

Penalized Spline Models for Functional Principal Component Analysis

Fang Yao[†] and Thomas C. M. Lee
Department of Statistics
Colorado State University
Fort Collins, CO, 80523

August 12, 2005

[†] Corresponding author, e-mail: fyao@stat.colostate.edu.

Summary

In this paper 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 amongst the repeated measurements 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 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, Lawton and Sylvestre (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), Fan and Zhang (2000), among others. For 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/longitudinal data. The main idea is to, via an iterative process, transform the original correlated data such that the resulting data are asymptotically equivalent to a set of independent data. During the iteration process, the mean function is updated with a popular smoothing technique, penalized spline regression, while the covariance surface, the variance of errors and the functional principal components 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 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 proposed procedure IPS, short for Iterative Penalized Spline fitting.

Through an analytic derivation of its asymptotic properties, IPS 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 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 model for functional data are available up to date, while kernel and smoothing spline approaches for clustered data are investigated by Lin and Carroll (2000), Wang (2003), Lin et al. (2004) and Wang, Carroll and Lin (2005). In most of these existing approaches, the covariance structure is modeled through a finite number of parameters using moment methods,

which inherits the feature of covariance estimation in classical longitudinal approaches. On the other hand 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 are also studied, via both a simulation study and an application to yeast cell cycle gene data, which suggests that IPS 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 proposed IPS procedure, together with its theoretical properties, are presented in Section 3. Simulation results that illustrate the effectiveness of the methodology are reported in Section 4. The application of IPS to yeast cell cycle gene expression data is provided in Section 5, while concluding remarks are offered in Section 6. Technical details are deferred to the appendices.

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 among repeated measurements from the same subject are ignored. Lastly, for completeness, we summarize the relevant results from Yao et al. (2003) that will be required for the rest of this article.

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 $G(s, t) = \text{cov}(X_i(s), X_i(t))$, 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} , while 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 $\{\phi_k\}_{k=1,2,\dots}$ and non-increasing eigenvalues $\{\lambda_k\}_{k=1,2,\dots}$: $G(s, t) = \sum_k \lambda_k \phi_k(s) \phi_k(t)$, $t, s \in \mathcal{T}$. Karhunen-Loève representation in the classical functional principal component analysis implies that the i th random curve can be expressed as $X_i(t) = \mu(t) + \sum_k \xi_{ik} \phi_k(t)$, $t \in \mathcal{T}$, where $\mu(t)$ is the mean function, the coefficients $\xi_{ik} = \int_{\mathcal{T}} \{X_i(t) - \mu(t)\} \phi_k(t) dt$ are uncorrelated random variables with zero mean and variances $E \xi_{ik}^2 = \lambda_k$, and $\sum_k \lambda_k < \infty$, $\lambda_1 \geq \lambda_2 \geq \dots$.

To model the noisy observations realistically, we incorporate uncorrelated measurement error ϵ_{ij} from a common distribution family with mean 0 and respective variances $\sigma^2(t_{ij})$ that may be heteroscedastic to reflect the additional noise, where $\sigma^2(t)$ is assumed to be bounded from zero and

infinity on \mathcal{T} ; i.e., $0 < \inf_{t \in \mathcal{T}} \sigma^2(t) \leq \sup_{t \in \mathcal{T}} \sigma^2(t) < \infty$. Let Y_{ij} denote the j th observation of $X_i(\cdot)$ at time t_{ij} with the additional noise ϵ_{ij} that is independent of ξ_{ik} , $i = 1, \dots, n, j = 1, \dots, n_i, k = 1, 2, \dots$, where n_i is the number of measurements made on the i th subject. Then we consider the model

$$Y_{ij} = X_i(t_{ij}) + \epsilon_{ij} = \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t_{ij}) + \epsilon_{ij}, \quad t_{ij} \in \mathcal{T}, \quad (1)$$

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

2.2 Estimation of Mean Function Using Penalized Spline Regression

As the mean function $\mu(t)$ is assumed smooth, we can estimate $\mu(t)$ using penalized regression with spline basis. Due to its flexibility to capture nonlinear relationships, efficiency in computation, and capability of providing effective inferential tools, penalized spline regression has become a popular method for estimating smooth functions (see Ruppert, Wand and Carroll, 2003). Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})^T$ and $\mathbf{T}_i = (t_{i1}, \dots, t_{in_i})^T$. Also let $B_q(t) = (B_{q1}(t), \dots, B_{qq}(t))^T$ denote the q -vector of a spline basis evaluated at time t used to model the mean function $\mu(t)$. The mean function $\mu(t)$ is thus modeled by the penalized approximation $B_q^T(t)\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_q)^T$ is the coefficient vector. Let λ^* be the smoothing parameter, and \mathbf{D} some symmetric positive semidefinite matrix. Let $\mathbf{B}_{qi} = (B_q(t_{i1}), \dots, B_q(t_{in_i}))^T$ 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 following penalized least squares criterion

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

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

A typical choice of the spline basis is the truncated power basis of degree p ; i.e., $B_q(t) = (1, t, \dots, t^p, (t - \kappa_1)_+^p, \dots, (t - \kappa_k)_+^p)^T$ with knots at $\kappa_1, \dots, \kappa_k$, which implies $q = p + k + 1$, where $x_+ = \max(0, x)$. A common choice of \mathbf{D} is the block diagonal matrix $\text{diag}(\mathbf{0}_{(p+1) \times (p+1)}, \mathbf{I}_{k \times k})$, where $\mathbf{0}_{p \times p}$ is $p \times p$ matrix with all entries equal to zero, 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, due to the within-subject correlation, chosen by one-curve-leave-out 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 κ_j in our

simulation and application example can be achieved by selecting the 10th, 20th, . . . , 90th percentiles of the pooled observation times.

Shi, Weiss and Taylor (1996), Rice and Wu (2000) and James, Hastie and Sugar (2001) study the use of B-splines with no roughness penalties to model the individual curves with random coefficients through mixed effects models. Perhaps due to the complexity of their modeling approaches, they did not investigate the asymptotic properties of the estimated components in relation to the true values, such as the behaviors of the estimated mean, covariance structure and principal components. In contrast, in our spline-based modeling 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 are able to derive asymptotic properties of our proposed procedure.

2.3 Estimation of Covariance Surface and Functional Principal Components

In this paper we adopt the procedure 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 uni- and bi-variate 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 obtained from the previous step. The local linear smoother estimate $\hat{G}(s, t)$ for $G(s, t)$ is obtained by minimizing

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

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

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

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

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

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

$$\int_{\mathcal{T}} \hat{G}(s, t) \hat{\phi}_k(s) ds = \hat{\lambda}_k \hat{\phi}_k(t), \quad (6)$$

with orthonormal constraints on $\{\hat{\phi}_k\}_{k \geq 1}$ that are unique up to a sign change, 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 $\xi_{ik} = \int \{X_i(t) - \mu_{g(i)}(t)\} \phi_k(t) dt$ are estimated by numerical integration,

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

Finally, for the selection of the number of eigenfunctions K , one could use the AIC type criterion suggested by Yao, Müller and Wang (2005). Denote $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(t_{i1}), \dots, \hat{\mu}(t_{in_i}))^T$, $\hat{\boldsymbol{\phi}}_{ik} = (\hat{\phi}_k(t_{i1}), \dots, \hat{\phi}_k(t_{in_i}))^T$, and $\Sigma_i = \text{diag}\{\hat{\sigma}^2(t_{i1}), \dots, \hat{\sigma}^2(t_{in_i})\}$. Then, if the error terms ϵ_{ij} in (1) are assumed to be normal, K is chosen by minimizing

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

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

3 Iterative Penalized Spline Fitting for Within-Subject Measurement Correlation

This section presents the proposed IPS procedure, an iterative penalized spline 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 naive application of the penalized splines (e.g., solutions to (2)) for this problem will not lead to optimal estimates when within-subject correlation is present. Although Lin and Carroll (2000) show 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, Lin and Carroll, 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) show that smoothing spline estimator has the smallest variance when the unobservable true covariance function is used. However, it is not clear if this conclusion can be extended to penalized spline regression, as penalized splines are a low-rank smoothing method while smoothing splines are a full-rank smoothing method. These considerations suggest the need for a more sophisticated penalized spline estimation method extending (2). The proposed IPS procedure is designed for handling this issue.

3.1 Iterative Penalized Spline Procedure

Hall and Opsomer (2005) show 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 amongst observations made for the same subject so that after iteration the *empirical working data* (defined in (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 $l = 0, 1, 2, \dots$:

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 obtained at the diagonal of the surface are omitted, as these are contaminated with the residual variance $\sigma^2(t)$ (see (3)).
2. Use (6) to compute estimates $\hat{\phi}_k^{(l)}$ and $\hat{\lambda}_k^{(l)}$ for, respectively, the eigenfunctions and eigenvalues.
3. By using the empirical variances obtained on the diagonal of covariance surface, estimate the variance function $\sigma^2(t)$ by (5).
4. Use the integration approximation (7) to obtain estimate $\hat{\xi}_{ik}^{(l)}$ for the individual functional principal component scores.
5. For all i and j , define the *theoretical working data* as $Y_{ij}^* = Y_{ij} - \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t_{ij})$. Note that the Y_{ij}^* 's are independent, and estimate them by the following *empirical working data*

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

where $K^{(l)}$ is the number of eigenfunctions, chosen by the AIC criterion (8), used for approximation in the current iteration.

6. In (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 differences (RISD) between $\hat{\mu}^{(l)}$ and $\hat{\mu}^{(l+1)}$ is less than a pre-specified tolerance:

$$\text{RISD}_l = \int_{\mathcal{T}} [\hat{\mu}^{(l+1)}(t) - \hat{\mu}^{(l)}(t)]^2 dt / \int_{\mathcal{T}} [\hat{\mu}^{(l)}(t)]^2 dt. \quad (10)$$

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 (e.g., see (15)). For both cases the amount of smoothing is chosen by one-curve-leave-out cross-validation or its generalized version.

Remark 1: It is easy to extend the proposed approach to sparse functional data; i.e., when the number of repeated measurements available per subject is small. For sparse data, Yao et al. (2005) demonstrate that the best linear prediction, denoted by $\hat{\xi}_{ik}^P$, of ξ_{ik} given the data from the subject $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{in_i})$ outperforms the traditional integration approximation (7). Yao et al. (2005) termed $\hat{\xi}_{ik}^P$ the PACE estimate, short for Principal Component Analysis through Conditional Expectation. 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 PACE estimates $\hat{\xi}_{ik}^P$. Although the theory of coupling IPS with PACE has not been developed, simulation results to be reported in the next section demonstrate the promising practical performance of IPS with PACE.

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., $\mathbf{a}^T \boldsymbol{\beta}$ for any $\mathbf{a} \in \mathbb{R}^q$. First from (2) with \mathbf{Y}_i replaced by $\mathbf{Y}_i^* = (Y_{i1}^*, \dots, Y_{in_i}^*)^T$, it is straightforward to obtain the following closed form expression for $\hat{\boldsymbol{\beta}}$: $\hat{\boldsymbol{\beta}} = (\sum_{i=1}^n \mathbf{B}_{qi}^T \mathbf{B}_{qi} + \lambda^* \mathbf{D})^{-1} \sum_{i=1}^n \mathbf{B}_{qi} \mathbf{Y}_i^*$. Also, $\text{cov}(Y_{ij}^*, Y_{il}^*) = \delta_{jl} \sigma^2(t_{ij})$, where $\sigma^2(\cdot)$ is the variance function of measurement error, and $\delta_{kl} = 1$ for $k = l$ and 0 otherwise. Denote $\mathbf{R}_i = \text{diag}\{\sigma^2(t_{i1}), \dots, \sigma^2(t_{in_i})\}$. Direct calculations lead to the following covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}}$ for $\hat{\boldsymbol{\beta}}$

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

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

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

where $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is calculated by plugging in the corresponding estimates 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; e.g., see Rice and Wu (2000), and Chiou, Müller and Wang (2003).

3.2 Theoretical Properties of IPS

We have studied the theoretical properties of the proposed IPS procedure, 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 only consider the case of one-step iteration. In the sequel $g(x; t)$ 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 (15) in the appendices, and we expect that the 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 $\{\widehat{Y}_{ij}^{*(0)}\}$.

The central results towards the theoretical analysis of the empirical working data are presented in Theorem 1 that 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 (A1.1)-(A7), and appropriate regularity assumptions for $g(x; t)$ and $g_2(x_1, x_2; t_1, t_2)$,*

$$\sup_{1 \leq k \leq K} |\hat{\xi}_{ik}^{(0)} - \xi_{ik}| \xrightarrow{P} 0, \quad (12)$$

$$\sup_{1 \leq j \leq n_i} |\widehat{Y}_{ij}^{*(0)} - Y_{ij}^*| \xrightarrow{P} 0. \quad (13)$$

From Theorem 1 we conclude that the empirical working data $\{\widehat{Y}_{ij}^{*(0)}\}$ are asymptotically equivalent to their theoretical counterparts $\{Y_{ij}^*\}$, and in fact $\sup_{1 \leq j \leq n_i} |\widehat{Y}_{ij}^{*(0)} - Y_{ij}^*| = O_p(\theta_{in})$ where θ_{in} is defined in (22) and the $O_p(\cdot)$ term holds uniformly over all i 's. Since these theoretical working data $\{Y_{ij}^*\}$'s are independent, by applying penalized spline smoothing to the empirical working data $\widehat{Y}_{ij}^{*(0)}$'s, 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 \widehat{G} that is obtained as in Step 1. The central results are provided in Theorem 2 below.

Theorem 2 *Under (A1.1)-(A10), and appropriate regularity assumptions for $g(x; t)$ and $g_2(x_1, x_2; t_1, t_2)$,*

$$\sup_{t \in \mathcal{T}} |\hat{\mu}(t) - \mu(t)| \xrightarrow{P} 0, \quad \sup_{s, t \in \mathcal{T}} |\widehat{G}(s, t) - G(s, t)| \xrightarrow{P} 0. \quad (14)$$

Remark 1: We remark that the uniform convergence rate for the penalized spline estimator $\hat{\mu}$ is in fact obtained as $O_p(\omega_n + \theta_n^*)$, where ω_n and θ_n^* are as defined in (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 $O_p\{\omega_n + \theta_n^* + 1/(\sqrt{n}h_G^2)\}$, where h_G is the bandwidth used in (3). Based on Theorem 2, the asymptotic consistency for other model components, such as principal components, can also be obtained by analogy to Lemma 1.

Remark 2: One can see that the iterative approach proposed here is also applicable for other smoothing methods, not only restricted to penalized spline regression. For example, one could apply local polynomial smoothing or smoothing spline methods to the empirical working data in each iteration. Based on the established theory for those smoothing methods, the theoretical arguments for the iterative approach still hold, and similar consistency results can be obtained. Due to the computational efficiency, we focus on the penalized spline models, though the theoretical development for penalized splines is more challenging.

4 Simulation Studies

In order to assess the practical performance of the proposed IPS procedure, a simulation study was conducted. We generated 100 independently and identically distributed (i.i.d.) normal and 100 i.i.d. non-normal samples consisting of $n = 100$ random trajectories respectively. The simulated processes have a mean function $\mu(t) = t + \sin(t)$, $0 \leq t \leq 10$, and covariance function derived from two eigenfunctions $\phi_1(t) = -\cos(\pi t/10)/\sqrt{5}$, and $\phi_2(t) = \sin(\pi t/10)/\sqrt{5}$, $0 \leq t \leq 10$. We chose $\lambda_1 = 4$, $\lambda_2 = 1$ and $\lambda_k = 0$, $k \geq 3$, as eigenvalues, and $\sigma^2(t) \equiv 0.25$ as variance of the additional measurement errors ϵ_{ij} in (1), which are assumed to be normal with mean 0. For the 100 normal samples, the functional principal component scores ξ_{ik} were generated from $\mathcal{N}(0, \lambda_k)$, while the ξ_{ik} for the non-normal samples were generated from a mixture of two normals, $\mathcal{N}(\sqrt{\lambda_k/2}, \lambda_k/2)$ with probability 1/2 and $\mathcal{N}(-\sqrt{\lambda_k/2}, \lambda_k/2)$ with probability 1/2.

We also consider both sparse and non-sparse designs. For an equally spaced grid $\{c_0, \dots, c_{50}\}$ on $[0, 10]$ with $c_0 = 0$, $c_{50} = 10$, let $s_i = c_i + e_i$, where e_i are i.i.d. 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 sparse design, each curve was sampled at a random number of points, chosen from a discrete uniform distribution on $\{3, \dots, 6\}$, and the locations of the measurements were randomly chosen from $\{s_1, \dots, s_{49}\}$ without replacement, while for non-sparse design, the number of observations for each curve was randomly chosen from $\{20, \dots, 30\}$.

The following four different methods were compared.

M1 : The mean function $\mu(t)$ is estimated using the penalized spline model (2), and the covariance

and principal components are estimated by the method described in Section 2.3. Note that no iteration is performed.

M2 : Similar to M1, but the mean function $\mu(t)$ is estimated with local polynomial smoothing (15).

M3 : the proposed IPS procedure where the initial group mean estimates $\hat{\mu}^{(0)}$ are obtained by local polynomial smoothing (15).

M4 : The proposed IPS procedure where the initial group mean estimates $\hat{\mu}^{(0)}$ are obtained by the penalized spline model (2).

In the above, for all the local polynomial smoothing steps, either the univariate or the bivariate Epanechnikov kernel functions were used; i.e., $K_1(x) = 3/4(1-x^2)\mathbf{1}_{[-1,1]}(x)$ and $K_2(x, y) = 9/16(1-x^2)(1-y^2)\mathbf{1}_{[-1,1]}(x)\mathbf{1}_{[-1,1]}(y)$, where $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise for any set A , and for penalized regression, cubic spline basis was used; i.e., $p = 3$.

To demonstrate the superior performances of the proposed IPS procedure (methods M3 and M4) compared to non-iterative methods (M1 and M2), we report in Table 1 the Monte Carlo estimates obtained from 100 non-sparse/sparse & Normal/Mixture simulated datasets (in total 400 datasets) for integrated mean squared error (IMSE) of $\hat{\mu}(t)$ that consists of integrated squared bias (IBIAS) and integrated variance (IVAR); i.e., $\int_0^{10} E[\{\hat{\mu}(t) - \mu(t)\}^2]dt = \int_0^{10} \{\hat{\mu}(t) - E\hat{\mu}(t)\}^2 dt + \int_0^{10} \{E\hat{\mu}(t) - \mu(t)\}^2 dt$. Recall that the predicted individual trajectories using K eigenfunctions are denoted by $\hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t)$, where $\hat{\xi}_{ik}$ are obtained either by the integration method (7) for non-sparse data or by the PACE method for sparse data.

To avoid the bias in comparison possibly caused by inadequate choices of tuning parameters, such as λ^* , K , bandwidths and knots, we constructed two scenarios. First, the methods M1-M4 are compared at optimal tuning parameter values, denoted by ‘‘Optimal’’. Specifically, the optimal bandwidth h_μ for $\hat{\mu}^{(0)}(t)$ used in M2 and M3 (initial estimate in M3) is chosen by minimizing the L^2 distance between estimated and true mean functions; i.e., $\int_0^{10} \{\hat{\mu}^{(0)}(t; h_\mu) - \mu(t)\}^2 dt$. Other optimal smoothing parameters, including h_G , h_V , λ^* used in M1-M4 for the covariance function of $X(t)$, the variance function of $\epsilon(t)$ and the penalized spline estimate of $\mu(t)$, are also chosen by minimizing corresponding L^2 distances respectively. 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 ten-fold cross-validation to choose h_μ , h_G and h_V , which involving removing 10% of the individual curves as a test set, finding the estimates from the remaining data, and repeating the process nine more times, while λ^* was chosen by ten-fold generalized cross-validation. The number of eigenfunctions K in each run was chosen by the AIC criterion (8). Since Ruppert (2002) showed that the penalized spline estimators are relatively insensitive to the choice of basis functions as compared to the choice of λ^* , 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 M1-M4 for both scenarios.

From Table 1, one can see that the IPS procedures (M3 and M4) improved the integrated mean squared errors (IMSE) of mean estimates over non-iterative procedures M1 and M2 by around 25%-40% for sparse samples and 15%-35% for non-sparse samples in both Optimal and Model-Selected scenarios, while it is not surprising that the IMSE's obtained in Optimal scenario are slightly smaller than those in Model-Selected scenario. The procedures are "robust" regarding the distribution of random components ξ_{ik} , yielding similar amount of improvement for 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 PACE estimates as compared to non-iterative procedures. It is interesting that the improvement obtained by the IPS procedures for sparse samples is in fact more dramatic than that obtained for non-sparse samples, which suggests further investigation of such 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 criterion (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/sparse and Normal/Mixture). Regarding computational efficiency, the proposed IPS procedures (M3 and M4) usually converge very fast, with no more than 4 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 and 10 minutes respectively.

We also compared the Monte Carlo estimates of the integrated squared prediction error (IPE) of the true curves X_i obtained by M1-M4 from those simulated samples; i.e., $\text{IPE} = \sum_{i=1}^n \int_0^{10} \{X_i(t) - \hat{X}_i^K(t)\}^2 dt/n$, where $\hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^K \hat{\xi}_{ik} \hat{\phi}_k(t)$, reported in Table 1. It is seen that the IPS procedures (M3 and M4) improve the prediction errors by around 10%, and M3 and M4 gave comparable results. This is consistent with the previous observations obtained from comparing IMSE of the mean estimates regarding the superior performance of the proposed IPS procedures.

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 (α factor-based) by a chemical process. There are 6178 genes in total, and each gene expression profile consists of 18 data points, measured every seven minutes between 0 and 119 minutes, 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. For the purpose of demonstrating the proposed method, the 43 genes related to the G1 phase are used in the following analysis; see Figure 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 the right panel of Figure 1. The first mean estimate was obtained using the proposed IPS procedure with initial estimates given by local polynomial smoothing (15), Using (11) and the penalized spline approximation $\mu(t) \approx B_q^T(t)\boldsymbol{\beta}$, 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_μ and the smoothing parameter λ^* were chosen by one-curve-leave-out cross-validation and its generalized version. The iterative procedure converged in 3 iterations with tolerance (10) set to 10^{-4} . One finds that, when compared to the traditional non-iterative method, the mean function 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 minutes in the yeast cell cycle which can also be seen from the original data displayed in the left panel of Figure 1. This is detected by the mean estimate obtained using the proposed method, while the estimate obtained by the traditional approach with no iteration does not provide any information of this feature.

The smooth covariance surface estimate obtained using the proposed IPS procedure is displayed in the left panel of Figure 2, where the bandwidth h_G is chosen by one-curve-leave-out cross-validation in each iteration, as well as h_V as in (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 criterion (8) to approximate the expression profiles (right panel of Figure 2). 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 $\widehat{X}_i(t) = \widehat{\mu}(t) + \sum_{k=1}^K \widehat{\xi}_{ik} \widehat{\phi}_k(t)$, where $\widehat{\xi}_{ik}$ are as in (7), in Figure 3. The predicted trajectories are obtained using IPS with initial mean estimated by local polynomial smoothing (15), and also the traditional method using local polynomial estimation for mean and covariance as described in Step 1 to 4 of Section 3.1. One finds that the proposed IPS fitting provides better prediction compared to 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 $\text{MPE} = (1/n) \sum_{i=1}^n \sum_{j=1}^{n_i} \{Y_{ij} - \widehat{Y}_i(t_{ij})\}^2/n_i$. The proposed IPS procedure gives $\text{MPE}=0.120$, while the traditional approach yields $\text{MPE}=0.138$, which indicates about 15% reduction. From the above evidence, we conclude that the proposed iterative IPS procedure indeed improves upon the traditional non-iterative approach for the modeling of functional data.

6 Concluding Remarks

In this paper a new method for performing functional principal component analysis that uses penalized spline regression is presented. For the purpose of reducing the within-subject correlation commonly found in functional/longitudinal data, an iterative estimation procedure is proposed to improve the estimation of the mean function. Through an analytic derivation of its asymptotic properties, IPS is 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 obtained by simulations and the real data example, one can see that, when comparing to traditional non-iterative methods, significant improvement can be achieved by the proposed approach. Another attractive property of the proposed method is that it allows simple covariate incorporation and straightforward approximate inference.

Acknowledgement

The authors are most grateful to the reviewers and the associate editor for their most constructive comments. The work of Lee was supported in part by U.S. National Science Foundation grants DMS-0203901.

Appendix

A.1 Assumptions and Notations

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

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

with respect to β_0 and β_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 $h_\mu = h_\mu(n)$, $h_G = h_G(n)$, $h_V = h_V(n)$ are the bandwidths for estimating $\hat{\mu}^{(0)}$ in (15), $\hat{G}^{(0)}$ in (3) and $\hat{V}^{(0)}$ in (4). We develop asymptotics as the number of subjects $n \rightarrow \infty$, and require

$$(A1.1) \quad h_\mu \rightarrow 0, h_V \rightarrow 0, nh_\mu^4 \rightarrow \infty, nh_V^4 \rightarrow \infty, nh_\mu^6 < \infty, \text{ and } nh_V^6 < \infty.$$

$$(A1.2) \quad h_G \rightarrow 0, nh_G^6 \rightarrow \infty, \text{ and } nh_G^8 < \infty.$$

The time points $\{t_{ij}\}_{i=1,\dots,n;j=1,\dots,n_i}$ here are considered deterministic. Denote the sorted time points across all subjects as $a_X \leq t_{(1)} \leq \dots \leq t_{(N_n)} \leq b_X$, and $\Delta_n = \max\{t_{(k)} - t_{(k-1)} : k = 1, \dots, N+1\}$, where $N_n = \sum_{i=1}^n n_i$, $\mathcal{T} = [a_X, b_X]$, $t_{(0)} = a_X$, and $t_{(N+1)} = b_X$. For the i th subject, suppose that the time points t_{ij} have been ordered non-decreasingly. Let $\Delta_{in} = \max\{t_{ij} - t_{i,j-1} : j = 1, \dots, n_i+1\}$ and $\Delta_n^* = \max\{\Delta_{in} : i = 1, \dots, n\}$, where $t_{i0} = a_X$ and $t_{i,n_i+1} = b_X$. Also denote $\bar{n} = n^{-1} \sum_{i=1}^n n_i$. To obtain the 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

$$(A2.1) \quad \Delta_n = O(\min\{n^{-1/2}h_\mu^{-1}, n^{-1/2}h_V^{-1}, n^{-1/4}h_G^{-1}\}).$$

$$(A2.2) \quad \bar{n} \rightarrow \infty, \max\{n_i : i = 1, \dots, n\} \leq C\bar{n} \text{ for some } C > 0, \text{ and } \Delta_n^* = O(1/\bar{n}), \text{ as } n \rightarrow \infty.$$

Fourier transforms of $K_1(u)$ and $K_2(u, v)$ are denoted by $\kappa_1(t) = \int e^{-iut} K_1(u) du$ and $\kappa_2(t, s) = \int \int e^{-(iut+ivs)} K_2(u, v) du dv$ respectively. They satisfy

$$(A3.1) \quad \kappa_1(t) \text{ is absolutely integrable; i.e., } \int |\kappa_1(t)| dt < \infty.$$

$$(A3.2) \quad \kappa_2(t, s) \text{ is absolutely integrable; i.e., } \int \int |\kappa_2(t, s)| dt ds < \infty.$$

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

$$(A4) \quad \sup_{t \in \mathcal{T}} E[Y^4(t)] < \infty.$$

Define the rank one 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 $\langle T_1, T_2 \rangle_F = \text{tr}(T_1 T_2^*) = \sum_j \langle T_1 u_j, T_2 u_j \rangle_H$ and $\|T\|_F^2 = \langle T, T \rangle_F$, where $T_1, T_2, T \in F$, and $\{u_j : j \geq 1\}$ is any complete orthonormal system in H . The covariance operator \mathbf{G} (respectively, $\widehat{\mathbf{G}}$) is generated by the kernel G (respectively, \widehat{G}); i.e., $\mathbf{G}(f) = \int_{\mathcal{T}} G(s, t) f(s) ds$, $\widehat{\mathbf{G}}(f) = \int_{\mathcal{T}} \widehat{G}(s, t) f(s) ds$.

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

$$\delta_j = \frac{1}{2} \min\{|\lambda_l - \lambda_j| : l \notin \mathcal{I}_j\}, \quad (16)$$

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 resolvent of \mathbf{G} (respectively, $\widehat{\mathbf{G}}$) is denoted by \mathbf{R} (respectively, $\widehat{\mathbf{R}}$); i.e., $\mathbf{R}(z) = (\mathbf{G} - zI)^{-1}$ (respectively, $\widehat{\mathbf{R}}(z) = (\widehat{\mathbf{G}} - zI)^{-1}$). Let

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

Let $K = K(n)$ denote the numbers of leading eigenfunctions included to approximate $X(t)$; i.e., $\hat{X}_i(t) = \hat{\mu}^{(0)}(t) + \sum_{k=1}^K \hat{\xi}_{ik}^{(0)} \hat{\phi}_k^{(0)}(t)$, 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 included eigenfunctions depends on the sample size n , such that, as $n \rightarrow \infty$,

$$(A5) \quad K \rightarrow \infty, \text{ and } v_n = \sum_{k=1}^K \delta_k A_{\delta_k} \|\phi_k\|_\infty / (\sqrt{nh_G^2} - A_{\delta_k}) \rightarrow 0.$$

$$(A6) \quad \sum_{k=1}^K \|\phi_k\|_\infty = o(\min\{\sqrt{nh_\mu}, \sqrt{\bar{n}}\}), \text{ and } \sum_{k=1}^K \|\phi_k\|_\infty \|\phi_k'\|_\infty = o(\bar{n}).$$

The assumptions (A5) and (A6) describe how the number of included eigenfunctions K increases when n tends to infinity. The quantities δ_k reflects the decay of the eigenvalues of the covariance operators, while A_{δ_k} depend on the local properties of the covariance operator \mathbf{G} around the eigenvalues λ_k . In practice, the eigenvalues usually decrease rapidly to zero, the number of included eigenfunctions K is much less than n ; i.e., $n \gg K$, which suggests the assumptions (A5) and (A6) can be easily fulfilled for such processes. Moreover, the process X is assumed to process the following property,

$$(A7) \quad E(\|X\|_\infty^2 + \|X'\|_\infty^2) < \infty, \quad E[\{\sup_{t \in \mathcal{T}} |X(t) - X^K(t)|\}^2] = o(n), \text{ where } X^K(t) = \mu(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t).$$

To apply the asymptotic results for penalized spline regression developed for independent data in Hall and Opsomer (2005), we adopt the following notations. Recall that the independent theoretical pseudo-data $Y_{ij}^* = Y_{ij} - \sum_{k=1}^\infty \xi_{ik} \phi_k(t)$ can also be written as $Y_{ij}^* = \mu(t) + \epsilon_{ij}$. Denote the penalized spline approximation of $\mu(t)$ by $\mu(t; \beta, q) = \sum_{\ell=1}^q \beta_\ell b_\ell(t)$, A typical example is the power basis of degree p with k knots; i.e., $b_\ell(t) = t^\ell$, for $0 \leq \ell \leq p$, and $b_\ell(t) = (t - \kappa_{\ell-p})_+^p$, for $p+1 \leq \ell \leq k$, where $q = p+k+1$. If the theoretical pseudo-data are used, the penalized spline estimator of $\mu(t)$ is obtained by minimizing (2) with Y_{ij} replaced by Y_{ij}^* , denoted by $\tilde{\mu}(t)$, while the estimator obtained by fitting the empirical pseudo-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 power basis, where $\kappa \in \mathcal{T}$, so that $b_\ell(t) = b(t|\kappa_{\ell-p})$ for $\ell \geq p+1$. Let $a(t)$ be the asymptotic value of the proportion of knots κ_j , $j \leq q-p$, which are distributed in a neighborhood of $t \in \mathcal{T}$, as q increase. The following assumptions (A8)-(A10) are sufficient to derive the uniform convergence of the hypothetical penalized spline estimator $\tilde{\mu}$, as shown in Hall and Opsomer (2005). Assume that

$$(A8) \quad \text{The number of knots tends to infinity for fixed degree } p, \text{ as } n \rightarrow \infty, \text{ such that } a(t) \text{ is bounded away from zero and infinity on } \mathcal{T}.$$

For the spline basis function $b(t|\kappa)$, define a functional operator ψ by letting $\psi(u, v) = \int_{\mathcal{T}} b(t|u)b(t|v)dv$ and taking the operator to be the functional which maps any square-integrable function α to $\psi\alpha$,

defined by $(\psi\alpha)(u) = \int_{\mathcal{T}} \psi(u, v)\alpha(v)dt$. In what follows, we use the same symbol for both the operator and its “kernel”. Let $\mu^*(t) = \mu(t) - \sum_{\ell=1}^p b_{\ell}(t)$, and define the function β^* to be solution of $\mu^*(t) = \int_{\mathcal{T}} \beta^*(s)b(t|s)a(s)ds$ for all $t \in \mathcal{T}$.

(A9) $\sup_{t \in \mathcal{T}} \int_{\mathcal{T}} b(t|s)^2 ds < \infty$, the operator ψ is nonsingular, and β^* is square integrable; i.e., $\int_{\mathcal{T}} \beta^*(t)^2 dt < \infty$.

Let $\{\rho_j\}_{j=1, \dots}$ and $\{\psi_j\}_{j=1, \dots}$ be the nondecreasing eigenvalues and corresponding eigenfunctions of the operator ψ . One requires

(A10) $\sum_{j=1}^{\infty} |\int_{\mathcal{T}} \beta^*(t)\psi_j(t)dt| + \sum_{j=1}^{\infty} \sqrt{\rho_j \log j} < \infty$, and $\lambda^* \rightarrow 0$ sufficiently slowly as $n \rightarrow n$, such that $n^{-1/2} \sum_{j=1}^{\infty} \sqrt{\rho_j \log j} / (\rho_j + \lambda^*) \rightarrow 0$, 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 in analogy to Lemma 1 in Yao et al. (2005). This lemma is particularly derived for the case of deterministic design points t_{ij} , while the random design was discussed in Yao et al. (2005). For simplicity, we only address the univariate case. The assumptions (B1)-(B4) only required for this lemma are listed as follows. Let ν, ℓ be given integers, with $0 \leq \nu < \ell$.

(B1) $(d^{\ell}/dt^{\ell})g(y; t)$ exists and is uniformly continuous on $\mathfrak{R} \times \mathcal{T}$;

We say that a univariate kernel function K_1 is of order (ν, ℓ) , if $\int u^q K_1(u)du$ equals $(-1)^{\nu} \nu!$ for $q = \nu$, a nonzero constant for $q = \ell$, and 0 otherwise. The assumptions for the kernel function $K_1 : \mathfrak{R} \rightarrow \mathfrak{R}$ are as follows,

(B2) K_1 is compactly supported kernel function of order (ν, ℓ) , and $\|K_1\|^2 = \int K_1^2(u)du < \infty$.

The following auxiliary results provide the weak uniform convergence rate for a general form of univariate weighted averages defined below, compare Bhattacharya and Müller (1993) and Yao et al. (2005). For a positive integer $q \geq 1$, let $(\psi_p)_{p=1, \dots, q}$ be a collection of real functions $\psi_p : \mathfrak{R}^2 \rightarrow \mathfrak{R}$, which satisfy:

(B3.1) ψ_p are uniformly continuous on $\mathcal{T} \times \mathfrak{R}$;

(B3.2) The functions $(d^{\ell}/dt^{\ell})\psi_p(t, x)$ exist for all arguments (t, x) and are uniformly continuous on $\mathcal{T} \times \mathfrak{R}$;

$$(B3.3) \quad \sup_{t \in \mathcal{T}} \int \psi_p^2(t, x) g(x; t) dx dt < \infty.$$

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

$$(B4) \quad h_\mu \rightarrow 0, nh_\mu^{\nu+1} \rightarrow \infty, nh_\mu^{2\ell+2} < \infty, \Delta_n = O(1/(\sqrt{n}h_\mu^{\nu+1})), \text{ and } \max\{n_i : i = 1, \dots, n\} \leq C\bar{n},$$

as $n \rightarrow \infty$.

Define the weighted averages

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

and the quantity

$$\mu_p = \mu_p(t) = \frac{d^\nu}{dt^\nu} \int \psi_p(t, x) g(x; t) dx, \quad p = 1, \dots, q.$$

A.2 Auxiliary Results and Proofs of Main Theorems

Lemma 1 Under (A3.1) and (B1)-(B4), $\tau_{pn} = \sup_{t \in \mathcal{T}} |\Psi_{pn}(t) - \mu_p| = O_p(1/(\sqrt{n}h_\mu^{\nu+1}))$.

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 (A2).

Following the arguments used in the proofs of Theorems 1 and 2 of Yao et al. (2005) with slight modifications and extending Lemma 1 to two-dimensional smoother, leads to Lemma 2.

Lemma 2 Let h_μ , h_G and h_V be the bandwidths used in the local polynomial smoothing steps for $\hat{\mu}^{(0)}(t)$ in (15), $\hat{G}^{(0)}(s, t)$ in (3) and $\hat{V}^{(0)}(t)$ in (4). Under (A1.1)-(A2.1), (A3.1)-(A5), and appropriate regularity assumptions for $g(y; t)$ and $g_2(y_1, y_2; t_1, t_2)$,

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

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

$$\sup_{t \in \mathcal{T}} |\hat{\phi}_k^{(0)}(t) - \phi_k(t)| = O_p\left(\frac{\delta_k A_{\delta_k}}{\sqrt{n}h_G^2 - A_{\delta_k}}\right), \quad \hat{\lambda}_k^{(0)} - \lambda_k = O_p\left(\frac{\delta_k A_{\delta_k}}{\sqrt{n}h_G^2 - A_{\delta_k}}\right), \quad (19)$$

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

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

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., fixed design, while the results in Yao et al. (2005) are only valid for random observation times t_{ij} that are required to be i.i.d..

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 λ^* be the smoothing parameter used for obtaining the hypothetical penalized spline estimator $\tilde{\mu}(t)$. Under (A8)-(A10), and appropriate regularity assumptions for $g(y; t)$,*

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

We now consider the proof of Theorem 1. With v_n as in (A5), we define the quantities θ_{in} and θ_n^* that are related to the rate of convergence of $\sup_{1 \leq j \leq n_i} |Y_{ij}^* - \hat{Y}_{ij}^{*(0)}|$ as follows. Let

$$\begin{aligned} \theta_{in} &= v_n \left\{ \|X_i\|_{\infty} \|X_i'\|_{\infty} \Delta_n^* + \sum_{j=2}^{n_i} |\epsilon_{ij}| (t_{ij} - t_{i,j-1}) \right\} + \left(\frac{1}{\sqrt{n} h_{\mu}} + \sqrt{\Delta_n^*} \right) \sum_{k=1}^K \|\phi_k\|_{\infty} \\ &+ \sum_{k=1}^K \frac{\delta_k A_{\delta_k} |\xi_{ik}|}{\sqrt{n} h_G^2 - A_{\delta_k}} + \Delta_n^* \sum_{k=1}^K \|\phi_k\|_{\infty} \|\phi_k'\|_{\infty} (\|X_i\|_{\infty} + \|X_i'\|_{\infty}) + \sup_{t \in \mathcal{T}} |X_i(t) - X_i^K(t)|, \end{aligned} \quad (22)$$

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

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

Proof of Theorem 1. One notes that the observation times t_{ij} for the i th subject are deterministic and non-decreasingly ordered. We first prove (12). Let

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

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

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

Without loss of generality, assume $\|\phi_k\|_{\infty} \geq 1$, $\|\phi_k'\|_{\infty} \geq 1$, $\|X_i\|_{\infty} \geq 1$ and $\|X_i'\|_{\infty} \geq 1$. Then (A5) implies $\tilde{v}_n = \sup_{1 \leq k \leq K} \delta_k A_{\delta_k} / (\sqrt{n} h_G^2 - A_{\delta_k}) \rightarrow 0$. Note that $\sum_{k=1}^K \|\phi_k\|_{\infty} \|\phi_k'\|_{\infty} / \bar{n} \rightarrow 0$

implies $\sup_{1 \leq k \leq K} \|\phi_k\|_\infty \|\phi'_k\|_\infty \Delta_n^* \rightarrow 0$. The first term in the right hand side (r.h.s.) of (24) is thus bounded in probability by,

$$\begin{aligned}
& \sup_{1 \leq k \leq K} \left\{ \sum_{j=2}^{n_i} [|X_i(t_{ij}) - \hat{\mu}^{(0)}(t_{ij})| \cdot |\hat{\phi}_k^{(0)}(t_{ij}) - \phi_k(t_{ij})| + |\hat{\mu}^{(0)}(t_{ij}) - \mu(t_{ij})| \cdot |\phi_k(t_{ij})|] (t_{ij} - t_{i,j-1}) \right\} \\
& \leq \left\{ \sum_{j=1}^{n_i} [|X_i(t_{ij})| + |\mu(t_{ij})| + 1]^2 (t_{ij} - t_{i,j-1}) \right\}^{1/2} \sup_{1 \leq k \leq K} \left\{ \sum_{j=2}^{n_i} [\hat{\phi}_k^{(0)}(t_{ij}) - \phi_k(t_{ij})]^2 (t_{ij} - t_{i,j-1}) \right\}^{1/2} \\
& \quad + \left\{ \sum_{j=1}^{n_i} [\hat{\mu}^{(0)}(t_{ij}) - \mu(t_{ij})]^2 (t_{ij} - t_{i,j-1}) \right\}^{1/2} \sup_{1 \leq k \leq K} \left\{ \sum_{j=2}^{n_i} \phi_k^2(t_{ij}) (t_{ij} - t_{i,j-1}) \right\}^{1/2} \\
& \leq [c_1 (\|X_i\|_{L^2} + \|X_i\|_\infty \|X'_i\|_\infty \Delta_n^*) + c_2] \tilde{v}_n + (1 + \sup_{1 \leq k \leq K} \|\phi_k\|_\infty \|\phi'_k\|_\infty \Delta_n^*) \frac{1}{\sqrt{n} h_\mu} \xrightarrow{p} 0, \quad (25)
\end{aligned}$$

where $\|X_i\|_{L^2} = \sqrt{\int_{\mathcal{T}} X_i^2(t) dt}$, for some constants c_1 and c_2 that do not depend on i and k , given (A2.2), (A5), (A6). The second term in the r.h.s. of (24) has an upper bound in probability,

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

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

For the third term in the r.h.s. of (24), it is sufficient to show $\sum_{k=1}^K |\hat{\tau}_{ik}| \cdot \|\phi_k\|_\infty \xrightarrow{p} 0$. Note that $|\hat{\tau}_{ik}| \leq |\tilde{\tau}_{ik}| + \sum_{j=2}^{n_i} |\epsilon_{ij}| \cdot |\hat{\phi}_k^{(0)}(t_{ij}) - \phi_k(t_{ij})| (t_{ij} - t_{i,j-1})$. One has $E[\tilde{\tau}_{ik}] = 0$ and

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

which implies that, in probability, $\sum_{k=1}^K |\tilde{\tau}_{ik}| \cdot \|\phi_k\|_\infty \leq \sqrt{2 \sup_{t \in \mathcal{T}} \sigma^2(t) \Delta_n^*} \sum_{k=1}^K \|\phi_k\|_\infty \rightarrow 0$ by (A6). Also observing that

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

and $E[\sum_{j=2}^{n_i} |\epsilon_{ij}| (t_{ij} - t_{i,j-1})] \leq |\mathcal{T}| \sup_{t \in \mathcal{T}} \sqrt{\sigma^2(t)}$, this implies that $\sum_{j=2}^{n_i} |\epsilon_{ij}| (t_{ij} - t_{i,j-1}) = O_p(1)$. Then we have $\sum_{k=1}^K |\hat{\tau}_{ik}| \cdot \|\phi_k\|_\infty \xrightarrow{p} 0$. Then the result (12) follows.

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

$$\begin{aligned}
& \sup_{t \in \mathcal{T}} \left| \sum_{k=1}^K \hat{\xi}_{ik}^{(0)} \hat{\phi}_k^{(0)}(t) - \sum_{k=1}^\infty \xi_{ik} \phi_k(t) \right| \\
& \leq \sup_{t \in \mathcal{T}} \left| \sum_{k=1}^K \{ \hat{\xi}_{ik}^{(0)} \hat{\phi}_k^{(0)}(t) - \xi_{ik} \phi_k(t) \} \right| + \sup_{t \in \mathcal{T}} \left| \sum_{k=K+1}^\infty \xi_{ik} \phi_k(t) \right| \xrightarrow{p} 0. \quad (27)
\end{aligned}$$

The second term converging to zero in probability is guaranteed by Karhunen-Loève Theorem, provided $K \rightarrow \infty$ as $n \rightarrow \infty$. We now focus on the first term,

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

Observing $E|Q_2(n)| \leq \sum_{k=1}^K \delta_k A_{\delta_k} E|\xi_{ik}| / (\sqrt{n}h_G^2 - A_{\delta_k}) \leq \sum_{k=1}^K \delta_k A_{\delta_k} \sqrt{\lambda_k} / (\sqrt{n}h_G^2 - A_{\delta_k})$ and $\lambda_k \rightarrow 0$, one has $E|Q_2(n)| = O(v_n)$; i.e., $Q_2(n) = O_p(v_n)$. It is easy to see that $Q_1(n) \leq 2 \sum_{k=1}^K |\hat{\xi}_{ik}^{(0)} - \xi_{ik}| \cdot \|\phi_k\|_\infty$ for large n . Note that

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

In analogy to (25), given (A2.2), (A5) and (A6), the first term in the r.h.s. of (28) is bounded in probability by

$$[c_1(\|X_i\|_{L^2} + \|X_i\|_\infty \|X'_i\|_\infty \Delta_n^*) + c_2]v_n + (1 + \sum_{k=1}^K \|\phi_k\|_\infty \|\phi'_k\|_\infty \Delta_n^*) \frac{\sum_{k=1}^K \|\phi_k\|_\infty}{\sqrt{n}h_\mu} \xrightarrow{p} 0.$$

The second term in the r.h.s. of (28) is bounded in probability by

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

For the third term in the r.h.s. of (28), we have already shown that $\sum_{k=1}^K |\hat{\tau}_{ik}| \cdot \|\phi_k\|_\infty \xrightarrow{p} 0$.

From the above proof, one can see that $\sup_{1 \leq j \leq n_i} |Y_{ij}^* - Y_{ij}^{*(0)}| = O_p(\theta_{in})$ where the $O_p(\cdot)$ holds uniformly over all i 's, and θ_{in} is defined in (22). Based on Theorem 1 and Lemma 3, we are able to derive the uniform convergence of the penalized spline estimator $\hat{\mu}(t)$ and thus the covariance estimator $\hat{G}(s, t)$ obtained by (3). Then the uniform consistency of other model components, including eigenfunctions and eigenvalues, can be obtained in similar manner as in Lemma 1.

Proof of Theorem 2. Recall that the hypothetical penalized spline estimator $\tilde{\mu}(t)$ is obtained by fitting the theoretical pseudo-data Y_{ij}^* , while $\hat{\mu}(t)$ is obtained using $\hat{Y}_{ij}^{*(0)}$ as input. Let \tilde{G} denote the hypothetical covariance estimator obtained by (3) using $\tilde{\mu}(t)$ as mean estimate, while \hat{G} is obtained by (3) using $\hat{\mu}(t)$ as mean estimate. Since linear smoothers, including penalized spline fitting, are weighted averages, and as (14) implies $\hat{Y}_{ij}^{*(0)} = Y_{ij}^* + O_p(\theta_{in})$, 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 $\bar{\theta}_n = \sum_{i=1}^n \theta_{in}$. Observing (A7), and the following $E(\|X\|_\infty \|X'\|_\infty) \leq \sqrt{E(\|X\|_\infty^2) E(\|X'\|_\infty^2)} < \infty$, $E\{\sum_{j=2}^{n_i} |\epsilon_{ij}| (t_{ij} - t_{i,j-1})\} \leq |\mathcal{T}| \sup_{t \in \mathcal{T}} \sqrt{\sigma^2(t)} < \infty$, and $E\{\sum_{k=1}^K \delta_k A_{\delta_k} |\xi_{ik}| / (\sqrt{n}h_G^2 - A_{\delta_k})\} \leq \sum_{k=1}^K \delta_k A_{\delta_k} \sqrt{\lambda_k} / (\sqrt{n}h_G^2 - A_{\delta_k}) \leq v_n$, one has $\bar{\theta}_n = O_p(\theta_n^*) \xrightarrow{p} 0$, where θ_n^* is defined in (23). In

view the convergence results in Lemma 3, this leads to the results (14). In fact, one has the uniform convergence rate of $\hat{\mu}(t)$ and $\widehat{G}(t)$ as follows,

$$\sup_{t \in \mathcal{I}} |\hat{\mu}(t) - \mu(t)| = O_p(\omega_n + \theta_n^*), \quad \sup_{s, t \in \mathcal{I}} |\widehat{G}(s, t) - G(s, t)| = O_p(\omega_n + \theta_n^* + \frac{1}{\sqrt{nh_G^2}}), \quad (29)$$

where ω_n is as in (21), θ_n^* is as in (23) and h_G is the bandwidth 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. *Sankhyā* **55**, 420-441.
- Boente, G. and Fraiman, R. (2000) Kernel-based functional principal components. *Statistics and Probability Letters* **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). *Journal of the American Statistical Association* **93**, 961-1006.
- Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Inference: A Practical Information-Theoretic Approach. Second Edition.* Springer-Verlag New York Inc.
- Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003) Testing hypotheses in the functional linear model. *Scandinavian Journal of Statistics* **30**, 241-255.
- 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. *Journal of the Royal Statistical Society, Series B* **65**, 405-423.
- Fan, J. and Zhang, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series 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. *The American Statistician* **46**, 140-145.
- Lin, X. and Carroll, R. J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association* **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. *Statistical Science* **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. 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.
- 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. *Journal of Computational and Graphical Statistics* **11**, 735-757.
- Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. 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. *Applied Statistics* **45**, 151-163.
- Silverman, B. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics* **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. *Molecular Biology of the 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. *Journal of the American Statistical Association* **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. *Journal of the American Statistical Association* **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. *Journal of the American Statistical Association* **100**, 577-590.

Table 1: Simulation results for comparing mean estimates obtained by Methods M1 to M4 from 100 Monte Carlo runs with $n = 100$ random trajectories per sample. The functional principal component scores were calculated using either the Integration or the PACE methods described in Section 3. Shown in the table are the Monte Carlo estimates of integrated mean squared error (IMSE), integrated squared bias (IBIAS), integrated variance (IVAR), and integrated squared prediction error (IPE). See Section 4 for details.

			Normal				Mixture			
	Design	Method	IBIAS	IVAR	IMSE	IPE	IBIAS	IVAR	IMSE	IPE
Optimal	non-sparse (Integration)	M1	.003	.066	.069	.242	.004	.065	.069	.243
		M2	.007	.072	.079	.246	.008	.072	.080	.247
		M3	.003	.055	.058	.223	.003	.056	.059	.224
		M4	.003	.056	.059	.224	.003	.057	.060	.225
	sparse (PACE)	M1	.006	.154	.160	1.77	.008	.166	.174	1.75
		M2	.030	.173	.203	1.79	.034	.178	.212	1.79
		M3	.004	.116	.120	1.62	.003	.123	.126	1.63
		M4	.004	.118	.122	1.61	.004	.125	.129	1.60
Model- Selected	non-sparse (Integration)	M1	.004	.070	.074	.245	.003	.069	.073	.245
		M2	.008	.077	.085	.248	.007	.077	.084	.251
		M3	.003	.058	.061	.227	.003	.058	.061	.228
		M4	.003	.059	.062	.226	.003	.059	.062	.227
	sparse (PACE)	M1	.004	.205	.209	1.84	.005	.198	.203	1.85
		M2	.028	.218	.246	1.86	.034	.208	.242	1.84
		M3	.003	.148	.151	1.69	.002	.154	.156	1.68
		M4	.003	.145	.148	1.68	.003	.148	.151	1.67

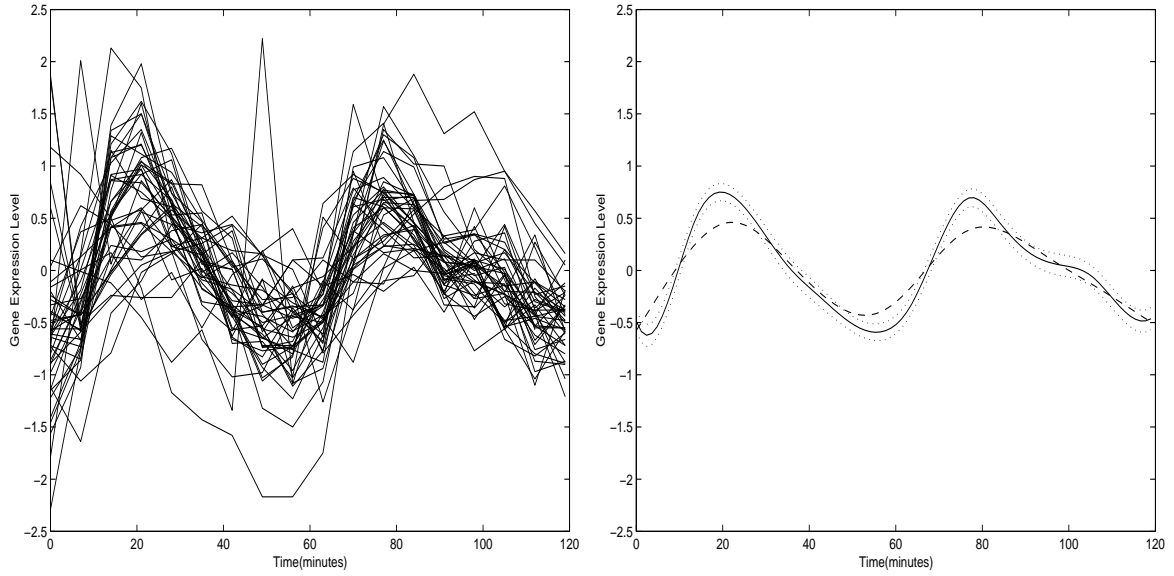


Figure 1: Left: Gene expression profiles of 43 genes from the G1 phase. Right: Estimated mean functions obtained with the traditional non-iterative local polynomial smoothing (15) (dashed) and the proposed IPS procedure (solid) with initial mean estimates given by (15) for the 43 G1 phase genes as well as an approximate pointwise 95% confidence interval (dotted) obtained by (11).

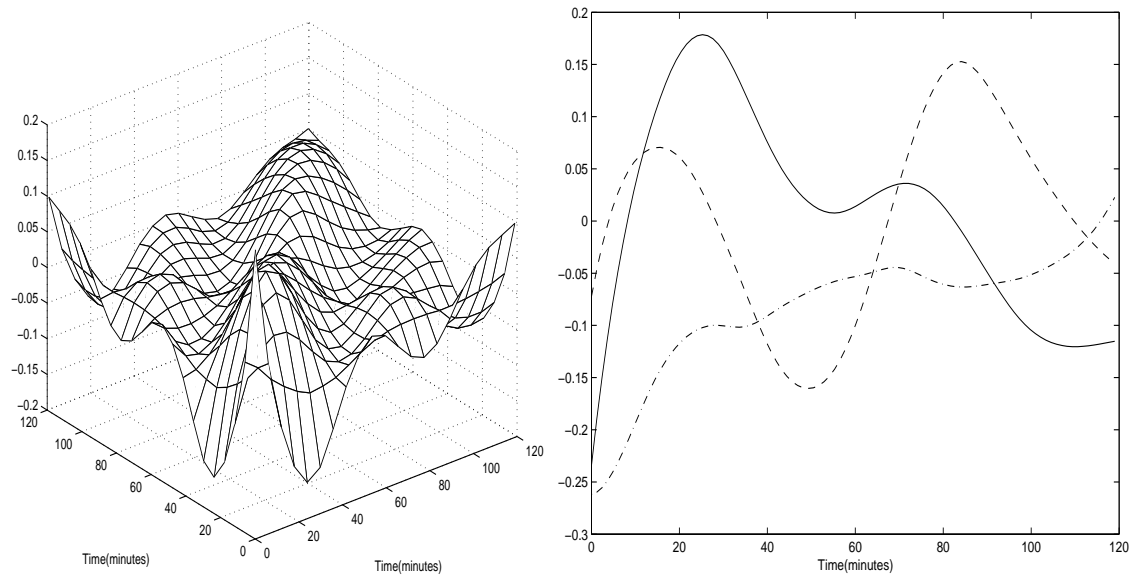


Figure 2: Smooth estimate of the covariance surface (left panel) and three eigenfunctions (right panel) obtained using the proposed IPS procedure for the G1 phase yeast cell cycle gene expression profiles.

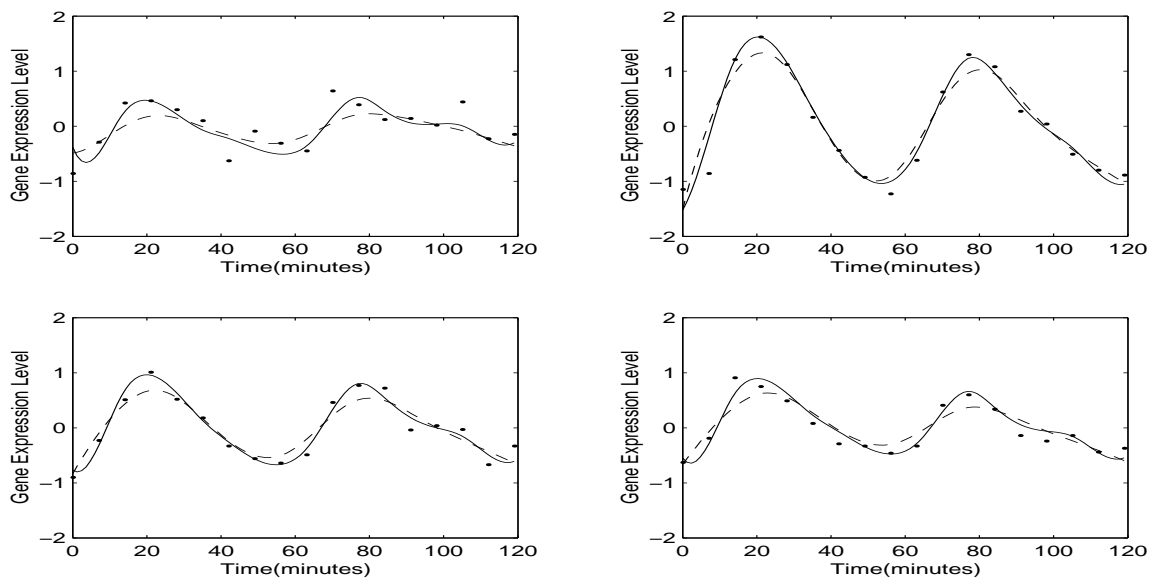


Figure 3: Observed (dot) and estimated gene expression profiles obtained using the proposed IPS procedure (solid) and the traditional non-iterative functional principal component analysis combined with local polynomial smoothing (dashed) for four randomly selected genes related to the G1 phase.