*J. R. Statist. Soc.* B (2006)
**68**, *Part* 1, *pp.* 3–25

# Penalized spline models for functional principal component analysis

Fang Yao and Thomas C. M. Lee

*Colorado State University, Fort Collins, USA*

[Received November 2004. Final revision August 2005]

**Summary.** We propose an iterative estimation procedure for performing functional principal component analysis. The procedure aims at functional or longitudinal data where the repeated measurements from the same subject are correlated. An increasingly popular smoothing approach, penalized spline regression, is used to represent the mean function. This allows straightforward incorporation of covariates and simple implementation of approximate inference procedures for coefficients. For the handling of the within-subject correlation, we develop an iterative procedure which reduces the dependence between the repeated measurements that are made for the same subject. The resulting data after iteration are theoretically shown to be asymptotically equivalent (in probability) to a set of independent data. This suggests that the general theory of penalized spline regression that has been developed for independent data can also be applied to functional data. The effectiveness of the proposed procedure is demonstrated via a simulation study and an application to yeast cell cycle gene expression data.

*Keywords*: Asymptotics; Functional data; Penalized spline regression; Principal components; Smoothing; Within-subject correlation

## 1. Introduction

Advances in modern technology, including computational genomics, have facilitated the collection and analysis of high dimensional data, or data that are repeatedly measured for the same subject or cluster. When the observed data are in the forms of random curves, rather than scalars or vectors, dimension reduction is necessary. Therefore functional principal component analysis has become a useful tool, as it achieves this by reducing random trajectories to a set of functional principal component scores. Besides dimension reduction, functional principal component analysis attempts to characterize the dominant modes of variation of a sample of random trajectories around their mean trend(s). There is an extensive literature on functional principal component analysis. Rao (1958) introduced the method for growth curves, and earlier work includes Besse and Ramsay (1986), Castro *et al.* (1986) and Berkey *et al.* (1991). Since then there has emerged a central tool of functional data analysis; for examples, see Rice and Silverman (1991), Jones and Rice (1992), Silverman (1996), Brumback and Rice (1998), Boente and Fraiman (2000) and Fan and Zhang (2000), among others. For an introduction and summary, see Ramsay and Silverman (1997).

In this paper a new iterative procedure for fitting functional principal component models is proposed. Attractive properties of this new procedure include that it addresses the within-subject (cluster) correlation in functional or longitudinal data. The main idea is, via an iterative process, to transform the original correlated data such that the resulting data are asymptotically

*Address for correspondence*: Fang Yao, Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.
E-mail: fyao@stat.colostate.edu

 1369–7412/06/68003

equivalent to a set of independent data. During the iteration process, the mean function is updated with a popular smoothing technique, penalized spline regression, whereas the covariance surface, the variance of errors and the functional principal components that are described in model (1) below are estimated by the local polynomial method of Yao *et al.* (2003). The use of penalized splines provides an easy and straightforward way to incorporate covariates and to make inference for covariate effects, and the coupling with the method of Yao *et al.* (2003) facilitates a theoretical study of the asymptotic properties of the overall proposed procedure. We term the procedure iterative penalized spline (IPS) fitting.

Through an analytic derivation of its asymptotic properties, IPS fitting is shown to provide a sample of transformed data which are asymptotically equivalent (in probability) to a set of independent data. Therefore the theory of penalized spline regression that has been developed for independent data can be applied to the transformed data and uniform consistency of the mean estimate as well as other model components is obtained as a consequence. To the best of our knowledge, no asymptotic consistency results of penalized spline models for functional data are available so far, whereas kernel and smoothing spline approaches for clustered data have been investigated by Lin and Carroll (2000), Wang (2003), Lin *et al.* (2004) and Wang *et al.* (2005). In most of these existing approaches, the covariance structure is modelled through a finite number of parameters by using moment methods, which inherits the feature of covariance estimation in classical longitudinal approaches. In contrast we use a nonparametric smoothing approach to model the covariance surface without assuming any parametric form, which makes the theoretical development more challenging. The empirical properties of IPS fitting are also studied, via both a simulation study and an application to yeast cell cycle gene data, which suggest that IPS fitting is superior to other existing methods.

The remainder of the paper is organized as follow. In Section 2 we introduce the principal component models and penalized spline regression for functional data. The IPS procedure proposed, together with its theoretical properties, is presented in Section 3. Simulation results that illustrate the effectiveness of the methodology are reported in Section 4. The application of IPS fitting to yeast cell cycle gene expression data is provided in Section 5, and concluding remarks are offered in Section 6. Technical details are deferred to Appendix A.

## 2. Background

This section provides some background material for the development of the IPS procedure. First, a general description of the classical functional principal component models is given in Section 2.1. Then Section 2.2 demonstrates how the penalized splines can be straightforwardly applied to model the mean function when the within-subject correlation between repeated measurements from the same subject is ignored. Lastly, for completeness, we summarize the relevant results from Yao *et al.* (2003) that will be required for the rest of this paper.

### 2.1. Model with measurement error

We model the functional data as noisy repeated measurements from a collection of curves with the common unknown covariance function $G(s,t) = \mathrm{cov}\{X_i(s), X_i(t)\}$, where $X_i$ is the smooth random trajectory of the $i$th subject. The domain of $X_i(\cdot)$ typically is a bounded and closed time interval $\mathcal{T}$, although it could also be a spatial variable, such as in image or geoscience applications. We assume that there is an orthogonal expansion (in the $L^2$-sense) of $G$ in terms of eigenfunctions $\{\phi_k\}_{k=1,2,\ldots}$ and non-increasing eigenvalues $\{\lambda_k\}_{k=1,2,\ldots}$:

$$G(s,t)=\sum_{k=1}^{\infty}\lambda_k\,\phi_k(s)\,\phi_k(t), \qquad t,s\in\mathcal{T}.$$

Karhunen–Loève representation in the classical functional principal component analysis implies that the $i$th random curve can be expressed as

$$X_i(t)=\mu(t)+\sum_{k=1}^{\infty}\xi_{ik}\,\phi_k(t), \qquad t\in\mathcal{T},$$

where $\mu(t)$ is the mean function, the coefficients

$$\xi_{ik}=\int_{\mathcal{T}}\{X_i(t)-\mu(t)\}\,\phi_k(t)\,\mathrm{d}t$$

are uncorrelated random variables with zero mean and variances $E(\xi_{ik}^2)=\lambda_k$, and $\Sigma_k\,\lambda_k<\infty$, $\lambda_1\geqslant\lambda_2\geqslant\ldots$.

To model the noisy observations realistically, we incorporate uncorrelated measurement error $\varepsilon_{ij}$ from a common distribution family with mean 0 and variances $\sigma^2(t_{ij})$ that may be heteroscedastic to reflect the additional noise, where $\sigma^2(t)$ is assumed to be bounded from 0 and $\infty$ on $\mathcal{T}$, i.e. $0<\inf_{t\in\mathcal{T}}\{\sigma^2(t)\}\leqslant\sup_{t\in\mathcal{T}}\{\sigma^2(t)\}<\infty$. Let $Y_{ij}$ denote the $j$th observation of $X_i(\cdot)$ at time $t_{ij}$ with additional noise $\varepsilon_{ij}$ that is independent of $\xi_{ik}$, $i=1,\ldots,n,\ j=1,\ldots,n_i, k=1,2,\ldots$, where $n_i$ is the number of measurements that are made on the $i$th subject. Then we consider the model

$$\begin{aligned}Y_{ij}&=X_i(t_{ij})+\varepsilon_{ij}\\&=\mu(t_{ij})+\sum_{k=1}^{\infty}\xi_{ik}\,\phi_k(t_{ij})+\varepsilon_{ij}, \qquad t_{ij}\in\mathcal{T},\end{aligned} \tag{1}$$

where $E(\varepsilon_{ij})=0$ and $E(\varepsilon_{ij}^2)=\sigma^2(t_{ij})$.

### 2.2. Estimation of mean function using penalized spline regression

As the mean function $\mu(t)$ is assumed smooth, we can estimate $\mu(t)$ by using penalized regression with a spline basis. Owing to its flexibility to capture non-linear relationships, efficiency in computation, and its ability to provide effective inferential tools, penalized spline regression has become a popular method for estimating smooth functions (see Ruppert *et al.* (2003)). Let $\mathbf{Y}_i=(Y_{i1},\ldots,Y_{in_i})^{\mathrm{T}}$ and $\mathbf{T}_i=(t_{i1},\ldots,t_{in_i})^{\mathrm{T}}$. Also let $B_q(t)=(B_{q1}(t),\ldots,B_{qq}(t))^{\mathrm{T}}$ denote the $q$-vector of a spline basis evaluated at time $t$ that is used to model the mean function $\mu(t)$. The mean function $\mu(t)$ is thus modelled by the penalized approximation $B_q^{\mathrm{T}}(t)\boldsymbol{\beta}$, where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_q)^{\mathrm{T}}$ is the coefficient vector. Let $\lambda^*$ be the smoothing parameter and $\mathbf{D}$ some symmetric positive semidefinite matrix. Let $\mathbf{B}_{qi}=(B_q(t_{i1}),\ldots,B_q(t_{in_i}))^{\mathrm{T}}$ denote the $n_i\times q$ spline basis matrices evaluated at design points $\mathbf{T}_i$. Then the coefficient vector $\boldsymbol{\beta}$ is estimated by the minimizer of the penalized least squares criterion

$$\sum_{i=1}^{n}\|\mathbf{Y}_i-\mathbf{B}_{qi}\boldsymbol{\beta}\|^2+\lambda^*\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}, \tag{2}$$

where the roughness penalty is given by $\lambda^*\boldsymbol{\beta}^{\mathrm{T}}\mathbf{D}\boldsymbol{\beta}$. The idea of introducing such a penalty term can be dated back as early as O'Sullivan (1986). If there is correlation between repeated measurements that are made for the same subject, the estimates that are obtained by criterion (2) might not be optimal.

A typical choice of the spline basis is the truncated power basis of degree $p$, i.e. $B_q(t)=(1,t,\ldots,t^p,(t-\kappa_1)_+^p,\ldots,(t-\kappa_k)_+^p)^{\mathrm{T}}$ with knots at $\kappa_1,\ldots,\kappa_k$, which implies that $q=p+k+1$,

where $(x)_+ = \max(0, x)$. A common choice of $\mathbf{D}$ is the block diagonal matrix diag($\mathbf{0}_{(p+1)\times(p+1)}$, $\mathbf{I}_{k\times k}$), where $\mathbf{0}_{p\times p}$ is a $p \times p$ matrix with all entries equal to 0 and $\mathbf{I}_{p\times p}$ is the $p \times p$ identity matrix. Other spline bases can also be used to achieve low rank approximations, such as $B$-splines and radial basis functions. In our implementation the smoothing parameter is, owing to the within-subject correlation, chosen by leave out one curve generalized cross-validation (Rice and Silverman, 1991). In practice the choice of the number of knots is not as crucial as the choice of the smoothing parameter as long as an adequate number of knots are used (see Ruppert (2002)). A reasonable choice of knots $\kappa_j$ in our simulation and application example can be achieved by selecting the 10th, 20th, ..., 90th percentiles of the pooled observation times.

Shi *et al.* (1996), Rice and Wu (2000) and James *et al.* (2001) studied the use of $B$-splines with no roughness penalties to model the individual curves with random coefficients through mixed effects models. Perhaps because of the complexity of their modelling approaches, they did not investigate the asymptotic properties of the estimated components in relation to the true values, such as the behaviours of the estimated mean, covariance structure and principal components. In contrast, in our spline-based modelling approach we represent the trajectories directly through the Karhunen–Loève expansion, in which the eigenfunctions are determined from the data. With this simpler and more direct approach, as demonstrated below, we can derive asymptotic properties of our proposed procedure.

### 2.3. Estimation of covariance surface and functional principal components

In this paper we adopt the procedures of Yao *et al.* (2003) to estimate the covariance surface, the variance of errors and the functional principal components in model (1). This subsection provides a brief description of these estimation procedures.

Let $K_1(\cdot)$ and $K_2(\cdot, \cdot)$ be univariate and bivariate compactly supported kernel densities with zero means and finite variances that are used to estimate covariance $G(s, t)$ and $\{G(t, t) + \sigma^2(t)\}$. Let $h_G = h_G(n)$ and $h_V = h_V(n)$ be the corresponding bandwidths. Let

$$G_i(t_{ij}, t_{il}) = \{Y_{ij} - \hat{\mu}(t_{ij})\}\{Y_{il} - \hat{\mu}(t_{il})\},$$

where $\hat{\mu}(t)$ is the estimated mean function that is obtained from the previous step. The local linear smoother estimate $\hat{G}(s, t)$ for $G(s, t)$ is obtained by minimizing

$$\sum_{i=1}^{n} \sum_{1 \leqslant j \neq l \leqslant n_i} K_2\left(\frac{t_{ij} - s}{h_G}, \frac{t_{il} - t}{h_G}\right) [G_i(t_{ij}, t_{il}) - f\{\gamma, (s, t), (t_{ij}, t_{il})\}]^2 \tag{3}$$

where

$$f\{\gamma, (s, t), (t_{ij}, t_{il})\} = \gamma_0 + \gamma_{11}(s - t_{ij}) + \gamma_{12}(t - t_{il}).$$

To estimate $\sigma^2(t)$, a local linear fit is obtained in the directions of the diagonal, where

$$\sum_{i=1}^{n} \sum_{j=1}^{n_i} K_1\left(\frac{t_{ij} - t}{h_V}\right) \{G_i(t_{ij}, t_{ij}) - \alpha_0 - \alpha_1(t - t_{ij})\}^2 \tag{4}$$

is minimized. The resulting linear fit is denoted as $\hat{V}(t)$ and the estimate of $\sigma^2(t)$ is then

$$\hat{\sigma}^2(t) = \int_{\mathcal{T}} \{\hat{V}(t) - \hat{G}(t, t)\}_+ \, \mathrm{d}t, \tag{5}$$

where $(x)_+ = \max(0, x)$. The estimates of $\{\lambda_k, \phi_k\}_{k \geqslant 1}$ are obtained as the solutions $\{\hat{\lambda}_k, \hat{\phi}_k\}_{k \geqslant 1}$ of the eigenequations

$$\int_{\mathcal{T}} \hat{G}(s,t)\,\hat{\phi}_k(s)\,\mathrm{d}s = \hat{\lambda}_k\,\hat{\phi}_k(t), \tag{6}$$

with orthonormal constraints on $\{\hat{\phi}_k\}_{k \geqslant 1}$ that are unique up to a change in sign; see Yao *et al.* (2003) for details.

When the density of the grid of measurements for each subject is sufficiently large, the functional principal component scores $\xi_{ik} = \int \{X_i(t) - \mu_{g(i)}(t)\}\,\phi_k(t)\,\mathrm{d}t$ are estimated by numerical integration:

$$\hat{\xi}_{ik} = \sum_{j=2}^{n_i} \{Y_{ij} - \hat{\mu}(t_{ij})\}\,\hat{\phi}_k(t_{ij})(t_{ij} - t_{i,\,j-1}). \tag{7}$$

Finally, for the selection of the number of eigenfunctions $K$, we could use the Akaike information criterion (AIC) type of criterion that was suggested by Yao *et al.* (2005). Denote $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(t_{i1}), \ldots, \hat{\mu}(t_{in_i}))^{\mathrm{T}}$, $\hat{\boldsymbol{\phi}}_{ik} = (\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{in_i}))^{\mathrm{T}}$ and $\Sigma_i = \mathrm{diag}\{\hat{\sigma}^2(t_{i1}), \ldots, \hat{\sigma}^2(t_{in_i})\}$. Then, if the error terms $\varepsilon_{ij}$ in model (1) are assumed to be normal, $K$ is chosen by minimizing

$$\mathrm{AIC}(K) \propto \sum_{i=1}^{n} \left\{ -\tfrac{1}{2} \left( \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_{ik} \right)^{\mathrm{T}} \Sigma_i^{-1} \left( \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_{ik} \right) \right\} + K, \tag{8}$$

where the terms that do not depend on $K$ are eliminated. For a more general discussion on the AIC, see Burnham and Anderson (2002).

## 3. Iterative penalized spline fitting for within-subject measurement correlation

This section presents the proposed IPS procedure, a smoothing procedure for fitting functional principal component models. An advantage of adopting penalized splines for the estimation of the group mean functions $\mu_g(t)$ is that it allows easy incorporation of covariates. A naïve application of the penalized splines (e.g. solutions to expression (2)) for this problem will not lead to optimal estimates when within-subject correlation is present. Although Lin and Carroll (2000) showed that, for longitudinal data, it is reasonable to ignore the within-subject correlation when using kernel-based smoothing methods, the same is not true for splines (see Welsh *et al.* (2002)). Splines and conventional kernels are very different in local properties and thus behave differently in terms of accounting for the within-subject dependence. Lin *et al.* (2004) showed that the smoothing spline estimator has the smallest variance when the unobservable true covariance function is used. However, it is not clear whether this conclusion can be extended to penalized spline regression, as penalized splines are a low rank smoothing method whereas smoothing splines are a full rank smoothing method. These considerations suggest the need for a more sophisticated penalized spline estimation method extending expression (2). The proposed IPS procedure is designed for handling this issue.

### 3.1. Iterative penalized spline procedure

Hall and Opsomer (2005) showed that the penalized spline smoother is a uniformly consistent estimator for independent data. Motivated by this result, our strategy is to reduce the within-subject correlation between observations that are made for the same subject so that after iteration the *empirical working data* (defined in equation (9) below) are asymptotically equivalent (in probability) to a set of independent data. We first assume that the trajectories are observed on

a dense grid, i.e. the density of measurements for each subject is sufficiently large. The case for sparse functional data is briefly discussed later.

Given an initial mean estimate $\hat{\mu}^{(0)}$, the IPS procedure iterates, until convergence, the following steps for $l = 0, 1, 2, \ldots$.

*Step 1*: with the current mean function estimate $\hat{\mu}^{(l)}$ at the $l$th iteration, obtain an estimate $\hat{G}^{(l)}$ for the smooth covariance surface by two-dimensional local linear smoothing (3). Note that the empirical variances that are obtained at the diagonal of the surface are omitted, as these are contaminated with the residual variance $\sigma^2(t)$ (see expression (3)).
*Step 2*: use equation (6) to compute estimates $\hat{\phi}_k^{(l)}$ and $\hat{\lambda}_k^{(l)}$ for respectively the eigenfunctions and eigenvalues.
*Step 3*: by using the empirical variances that are obtained on the diagonal of the covariance surface, estimate the variance function $\sigma^2(t)$ by equation (5).
*Step 4*: use the integration approximation (7) to obtain estimate $\hat{\xi}_{ik}^{(l)}$ for the individual functional principal component scores.
*Step 5*: for all $i$ and $j$, define the *theoretical working data* as

$$Y_{ij}^* = Y_{ij} - \sum_{i=1}^{\infty} \xi_{ik}\, \phi_k(t_{ij}).$$

Note that the $Y_{ij}^*$s are independent, and estimate them by the empirical working data

$$\hat{Y}_{ij}^{*(l)} = Y_{ij} - \sum_{k=1}^{K^{(l)}} \hat{\xi}_{ik}^{(l)}\, \hat{\phi}_k^{(l)}(t_{ij}), \tag{9}$$

where $K^{(l)}$ is the number of eigenfunctions, chosen by the AIC (8), that are used for approximation in the current iteration.
*Step 6*: in equation (2) replace the real data $Y_{ij}$ with the estimated working data $\hat{Y}_{ij}^{*(l)}$ and compute the next iterative mean function estimate $\hat{\mu}^{(l+1)}$ as its minimizer.

In our implementation, convergence is declared if the following relative integrated squared difference RISD between $\hat{\mu}^{(l)}$ and $\hat{\mu}^{(l+1)}$ is less than a prespecified tolerance:

$$\mathrm{RISD}_l = \int_{\mathcal{T}} \{\hat{\mu}^{(l+1)}(t) - \hat{\mu}^{(l)}(t)\}^2 \,\mathrm{d}t \Big/ \int_{\mathcal{T}} \hat{\mu}^{(l)}(t)^2 \,\mathrm{d}t. \tag{10}$$

Also, for the initial estimates $\hat{\mu}^{(0)}$, we investigated the use of the penalized spline model (2) and the more traditional local polynomial smoothing (for example see expression (15) in Appendix A.1). For both cases the amount of smoothing is chosen by leave out one curve cross-validation or its generalized version.

*Remark 1.* It is easy to extend the approach proposed to sparse functional data, i.e. when the number of repeated measurements that are available per subject is small. For sparse data, Yao *et al.* (2005) demonstrated that the best linear prediction, denoted by $\hat{\xi}_{ik}^P$, of $\xi_{ik}$ given the data from the subject $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})$ outperforms the traditional integration approximation (7). Yao *et al.* (2005) termed $\hat{\xi}_{ik}^P$ the principal component analysis through conditional expectation estimate. Details on the construction of $\hat{\xi}_{ik}^P$ can be found in Yao *et al.* (2005). Hence, for sparse data, in step 4 of the IPS procedure we suggest replacing the integration estimates $\hat{\xi}_{ik}$ (7) by the principal component analysis through conditional expectation estimates $\hat{\xi}_{ik}^P$. Although the theory of coupling IPS with principal component analysis through conditional expectation has not been developed, simulation results to be reported in the next section demonstrate the promising practical performance of IPS with it.

*Remark 2.* As mentioned before, an advantage of using penalized spline regression to estimate the mean function is that it allows simple implementation of approximate inference procedures. Here we follow the approach of Ruppert *et al.* (2003) and demonstrate the construction of an approximate confidence interval for any contrast of the coefficients, i.e. $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$ for any $\mathbf{a} \in \Re^q$. First, from equation (2) with $\mathbf{Y}_i$ replaced by $\mathbf{Y}_i^* = (Y_{i1}^*, \ldots, Y_{in_i}^*)^{\mathrm{T}}$, it is straightforward to obtain the following closed form expression for $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}} = \left( \sum_{i=1}^{n} \mathbf{B}_{qi}^{\mathrm{T}} \mathbf{B}_{qi} + \lambda^* \mathbf{D} \right)^{-1} \sum_{i=1}^{n} \mathbf{B}_{qi} \mathbf{Y}_i^*.$$

Also, $\mathrm{cov}(Y_{ij}^*, Y_{il}^*) = \delta_{jl}\sigma^2(t_{ij})$, where $\sigma^2(\cdot)$ is the variance function of measurement error and $\delta_{kl} = 1$ for $k = l$ and $\delta_{kl} = 0$ otherwise. Denote $\mathbf{R}_i = \mathrm{diag}\{\sigma^2(t_{i1}), \ldots, \sigma^2(t_{in_i})\}$. Direct calculations lead to the following covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$ for $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \Sigma_{\hat{\boldsymbol{\beta}}} &= \mathrm{cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}) \\ &= \left( \sum_{i=1}^{n} \mathbf{B}_{qi}^{\mathrm{T}} \mathbf{B}_{qi} + \lambda^* \mathbf{D} \right)^{-1} \left( \sum_{i=1}^{n} \mathbf{B}_{qi} \mathbf{R}_i \mathbf{B}_{qi}^{\mathrm{T}} \right) \left( \sum_{i=1}^{n} \mathbf{B}_{qi}^{\mathrm{T}} \mathbf{B}_{qi} + \lambda^* \mathbf{D} \right)^{-1}. \end{aligned}$$

Then an approximate $100(1-\alpha)\%$ confidence interval of $\mathbf{a}^{\mathrm{T}}\boldsymbol{\beta}$ can be obtained by

$$\mathbf{a}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm \Phi(1-\alpha/2)(\mathbf{a}^{\mathrm{T}}\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}\mathbf{a})^{1/2}, \tag{11}$$

where $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is calculated by plugging in the corresponding estimates that are obtained from the last iteration of the IPS fitting, and $\Phi(\cdot)$ is the standard Gaussian distribution function. Possible fixed covariates can be included by adding columns to the design matrices $\mathbf{B}_{qi}$ under appropriate model assumptions. Our framework also provides a natural way for examining the time-varying effect of any time-independent random covariate; for example, see Rice and Wu (2000) and Chiou *et al.* (2003).

### 3.2. Theoretical properties of iterative penalized splines

We have studied the theoretical properties of the IPS procedure proposed, and we summarize the results in the following two theorems. Assumptions and proofs are deferred to Appendices A.1 and A.2. For simplicity, we consider only the case of one-step iteration. In what follows $g(x;t)$ denotes the density function of $Y(t)$ and $g_2(y_1, y_2; t_1, t_2)$ denotes the density of $(Y(t_1), Y(t_2))$. It is assumed that these density functions satisfy appropriate regularity conditions.

We assume that the initial estimates $\hat{\mu}^{(0)}$ are obtained by local polynomial smoothing, as described by expression (15) in Appendix A.1, and we expect that similar theoretical results can be obtained if $\hat{\mu}^{(0)}$ were computed by the penalized spline model (2). We further require that the repeated measurements from each subject are sufficiently dense; see the precise description in Appendix A.1. Under these conditions, we obtain uniform consistency of the estimates of the local polynomial estimates of the mean and the covariance functions of the process $X(t)$. We also obtain convergence results for the estimated principal components, with the rate depending on the specific property of the process $X$ as stated in lemma 2 (see Appendix A.2). These results are the first step to the asymptotic analysis of the empirical working data $\{\hat{Y}_{ij}^{*(0)}\}$.

The central results towards the theoretical analysis of the empirical working data are presented in theorem 1, which provides uniform consistency of the estimated principal component scores $\hat{\xi}_{ik}^{(0)}$ and thus the empirical working data $Y_{ij}^*$ over $j$s. These results form the basis for the consistent estimation of model components using empirical working data in the iterative steps.

*Theorem 1.* Under assumptions (a)–(j) in Appendix A.1 and appropriate regularity assumptions for $g(x;t)$ and $g_2(x_1,x_2;t_1,t_2)$,

$$\sup_{1\leqslant k\leqslant K}|\hat{\xi}_{ik}^{(0)}-\xi_{ik}|\xrightarrow{\mathrm{p}}0, \tag{12}$$

$$\sup_{1\leqslant j\leqslant n_i}|\hat{Y}_{ij}^{*(0)}-Y_{ij}^{*}|\xrightarrow{\mathrm{p}}0. \tag{13}$$

From theorem 1 we conclude that the empirical working data $\{\hat{Y}_{ij}^{*(0)}\}$ are asymptotically equivalent to their theoretical counterparts $\{Y_{ij}^{*}\}$, and in fact $\sup_{1\leqslant j\leqslant n_i}|\hat{Y}_{ij}^{*(0)}-Y_{ij}^{*}|=O_p(\theta_{in})$ where $\theta_{in}$ is defined in equation (22) in Appendix A.2 and the $O_p(\cdot)$ term holds uniformly over all $i$s. Since these theoretical working data $\{Y_{ij}^{*}\}$ are independent, by applying penalized spline smoothing to the empirical working data $\hat{Y}_{ij}^{*(0)}$, we obtain the uniform consistency for the penalized spline estimate of the mean function, by using the results in section 4.2 of Hall and Opsomer (2005) under appropriate conditions. Then the uniform consistency can also be shown for the covariance estimator $\hat{G}$ that is obtained as in step 1. The central results are provided in theorem 2 below.

*Theorem 2.* Under assumptions (a)–(m) in Appendix A.1 and appropriate regularity assumptions for $g(x;t)$ and $g_2(x_1,x_2;t_1,t_2)$,

$$\begin{aligned}&\sup_{t\in\mathcal{T}}|\hat{\mu}(t)-\mu(t)|\xrightarrow{\mathrm{p}}0,\\ &\sup_{s,t\in\mathcal{T}}|\hat{G}(s,t)-G(s,t)|\xrightarrow{\mathrm{p}}0.\end{aligned} \tag{14}$$

*Remark 3.* We remark that the uniform convergence rate for the penalized spline estimator $\hat{\mu}$ is obtained as $O_p(\omega_n+\theta_n^*)$, where $\omega_n$ and $\theta_n^*$ are as defined in expressions (21) and (23), and thus the uniform convergence rate of the covariance estimator $\hat{G}$ (3), in which $\hat{\mu}(t)$ is used, can also be expressed explicitly as $O_p(\omega_n+\theta_n^*+1/n^{1/2}h_G^2)$, where $h_G$ is the bandwidth that is used in expression (3). On the basis of theorem 2, the asymptotic consistency for other model components, such as principal components, can also be obtained by analogy with lemma 1.

*Remark 4.* We can see that the iterative approach that is proposed here is also applicable to other smoothing methods and is not only restricted to penalized spline regression. For example, we could apply local polynomial smoothing or smoothing spline methods to the empirical working data in each iteration. On the basis of the established theory for those smoothing methods, the theoretical arguments for the iterative approach still hold, and similar consistency results can be obtained. Because of the computational efficiency, we focus on the penalized spline models, though the theoretical development for penalized splines is more challenging.

## 4. Simulation studies

To assess the practical performance of the IPS procedure proposed, a simulation study was conducted. We generated 100 independently and identically distributed normal and 100 independently and identically distributed non-normal samples consisting of $n=100$ random trajectories. The simulated processes have a mean function $\mu(t)=t+\sin(t)$, $0\leqslant t\leqslant 10$, and covariance function derived from two eigenfunctions $\phi_1(t)=-\cos(\pi t/10)/\sqrt{5}$ and $\phi_2(t)=\sin(\pi t/10)/\sqrt{5}$, $0\leqslant t\leqslant 10$. We chose $\lambda_1=4$, $\lambda_2=1$ and $\lambda_k=0$, $k\geqslant 3$, as eigenvalues and $\sigma^2(t)\equiv 0.25$ as variance of the additional measurement errors $\varepsilon_{ij}$ in model (1), which are assumed to be normal

with mean 0. For the 100 normal samples, the functional principal component scores $\xi_{ik}$ were generated from the $\mathcal{N}(0, \lambda_k)$ distribution, whereas the $\xi_{ik}$ for the non-normal samples were generated from a mixture of two normal distributions, $\mathcal{N}\{(\lambda_k/2)^{1/2}, \lambda_k/2\}$ with probability $\frac{1}{2}$ and $\mathcal{N}\{-(\lambda_k/2)^{1/2}, \lambda_k/2\}$ with probability $\frac{1}{2}$.

We also consider both sparse and non-sparse designs. For an equally spaced grid $\{c_0, \ldots, c_{50}\}$ on $[0, 10]$ with $c_0 = 0$ and $c_{50} = 10$, let $s_i = c_i + e_i$, where $e_i$ are independently and identically distributed with $\mathcal{N}(0, 0.1^2)$, $s_i = 0$ if $s_i < 0$ and $s_i = 10$ if $s_i > 10$, allowing for non-equidistant 'jittered' designs. For the sparse design, each curve was sampled at a random number of points, chosen from a discrete uniform distribution on $\{3, \ldots, 6\}$, and the locations of the measurements were randomly chosen from $\{s_1, \ldots, s_{49}\}$ without replacement, whereas, for the non-sparse design, the number of observations for each curve was randomly chosen from $\{20, \ldots, 30\}$.

The following four different methods were compared.

(a) The mean function $\mu(t)$ is estimated by using the penalized spline model (2), and the covariance and principal components are estimated by the method that was described in Section 2.3. Note that no iteration is performed (method 1).
(b) Method 2 is similar to method 1, but the mean function $\mu(t)$ is estimated with local polynomial smoothing (15).
(c) Method 3 is the IPS procedure proposed where the initial group mean estimates $\hat{\mu}^{(0)}$ are obtained by local polynomial smoothing (15).
(d) Method 4 is the IPS procedure proposed where the initial group mean estimates $\hat{\mu}^{(0)}$ are obtained by the penalized spline model (2).

In these methods, for all the local polynomial smoothing steps, either the univariate or the bivariate Epanechnikov kernel functions were used, i.e.

$$K_1(x) = \frac{3}{4}(1 - x^2)\,\mathbf{1}_{[-1,1]}(x)$$

and

$$K_2(x, y) = \frac{9}{16}(1 - x^2)(1 - y^2)\,\mathbf{1}_{[-1,1]}(x)\,\mathbf{1}_{[-1,1]}(y),$$

where $\mathbf{1}_A(x) = 1$ if $x \in A$ and $\mathbf{1}_A(x) = 0$ otherwise for any set $A$ and, for the penalized spline regression, a cubic spline basis was used, i.e. $p = 3$.

To demonstrate the superior performances of the IPS procedure proposed (methods 3 and 4) compared with the non-iterative methods (1 and 2), we report in Table 1 the Monte Carlo estimates that were obtained from 100 non-sparse or sparse and normal or mixture simulated data sets (in total 400 data sets) for the integrated mean-squared error IMSE of $\hat{\mu}(t)$ that consists of the integrated squared bias IBIAS and integrated variance IVAR, i.e.

$$\int_0^{10} E[\{\hat{\mu}(t) - \mu(t)\}^2]\,\mathrm{d}t = \int_0^{10} [\hat{\mu}(t) - E\{\hat{\mu}(t)\}]^2\,\mathrm{d}t + \int_0^{10} [E\{\hat{\mu}(t)\} - \mu(t)]^2\,\mathrm{d}t.$$

Recall that the predicted individual trajectories using $K$ eigenfunctions are denoted by $\hat{X}_i^K(t) = \hat{\mu}(t) + \Sigma_{k=1}^K \hat{\xi}_{ik}\,\hat{\phi}_k(t)$, where $\hat{\xi}_{ik}$ are obtained either by the integration method (7) for non-sparse data or by the principal component analysis through conditional expectation method for sparse data.

**Table 1.** Simulation results for comparing mean estimates obtained by methods 1–4 from 100 Monte Carlo runs with $n = 100$ random trajectories per sample†

| *Design* | *Method* | *Results for the normal distribution* | | | | *Results for the mixture distribution* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *IBIAS* | *IVAR* | *IMSE* | *IPE* | *IBIAS* | *IVAR* | *IMSE* | *IPE* |
| *Optimal* | | | | | | | | | |
| Non-sparse (integration) | 1 | 0.003 | 0.066 | 0.069 | 0.242 | 0.004 | 0.065 | 0.069 | 0.243 |
| | 2 | 0.007 | 0.072 | 0.079 | 0.246 | 0.008 | 0.072 | 0.080 | 0.247 |
| | 3 | 0.003 | 0.055 | 0.058 | 0.223 | 0.003 | 0.056 | 0.059 | 0.224 |
| | 4 | 0.003 | 0.056 | 0.059 | 0.224 | 0.003 | 0.057 | 0.060 | 0.225 |
| Sparse (principal component analysis through conditional expectation) | 1 | 0.006 | 0.154 | 0.160 | 1.77 | 0.008 | 0.166 | 0.174 | 1.75 |
| | 2 | 0.030 | 0.173 | 0.203 | 1.79 | 0.034 | 0.178 | 0.212 | 1.79 |
| | 3 | 0.004 | 0.116 | 0.120 | 1.62 | 0.003 | 0.123 | 0.126 | 1.63 |
| | 4 | 0.004 | 0.118 | 0.122 | 1.61 | 0.004 | 0.125 | 0.129 | 1.60 |
| *Model selected* | | | | | | | | | |
| Non-sparse (integration) | 1 | 0.004 | 0.070 | 0.074 | 0.245 | 0.003 | 0.069 | 0.073 | 0.245 |
| | 2 | 0.008 | 0.077 | 0.085 | 0.248 | 0.007 | 0.077 | 0.084 | 0.251 |
| | 3 | 0.003 | 0.058 | 0.061 | 0.227 | 0.003 | 0.058 | 0.061 | 0.228 |
| | 4 | 0.003 | 0.059 | 0.062 | 0.226 | 0.003 | 0.059 | 0.062 | 0.227 |
| Sparse (principal component analysis through conditional expectation) | 1 | 0.004 | 0.205 | 0.209 | 1.84 | 0.005 | 0.198 | 0.203 | 1.85 |
| | 2 | 0.028 | 0.218 | 0.246 | 1.86 | 0.034 | 0.208 | 0.242 | 1.84 |
| | 3 | 0.003 | 0.148 | 0.151 | 1.69 | 0.002 | 0.154 | 0.156 | 1.68 |
| | 4 | 0.003 | 0.145 | 0.148 | 1.68 | 0.003 | 0.148 | 0.151 | 1.67 |

†The functional principal component scores were calculated by using either the integration or the principal component analysis through conditional expectation methods that were described in Section 3. Shown are the Monte Carlo estimates of the integrated mean-squared error IMSE, the integrated squared bias IBIAS, the integrated variance IVAR and the integrated squared prediction error IPE. See Section 4 for details.

To avoid the bias in comparison that is possibly caused by inadequate choices of tuning parameters, such as $\lambda^*$, $K$, bandwidths and knots, we constructed two scenarios. First, methods 1–4 are compared at optimal tuning parameter values, denoted by 'optimal'. Specifically, the optimal bandwidths $h_\mu$ for $\hat{\mu}^{(0)}(t)$ that were used in methods 2 and 3 (initial estimate in method 3) is chosen by minimizing the $L^2$-distance between the estimated and true mean functions, i.e. $\int_0^{10} \{\hat{\mu}^{(0)}(t; h_\mu) - \mu(t)\}^2 \, \mathrm{d}t$. Other optimal smoothing parameters, including $h_G$, $h_V$ and $\lambda^*$ that were used in methods 1–4 for the covariance function of $X(t)$, the variance function of $\varepsilon(t)$ and the penalized spline estimate of $\mu(t)$ are also chosen by minimizing the corresponding $L^2$-distances. The number of eigenfunctions $K$ was fixed at the true value 2. Second, the tuning parameters are chosen by model-based procedures, denoted by 'model selected'. For computational convenience, here we used tenfold cross-validation to choose $h_\mu$, $h_G$ and $h_V$, which involved removing 10% of the individual curves as a test set, finding the estimates from the remaining data and repeating the process nine more times, whereas $\lambda^*$ was chosen by tenfold generalized cross-validation. The number of eigenfunctions $K$ in each run was chosen by the AIC (8). Since Ruppert (2002) showed that the penalized spline estimators are relatively insensitive to the choice of basis functions compared with the choice of $\lambda^*$, as long as enough of them are used, here an adequate choice of knots (10th, ..., 90th percentiles of the pooled observation times) was used in methods 1–4 for both scenarios.

From Table 1, we can see that the IPS procedures (methods 3 and 4) improved the integrated mean-squared errors IMSE of mean estimates over non-iterative procedures 1 and 2 by around

25–40% for sparse samples and 15–35% for non-sparse samples in both the optimal and the model-selected scenarios, although it is not surprising that the IMSEs that were obtained in the optimal scenario are slightly smaller than those in the model-selected scenario. The procedures are 'robust' regarding the distribution of random components $\xi_{ik}$, yielding similar amounts of improvement for the normal and mixture samples. This also provides empirical justifications for the use of IPS in the sparse data situation where the functional principal component scores are obtained by principal component analysis through conditional expectation estimates compared with the non-iterative procedures. It is interesting that the improvement that is obtained by the IPS procedures for sparse samples is more dramatic than that obtained for non-sparse samples, which suggests that further investigation of such a phenomenon is worthwhile (and also beyond the scope of this paper). The comparison also suggests that the bias is not of concern and the variance is a dominating factor when comparing the IMSEs. The proposed AIC (8) chose the correct number of principal components, $K = 2$, for around 95 out of 100 samples in each situation of the model-selected scenario (a total of 400 samples: non-sparse or sparse and normal or mixture). Regarding computational efficiency, the IPS procedures proposed (methods 3 and 4) usually converge very quickly, with no more than four iterations with tolerance (10) equal to $10^{-4}$ in all the simulation runs. In addition, the computational times for a sparse and a non-sparse sample are about, on a Pentium-M 1.6G laptop, 1 min and 10 min respectively.

We also compared the Monte Carlo estimates of the integrated squared prediction error IPE of the true curves $X_i$ obtained by methods 1–4 from those simulated samples, i.e.

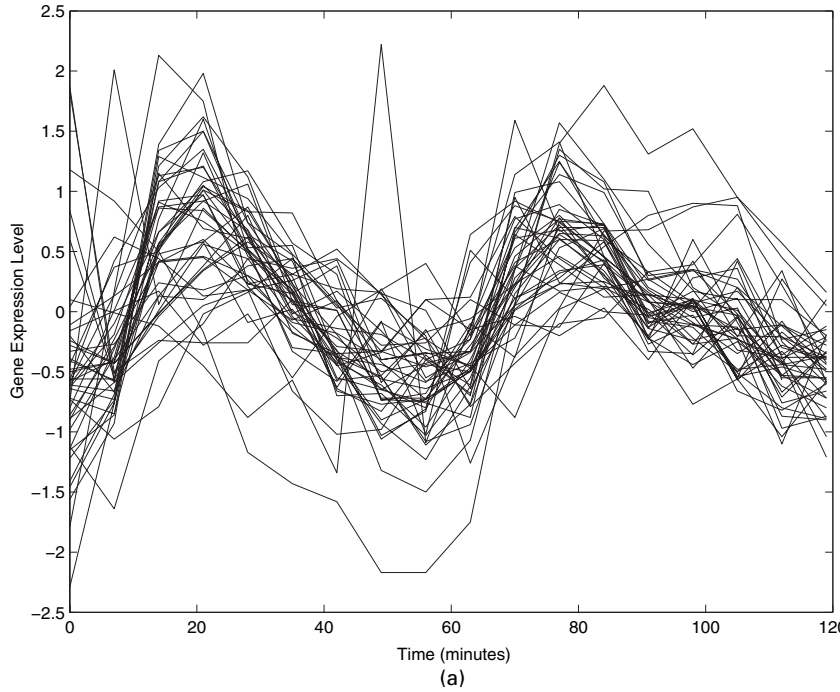
$$\mathrm{IPE} = \sum_{i=1}^{n} \int_0^{10} \{X_i(t) - \hat{X}_i^K(t)\}^2 \, \mathrm{d}t/n,$$

where $\hat{X}_i^K(t) = \hat{\mu}(t) + \Sigma_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t)$, reported in Table 1. It is seen that the IPS procedures (methods 3 and 4) improve the prediction errors by around 10%, and methods 3 and 4 gave comparable results. This is consistent with the previous observations which were obtained from comparing the IMSEs of the mean estimates regarding the superior performance of the IPS procedures proposed.

## 5. Application to yeast cell cycle gene expression data

Time course gene expression data (factor synchronized) for the yeast cell cycle were obtained by Spellman *et al.* (1998). The experiment started with a collection of yeast cells, whose cycles were synchronized ($\alpha$-factor based) by a chemical process. There are 6178 genes in total, and each gene expression profile consists of 18 data points, measured every 7 min between 0 and 119 min, covering two cell cycles. Of these genes, 92 had sufficient data and were identified by traditional biological methods, of which 43 are known to be related to the G1 phase regulation that is of interest. To demonstrate the method proposed, the 43 genes related to the G1 phase are used in the following analysis; Fig. 1. The gene expression level measurement at each time point is obtained as a logarithm of the expression level ratio.

Two estimates of the mean function are shown in Fig. 1(b). The first mean estimate was obtained by using the proposed IPS procedure with initial estimates given by local polynomial smoothing (15). Using expression (11) and the penalized spline approximation $\mu(t) \approx B_q^{\mathrm{T}}(t)\boldsymbol{\beta}$, an approximate 95% pointwise confidence interval was also constructed. The second mean estimate was obtained by traditional local polynomial smoothing (15) (i.e. with no iteration), where the bandwidth $h_\mu$ and the smoothing parameter $\lambda^*$ were chosen by leave out one curve



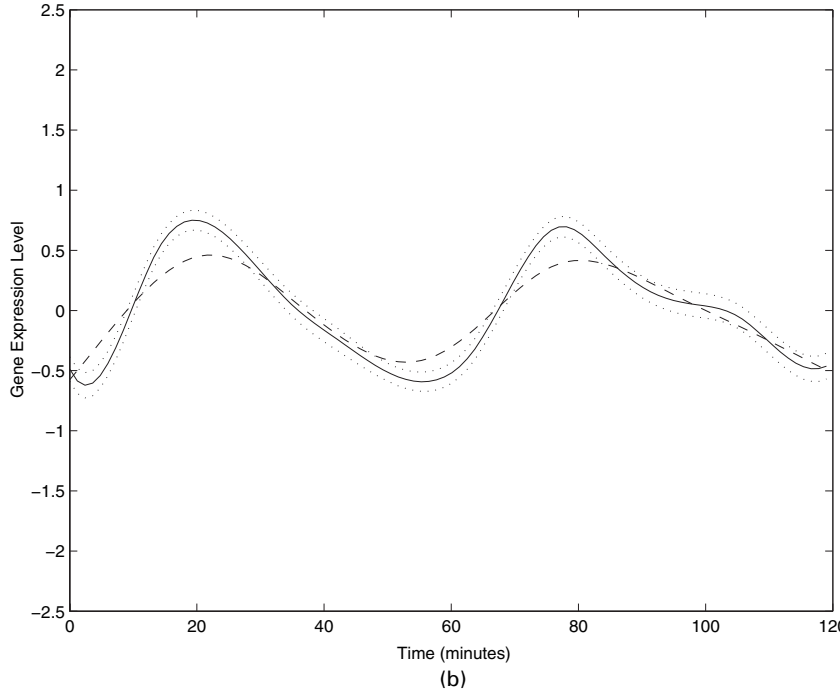




**Fig. 1.** (a) Gene expression profiles of 43 genes from the G1 phase and (b) estimated mean functions obtained with traditional non-iterative local polynomial smoothing (15) (-------) and the proposed IPS procedure (———) with initial mean estimates given by expression (15) for the 43 G1 phase genes as well as an approximate pointwise 95% confidence interval (· · · · · · ·) obtained by expression (11)

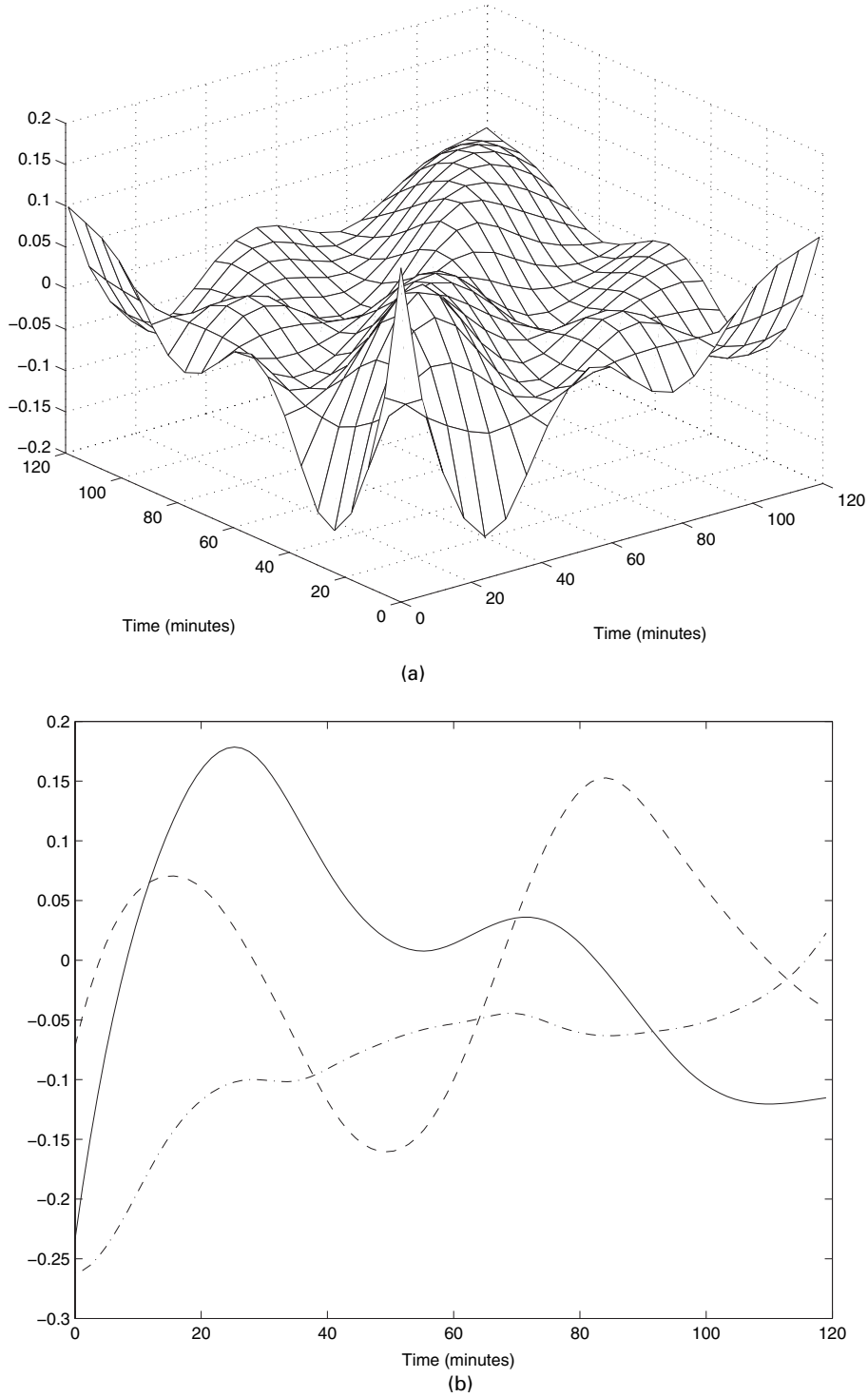



**Fig. 2.** (a) Smooth estimate of the covariance surface and (b) three eigenfunctions obtained using the IPS procedure proposed, for the G1 phase yeast cell cycle gene expression profiles

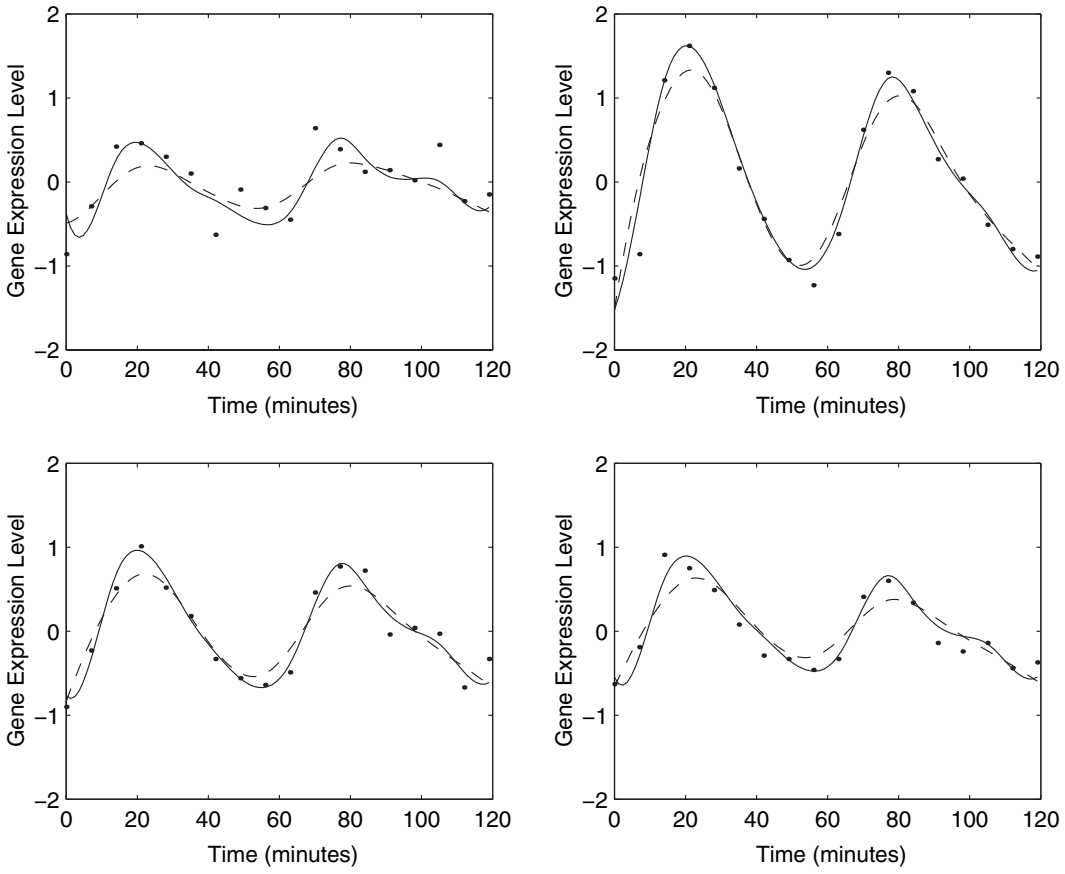



**Fig. 3.** Observed (·) and estimated gene expression profiles obtained by using the IPS procedure proposed (———) and the traditional non-iterative functional principal component analysis combined with local polynomial smoothing (– – –) for four randomly selected genes related to the G1 phase

cross-validation and its generalized version. The iterative procedure converged in three iterations with tolerance (10) set to $10^{-4}$. We find that, when compared with the traditional non-iterative method, the mean function that is estimated by IPS reveals more clearly the features of the regions with high curvature, i.e. peaks or valleys for the G1 phase genes. Another feature of the mean pattern for the G1 phase genes is a delay after the second peak around 100 min in the yeast cell cycle which can also be seen from the original data that are displayed in Fig. 1(a). This is detected by the mean estimate that is obtained by using the method proposed, whereas the estimate that is obtained by the traditional approach with no iteration does not provide any information for this feature.

The smooth covariance surface estimate that is obtained by using the IPS procedure proposed is displayed in Fig. 2(a), where the bandwidth $h_G$ is chosen by leave out one curve cross-validation in each iteration, as well as $h_V$ as in expression (4). This surface estimate reveals the periodic structure of variation patterns of the underlying process for the yeast cell cycle. We use the first three eigenfunctions chosen by AIC (8) to approximate the expression profiles (Fig. 2(b)). The estimates of these three leading eigenfunctions also reflect periodicity as well as an overall shift, explaining around 91% of the total variation.

We randomly select four genes and present the predicted profiles

$$\hat{X}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\, \hat{\phi}_k(t),$$

where $\hat{\xi}_{ik}$ are as in equation (7), in Fig. 3. The predicted trajectories are obtained by using IPS with initial mean estimated by local polynomial smoothing (15), and also the traditional method using local polynomial estimation for the mean and covariance as described in steps 1–4 of Section 3.1. We find that the IPS fitting proposed provides better prediction compared with the traditional non-iterative method, as the observed data are more effectively recovered particularly for the regions with high curvature (peak or valley). We also compare the mean prediction error, which is a global measure of discrepancy defined as

$$\text{MPE} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{\{Y_{ij} - \hat{Y}_i(t_{ij})\}^2}{n_i}.$$

The IPS procedure proposed gives MPE = 0.120, whereas the traditional approach yields MPE = 0.138, which indicates a reduction of about 15%. From the above evidence, we conclude that the iterative IPS procedure proposed indeed improves on the traditional non-iterative approach for the modelling of functional data.

## 6. Concluding remarks

In this paper a new method for performing functional principal component analysis that uses penalized spline regression has been presented. For reducing the within-subject correlation that is commonly found in functional or longitudinal data, an iterative estimation procedure was proposed to improve the estimation of the mean function. Through an analytic derivation of its asymptotic properties, IPS fitting was shown to provide a sample of transformed data which are asymptotically equivalent to independent data. From the investigation of theoretical properties, and the encouraging numerical results that were obtained by simulations and the real data example, we can see that, when comparing with traditional non-iterative methods, significant improvements can be achieved by the approach proposed. Another attractive property of the method is that it allows simple covariate incorporation and straightforward approximate inference.

## Acknowledgements

The authors are most grateful to the reviewers and the Associate Editor for their very constructive comments. The work of Lee was supported in part by US National Science Foundation grant DMS-0203901.

## Appendix A

### *A.1. Assumptions and notation*

Define the local linear scatterplot smoothers for $\mu(t)$ through minimizing

$$\sum_{i=1}^{n} \sum_{l=1}^{n_i} K_1\left(\frac{t_{ij} - t}{h_\mu}\right) \{Y_{ij} - \beta_0 - \beta_1(t - t_{ij})\}^2, \tag{15}$$

with respect to $\beta_0$ and $\beta_1$, leading to $\hat{\mu}^{(0)}(t) = \hat{\beta}_0(t)$.

Without loss of generality, we consider the case of a single group throughout the appendix, i.e. $g = 1$. Recall that $K_1$ and $K_2$ are compactly supported densities with zero means and finite variances, and that $h_\mu = h_\mu(n)$, $h_G = h_G(n)$ and $h_V = h_V(n)$ are the bandwidths for estimating $\hat{\mu}^{(0)}$ in expression (15), $\hat{G}^{(0)}$ in

expression (3) and $\hat{V}^{(0)}$ in expression (4). We develop asymptotics as the number of subjects $n \to \infty$, and require

(a) $h_\mu \to 0$, $h_V \to 0$, $nh_\mu^4 \to \infty$, $nh_V^4 \to \infty$, $nh_\mu^6 < \infty$ and $nh_V^6 < \infty$, and
(b) $h_G \to 0$, $nh_G^6 \to \infty$ and $nh_G^8 < \infty$.

The time points $\{t_{ij}\}_{i=1,\ldots,n;\, j=1,\ldots,n_i}$ here are considered deterministic. Denote the sorted time points across all subjects as $a_X \leqslant t_{(1)} \leqslant \ldots \leqslant t_{(N_n)} \leqslant b_X$, and $\Delta_n = \max\{t_{(k)} - t_{(k-1)} : k = 1, \ldots, N+1\}$, where $N_n = \Sigma_{i=1}^n n_i$, $\mathcal{T} = [a_X, b_X]$, $t_{(0)} = a_X$ and $t_{(N+1)} = b_X$. For the $i$th subject, suppose that the time points $t_{ij}$ have been ordered non-decreasingly. Let $\Delta_{in} = \max\{t_{ij} - t_{i,j-1} : j = 1, \ldots, n_i + 1\}$ and $\Delta_n^* = \max\{\Delta_{in} : i = 1, \ldots, n\}$, where $t_{i0} = a_X$ and $t_{i,n_i+1} = b_X$. Also denote $\bar{n} = n^{-1}\Sigma_{i=1}^n n_i$. To obtain uniform consistency, we require both the pooled data across all subjects and also the data from each subject to be dense in the time domain $\mathcal{T}$. Assume that

(c) $\Delta_n = O(\min\{n^{-1/2}h_\mu^{-1}, n^{-1/2}h_V^{-1}, n^{-1/4}h_G^{-1}\})$ and
(d) $\bar{n} \to \infty$, $\max\{n_i : i = 1, \ldots, n\} \leqslant C\bar{n}$ for some $C > 0$, and $\Delta_n^* = O(1/\bar{n})$, as $n \to \infty$.

Fourier transforms of $K_1(u)$ and $K_2(u, v)$ are denoted by $\kappa_1(t) = \int \exp(-\mathrm{i}ut)\, K_1(u)\, \mathrm{d}u$ and $\kappa_2(t, s) = \int \exp\{-(\mathrm{i}ut + \mathrm{i}vs)\}\, K_2(u, v)\, \mathrm{d}u\, \mathrm{d}v$ respectively. They satisfy the conditions that

(e) $\kappa_1(t)$ is absolutely integrable, i.e. $\int |\kappa_1(t)|\, \mathrm{d}t < \infty$, and
(f) $\kappa_2(t, s)$ is absolutely integrable, i.e. $\int\int |\kappa_2(t, s)|\, \mathrm{d}t\, \mathrm{d}s < \infty$.

Assume that the fourth moment of $Y(t)$ is uniformly bounded for all $t \in \mathcal{T}$, i.e. that

(g) $\sup_{t \in \mathcal{T}}[E\{Y^4(t)\}] < \infty$.

Define the rank 1 operator $f \otimes g = \langle f, h\rangle y$, for $f, h \in H$, and denote the separable Hilbert space of Hilbert–Schmidt operators on $H$ by $F \equiv \sigma_2(H)$, endowed by $\langle T_1, T_2\rangle_F = \mathrm{tr}(T_1 T_2^*) = \Sigma_j \langle T_1 u_j, T_2 u_j\rangle_H$ and $\|T\|_F^2 = \langle T, T\rangle_F$, where $T_1, T_2, T \in F$, and $\{u_j : j \geqslant 1\}$ is any complete orthonormal system in $H$. The covariance operators $\mathbf{G}$ and $\hat{\mathbf{G}}$ respectively are generated by the kernels $G$ and $\hat{G}$, i.e. $\mathbf{G}(f) = \int_{\mathcal{T}} G(s, t)\, f(s)\, \mathrm{d}s$ and $\hat{\mathbf{G}}(f) = \int_{\mathcal{T}} \hat{G}(s, t)\, f(s)\, \mathrm{d}s$.

Let $\mathcal{I}_i = \{j : \lambda_j = \lambda_i\}$ and $\mathcal{I}' = \{i : |\mathcal{I}_i| = 1\}$, where $|\mathcal{I}_i|$ denotes the number of elements in $\mathcal{I}_i$. Let $\mathbf{P}_j = \Sigma_{k \in \mathcal{I}_j} \phi_k \otimes \phi_k$ and $\hat{\mathbf{P}}_j = \Sigma_{k \in \mathcal{I}_j} \hat{\phi}_k \otimes \hat{\phi}_k$ denote the true and estimated orthogonal projection operators from $H$ to the subspace that is spanned by $\{\phi_k : k \in \mathcal{I}_j\}$. For fixed $j$, let

$$\delta_j = \tfrac{1}{2} \min\{|\lambda_l - \lambda_j| : l \notin \mathcal{I}_j\}, \tag{16}$$

and let $\mathbf{\Lambda}_{\delta_j} = \{z \in \mathcal{C} : |z - \lambda_j| = \delta_j\}$, where $\mathcal{C}$ stands for the set of complex numbers. The resolvents of $\mathbf{G}$ and $\hat{\mathbf{G}}$ respectively are denoted by $\mathbf{R}$ and $\hat{\mathbf{R}}$, i.e. $\mathbf{R}(z) = (\mathbf{G} - zI)^{-1}$ and $\hat{\mathbf{R}}(z) = (\hat{\mathbf{G}} - zI)^{-1}$. Let

$$A_{\delta_j} = \sup\{\|\mathbf{R}(z)\|_F : z \in \mathbf{\Lambda}_{\delta_j}\}. \tag{17}$$

Let $K = K(n)$ denote the numbers of leading eigenfunctions that are included to approximate $X(t)$:

$$\hat{X}_i(t) = \hat{\mu}^{(0)}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}^{(0)}\, \hat{\phi}_k^{(0)}(t),$$

suppressing the notation of the first iteration of $K$ for simplicity, i.e. $K = K^{(0)}$. Denote $\|\pi\|_\infty = \sup_{t \in \mathcal{T}}\{|\pi(t)|\}$ for an arbitrary function $\pi(\cdot)$ with support $\mathcal{T}$. We assume that the number $K$ of eigenfunctions included depends on the sample size $n$, such that, as $n \to \infty$,

(h) $K \to \infty$ and $\upsilon_n = \Sigma_{k=1}^K \delta_k A_{\delta_k} \|\phi_k\|_\infty / (n^{1/2}h_G^2 - A_{\delta_k}) \to 0$ and
(i) $\Sigma_{k=1}^K \|\phi_k\|_\infty = o(\min\{n^{1/2}h_\mu, \bar{n}^{1/2}\})$ and $\Sigma_{k=1}^K \|\phi_k\|_\infty \|\phi_k'\|_\infty = o(\bar{n})$.

Assumptions (h) and (i) describe how the number of included eigenfunctions $K$ increases when $n \to \infty$. The quantities $\delta_k$ reflect the decay of the eigenvalues of the covariance operators, whereas $A_{\delta_k}$ depend on the local properties of the covariance operator $\mathbf{G}$ around the eigenvalues $\lambda_k$. In practice, the eigenvalues usually decrease rapidly to 0; the number of included eigenfunctions $K$ is much less than $n$, i.e. $n \gg K$, which suggests that assumptions (h) and (i) can be easily fulfilled for such processes. Moreover, the process $X$ is assumed to process the property

(j) $E(\|X\|_\infty^2 + \|X'\|_\infty^2) < \infty$ and $E[\{\sup_{t \in \mathcal{T}} |X(t) - X^K(t)|\}^2] = o(n)$, where $X^K(t) = \mu(t) + \Sigma_{k=1}^K \xi_{ik}\, \phi_k(t)$.

To apply the asymptotic results for penalized spline regression that were developed for independent data in Hall and Opsomer (2005), we adopt the following notation. Recall that the independent theoretical working data $Y_{ij}^* = Y_{ij} - \Sigma_{k=1}^{\infty} \xi_{ik}\, \phi_k(t)$ can also be written as $Y_{ij}^* = \mu(t) + \varepsilon_{ij}$. Denote the penalized spline approximation of $\mu(t)$ by $\mu(t; \beta, q) = \Sigma_{l=1}^{q} \beta_l\, b_l(t)$. A typical example is the power basis of degree $p$ with $k$ knots, i.e. $b_l(t) = t^l$, for $0 \leqslant l \leqslant p$, and $b_l(t) = (t - \kappa_{l-p})_+^p$, for $p+1 \leqslant l \leqslant k$, where $q = p + k + 1$. If the theoretical working data are used, the penalized spline estimator of $\mu(t)$ is obtained by minimizing expression (2) with $Y_{ij}$ replaced by $Y_{ij}^*$, denoted by $\tilde{\mu}(t)$, whereas the estimator that is obtained by fitting the empirical working data is denoted by $\hat{\mu}(t)$. According to Hall and Opsomer (2005), the basis functions involving knots are written as continuous functions of the knots, e.g. $b(t|\kappa) = (t - \kappa)_+^p$ for a power basis, where $\kappa \in \mathcal{T}$, so that $b_l(t) = b(t|\kappa_{l-p})$ for $l \geqslant p+1$. Let $a(t)$ be the asymptotic value of the proportion of knots $\kappa_j$, $j \leqslant q - p$, which are distributed in a neighbourhood of $t \in \mathcal{T}$, as $q$ increase. The following assumptions (k)–(m) are sufficient to derive the uniform convergence of the hypothetical penalized spline estimator $\tilde{\mu}$, as shown in Hall and Opsomer (2005). Assume that

(k) the number of knots tends to $\infty$ for fixed degree $p$, as $n \to \infty$, such that $a(t)$ is bounded away from 0 and $\infty$ on $\mathcal{T}$.

For the spline basis function $b(t|\kappa)$, define a functional operator $\psi$ by letting

$$\psi(u, v) = \int_{\mathcal{T}} b(t|u)\, b(t|v)\, \mathrm{d}v$$

and taking the operator to be the functional which maps any square integrable function $\alpha$ to $\psi\alpha$, defined by

$$(\psi\alpha)(u) = \int_{\mathcal{T}} \psi(u, v)\, \alpha(v)\, \mathrm{d}t.$$

In what follows, we use the same symbol for both the operator and its 'kernel'. Let $\mu^*(t) = \mu(t) - \Sigma_{l=1}^{p} b_l(t)$, and define the function $\beta^*$ to be the solution of

$$\mu^*(t) = \int_{\mathcal{T}} \beta^*(s)\, b(t|s)\, a(s)\, \mathrm{d}s$$

for all $t \in \mathcal{T}$.

(l) $\sup_{t \in \mathcal{T}} \{\int_{\mathcal{T}} b(t|s)^2\, \mathrm{d}s\} < \infty$, the operator $\psi$ is non-singular and $\beta^*$ is square integrable, i.e. $\int_{\mathcal{T}} \beta^*(t)^2\, \mathrm{d}t < \infty$.

Let $\{\rho_j\}_{j=1,\ldots}$ and $\{\psi_j\}_{j=1,\ldots}$ be the non-decreasing eigenvalues and corresponding eigenfunctions of the operator $\psi$. We require that

(m) $\Sigma_{j=1}^{\infty} |\int_{\mathcal{T}} \beta^*(t)\, \psi_j(t)\, \mathrm{d}t| + \Sigma_{j=1}^{\infty} \sqrt{\{\rho_j \log(j)\}} < \infty$, and $\lambda^* \to 0$ sufficiently slowly as $n \to n$, such that $n^{-1/2} \Sigma_{j=1}^{\infty} \sqrt{\{\rho_j \log(j)\}}/(\rho_j + \lambda^*) \to 0$, where $\lambda^* = \lambda^*(n)$ is the smoothing parameter for obtaining $\tilde{\mu}(t)$.

Recall that $g(y; t)$ is the density function of $Y(t)$ and $g_2(y_1, y_2; t_1, t_2)$ is the density of $(Y(t_1), Y(t_2))$. Appropriate regularity assumptions will be imposed for these density functions.

We first derive a lemma that is useful to obtain uniform consistency of the mean and covariance estimates by analogy with lemma 1 in Yao *et al.* (2005). This lemma is particularly derived for the case of deterministic design points $t_{ij}$, whereas the random design was discussed in Yao *et al.* (2005). For simplicity, we address only the univariate case. The following assumptions (n)–(s) which are only required for this lemma are listed as follows. Let $\nu$ and $l$ be given integers, with $0 \leqslant \nu < l$.

(n) $(\mathrm{d}^l/\mathrm{d}t^l)\, g(y; t)$ exists and is uniformly continuous on $\Re \times \mathcal{T}$.

We say that a univariate kernel function $K_1$ is of order $(\nu, l)$, if $\int u^q\, K_1(u)\, \mathrm{d}u$ equals $(-1)^{\nu}\, \nu!$ for $q = \nu$, a non-zero constant for $q = l$ and 0 otherwise. The assumptions for the kernel function $K_1 : \Re \to \Re$ are as follows.

(o) $K_1$ is a compactly supported kernel function of order $(\nu, l)$, and $\|K_1\|^2 = \int K_1^2(u)\, \mathrm{d}u < \infty$.

The following auxiliary results provide the weak uniform convergence rate for a general form of univariate weighted averages defined below; see Bhattacharya and Müller (1993) and Yao *et al.* (2005). For a positive integer $q \geqslant 1$, let $(\psi_p)_{p=1,\ldots,q}$ be a collection of real functions $\psi_p : \Re^2 \to \Re$ which satisfy the conditions

(p) $\psi_p$ are uniformly continuous on $\mathcal{T} \times \Re$,
(q) the functions $(\mathrm{d}^l/\mathrm{d}t^l)\,\psi_p(t,x)$ exist for all arguments $(t,x)$ and are uniformly continuous on $\mathcal{T} \times \Re$ and
(r) $\sup_{t\in\mathcal{T}}\{\int \psi_p^2(t,x)\, g(x;t)\,\mathrm{d}x\,\mathrm{d}t\} < \infty$.

Bandwidths $h_\mu = h_\mu(n)$ used for one-dimensional smoothers are assumed to satisfy

(s) $h_\mu \to 0, nh_\mu^{\nu+1} \to \infty, nh_\mu^{2l+2} < \infty, \Delta_n = O\{1/(n^{1/2}h_\mu^{\nu+1})\}$ and $\max\{n_i : i = 1,\ldots,n\} \leqslant C\bar{n}$, as $n \to \infty$.

Define the weighted averages

$$\begin{aligned}\Psi_{pn} &= \Psi_{pn}(t)\\ &= \frac{1}{nh_\mu^{\nu+1}} \sum_{i=1}^{n} \frac{1}{\bar{n}} \sum_{j=1}^{n_i} \psi_p(t_{ij}, Y_{ij})\, K_1\left(\frac{t-t_{ij}}{h_\mu}\right), \qquad p = 1,\ldots,q,\end{aligned}$$

and the quantity

$$\begin{aligned}\mu_p &= \mu_p(t)\\ &= \frac{\mathrm{d}^\nu}{\mathrm{d}t^\nu} \int \psi_p(t,x)\, g(x;t)\,\mathrm{d}x, \qquad p = 1,\ldots,q.\end{aligned}$$

## A.2. Auxiliary results and proofs of main theorems

*Lemma 1.* Under assumptions (e) and (n)–(s), $\tau_{pn} = \sup_{t\in\mathcal{T}} |\Psi_{pn}(t) - \mu_p| = O_p\{1/(n^{1/2}h_\mu^{\nu+1})\}$.

This can be shown by essentially following the proof of lemma 1 in Yao *et al.* (2005), with modifications for deterministic time points $t_{ij}$ using assumptions (c) and (d).

Following the arguments that were used in the proofs of theorems 1 and 2 of Yao *et al.* (2005) with slight modifications and extending lemma 1 to a two-dimensional smoother lead to lemma 2.

*Lemma 2.* Let $h_\mu$, $h_G$ and $h_V$ be the bandwidths that are used in the local polynomial smoothing steps for $\hat{\mu}^{(0)}(t)$ in expression (15), $\hat{G}^{(0)}(s,t)$ in expression (3) and $\hat{V}^{(0)}(t)$ in expression (4). Under assumptions (a)–(c) and (e)–(h) and appropriate regularity assumptions for $g(y;t)$ and $g_2(y_1, y_2; t_1, t_2)$,

$$\begin{aligned}\sup_{t\in\mathcal{T}} |\hat{\mu}^{(0)}(t) - \mu(t)| &= O_p\left(\frac{1}{n^{1/2}h_\mu}\right),\\ \sup_{s,t\in\mathcal{T}} |\hat{G}^{(0)}(s,t) - G(s,t)| &= O_p\left(\frac{1}{n^{1/2}h_G^2}\right).\end{aligned} \tag{18}$$

Considering eigenvalues $\lambda_k$ of multiplicity 1, $\hat{\phi}_k$ can be chosen such that

$$\begin{aligned}\sup_{t\in\mathcal{T}} |\hat{\phi}_k^{(0)}(t) - \phi_k(t)| &= O_p\left(\frac{\delta_k A_{\delta_k}}{n^{1/2}h_G^2 - A_{\delta_k}}\right),\\ \hat{\lambda}_k^{(0)} - \lambda_k &= O_p\left(\frac{\delta_k A_{\delta_k}}{n^{1/2}h_G^2 - A_{\delta_k}}\right),\end{aligned} \tag{19}$$

where the $O_p(\cdot)$ terms in equations (19) hold uniformly over all $k$, and $\delta_k$ and $A_{\delta_k}$ are defined respectively by equations (16) and (17) in Appendix A.1. As a consequence of equations (18),

$$\sup_{t\in\mathcal{T}} |\hat{\sigma}^{2,(0)}(t) - \sigma^{2,(0)}(t)| = O_p\left\{\max\left(\frac{1}{n^{1/2}h_G^2}, \frac{1}{n^{1/2}h_V}\right)\right\}. \tag{20}$$

We remark that, though lemma 2 is similar to theorems 1 and 2 of Yao *et al.* (2005), the results in this paper are developed for deterministic observation times, i.e. a fixed design, whereas the results in Yao *et al.* (2005) are valid only for random observation times $t_{ij}$ that are required to be independently and identically distributed.

The uniform convergence of the hypothetical penalized spline estimator $\tilde{\mu}$ was derived in section 4.2 and the appendix of Hall and Opsomer (2005). Here we put this result in lemma 3.

*Lemma 3.* Let $\lambda^*$ be the smoothing parameter that is used for obtaining the hypothetical penalized spline estimator $\tilde{\mu}(t)$. Under assumptions (k)–(m) and appropriate regularity assumptions for $g(y;t)$,

$$\sup_{t\in\mathcal{T}}|\mu^*(t)-\mu(t)|=O_p(\omega_n), \qquad \text{where } \omega_n=\frac{1}{n^{1/2}}\sum_{j=1}^{\infty}\frac{\sqrt{\{\rho_j\log(j)\}}}{\rho_j+\lambda^*}+\sum_{j=1}^{\infty}\frac{\lambda^*|\int_{\mathcal{T}}\beta^*(t)\,\psi_j(t)\,\mathrm{d}t|}{\rho_j+\lambda^*}. \tag{21}$$

We now consider the proof of theorem 1. With $\upsilon_n$ as in assumption (h) we define the quantities $\theta_{in}$ and $\theta_n^*$ that are related to the rate of convergence of $\sup_{1\leqslant j\leqslant n_i}|Y_{ij}^*-\hat{Y}_{ij}^{*(0)}|$ as follows. Let

$$\theta_{in}=\upsilon_n\{\|X_i\|_\infty\|X_i'\|_\infty\Delta_n^*+\sum_{j=2}^{n_i}|\varepsilon_{ij}|(t_{ij}-t_{i,j-1})\}+\left(\frac{1}{n^{1/2}h_\mu}+{\Delta_n^*}^{1/2}\right)\sum_{k=1}^{K}\|\phi_k\|_\infty+\sum_{k=1}^{K}\frac{\delta_k A_{\delta_k}|\xi_{ik}|}{n^{1/2}h_G^2-A_{\delta_k}}$$
$$+\Delta_n^*\sum_{k=1}^{K}\|\phi_k\|_\infty\|\phi_k'\|_\infty(\|X_i\|_\infty+\|X_i'\|_\infty)+\sup_{t\in\mathcal{T}}|X_i(t)-X_i^K(t)|, \tag{22}$$

$$\theta_n^*=\upsilon_n+\sum_{k=1}^{K}\|\phi_k\|_\infty\left(\frac{1}{n^{1/2}h_\mu}+{\Delta_n^*}^{1/2}\right)+\Delta_n^*\sum_{k=1}^{K}\|\phi_k\|_\infty\|\phi_k'\|_\infty+n^{-1/2}\,E^{1/2}[\{\sup_{t\in\mathcal{T}}|X(t)-X^K(t)|\}^2], \tag{23}$$

where

$$X^K(t)=\mu(t)+\sum_{k=1}^{K}\xi_k\,\phi_k(t).$$

### *A.2.1. Proof of theorem 1*

We note that the observation times $t_{ij}$ for the $i$th subject are deterministic and non-decreasingly ordered. We first prove result (12). Let

$$\hat{\eta}_{ik}=\sum_{j=2}^{n_i}\{X_i(t_{ij})-\hat{\mu}^{(0)}(t_{ij})\}\,\hat{\phi}_k^{(0)}(t_{ij})(t_{ij}-t_{i,j-1}),$$
$$\tilde{\eta}_{ik}=\sum_{j=2}^{n_i}\{X_i(t_{ij})-\mu(t_{ij})\}\,\phi_k(t_{ij})(t_{ij}-t_{i,j-1}),$$
$$\hat{\tau}_{ik}=\sum_{j=2}^{n_i}\varepsilon_{ij}\,\hat{\phi}_k^{(0)}(t_{ij})(t_{ij}-t_{i,j-1}),$$
$$\tilde{\tau}_{ik}=\sum_{j=2}^{n_i}\varepsilon_{ij}\,\phi_k(t_{ij})(t_{ij}-t_{i,j-1}),$$

and obviously $\hat{\xi}_{ik}^{(0)}=\hat{\eta}_{ik}+\hat{\tau}_{ik}$. Let $\|\phi_k\|_\infty^K=\max_{1\leqslant k\leqslant K}(\|\phi_k\|_\infty)$. Note that

$$\sup_{1\leqslant k\leqslant K}|\hat{\xi}_{ik}^{(0)}-\xi_{ik}|\leqslant\sup_{1\leqslant k\leqslant K}(|\hat{\eta}_{ik}-\tilde{\eta}_{ik}|+|\tilde{\eta}_{ik}-\xi_{ik}|+|\hat{\tau}_{ik}|). \tag{24}$$

Without loss of generality, assume that $\|\phi_k\|_\infty\geqslant 1$, $\|\phi_k'\|_\infty\geqslant 1$, $\|X_i\|_\infty\geqslant 1$ and $\|X_i'\|_\infty\geqslant 1$. Then assumption (h) implies that $\tilde{\upsilon}_n=\sup_{1\leqslant k\leqslant K}\{\delta_k A_{\delta_k}/(n^{1/2}h_G^2-A_{\delta_k})\}\to 0$. Note that $\Sigma_{k=1}^K\|\phi_k\|_\infty\|\phi_k'\|_\infty/\bar{n}\to 0$ implies that $\sup_{1\leqslant k\leqslant K}(\|\phi_k\|_\infty\|\phi_k'\|_\infty\Delta_n^*)\to 0$. The first term on the right-hand side of inequality (24) is thus bounded in probability by

$$\sup_{1\leqslant k\leqslant K}\left[\sum_{j=2}^{n_i}\{|X_i(t_{ij})-\hat{\mu}^{(0)}(t_{ij})||\hat{\phi}_k^{(0)}(t_{ij})-\phi_k(t_{ij})|+|\hat{\mu}^{(0)}(t_{ij})-\mu(t_{ij})||\phi_k(t)|\}(t_{ij}-t_{i,j-1})\right]$$
$$\leqslant\left[\sum_{j=1}^{n_i}\{|X_i(t_{ij})|+|\mu(t_{ij})|+1\}^2(t_{ij}-t_{i,j-1})\right]^{1/2}\sup_{1\leqslant k\leqslant K}\left[\sum_{j=2}^{n_i}\{\hat{\phi}_k^{(0)}(t_{ij})-\phi_k(t_{ij})\}^2(t_{ij}-t_{i,j-1})\right]^{1/2}$$
$$+\left[\sum_{j=1}^{n_i}\{\hat{\mu}^{(0)}(t_{ij})-\mu(t_{ij})\}^2(t_{ij}-t_{i,j-1})\right]^{1/2}\sup_{1\leqslant k\leqslant K}\left\{\sum_{j=2}^{n_i}\phi_k^2(t_{ij})(t_{ij}-t_{i,j-1})\right\}^{1/2}$$
$$\leqslant\{c_1(\|X_i\|_{L^2}+\|X_i\|_\infty\|X_i'\|_\infty\Delta_n^*)+c_2\}\tilde{\upsilon}_n+\{1+\sup_{1\leqslant k\leqslant K}(\|\phi_k\|_\infty\|\phi_k'\|_\infty\Delta_n^*)\}\frac{1}{n^{1/2}h_\mu}\xrightarrow{\mathrm{p}}0, \tag{25}$$

where $\|X_i\|_{L^2} = \{\int_{\mathcal{T}} X_i^2(t)\,\mathrm{d}t\}^{1/2}$, for some constants $c_1$ and $c_2$ that do not depend on $i$ and $k$, given assumptions (d), (h) and (i). The second term on the right-hand side of inequality (24) has an upper bound in probability,

$$\begin{aligned}\sup_{1\leqslant k\leqslant K}|\tilde{\eta}_{ij}-\xi_{ik}| &\leqslant \sup_{1\leqslant k\leqslant K}\{\|(X_i+\mu)'\phi_k+(X_i+\mu)\phi_k'\|_\infty\,\Delta_n^*\}\\ &\leqslant \sup_{1\leqslant k\leqslant K}(\|X_i\|_\infty\,\|\phi_k\|_\infty+\|X_i'\|_\infty\,\|\phi_k\|+c_3\|\phi_k\|_\infty+c_4\|\phi_k\|_\infty)\Delta_n^*\\ &\leqslant (c_5\|X_i\|_\infty+c_6\|X_i'\|_\infty+c_7)\sup_{1\leqslant k\leqslant K}(\|\phi_k'\|_\infty\,\Delta_n^*)\xrightarrow{\mathrm{p}}0,\end{aligned}\tag{26}$$

for some constants $c_3,\ldots,c_7$ that do not depend on $i$ and $k$.

For the third term on the right-hand side of inequality (24), it is sufficient to show that

$$\sum_{k=1}^{K}|\hat{\tau}_{ik}|\,\|\phi_k\|_\infty\xrightarrow{\mathrm{p}}0.$$

Note that

$$|\hat{\tau}_{ik}|\leqslant|\tilde{\tau}_{ik}|+\sum_{j=2}^{n_i}|\varepsilon_{ij}|\,|\hat{\phi}_k^{(0)}(t_{ij})-\phi_k(t_{ij})|(t_{ij}-t_{i,j-1}).$$

We have $E(\tilde{\tau}_{ik})=0$ and

$$\begin{aligned}\mathrm{var}(\tilde{\tau}_{ik}) &= \sum_{j=2}^{n_i}\sigma^2(t_{ij})\,\phi_k^2(t_{ij})(t_{ij}-t_{i,j-1})^2\\ &\leqslant \sup_{t\in\mathcal{T}}\{\sigma^2(t)(1+2\|\phi_k\|_\infty\,\|\phi_k'\|_\infty\,\Delta_n^*)\Delta_n^*\}\\ &\leqslant 2\sup_{t\in\mathcal{T}}\{\sigma^2(t)\Delta_n^*\},\end{aligned}$$

which implies that, in probability,

$$\sum_{k=1}^{K}|\tilde{\tau}_{ik}|\,\|\phi_k\|_\infty\leqslant[2\sup_{t\in\mathcal{T}}\{\sigma^2(t)\Delta_n^*\}]^{1/2}\sum_{k=1}^{K}\|\phi_k\|_\infty\to 0$$

by assumption (i). Also observing that

$$\sum_{k=1}^{K}\sum_{j=2}^{n_i}|\varepsilon_{ij}||\hat{\phi}_k^{(0)}(t_{ij})-\phi_k(t_{ij})|(t_{ij}-t_{i,j-1})\,\|\phi_k\|_\infty\leqslant\upsilon_n\sum_{j=2}^{n_i}|\varepsilon_{ij}|(t_{ij}-t_{i,j-1}),$$

and

$$E\left\{\sum_{j=2}^{n_i}|\varepsilon_{ij}|(t_{ij}-t_{i,j-1})\right\}\leqslant|\mathcal{T}|\sup_{t\in\mathcal{T}}\{\sigma(t)\},$$

this implies that $\Sigma_{j=2}^{n_i}|\varepsilon_{ij}|(t_{ij}-t_{i,j-1})=O_p(1)$. Then we have

$$\sum_{k=1}^{K}|\hat{\tau}_{ik}|\,\|\phi_k\|_\infty\xrightarrow{\mathrm{p}}0.$$

Then result (12) follows.

To prove result (13), it is sufficient to show that

$$\sup_{t\in\mathcal{T}}\left|\sum_{k=1}^{K}\hat{\xi}_{ik}^{(0)}\,\hat{\phi}_k^{(0)}(t)-\sum_{k=1}^{\infty}\xi_{ik}\,\phi_k(t)\right|\leqslant\sup_{t\in\mathcal{T}}\left|\sum_{k=1}^{K}\{\hat{\xi}_{ik}^{(0)}\,\hat{\phi}_k^{(0)}(t)-\xi_{ik}\,\phi_k(t)\}\right|+\sup_{t\in\mathcal{T}}\left|\sum_{k=K+1}^{\infty}\xi_{ik}\,\phi_k(t)\right|\xrightarrow{\mathrm{p}}0.\tag{27}$$

The second term converging to 0 in probability is guaranteed by the Karhunen–Loève theorem, provided that $K\to\infty$ as $n\to\infty$. We now focus on the first term,

$$\sup_{t\in\mathcal{T}}\left|\sum_{k=1}^{K}\{\hat{\xi}_{ik}^{(0)}\,\hat{\phi}_k^{(0)}(t)-\xi_{ik}\,\phi_k(t)\}\right|\leqslant\sum_{k=1}^{K}|\hat{\xi}_{ik}^{(0)}-\xi_{ik}|(\|\phi_k\|_\infty+\tilde{\upsilon}_n)+\left|\sum_{k=1}^{K}\xi_{ik}\{\hat{\phi}_k^{(0)}(t)-\phi_k(t)\}\right|$$
$$\equiv Q_1(n)+Q_2(n).$$

Observing that

$$E|Q_2(n)|\leqslant\sum_{k=1}^{K}\delta_k A_{\delta_k}\,E|\xi_{ik}|/(n^{1/2}h_G^2-A_{\delta_k})\leqslant\sum_{k=1}^{K}\delta_k A_{\delta_k}\lambda_k^{1/2}/(n^{1/2}h_G^2-A_{\delta_k})$$

and $\lambda_k\to 0$, we have $E|Q_2(n)|=O(\upsilon_n)$, i.e. $Q_2(n)=O_p(\upsilon_n)$. It is easy to see that

$$Q_1(n)\leqslant 2\sum_{k=1}^{K}|\hat{\xi}_{ik}^{(0)}-\xi_{ik}|\|\phi_k\|_\infty$$

for large $n$. Note that

$$\sum_{k=1}^{K}|\hat{\xi}_{ik}^{(0)}-\xi_{ik}|\|\phi_k\|_\infty\leqslant\sum_{k=1}^{K}|\hat{\eta}_{ik}-\tilde{\eta}_{ik}|\|\phi_k\|_\infty+\sum_{k=1}^{K}|\tilde{\eta}_{ik}-\xi_{ik}|\|\phi_k\|_\infty+\sum_{k=1}^{K}|\hat{\tau}_{ik}|\|\phi_k\|_\infty. \tag{28}$$

By analogy with inequality (25), given assumptions (d), (h) and (i), the first term on the right-hand side of inequality (28) is bounded in probability by

$$\{c_1(\|X_i\|_{L^2}+\|X_i\|_\infty\|X_i'\|_\infty\Delta_n^*)+c_2\}\upsilon_n+\left(1+\sum_{k=1}^{K}\|\phi_k\|_\infty\|\phi_k'\|_\infty\Delta_n^*\right)\frac{\sum_{k=1}^{K}\|\phi_k\|_\infty}{n^{1/2}h_\mu}\xrightarrow{\text{p}}0.$$

The second term on the right-hand side of inequality (28) is bounded in probability by

$$(c_5\|X_i\|_\infty+c_6\|X_i'\|_\infty+c_7)\sum_{k=1}^{K}\|\phi_k\|_\infty\|\phi_k'\|_\infty\Delta_n^*\xrightarrow{\text{p}}0.$$

For the third term on the right-hand side of inequality (28), we have already shown that

$$\sum_{k=1}^{K}|\hat{\tau}_{ik}|\|\phi_k\|_\infty\xrightarrow{\text{p}}0.$$

From the above proof, we can see that $\sup_{1\leqslant j\leqslant n_i}|Y_{ij}^*-Y_{ij}^{*(0)}|=O_p(\theta_{in})$ where the $O_p(\cdot)$ holds uniformly over all $i$s, and $\theta_{in}$ is defined in equation (22). On the basis of theorem 1 and lemma 3, we can derive the uniform convergence of the penalized spline estimator $\hat{\mu}(t)$ and thus the covariance estimator $\hat{G}(s,t)$ obtained by expression (3). Then the uniform consistency of other model components, including eigenfunctions and eigenvalues, can be obtained in a similar manner to that in lemma 1.

### *A.2.2. Proof of theorem 2*

Recall that the hypothetical penalized spline estimator $\tilde{\mu}(t)$ is obtained by fitting the theoretical working data $Y_{ij}^*$, whereas $\hat{\mu}(t)$ is obtained by using $\hat{Y}_{ij}^{*(0)}$ as input. Let $\tilde{G}$ denote the hypothetical covariance estimator that is obtained by expression (3) using $\tilde{\mu}(t)$ as mean estimate, whereas $\hat{G}$ is obtained by expression (3) using $\hat{\mu}(t)$ as mean estimate. Since linear smoothers, including penalized spline fitting, are weighted averages, and as expression (14) implies that $\hat{Y}_{ij}^{*(0)}=Y_{ij}^*+O_p(\underline{\theta}_{in})$, where the $O_p(\cdot)$ is uniform over $j$, it follows that $\sup_{t\in\mathcal{T}}|\hat{\mu}(t)-\tilde{\mu}(t)|=O_p(\bar{\theta}_n)$ and $\sup_{s,t\in\mathcal{T}}|\hat{G}(s,t)-\tilde{G}(s,t)|=O_p(\bar{\theta}_n)$, where $\bar{\theta}_n=\Sigma_{i=1}^{n}\theta_{in}$. Observing assumption ($j$) and

$$E(\|X\|_\infty\|X'\|_\infty)\leqslant\{E(\|X\|_\infty^2)\,E(\|X'\|_\infty^2)\}^{1/2}<\infty,$$
$$E\left\{\sum_{j=2}^{n_i}|\varepsilon_{ij}|(t_{ij}-t_{i,j-1})\right\}\leqslant|\mathcal{T}|\sup_{t\in\mathcal{T}}\{\sigma(t)\}<\infty$$

and

$$E\left\{\sum_{k=1}^{K}\delta_k A_{\delta_k}|\xi_{ik}|/(n^{1/2}h_G^2-A_{\delta_k})\right\}\leqslant\sum_{k=1}^{K}\delta_k A_{\delta_k}\lambda_k^{1/2}/(n^{1/2}h_G^2-A_{\delta_k})\leqslant\upsilon_n,$$

we have that

$$\bar{\theta}_n = O_p(\theta_n^*) \xrightarrow{\mathrm{p}} 0,$$

where $\theta_n^*$ is defined in equation (23). In view of the convergence results in lemma 3, this leads to the results (14). In fact, we have the uniform convergence rate of $\hat{\mu}(t)$ and $\hat{G}(t)$ as follows:

$$\begin{aligned} \sup_{t\in\mathcal{T}} |\hat{\mu}(t) - \mu(t)| &= O_p(\omega_n + \theta_n^*), \\ \sup_{s,t\in\mathcal{T}} |\hat{G}(s,t) - G(s,t)| &= O_p\left(\omega_n + \theta_n^* + \frac{1}{n^{1/2}h_G^2}\right), \end{aligned} \tag{29}$$

where $\omega_n$ is as in expression (21), $\theta_n^*$ is as in equation (23) and $h_G$ is the bandwidth that is used for the covariance smoothing (3).

## References


Berkey, C. S., Laird, N. M., Valadian, I. and Gardner, J. (1991) Modelling adolescent blood pressure patterns and their prediction of adult pressures. *Biometrics*, **47**, 1005–1018.

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

Bhattacharya, P. K. and Müller, H. G. (1993) Asymptotics for nonparametric regression. *Sankhya* A, **55**, 420–441.

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

Brumback, B. and Rice, J. A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Am. Statist. Ass.*, **93**, 961–1006.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Inference: a Practical Information-theoretic Approach*, 2nd edn. New York: Springer.

Castro, P. E., Lawton, W. H. and Sylvestre, E. A. (1986) Principal modes of variation for processes with continuous sample curves. *Technometrics*, **28**, 329–337.

Chiou, J.-M., Müller, H. G. and Wang, J.-L. (2003) Functional quasi-likelihood regression models with smooth random effects. *J. R. Statist. Soc.* B, **65**, 405–423.

Fan, J. and Zhang, J.-T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc.* B, **62**, 303–322.

Hall, P. and Opsomer, J. D. (2005) Theory for penalised spline regression. *Biometrika*, **92**, 105–118.

James, G., Hastie, T. G. and Sugar, C. A. (2001) Principal component models for sparse functional data. *Biometrika*, **87**, 587–602.

Jones, M. C. and Rice, J. (1992) Displaying the important features of large collections of similar curves. *Am. Statistn*, **46**, 140–145.

Lin, X. and Carroll, R. J. (2000) Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Am. Statist. Ass.*, **95**, 520–534.

Lin, X., Wang, N., Welsh, A. H. and Carroll, R. J. (2004) Equivalent kernels of smoothing splines in nonparametric regression for clustered/longitudinal data. *Biometrika*, **91**, 177–193.

O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, **1**, 502–518.

Ramsay, J. and Silverman, B. (1997) *Functional Data Analysis*. New York: Springer.

Rao, C. R. (1958) Some statistical methods for comparison of growth curves. *Biometrics*, **14**, 1–17.

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.

Rice, J. and Wu, C. (2000) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.

Ruppert, D. (2002) Selecting the number of knots for penalized splines. *J. Computnl Graph. Statist.*, **11**, 735–757.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Shi, M., Weiss, R. E. and Taylor, J. M. G. (1996) An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Appl. Statist.*, **45**, 151–163.

Silverman, B. (1996) Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, **68**, 45–54.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Tyer, V. R., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molec. Biol. Cell*, **9**, 3273–3297.

Wang, N. (2003) Marginal nonparametric kernel regression accounting for within-subject correlation. *Biometrika*, **90**, 43–52.

Wang, N., Carroll, R. J. and Lin, X. (2005) Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Am. Statist. Ass.*, **100**, 147–157.

Welsh, A. H., Lin, X. and Carroll, R. J. (2002) Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *J. Am. Statist. Ass.*, **97**, 482–493.

Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A. and Vogel, J. S. (2003) Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, **59**, 676–685.

Yao, F., Müller, H. G. and Wang, J. L. (2005) Functional data analysis for sparse longitudinal data. *J. Am. Statist. Ass.*, **100**, 577–590.