Functional Data Analysis for Sparse Longitudinal Data

Fang Yao, Hans-Georg Müller[†], and Jane-Ling Wang
Department of Statistics
University of California
Davis, CA 95616

Second Revision

August 31, 2004

[†] Corresponding author, e-mail: *mueller@wald.ucdavis.edu*. Research supported in part by NSF grants DMS98-03637, DMS99-71602, DMS02-04869, DMS03-54448 and DMS04-06430. We wish to thank an Associate Editor and two referees for insightful comments on a previous version of this paper which led to many improvements.

ABSTRACT

We propose a nonparametric method to perform functional principal components analysis for the case of sparse longitudinal data. The method aims at irregularly spaced longitudinal data, where the number of repeated measurements available per subject is small. In contrast, classical functional data analysis requires a large number of regularly spaced measurements per subject. We assume that the repeated measurements are randomly located with a random number of repetitions for each subject, and are determined by an underlying smooth random (subject-specific) trajectory plus measurement errors. Basic elements of our approach are the parsimonious estimation of the covariance structure and mean function of the trajectories, and the estimation of the variance of the measurement errors. The eigenfunction basis is estimated from the data, and functional principal component score estimates are obtained by a conditioning step. This conditional estimation method is conceptually simple and straightforward to implement. A key step is the derivation of asymptotic consistency and distribution results under mild conditions, using tools from functional analysis. Functional data analysis for sparse longitudinal data enables prediction of individual smooth trajectories even if only one or few measurements are available for a subject. Asymptotic pointwise and simultaneous confidence bands are obtained for predicted individual trajectories, based on asymptotic distributions, for simultaneous bands under the assumption of a finite number of components. We implement model selection techniques, such as the Akaike information criterion, to choose the model dimension corresponding to the number of eigenfunctions in the model. The methods are illustrated with a simulation study, longitudinal CD4 data for a sample of AIDS patients, and time-course gene expression data for the yeast cell cycle.

KEY WORDS: Asymptotics, Conditioning, Confidence Bands, Measurement Error, Principal Components, Simultaneous Inference, Smoothing.

1. INTRODUCTION

We develop a version of functional principal components (FPC) analysis, in which the functional principal component scores are framed as conditional expectations. We demonstrate that this extends the applicability of functional principal components analysis to situations in longitudinal data analysis, where only few repeated and sufficiently irregularly spaced measurements are available per subject, and refer to this approach as Principal Components Analysis through Conditional Expectation (PACE) for longitudinal data.

When the observed data are in the form of random curves, rather than scalars or vectors, dimension reduction is mandatory, and functional principal components analysis has become a common tool to achieve this, by reducing random trajectories to a set of functional principal component scores. This method however encounters difficulties when applied to longitudinal data with only few repeated observations per subject.

Beyond dimension reduction, functional principal components analysis attempts to characterize the dominant modes of variation of a sample of random trajectories around an overall mean trend function. There exists an extensive literature on functional principal components analysis when individuals are measured at a dense grid of regularly spaced time points. The method was introduced in Rao (1958) for growth curves, and the basic principle has been studied by Besse and Ramsay (1986), Castro, Lawton and Sylvestre (1986), and Berkey et al. (1991). Rice and Silverman (1991) discussed smoothing and smoothing parameter choice in this context, while Jones and Rice (1992) emphasized applications. Various theoretical properties were studied by Silverman (1996), Boente and Fraiman (2000), and Kneip and Utikal (2001). For introduction and summary, see Ramsay and Silverman (1997). Staniswalis and Lee (1998) proposed kernel-based functional principal components analysis for repeated measurements with an irregular grid of time points. The case of irregular grids was also studied by Besse, Cardot and Ferraty (1997) and Boularan, Ferré and Vieu (1993). However, when the time points vary widely across subjects and are sparse, down to one or two measurements, the functional principal component scores defined through the *Karhunen-Loève* expansion are not well approximated by the usual integration method.

Shi, Weiss and Taylor (1996), Rice and Wu (2000), James, Hastie and Sugar (2001), and James and Sugar (2003) proposed B-splines to model the individual curves with random coefficients through mixed effects models. James et al. (2001) and James and Sugar (2003) emphasized the case of sparse data, postulating a reduced rank mixed-effects model through a B-spline basis for the underlying random trajectories. In contrast, we represent the trajectories directly through the *Karhunen-Loève* expansion, determining the eigenfunctions from the data. Perhaps owing to the complexity of their modeling approach, James et al. (2001) did not investigate the asymptotic properties of the estimated

components in relation to the true components, such as the behavior of the estimated covariance structure, eigenvalues and eigenfunctions, especially for the sparse situation. They constructed pointwise confidence intervals for the individual curves using bootstrap. With our simpler and more direct approach, we are able to derive asymptotic properties, using tools from functional analysis. We also derive both pointwise and simultaneous bands for predicted individual trajectories. This requires to obtain first the uniform convergence results for nonparametric function and surface estimates under dependence structure that follows from the longitudinal nature of the data. The dependence is a consequence of the assumed random nature of the observed sample of trajectories, which sets our work apart from previous results where either the observed functions are non-random with independent measurements (Kneip, 1994), are random vectors of large but fixed dimensions (Ferré, 1995) or are random trajectories sampled on dense and regular grids (Cardot, Ferraty and Sarda, 1999).

The contributions of this paper are as follows: First, we provide a new technique, Principal Components Analysis through Conditional Expectation (PACE) for longitudinal and functional data, a method designed to handle sparse and irregular longitudinal data for which the pooled time points are sufficiently dense. Second, the presence of additional measurement errors is taken into account, extending previous approaches of Staniswalis and Lee (1998) and Yao et al. (2003). Third, an emphasis is the derivation of asymptotic consistency properties, by first establishing uniform convergence for smoothed estimates of the mean and covariance functions under mild assumptions. These uniform consistency results are developed for smoothers in the situation where repeated and thus dependent measurements are obtained for the same subject. Then we couple these results with the theory of eigenfunctions and eigenvalues of compact linear operators, to obtain uniform convergence of estimated eigenfunctions and eigenvalues. To our knowledge, there exist only few published asymptotic results for functional principal components (Dauxois, Pousse and Romain 1982; Bosq 1991; Silverman 1996), and none for functional data analysis in the sparse situation. Fourth, we derive the asymptotic distribution that is needed to obtain pointwise confidence intervals for individual trajectories, and obtain asymptotic simultaneous bands for these trajectories.

The main novelty of our work is that we establish the conditional method for the case of sparse and irregular data, show that this provides a straightforward and simple tool for the modeling of longitudinal data, and derive asymptotic results for this method. Under Gaussian assumptions, the proposed estimation of individual functional principal component scores in principal components analysis through conditional expectation corresponds to the best prediction, combining the data from the individual subject to be predicted with data from the entire collection of subjects. In the non-Gaussian case, it provides an estimate for the best linear prediction. The proposed principal components analysis through conditional expectation method extends to the case of sparse and irreg-

ular data, provided that as the number of subject increases, the pooled time points from the entire sample become dense in the domain of the data. We suggest one-curve-leave-out cross-validation for choosing auxiliary parameters such as the degree of smoothing and the model dimension, corresponding to the number of eigenfunctions to be included, similar to Rice and Silverman (1991). For faster computing, we also consider an AIC criterion to select the number of eigenfunctions.

The remainder of the paper is organized as follows: In Section 2 we introduce the principal components analysis through conditional expectation approach, i.e., the proposed conditional estimates for the functional principal component scores. Asymptotic results for the proposed method are presented in Section 3, with proofs in the Appendix. Simulation results that illustrate the usefulness of the methodology are discussed in Section 4. Applications of principal components analysis through conditional expectation (PACE) to longitudinal CD4 data and time-course gene expression data for yeast cell cycle genes are the theme of Section 5, followed by concluding remarks (Section 6) and an Appendix with proofs and theoretical results.

2. FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS FOR SPARSE DATA

2.1 Model with Measurement Errors

We model sparse functional data as noisy sampled points from a collection of trajectories that are assumed to be independent realizations of a smooth random function, with unknown mean function $EX(t) = \mu(t)$ and covariance function $\mathrm{cov}(X(s), X(t)) = G(s,t)$. The domain of $X(\cdot)$ typically is a bounded and closed time interval $\mathcal{T}$. While we refer to the index variable as time, 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 ϕ_k and non-increasing eigenvalues λ_k: $G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t)$, $t, s \in \mathcal{T}$. In classical functional principal components (FPC) analysis it is assumed that the ith random curve can be expressed as $X_i(t) = \mu(t) + \sum_k \xi_{ik}\phi_k(t)$, $t \in \mathcal{T}$, where the ξ_{ik} are uncorrelated random variables with zero mean and variances $E\xi_{ik}^2 = \lambda_k$, where $\sum_k \lambda_k < \infty$, $\lambda_1 \geq \lambda_2 \geq \ldots$.

We consider an extended version of the model that incorporates uncorrelated measurement errors with mean zero and constant variance σ^2 to reflect additive measurement errors (see also Rice and Wu, 2000). Let Y_{ij} be the jth observation of the random function $X_i(\cdot)$, made at a random time T_{ij}, and ϵ_{ij} the additional measurement errors that are assumed to be i.i.d. and independent of the random coefficients ξ_{ik}, where $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, $k = 1, 2, \ldots$. Then the model we consider

is

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

where $E\epsilon_{ij} = 0$, $\mathrm{var}(\epsilon_{ij}) = \sigma^2$, and the number of measurements N_i made on the ith subject is considered random, reflecting sparse and irregular designs. The r.v.'s N_i are assumed to be i.i.d. and independent of all other random variables.

2.2 Estimation of the Model Components

Mean, covariance, and eigenfunctions are assumed to be smooth. We use local linear smoothers (Fan and Gijbels, 1996) for function and surface estimation, fitting local lines in one dimension and local planes in two dimensions by weighted least squares. In a first step, we estimate the mean function μ based on the pooled data from all individuals. The formula for this local linear smoother is in (26) below. Data-adaptive methods for bandwidth choice are available [see Müller and Prewitt (1993) for surface smoothing and Rice and Silverman (1991) for one-curve-leave-out cross-validation]; subjective choices are often adequate. For issues of smoothing dependent data, compare Lin and Carroll (2000). Adapting to estimated correlations when estimating the mean function did not lead to improvements (simulations not reported), therefore we do not incorporate such adjustments.

Note that in model (1), $\mathrm{cov}(Y_{ij}, Y_{il}|T_{ij}, T_{il}) = \mathrm{cov}(X(T_{ij}), X(T_{il})) + \sigma^2\delta_{jl}$, where δ_{jl} is 1 if $j = l$ and 0 otherwise. Let $\mathrm{G}_i(T_{ij}, T_{il}) = (Y_{ij} - \hat{\mu}(T_{ij}))(Y_{il} - \hat{\mu}(T_{il}))$ be the "raw" covariances, where $\hat{\mu}(t)$ is the estimated mean function obtained from the previous step. It is easy to see that $E[\mathrm{G}_i(T_{ij}, T_{il})|T_{ij}, T_{il}] \approx \mathrm{cov}(X(T_{ij}), X(T_{il})) + \sigma^2\delta_{jl}$. Therefore the diagonal of the raw covariances should be removed, i.e., only $\mathrm{G}_i(T_{ij}, T_{il})$, $j \neq l$, should be included as input data for the covariance surface smoothing step (as previously observed in Staniswalis and Lee, 1998). We use one-curve-leave-out cross-validation to choose the smoothing parameter for this surface smoothing step.

The variance σ^2 of the measurement errors is of interest in model (1). Let $\widehat{G}(s,t)$ be a smooth surface estimate (see (27) below) of $G(s,t) = \mathrm{cov}(X(s), X(t))$. Following Yao et al. (2003), since the covariance of $X(t)$ is maximal along the diagonal, a local quadratic rather than a local linear fit is expected to approximate the shape of the surface in the direction orthogonal to the diagonal better. We thus fit a local quadratic component along the direction perpendicular to the diagonal, and a local linear component in the direction of the diagonal; implementation of this local smoother is achieved by rotating the coordinates by 45° and then minimizing weighted least squares (similar to (27)) in rotated coordinates with local quadratic and linear components, see (28) below for details.

Denote the diagonal of the resulting surface estimate by $\widetilde{G}(t)$, and a local linear smoother focusing on diagonal values $\{G(t,t) + \sigma^2\}$ by $\widehat{V}(t)$, obtained by (26) with $\{G_i(T_{ij}, T_{ij})\}$ as input. To mitigate boundary effects, we cut off the two ends of the interval to get a more stable estimate, following a

suggestion of Staniswalis and Lee (1998). Let $|\mathcal{T}|$ denote the length of $\mathcal{T}$, and $\mathcal{T}_1$ be the interval $\mathcal{T}_1 = [\inf\{x : x \in \mathcal{T}\} + |\mathcal{T}|/4,\ \sup\{x : x \in \mathcal{T}\} - |\mathcal{T}|/4]$. The proposed estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{2}{|\mathcal{T}|} \int_{\mathcal{T}_1} \{\widehat{V}(t) - \widetilde{G}(t)\}\, dt, \tag{2}$$

if $\hat{\sigma}^2 > 0$ and $\hat{\sigma}^2 = 0$ otherwise.

The estimates of eigenfunctions and eigenvalues correspond to the solutions $\hat{\phi}_k$ and $\hat{\lambda}_k$ of the eigen-equations,

$$\int_{\mathcal{T}} \widehat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda}_k\hat{\phi}_k(t), \tag{3}$$

where the $\hat{\phi}_k$ are subject to $\int_{\mathcal{T}} \hat{\phi}_k(t)^2 dt = 1$ and $\int_{\mathcal{T}} \hat{\phi}_k(t)\hat{\phi}_m(t)dt = 0$ for $m < k$. We estimate the eigenfunctions by discretizing the smoothed covariance, as previously described in Rice and Silverman (1991) and Capra and Müller (1997).

2.3 Functional Principal Components Analysis through Conditional Expectation

The functional principal component scores $\xi_{ik} = \int (X_i(t) - \mu(t))\phi_k(t)dt$ have traditionally been estimated by numerical integration which works well when the density of the grid of measurements for each subject is sufficiently large. Since in our model the Y_{ij} are only available at discrete random times T_{ij}, reflecting sparseness of the data, the integrals in the definition of the FPC scores ξ_{ik} would accordingly be approximated by sums, substituting Y_{ij} as defined in (1) for $X_i(T_{ij})$, and estimates $\hat{\mu}(t_{ij})$ for $\mu(t_{ij})$ and $\hat{\phi}_k(t_{ij})$ for $\phi_k(t_{ij})$, leading to $\hat{\xi}_{ik}^S = \sum_{j=1}^{N_i}(Y_{ij} - \hat{\mu}(T_{ij}))\hat{\phi}_k(T_{ij})(T_{ij} - T_{i,j-1})$, setting $T_{i0} = 0$. For sparse functional data, $\hat{\xi}_{ik}^S$ will not provide reasonable approximations to ξ_{ik}, for example when one has only two observations per subject. Moreover, when the measurements are contaminated with errors, the underlying random process X cannot be directly observed. Substituting Y_{ij} for $X_i(T_{ij})$ then leads to biased FPC scores. These considerations motivate the alternative PACE method to obtain the FPC scores.

Assume that in (1), ξ_{ik} and ϵ_{ij} are jointly Gaussian. In all of the following, the results pertaining to expectations are always conditional on the observation times T_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, N_i$. For simplicity, the dependence on T_{ij} is suppressed. Write $\widetilde{\boldsymbol{X}}_i = (X_i(T_{i1}), \ldots, X_i(T_{iN_i}))^T$, $\widetilde{\boldsymbol{Y}}_i = (Y_{i1}, \ldots, Y_{iN_i})^T$, $\boldsymbol{\mu}_i = (\mu(T_{i1}), \ldots, \mu(T_{iN_i}))^T$, and $\boldsymbol{\phi}_{ik} = (\phi_k(T_{i1}), \ldots, \phi_k(T_{iN_i}))^T$. The best prediction of the FPC scores for the ith subject, given the data from that individual, is the conditional expectation, which under Gaussian assumptions (also given in (A5) below) is found to be (see, e.g., Theorem 3.2.4 in Mardia, Kent and Bibby, 1979)

$$\tilde{\xi}_{ik} = E[\xi_{ik}|\widetilde{\boldsymbol{Y}}_i] = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{Y_i}^{-1}(\widetilde{\boldsymbol{Y}}_i - \boldsymbol{\mu}_i), \tag{4}$$

where $\boldsymbol{\Sigma}_{Y_i} = \mathrm{cov}(\widetilde{\boldsymbol{Y}}_i, \widetilde{\boldsymbol{Y}}_i) = \mathrm{cov}(\widetilde{\boldsymbol{X}}_i, \widetilde{\boldsymbol{X}}_i) + \sigma^2 \boldsymbol{I}_{N_i}$, i.e, the (j,l) entry of the $N_i \times N_i$ matrix $\boldsymbol{\Sigma}_{Y_i}$ is $(\boldsymbol{\Sigma}_{Y_i})_{j,l} = G(T_{ij}, T_{il}) + \sigma^2\delta_{jl}$ with $\delta_{jl} = 1$ if $j = l$ and 0 if $j \neq l$.

Estimates for the FPC scores ξ_{ik} are obtained from (4), by substituting estimates of $\boldsymbol{\mu}_i$, λ_k and $\boldsymbol{\phi}_{ik}$, $\boldsymbol{\Sigma}_{Y_i}$ obtained from the entire data ensemble, leading to

$$\hat{\xi}_{ik} = \widehat{E}[\xi_{ik}|\widetilde{\boldsymbol{Y}}_i] = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \widehat{\boldsymbol{\Sigma}}_{Y_i}^{-1} (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i), \tag{5}$$

where the (j,l) element of $\widehat{\boldsymbol{\Sigma}}_{Y_i}$ is $(\widehat{\boldsymbol{\Sigma}}_{Y_i})_{j,\,l} = \widehat{G}(T_{ij}, T_{il}) + \hat{\sigma}^2 \delta_{jl}$. Assume the infinite-dimensional processes under consideration are well approximated by the projection on the function space spanned by the first K eigenfunctions. The choice of K will be discussed in Section 2.5. In practice, the prediction for the trajectory $X_i(t)$ for the ith subject is then as follows, using the first K eigenfunctions,

$$\widehat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\phi}_k(t). \tag{6}$$

This conditioning method is simple, provides the best predictors under Gaussian assumptions, and works in the presence of both measurement errors and sparsity. The quantities λ_k and $\boldsymbol{\Sigma}_{Y_i}$ are estimated from the entire data set, borrowing strength from the data on all subjects. We note that $\tilde{\xi}_{ik}$ in (4) is the best linear prediction of ξ_{ik}, given the information from the ith subject, irrespective of whether the Gaussian assumption holds or not. Simulation results, reported in Section 4, indicate that the proposed method is robust in regard to violations of the Gaussian assumption.

2.4 Asymptotic Confidence Bands for Individual Trajectories

To obtain confidence intervals for the FPC scores, for an arbitrary integer $K \geq 1$, write $\boldsymbol{\xi}_{K,\,i} = (\xi_{i1}, \ldots, \xi_{iK})^T$ and $\tilde{\boldsymbol{\xi}}_{K,\,i} = (\tilde{\xi}_{i1}, \ldots, \tilde{\xi}_{iK})^T$. The covariance matrix of $\tilde{\boldsymbol{\xi}}_{K,\,i}$ is $\mathrm{var}(\tilde{\boldsymbol{\xi}}_{K,\,i}) = \boldsymbol{H}\boldsymbol{\Sigma}_{Y_i}^{-1}\boldsymbol{H}^T$, for the $K \times N_i$ matrix $\boldsymbol{H} = \mathrm{cov}(\boldsymbol{\xi}_{K,\,i}, \widetilde{\boldsymbol{Y}}_i) = (\lambda_1\boldsymbol{\phi}_{i1}, \ldots, \lambda_K\boldsymbol{\phi}_{iK})^T$, since $\tilde{\xi}_{K,\,i}$ is a linear function of $\widetilde{\boldsymbol{Y}}_i$. To take into account the variation of $\boldsymbol{\xi}_{K,\,i}$, we use $\mathrm{var}(\tilde{\boldsymbol{\xi}}_{K,\,i} - \boldsymbol{\xi}_{K,\,i})$ to assess the estimation error of $\tilde{\boldsymbol{\xi}}_{K,\,i}$. Because $\tilde{\boldsymbol{\xi}}_{K,\,i} = E[\boldsymbol{\xi}_{K,\,i}|\widetilde{\boldsymbol{Y}}_i]$ is the projection of $\boldsymbol{\xi}_{K,\,i}$ on the space spanned by the linear functions of $\widetilde{\boldsymbol{Y}}_i$, we have $E[\tilde{\boldsymbol{\xi}}_{K,\,i}\boldsymbol{\xi}_{K,\,i}^T] = E[\tilde{\boldsymbol{\xi}}_{K,\,i}\tilde{\boldsymbol{\xi}}_{K,\,i}^T]$, i.e., $\mathrm{var}(\tilde{\boldsymbol{\xi}}_{K,\,i} - \boldsymbol{\xi}_{K,\,i}) = \mathrm{var}(\boldsymbol{\xi}_{K,\,i}) - \mathrm{var}(\tilde{\boldsymbol{\xi}}_{K,\,i}) = \boldsymbol{\Omega}_K$, where $\boldsymbol{\Omega}_K = \boldsymbol{\Lambda} - \boldsymbol{H}\boldsymbol{\Sigma}_{Y_i}^{-1}\boldsymbol{H}^T$, and $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_K\}$. Under Gaussian assumptions, then $(\tilde{\boldsymbol{\xi}}_{K,\,i} - \boldsymbol{\xi}_{K,\,i}) \sim \mathcal{N}(0, \boldsymbol{\Omega}_K)$.

We construct asymptotic pointwise confidence intervals for individual trajectories as follows. Let $\widehat{\boldsymbol{\Omega}}_K = \widehat{\boldsymbol{\Lambda}} - \widehat{\boldsymbol{H}}\widehat{\boldsymbol{\Sigma}}_{Y_i}^{-1}\widehat{\boldsymbol{H}}^T$, where $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}\{\hat{\lambda}_1, \ldots, \hat{\lambda}_K\}$, and $\widehat{\boldsymbol{H}} = (\hat{\lambda}_1\hat{\boldsymbol{\phi}}_{i1}, \ldots, \hat{\lambda}_K\hat{\boldsymbol{\phi}}_{iK})^T$. For $t \in \mathcal{T}$, let $\boldsymbol{\phi}_{K,\,t} = (\phi_1(t), \ldots, \phi_K(t))^T$, $\hat{\boldsymbol{\phi}}_{K,\,t} = (\hat{\phi}_1(t), \ldots, \hat{\phi}_K(t))^T$, and $\widehat{X}_i^K(t) = \hat{\mu}(t) + \hat{\boldsymbol{\phi}}_{K,\,t}^T\hat{\xi}_{K,\,i}$. Theorem 4 below establishes that the distribution of $\{\widehat{X}_i^K(t) - X_i(t)\}$ may be asymptotically approximated by $\mathcal{N}(0,\, \hat{\boldsymbol{\phi}}_{K,\,t}^T\widehat{\boldsymbol{\Omega}}_K\hat{\boldsymbol{\phi}}_{K,\,t})$. Since we assume that X_i can be approximated sufficiently well by the first K eigenfunctions, we may construct the $(1-\alpha)$ asymptotic pointwise interval for $X_i(t)$,

$$\widehat{X}_i^K(t) \pm \Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\hat{\boldsymbol{\phi}}_{K,\,t}^T\widehat{\boldsymbol{\Omega}}_K\hat{\boldsymbol{\phi}}_{K,\,t}}\ , \tag{7}$$

where Φ is the standard Gaussian c.d.f.. These confidence intervals are constructed by ignoring the bias that results from the truncation at K in $\widehat{X}_i^K$.

Next consider the construction of asymptotic simultaneous confidence bands. Let $X_i^K(t) = \mu(t) + \sum_{k=1}^K \xi_{ik}\phi_k(t)$. Theorem 5 provides the asymptotic simultaneous band for $\{\widehat{X}_i^K(t) - X_i^K(t)\}$, for a given fixed K. The *Karhunen-Loève* Theorem implies that $\sup_{t\in\mathcal{T}} E[X_i^K(t) - X_i(t)]^2$ is small for fixed and sufficiently large K. Therefore, ignoring a remaining approximation error that may interpreted as a bias, we may construct $(1-\alpha)$ asymptotic simultaneous bands for $X_i(t)$ through

$$\widehat{X}_i^K(t) \pm \sqrt{\chi^2_{K,\,1-\alpha}\hat{\boldsymbol{\phi}}_{K,\,t}^T\widehat{\boldsymbol{\Omega}}_K\hat{\boldsymbol{\phi}}_{K,\,t}}\ , \tag{8}$$

where $\chi^2_{K,\,1-\alpha}$ is the $100(1-\alpha)$th percentile of the Chi-square distribution with K degrees of freedom. Since $\sqrt{\chi^2_{K,\,1-\alpha}} > \Phi^{-1}(1-\alpha/2)$ for all $K \geq 1$, the asymptotic simultaneous band is always wider than the corresponding asymptotic pointwise confidence intervals.

Analogously, one obtains simultaneous intervals for all linear combinations of the FPC scores. Given K, let $\mathcal{A} \subseteq \Re^K$ be a linear space with dimension $d \leq K$. Then asymptotically, it follows from the uniform result in Corollary 2 below that, for all linear combinations $\boldsymbol{l}^T\boldsymbol{\xi}_{K,\,i}$ simultaneously, where $\boldsymbol{l} \in \mathcal{A}$,

$$\boldsymbol{l}^T\boldsymbol{\xi}_{K,\,i} \in \boldsymbol{l}^T\hat{\boldsymbol{\xi}}_{K,\,i} \pm \sqrt{\chi^2_{d,\,1-\alpha}\boldsymbol{l}^T\widehat{\boldsymbol{\Omega}}_K\boldsymbol{l}}\,, \tag{9}$$

with approximate probability $(1-\alpha)$.

2.5 Selection of the Number of Eigenfunctions

In order to choose the number of eigenfunctions that provide a reasonable approximation to the infinite-dimensional process, one may use the cross-validation score based on the one-curve-leave-out prediction error (Rice and Silverman, 1991). Let $\hat{\mu}^{(-i)}$ and $\hat{\phi}_k^{(-i)}$ be the estimated mean and eigenfunctions after removing the data for the ith subject. Then we choose K so as to minimize the cross-validation score based on the squared prediction error

$$\mathrm{CV}(K) = \sum_{i=1}^{n}\sum_{j=1}^{N_i}\{Y_{ij} - \widehat{Y}_i^{(-i)}(T_{ij})\}^2, \tag{10}$$

where $\widehat{Y}_i^{(-i)}$ is the predicted curve for the ith subject, computed after removing the data for this subject, i.e., $\widehat{Y}_i^{(-i)}(t) = \hat{\mu}^{(-i)}(t) + \sum_{k=1}^K \hat{\xi}_{ik}^{(-i)}\hat{\phi}_k^{(-i)}(t)$, where $\hat{\xi}_{ik}$ is obtained by (5).

One can also adapt AIC type criteria (Shibata, 1991) to this situation, and in simulations not reported here, we found that AIC is computationally more efficient while the results are similar to those obtained by cross-validation. A pseudo-Gaussian log-likelihood, summing the contributions

from all subjects, conditional on the estimated FPC scores $\hat{\xi}_{ik}$ (5), is given by

$$\widehat{L} = \sum_{i=1}^{n} \{ -\frac{N_i}{2} \log(2\pi) - \frac{N_i}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_{ik})^T (\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i - \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_{ik}) \}, \quad (11)$$

where we define AIC $= -\widehat{L} + K$.

3. ASYMPTOTIC PROPERTIES

We derive consistency and distribution results, demonstrating the consistency of the estimated FPC scores $\hat{\xi}_{ik}$ (5) for the true conditional expectations $\tilde{\xi}_{ik}$ (4). Uniform convergence of the local linear estimators of mean and covariance functions on bounded intervals plays a central role in obtaining these results and is therefore established first (Theorem 1). Proofs are deferred to the Appendix.

The data (T_{ij}, Y_{ij}), $i = 1, \ldots, n$, $j = 1, \ldots, N_i$, coming from model (1), are assumed to have the same distribution as (T, Y), with joint density $g(t, y)$. Assume that the observation times T_{ij} are i.i.d. with marginal density $f(t)$, but that dependence is allowed between observations Y_{ij} and Y_{ik}, coming from the same subject or cluster. The following assumptions pertain to the number of observations N_i made on the ith subject or cluster.

(A1.1) The number of observations N_i made for the ith subject or cluster is a r.v. with $N_i \overset{\text{i.i.d}}{\sim} N$, where $N > 0$ is a positive discrete random variable, with $EN < \infty$ and $P\{N > 1\} > 0$.

The observation times and measurements are assumed to be independent of the number of measurements, i.e., for any subset $J_i \subseteq \{1, \ldots, N_i\}$ and for all $i = 1, \ldots, n$,

(A1.2) $(\{T_{ij} : j \in J_i\}, \{Y_{ij} : j \in J_i\})$ is independent of N_i.

Writing $\widetilde{\boldsymbol{T}}_i = (T_{i1}, \ldots, T_{iN_i})^T$ and $\widetilde{\boldsymbol{Y}}_i = (Y_{i1}, \ldots, Y_{iN_i})^T$ as before, it is easy to see that the triples $\{\widetilde{\boldsymbol{T}}_i, \widetilde{\boldsymbol{Y}}_i, N_i\}$ are i.i.d.. Let T_1, T_2 be i.i.d. as T, and let Y_1 and Y_2 be two measurements made on the same subject at times T_1 and T_2. Assume $(T_{ij}, T_{il}, Y_{ij}, Y_{il})$, $j, l \in J_i, j \neq l$, is distributed as (T_1, T_2, Y_1, Y_2) with joint density function $g_2(t_1, t_2, y_1, y_2)$. We assume regularity conditions for the marginal and joint densities, $f(t)$, $g(t, y)$ and $g_2(t_1, t_2, y_1, y_2)$, which are listed as (B1.1)-(B1.3) in the Appendix.

Let $\kappa_1(\cdot)$ and $\kappa_2(\cdot, \cdot)$ be nonnegative univariate and bivariate kernel functions that are used in the smoothing steps for the mean μ and covariance G in Section 2.2 (see (26), (27) for the definition of these smoothers). Kernel $\kappa_1(\cdot)$ is also used for obtaining the estimate $\widehat{V}$ for $\{G(t,t) + \sigma^2\}$ with the local linear smoother. Let h_μ, h_G and h_V be the bandwidths for estimating $\hat{\mu}$, $\widehat{G}$ and $\widehat{V}$. Assume that κ_1 and κ_2 are compactly supported densities with properties (B2.1a)-(B2.2a) and (B2.1b)-(B2.2b) below. We develop asymptotics as the number of subjects $n \to \infty$, and require

(A2.1) $h_\mu \to 0$, $nh_\mu^4 \to \infty$, and $nh_\mu^6 < \infty$.

(A2.2) $h_G \to 0$, $nh_G^6 \to \infty$, and $nh_G^8 < \infty$.

(A2.3) $h_V \to 0$, $nh_V^4 \to \infty$, and $nh_V^6 < \infty$.

Define the Fourier transforms of $\kappa_1(u)$, $\kappa_2(u,v)$ by $\zeta_1(t) = \int e^{-iut}\kappa_1(u)du$ and $\zeta_2(t,s) = \int e^{-(iut+ivs)}\kappa_2(u,v)dudv$. They satisfy

(A3.1) $\zeta_1(t)$ is absolutely integrable, i.e., $\int |\zeta_1(t)|dt < \infty$.

(A3.2) $\zeta_2(t,s)$ is absolutely integrable, i.e., $\int\int |\zeta_2(t,s)|dtds < \infty$.

Assume that the fourth moment of Y centered at $\mu(T)$ is finite, i.e.,

(A4) $E[(Y - \mu(T))^4] < \infty$.

Then we obtain uniform convergence rates for local linear estimators $\hat{\mu}(t)$ of $\mu(t)$ and $\widehat{G}(s,t)$ of $G(s,t)$ on compact sets $\mathcal{T}$ and $\mathcal{T}^2$.

Theorem 1 *Under (A1.1)-(A4) and (B1.1)-(B2.2b) with $\nu = 0$, $\ell = 2$ in (B2.2a) and $\nu = (0,0)$, $\ell = 2$ in (B2.2b),*

$$\sup_{t\in\mathcal{T}} |\hat{\mu}(t) - \mu(t)| = O_p(\frac{1}{\sqrt{n}h_\mu}), \tag{12}$$

$$\sup_{t,s\in\mathcal{T}} |\widehat{G}(s,t) - G(s,t)| = O_p(\frac{1}{\sqrt{n}h_G^2}), \tag{13}$$

The consistency of $\hat{\sigma}^2$ (2) is obtained as a consequence.

Corollary 1 *Under (A1.1)-(A4) and (B1.1)-(B2.2b) with $\nu = 0$, $\ell = 2$ in (B2.2a) and $\nu = (0,0)$, $\ell = 2$ in (B2.2b),*

$$|\hat{\sigma}^2 - \sigma^2| \quad = \quad O_p(\frac{1}{\sqrt{n}}(\frac{1}{h_G^2} + \frac{1}{h_V})). \tag{14}$$

We note that the rates of convergence provided in (12) and (13) are slower than the optimal ones known for the case of smoothing functions or surfaces from sufficiently dense spaced independent measurements. These rates would be of order $O_p(\sqrt{\log n/(nh_\mu)})$ for function estimates and $O_p(\sqrt{\log n/(nh_G^2)})$ for surface estimates. It is an interesting question whether these rates remain optimal for the present dependent data setting and whether they can be attained in the situation of dependent and sparse data that we are dealing with.

Next consider the real separable Hilbert space $L^2(\mathcal{T}) \equiv H$ endowed with inner product $\langle f,g\rangle_H = \int_\mathcal{T} f(t)g(t)dt$ and norm $\|f\|_H = \sqrt{\langle f,f\rangle_H}$ (Courant and Hilbert, 1953). Let $\mathcal{I}'$ denote the set of

indices of the eigenfunctions ϕ_k corresponding to eigenvalues λ_k of multiplicity one. We obtain the consistency of the $\hat{\lambda}_k$ in (3) for λ_k, the consistency of $\hat{\phi}_k$ in (3) for ϕ_k in the L^2 norm $\|\cdot\|_H$, by choosing $\hat{\phi}_k$ appropriately when λ_k is of multiplicity one, and furthermore the uniform consistency of $\hat{\phi}_k$ for ϕ_k on the bounded interval $\mathcal{T}$.

Theorem 2 *Under (A1.1)-(A4) and (B1.1)-(B2.2b) with $\nu = 0$, $\ell = 2$ in (B2.2a) and $\nu = (0,0)$, $\ell = 2$ in (B2.2b),*

$$|\hat{\lambda}_k - \lambda_k| = O_p(\frac{1}{\sqrt{n}h_G^2}), \tag{15}$$

$$\|\hat{\phi}_k - \phi_k\|_H = O_p(\frac{1}{\sqrt{n}h_G^2}), \qquad k \in \mathcal{I}', \tag{16}$$

$$\sup_{t\in\mathcal{T}} |\hat{\phi}_k(t) - \phi_k(t)| = O_p(\frac{1}{\sqrt{n}h_G^2}), \qquad k \in \mathcal{I}'. \tag{17}$$

We remark that the rates (15)-(17) are direct consequences of the rates (12) and (13), as is evident from the proofs. If the rates in (12) and (13) are both $O_p(\alpha_n)$ respectively, then the rates in (15)-(17) will also be $O_p(\alpha_n)$.

For the following results we require Gaussian assumptions.

(A5) The FPC scores ξ_{ik} and measurement errors ϵ_{ij} in (1) are jointly Gaussian.

We also assume that the data asymptotically follow a linear scheme:

(A6) The number, location and values of measurements for a given subject or cluster remain unaltered as $n \to \infty$.

The target trajectories that we aim to predict are

$$\widetilde{X}_i(t) = \mu(t) + \sum_{k=1}^{\infty} \tilde{\xi}_{ik}\phi_k(t), \qquad i = 1, \ldots, n, \tag{18}$$

with $\tilde{\xi}_{ik}$ as defined in (4). We note that $\widetilde{X}_i$ may be defined as a limit of random functions $\tilde{X}_i^K(t) = \mu(t) + \sum_{k=1}^K \tilde{\xi}_{ik}\phi_k(t)$, as $\sup_{t\in\mathcal{T}} E[\widetilde{X}_i^K(t) - \widetilde{X}_i(t)]^2 \to 0$ (see Lemma 3 in the Appendix). For any $K \geq 1$, the target curve $\widetilde{X}_i(t)$ is then estimated by

$$\widehat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t), \tag{19}$$

with $\hat{\xi}_{ik}$ as in (5).

Theorem 3 *Assume (A1.1)-(A6) and (B1.1)-(B2.2b) with $\nu = 0$, $\ell = 2$ in (B2.2a) and $\nu = (0,0)$, $\ell = 2$ in (B2.2b). Then*

$$\lim_{n\to\infty} \hat{\xi}_{ik} = \tilde{\xi}_{ik}, \qquad \textit{in probability}, \tag{20}$$

and for all $t \in \mathcal{T}$,

$$\lim_{K\to\infty} \lim_{n\to\infty} \widehat{X}_i^K(t) = \widetilde{X}_i(t), \qquad \textit{in probability}. \tag{21}$$

We note that the conclusion is still valid for the best linear prediction of ξ_{ik}, given the data vector $\widetilde{\boldsymbol{Y}}_i$, irrespective of whether the Gaussian assumption (A5) holds or not.

For the ith subject and any integer $K \geq 1$, recall that $\boldsymbol{\Omega}_K = \boldsymbol{\Lambda} - \boldsymbol{H}\boldsymbol{\Sigma}_{Y_i}^{-1}\boldsymbol{H}^T$, $\widehat{\boldsymbol{\Omega}}_K = \widehat{\boldsymbol{\Lambda}} - \widehat{\boldsymbol{H}}\widehat{\boldsymbol{\Sigma}}_{Y_i}^{-1}\widehat{\boldsymbol{H}}^T$, $\widehat{X}_i^K(t) = \hat{\mu}(t) + \hat{\boldsymbol{\phi}}_{K,t}^T \hat{\boldsymbol{\xi}}_{K,i}$, $\hat{\boldsymbol{\phi}}_{K,t} = (\hat{\phi}_1(t), \ldots, \hat{\phi}_K(t))^T$, and $\hat{\boldsymbol{\xi}}_{K,i} = (\hat{\xi}_{i1}, \ldots, \hat{\xi}_{iK})^T$. Let $\omega_K(s,t) = \boldsymbol{\phi}_{K,s}^T \boldsymbol{\Omega}_K \boldsymbol{\phi}_{K,t}$ for $t, s \in \mathcal{T}$ and $\hat{\omega}_K(s,t) = \hat{\boldsymbol{\phi}}_{K,s}^T \widehat{\boldsymbol{\Omega}}_K \hat{\boldsymbol{\phi}}_{K,t}$. Then $\{\omega_K(s,t)\}$ is a sequence of continuous positive definite functions. Assume that

(A7) There exists a continuous positive definite function $\omega(s,t)$ such that $\omega_K(s,t) \to \omega(s,t)$, as $K \to \infty$.

Applying Theorem 1 and Theorem 2, the estimate $\hat{\omega}_K(s,t)$ is consistent for $\omega(s,t)$ for all $t, s \in \mathcal{T}$, i.e., $\lim_{K\to\infty} \lim_{n\to\infty} \hat{\omega}_K(s,t) = \omega(s,t)$ in probability.

Theorem 4 *Assume (A1.1)-(A7) and (B1.1)-(B2.2b) with $\nu = 0$, $\ell = 2$ in (B2.2a) and $\nu = (0,0)$, $\ell = 2$ in (B2.2b). For all $t \in \mathcal{T}$ and $x \in \Re$,*

$$\lim_{K\to\infty} \lim_{n\to\infty} P\{\frac{\widehat{X}_i^K(t) - X_i(t)}{\sqrt{\hat{\omega}_K(t,t)}} \leq x\} = \Phi(x), \tag{22}$$

where Φ is the standard Gaussian c.d.f..

The number of random components and eigenfunctions K that are needed in Theorem 3 and 4 to approximate the trajectory $\tilde{X}_i(t)$ depends primarily on the complexity of the covariance structure $G(s,t)$, and number and location of the measurements that are observed for a given subject. It also depends on the sample size n, through the eigenfunction and covariance estimates. While data-based choices for K are available through (10), (11) and are successful in practical applications, results (21), (22) indicate that for large n, the number of components K needs to be increased in order to obtain consistency, but do not provide further guidance as to how K should be chosen in dependence on n.

We next establish $(1-\alpha)$ asymptotic simultaneous inference for $\{\widehat{X}_i^K(t) - X_i^K(t)\}$ on the domain $\mathcal{T}$, where $X_i^K(t) = \mu(t) + \sum_{k=1}^K \xi_{ik}\phi_k(t)$. For these results, we are not providing functional asymptotics but instead finite-dimensional asymptotics, as the number of included components K is considered fixed, while the sample size $n \to \infty$ as before. If K is chosen such that only trajectories

truncated at the first K components $X_i^K(\cdot)$ of their expansion are of interest, then the following two results provide simultaneous confidence bands, as well as simultaneous confidence sets for the first K random effects. Corollary 2 below is a variation of Scheffé's method.

Theorem 5 *Under (A1.1)-(A7) and (B1.1)-(B2.2b) with $\nu = 0$, $\ell = 2$ in (B2.2a) and $\nu = (0,0)$, $\ell = 2$ in (B2.2b), for fixed number of components K,*

$$\lim_{n\to\infty} P\{\sup_{t\in\mathcal{T}} \frac{|\widehat{X}_i^K(t) - X_i^K(t)|}{\sqrt{\hat{\omega}_K(t,t)}} \leq \sqrt{\chi^2_{K,\,1-\alpha}}\} \geq 1-\alpha, \tag{23}$$

where $\chi^2_{K,\,1-\alpha}$ is the $(1-\alpha)$th percentile of the Chi-square distribution with K degrees of freedom.

Assuming K components, let $\mathcal{A} \subseteq \Re^K$ be a linear space with dimension $d \leq K$. By arguments analogous to the proof of Theorem 5, we obtain the asymptotic simultaneous $(1-\alpha)$ confidence region for all linear combinations $\boldsymbol{l}^T\hat{\boldsymbol{\xi}}_{K,\,i}$, where $\boldsymbol{l} \in \mathcal{A}$,

Corollary 2 *Under the assumptions of Theorem 5,*

$$\lim_{n\to\infty} P\{\sup_{\boldsymbol{l}\in\mathcal{A}} \frac{|\boldsymbol{l}^T(\hat{\boldsymbol{\xi}}_{K,\,i} - \boldsymbol{\xi}_{K,\,i})|}{\sqrt{\boldsymbol{l}^T\widehat{\boldsymbol{\Omega}}_K\boldsymbol{l}}} \leq \sqrt{\chi^2_{d,\,1-\alpha}}\} \geq 1-\alpha, \tag{24}$$

where $\chi^2_{d,\,1-\alpha}$ is the $(1-\alpha)$th percentile of the Chi-square distribution with d degrees of freedom.

4. SIMULATION STUDIES

To illustrate the implementation of sparse FPC analysis by PACE, we construct 100 i.i.d. normal and 100 i.i.d. non-normal samples consisting of $n = 100$ random trajectories each. The simulated processes have mean function $\mu(t) = t + \sin(t)$, 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 $\sigma^2 = 0.25$ as variance of the additional measurement errors ϵ_{ij} in (1), which are assumed to be normal with mean 0. For the smoothing steps, univariate and bivariate Epanechnikov kernel functions are used, i.e., $\kappa_1(x) = 3/4(1-x^2)\mathbf{1}_{[-1,1]}(x)$ and $\kappa_2(x,y) = 9/16(1-x^2)(1-y^2)\mathbf{1}_{[-1,1]}(x)\mathbf{1}_{[-1,1]}(y)$, where $\mathbf{1}_A(x) = 1$ if $x \in A$ and 0 otherwise for any set A. For an equally spaced grid $\{c_0, \ldots, c_{50}\}$ 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 $\mathcal{N}(0, 0.1)$, $s_i = 0$ if $s_i < 0$ and $s_i = 10$ if $s_i > 10$, allowing for non-equidistant "jittered" designs. Each curve was sampled at a random number of points, chosen from a discrete uniform distribution on $\{1, \ldots, 4\}$, and the locations of the measurements were randomly chosen from $\{s_1, \ldots, s_{49}\}$ without replacement. For the 100 normal samples, the FPC 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.

To demonstrate the superior performance of the conditional method, we report in Table 1 the average mean squared error for the true curves X_i, $\text{MSE} = \sum_{i=1}^{n} \int_0^{10} \{X_i(t) - \widehat{X}_i^K(t)\}^2 dt/n$, where $\widehat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k(t)$, and $\hat{\xi}_{ik}$ were obtained either by using the proposed principal components analysis through conditional expectation (PACE) method (5) or by using the integration method. The number of eigenfunctions K in each run was chosen by the AIC criterion (11). In each simulation consisting of 100 Monte Carlo runs (for a total of 400 runs: Normal/Mixture and Sparse/Non-sparse), there were always more than 95 runs in which two eigenfunctions were chosen.

Another outcome measure of interest is the average squared error for the two FPC scores, $\text{ASE}(\xi_k) = \sum_{i=1}^{n} (\hat{\xi}_{ik} - \xi_{ik})^2/n$, $k = 1, 2$, also listed in Table 1. We also compared the two methods for irregular but non-sparse simulated data, where the number of observations for each curve was randomly chosen from $\{30, \ldots, 40\}$, with results in Table 1. We find that the gains in the sparse situation are dramatic when switching from the traditional to the PACE method. For the case of an underlying normal distribution, the MSE was reduced by 43% using the PACE method (5) as compared to the traditional method; the $\text{ASE}(\xi_k)$ were reduced by 52%/27% ($k = 1, 2$). For the mixture distribution case, the decreases were still 42% for MSE, and 52%/28% for $\text{ASE}(\xi_k)$ ($k = 1, 2$). In non-sparse situations, the traditional estimates provide reasonable approximations to the underlying integrals, but nevertheless PACE still produces better estimates, with improvements of 10%/10% for MSE and of 20%/21%, 5%/8% for $\text{ASE}(\xi_k)$, $k = 1, 2$, for normal/non-normal samples. We conclude that the gains obtainable by using PACE are substantial for sparse data, and also extend to the case of dense and non-Gaussian data.

5. APPLICATIONS

5.1 Longitudinal CD4 Counts

Since CD4 counts constitute a critical assessment of the status of the immune system and are used as an important marker in describing the progress to AIDS in adults, CD4 cell counts and CD4 percentages, i.e., CD4 counts divided by the total number of lymphocytes, are commonly used markers for the health status of HIV infected persons. The data set considered here is from the Multicenter AIDS Cohort Study, which includes repeated measurements of physical exams, laboratory results and CD4 percentages for 283 homosexual men who became HIV positive between 1984 and 1991. All individuals were scheduled to have their measurements made at semi-annual visits. However, since many individuals missed scheduled visits and the HIV infections happened randomly during the study, the data are sparse with unequal numbers of repeated measurements per subject and different measurement times T_{ij} per individual. The number of observations per subject ranged between 1 and 14, with a median of 6 measurements. The trajectories in their entirety are assembled

in the left panel of Figure 1.

That the data from such a classical longitudinal study, with measurements intended to be spaced at regular 6-months time intervals, are quite well suitable for the analysis by PACE can be seen from Figure 2. As this figure illustrates, the assembled pairs (T_{ij}, T_{ik}) are sufficiently dense in the domain plane and the estimation of the covariance function (27) is feasible for these data. Further details about design, methods and medical implications of the study can be found in Kaslow et al. (1987). Fan and Zhang (2000) and Wu and Chiang (2000) have analyzed these data with varying coefficient models adapted to longitudinal data, while Diggle, Zeger and Liang (1994) discuss classical longitudinal approaches for these data.

The objective of our analysis is estimating the overall trend over time, uncovering subject-specific variation patterns, extracting the dominant modes of variation, and recovering individual trajectories from sparse measurements. This includes predicting the time course for an individual, given only few observations, and constructing pointwise and simultaneous bands for an individual's trajectory. The estimate of the mean function using local linear smoothing is in the right panel of Figure 1, revealing the overall decreasing trend in CD4 cell counts. Estimates of variance and correlation functions are shown in Figure 3; the variance is clearly non-stationary, with high variability at very early times, decreasing until about one year and then increasing again. Measurements made on the same subject are strongly correlated, irrespective of the time difference. However, the correlation between very early and late counts dies off relatively rapidly, whereas for middle and later times, the dependence patterns persist more strongly. These features would be difficult to anticipate in a traditional parametric model; they would not be produced, e.g., by linear random effects models.

Next consider the eigenfunction decomposition of the estimated covariance surface. Three eigenfunctions shown in the upper panels of Figure 4 are used to approximate the infinite-dimensional process. The choice $K = 3$ emerges as a reasonable choice, supported both by the AIC criterion (11) and one-curve-leave-out cross-validation. The first eigenfunction is somewhat similar to the mean function, the second corresponds to a contrast between very early and very late times, and the third to a contrast early and medium plus later times. These eigenfunctions account for 76.9%, 12.3% and 8.1%, respectively, of the total variation. Most of the variability is thus in the direction of overall CD4 percentage level. In exploring such data, extreme individual cases are difficult to detect by visual examination due to irregular sampling and substantial noise. One way to explore the variability in the sample and to single out extreme cases is to identify cases that exhibit large principal component scores in the directions of a few leading eigenfunctions (Jones and Rice, 1992). Three such cases, corresponding to the largest absolute values of the projections on the first three eigenfunctions, are shown in the lower panels of Figure 4.

The predicted curves, 95% pointwise and simultaneous confidence bands for four randomly chosen individuals are displayed in Figure 5, where the principal component scores of each subject are estimated using the PACE method. The predicted curves are seen to be reasonably close to the observations. Even for individuals with very sparse measurements, one is still able to effectively recover their random trajectories, combining the information from that individual and the entire collection. For example, the PACE principle of borrowing strength from the entire sample for predicting individual trajectories makes it feasible to predict trajectories and construct corresponding prediction bands for those cases where only one observation is available per subject, as exemplified in the lower left panel of Figure 5. The predictions based on only one observation per subject work reasonably well as is demonstrated in the second example described below in Section 5.2 (see lower right panel in Figure 9 where only one single measurement enclosed in the circle is used for the prediction of the trajectory). Since we need to be able to consistently estimate the covariance structure, it is however not feasible to apply the method if there is only one observation available per subject for all subjects. Note that the 95% simultaneous bands show a widening near the endpoints due to end effects and increased variance near the ends, and that all observed data fall within these bands.

5.2 Yeast Cell Cycle Gene Expression Profiles

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 methods, of which 43 are known to be related to G1 phase regulation and 49 to non-G1 phase regulation (i.e. S, S/G2, G2/M and M/G1 phases) of the yeast cell cycle; these genes serve as training set. The gene expression level measurement at each time point is obtained as a logarithm of the expression-level ratio.

In order to demonstrate the usefulness of the PACE method for sparse functional data, we artificially "sparsify" the measurements made for the genes in the training data, and compare the results obtained from this "sparsified" data with those obtained from the complete data. To sparsify the expression measurements made for the ith gene expression profile, the number of measurements N_i is randomly chosen from 1 to 6 with equal probability, and the measurement locations are then randomly selected from the 18 recorded gene expression measurements per profile. The median number of observations per gene expression profile for the resulting sparse data is just 3.

Analyses of both complete and sparsified yeast cell cycle profile data are presented in Figures

6–8. The two mean function estimates for the sparse and complete data, obtained by local linear smoothing of the pooled data, are close to each other and show periodicity (see the left panel of Figure 8, presenting two cell cycles). The two smooth covariance surface estimates revealing the structure of the underlying process are displayed in Figure 7. Both surfaces are very similar and exhibit periodic features. We use the first two eigenfunctions to approximate the expression profiles (middle and right panels of Figure 8). The estimates of the first two eigenfunctions obtained from both sparse and complete data are also close and reflect periodicity, explaining around 75% of the total variation.

We randomly select four genes, and present the predicted profiles obtained from both sparse and complete data and the confidence bands using only the sparse data in Figure 9. We note that the trajectories obtained for the complete data are found to be enclosed in the simultaneous 95% confidence bands constructed from the sparse data. The predictions obtained from the sparse data are similar to those from the complete data, and are reasonable when compared with the complete measurements. This demonstrates that the PACE method allows us to effectively recover entire individual trajectories from fragmental data.

6. CONCLUDING REMARKS

Besides the general application to perform functional principal components analysis for sparse and irregular data, an application of the proposed PACE method to impute missing data in longitudinal studies is also feasible. Consider a regular design where for some subjects many data are missing. The PACE method can then be used to impute the missing data from predicted trajectories.

An interesting finding from the simulation study is that the PACE method improves upon traditional functional principal components analysis even under dense and regular designs. This improvement is due to replacing integrals by conditional expectations when determining functional principal component scores. The conditioning step can be interpreted as shrinkage of these random effects towards zero. The observed improvement indicates that the PACE can also be used to advantage for regularly spaced data which enhances the appeal of this method. We conclude that the underlying principle of borrowing strength from an entire sample of curves to predict individual trajectories shows promise in applications.

APPENDIX: PROOFS AND AUXILIARY RESULTS

We assume regularity conditions for the marginal and joint densities $f(t)$, $g(t, y)$ and $g_2(t_1, t_2, y_1, y_2)$. Let ν_1, ν_2, ℓ be given integers, with $0 \leq \nu_1 + \nu_2 < \ell$.

(B1.1) $(d^\ell/dt^\ell)f(t)$ exists and is continuous on $\mathcal{T}$ with $f(t) > 0$ on $\mathcal{T}$;

(B1.2) $(d^\ell/dt^\ell)g(t, y)$ exists and is uniformly continuous on $\mathcal{T} \times \Re$;

(B1.3) $(d^\ell/(dt_1^{\ell_1} dt_2^{\ell_2}))g_2(t_1, t_2, y_1, y_2)$ exists and is uniformly continuous on $\mathcal{T}^2 \times \Re^2$, for $\ell_1 + \ell_2 = \ell$, $0 \leq \ell_1, \ell_2 \leq \ell$.

The assumptions for kernel functions $\kappa_1 : \Re \to \Re$ and $\kappa_2 : \Re^2 \to \Re$ are as follows. We say that a bivariate kernel function κ_2 is of order (ν, ℓ), where ν is a multi-index $\nu = (\nu_1, \nu_2)$, if

$$\int\int u^{\ell_1} v^{\ell_2} \kappa_2(u, v) du dv = \begin{cases} 0 & 0 \leq \ell_1 + \ell_2 < \ell, \ell_1 \neq \nu_1, \ell_2 \neq \nu_2, \\ (-1)^{|\nu|} |\nu|! & \ell_1 = \nu_1, \ell_2 = \nu_2, \\ \neq 0 & \ell_1 + \ell_2 = \ell, \end{cases} \tag{25}$$

where $|\nu| = \nu_1 + \nu_2$. A univariate kernel κ_1 is of order (ν, ℓ) for a univariate $\nu = \nu_1$, if (25) holds with $\ell_2 = 0$ on the right hand side, integrating only over the argument u on the left hand side.

(B2.1a) κ_1 is compactly supported, $\|\kappa_1\|^2 = \int \kappa_1^2(u) du < \infty$;

(B2.2a) κ_1 is a kernel function of order (ν, ℓ).

(B2.1b) κ_2 is compactly supported, $\|\kappa_2\|^2 = \int\int \kappa_2^2(u, v) du dv < \infty$;

(B2.2b) κ_2 is a kernel function of order (ν, ℓ).

We define the local linear scatterplot smoother for $\mu(t)$ by minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{N_i} \kappa_1(\frac{T_{ij} - t}{h_\mu}) \{Y_{ij} - \beta_0 - \beta_1(t - T_{ij})\}^2 \tag{26}$$

with respect to β_0, β_1. The estimate of $\mu(t)$ is then $\hat{\mu}(t) = \hat{\beta}_0(t)$. The local linear surface smoother for $G(s, t)$ is defined by minimizing

$$\sum_{i=1}^{n} \sum_{1 \leq j \neq l \leq N_i} \kappa_2(\frac{T_{ij} - s}{h_G}, \frac{T_{il} - t}{h_G}) \{\mathrm{G}_i(T_{ij}, T_{il}) - f(\beta, (s, t), (T_{ij}, T_{il}))\}^2 \tag{27}$$

where $f(\beta, (s, t), (T_{ij}, T_{il})) = \beta_0 + \beta_{11}(s - T_{ij}) + \beta_{12}(t - T_{il})$. Minimization is with regard to $\beta = (\beta_0, \beta_{11}, \beta_{12})$, yielding the estimate $\widehat{G}(s, t) = \hat{\beta}_0(s, t)$. To obtain the adjusted estimate of $G(s, t)$ on the diagonal, i.e., $\widetilde{G}(t)$, we first rotate both x-axis and y-axis by 45° clockwise and obtain the coordinates of (T_{ij}, T_{ik}) in the rotated axes, denoted by (T_{ij}^*, T_{ik}^*), i.e., $\begin{pmatrix} T_{ij}^* \\ T_{ik}^* \end{pmatrix} =$

$\begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ -\sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix} \begin{pmatrix} T_{ij} \\ T_{ik} \end{pmatrix}$. We then define the surface estimate $\bar{G}(s,t)$ by minimizing the weighted least squares,

$$\sum_{i=1}^{n} \sum_{1 \le j \ne l \le N_i} \kappa_2(\frac{T^*_{ij}-s}{h_G}, \frac{T^*_{il}-t}{h_G})\{\mathrm{G}_i(T^*_{ij}, T^*_{il}) - g(\gamma, (s,t), (T^*_{ij}, T^*_{il}))\}^2, \tag{28}$$

where $g(\gamma, (s,t), (T^*_{ij}, T^*_{il})) = \gamma_0 + \gamma_1(s - T^*_{ij}) + \gamma_2(t - T^*_{ik})^2$. Minimization is with respect to $\gamma = (\gamma_1, \gamma_2, \gamma_3)^T$, leading to $\bar{G}(s,t) = \hat{\gamma}_0(s,t)$. Because of the rotation, the estimate of the covariance surface on the diagonal, $\widetilde{G}(t)$, is now indeed $\bar{G}(0, t/\sqrt{2})$ obtained with the rotated coordinates.

The following auxiliary results provide the weak uniform convergence rate for univariate weighted averages defined below, compare Bhattacharya and Müller (1993). For a positive integer $l \ge 1$, let $(\psi_p)_{p=1,\ldots,l}$ be a collection of real functions $\psi_p : \Re^2 \to \Re$, which satisfy:

(C1.1a) ψ_p are uniformly continuous on $\mathcal{T} \times \Re$;

(C1.2a) The functions $(d^\ell/dt^\ell)\psi_p(t,y)$ exist for all arguments (t,y) and are uniformly continuous on $\mathcal{T} \times \Re$;

(C1.3a) $\int \psi_p^2(t,y) g(t,y) dy dt < \infty$.

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

(C2.1a) $h_\mu \to 0$, $nh_\mu^{\nu+1} \to \infty$, $nh_\mu^{2\ell+2} < \infty$, as $n \to \infty$.

Define the weighted averages

$$\Psi_{pn} = \Psi_{pn}(t) = \frac{1}{nh_\mu^{\nu+1}} \sum_{i=1}^{n} \frac{1}{EN} \sum_{j=1}^{N_i} \psi_p(T_{ij}, Y_{ij}) \kappa_1(\frac{t - T_{ij}}{h_\mu}), \qquad p = 1, \ldots, l,$$

and the quantity

$$\mu_p = \mu_p(t) = \frac{d^\nu}{dt^\nu} \int \psi_p(t,y) g(t,y) dy, \qquad p = 1, \ldots, l.$$

Lemma 1 *Under (A1.1), (A1.2), (A3.1), (B1.1), (B1.2), (B2.1a), (B2.2a), (C1.1a)-(C1.3a), and C(2.1a),* $\tau_{pn} = \sup_{t \in \mathcal{T}} |\Psi_{pn} - \mu_p| = O_p(1/(\sqrt{n} h_\mu^{\nu+1}))$.

Proof. Note $E|\tau_{pn}| \le \sup_t |E\Psi_{pn} - \mu_p| + E\{\sup_t |\Psi_{pn} - E\Psi_{pn}|\}$, where t takes values in $\mathcal{T}$, and $E|\tau_{pn}| = O(1/(\sqrt{n} h_\mu^{\nu+1}))$ implies $\tau_{pn} = O_p(1/(\sqrt{n} h_\mu^{\nu+1}))$.

Using a Taylor expansion to order ℓ, it is easy to show that $E\Psi_{pn} = \mu_p + O(h_\mu^{\ell-\nu})$, where the remainder term is uniform in t, observing that $(d^\ell/dt^\ell)\psi_p(t,y)$ and $(d^\ell/dt^\ell)g(t,y)$ are uniformly continuous. It remains to show that $E\{\sup_t |\Psi_{pn} - E\Psi_{pn}|\} = O(1/(\sqrt{n} h_\mu^{\nu+1}))$. Recall that the inverse Fourier

transform is $\zeta_1(t) = \int e^{-iut}\kappa_1(u)du$. We may insert $\kappa_1((t-T_{ij})/h_\mu) = \int e^{iv(t-T_{ij})/h_\mu}\zeta_1(v)dv/(2\pi)$ into Ψ_{pn}. Letting

$$\varphi_{pn}(u) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{EN}\sum_{j=1}^{N_i} e^{iuT_{lj}}\psi_p(T_{lj}, Y_{lj}),$$

one obtains

$$\Psi_{pn} = \frac{1}{nh_\mu^{\nu+1}}\sum_{i=1}^{n}\frac{1}{EN}\sum_{j=1}^{N_i}\kappa_1(\frac{t-T_{lj}}{h_\mu})\psi_p(T_{lj}, Y_{lj}) = \frac{1}{2\pi h_\mu^\nu}\int \varphi_{pn}(u)e^{-itu}\zeta_1(uh_\mu)du,$$

and thus

$$\sup_t|\Psi_{pn} - E\Psi_{pn}| \le \frac{1}{2\pi h_\mu^\nu}\int |\varphi_{pn}(u) - E\varphi_{pn}(u)|\cdot|\zeta_1(uh_\mu)|du.$$

Note that $E|\varphi_{pn}(u) - E\varphi_{pn}(u)| \le \sqrt{E[\varphi_{pn}(u) - E\varphi_{pn}(u)]^2}$, and because $\{\widetilde{\boldsymbol{T}}_i, \widetilde{\boldsymbol{Y}}_i, N_i\}$ are i.i.d., using the *Cauchy-Schwarz* inequality,

$$\mathrm{var}(\varphi_{pn}(u)) = \frac{1}{n}\mathrm{var}\{\frac{1}{EN}\sum_{j=1}^{N} e^{iuT_j}\psi_p(T_j, Y_j)\} \le \frac{1}{n}E\{(\frac{1}{EN}\sum_{j=1}^{N} e^{iuT_j}\psi_p(T_j, Y_j))^2\}$$

$$\le \frac{1}{n}E\{\frac{1}{(EN)^2}(\sum_{j=1}^{N} e^{i2uT_j})(\sum_{j=1}^{N}\psi_p^2(T_j, Y_j))\} \le \frac{1}{n}E\{\frac{N}{(EN)^2}\sum_{j=1}^{N}E(\psi_p^2(T_j, Y_j)|N)\} = \frac{1}{n}E\psi_p^2(T, Y),$$

implying

$$\begin{aligned} E\{\sup_t|\Psi_{pn} - E\Psi_{pn}|\} &\le \frac{1}{2\pi h_\mu^\nu}\int|\varphi_{pn}(u) - E\varphi_{pn}(u)|\cdot|\zeta_1(uh_\mu)|du \\ &\le \frac{\sqrt{E\psi_p^2(T,Y)}\int|\zeta_1(u)|du}{2\pi}\frac{1}{\sqrt{n}h_\mu^{\nu+1}}. \end{aligned}$$

Since $nh_\mu^{2\ell+2} < \infty$ implies $h_\mu^{\ell-\nu} = O(1/(\sqrt{n}h_\mu^{\nu+1}))$, the result follows.

Analogous to Lemma 1, we obtain the rate of uniform convergence in the two-dimensional situation. Let $\{\theta_p(t, s, y_1, y_2)\}_{p=1,\ldots,l}$ be a collection of real functions $\theta_p : \Re^4 \to \Re$ with the following properties:

(C1.1b) θ_p are uniformly continuous on $\mathcal{T}^2 \times \Re^2$;

(C1.2b) the functions $(d^\ell/(dt^{\ell_1}ds^{\ell_2}))\theta_p(t, s, y_1, y_2)$ exist for all arguments (t, s, y_1, y_2) and are uniformly continuous on $\mathcal{T}^2 \times \Re^2$, for $\ell_1 + \ell_2 = \ell$, $0 \le \ell_1, \ell_2 \le \ell$;

(C1.3b) $\int\int\int\int \theta_p^2(t, s, y_1, y_2)g_2(t, s, y_1, y_2)dy_1dy_2dtds < \infty$.

The sequence of bandwidths $h_G = h_G(n)$ for the two-dimensional smoothers satisfies

(C2.1b) $h_G \to 0$, $nh_G^{|\nu|+2} \to \infty$, $nh_G^{2\ell+4} < \infty$, as $n \to \infty$

Define the weighted averages,

$$\Theta_{pn} = \Theta_{pn}(t,s) = \frac{1}{nh_G^{|\nu|+2}} \sum_{i=1}^{n} \frac{1}{EN(EN-1)} \sum_{1 \le j \ne k \le N_i} \theta_p(T_{ij}, T_{ik}, Y_{ij}, Y_{ik}) \kappa_2(\frac{t-T_{ij}}{h_G}, \frac{s-T_{ik}}{h_G}),$$

and

$$\varrho_p = \varrho_p(t,s) = \sum_{\ell_1+\ell_2=|\nu|} \frac{d^{|\nu|}}{dt^{\ell_1} ds^{\ell_2}} \int\int \theta_p(t,s,y_1,y_2) g_2(t,s,y_1,y_2) dy_1 dy_2, \qquad p = 1, \ldots, l.$$

Lemma 2 *Under (A1.1), (A1.2), (A3.2), (B1.1b), (B1.2b), (B2.1)-(B2.3), (C1.1b)-(C1.3b), and C(2.1b),* $\vartheta_{pn} = \sup_{t,s \in \mathcal{T}} |\Theta_{pn} - \varrho_p| = O_p(1/(\sqrt{n} h^{|\nu|+2}))$.

Proof. Analogous to the proof of Lemma 1.

Proof of Theorem 1. From (26), the local linear estimator $\hat{\mu}(t)$ of the mean function $\mu(t)$ can be explicitly written as

$$\hat{\mu}(t) = \hat{\beta}_0(t) = \frac{\sum_i \frac{1}{EN} \sum_j w_{ij} Y_{ij}}{\sum_i \frac{1}{EN} \sum_j w_{ij}} - \frac{\sum_i \frac{1}{EN} \sum_j w_{ij}(T_{ij}-t)}{\sum_i \frac{1}{EN} \sum_j w_{ij}} \hat{\beta}_1(t) \tag{29}$$

where

$$\hat{\beta}_1(t) = \frac{\sum_i \frac{1}{EN} \sum_j w_{ij}(T_{ij}-t) Y_{ij} - (\sum_i \frac{1}{EN} \sum_j w_{ij}(T_{ij}-t) \sum_i \frac{1}{EN} \sum_j w_{ij} Y_{ij}) / (\sum_i \frac{1}{EN} \sum_j w_{ij})}{\sum_i \frac{1}{EN} \sum_j w_{ij}(T_{ij}-t)^2 - (\sum_i \frac{1}{EN} \sum_j w_{ij}(T_{ij}-t))^2 / (\sum_i \frac{1}{EN} \sum_j w_{ij})}. \tag{30}$$

Here $w_{ij} = \kappa_1((t-T_{ij})/h_\mu)/(nh_\mu)$, where κ_1 is a kernel function of order $(0,2)$, satisfying (B2.1a) and (B2.2a), and $\hat{\beta}_1(t)$ is an estimator for the first derivative $\mu'(t)$ of μ at t.

Considering the Nadaraya-Watson estimator of μ, $\hat{\mu}_{NW}(t) = (\sum_i \sum_j w_{ij} Y_{ij}/EN)/(\sum_i \sum_j w_{ij}/EN)$, and $\hat{f}(t) = \sum_i \sum_j w_{ij}/EN$, we choose $\nu = 0$, $\ell = 2$, $l = 2$, $\psi_1(t,y) = y$, and $\psi_2(t,y) \equiv 1$ in Lemma 1. Obviously $\hat{\mu}_{NW}(t) = H(\Psi_{1n}, \Psi_{2n})$ with $H(x_1,x_2) = x_1/x_2$, and $\hat{f}(t) = \Psi_{2n}$. Using *Slutsky's* Theorem and Lemma 1, it follows that $\sup_{t \in \mathcal{T}} |\hat{\mu}_{NW}(t) - \mu(t)| = O_p(1/(\sqrt{n} h_\mu))$, and $\sup_{t \in \mathcal{T}} |\hat{f}(t) - f(t)| = O_p(1/(\sqrt{n} h_\mu))$.

For the uniform consistency of $\hat{\beta}_1$ as estimator of the derivative μ', define Ψ_{pn}, $1 \le p \le 3$, $\sigma_{\kappa_1}^2 = \int u^2 \kappa_1(u) du$, and the kernel function $\tilde{\kappa}_1(t) = -t\kappa_1(t)/\sigma_{\kappa_1}^2$, furthermore $\psi_1(u,y) = y$, $\psi_2(u,y) \equiv 1$, $\psi_3(u,y) = u - t$. Observe that $\tilde{\kappa}_1$ is of order $(1,3)$, $\sup_{t \in \mathcal{T}} |\hat{f}(t) - f(t)| = O_p(1/(\sqrt{n} h_\mu))$, and define

$$\tilde{H}(x_1,x_2,x_3) = \frac{x_1 - x_2 \hat{\mu}_{NW}(t)}{x_3 - h_\mu^2 x_2^2 / \hat{f}(t) \cdot \sigma_{\kappa_1}^2}, \quad \text{and} \quad H(x_1,x_2,x_3) = \frac{x_1 - x_2 \mu(t)}{x_3}.$$

Then

$$\hat{\beta}_1(t) = \tilde{H}(\Psi_{1n}, \Psi_{2n}, \Psi_{3n}) = [H(\Psi_{1n}, \Psi_{2n}, \Psi_{3n}) + \frac{\Psi_{2n}(\mu(t) - \hat{\mu}_{NW}(t))}{\Psi_{3n}}] \frac{\Psi_{3n}}{\Psi_{3n} + h_\mu^2 \Psi_{2n}^2 / \hat{f}(t) \cdot \sigma_{\kappa_1}^2}.$$

Note that $\mu_1 = (\mu' f + m f')(t)$, $\mu_2 = f'(t)$, and $\mu_3 = f(t)$, implying $\sup_{t\in\mathcal{T}} |\Psi_{pn} - \mu_p| = O_p(1/(\sqrt{n}h_\mu^2))$, for $p = 1, 2, 3$, by Lemma 1. Using the uniform version of *Slutsky's* Theorem, $\sup_{t\in\mathcal{T}} |H(\Psi_{1n}, \Psi_{2n}, \Psi_{3n}) - \mu'(t)| = O_p(1/(\sqrt{n}h_\mu^2))$ follows.

Considering the uniform convergence of $\hat{\beta}_0$ for μ, note that

$$\hat{\beta}_0(t) = \hat{\mu}_{NW}(t) + \frac{\Psi_{2n}\hat{\beta}_1(t)}{\hat{f}(t)} h_\mu^2,$$

Because $\sup_{t\in\mathcal{T}} |\Psi_{2n} - f'(t)| = O_p(1/(\sqrt{n}h_\mu^2))$, $\sup_{t\in\mathcal{T}} |\hat{\beta}_1(t) - \mu'(t)| = O_p(1/(\sqrt{n}h_\mu^2))$, and $\sup_{t\in\mathcal{T}} |\hat{f}'(t) - f(t)| = O_p(1/(\sqrt{n}h_\mu))$, we have $\sup_{t\in\mathcal{T}} |\Psi_{2n}\hat{\beta}_1(t)h_\mu^2/\hat{f}(t)| = O_p(h_\mu^2) = O_p(1/(\sqrt{n}h_\mu))$, as $nh_\mu^6 < \infty$. As $\sup_{t\in\mathcal{T}} |\hat{\mu}_{NW}(t) - \mu(t)| = O_p(1/(\sqrt{n}h_\mu))$, the result (12) follows.

We proceed to show (13). In the local linear estimator for the covariance $G(s,t)$, we used the raw observations, $G_i(T_{ij}, T_{ik}) = (Y_{ij} - \hat{\mu}(T_{ij}))(Y_{ik} - \hat{\mu}(T_{ik}))$, instead of $\tilde{G}_i(T_{ij}, T_{ik}) = (Y_{ij} - \mu(T_{ij}))(Y_{ik} - \mu(T_{ik}))$. Note that

$$\begin{aligned} G_i(T_{ij}, T_{ik}) &= \tilde{G}_i(T_{ij}, T_{ik}) + (Y_{ij} - \mu(T_{ij}))(\mu(T_{ik}) - \hat{\mu}(T_{ik})) + (Y_{ik} - \mu(T_{ik}))(\mu(T_{ij}) - \hat{\mu}(T_{ij})) \\ &\quad + (\mu(T_{ij}) - \hat{\mu}(T_{ij}))(\mu(T_{ik}) - \hat{\mu}(T_{ik})) \end{aligned}$$

Since $\sup_{t\in\mathcal{T}} |\hat{\mu}(t) - \mu(t)| = O_p(1/(\sqrt{n}h_\mu))$ by (12), letting $\theta_1(t_1, t_2, y_1, y_2) = (y_1 - \mu(t_1))(y_2 - \mu(t_2))$, $\theta_2(t_1, t_2, y_1, y_2) = y_1 - \mu(t_1)$, and $\theta_3(t_1, t_2, y_1, y_2) \equiv 1$, then $\sup_{t,s\in\mathcal{T}} |\Theta_{pn}| = O_p(1)$, for $p = 1, 2, 3$, by Lemma 2. This implies that $\sup_{t,s\in\mathcal{T}} |\Theta_{2n}| O_p(1/(\sqrt{n}h_\mu)) = O_p(1/(\sqrt{n}h_\mu))$ and $\sup_{t,s\in\mathcal{T}} |\Theta_{3n}| O_p(1/(\sqrt{n}h_\mu)) = O_p(1/(\sqrt{n}h_\mu))$. Since $\sup_{t\in\mathcal{T}} |\hat{\mu}(t) - \mu(t)|^2 = O_p(1/(nh_\mu^2))$ are negligible compared to Θ_{1n}, the local linear estimator, $\widehat{G}(s,t)$, of $G(s,t)$ obtained from $G_i(T_{ij}, T_{ik})$ is asymptotically equivalent to that obtained from $\tilde{G}_i(T_{ij}, T_{ik})$, denoted by $\tilde{G}(t,s)$. Analogously to the proof of (12), using Lemma 2, and the uniform version of *Slutsky's* Theorem, we obtain the uniform consistency of the local linear estimator $\widehat{G}(s,t)$.

Proof of Corollary 1. Since $\widehat{V}(t)$ is a uniformly consistent estimator of $\{G(t,t) + \sigma^2\}$, analogously to (12), (14) follows by applying (13).

Proof of Theorem 2. 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 $\langle T_1, T_2\rangle_F = \mathrm{tr}(T_1 T_2^*) = \sum_j \langle T_1 u_j, T_2 u_j\rangle_H$ and $\|T\|_F^2 = \langle T, T\rangle_F$, where $T_1, T_2, T \in F$, T_2^* is the adjoint of T_2, and $\{u_j : j \geq 1\}$ is any complete orthonormal system in H. The covariance operator $\boldsymbol{G}$ (resp. $\widehat{\boldsymbol{G}}$) is generated by the kernel G (resp. $\widehat{G}$), i.e., $\mathbf{G}(f) = \int_{\mathcal{T}} G(s,t) f(s) ds$, $\widehat{\mathbf{G}}(f) = \int_{\mathcal{T}} \widehat{G}(s,t) f(s) ds$. It is obvious that $\mathbf{G}$ and $\widehat{\mathbf{G}}$ are Hilbert-Schmidt operators, and (13) implies $\|\widehat{\mathbf{G}} - \mathbf{G}\|_F = O_p(1/(\sqrt{n}h_G^2))$.

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 $\mathcal{I}_i$. To obtain (16), let $\mathbf{P}_j = \sum_{k\in\mathcal{I}_j} \phi_k \otimes \phi_k$, and $\widehat{\mathbf{P}}_j = \sum_{k\in\mathcal{I}_j} \hat{\phi}_k \otimes \hat{\phi}_k$ denote the true and estimated

orthogonal projection operators from H to the subspace spanned by $\{\phi_k : k \in \mathcal{I}_j\}$. For fixed $0 < \rho < \min\{|\lambda_l - \lambda_j| : l \notin \mathcal{I}_j\}$, let $\mathbf{\Lambda}_{\rho,j} = \{z \in \mathcal{C} : |z - \lambda_j| = \rho\}$, where $\mathcal{C}$ stands for the complex numbers. The resolvent of $\mathbf{G}$ (resp. $\widehat{\mathbf{G}}$) is denoted by $\mathbf{R}$ (resp. $\widehat{\mathbf{R}}$), i.e., $\mathbf{R}(z) = (\mathbf{G} - zI)^{-1}$ (resp. $\widehat{\mathbf{R}}(z) = (\widehat{\mathbf{G}} - zI)^{-1}$). As $\widehat{\mathbf{R}}(z) = \mathbf{R}(z)[I + (\widehat{\mathbf{G}} - \mathbf{G})\mathbf{R}(z)]^{-1} = \mathbf{R}(z)\sum_{l=0}^{\infty}[(\widehat{\mathbf{G}} - \mathbf{G})\mathbf{R}(z)]^l$, $\|\widehat{\mathbf{R}}(z) - \mathbf{R}(z)\|_F \leq (\|\widehat{\mathbf{G}} - \mathbf{G}\|_F\|\mathbf{R}(z)\|_F)/(1 - \|\widehat{\mathbf{G}} - \mathbf{G}\|_F\|\mathbf{R}(z)\|_F)$. Note that $\mathbf{P}_j = (2\pi i)^{-1}\int_{\mathbf{\Lambda}_{\rho,j}} \mathbf{R}(z)dz$, $\widehat{\mathbf{P}}_j = (2\pi i)^{-1}\int_{\mathbf{\Lambda}_{\rho,j}} \widehat{\mathbf{R}}(z)dz$. Let $M_{\rho,j} = \sup\{\|\mathbf{R}(z)\|_F : z \in \mathbf{\Lambda}_{\rho,j}\} < \infty$, and let ϵ be such that $0 < \epsilon < 1/(2M_{\rho,j})$, then

$$\|\widehat{\mathbf{P}}_j - \mathbf{P}_j\|_F \leq \int_{\mathbf{\Lambda}_{\rho,j}} \|\widehat{\mathbf{R}}(z) - \mathbf{R}(z)\|_F dz/(2\pi) \leq \rho\frac{\|\widehat{\mathbf{G}} - \mathbf{G}\|_F M_{\rho,j}}{1 - \|\widehat{\mathbf{G}} - \mathbf{G}\|_F M_{\rho,j}} \leq 2\rho M_{\rho,j}\epsilon.$$

Considering ϕ_k corresponding to $k \in \mathcal{I}'$, choose $\hat{\phi}_k$ such that $\langle\hat{\phi}_k, \phi_k\rangle_H > 0$. Then

$$\|\widehat{\mathbf{P}}_k - \mathbf{P}_k\|_F^2 = 2(1 - \langle\hat{\phi}_k \otimes \hat{\phi}_k, \phi_k \otimes \phi_k\rangle_H) = 2(1 - \langle\hat{\phi}_k, \phi_k\rangle_H^2) \geq \|\hat{\phi}_k - \phi_k\|_H^2,$$

and (16) follows. Note that $\lambda_k = \langle\phi_k, \mathbf{G}(\phi_k)\rangle_H$ and $\hat{\lambda}_k = \langle\hat{\phi}_k, \widehat{\mathbf{G}}(\hat{\phi}_k)\rangle_H$, then (15) follows by applying *Slusky's* Theorem. To obtain (17), for fixed $k \in \mathcal{I}'$,

$$\begin{aligned}
&|\hat{\lambda}_k\hat{\phi}_k(t) - \lambda_k\phi_k(t)| = |\int_0^{\mathcal{T}} \widehat{G}(s,t)\hat{\phi}_k(s)ds - \int_0^{\mathcal{T}} G(s,t)\phi_k(s)ds| \\
&\leq \int_0^{\mathcal{T}} |\widehat{G}(s,t) - G(s,t)| \cdot |\hat{\phi}_k(s)|ds + \int_0^{\mathcal{T}} |G(s,t)| \cdot |\hat{\phi}_k(s) - \phi_k(s)|ds \\
&\leq \sqrt{\int_0^{\mathcal{T}} (\widehat{G}(s,t) - G(s,t))^2 ds} + \sqrt{\int_0^{\mathcal{T}} G^2(s,t)ds}\,\|\hat{\phi}_k - \phi_k\|_H
\end{aligned}$$

Due to (13) and (16), assuming $\lambda_k > 0$ without loss of generality, we have $|\hat{\lambda}_k\hat{\phi}_k(t)/\lambda_k - \phi_k(t)| = O_p(1/(\sqrt{n}h^2))$, uniformly in $t \in \mathcal{T}$. Then (17) follows by applying (15).

The next result ensures that the target trajectory $\widetilde{X}_i$ is well defined.

Lemma 3 *For the positive definite covariance operator* $\mathbf{G}$ *generated by the continuous symmetric function* $G(s,t)$ *on* $\mathcal{T}^2$*, as* $K \to \infty$*,*

$$\sup_{t\in\mathcal{T}} E[\widetilde{X}_i^K(t) - \widetilde{X}_i(t)]^2 \longrightarrow 0. \tag{31}$$

Proof. Since the covariance operator $\mathbf{G}$ generated by the continuous symmetric function $G(s,t)$ is positive definite, by *Mercer's* Theorem, $\sum_{k=K}^{\infty}\lambda_k\phi_k(s)\phi_k(t)$ converges to 0 uniformly in $(s,t) \in \mathcal{T}^2$. Note that $\widetilde{X}_{i,K}(t) - \widetilde{X}_i(t) = E[\sum_{k=K+1}^{\infty}\xi_{ik}\phi_k(t)|\widetilde{\boldsymbol{Y}}_i]$. From

$$\begin{aligned}
\sup_{t\in\mathcal{T}} \mathrm{var}(\sum_{k=K+1}^{\infty}\xi_{ik}\phi_k(t)) &= \sup_{t\in\mathcal{T}}\{E[E[\sum_{k=K+1}^{\infty}\xi_{ik}\phi_k(t)|\widetilde{\boldsymbol{Y}}_i]^2] + E[\mathrm{var}(\sum_{k=K+1}^{\infty}\xi_{ik}\phi_k(t)|\widetilde{\boldsymbol{Y}}_i)]\} \\
&= \sup_{t\in\mathcal{T}}\sum_{k=K+1}^{\infty}\lambda_k\phi_k^2(t) \longrightarrow 0,
\end{aligned}$$

and $E[\text{var}(\sum_{k=K+1}^{\infty}\xi_{ik}\phi_k(t)|\widetilde{\boldsymbol{Y}}_i)] \geq 0$, (31) follows.

Proof of Theorem 3. Recall that $\hat{\xi}_{ik} = \hat{\lambda}_k\hat{\boldsymbol{\phi}}_{ik}^T\widehat{\boldsymbol{\Sigma}}_{Y_i}^{-1}(\widetilde{\boldsymbol{Y}}_i - \hat{\boldsymbol{\mu}}_i)$, where the (j,l)th entry of the $N_i \times N_i$ matrix $\widehat{\boldsymbol{\Sigma}}_{Y_i}$ is $(\widehat{\boldsymbol{\Sigma}}_{Y_i})_{j,l} = \widehat{G}(T_{ij}, T_{il}) + \hat{\sigma}^2\delta_{jl}$ with $\delta_{jl} = 1$ if $j = l$ and 0 if $j \neq l$. Applying Theorem 1, Theorem 2, Corollary 1 and *Slutsky's* Theorem, (20) follows. We next prove (21) for each fixed $t \in \mathcal{T}$. Let $\widetilde{X}_i^K(t) = \mu(t) + \sum_{k=1}^{K}\tilde{\xi}_{ik}\phi_k(t)$, where $\tilde{\xi}_{ik}$ is defined in (4). Note that

$$|\widehat{X}_i^K(t) - \widetilde{X}_i(t)| \leq |\widehat{X}_i^K(t) - \widetilde{X}_i^K(t)| + |\widetilde{X}_i^K(t) - \widetilde{X}_i(t)|$$

Lemma 3 implies $\widetilde{X}_i^K(t) \xrightarrow{p} \widetilde{X}_i(t)$ as $K \to \infty$. For fixed K, observing that $\hat{\xi}_{ik} \xrightarrow{p} \tilde{\xi}_{ik}$ as $n \to \infty$, then $\sup_{t\in\mathcal{T}}|\widehat{X}_i^K(t) - \widetilde{X}_i^K(t)| \xrightarrow{p} 0$ as $n \to \infty$, by (12), (17) and *Slutsky's* Theorem. This implies that for given $\epsilon, \delta > 0$, there exists K_0 such that for $K \geq K_0$, $P\{|\widetilde{X}_i^K(t) - \widetilde{X}_i(t)| > \epsilon/2\} \leq \delta/2$. For each K, there exists $n_0(K)$ such that for $n \geq n_0(K)$, $P\{|\widehat{X}_i^K(t) - \widetilde{X}_i^K(t)| \geq \epsilon/2\} \leq \delta/2$. Thus, for $K \geq K_0$ and $n \geq n_0(K)$, $P\{|\widehat{X}_i^K(t) - \widetilde{X}_i(t)| \geq \epsilon\} \leq P\{|\widehat{X}_i^K(t) - \widetilde{X}_i^K(t)| > \epsilon/2\} + P\{|\widetilde{X}_i^K(t) - \widetilde{X}_i(t)| \geq \epsilon/2\} \leq \delta$, which leads to (21).

Proof of Theorem 4. Under the Gaussian assumption, for any fixed $K \geq 1$, from Section 2.4, one has $(\tilde{\boldsymbol{\xi}}_{K,i} - \boldsymbol{\xi}_{K,i}) \sim \mathcal{N}(0, \boldsymbol{\Omega}_K)$. Observing (12), (17) and (20), $\lim_{n\to\infty}\sup_{t\in\mathcal{T}}|\widehat{X}_i^K(t) - \widetilde{X}_i^K(t)| \xrightarrow{p} 0$. Since $\widehat{X}_i^K(t) - X_i^K(t) = \widehat{X}_i^K(t) - \widetilde{X}_i^K(t) + \widetilde{X}_i^K(t) - X_i^K(t)$, for fixed K, it follows that $\{\widehat{X}_i^K(t) - X_i^K(t)\} \xrightarrow{\mathcal{D}} Z_K \sim \mathcal{N}(0, \omega_K(t,t))$. Under (A7), letting $K \to \infty$ leads to $Z_K \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0, \omega(t,t))$. From the *Karhunen-Loève* Theorem, $|X_i^K(t) - X_i(t)| \xrightarrow{p} 0$, as $K \to \infty$. Therefore $\lim_{K\to\infty}\lim_{n\to\infty}\{\widehat{X}_i^K(t) - X_i(t)\} \stackrel{\mathcal{D}}{=} Z$. From Theorem 1 and Theorem 2, it can be shown that $\hat{\omega}_K(t,t) \xrightarrow{p} \omega_K(t,t)$ as $n \to \infty$. Under (A7), it follows that $\lim_{K\to\infty}\lim_{n\to\infty}\hat{\omega}_K(t,t) = \omega(t,t)$ in probability. Applying *Slutsky's* Theorem, (22) follows.

Proof of Theorem 5. We first prove

$$P\{\sup_{t\in\mathcal{T}}\frac{|\widetilde{X}_i^K(t) - X_i^K(t)|}{\sqrt{\omega_K(t,t)}} \leq \sqrt{\chi^2_{K,1-\alpha}}\} \geq 1 - \alpha. \tag{32}$$

It is obvious that $\widetilde{X}_i^K(t) - X_i^K(t) = \boldsymbol{\phi}_{K,t}^T(\tilde{\boldsymbol{\xi}}_{K,i} - \boldsymbol{\xi}_{K,i})$. Due to orthogonality, $\mathcal{F} = \{\boldsymbol{\phi}_{K,t} : t \in \mathcal{T}\}$ is an K-dimensional compact set. Since $\boldsymbol{\Omega}_K$ is positive definite, there exists a $K \times K$ non-singular matrix $\boldsymbol{U}$ such that $\boldsymbol{U}\boldsymbol{\Omega}_K\boldsymbol{U}^T = \boldsymbol{I}_K$. Let $\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{\xi}_{K,i}$ and $\tilde{\boldsymbol{\theta}} = \boldsymbol{U}\tilde{\boldsymbol{\xi}}_{K,i}$, then $(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \mathcal{N}(0, \boldsymbol{I}_K)$. This leads to $(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \chi^2_K$, and $P\{(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\} = 1 - \alpha$. We use the following result known from linear algebra.

Lemma 4 *For a fixed p-vector $\boldsymbol{x}$ and a constant $c > 0$, $\boldsymbol{x}^T\boldsymbol{x} \leq c^2$ if and only if $|\boldsymbol{a}^T\boldsymbol{x}| \leq c\sqrt{\boldsymbol{a}^T\boldsymbol{a}}$, for all $\boldsymbol{a} \in \Re^p$.*

Hence, $P\{|\boldsymbol{a}^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})| \leq \sqrt{\chi^2_{K,1-\alpha}\boldsymbol{a}^T\boldsymbol{a}} : \text{ for all } \boldsymbol{a} \in \Re^K\} = 1 - \alpha$. Let $\mathcal{E} = \{\boldsymbol{a} \in \Re^K : \boldsymbol{\phi}_{K,t} = \boldsymbol{U}^T\boldsymbol{a}, t \in \mathcal{T}\}$, which is a compact subset of $\Re^K$. Then $P\{|\boldsymbol{a}^T(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})| \leq \sqrt{\chi^2_{K,1-\alpha}\boldsymbol{a}^T\boldsymbol{a}} : \text{ for all } \boldsymbol{a} \in \mathcal{E}\} \geq 1 - \alpha$,

i.e.,

$$P\{|\boldsymbol{\phi}_{K,t}^T(\tilde{\boldsymbol{\xi}}_{K,i}-\boldsymbol{\xi}_{K,i})| \le \sqrt{\chi^2_{K,1-\alpha}\boldsymbol{\phi}_{K,t}\boldsymbol{U}^{-1}(\boldsymbol{U}^T)^{-1}\boldsymbol{\phi}_{K,t}} : \text{ for all } t\in\mathcal{T}\} \ge 1-\alpha.$$

Observing that $\boldsymbol{U}\boldsymbol{\Omega}_K\boldsymbol{U}^T=\boldsymbol{I}_K$, (32) follows.

To prove (23), note that

$$\sup_{t\in\mathcal{T}}\frac{|\widehat{X}_i^K(t)-X_i^K(t)|}{\sqrt{\omega_K(t,t)}} \le \left(\sup_{t\in\mathcal{T}}\frac{|\widehat{X}_i^K(t)-\widetilde{X}_i^K(t)|}{\sqrt{\omega_K(t,t)}}+\sup_{t\in\mathcal{T}}\frac{|\widetilde{X}_i^K(t)-X_i^K(t)|}{\sqrt{\omega_K(t,t)}}\right)\sup_{t\in\mathcal{T}}\sqrt{\frac{\omega_K(t,t)}{\hat{\omega}_K(t,t)}}.$$

Let $A=\sup_{t\in\mathcal{T}}|\widehat{X}_i^K(t)-\widetilde{X}_i^K(t)|/\sqrt{\omega_K(t,t)}$, $B=\sup_{t\in\mathcal{T}}|\widetilde{X}_i^K(t)-X_i^K(t)|/\sqrt{\omega_K(t,t)}$, and $C=\sup_{t\in\mathcal{T}}\sqrt{\omega_K(t,t)/\hat{\omega}_K(t,t)}$. Since $\omega_K(t,t)$ is a continuous positive definite function on the bounded interval $\mathcal{T}$, it is bounded from above and below, say $0<a\le\omega_K(t,t)\le b<\infty$. Because $\sup_{t\in\mathcal{T}}|\widehat{X}_i^K(t)-\widetilde{X}_i^K(t)| \xrightarrow{p} 0$, as $n\to\infty$, we have $A\xrightarrow{p}0$ as $n\to\infty$. In the proof of (22), it was established that $\hat{\omega}_K(t,t)\xrightarrow{p}\omega_K(t,t)$, as $n\to\infty$, implying that $C\xrightarrow{p}1$. We now will show

$$\lim_{n\to\infty}P\{(A+B)C\ge(\epsilon+\sqrt{\chi^2_{K,1-\alpha}})(1+\epsilon)\}\le\alpha. \tag{33}$$

Note that

$$\begin{aligned}\{(A+B)C\ge(\epsilon+\sqrt{\chi^2_{K,1-\alpha}})(1+\epsilon)\} &\subseteq \{(A+B)\ge(\epsilon+\sqrt{\chi^2_{K,1-\alpha}})\}\cup\{C\ge(1+\epsilon)\}\\ &\subseteq \{A\ge\epsilon\}\cup\{B\ge\sqrt{\chi^2_{K,1-\alpha}}\}\cup\{C\ge(1+\epsilon)\}.\end{aligned}$$

Since $A\xrightarrow{p}0$ and $C\xrightarrow{p}1$ as $n\to\infty$, for sufficiently large n, $P(A\ge\epsilon)\le\tau/3$ and $P(C-1\ge\epsilon)\le\tau/3$. We have shown $P(B\ge\sqrt{\chi^2_{K,1-\alpha}})\le\alpha$ in (32). This implies (33), and then (23) by letting $\epsilon\to0$.

Proof of Corollary 2. There exists an $K\times d$ matrix $\boldsymbol{Q}^T$ with rank $d\le K$, such that $\mathcal{F}$ is spanned by the column vectors of $\boldsymbol{Q}^T$. Letting $\boldsymbol{\delta}=\boldsymbol{Q}\boldsymbol{\xi}_{K,i}$ and $\tilde{\boldsymbol{\delta}}=\boldsymbol{Q}\,\tilde{\boldsymbol{\xi}}_{K,i}$, for any $\boldsymbol{l}\in\mathcal{A}$, where $\mathcal{A}\subseteq\Re^K$ is a linear space with dimension d, there exists a vector $\boldsymbol{\lambda}\in\Re^d$ such that $\boldsymbol{l}=\boldsymbol{Q}^T\boldsymbol{\lambda}$. Then

$$\boldsymbol{l}^T\tilde{\boldsymbol{\xi}}_{K,i}-\boldsymbol{l}^T\boldsymbol{\xi}_{K,i}=\boldsymbol{\lambda}^T\tilde{\boldsymbol{\delta}}-\boldsymbol{\lambda}^T\boldsymbol{\delta}\sim\mathcal{N}(0,\boldsymbol{\lambda}^T\boldsymbol{Q}\,\boldsymbol{\Omega}_K\boldsymbol{Q}^T\boldsymbol{\lambda}).$$

Since $\boldsymbol{Q}$ is of rank d and $\boldsymbol{\Omega}_K$ is positive definite, which implies that $\boldsymbol{Q}\,\boldsymbol{\Omega}_K\boldsymbol{Q}^T$ is also positive definite, there exists a non-singular $d\times d$ matrix $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{Q}\,\boldsymbol{\Omega}_K\boldsymbol{Q}^T\boldsymbol{P}^T=\boldsymbol{I}_d$, where $\boldsymbol{I}_d$ is the $d\times d$ identity matrix. Letting $\boldsymbol{\eta}=\boldsymbol{P}\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\eta}}=\boldsymbol{P}\tilde{\boldsymbol{\delta}}$, we have $(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta})\sim\mathcal{N}(0,\boldsymbol{I}_d)$, i.e, $(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta})^T(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta})\sim\chi^2_d$. Therefore $P\{(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta})^T(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta})\le\chi^2_{d,1-\alpha}\}=1-\alpha$. Applying Lemma 4, we obtain $P\{|\boldsymbol{a}^T(\tilde{\boldsymbol{\eta}}-\boldsymbol{\eta})|\le\sqrt{\chi^2_{d,1-\alpha}\boldsymbol{a}^T\boldsymbol{a}}:\ \text{for all } \boldsymbol{a}\in\Re^d\}=1-\alpha$. Since $\boldsymbol{P}$ is non-singular and $\boldsymbol{Q}$ is of rank d, there exists $\boldsymbol{\lambda}\in\Re^d$ and $l\in\mathcal{A}$, such that $\boldsymbol{\lambda}=\boldsymbol{P}^T\boldsymbol{a}$ and $\boldsymbol{l}=\boldsymbol{Q}^T\boldsymbol{\lambda}$. If $\boldsymbol{a}$ takes all values in $\Re^d$, then $\boldsymbol{l}$ will also take all values in $\mathcal{A}$, i.e.,

$$P\{|\boldsymbol{l}^T(\tilde{\boldsymbol{\xi}}_{K,i}-\boldsymbol{\xi}_{K,i})|\le\sqrt{\chi^2_{d,1-\alpha}\boldsymbol{l}^T(\boldsymbol{P}\boldsymbol{Q})^{-1}(\boldsymbol{Q}^T\boldsymbol{P}^T)^{-1}\boldsymbol{l}}:\ \text{for all } \boldsymbol{l}\in\mathcal{A}\}=1-\alpha.$$

Since $\boldsymbol{P}\boldsymbol{Q}\,\boldsymbol{\Omega}_K\boldsymbol{Q}^T\boldsymbol{P}^T=\boldsymbol{I}_d$, the result (24) follows.

REFERENCES

Berkey, C. S., Laird, N. M., Valadian, I., and Gardner, J. (1991), "Modeling Adolescent Blood Pressure Patterns and Their Prediction of Adult Pressures," *Biometrics*, 47, 1005-1018.

Besse, P., Cardot, H., and Ferraty, F. (1997), "Simultaneous Nonparametric Regression of Unbalanced Longitudinal Data," *Computational Statistics and Data Analysis*, 24, 255-270.

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.

Boularan, J., Ferré, L., and Vieu, P. (1993), "Growth Curves: A Two-stage Nonparametric Approach," *Journal of Statistical Planning and Inference*, 38, 327-350.

Bosq, D. (1991), "Modelization, Nonparametric Estimation and Prediction for Continuous Time Processes," in *Nonparametric Functional Estimation and Related Topics* (1991 ed.), ed. G. Roussas, Dordrecht, Netherlands: Kluwer Academic, pp 509-529.

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

Cardot, H., Ferraty, F., and Sarda, P. (1999), "Functional Linear Model," *Statistics and Probability Letters* 45, 11-22.

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.

Courant, R., and Hilbert, D. (1953), *Methods of Mathematical Physics* (1989 ed.), New York: Wiley.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford University Press.

Dauxois, J., Pousse, A., and Romain, Y. (1982), "Asymptotic Theory for the Principal Component Analysis of a Vector Random Function: Some Applications to Statistical Inference," *Journal of Multivariate Analysis*, 12, 136-154.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J., and Zhang, J. T. (2000), "Two-step Estimation of Functional Linear Models with Applications to Longitudinal Data," *Journal of Royal Society of Statistics, Series B*, 62, 303-322.

Ferré, L. (1995), "Improvement of Some Multivariate Estimates by Reduction of Dimensionality," *Journal of Multivariate Analysis*, 54, 147-162.

James, G., Hastie, T. G., and Sugar, C. A. (2001), "Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587-602.

James, G., and Sugar, C. A. (2003), "Clustering for Sparsely Sampled Functional Data," *Journal of the American Statistical Association*, 98, 397-408.

Jones, M. C., and Rice, J. (1992), "Displaying the Important Features of Large Collections of Similar Curves," *The American Statistician*, 46, 140-145.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. (1987), " The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126, 310-318.

Kneip, A. (1994), "Nonparametric Estimation of Common Regressors for Similar Curve Data," *The Annals of Statistics*, 22, 1386-1472.

Kneip, A., and Utikal, K. (2001), "Inference for Density Families Using Functional Principal Component Analysis," *Journal of the American Statistical Association*, 96, 519-532.

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.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, London: Academic Press.

Müller, H. G., and Prewitt, K. (1993)," Multiparameter Bandwidth Processes and Adaptive Surface Smoothing," *Journal of Multivariate Analysis*, 47, 1-21.

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*, Ser. B, 53, 233-243.

Rice, J., and Wu, C. (2000), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253-259.

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.

Shibata, R. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45-54.

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 Microarry Hybridization," *Molecular Biology of the Cell*, 9, 3273-3297.

Staniswalis, J. G., and Lee, J. J. (1998), "Nonparametric Regression Analysis of Longitudinal Data," *Journal of the American Statistical Association*, 93, 1403-1418.

Wu, C., and Chiang, C. (2000). "Kernel Smoothing on Varying Coefficient Models with Longitudinal Dependent Variable," *Statistica Sinica*, 10, 433-456.

Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., Vogel, J. S. (2003), "Shrinkage Estimation for Functional Principal Component Scores with Application to the Population Kinetics of Plasma Folate," *Biometrics*, 59, 676-685.

Table 1: Results for Functional Principal Components Analysis (FPC) using conditional expectation (CE) and integration (IN) methods for 100 Monte Carlo runs with $N = 100$ random trajectories per sample, generated with two random components. Shown are the averages of estimated mean squared prediction error MSE and average squared error $\text{ASE}(\xi_k)$, $k = 1, 2$, as described in Section 4. The number of components for each Monte Carlo run is chosen by the AIC criterion (11).

N=100		Normal			Mixture		
FPC		MSE	$\text{ASE}(\xi_1)$	$\text{ASE}(\xi_2)$	MSE	$\text{ASE}(\xi_1)$	$\text{ASE}(\xi_2)$
Sparse	CE	1.33	.762	.453	1.30	.737	.453
	IN	2.32	1.58	.622	2.25	1.53	.631
Non-Sparse	CE	.259	.127	.110	.256	.132	.105
	IN	.286	.159	.115	.286	.168	.114

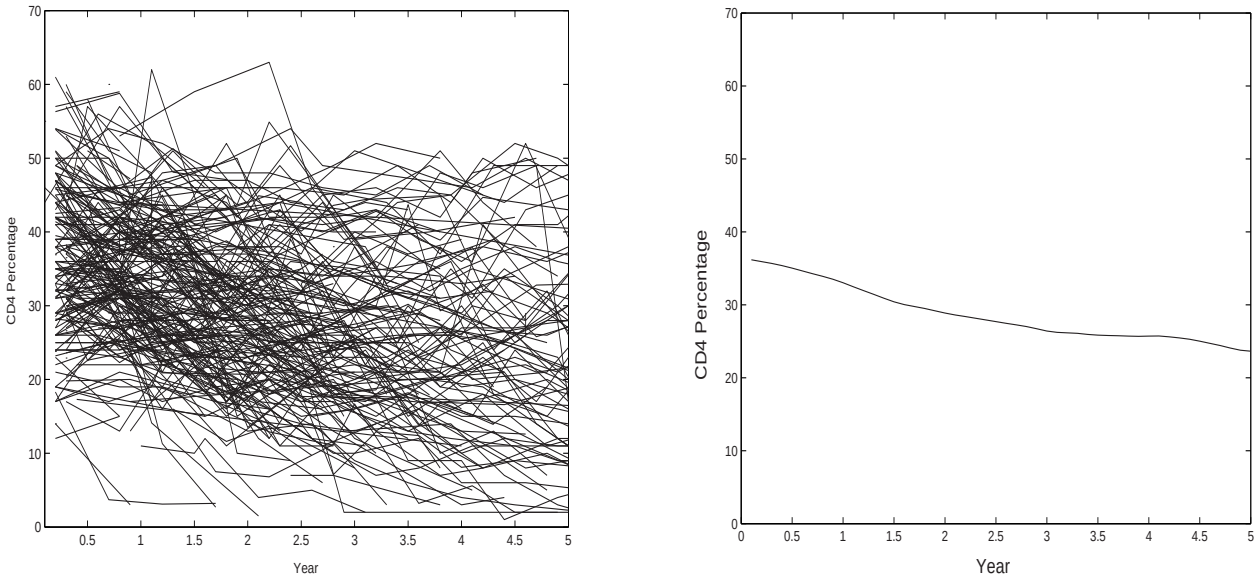

Figure 1: Left panel: Observed individual trajectories of 283 sequences of CD4 percentages. Right panel: Smooth estimate of the mean function.

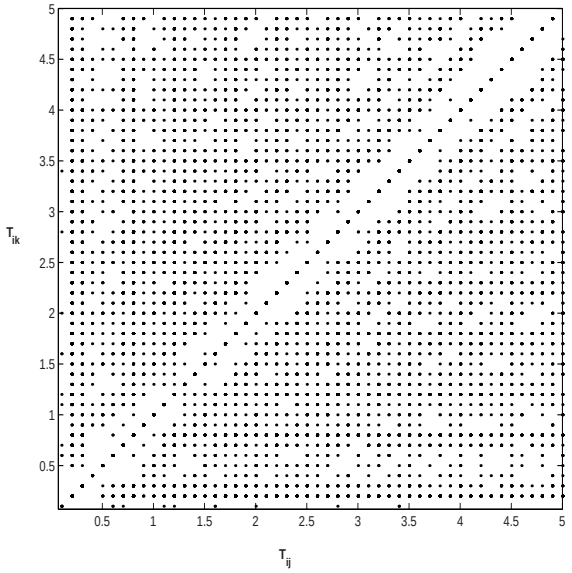

Figure 2: Assembled pairs (T_{ij}, T_{ik}) of all subjects, $i = 1, \ldots, n$, $j, k = 1, \ldots, N_i$, for the CD4 count data. While the data available per subject are sparse, the assembled data fill the domain of the covariance surface quite densely.

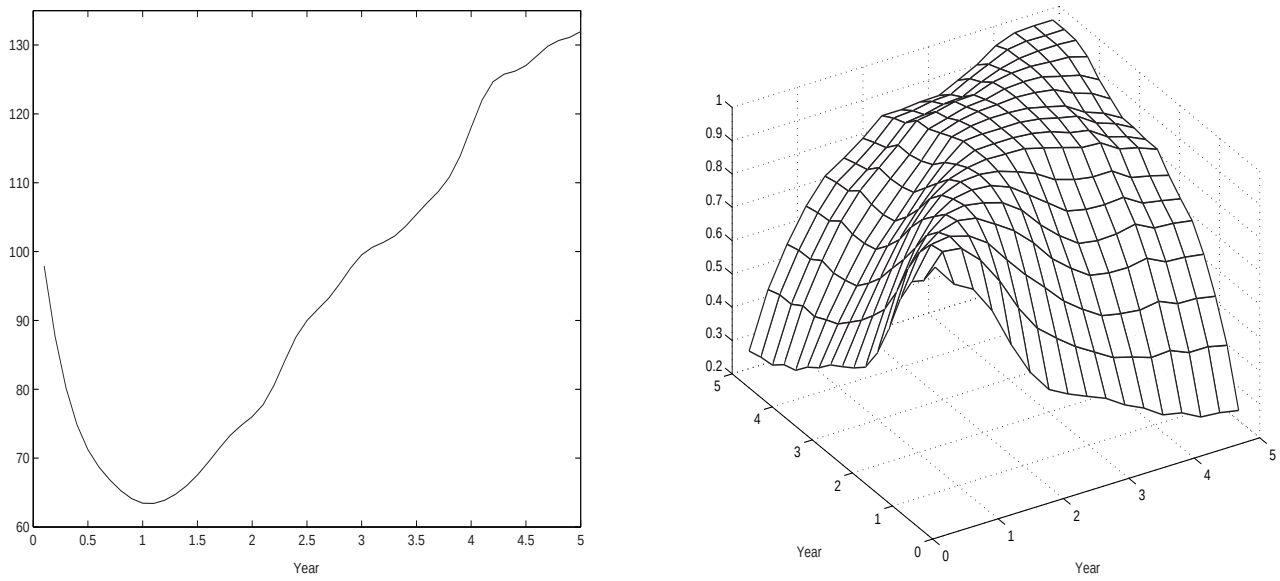

Figure 3: Left panel: Smooth estimate of the variance function for CD4 count data. Right panel: Smooth estimate of the correlation function, eliminating the "raw" data falling on the diagonal.

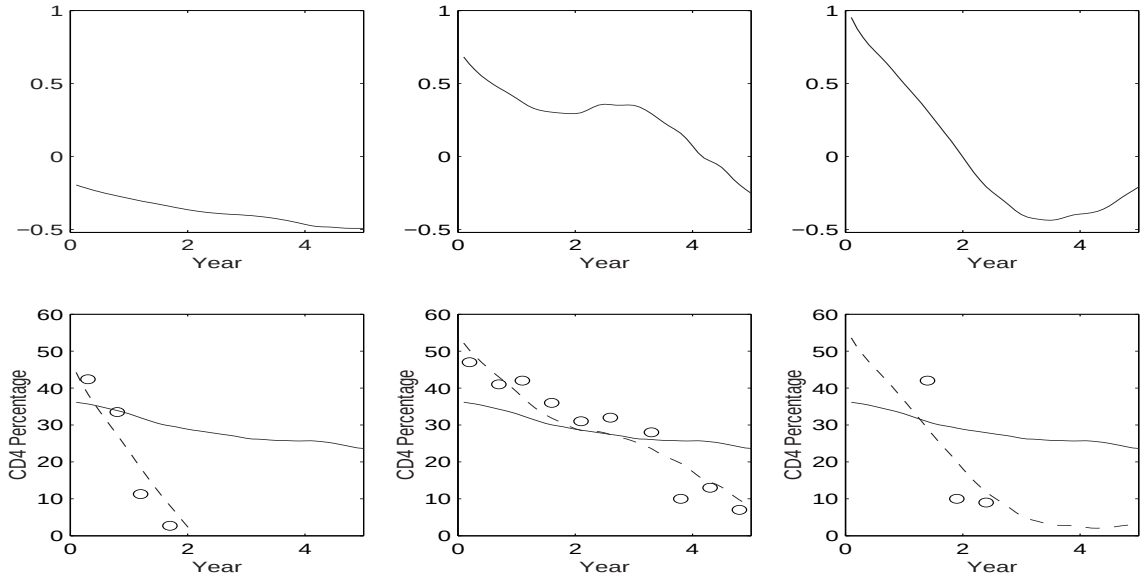

Figure 4: Top panels: Smooth estimates of the first three eigenfunctions, from left to right, for CD4 count data. Bottom panels: Observations (circles) and predicted trajectories (dashed) for the three individuals with the largest projections on the respective eigenfunctions above, overlaid with the overall estimated mean function (solid).

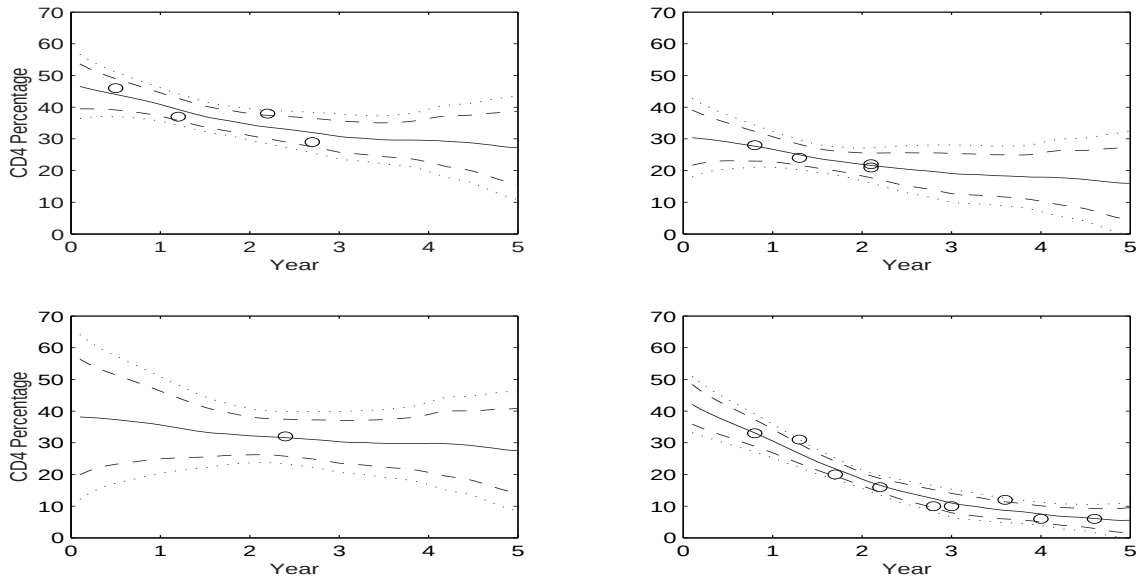

Figure 5: Observations (circles), predicted (solid) trajectories, 95% pointwise (dashed) and simultaneous (dotted) bands for four randomly chosen individuals, for CD4 count data.

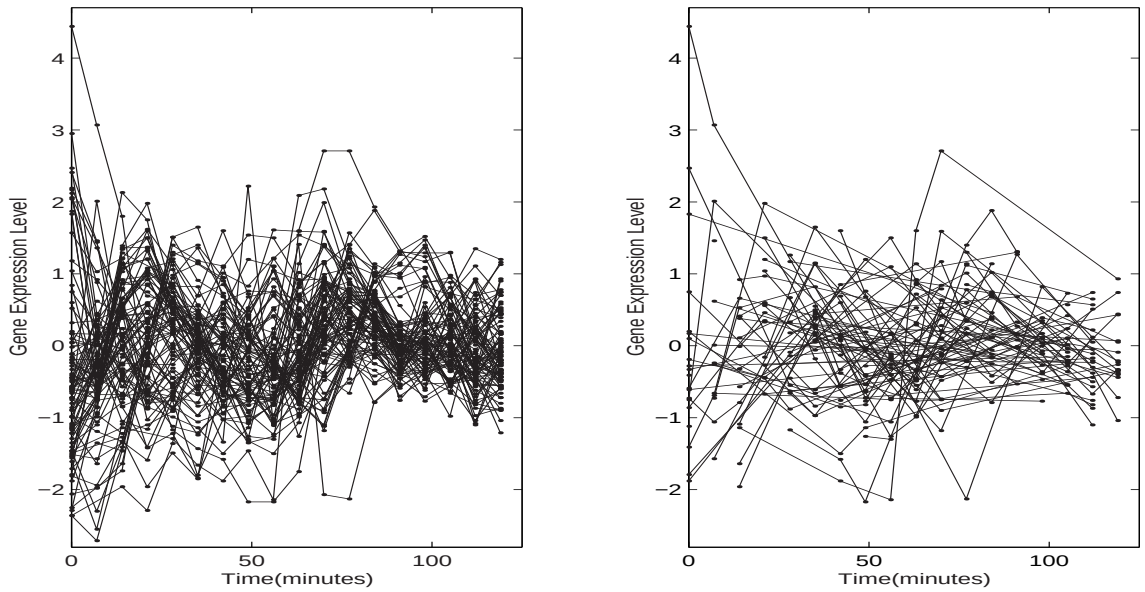

Figure 6: Complete measurements (left panel) of gene expression profiles and a randomly "sparsified" subset (right panel) for 92 yeast cell cycles.

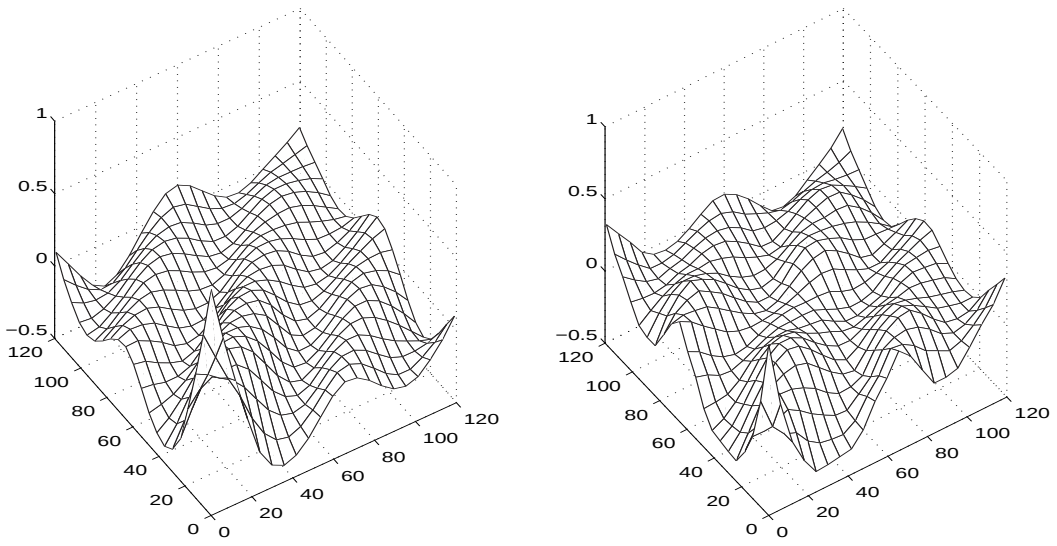

Figure 7: Smooth surface estimates $\widehat{G}$ (27) of the covariance functions obtained from the complete data (left panel) and from the sparsified data (right panel) for yeast cell cycle gene expression profiles.

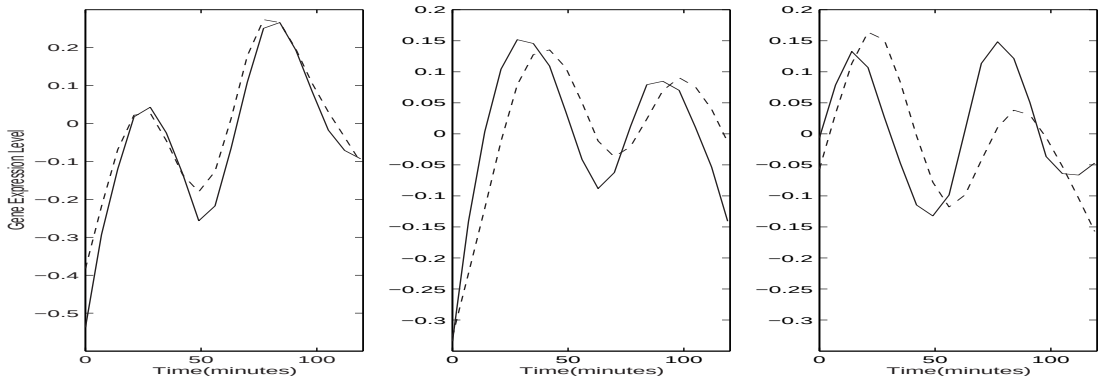

Figure 8: Smooth estimates of the mean function (left panel), the first (middle panel) and second (right panel) eigenfunctions, obtained from sparse (solid) and complete (dashed) gene expression data.

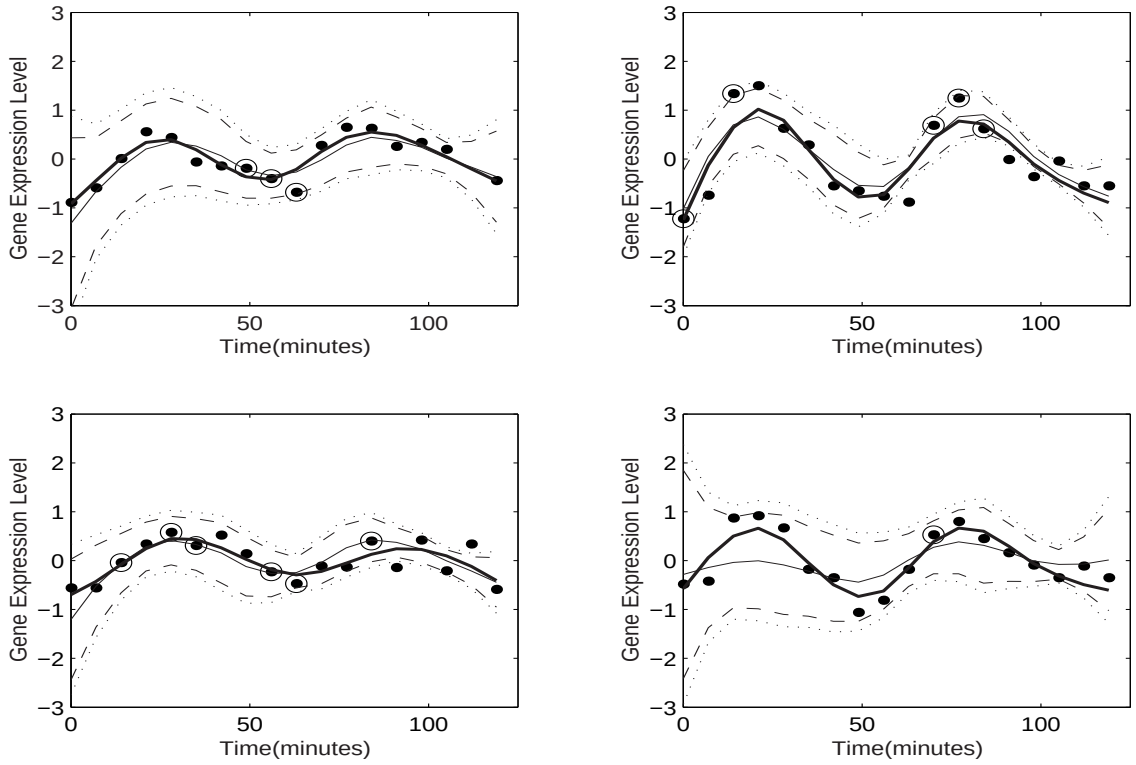

Figure 9: Predicted gene expression profiles obtained from complete measurements (thick solid) and from sparse measurements (solid) for four randomly selected genes. Also shown are 95% pointwise (dashed) and simultaneous (dotted) bands obtained exclusively from the sparse data. Solid circles indicate the measurements for the complete data, and solid circles enclosed by an open circle correspond to the randomly sampled sparse data.