.835	0.140	0.025
	PC$_{p1}$	0.000	0.000	1.000	0.000	0.000
	IC$_{p1}$	0.000	0.000	1.000	0.000	0.000

Compared to BIC, the performance of the proposed AIC method is even more impressive. Although BIC is designed to be a consistent model selector, the AIC method selects the right number of principal component with a higher percentage in most of the cases we considered. This is partially because AIC makes more use of the information from the likelihood. Even though the data are non-Gaussian in Scenario II and IV, the AIC still performs better than the BIC, and it shows that both the PACE method and the AIC method are quite robust against mild violation of the Gaussian assumption. Even though the motivation and theoretical development for the AIC method described in Section 3.2 are for dense functional data, it performs surprisingly well for sparse data, such as the case $m = 5$.

There are six criteria in (20), and we find that the PC$_p$'s and the IC$_p$'s tend to perform similarly. To save journal space, we only provide the results for PC$_{p1}$ and IC$_{p1}$, and the results for the remaining criteria in (20) can be find in the expanded versions of Tables 1–4 in the supplementary material. As we can see,

Table 2. When $m = 10$, displayed are the distributions of the number of selected principal components $\widehat{p}$ for all methods and across Scenarios I–IV. The true number of principal components is 3

Scenario	Method	$\widehat{p} \le 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \ge 5$
I	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.005	0.980	0.015	0.000
	BIC	0.000	0.040	0.670	0.255	0.035
	PC$_{p1}$	0.000	0.040	0.955	0.000	0.005
	IC$_{p1}$	0.000	0.005	0.985	0.010	0.000
II	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.005	0.995
	AIC	0.000	0.000	0.710	0.260	0.030
	BIC	0.000	0.170	0.665	0.135	0.030
	PC$_{p1}$	0.000	0.000	0.570	0.355	0.075
	IC$_{p1}$	0.000	0.000	0.805	0.185	0.010
III	AIC$_{\text{PACE}}$	0.000	0.015	0.000	0.000	0.985
	AIC	0.000	0.000	0.580	0.400	0.020
	BIC	0.005	0.035	0.770	0.145	0.045
	PC$_{p1}$	0.000	0.000	0.965	0.030	0.005
	IC$_{p1}$	0.000	0.000	0.665	0.320	0.015
IV	AIC$_{\text{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.830	0.150	0.020
	BIC	0.010	0.005	0.775	0.190	0.020
	PC$_{p1}$	0.000	0.000	0.920	0.045	0.035
	IC$_{p1}$	0.000	0.000	0.900	0.085	0.015

Table 4. Distributions of the number of selected principal components $\widehat{p}$ for Scenario V. The true number of principal components is 6

Scenario	Method	$\widehat{p} \le 4$	$\widehat{p} = 5$	$\widehat{p} = 6$	$\widehat{p} = 7$	$\widehat{p} \ge 8$
$m = 5$	AIC$_{\text{PACE}}$	0.005	0.005	0.705	0.245	0.040
	AIC	0.165	0.330	0.470	0.035	0.000
	BIC	0.835	0.020	0.090	0.050	0.005
	PC$_{p1}$	0.580	0.345	0.070	0.005	0.000
	IC$_{p1}$	0.060	0.335	0.545	0.060	0.000
$m = 10$	AIC$_{\text{PACE}}$	0.005	0.000	0.065	0.475	0.455
	AIC	0.000	0.000	0.570	0.280	0.15
	BIC	0.250	0.030	0.525	0.165	0.030
	PC$_{p1}$	0.000	0.145	0.775	0.020	0.060
	IC$_{p1}$	0.000	0.000	0.705	0.185	0.110
$m = 50$	AIC$_{\text{PACE}}$	0.000	0.065	0.000	0.000	0.935
	AIC	0.000	0.000	0.260	0.405	0.335
	BIC	0.005	0.000	0.590	0.325	0.080
	PC$_{p1}$	0.000	0.000	0.980	0.010	0.010
	IC$_{p1}$	0.000	0.000	0.965	0.035	0.000

these criteria behave similar to the AIC, and they tend to do better only in a few occasions when AIC overestimates p.

For almost all scenarios considered, Yao's AIC hardly ever picks the correct model, with the exception of Scenario V, $m = 5$, which will be discussed in more detail below. When the true number of principal components is 3, Yao's AIC will normally chose a number greater than 5. This phenomenon becomes more severe when the data are dense. For example, when $m = 50$, Yao's AIC almost always pick the maximum order considered, which is 15 in our simulations. The behavior of the MDL by Poskitt and Sengarapillai (2013) is similar to Yao's AIC, and hence these results are only provided in Tables S.2– S.5 in the supplementary material.

Scenario V, Table 4 is specially designed to check the performance of the proposed information criteria under the situations where we have a relatively large number of principal components. The proposed criteria worked reasonably well for $m = 10$ and 50, and performed much better than Yao's AIC. The case of $m = 5$ under Scenario V is the only case in all of our simulations that Yao's AIC picks the correct model more often than our criteria. With a closer look at the results, we find an explanation. The true covariance function under Scenario V is quite rough, and the GCV criterion in the PACE package chose a large bandwidth so that the local fluctuations on the true covariance surface are smoothed out. In other words, high-frequency signals are smoothed out and treated as noise. In a typical run, the PACE estimates for the eigenvalues are (4.1736, 2.1350, 1.6697, 1.0009, 0.3978, 0.0476), which are far from the truth, (4.0, 3.5, 3.0, 2.5, 2.0, 1.5), and the estimated error variance is 6.519 in contrast to the truth $\sigma_u^2 = 0.5$. It is the combination of seriously underestimating the high-order eigenvalues and small penalty in AIC that makes Yao's criterion pick the correct number of principal components. Switching to our undersmoothing bandwidths, these estimates are improved but then Yao's AIC will choose much larger values for p. This case also highlights the difficulty of FPCA when p is large but the data are sparse. Unless we have a very large sample size, estimation of these principal components is very difficult, and comparing the model selection procedures in such a case would not be meaningful.

4.2 Further Simulations

The supplementary material, Section S.4 contains further simulations, including (a) expanded results with other model selectors in Tables S.2– S.5; (b) an examination of the sensitivity of the results to the bandwidth (supplementary Table S.6); (c) the behavior of BIC with much larger sample size (supplementary Table S.7); and (d) results when the value of m is not constant, that is, $m_i \neq m$ for all i (supplementary Table S.8).

5. DATA ANALYSIS

The colon carcinogenesis data in our study have been analyzed in Li, Wang, and Carroll (2010), Li et al. (2007), and Baladandayuthapani et al. (2008). The biomarker of interest in this experiment is *p27*, which is a protein that inhibits cell cycle. We have 12 rats injected with carcinogen and sacrificed 24 hr after the injection. Beneath the colon tissue of the rats, there are pore structures called "colonic crypts." A crypt typically contains 25–30 cells, lined up from the bottom to the top. The stem cells are at the bottom of the crypt, where daughter cells are generated. These daughter cells move toward the top as they mature. We sampled about 20 crypts from each of the 12 rats. The *p27* expression level was measured for each cell within the sampled crypts. As previously noted in the literature (Morris et al. 2001, 2003), the *p27* measurements, indexed by the relative cell location within the crypt, are natural functional data. We have $m = 25$–30 observations (cells) on each function. As in the previous analyses, we consider *p27* in the logarithmic scale. By pooling data from the 12 rats, we have a total of $n = 249$ crypts (functions). In the literature, it has been noted that there is spatial correlation among the crypts within the same rat (Li et al. 2007; Baladandayuthapani et al. 2008). In this experiment, we sampled crypts sufficiently far apart so that the spatial correlations are negligible, and thus we can assume that the crypts are independent.

We perform the FPCA procedure as described in Section 2, with the bandwidths chosen by leave one curve out cross-validation. The estimated covariance function is given in the top panel of Figure 1. The estimated variance of measurement error by integration is $\widetilde{\sigma}_{u,\mathrm{I}} = 0.103$. In contrast, the top three eigenvalues are 0.8711, 0.0197, and 0.0053. Let $k_n = \max\{k;\ \widehat{\omega}_k > 0\}$, then the percentage of variation explained by the kth principal component is estimated by $\widehat{\omega}_k/(\sum_{j=1}^{k_n}\widehat{\omega}_j)$. The percentage of variation explained by the first seven principal components is (0.966, 0.022, 0.006, 0.003, 0.002, 0.001, 0.000).

We apply the proposed AIC, adaptive BIC, the Bai and Ng criteria (20), and Yao's AIC to the data. All of the proposed methods lead to $p = 3$ principal components, for which the corresponding eigenfunctions are shown in the middle panel of Figure 1. As we can see, the first principal component is a constant over time, and the second and third eigenfunctions are essentially linear and quadratic functions. Eigenfunctions 4–7 are shown in the bottom panel of Figure 1, and they are basically noises and are hard to interpret. We therefore can see that the variation among different crypts can be explained by random quadratic polynomials. Yao's AIC, on the other hand, picked a much large number of principal components, with $p = 9$. This is because a much smaller penalty is used in Yao's AIC criterion. We have repeated the data analysis using other choices of bandwidths, and the results are the same.

6. SUMMARY

6.1 Basic Summary

Choosing the number of principal components is a crucial step in FDA. There have been some data-driven procedures proposed in the literature that can be used to choose the number of principal components, but these procedures have not been studied theoretically nor were they tested numerically as extensively as in this article.

To promote practically useful model selection criteria, we have assumed that there exists a finite-dimensional true model. We found that the consistency of the model selection criteria depends on both the sample size n and the number of repeated measurements m on each curve. We proposed a marginal BIC criterion that is consistent for both dense and sparse functional data, which means m can be of any rate relative to n. In the

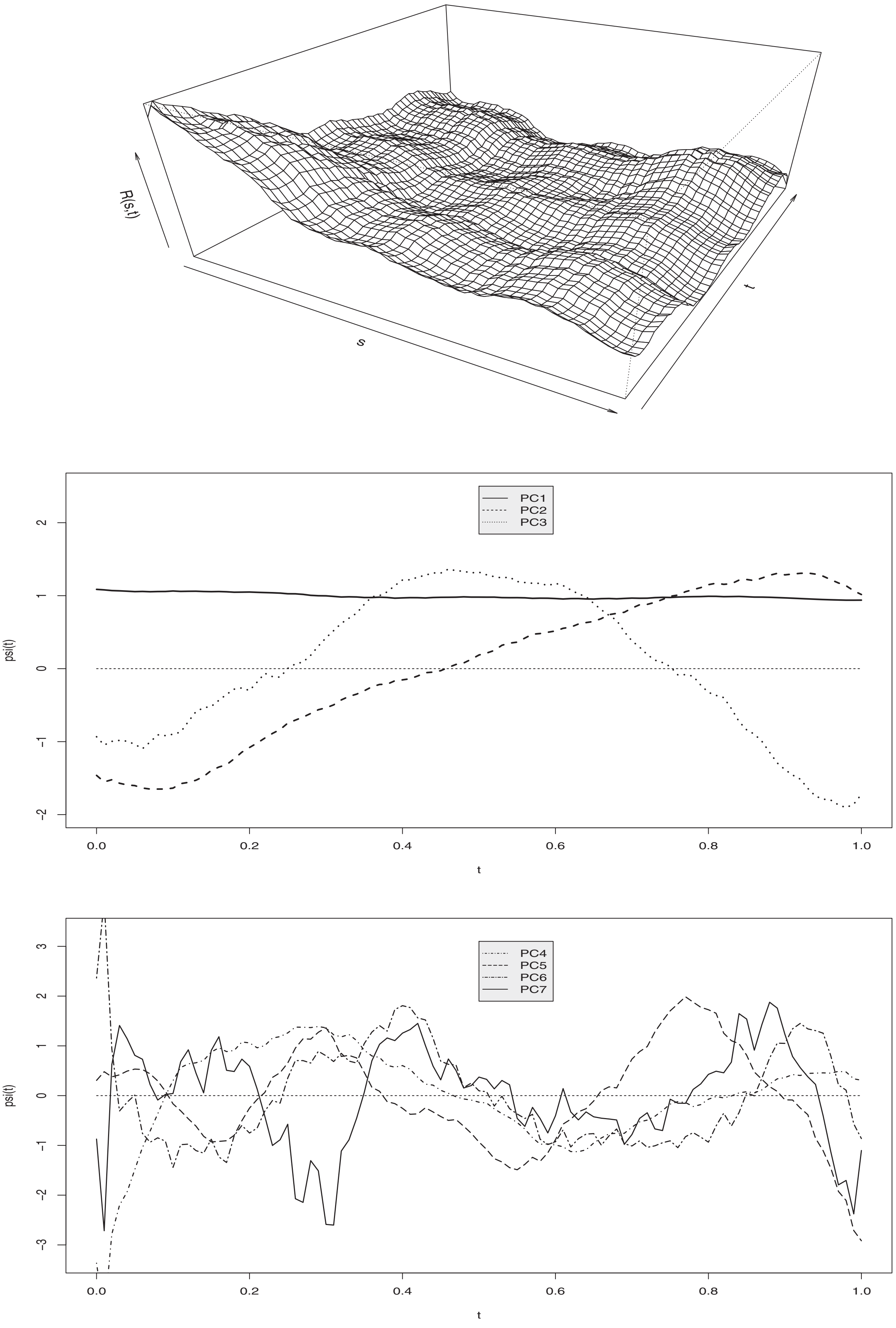

Figure 1. Functional principal component analysis for the colon carcinogenesis *p27* data. Top panel: estimated covariance function; middle panel: the first three eigenfunctions; and lower panel: eigenfunctions 4–7.

framework of dense functional data, where both n and m diverge to infinity, we proposed a conditional AIC, which is motivated by an asymptotic study of the expected Kullback–Leibler distance under Gaussian assumption.

Following the standard approach of Hurvich, Simonoff, and Tsai (1998), we ignored smoothing biases in developing AIC. Our intensive simulation studies also confirm that bias plays a very small role in model selection. In our simulations in Section 4.2, we tried a wide range of bandwidths and thus increase or decrease the biases in the estimators, but the performance of AIC is almost the same. Intuitively, the models under different numbers of principal components are nested, for a fixed bandwidth the smoothing bias exists in all models that we compare, and therefore variation is a more decisive factor in model selection.

In view of the connection of FPCA with factor analysis in multivariate time series data, we revisited the information criteria proposed by Bai and Ng (2002). Even though our setting is fundamentally different, since we assumed that the observational times are random, and the FPCA estimators depend heavily on nonparametric smoothing and are much more complex than those in Bai and Ng, we show essentially similar information criteria can be constructed. Using perturbation theory of random operators and matrices, and under an undersmoothing scheme prescribed in Section 3.3, we showed that these information criteria are consistent when both n and m go to infinity.

6.2 Discussion of the Case $p_0 \to \infty$

Some processes considered as functional data are intrinsically infinite dimensional. In those cases, the assumption of p_0 being finite is a finite sample approximation. As the sample size n increases, we can afford to include more principal components in the model and data analysis. It is helpful to consider that the true dimension p_{0n} increases to infinity as a function of n. This setting was considered in the estimation of a functional linear model (Cai and Hall 2006). To the best of our knowledge, no information criteria have been previously studied under this setting.

While allowing $p_{0n} \to \infty$, the convergence rates for $\widehat{\mu}(t)$ and $\widehat{R}(s,t)$ remain the same as those given in Lemma S.1.1 in the supplementary material, but the convergence rates for $\widehat{\psi}_j(t)$ are affected by the spacing of the true eigenvalues. Following condition (4.2) in Cai and Hall (2006), we assume that for some positive constants C and α,

$$C^{-1} j^{-\alpha} \le \omega_j \le C j^{-\alpha}, \quad \omega_j - \omega_{j+1} \ge C^{-1} j^{-1-\alpha}, \\ j = 1, \ldots, p_{0n}. \qquad (21)$$

To ensure that $\sum_j^{p_{0n}} \omega_j < \infty$, we assume that $\alpha > 1$. Define the distances between the eigenvalues, $\delta_j = \min_{k \le j}(\omega_k - \omega_{k+1})$, which is no less than $C^{-1} j^{-1-\alpha}$ under condition (21). By the asymptotic expansion of $\widehat{\psi}_j(t)$, see (2.8) in Hall and Hosseini-Nasab (2006), one can show that the convergence rate of $\widehat{\psi}_j$ is δ_j^{-1} times those in Lemma S.1.2 in the supplementary material, that is,

$$\widehat{\psi}_j(t) - \psi_j(t) = O_p\big[j^{\alpha+1} \times \{h_\mu^2 + \delta_{n1}(h_\mu) + h_C^2 + \delta_{n1}(h_C) \\ + \delta_{n2}^2(h_C)\}\big], \quad j = 1, \ldots, p_{0n}.$$

Assume that $n, m, p_{0n} \to \infty$, $p_{0n}^{\alpha+1}\varrho_n \to 0$, and $p_{0n}^{\alpha+3}/\min(n,m) \to 0$. Following the proof of Theorem 2, while taking into account the increasing estimation error in $\widehat{\psi}_j(t)$ as j increases and the increasing dimensionality of the design matrix Ψ_i, we can show that

$$\widehat{\sigma}^2_{[p]} = \begin{cases} \sigma_u^2 + \tau_p + O_p(pm^{-1} + N^{-1/2}) & \\ \quad + o_p\big(\tau_p + p^{\alpha+3}\varrho_n^2\big), & \text{for } p < p_{0n}; \\ \widehat{\sigma}^2_{[p_{0n}]} + O_p\big(m^{-1} + p_{0n}^{\alpha+3}\varrho_n^2\big), & \text{for } p \ge p_{0n}, \end{cases} \qquad (22)$$

where $\tau_p \asymp \mathrm{tr}(\Lambda_{[p+1:p_{0n}]})$ is analogous to (19) and ϱ_n is as defined in Theorem 2. Since the eigenvalues are decaying to 0, the size of the signal $\tau_p \asymp p^{-\alpha}$ as p increases to p_{0n}. To have some hope of choosing p_{0n} correctly, we need τ_p to be greater than the size of the estimation error, which implies that $p_{0n}^{2\alpha+3}\varrho_n^2 \to 0$.

Now, consider the class of information criteria in Section 3.3. Suppose that p_{0n} increases slowly enough so that $p_{0n}^{2\alpha+3}/\min(n,m) \to 0$, and that the penalty term satisfies $\tau_p/(pg_n) \to \infty$ for $p < p_{0n}$ and $pg_n/(m^{-1} + p^{\alpha+3}\varrho_n^2) \to \infty$ for $p > p_{0n}$. Then we can show that the $\widehat{p}$ that minimizes PC(p) or IC(p) is consistent. These conditions translate to

$$p_{0n}^{\alpha+1} g_n \to 0, \quad g_n/\big(p_{0n}^{-1}m^{-1} + p_{0n}^{\alpha+2}\varrho_n^2\big) \to \infty. \qquad (23)$$

If $p_{0n} = \{\min(m,n)\}^\beta$, where $0 < \beta < 1/(2\alpha+3)$, one can see that the criteria in (20) do not satisfy the conditions in (23) automatically and hence are not guaranteed to be consistent. An information criterion satisfying condition (23) requires a priori knowledge of the decay rate of the eigenvalues. Developing a data-adaptive information criterion that does not require such a priori knowledge is a challenging topic for future research.

7. SUPPLEMENTARY MATERIALS

The online supplementary material contains the technical proofs and additional simulation results.

[Received December 2011. Revised December 2012.]

REFERENCES

Bai, J., and Ng, S. (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrika*, 70, 191–221. [1285,1288,1289,1293]

Baladandayuthapani, V., Mallick, B., Hong, M., Lupton, J., Turner, N., and Carroll, R. J. (2008), "Bayesian Hierarchical Spatially Correlated Functional Data Analysis With Application to Colon Carcinogenesis," *Biometrics*, 64, 64–73. [1291]

Cai, T., and Hall, P. (2006), "Prediction in Functional Linear Regression," *The Annals of Statistics*, 34, 2159–2179. [1284,1293]

Capra, W. B., and Müller, H. G. (1997), "An Accelerated-Time Model for Response Curves," *Journal of the American Statistical Association*, 92, 72–83. [1286]

Claeskens, G., and Hjort, N. L. (2008), *Model Selection and Model Averaging*, New York: Cambridge University Press. [1287]

Hall, P., and Hosseini-Nasab, M. (2006), "On Properties of Functional Principal Components Analysis," *Journal of the Royal Statistical Society,* Series B, 68, 109–126. [1284,1293]

Hall, P., Müller, H. -G., and Wang, J. -L. (2006), "Properties of Principal Component Methods for Functional and Longitudinal Data Analysis," *The Annals of Statistics*, 34, 1493–1517. [1284,1286,1289]

Hall, P., and Vial, C. (2006), "Assessing the Finite Dimensionality of Functional Data," *Journal of the Royal Statistical Society,* Series B, 68, 689–705. [1285]

Hurvich, C. M., Simonoff, J. S., and Tsai, C. L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society,* Series B, 60, 271–293. [1287,1288,1293]

Hurvich, C. M., and Tsai, C. L. (1989), "Regression of Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307. [1287]

James, G., and Hastie, T. (2001), "Functional Linear Discriminant Analysis for Irregularly Sampled Curves," *Journal of the Royal Statistical Society, Series B*, 63, 533–550. [1284]

Li, Y., and Hsing, T. (2010a), "Deciding the Dimension of Effective Dimension Reduction Space for Functional and High-Dimensional Data," *The Annals of Statistics*, 38, 3028–3062. [1284]

——— (2010b), "Uniform Convergence Rates for Nonparametric Regression and Principal Component Analysis in Functional/Longitudinal Data," *The Annals of Statistics*, 38, 3321–3351. [1285,1289]

Li, Y., Wang, N., and Carroll, R. J. (2010), "Generalized Functional Linear Models With Semiparametric Single-Index Interactions," *Journal of the American Statistical Association*, 105, 621–633. [1284,1291]

Li, Y., Wang, N., Hong, M., Turner, N., Lupton, J., and Carroll. R. J. (2007), "Nonparametric Estimation of Correlation Functions in Spatial and Longitudinal Data, With Application to Colon Carcinogenesis Experiments," *The Annals of Statistics*, 35, 1608–1643. [1291]

Morris, J. S., Vannucci, M., Brown, P. J., and Carroll, R. J. (2003), "Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis," *Journal of the American Statistical Association*, 98, 573–583. [1291]

Morris, J. S., Wang, N., Lupton, J. R., Chapkin, R. S., Turner, N. D., Hong, M. Y., and Carroll, R. J. (2001), "Parametric and Nonparametric Methods for Understanding the Relationship Between Carcinogen-Induced DNA Adduct Levels in Distal and Proximal Regions of the Colon," *Journal of the American Statistical Association*, 96, 816–826. [1291]

Müller, H.-G., and Stadtmüller, U. (2005), "Generalized Functional Linear Models," *The Annals of Statistics*, 33, 774–805. [1284]

Poskitt, D. S., and Sengarapillai, A. (2013), "Description Length and Dimensionality Reduction in Functional Data Analysis," *Computational Statistics & Data Analysis*, 58, 98–113. [1285,1289,1291]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer-Verlag. [1284]

Rice, J., and Silverman, B. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves," *Journal of the Royal Statistical Society,* Series B, 53, 233–243. [1284,1286]

Yao, F., Müller, H. G., and Wang, J. L. (2005a), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [1284,1285,1287]

——— (2005b), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903. [1284]

Zhou, L., Huang, J. Z., and Carroll, R. J. (2008), "Joint Modelling of Paired Sparse Functional Data Using Principal Components," *Biometrika*, 95, 601–619. [1284]