11-9-2011

CORRECTED CONFIDENCE BANDS FOR FUNCTIONAL DATA USING PRINCIPAL COMPONENTS

Jeff Goldsmith
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, jgoldsmi@jhsph.edu

Sonja Greven
Department of Statistics, Ludwig-Maximilians-University Munich

Ciprian M. Crainiceanu
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Corrected Confidence Bands for Functional Data Using Principal Components

BY JEFF GOLDSMITH, SONJA GREVEN AND CIPRIAN CRAINICEANU

November 8, 2011

Abstract

Functional principal components (FPC) analysis is widely used to decompose and express functional observations. Curve estimates implicitly condition on basis functions and other quantities derived from FPC decompositions; however these objects are unknown in practice. In this paper, we propose a method for obtaining correct curve estimates by accounting for uncertainty in FPC decompositions. Additionally, pointwise and simultaneous confidence intervals that account for both model-based and decomposition-based variability are constructed. Standard mixed-model representations of functional expansions are used to construct curve estimates and variances conditional on a specific decomposition. A bootstrap procedure is implemented to understand the uncertainty in principal component decomposition quantities. Iterated expectation and variance formulas combine both sources of uncertainty by combining model-based conditional estimates across the distribution of decompositions. Our method compares favorably to competing approaches in simulation studies that include both densely- and sparsely-observed functions. We apply our method to sparse observations of CD4 cell counts and to dense white-matter tract profiles. Code for the analyses and simulations is publicly available, and our method is implemented as the `IVfpca()` function in the `R` package `refund` on CRAN.

Keywords: Bootstrap; functional principal components analysis; iterated expectation and variance; simultaneous bands.

1 Introduction

Functional principal components (FPC) analysis, in which individual curves are expressed as a linear combination of population-level basis functions and curve-specific scores, is a key tool in functional data analysis. FPC analysis has applications in dimension reduction, estimation of curves in the presence of measurement error, and the expression of traditional longitudinal data as sparsely observed functional data. Despite the ubiquity of FPC analysis in the functional data literature, principal components expansions are poorly understood in terms of their sources of variability. In this paper, we develop a method for the estimation of curves that accounts for uncertainty in the model-based estimation of curve-specific scores for a particular FPC decomposition as well as uncertainty in the FPC decomposition itself. Further, we construct confidence intervals for the curves that include both model-based and decomposition-based sources of variability.

Using FPC analysis, individual curves are expressed as a linear combination of population-level basis functions and curve-specific scores. These scores can be predicted as the best linear unbiased predictors (BLUPs) in a mixed model framework for a particular FPC decomposition, which includes the basis functions and score variances. Thus FPC analysis allows a parsimonious, nonparametric representation of functional data. Additionally, mixed model results allow the construction of confidence intervals for estimated curves, again for a particular decomposition. However, in practice the decomposition is based on the estimated covariance structure of the population of curves and is subject to variability. This added variability can be substantial in many settings, including when a small number of curves is observed and when the curves are observed sparsely at the subject level, yet is often overlooked. Accurate estimation and inference for functional expansions is important is many scientific settings, such as understanding CD4 cell count trajectories in HIV positive individuals based on infrequent measurements or quantifying intracranial white-matter structure using diffusion tensor imaging.

Our proposed method explicitly conditions on a particular FPC decomposition when predicting curve-specific scores and estimating the associated variability. Next, we use iterated expectation and variance formulas to combine uncertainty in the conditional estimates across the distribution of FPC decomposition objects (the basis functions, mean function, measurement error and score variances, and the truncation lag). We use the bootstrap to derive an empirical distribution of decomposition objects by resampling

curves with replacement and, for each sample, constructing an FPC decomposition. Thus, our method accounts for two major sources of variability: 1) the model-based uncertainty conditional on a particular decomposition; and 2) the uncertainty in the estimated FPC decomposition objects. Moreover, it is applicable when curves are sparsely or densely sampled, for small and large sample sizes and when the curves are measured with error.

FPC analysis has a long history in the statistical literature, dating at least to the treatment of growth curves in Rao (1958). For a modern introduction to the topic, see Ramsay and Silverman (2005), Chapter 8, or Jolliffe (2002), Section 12.3. Recently, Yao et al. (2005) proposed the use of BLUPs to predict curve-specific scores for FPC expansions in a mixed model framework; this approach is related to that of Shi et al. (1996) and Rice and Wu (2001), in which curves are expressed using penalized spline bases in a mixed model, but can be used when curves are sparsely observed. BLUPs have also been used in multilevel and longitudinal functional principal components analysis by Di et al. (2009) and Greven et al. (2010), respectively, to predict scores in more complex decompositions. The mixed model framework has allowed the use of Bayesian techniques for FPC analysis (Crainiceanu and Goldsmith, 2010) and the joint modeling of curves and scalar outcomes in functional regression models (Goldsmith et al., Accepted). Asymptotic consistency and distributions of scores predicted in the mixed model framework are developed by Yao et al. (2005), and asymptotic properties of FPC decompositions are considered in Hall and Hosseini-Nasab (2006). James et al. (2000) use a penalized spline approach to estimate FPC basis functions directly, rather than through decomposing an estimated covariance matrix. This paper also uses the bootstrap to generate pointwise confidence intervals by taking the empirical quantiles of estimated functions; importantly, this approach accounts for uncertainty in the FPC decompositions, but not for model-based uncertainty. The bootstrap is similarly used with functional data by Hall and Hosseini-Nasab (2006) to construct confidence intervals for FPC basis functions, and by Crainiceanu et al. (2011b) for inference on the mean difference curve between two groups.

The remainder of the paper is organized as follows. In Section 2 we review FPC analysis and the estimation of curves in a mixed model framework. Section 3 introduces our method for including uncertainty in principal component expansions in the estimation of curves and understanding the associated variability. Simulation results comparing several methods are presented and discussed in Section 4. Two

statistically and scientifically distinct applications are considered in Section 5. We close with a discussion in Section 6.

2 Mixed-model Framework for FPC Analysis

Given curves $Y_i(d)$, $d \in [0,1]$, for $1 \leq i \leq I$, define the covariance operator $\Sigma^{\boldsymbol{Y}}(d,d') = \text{Cov}[Y_i(d), Y_i(d')]$ so that $\Sigma^{\boldsymbol{Y}}(d,d')$ is a bivariate function giving the covariance between two locations on the same curve. Let $\sum_{k=1}^{\infty} \lambda_k \phi_k(d)\phi_k(d')$ be the spectral decomposition of $\Sigma^{\boldsymbol{Y}}(d,d')$, where $\boldsymbol{\phi}(d) = \{\phi_k(d) : k \in \mathbb{Z}^+\}$ are orthonormal eigenfunctions and $\lambda_1 \geq \lambda_2 \geq \ldots$ are the corresponding non-increasing eigenvalues. Based on this decomposition, a Karhunen-Loève expansion for $Y_i(d)$ is $Y_i(d) = \mu(d) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(d)$, where the $\xi_{ik} = \int_0^1 \{Y_i(d) - \mu(d)\}\, \phi_k(d) dd$ are uncorrelated random variables with variance λ_k, and $\mu(d) = \text{E}[Y(d)]$.

In practice, we observe the functions $Y_i(d)$ with error – that is, we observe $\tilde{Y}_i(d) = Y_i(d) + \epsilon_i(d)$ where $\epsilon_i(d) \sim \text{N}\left[0, \sigma^2\right]$. Moreover, functions $Y_i(d)$ are observed on finite grids $\{d_{ij}\}_{j=1}^{J_i}$ that are often sparse at the subject level. These practical concerns impact the estimation of the mean function, the FPC decomposition, and the Karhunen-Loève expansion. First, the mean function is estimated using penalized splines fit to the pooled observations under working independence; the smoothing parameter is selected via restricted maximum likelihood (REML). Next, we use a method of moments approach to construct a raw estimate of the covariance matrix and smooth the off-diagonal elements to obtain $\tilde{\Sigma}^{\boldsymbol{Y}}(d,d')$ (Staniswalis and Lee, 1998; Yao et al., 2003). Smoothing of the raw covariance matrix is performed using bivariate penalized splines, again with smoothing parameters chosen via REML. For both the mean function and the covariance matrix, other smoothing approaches could be used without modification of the underlying technique. The retained estimated principal component basis functions $\hat{\boldsymbol{\phi}}(d) = \{\hat{\phi}_k(d) : k \in \{1, \ldots, \hat{K}\}\}$ and score variances $\{\hat{\lambda}_k : k \in \{1, \ldots, \hat{K}\}\}$ are based on the decomposition of $\tilde{\Sigma}^{\boldsymbol{Y}}(d,d')$; the estimate $\hat{K}$ of the truncation K lag is the minimum number of components needed to explain 99% of the variability in the observed curves, although other methods for estimating K are possible. The estimate of the covariance matrix $\Sigma^{\boldsymbol{Y}}(d,d')$ is given by

$$\hat{\Sigma}^{\boldsymbol{Y}}(d,d') = \sum_{k=1}^{\hat{K}} \hat{\lambda}_k \hat{\phi}_k(d)\hat{\phi}_k(d') = \hat{\boldsymbol{\phi}}(d)\hat{\Lambda}\hat{\boldsymbol{\phi}}^T(d') \tag{1}$$

where $\hat{\Lambda}$ is a diagonal matrix with elements $\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{K}}$ based on the retained basis functions and score variances. The measurement error variance σ^2 is estimated as the average difference between the middle 60% of diagonal elements of the raw covariance matrix and $\hat{\Sigma}^{\boldsymbol{Y}}(d, d')$; omitting the extreme diagonal elements provides more stable estimates of the measurement error variance.

Numeric integration for the estimation of the scores may be inaccurate when the curves are sparsely observed. With this in mind, Yao et al. (2005) predict scores $\boldsymbol{\xi}_i = \{\xi_{ik} : k \in \hat{K}\}$ using best linear unbiased predictors, effectively basing inference on the the mixed model

$$\begin{aligned} Y_i(d_{ij}) &= \mu(d_{ij}) + \sum_{k=1}^{K} \xi_{ik}\phi_k(d_{ij}) + \epsilon_i(d_{ij}) \\ \boldsymbol{\xi}_i &\overset{iid}{\sim} \mathrm{N}\left[0, \Lambda\right], \epsilon_i(d_{ij}) \overset{iid}{\sim} \mathrm{N}\left[0, \sigma^2\right] \end{aligned} \tag{2}$$

where $\boldsymbol{\xi}_i$ and $\boldsymbol{\epsilon}_i = \{\epsilon_i(d_{i1}), \ldots, \epsilon_i(d_{iJ_i})\}$ are independent. Notationally, let $\boldsymbol{\theta} = \{\boldsymbol{\phi}, \mu, \Lambda, \sigma^2, K\}$ be the collection of unobserved FPC decomposition objects: the basis functions $\boldsymbol{\phi}$ evaluated over the domain of d; the mean function μ evaluated over the domain of d; the matrix of score variances $\Lambda = \mathrm{diag}\{\lambda_1, \ldots \lambda_K\}$; the measurement error variance σ^2; and the truncation lag K. Similarly, let $\hat{\boldsymbol{\theta}} = \{\hat{\boldsymbol{\phi}}, \hat{\mu}, \hat{\Lambda}, \hat{\sigma}^2, \hat{K}\}$ be the estimated FPC decomposition objects. Also, let $\boldsymbol{d}_i = \{d_{ij}\}_{j=1}^{J_i}$ and $\boldsymbol{d}_g$ be a dense grid on the domain of d, often taken as the union of the $\boldsymbol{d}_i$. In the following, $\tilde{Y}_i(\boldsymbol{d}_i)$ will represent the vector of observations for curve i, the vector $\phi_k(\boldsymbol{d}_i)$ will be the kth basis function evaluated at locations $\boldsymbol{d}_i$, and the $J_i \times K$ matrix $\boldsymbol{\phi}(\boldsymbol{d}_i)$ will be the collection of basis functions evaluated at locations $\boldsymbol{d}_i$. Analogous notation will represent functions evaluated over the dense grid $\boldsymbol{d}_g$. We emphasize the conditioning on the FPC decomposition objects notationally by conditioning on $\hat{\boldsymbol{\theta}}$.

The estimated best linear unbiased predictors for the loadings are

$$\hat{\boldsymbol{\xi}}_{\hat{\boldsymbol{\theta}},i} = \mathrm{E}\left[\boldsymbol{\xi}_i | \tilde{Y}_i(\boldsymbol{d}_i), \hat{\boldsymbol{\theta}}\right] = \left(\hat{\boldsymbol{\phi}}^T(\boldsymbol{d}_i)\hat{\boldsymbol{\phi}}(\boldsymbol{d}_i) + \hat{\sigma}^2\hat{\Lambda}^{-1}\right)^{-1} \hat{\boldsymbol{\phi}}^T(\boldsymbol{d}_i)\left(\tilde{Y}_i(\boldsymbol{d}_i) - \hat{\mu}(\boldsymbol{d}_i)\right). \tag{3}$$

A truncated Karhunen-Loève expansion for the ith curve over the dense grid $\boldsymbol{d}_g$, based on the objects

estimated in the FPC decomposition and the predicted scores, is

$$\begin{aligned}\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) = \mathrm{E}\left[Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\xi}}_{\hat{\boldsymbol{\theta}},i}, \hat{\boldsymbol{\theta}}\right] &= \hat{\mu}(\boldsymbol{d}_g) + \sum_{k=1}^{\hat{K}} \hat{\xi}_{\hat{\boldsymbol{\theta}},ik}\hat{\phi}_k(\boldsymbol{d}_g) \\ &= \hat{\mu}(\boldsymbol{d}_g) + \hat{\boldsymbol{\phi}}(\boldsymbol{d}_g)\hat{\boldsymbol{\xi}}_{\hat{\boldsymbol{\theta}},i}. \end{aligned} \tag{4}$$

Model-based covariances for the predicted scores are used to obtain the covariance operator for the estimated curve

$$\mathrm{Var}\left[\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right] = \hat{\boldsymbol{\phi}}(\boldsymbol{d}_g)\left(\frac{1}{\hat{\sigma}^2}\hat{\boldsymbol{\phi}}^T(\boldsymbol{d}_i)\hat{\boldsymbol{\phi}}(\boldsymbol{d}_i) + \hat{\Lambda}^{-1}\right)^{-1}\hat{\boldsymbol{\phi}}^T(\boldsymbol{d}_g); \tag{5}$$

the curve specific covariance in (5) is similar to the estimate of $\Sigma^{\boldsymbol{Y}}$ in (1), but is decreased according to the term $\frac{1}{\hat{\sigma}^2}\hat{\boldsymbol{\phi}}(\boldsymbol{d}_i)^T\hat{\boldsymbol{\phi}}(\boldsymbol{d}_i)$. This covariance focuses on the variability in the curve-specific expansion, and does not include uncertainty in the estimated overall mean function $\hat{\mu}(\boldsymbol{d}_g)$. Next, approximate $(1-\alpha)\%$ level pointwise confidence intervals for the predicted curves are given by

$$\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathrm{diag}\left\{\mathrm{Var}\left[\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right]\right\}} \tag{6}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. Similarly, approximate $(1-\alpha)\%$ level simultaneous intervals are given by

$$\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) \pm \sqrt{\chi^2_{K,(1-\alpha)}\mathrm{diag}\left\{\mathrm{Var}\left[\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right]\right\}} \tag{7}$$

where $\chi^2_{K,(1-\alpha)}$ is the $(1-\alpha)$ quantile of a chi-squared distribution with K degrees of freedom. The intervals in (6) and (7) are justified by asymptotic arguments in Yao et al. (2005).

3 Correct Expansions and Inference

The mixed model framework developed in Section 2 explicitly conditions on the FPC decomposition objects $\boldsymbol{\theta}$ and, given the basis functions, mean function, score and measurement error variances, and truncation lag, derives best linear predictors for curve-specific loadings and variance estimates for functional

expansions. When these elements are fixed at their true values, standard mixed model results support the construction described above (Ruppert et al., 2003). In practice, the elements of the FPC decomposition are unknown and must be estimated from the data. As noted, asymptotic work in Yao et al. (2005) can be used to justify the above construction for large samples, but this approach can lead to a substantial understatement of the total variability in the estimated curves when the uncertainty in the decomposition is high.

Our proposed method accounts for both model-based and decomposition-based uncertainty in estimating curves and in assessing the variability of these estimates. Briefly, the framework described in Section 2 is used to construct estimates of curves and their associated variability given a particular decomposition. Uncertainty in the distribution of decompositions is examined through a bootstrap procedure; that is, we resample curves with replacement and construct FPC decompositions based on the bootstrap sample. Finally, estimated curves are given by averaging the model-based estimates across the distribution of decompositions using iterated expectation, and total variability is understood using the iterated variance formula. Because we estimate all objects of the FPC decomposition $\boldsymbol{\theta} = \{\phi, \mu, \Lambda, \sigma^2, K\}$ within each bootstrap sample, uncertainty in all of these objects is accounted for by our method. This holds if other methods are used to estimate the individual components of $\boldsymbol{\theta}$ so that, for example, other smoothing approaches can be used for the covariance operator and other rules can be used to select the truncation lag K.

More specifically, we draw a sample $\boldsymbol{Y}_B$ of I functions from the full data $\boldsymbol{Y}$ with replacement. Based on this sample, we derive an estimate of the covariance operator $\hat{\Sigma}_B^{\boldsymbol{Y}}$ by smoothing the off-diagonal elements of the raw covariance matrix and obtain an estimate of the FPC decomposition objects $\hat{\boldsymbol{\theta}}_B = \{\hat{\phi}_B, \hat{\mu}_B, \hat{\Lambda}_B, \hat{\sigma}_B^2, \hat{K}_B\}$. Conditioning on $\hat{\boldsymbol{\theta}}_B$, we proceed as in Section 2 and estimate the scores $\hat{\boldsymbol{\xi}}_{\hat{\boldsymbol{\theta}}_B,i}$, curves $\hat{Y}_{\hat{\boldsymbol{\theta}}_B,i}(\boldsymbol{d}_g)$, and variances $\text{Var}\left[\hat{Y}_{\hat{\boldsymbol{\theta}}_B,i}(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}_B\right]$ for each element of the full dataset $\boldsymbol{Y}$. Thus, resampling is used only to assess uncertainty in the decomposition – given each bootstrap decomposition, estimates and variances are constructed for each curve Y_i. Model-based expectations and variances condition on a particular decomposition, but can be estimated for curves Y_i that are not contained in the bootstrap sample $\boldsymbol{Y}_B$.

Next, we combine information across bootstrap samples to estimate curves and construct variability

estimates that account for uncertainty in the FPC decompositions. First, we use iterated expectations to estimate curves over the dense grid $\boldsymbol{d}_g$:

$$\hat{Y}_i(\boldsymbol{d}_g) \quad = \quad \mathrm{E}\left[Y_i(\boldsymbol{d}_g)\right] = \mathrm{E}_{\hat{\boldsymbol{\theta}}}\left\{\mathrm{E}_{Y_i|\hat{\boldsymbol{\theta}}}\left[Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\xi}}_{\hat{\boldsymbol{\theta}},i}, \hat{\boldsymbol{\theta}}\right]\right\}. \tag{8}$$

Thus, estimated curves are the average of model-based estimates taken over the distribution of decomposition objects in $\boldsymbol{\theta}$. This is estimated taking the average of the model-based estimates across our bootstrap samples. The total covariance operator of the estimated curves is based on the iterated variance formula:

$$\mathrm{Var}\left[\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)\right] \quad = \quad \mathrm{E}_{\hat{\boldsymbol{\theta}}}\left[\mathrm{Var}_{\hat{Y}|\hat{\boldsymbol{\theta}}}\left(\hat{Y}_{\hat{\boldsymbol{\theta}},i}(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right)\right] + \mathrm{Var}_{\hat{\boldsymbol{\theta}}}\left[\mathrm{E}_{\hat{Y}|\hat{\boldsymbol{\theta}}}\left(\hat{Y}(\boldsymbol{d}_g)|\hat{\boldsymbol{\xi}}_{\hat{\boldsymbol{\theta}},i}, \hat{\boldsymbol{\theta}}\right)\right]. \tag{9}$$

The conditional mean and variance are the model-based quantities derived for a given FPC decomposition. Therefore, the left term appearing in (9) gives the average model-based uncertainty across the distribution of decompositions, and is estimated for each curve i by taking the average of covariance matrices estimated within each bootstrap sample. The right term inflates this by the variance in the estimated functions taken over the distribution of decompositions, and is estimated using the covariance matrix of the estimates of curve i across bootstrap samples. Pointwise confidence intervals for the predicted curves are given by

$$\hat{Y}_i(\boldsymbol{d}_g) \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\mathrm{diag}\left\{\mathrm{Var}\left[\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)\right]\right\}}. \tag{10}$$

Simultaneous confidence intervals for FPC expansions are constructed similarly to intervals for penalized splines (Ruppert et al., 2003). Specifically, a $100(1-\alpha)\%$ simultaneous confidence interval for $Y_i(\boldsymbol{d}_g)$ is given by

$$\hat{Y}_i(\boldsymbol{d}_g) \pm m_{(1-\alpha)}\sqrt{\mathrm{diag}\left\{\mathrm{Var}\left[\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)\right]\right\}}. \tag{11}$$

where $m_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of the random variable

$$\max_{d \in \boldsymbol{d}_g} \left| \frac{\hat{Y}_i(d) - Y_i(d)}{\sqrt{\text{diag}\left\{\text{Var}\left[\hat{Y}_i(d) - Y_i(d)\right]\right\}}} \right|. \tag{12}$$

To estimate $m_{(1-\alpha)}$, we note that $\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)$ is approximately multivariate normal centered at $\mathbf{0}$ and with the total covariance found in (9):

$$\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g) \overset{\text{approx.}}{\sim} \text{N}\left[\mathbf{0}, \text{Var}\left(\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)\right)\right] \tag{13}$$

We repeatedly draw functions with the estimated covariance from this distribution, and for each draw calculate the resulting value of (12); the $(1-\alpha)$ quantile of these values is used as an estimate of $m_{(1-\alpha)}$. Because this sampling is done only once after all bootstrap samples have been used, estimating simultaneous intervals does not add considerably to the total computation time of this approach.

The construction of pointwise intervals in (10) and simultaneous intervals in (11) assumes that the estimated curve averaged over the distribution of FPC decompositions is approximately Normally distributed around the true curve. In practice, the distribution of curves may exhibit skewness, overdispersion, or multimodality; in this case, the expected value and variance defined in (8) and (9) may not completely specify the distribution. Alternatively, one might pose more flexible parametric distributions or use empirical distributions to describe the behavior of curves accounting for variability in FPC decompositions. However, asymptotic results in Yao et al. (2005) justify the Normal distribution for large sample sizes, and simulation studies in Section 4 indicate that the Normality assumption is reasonable and provides confidence intervals that achieve the nominal coverage in a variety of situations, including small sample sizes and sparse observations.

4 Simulations

In this section, we conduct a simulation study to examine the accuracy of the estimated curves and the coverage probabilities of confidence intervals based on our proposed method.

We construct datasets using the following model

$$Y_i(d) \quad = \quad \mu(d) + \sum_{k=1}^{4} \xi_{ik}\phi_k(d) + \epsilon_i(d)$$

for d observed on the dense equally-spaced grid $\boldsymbol{d}_g = \{(t-.5)/50, t = 1, \ldots, 50\}$, $\mu(d) = \frac{1}{4}d$ and $\epsilon_i(d) \sim \mathrm{N}\left[0, \sigma^2\right]$; two scenarios for the distribution of scores $\boldsymbol{\xi}_i$ are given below. The eigenfunctions are taken to be $\boldsymbol{\phi} = \{\phi_1(d) = 1; \phi_2(d) = \sqrt{3}(2d-1); \phi_3(d) = \sqrt{5}(6d^2-6d+1); \phi_4(d) = \sqrt{7}(20d^3-30d^2+12d-1)\}$ and the variances are $\lambda_k = \{.75^{k-1}\}_{k=1}^{4}$. Five hundred such datasets are constructed for all possible combinations of the scenarios below:

1. Sample sizes (a) $I = 20$; (b) $I = 50$; (c) $I = 100$;

2. Measurement error variances (a) $\sigma^2 = .0025$; (b) $\sigma^2 = .01$;

3. Number of points observed per curve (a) $J_i = 50$; (b) $J_i = 20$; (c) an unbalanced design. For the unbalanced design, the number of points observed per curve follow a Pois[$\lambda = 15$] distribution; for sparse designs (b) and (c), the locations of observed points are uniformly distributed within $\boldsymbol{d}_g$.

4. a) Normally distributed scores: $\xi_{ik} \sim \mathrm{N}\left[0, \lambda_k\right]$;

   b) Non-normally distributed scores (mixture of normals): ξ_{ik} drawn with equal probability from either $\mathrm{N}\left[-\sqrt{\lambda_k/2}, \lambda_k/2\right]$ or $\mathrm{N}\left[\sqrt{\lambda_k/2}, \lambda_k/2\right]$.

This gives a total of 36 possible designs.

For each dataset, we estimate curves and construct 95% pointwise confidence intervals using the method described in Section 3 with 100 bootstrap samples per dataset; additional unreported simulations indicate that this number of bootstrap samples is sufficient to understand the uncertainty in the estimation of FPC decomposition objects in the scenarios we consider. We estimate curves and confidence intervals using three methods: (i) the iterated expectation and variance method proposed in Section 3, labeled "IV" in Figures, Tables and the text below; (ii) the mixed model approach estimating the FPC decomposition objects $\boldsymbol{\theta}$ from the data, labeled "MM - $\hat{\boldsymbol{\theta}}$"; and (iii) the mixed model approach fixing the FPC decomposition objects $\boldsymbol{\theta}$ at their true values, labeled "MM - $\boldsymbol{\theta}$". In addition, when examining pointwise intervals we consider using the 2.5% and 97.5% quantiles of the curves estimated within each bootstrap sample,

labeled "Bootstrap". For simulated datasets and bootstrap samples in which the measurement error variance σ^2 is estimated to be zero, the model-based variance estimates are also zero. This is most common for scenarios in which $\sigma^2 = .0025$, $J_i = 20$ or the design is unbalanced, and $I = 20$. Such datasets are omitted from the results below.

Table 1 provides the average mean squared error AMSE $= \frac{1}{I}\sum_{i=1}^{I}\left[\frac{1}{50}\sum_{d\in \boldsymbol{d}_g}\left(\hat{Y}_i(d) - Y_i(d)\right)^2\right]$ of the curves estimated from IV, MM - $\hat{\boldsymbol{\theta}}$, and MM - $\boldsymbol{\theta}$ approaches using non-normally distributed scores (results for normally distributed scores are similar and reported in an online supplement). For reference on the values in this table, Figure 1 displays two examples of simulated curves, the observed data points, and curve estimates with the MSE for each estimate. The example in the left panel comes from a simulation

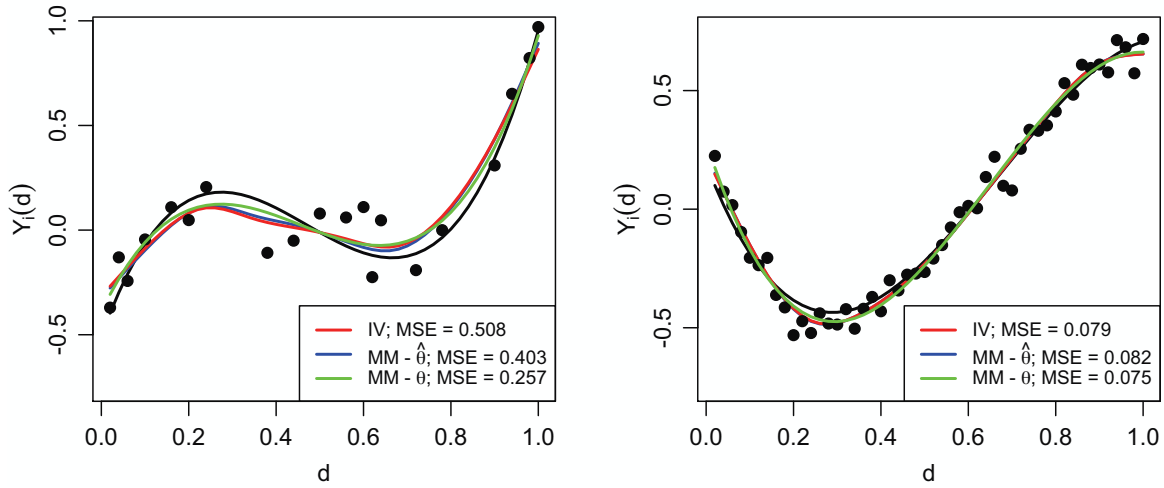

Figure 1: Comparison of three methods for estimating curves: the proposed iterated expectation method ("IV"), the model-only approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach fixing $\boldsymbol{\theta}$ at the truth ("MM - $\boldsymbol{\theta}$"). The left panel displays a curve from a simulated dataset with $I = 20$, $J_i = 20$, and $\sigma^2 = .01$; the right panel comes from a simulation with $I = 50$, $J_i = 50$, and $\sigma^2 = .0025$. MSEs are provided for each estimate.

with $I = 20$, $J_i = 20$, and $\sigma^2 = .01$, while the example in the right comes from a simulation with $I = 50$, $J_i = 50$, and $\sigma^2 = .0025$. For densely observed curves, Table 1 indicates that the IV and MM - $\hat{\boldsymbol{\theta}}$ approaches have similar performances, but that the IV method modestly outperforms the model-based approach when curves are sparsely observed. Moreover, these methods are most similar to the MM - $\boldsymbol{\theta}$ approach when curves are fully observed. As expected, performance improves as sample size increases and when the measurement error variance is smaller.

Next we consider inference on estimated curves. Table 2 displays the average coverage, taken across

	IV			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	0.026	0.259	0.509	0.026	0.268	0.561	0.020	0.065	0.138
$\sigma^2 = .01$	0.100	0.475	0.814	0.099	0.485	0.834	0.078	0.242	0.416
$I = 50; \sigma^2 = .0025$	0.023	0.145	0.303	0.023	0.166	0.342	0.020	0.064	0.143
$\sigma^2 = .01$	0.088	0.340	0.600	0.088	0.355	0.627	0.079	0.240	0.423
$I = 100; \sigma^2 = .0025$	0.022	0.109	0.233	0.022	0.121	0.266	0.020	0.064	0.140
$\sigma^2 = .01$	0.083	0.295	0.519	0.083	0.305	0.535	0.078	0.240	0.418

Table 1: $100 \times$ AMSE, taken over all grid points and subjects, of estimated curves constructed using the proposed iterated expectation method ("IV"), the mixed model approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach with $\boldsymbol{\theta}$ fixed at the true value ("MM - $\boldsymbol{\theta}$"). Curves are sampled with $J_i = 50$, $J_i = 20$, and using an unbalanced design ("Unbal").

all grid points and all subjects, of pointwise confidence intervals for simulation scenarios with non-normally distributed scores (again, results for normally distributed scores are similar and reported in an online supplement). The results in Table 2 demonstrate the reliability of the IV method, with coverages

	IV			Bootstrap			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	95.3	97.3	97.0	63.6	86.2	83.8	90.3	68.4	72.0	95.0	94.9	95.0
$\sigma^2 = .01$	95.2	94.7	93.8	61.2	75.7	73.9	90.9	81.2	80.0	95.0	94.9	95.0
$I = 50; \sigma^2 = .0025$	94.8	95.3	96.6	46.4	77.2	75.9	93.0	75.6	78.5	95.0	94.9	95.0
$\sigma^2 = .01$	94.8	94.3	93.9	43.9	61.4	61.8	93.2	88.6	86.8	95.0	94.9	95.0
$I = 100; \sigma^2 = .0025$	94.9	93.8	95.0	35.3	70.4	68.7	94.1	79.9	79.2	95.0	95.0	95.0
$\sigma^2 = .01$	94.9	94.7	94.2	33.6	52.3	51.6	94.1	91.5	90.5	95.0	95.0	95.0

Table 2: Average pointwise coverage, taken over all grid points and subjects, of 95% confidence intervals constructed using the proposed iterated variance method ("IV"), the approach that uses the quantiles of bootstrapped curve estimates ("Bootstrap"), the mixed model approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach with $\boldsymbol{\theta}$ fixed at its true value ("MM - $\boldsymbol{\theta}$"). Coverages are expressed as percents.

between 93.8 and 97.3 for all simulation scenarios. Lower coverages are observed for both the bootstrap method and the MM - $\hat{\boldsymbol{\theta}}$ approach, particularly for smaller sample sizes and sparser observations.

The bootstrap and MM - $\hat{\boldsymbol{\theta}}$ approaches each consider only one important source of variability: the bootstrap approach captures uncertainty in the FPC decomposition objects $\boldsymbol{\theta}$ but does not consider the model-based uncertainty in the estimation of the scores, while the MM - $\hat{\boldsymbol{\theta}}$ approach includes the model-based variability in the subject-specific scores but does not consider uncertainty in the decomposition objects. As the sample size increases, $\boldsymbol{\theta}$ is better estimated, so the coverage of the MM - $\hat{\boldsymbol{\theta}}$ approach that conditions on this estimate improves. For the bootstrap interval, coverage decreases as sample size increases because $\boldsymbol{\theta}$ is better estimated within each bootstrap sample, which provides more stable estimates

of individual curves across samples. Thus, the bootstrap estimates themselves show lower variability, but do not reflect the uncertainty in the predicted curve-specific scores. Surprisingly, coverage for the MM - $\hat{\boldsymbol{\theta}}$ method is worse when the measurement error variance is low - accurate estimation of this parameter may be more difficult when it is small, which would impact the coverage of resulting confidence intervals. Finally, the MM - $\boldsymbol{\theta}$ method attains the nominal coverage in all situations, which illustrates that the low coverages in the asymptotic approach are not caused by the mixed-model framework, but by uncertainty in the FPC decomposition objects.

Table 3 provides the average coverage of simultaneous confidence intervals taken over all grid points and subjects. For the IV approach, these intervals are constructed using the quantile estimate procedure described in equations (11) - (13) in Section 3. For the MM - $\hat{\boldsymbol{\theta}}$ and MM - $\boldsymbol{\theta}$ methods, simultaneous intervals are given by equation (7) in Section 2. The coverage of the IV method is somewhat lower for simultaneous

	IV			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	94.1	96.5	95.0	86.4	34.4	41.6	98.2	98.2	98.3
$\sigma^2 = .01$	94.3	92.2	89.3	88.6	59.8	56.8	98.2	98.3	98.3
$I = 50; \sigma^2 = .0025$	93.8	93.2	96.1	94.1	49.0	56.4	98.3	98.2	98.3
$\sigma^2 = .01$	94.1	92.6	91.0	94.6	82.4	78.2	98.3	98.2	98.3
$I = 100; \sigma^2 = .0025$	94.1	89.9	92.8	96.4	59.6	59.4	98.3	98.3	98.2
$\sigma^2 = .01$	94.4	93.7	92.1	96.7	91.4	88.3	98.3	98.3	98.3

Table 3: Average coverage of 95% simultaneous confidence intervals constructed using the proposed iterated variance method ("IV"), the mixed model approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach with $\boldsymbol{\theta}$ fixed at its true value ("MM - $\boldsymbol{\theta}$"). Coverages are expressed as percents.

intervals than for pointwise intervals, with values ranging between 89.3% and 96.5%; in general, however, the intervals approach nominal coverage. On the other hand, the MM - $\hat{\boldsymbol{\theta}}$ approach has substantially worse performance in many scenarios with coverages as low as 34.4%. The effect of low measurement error variance on decreasing coverage probabilities using the mixed model approach is more evident for simultaneous intervals than for pointwise intervals, particularly for sparse data scenarios. Results for the MM - $\boldsymbol{\theta}$ method indicate that intervals constructed using equation (7) are in fact conservative, with coverages of approximately 98.2% for all simulation scenarios. In an online appendix, we discuss the adaptation of the quantile estimation procedure used to construct simultaneous intervals for the IV method to the MM - $\hat{\boldsymbol{\theta}}$ and MM - $\boldsymbol{\theta}$ methods. Additional simulations indicate that using the quantile estimation procedure, rather than the χ^2 quantile in equation (7), produces confidence intervals with

nominal coverage for the MM - $\boldsymbol{\theta}$ approach, although coverages for the MM - $\hat{\boldsymbol{\theta}}$ method drop to as low as 27.3%.

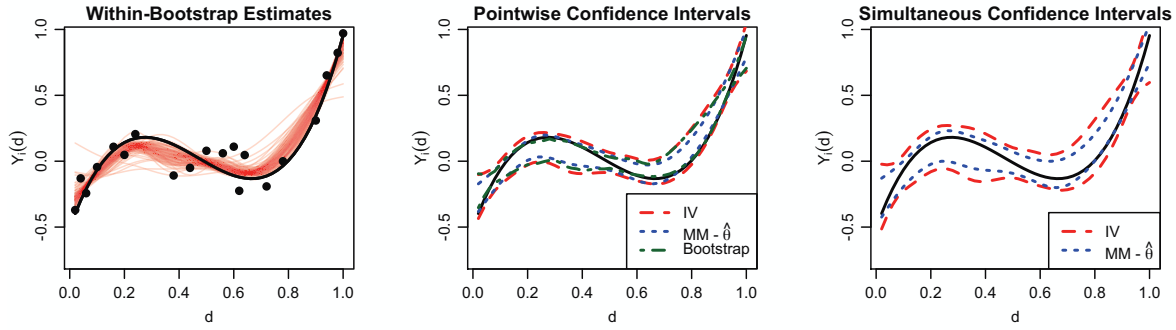

Figure 2: Illustration of the proposed inference method using a single function. The left panel shows the true curve in black, the observed points as black dots, and the within bootstrap sample estimated curves in red. The middle panel shows pointwise confidence intervals based on the iterated variance ("IV") approach, the mixed model method ("MM - $\hat{\boldsymbol{\theta}}$") and the quantiles of the bootstrapped estimates ("Bootstrap"). In the right panel are simultaneous confidence intervals from the IV and MM - $\hat{\boldsymbol{\theta}}$ approaches.

Figure 2 illustrates the results in Tables 2 and 3 by examining a single curve in a dataset with $I = 20$, $J_i = 20$, and $\sigma^2 = .01$. Shown in the left panel of this Figure is the true curve $Y_i(d)$, the observed points and the estimated curves within each bootstrap sample. This panel illustrates the amount of variability in the estimated curve based on uncertainty in the FPC decomposition objects $\boldsymbol{\theta}$. Displayed in the middle panel for the same curve are the 95% pointwise confidence intervals based on the IV approach, the MM - $\hat{\boldsymbol{\theta}}$ method, and the quantiles of the bootstrapped curve estimates. The bootstrap approach does not consider model-based variance in the estimated scores, while the MM - $\hat{\boldsymbol{\theta}}$ approach ignores variability in $\hat{\boldsymbol{\theta}}$. In the right panel are the simultaneous confidence intervals given by the IV and MM - $\hat{\boldsymbol{\theta}}$ approaches. Pointwise and simultaneous confidence intervals based on the model-only approach are narrower than those based on the iterated variance approach, and do not achieve nominal coverage level.

The improvements in average mean squared errors and coverage probabilities for the IV approach over the MM - $\hat{\boldsymbol{\theta}}$ method incur computational costs. Estimation of decomposition objects requires bivariate smoothing of a covariance matrix, which can be computationally expensive and must be repeated for each bootstrap sample. The time needed to complete this step will increase with the number of points in $\boldsymbol{d}_g$ and with the number of knots used in the bivariate smoothing. However, in the scenarios we consider estimating the decomposition objects takes between 4 and 6 seconds for each iteration, making the

proposed IV approach computationally feasible.

5 Applications

In this section we consider two statistically and scientifically distinct applications. First, we treat longitudinal observations of CD4 cell counts as sparsely observed functional data; second, we consider densely observed white-matter tract profiles. Both datasets and the code implementing our analyses are available online.

5.1 CD4 Data

Human immune deficiency virus (HIV) attacks an immune cell called the CD4 cell and thereby decreases an affected individual's response to infectious agents. Because of this, CD4 cell count per milliliter of blood is a useful surrogate measure for the progression of HIV. In this dataset from the Multicenter AIDS Cohort Study (MACS), we observe CD4 cell counts for 366 infected individuals in a traditional longitudinal study (Kaslow et al., 1987). That is, subjects were observed at roughly semi-annual visits with varying numbers of visits per subjects: there are a total of 1888 data points, with between 1 and 11 observations per subject and a median of 5. CD4 counts are plotted against months since seroconversion (the time at which HIV becomes detectable) in the top-left panel of Figure 3, with a randomly selected subset of patients highlighted in red. Longitudinal data analysis approaches to this study are presented in Diggle et al. (1994), and the treatment of a related dataset as sparsely observed functional data is given in Yao et al. (2005).

In this application, we wish to estimate the continuous trajectory of CD4 cell counts over time based on a limited number of observations for each subject. FPC analysis offers a powerful tool for estimating population-level trends by using the dense collection of observations across subjects and predicting subject-specific functions using a mixed model framework. The predicted functions are subject to model-based variability as well as uncertainty in the FPC decomposition. This second source of variability can be particularly important when functions are sparsely observed at the subject level. Here we analyze the CD4 cell count data using both the IV method developed in Section 3 and the MM - $\hat{\theta}$ method. Both approaches are implemented on the full dataset as well as the randomly selected subset of 25 individuals

to demonstrate the effect of sample size in the analysis of sparse functional data.

Results of the CD4 cell count application are shown in Figure 3. The top-left panel plots the CD4 cell

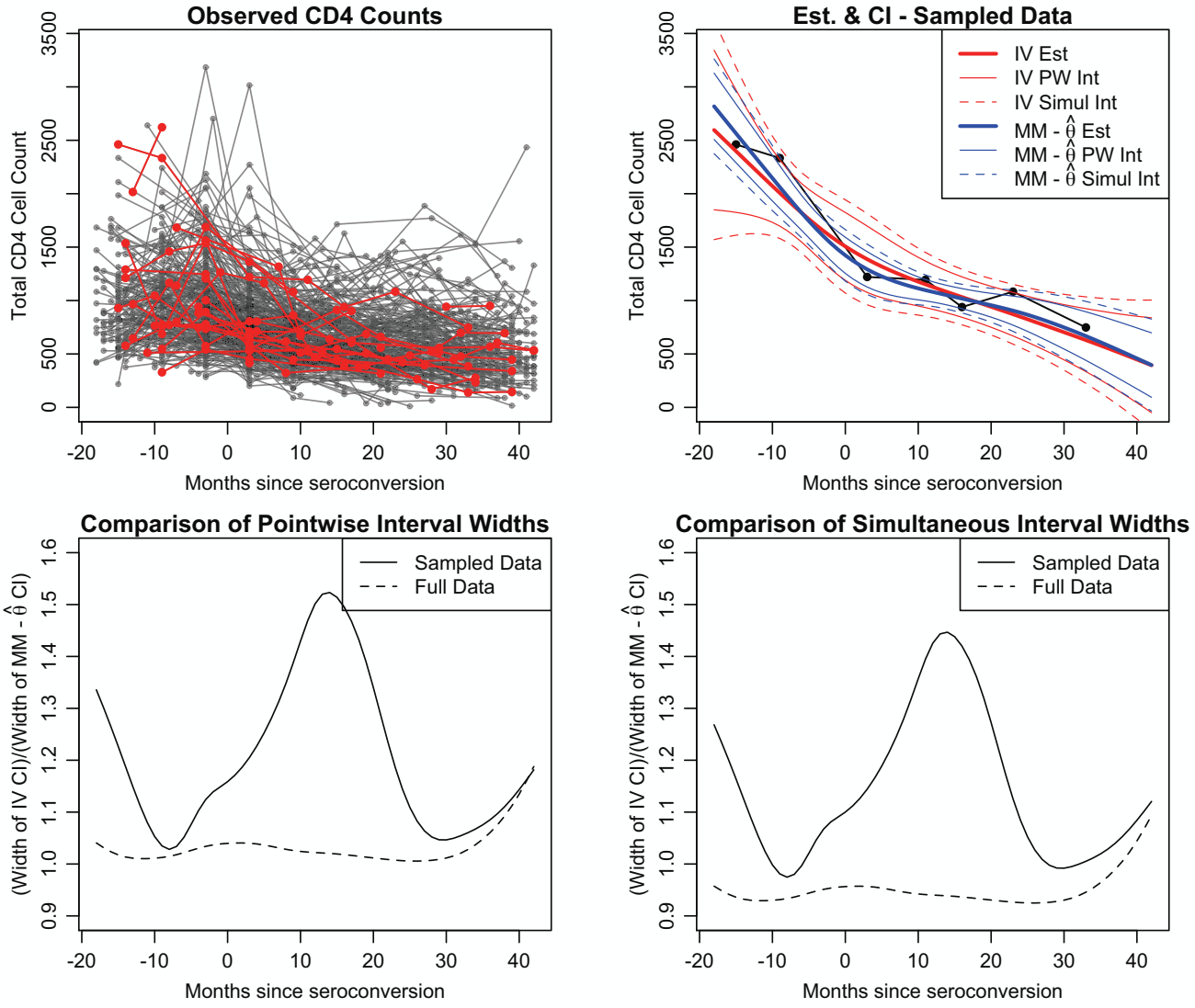

Figure 3: Results from analysis of CD4 cell count data. The top-left panel shows the observed CD4 counts; a random sample of 25 subjects is highlighted in red. The top-right panel shows the estimated curves ("Est") for a single CD4 trajectory based on the sampled subjects and full dataset using the iterated approach ("IV") and the mixed model approach ("MM - $\hat{\theta}$"); also shown are pointwise ("PW Int") and simultaneous ("Simul Int") confidence intervals. The bottom-left and bottom-right panels compare the width of pointwise and simultaneous confidence intervals, respectively, for the IV and MM - $\hat{\theta}$ approaches averaged over all subjects; comparisons for the full and sampled data are given. Simultaneous intervals are constructed using the resampling procedure in Section 3 for the IV method and using the interval in (7) for the MM - $\hat{\theta}$ method

counts against months since seroconversion, highlighting the sample used in the subset analysis. Curve estimates and intervals for a single subject in the subset analysis are provided in the top-right panel of Figure 3. Simultaneous intervals are constructed using resampling procedure for the IV approach and

using the asymptotic interval given in (7) for the MM - $\hat{\theta}$ method. For this subject, the estimated CD4 cell count trajectory using the IV method is lower for months -18 to -10 and higher for months 0 to 12 than the trajectory estimated using the mixed model approach. In these times, the pointwise and simultaneous confidence intervals constructed using the IV approach are much wider than those constructed using the MM - $\hat{\theta}$ framework. For example, using the method proposed in Section 3, at month -18 this subject's CD4 cell count is estimated to be approximately 2600 CD4 cells per milliliter of blood with upper and lower bounds of 3300 and 1900; using the mixed model approach, the CD4 cell count is estimated to be 2800 with bounds 3100 and 2500. Further, using the upper bound of the simultaneous interval to estimate the month in which CD4 cell count drops below 1500 per milliliter would yield month 10 under the IV method and month 4 under the MM - $\hat{\theta}$ approach. Thus, inference for a subject's CD4 trajectory can change dramatically depending on the procedure used to construct mean and variance estimates.

In the bottom-left panel, we compare the widths of pointwise confidence intervals yielded by the IV and MM - $\hat{\theta}$ approaches. For the full data analysis, intervals resulting from the proposed method have a similar width as those from the mixed model approach, but are up to 1.18 times as wide for months (36, 42) where there are fewer observations across subjects. In the analysis of the randomly selected 25 subjects, pointwise confidence intervals resulting from the IV approach are up to 1.52 times wider than those from the MM - $\hat{\theta}$ method, and are on average 1.21 times wider. Similarly, a comparison of the average widths of simultaneous confidence intervals given by the two methods is shown in the bottom-right panel of Figure 3. The relationship between the widths of simultaneous intervals is similar to that for pointwise intervals for the sampled data. For the full data analysis, the intervals given by the IV approach are in fact slightly narrower than those for the MM - $\hat{\theta}$ method; recall that simulations indicated that the simultaneous intervals in (7) may in fact be conservative asymptotically, while the IV simultaneous intervals could attain coverages slightly under the nominal level. Generally, these results indicate that ignoring the variability in the FPC decomposition can substantially understate the total variability in estimated curves for small samples and sparse data.

5.2 White-matter Tract Data

Next, we consider a neuroimaging study of white matter tracts in multiple sclerosis (MS) patients and healthy controls. Intracranial white-matter consists of tracts or bundles of myelinated axons that carry electrical signals between areas of the brain. Major examples of white matter tracts are the corticospinal tracts, which connect the motor cortex to the spinal cord, and the corpus callosum, which bridges the left and right hemispheres. MS is an autoimmune disorder that results in white matter lesions, axonal damage and demyelination, and can lead to severe patient disability. Diffusion tensor imaging (DTI) is an MRI modality able to resolve individual tracts by tracing the diffusion of water, which tends to be anisotropic within white matter tracts and isotropic elsewhere in the brain (Basser et al., 1994, 2000; LeBihan et al., 2001; Mori and Barker, 1999). Fractional anisotropy (FA) is one of several quantities derived from DTI scans that measures overall anisotropy, and may be decreased in MS patients.

Our dataset consists of 42 healthy controls and 100 MS patients, for whom we observe FA values sampled densely at 55 locations along the right corticospinal tract. The top-left panel of Figure 4 displays these FA profiles, separating MS patients and healthy controls. There is significant variability in tract profiles comparing subjects, as well as apparent measurement error in the observed FA values. In addition, 35% of subjects are missing observations for at least one tract location, and 17% of subjects are missing observations for 5 or more locations. For all subjects, any missing values are consecutive and begin at tract location 0, so that missingness is clustered at the beginning of the tract.

Because tract profiles are affected by disease progression, they are useful biomarkers and have been related to patient disability (Goldsmith et al., 2011). The goal of this analysis is to understand the total variability in tract profile estimates using FPCA. We compare the proposed IV method to the standard MM - $\hat{\boldsymbol{\theta}}$ approach to assess the contribution of decomposition-based uncertainty on overall variability in estimated curves. Additionally, to examine the effect of sample size when curves are densely sampled, we conduct our analyses on both the full data set and on the subset of healthy controls.

Figure 4 shows the results from the analysis of the DTI tract data. A comparison of the relative widths of pointwise confidence intervals constructed using the IV and MM - $\hat{\boldsymbol{\theta}}$ methods is given in the top-right panel of Figure 4. In the full data analysis, the IV method results in confidence intervals that are on average 1.12 times wider than those given by the MM - $\hat{\boldsymbol{\theta}}$ approach. When considering the 42 healthy

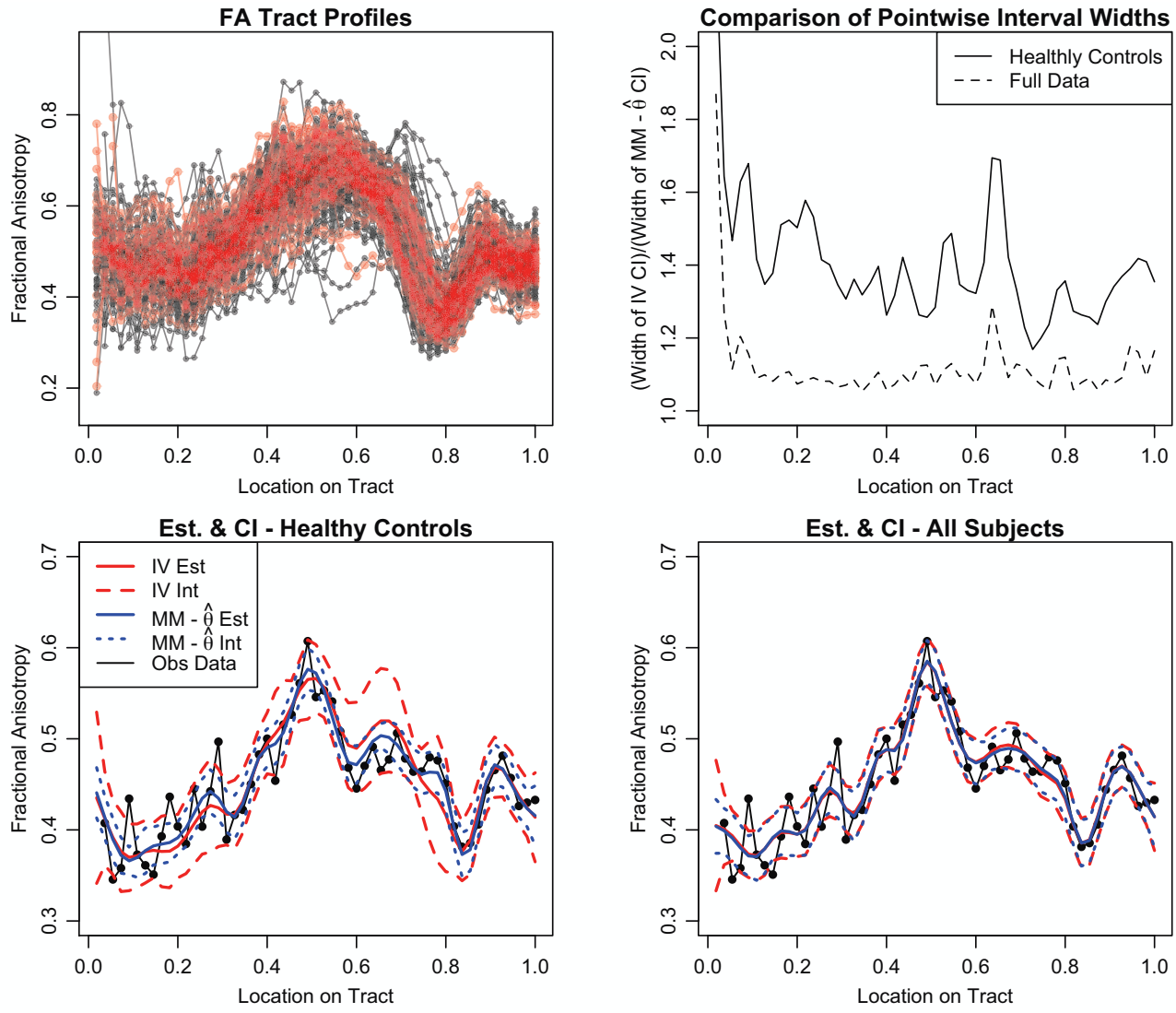

Figure 4: Results from analysis of DTI tract data. The top-left panel shows the observed profiles, with MS patients in black and healthy controls in red. The top-right panel compares the width of confidence intervals for the MM - $\hat{\boldsymbol{\theta}}$ and IV approaches averaged over all subjects; analyses of the full data and healthy controls only are provided. The bottom-left and -right panels show the estimated curves and pointwise confidence intervals for a single function based on the healthy controls and full dataset, respectively; estimates are constructed using both the iterated approach ("IV") and mixed model ("MM - $\hat{\boldsymbol{\theta}}$") approaches.

controls only, the IV approach produces intervals that are on average 1.40 times wider than those for the MM - $\hat{\boldsymbol{\theta}}$ approach. The comparison of the intervals yielded by these methods illustrates that, even for densely sampled observations, the mixed-model approach that conditions on a single decomposition can significantly understate the total variability by neglecting the uncertainty in the FPC decomposition objects.

A demonstration of the difference between the IV and MM - $\hat{\boldsymbol{\theta}}$ approaches is given in the bottom

panels of Figure 4 using the tract profile of a single healthy control. On the left are the estimated curves and pointwise confidence intervals from the analysis based on the 42 healthy controls only; the interval based on the IV method is substantially wider at several locations on the tract, particularly near 0.0, 0.6, and 0.8. In addition, the curve estimated using iterated expectations falls outside the pointwise confidence interval given by the MM - $\hat{\theta}$ approach at several locations between 0.5 and 0.8. On the right are the estimates and pointwise intervals from the full data analysis, showing that the two approaches provide more similar intervals when a larger sample size is used. Thus for small sample sizes, accounting for the uncertainty in the FPC decomposition by using the IV method results in noticeably wider confidence intervals; as the sample size increases, the decomposition uncertainty decreases and the IV and MM - $\hat{\theta}$ methods yield more similar results.

6 Discussion

Functional principal components analysis is ubiquitous in functional data analysis. FPC decompositions are estimated from the data, and curves are expressed using a truncated Karhunen-Loève expansion with curve-specific scores predicted from a mixed model that conditions (often implicitly) on the estimated decomposition. Previous work on estimation and inference for FPC expansions has focused on the model-based prediction of scores and the associated variability without including the uncertainty in the FPC decompositions. In many cases, including small samples or sparsely observed curves, this second source of uncertainty is non-negligible.

In this paper, we propose a method for estimating curves and understanding the variability of those estimates that accounts for both model-based and decomposition-based uncertainty. Specifically, we use iterated expectations to average model-based estimates that condition on specific decompositions across the distribution of those decompositions; similarly, the iterated variance formula provides correct inference by combining both sources of uncertainty. A bootstrap procedure provides an empirical distribution of the FPC decompositions. Simulations demonstrate the accuracy of our method, both in terms of the MSEs of estimated curves and in terms of the coverage of associated confidence intervals. We apply our method to two statistically and scientifically distinct data sets, one using longitudinally observed CD4 cell counts as sparse functional data and the other using partially-unobserved tract profiles from a neu-

roimaging study. An `R` implementation of our method is given by the `iv.fpca()` function in the `refund` package on `CRAN` (Crainiceanu et al., 2011a).

Future work may proceed in several directions. The proposed method assesses uncertainty in the FPC decomposition through the bootstrap; alternatively, one could use a probabilistic model for the unknown objects in the decomposition. Such an approach could be integrated seamlessly with the methods developed here, but the specification of an appropriate model for the unknown objects is a difficult challenge. Inference for more complex FPC-based methods may be improved through the extension of iterated expectation and variance formulas. For instance, longitudinal functional principal components analysis (LFPCA) decomposes repeated observations of curves into baseline subject-specific components, time-varying subject-specific components, and subject-visit-specific components (Greven et al., 2010); including model-based and decomposition-based variability will improve the understanding of uncertainty in this scientifically useful setting. Additionally, the proposed method has focused on uncertainty in the Karhunen-Loève expansion arising from both model-based and decomposition-based variability, but it does not address the model-based uncertainty in the estimation of the mean function. Correctly including this source of variability may increase the accuracy of the estimated curves and confidence intervals, particularly in cases where the estimate of the mean function may be poor.

7 Acknowledgments

The work of Goldsmith and Crainiceanu was supported by Award Number R01NS060910 from the National Institute Of Neurological Disorders And Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health. The work of Greven was funded by Emmy Noether grant GR 3793/1-1 from the German Research Foundation.

References

BASSER, P., MATTIELLO, J. and LEBIHAN, D. (1994). MR diffusion tensor spectroscopy and imaging. *Biophysical Journal*, **66** 259–267.

BASSER, P., PAJEVIC, S., PIERPAOLI, C. and DUDA, J. (2000). In vivo fiber tractography using DT-MRI data. *Magnetic Resonance in Medicine*, **44** 625–632.

CRAINICEANU, C., REISS, R., GOLDSMITH, J. and HUANG, L. (2011a). *refund: Regression with Functional Data*. R package version 0.1-5, URL http://CRAN.R-project.org/package=refund.

CRAINICEANU, C. M. and GOLDSMITH, J. (2010). Bayesian functional data analysis using WinBUGS. *Journal of Statistical Software*, **32** 1–33.

CRAINICEANU, C. M., STAICU, A.-M., RAY, S. and PUNJABI, N. (2011b). Statistical inference on the difference in the means of two correlated functional processes: an application to sleep EEG power spectra. *Technical Report*.

DI, C.-Z., CRAINICEANU, C. M., CAFFO, B. S. and PUNJABI, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics*, **4** 458–488.

DIGGLE, P. J., HEAGERTY, P., LIANG, K.-Y. and ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2011). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage*, **57** 431–439.

GOLDSMITH, J., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (Accepted). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C*.

GREVEN, S., CRAINICEANU, C. M., CAFFO, B. and REICH, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, **4** 1022–1054.

HALL, P. and HOSSEINI-NASAB, B. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B*, **68** 109–126.

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

JOLLIFFE, I. T. (2002). *Principal Component Analysis*. New York: Springer.

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.

LEBIHAN, D., MANGIN, J., POUPON, C. and CLARK, C. (2001). Diffusion tensor imaging: Concepts and applications. *Journal of Magnetic Resonance Imaging*, **13** 534–546.

MORI, S. and BARKER, P. (1999). Diffusion magnetic resonance imaging: its principle and applications. *The Anatomical Record*, **257** 102–109.

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. New York: Springer.

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

RICE, J. A. and WU, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57** 253–259.

RUPPERT, D., WAND, M. P. and CARROLL, R. J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

SHI, M., WEISS, R. and TAYLOR, J. (1996). An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *Journal of the Royal Statistical Society: Series C*, **45** 151–163.

STANISWALIS, J. and LEE, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, **444** 1403–1418.

YAO, F., MÜLLER, H. and WANG, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100** 577–590.

YAO, F., MÜLLER, H. G., CLIFFORD, A., DUEKER, S., FOLLETT, J., LIN, Y., BUCHHOLZ, B. and VOGEL, J. (2003). Shrinkage estimation for functional principal component scores with application to the population. *Biometrics*, **59** 676–685.

Online Appendix to "Corrected Confidence Bands for Functional Data Using Principal Components"

Jeff Goldsmith, Sonja Greven and Ciprian Crainiceanu

November 8, 2011

In this supplement, we provide simulation results in addition to those appearing in Section 4 in the main paper. In the following, we compare results given by the iterated variance method (IV), the mixed model approach that estimates the FPC decomposition objects $\boldsymbol{\theta}$ from the observed data (MM - $\hat{\boldsymbol{\theta}}$), and the mixed model approach that fixes the $\boldsymbol{\theta}$ at the truth (MM - $\boldsymbol{\theta}$).

A Additional Simulation Results

Table A.1 provides the average mean squared error AMSE $= \frac{1}{I}\sum_{i=1}^{I}\left[\frac{1}{50}\sum_{d\in \boldsymbol{d}_g}\left(\hat{Y}_i(d) - Y_i(d)\right)^2\right]$ of the curves estimated from IV, MM - $\hat{\boldsymbol{\theta}}$, and MM - $\boldsymbol{\theta}$ approaches using normally distributed scores (results for non-normally distributed scores appear in Table 1). When curves are densely observed, the IV and MM

	IV			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	0.026	0.266	0.524	0.026	0.277	0.574	0.020	0.064	0.134
$\sigma^2 = .01$	0.100	0.479	0.818	0.099	0.487	0.857	0.079	0.238	0.412
$I = 50; \sigma^2 = .0025$	0.023	0.150	0.307	0.023	0.169	0.347	0.020	0.064	0.143
$\sigma^2 = .01$	0.088	0.347	0.609	0.088	0.363	0.637	0.079	0.240	0.424
$I = 100; \sigma^2 = .0025$	0.022	0.114	0.238	0.022	0.126	0.270	0.020	0.064	0.141
$\sigma^2 = .01$	0.084	0.300	0.522	0.084	0.311	0.541	0.079	0.239	0.419

Table A.1: $100 \times$ AMSE, taken over all grid points and subjects for simulations with normally distributed scores, of estimated curves constructed using the proposed iterated expectation method ("IV"), the mixed model approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach with $\boldsymbol{\theta}$ fixed the true value ("MM - $\boldsymbol{\theta}$"). Curves are sampled with $J_i = 50$, $J_i = 20$, and using an unbalanced design ("Unbal").

- $\hat{\boldsymbol{\theta}}$ have similar results, but for sparsely observed curves the IV approach has lower AMSE indicating on average less bias in estimated curves. As expected, AMSEs increase as measurement error increases, and decrease as the number of observed curves increases. The results for the MM - $\boldsymbol{\theta}$ provide a reference for the performance of estimated curves.

Table A.2 displays the average coverage, taken across all grid points and all subjects, of pointwise confidence intervals for simulation scenarios with normally distributed scores (Table 2 provides coverages of pointwise intervals for simulations with non-normally distributed scores). The proposed IV method is

	IV			Bootstrap			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	95.3	97.2	97.0	63.8	86.0	83.5	90.3	69.3	72.7	94.9	95.1	94.9
$\sigma^2 = .01$	95.2	94.7	93.9	61.3	75.6	73.8	90.8	81.7	78.6	94.9	95.1	94.9
$I = 50; \sigma^2 = .0025$	94.8	95.3	96.8	45.9	77.0	75.9	93.0	75.0	77.7	95.0	95.0	95.0
$\sigma^2 = .01$	94.8	94.3	94.0	43.5	61.7	61.8	93.2	88.3	86.7	95.0	95.0	95.1
$I = 100; \sigma^2 = .0025$	94.9	93.7	95.3	34.9	70.1	68.8	94.0	78.7	79.2	95.0	95.1	95.0
$\sigma^2 = .01$	94.9	94.7	94.2	33.2	52.5	52.2	94.1	91.2	90.2	95.0	95.1	95.0

Table A.2: Average pointwise coverage for simulations with normally distributed scores, taken over all grid points and subjects, of 95% confidence intervals constructed using the proposed iterated variance method ("IV"), the approach that uses the quantiles of bootstrapped curve estimates ("Bootstrap"), the mixed model approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach with $\boldsymbol{\theta}$ fixed at its true value ("MM - $\boldsymbol{\theta}$"). Coverages are expressed as percents.

near the nominal coverage level for all simulation scenarios, with coverages between 93.7% and 97.2%. For densely observed curves with large sample sizes, the MM - $\hat{\boldsymbol{\theta}}$ method approaches nominal coverage, but for smaller sample sizes and especially sparse curves, coverages are much lower and can be as low as 69.3%. The MM - $\boldsymbol{\theta}$ method achieves nominal coverage in all scenarios, indicating that the lower coverages for MM - $\hat{\boldsymbol{\theta}}$ stem from failing to account for uncertainty in $\boldsymbol{\theta}$ rather than from model-based uncertainty.

Table A.3 provides the average coverage of simultaneous confidence intervals taken over all grid points and subjects constructed using the IV approach described in Section 3 and the MM - $\hat{\boldsymbol{\theta}}$ and MM - $\boldsymbol{\theta}$ methods described in Section 2 (results for simulations with non-normally distributed scores are given in Table 3). Coverages for the IV method approach the nominal level, with coverages between 89.6% and

	IV			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	94.5	96.4	95.2	86.1	35.8	41.4	98.4	98.3	98.3
$\sigma^2 = .01$	94.6	92.3	89.4	88.3	61.0	54.1	98.3	98.3	98.3
$I = 50; \sigma^2 = .0025$	93.7	93.3	96.1	93.8	48.7	56.1	98.2	98.3	98.3
$\sigma^2 = .01$	94.1	92.4	91.1	94.6	81.1	78.2	98.2	98.3	98.3
$I = 100; \sigma^2 = .0025$	94.0	89.6	93.4	96.3	57.1	60.8	98.3	98.2	98.2
$\sigma^2 = .01$	94.4	93.3	92.1	96.6	89.9	87.6	98.3	98.2	98.2

Table A.3: Average coverage of 95% simultaneous confidence intervals for simulations with normally distributed scores constructed using the proposed iterated variance method ("IV"), the mixed model approach estimating $\boldsymbol{\theta}$ from the data ("MM - $\hat{\boldsymbol{\theta}}$") and the mixed model approach with $\boldsymbol{\theta}$ fixed at its true value ("MM - $\boldsymbol{\theta}$"). Coverages are expressed as percents.

96.4%. For the MM - $\hat{\boldsymbol{\theta}}$, coverage for densely observed curves roughly achieves the nominal level, but for sparsely observed curves coverages are as low as 35.8%. Results for the MM - $\boldsymbol{\theta}$ approach indicate that the intervals are in fact conservative, with coverages of 98.2 and 98.3%.

The approach used to construct simultaneous intervals for the IV method can be adapted to the MM - $\hat{\boldsymbol{\theta}}$ and MM - $\boldsymbol{\theta}$ methods in the following way. A $100(1-\alpha)\%$ simultaneous confidence interval for $Y_i(\boldsymbol{d}_g)$ is given by

$$\hat{Y}_i(\boldsymbol{d}_g) \pm m_{(1-\alpha)} \sqrt{\text{diag}\left\{\text{Var}\left[\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right]\right\}}, \tag{A.1}$$

where $\text{Var}\left[\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right]$ is the variance conditional on the estimated decomposition $\hat{\boldsymbol{\theta}}$ given in equation (5). The critical value $m_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of the random variable

$$\max_{d \in \boldsymbol{d}_g} \left| \frac{\hat{Y}_i(d) - Y_i(d)}{\sqrt{\text{diag}\left\{\text{Var}\left[\hat{Y}_i(d) - Y_i(d)|\hat{\boldsymbol{\theta}}\right]\right\}}} \right|. \tag{A.2}$$

To estimate $m_{(1-\alpha)}$, we note that $\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)$ is approximately multivariate normal centered at $\mathbf{0}$ and with the conditional covariance in (5):

$$\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g) \overset{\text{approx.}}{\sim} \text{N}\left[\mathbf{0}, \text{Var}\left(\hat{Y}_i(\boldsymbol{d}_g) - Y_i(\boldsymbol{d}_g)|\hat{\boldsymbol{\theta}}\right)\right] \tag{A.3}$$

We repeatedly draw functions with the estimated covariance from this distribution, and for each draw calculate the resulting value of (A.2); the $(1-\alpha)$ quantile of these values is used as an estimate of $m_{(1-\alpha)}$.

Using the resampling approach, we construct simultaneous confidence intervals for the MM - $\hat{\boldsymbol{\theta}}$ and MM - $\boldsymbol{\theta}$ methods; coverages for these intervals are given in Table A.4. For reference, included in this Table are the coverages for simultaneous intervals from the IV approach. First, we note that the coverages for the MM - $\boldsymbol{\theta}$ method achieve the nominal coverage level, indicating that the resampling procedure may provide intervals that are less conservative than the asymptotic approach given in equation (7). Coverages for the MM - $\hat{\boldsymbol{\theta}}$ are lower for the resampling technique than for the asymptotic approach described in Table A.4, with coverages as low as 27.3% and no higher than 91.3%. As noted previously, coverages for the IV

	IV			MM - $\hat{\boldsymbol{\theta}}$			MM - $\boldsymbol{\theta}$		
	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal	$J_i = 50$	$= 20$	Unbal
$I = 20; \sigma^2 = .0025$	94.5	96.4	95.2	74.3	27.3	33.7	94.8	95.1	95.1
$\sigma^2 = .01$	94.6	92.3	89.4	77.6	48.1	43.4	94.9	95.0	94.8
$I = 50; \sigma^2 = .0025$	93.7	93.3	96.1	86.1	37.2	46.4	94.8	95.1	95.2
$\sigma^2 = .01$	94.1	92.4	91.1	87.5	68.4	66.1	94.8	95.1	95.1
$I = 100; \sigma^2 = .0025$	94.0	89.6	93.4	90.4	45.1	50.0	95.0	94.9	95.1
$\sigma^2 = .01$	94.4	93.3	92.1	91.3	79.6	76.8	94.9	95.0	95.1

Table A.4: Average coverage of 95% simultaneous confidence intervals for simulations with non-normally distributed scores constructed. All intervals are constructed using the resampling procedure to determine the critical value $m_{(1-\alpha)}$.

method approximate the nominal level, ranging between 89.6% and 96.4%.