Johns Hopkins University, Dept. of Biostatistics Working Papers

9-23-2008

GENERALIZED MULTILEVEL FUNCTIONAL REGRESSION

Ciprian M. Crainiceanu
Bloomberg School of Public Health, Department of Biostatistics, Johns Hopkins, ccrainic@jhsph.edu

Ana-Maria Staicu
Department of Mathematics, University of Bristol

Chongzhi Di
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

Generalized Multilevel Functional Regression

Ciprian M. Crainiceanu Ana-Maria Staicu Chongzhi Di

Abstract

We introduce Generalized Multilevel Functional Linear Models (GMFLM), a novel statistical framework motivated by and applied to the Sleep Heart Health Study (SHHS), the largest community cohort study of sleep. The primary goal of SHHS is to study the association between sleep disrupted breathing (SDB) and adverse health effects. An exposure of primary interest is the sleep electroencephalogram (EEG), which was observed for thousands of individuals at two visits, roughly 5 years apart. This unique study design led to the development of models where the outcome, e.g. hypertension, is in an exponential family and the exposure, e.g. sleep EEG, is multilevel functional data. We show that GMFLMs are, in fact, generalized multilevel mixed effect models. Two consequences of this result are that: 1) the mixed effects inferential machinery can be used for GMFLM and 2) functional regression models can be extended naturally to include, for example, additional covariates, random effects and nonparametric components. We propose and compare two inferential methods based on the parsimonious decomposition of the functional space.

Some key words: Functional principal components, Smoothing, Sleep EEG.

1 Introduction

The methodology described in this paper was motivated by our ongoing studies of the association of sleep and adverse health outcomes. For example, in the Sleep Heart Health Study (SHHS) we are interested in studying models where the health outcomes, such as Chronic Heart Disease (CHD) or Hypertension (HTN), are regressed on sleep electroencephalogram (EEG) data and other covariates. Because sleep-EEG data is recorded at two visits the exposure has a natural multilevel functional structure. SHHS contains the largest collection of sleep EEG data on an epidemiologic cohort, with more than 6000 subjects at the baseline visit and more than 4000 subjects at visit 2. This is just one example of modern research data that have become increasingly complex, raising non-traditional modeling and inferential challenges. In particular, advancements in technology and computation have made recording and processing of functional data possible. In this context, it has become increasingly necessary to develop models that describe the association between a functional measurement, such as a magnetic resonance image (MRI) or EEG, and outcomes, such as adverse health effects. Because functional data is now routinely collected at multiple visits these models have to be extended to incorporate the natural multilevel structure of the functional data.

An appealing statistical methodology for this type of problems is Functional Regression Analysis, which allows the outcomes or the regressors or both to be functions instead of scalars. Functional Regression Analysis is currently under intense methodological research [3, 9, 17, 23, 24, 27, 33] and is a particular case of Functional Data Analysis (FDA) [16, 14, 31, 32, 30]. Two comprehensive monographs that provide a broad overview of FDA with applications to curve and image analysis are [26, 27]. The fundamental notion of FDA methods is to decompose the space of curves into principal directions of variation. The main method for achieving this employs Principal Component Analysis (PCA) of the raw data or smoothed curves. PCA provides a simple recipe for dimensionality reduction by estimating the eigenvectors of the functional covariance operator. Furthermore, PCA estimates the subject-specific features as the coordinates of subject curves in the basis spanned by the functional principal components. There has been considerable recent effort to apply FDA to longitudinal data, e.g., [8, 28, 32, 36]. See [22] for a thorough review. Because longitudinal data are often multilevel, it might be assumed that this work on longitudinal FDA is multilevel. However, this work has assumed that one or more functions are observed only over a single time course, e.g., height is observed over childhood in growth studies. Thus, in all current FDA research, the term "longitudinal" represents single-level time series. FDA was extended to multilevel functional data [7], such as when subject-level curves are observed at several visits.

In this paper we present several novel methodological developments in the general area of functional regression based on functional PCA. First, we introduce the multilevel functional exposure to incorporate cases when functional data is observed at multiple time points. Second, we show that all regression models with functional predictors can be viewed as mixed effects models with two mixed effects sub-models: an outcome and an exposure model. This has important methodological and computational implications because the mixed effects inferential machinery can be used and models can be generalized within a well researched statistical framework. Third, we introduce a Bayesian inferential framework for the joint analysis of the outcome and exposure mixed effects models to account for the multi-layered variability and measurement errors. This method is contrasted with a simpler two-stage method that uses the predicted values of the random effects from the exposure model in the outcome model. Fourth, we present theoretical and simulation results that provide insight into when using a two-stage method is a reasonable alternative to the joint analysis and when it is expected to fail. This has important practical implications for researchers who would like to decide what method to use in a particular application. Fifth, we obtain the best linear unbiased predictors and their associated variability for the random effects in the functional exposure model. These theoretical results provide an appealing and computationally tractable platform for two-stage analyses. Sixth, we show how the mixed effects framework allows straightforward generalizations of functional regression models to incorporate covariates, random effects, smooth functions of other covariates, etc. Our methods are an evolutionary development in a growth area of research that build on and borrow strength from multiple methodological frameworks. Given the range of applications and methodological flexibility of our methods, we anticipate that they will become one of the standard tools of research in the area of functional regression.

The paper is organized as follows. Section 2 introduces our methodology for single-level functional regression. Section 3 discusses the specific challenges of a Bayesian analysis of the joint mixed effects model corresponding to functional regression. Section 4 generalizes the methods to account for multilevel functional data exposure. Section 5 provides extensions of functional regression models. Section 6 provides simulations. Section 7 describes an application to sleep EEG data from the SHHS. Section 8 summarizes our conclusions.

2 Single-level functional regression models

2.1 Joint mixed effects models

A particularly useful class of models that describe associations between non-gaussian outcomes and functional data is the class of generalized functional linear models (GFLM) [23]. The observed data for the ith subject in a GFLM is $[Y_i, \boldsymbol{Z}_i, \{W_i(t_{im}), t_{im} \in [0,1]\}]$, where Y_i is the continuous or discrete outcome, $\boldsymbol{Z}_i$ is a vector of covariates, and $W_i(t_{im})$ is a random curve in $L_2[0,1]$ observed at time t_{im}, which is the mth observation, $j = 1, \ldots, M_i$, for the ith subject, $i = 1, \ldots, n$. We assume that $W_i(t)$ is a proxy observation of the true underlying functional signal $X_i(t)$ and that $W_i(t) = \mu(t) + X_i(t) + \epsilon_i(t)$, where $\mu(t)$ is the population average and $\epsilon_i(t)$ is a mean zero white noise process with variance σ_ϵ^2. We also assume that the distribution of Y_i is in the exponential family with linear predictor η_i and dispersion parameter α, denoted here by $\mathrm{EF}(\eta_i, \alpha)$. The linear predictor is assumed to have the following form

$$\eta_i = \int_0^1 X_i(t)\beta(t)dt + \boldsymbol{Z}_i^t\boldsymbol{\gamma}, \tag{1}$$

where $\beta(\cdot) \in L_2[0,1]$ is a functional parameter and the main target of inference. Note that if $\{\psi_k(\cdot), k \geq 1\}$ is an orthonormal basis in $L_2[0,1]$ then both $X_i(\cdot)$ and $\beta(\cdot)$ have unique representations $X_i(t) = \sum_{k\geq 1}\xi_{ik}\psi_k(t)$, $\beta(t) = \sum_{k\geq 1}\beta_k\psi_k(t)$ and equation (1) can be rewritten as

$$\eta_i = \sum_{k\geq 1}\xi_{ik}\beta_k + \boldsymbol{Z}_i^t\boldsymbol{\gamma}. \tag{2}$$

In model (1) the functional parameter $\beta(t)$ does not depend on a basis, whereas the coefficients β_k are specific to a particular choice of orthonormal basis in $L_2[0,1]$. The coordinate version (2) of model (1) is intuitive because it provides a recipe for regressing an outcome, Y_i, on a function, $X_i(t)$, by regressing it on the coordinates, ξ_{ik}, of that function in an orthonormal basis, $\psi_k(\cdot)$. However, this form of the model is impractical because it involves an infinite number of regressors. Instead we will use the following truncated version $\eta_i^K = \sum_{k=1}^K \xi_{ik}\beta_k + \boldsymbol{Z}_i^t\boldsymbol{\gamma}$, where K is the truncation lag. Once $\psi_k(\cdot)$ and K are fixed, the functional regression model becomes a generalized linear model (GLM)

$$\left\{\begin{array}{lcl} Y_i & \sim & \mathrm{EF}(\eta_i^K, \alpha); \\ \eta_i^K & = & \sum_{k=1}^K \xi_{ik}\beta_k + \boldsymbol{Z}_i^t\boldsymbol{\gamma}. \end{array}\right. \tag{3}$$

For reference, we will call equations (3) the outcome model. Note that the outcome model (3) is not an ordinary GLM because the scores, ξ_{ik}, $k = 1, \ldots, K$, are indirectly observed through the random curves $X_i(t)$, $i = 1, \ldots, n$, which, in turn, are indirectly observed through the proxy functions $W_i(t)$.

From a theoretical perspective, the choice of the orthonormal basis, $\psi_k(\cdot)$, is not important. There are an infinite number of bases in $L_2[0,1]$, but some, including the

Fourier, wavelet and Hermite polynomials, are more popular. Each basis tends to work better in particular applications. For example, the Fourier basis works better when observed data are mixtures of sinusoidal signals, while polynomial bases work better when underlying signals are smooth. It is our practical experience that some bases are good for many applications and no basis is best for all.

In this paper we use Functional Principal Component Analysis (FPCA) [27] to obtain a basis that captures most of the functional variability of the space spanned by $X_i(t)$ with its first few dimensions. FPCA is based on the covariance operator $K_X(t,s) = \text{Cov}\{X_i(t), X_i(s)\}$. Mercer's theorem (see [15], Chapter 4) provides the following convenient spectral decomposition $K_X(t,s) = \sum_{k=1}^{\infty} \lambda_k \psi_k(t)\psi_k(s)$, where $\lambda_1 \geq \lambda_2 \geq \dots$ are the ordered eigenvalues and $\psi_k(\cdot)$ are the associated orthonormal eigenfunctions of $K_X(\cdot,\cdot)$ in the L^2 norm. The Karhunen-Loève (KL) decomposition [18, 21] of the subject level functions is $X_i(t) = \sum_{k=1}^{\infty} \xi_{ik}\psi_k(t)$ where $\xi_{ik} = \int_0^1 X_i(t)\psi_k(t)dt$ are the principal component scores with $E(\xi_{ik}) = 0$, $\text{Var}(\xi_{ik}) = \lambda_k$ and $\text{Cov}(\xi_{ik}, \xi_{ik'}) = 0$ for every i and $k \neq k'$.

The covariance operator of the observed data, $W_i(t)$, is $K_W(t,s) = K_X(t,s) + \sigma_\epsilon^2 \delta_{t,s}$, where $\delta_{t,s} = 1$ if $t = s$ and 0 otherwise. These equations suggest a natural solution for estimating the eigenvalues, eigenfunctions and the nugget variance, σ_ϵ^2. The first step of the procedure is to estimate the mean function $\mu(t)$ using, for example, penalized spline smoothing [29] under the working independence assumption. For issues on smoothing for dependent data, see discussion in [20]. The second step is to obtain the method of moment estimates of $K_W(t,s)$, denoted by $\hat{K}_W(t,s)$. The third step is to estimate $\hat{K}_X(t,s)$ by smoothing $\hat{K}_W(t,s)$ for $t \neq s$, as suggested by [31, 34]. We propose to use penalized thin plate because bivariate local polynomial smoothing would be prohibitively slow for the size of our sleep data. The fourth step is to predict the diagonal elements, $\hat{K}_X(t,t)$, and estimate the error variance σ_ϵ^2 as

$\hat{\sigma_\epsilon}^2 = \int\{\hat{K}_W(t,t) - \hat{K}_X(t,t)\}dt$. The fifth step is to estimate the eigenvalues and eigenfunctions of $\hat{K}_X(t,s)$.

Once the eigenfunctions $\psi_k(\cdot)$ and a truncation lag K are fixed, the model for observed functional data can be written as a linear mixed model. Indeed, by construction, ξ_{ik} are mutually uncorrelated with mean 0 and variance λ_k. By assuming a normal shrinkage distribution for scores and errors, the model can be rewritten as

$$\begin{cases} W_i(t) & = & \sum_{k=1}^K \xi_{ik}\psi_k(t) + \epsilon_i(t); \\ \xi_{ik} & \sim & \mathrm{N}(0,\lambda_k); \;\; \epsilon_i(t) \sim \mathrm{N}(0,\sigma_\epsilon^2). \end{cases} \tag{4}$$

For reference, we will call equations (4) the exposure model. A close inspection of the model will reveal that this is a linear mixed model [19] with the random effects ξ_{ik} being the quantities that are used in the outcome model (3). We propose to jointly estimate the two outcome and exposure mixed effects models (3) and (4). In Section 2.2 we show that two-stage estimation, that is predicting the random effects in model (4) and plugging them in the model (3), may lead to misspecified variability when the outcome, Y_i, is normally distributed and biased estimators and misspecified variability when it is not normally distributed.

2.2 BLUP plug-in versus joint estimation

To better understand the potential problems associated with two-stage estimation we describe the induced likelihood for the observed data. We introduce the following notations $\boldsymbol{\xi}_i = (\xi_{i1},\ldots,\xi_{iK})^t$ and $\boldsymbol{W}_i = \{W_i(t_{i1}),\ldots,W_i(t_{iM_i})\}^t$, where M_i is the total number of functional observations for subject i. With a slight abuse of notation $[Y_i|\boldsymbol{W}_i,\boldsymbol{Z}_i] = \int[Y_i,\boldsymbol{\xi}_i|\boldsymbol{W}_i,\boldsymbol{Z}_i]d\boldsymbol{\xi}_i$, where $[\cdot|\cdot]$ denotes the probability density function of the conditional distribution. The assumptions in models (3) and (4) imply that

$[Y_i, \boldsymbol{\xi}_i | \boldsymbol{W}_i, \boldsymbol{Z}_i] = [Y_i|\boldsymbol{\xi}_i, \boldsymbol{Z}_i][\boldsymbol{\xi}_i|\boldsymbol{W}_i]$, which, in turn, implies that

$$[Y_i|\boldsymbol{W}_i, \boldsymbol{Z}_i] = \int [Y_i|\boldsymbol{\xi}_i, \boldsymbol{Z}_i][\boldsymbol{\xi}_i|\boldsymbol{W}_i] d\boldsymbol{\xi}_i. \tag{5}$$

Under normality assumptions it is easy to prove that $[\boldsymbol{\xi}_i|\boldsymbol{W}_i] = \mathrm{N}\{m(\boldsymbol{W}_i), \boldsymbol{\Sigma}_i\}$, where $m(\boldsymbol{W}_i)$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance matrix of the conditional distribution of $\boldsymbol{\xi}$ given the observed functional data and model (4). In section 2.3 we provide the derivation of $m(\boldsymbol{W}_i)$ and $\boldsymbol{\Sigma}_i$ and more insight into their effect on inference.

For most nonlinear models the induced model for observed data (5) does not have an explicit form. A procedure to avoid this problem is to use a two-stage approach with the following components: 1) produce predictors of $\boldsymbol{\xi}_i$, say $\widehat{\boldsymbol{\xi}}_i$, based on the exposure model (4); and 2) estimate the parameters of the outcome model (3) by replacing $\boldsymbol{\xi}_i$ with $\widehat{\boldsymbol{\xi}}_i$. It is reasonable to use the best linear unbiased predictor (BLUP) of $\boldsymbol{\xi}_i$, $\widehat{\boldsymbol{\xi}}_i = m(\boldsymbol{W}_i)$, but other predictors could also be used. For example, Müller and Stadtmüller [23] used $\widehat{\xi}_{ik} = \int_0^1 W_i(t)\psi_k(t)dt$, which are unbiased predictors of ξ_{ik}. We will show that these predictors may lead to biased estimators even in normal linear models. Moreover, they have higher variance than the BLUPs, $m(\boldsymbol{W}_i)$, because they do not borrow strength across subjects. This problem is especially serious when the number of observations per subject is small, but may be negligible when it is large.

A two-stage estimation procedure is an appealing alternative to joint model estimation. In particular, it is intuitive, computationally tractable, and provides unbiased estimators under the normality assumption. A drawback of the two-stage procedure is that it ignores the effect of variability of predictors, $\widehat{\boldsymbol{\xi}}$. This may lead to misspecified variability when the distribution of the outcome is normal and estimation bias and misspecified variability when it is not. To illustrate these ideas we show the effects of the two-stage procedure in Normal/identity and a Bernoulli/probit models.

The Normal/identity model. Assume that $Y_i = \boldsymbol{\xi}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma} + e_i$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^t$, $e_i \sim N(0, \sigma_e^2)$ and $\boldsymbol{\xi}_i$ are mutually independent. It can be shown that $E(Y_i|\boldsymbol{W}_i, \boldsymbol{Z}_i) = m^t(\boldsymbol{W}_i)\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma}$ and $\mathrm{Var}(Y_i|\boldsymbol{W}_i, \boldsymbol{Z}_i) = \boldsymbol{\beta}^t\boldsymbol{\Sigma}_i\boldsymbol{\beta} + \sigma_e^2$. In Section 2.3 we show that, in typical applications, $\boldsymbol{\Sigma}_i$ does not depend on $\boldsymbol{W}_i$ or $\boldsymbol{Z}_i$ but depends on the sampling times, t_{im}, for the function $\boldsymbol{W}_i$. In the case when $M_i = M$ and $t_{im} = t_m$, for all i and m, $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ and $\sigma_\eta^2 = \boldsymbol{\beta}^t\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma_e^2$ is still arbitrary because σ_e^2 is arbitrary. In this case the induced model for observed data is equivalent to

$$Y_i = m^t(\boldsymbol{W}_i)\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma} + \eta_i \tag{6}$$

where $\eta_i \sim \mathrm{Normal}(0, \sigma_\eta^2)$ are mutually independent. Thus, in the balanced case the two-stage procedure leads to unbiased estimators of the model parameters and correctly specified variability if and only if $\boldsymbol{\xi}_i$ is replaced by $m(\boldsymbol{W}_i)$. However, if the number of observations per subject, M_i, or the sampling points, t_{im}, vary with the subject i then $\boldsymbol{\Sigma}_i$ is not constant. In this case the maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ based on model (6) would still be consistent. However, their standard errors would be incorrect because the homoscedastic model (6) would be used when the actual variances are heteroscedastic.

The Bernoulli/probit model. Consider the following outcome model $Y_i|\boldsymbol{\xi}_i, \boldsymbol{Z}_i \sim \mathrm{Bernoulli}(p_i)$, where $\Phi^{-1}(p_i) = \boldsymbol{\xi}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma}$, and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Under the normality assumption of the distribution of $\boldsymbol{\xi}_i$ it follows that the induced model for observed data is

$$\begin{cases} Y_i|\boldsymbol{W}_i, \boldsymbol{Z}_i & \sim \quad \mathrm{Bernoulli}(q_i); \\ \Phi^{-1}(q_i) & = \quad \{m^t(\boldsymbol{W}_i)\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma}\}/(1 + \boldsymbol{\beta}^t\boldsymbol{\Sigma}_i\boldsymbol{\beta})^{1/2}. \end{cases} \tag{7}$$

Thus, using the two-stage procedure, where $\boldsymbol{\xi}_i$ is simply replaced by $m^t(\boldsymbol{W}_i)$, leads to biased estimators with misspecified variability for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The size of these effects is controlled by $\boldsymbol{\beta}^t\boldsymbol{\Sigma}_i\boldsymbol{\beta}$.

2.3 The posterior distribution of scores

In the previous section we showed that a two-stage estimation procedure results in biased estimators and that the size of the bias is affected, if not determined, by the covariance matrix, $\boldsymbol{\Sigma}_i$, of the conditional distribution, $[\boldsymbol{\xi}_i|\boldsymbol{W}_i]$. This type of problem is also encountered in measurement error models, where the analog of the two stage-stage procedure is referred to as regression calibration [2]. While, in that context, regression calibration has been criticized for the same reasons we describe here for two-stage procedures, it remains a fast and robust first order bias correction strategy that often outperforms more sophisticated methods. Thus, it is reasonable to ask whether and how much would be gained in the functional regression context by switching from a two-stage to a joint model analysis. To answer this questions we take a closer look at the the conditional distribution $[\boldsymbol{\xi}_i|\boldsymbol{W}_i]$ and provide a simplified, but revealing, example at the end of this section. In Section 6 we provide further insight into the size of the bias and bias correction using both methods.

Under the assumptions in models (3) and (4) the joint distribution of $(\boldsymbol{W}_i^t, \boldsymbol{\xi}_i^t)^t$ is multivariate normal with zero mean. Because $\text{var}\{W_i(t)\} = \sigma_\epsilon^2 + \sum_{k=1}^K \lambda_k\psi_k^2(t)$, $\text{cov}\{W_i(t), W_i(s)\} = \sum_{k=1}^K \lambda_k\psi_k(t)\psi_k(s)$, $\text{var}\{\xi_{ik}\} = \lambda_k$ and $\text{cov}\{W_i(t), \xi_{ik}\} = \lambda_k\psi_k(t)$ it follows that $(\boldsymbol{W}_i^t, \boldsymbol{\xi}_i^t)^t \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_i)$ where

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_\epsilon^2\boldsymbol{I}_{M_i} + \boldsymbol{\Psi_i}\boldsymbol{\Lambda}\boldsymbol{\Psi_i}^t & \boldsymbol{\Psi_i}\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t & \boldsymbol{\Lambda} \end{pmatrix}, \tag{8}$$

$\boldsymbol{I}_{M_i}$ is the M_i dimensional identity matrix, $\boldsymbol{\Psi}_i$ is the $M_i \times K$ dimensional matrix with the jth row equal to $\boldsymbol{\psi}_{im}^t = \{\psi_1(t_{im}), \ldots, \psi_K(t_{im})\}$, and $\boldsymbol{\Lambda}$ is the $K \times K$ dimensional diagonal matrix with the diagonal equal to $(\lambda_1, \ldots, \lambda_K)$. It follows that $[\boldsymbol{\xi}_i|\boldsymbol{W}_i] = \mathrm{N}\{m(\boldsymbol{W}_i), \boldsymbol{\Sigma}_i\}$, where

$$\begin{cases} m(\boldsymbol{W}_i) & = & \boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t(\sigma_\epsilon^2 \boldsymbol{I}_{M_i} + \boldsymbol{\Psi}_i\boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t)^{-1}\boldsymbol{W}_i; \\ \boldsymbol{\Sigma}_i & = & \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t(\sigma_\epsilon^2 \boldsymbol{I}_{M_i} + \boldsymbol{\Psi}_i\boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t)^{-1}\boldsymbol{\Psi}_i\boldsymbol{\Lambda}. \end{cases} \tag{9}$$

The first equation in (9) provides the recipe for calculating the BLUPs of the functional scores based on the exposure model (4). A careful inspection of the second equation in (9) reveals important characteristics of $\boldsymbol{\Sigma}_i$. First, $\boldsymbol{\Sigma}_i \leq \boldsymbol{\Lambda}$, that is, $\boldsymbol{\Lambda} - \boldsymbol{\Sigma}_i$ is positive-semidefinite, where $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Lambda}$ are the conditional and prior covariance matrices of $\boldsymbol{\xi}_i$, respectively. This is a quantification of the natural reduction of variability after conditioning on observed data. Despite this reduction, $\boldsymbol{\Sigma}_i$ is not zero. Second, the amount of variability described by $\boldsymbol{\Sigma}_i$ depends essentially on the prior covariance, $\boldsymbol{\Lambda}$, and the variance, σ_ϵ^2, of the error process $\epsilon_i(t)$. In particular, when σ_ϵ^2 approaches infinity, $\boldsymbol{\Sigma}_i$ approaches $\boldsymbol{\Lambda}$ at the rate $O(\sigma_\epsilon^{-2})$, indicating that large noise levels will correspond to little or no reduction of variability. In practice, such extreme cases rarely occur, but a wide spectrum of noise levels might be expected. Depending on the noise levels, the matrix $\boldsymbol{\Sigma}_i$ will be closer to one of the extremes, $\boldsymbol{\Lambda}$ or $\boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t(\boldsymbol{\Psi}_i\boldsymbol{\Lambda}\boldsymbol{\Psi}_i^t)^{-}\boldsymbol{\Psi}_i\boldsymbol{\Lambda}$, corresponding to no or maximum variability reduction, respectively. Here A^- is a generalized inverse of A. Another way to gain insight into the problem is to write the matrix $\boldsymbol{\Sigma}_i$ in terms of the signal-to-noise ratio matrix, $\boldsymbol{\Lambda}/\sigma_\epsilon^2$, as $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}[\boldsymbol{I}_K - \boldsymbol{\Psi}_i^t\{\boldsymbol{I}_{M_i} + \boldsymbol{\Psi}_i(\boldsymbol{\Lambda}/\sigma_\epsilon^2)\boldsymbol{\Psi}_i^t\}^{-1}\boldsymbol{\Psi}_i(\boldsymbol{\Lambda}/\sigma_\epsilon^2)]$. Thus, when $\boldsymbol{\Lambda}/\sigma_\epsilon^2$ is close to zero $\boldsymbol{\Sigma}_i$ is close to $\boldsymbol{\Lambda}$.

To better understand the problem consider the simple example when $K = 1$ and $\psi(t) = 1$ for all t. The functional exposure model (4) becomes $W_i(t) = \xi_i + \epsilon_i(t)$, where $\xi_i \sim \mathrm{N}(0, \lambda)$ and $\epsilon_i(t) \sim \mathrm{N}(0, \sigma_\epsilon^2)$. For simplicity, we denoted by $\xi_i = \xi_{i1}$ and by $\lambda = \lambda_1$. This is exactly the classical measurement error model where $W_i(t)$ are viewed as M_i unbiased proxies of the variable measured with error, ξ_i, σ_ϵ^2 is the variance of the measurement error, and $\lambda/(\lambda+\sigma_\epsilon^2)$ is the reliability of the measurement mechanism.

In this case $\boldsymbol{\Sigma}_i$ is a scalar and using the results from equation (9) we obtain

$$\boldsymbol{\Sigma}_i = \frac{\lambda(\sigma_\epsilon^2/M_i)}{\lambda+\sigma_\epsilon^2/M_i} = \frac{\lambda}{1+(\lambda/\sigma_\epsilon^2)M_i} \leq \min(\lambda, \sigma_\epsilon^2/M_i).$$

These expressions reveal the various factors that affect the size of the conditional variance, $\boldsymbol{\Sigma}_i$, in this simplified context. First, a large number of observations, M_i, or a small value of the error variance, σ_ϵ^2, correspond to a small $\boldsymbol{\Sigma}_i$. Second, a large σ_ϵ^2 relative to λ, or a small signal-to-noise ratio, $\lambda/\sigma_\epsilon^2$, correspond to $\boldsymbol{\Sigma}_i \approx \lambda$. Third, $\boldsymbol{\Sigma}_i \leq \sigma_\epsilon^2/M_i$, which implies that in applications with a large number of observations per subject the bias induced by using a two-stage procedure might be negligible. However, applications with small to moderate number of observations per subject, large measurement error, or small signal-to-noise ratios require special attention.

An important particular case is when the functions are perfectly observed, that is when $\sigma_\epsilon^2 = 0$. In this case, if the functional data are single-level, as assumed in this section, then the two-stage procedure does not induce bias either for the linear or nonlinear models. This result cannot be generalized to the case when functional data are observed at multiple levels, as is the case in the SHHS application. Indeed, with the exception of exotic examples, functional data exhibits sizeable within-subject/between-visit variability, even when the functions are perfectly measured. In Section 4 we discuss the specific problems induced by two-stage procedures when functional data has a multi-level structure.

Focusing on $\boldsymbol{\Sigma}_i$ as a source of bias in a two-stage procedure provides important insights, but may also be slightly misleading. Indeed, the bias is more directly affected by the relative size of $\boldsymbol{\beta}^t\boldsymbol{\Sigma}_i\boldsymbol{\beta}$ and $m^t(\boldsymbol{W}_i)\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma}$ than by the absolute size of $\boldsymbol{\Sigma}_i$.

However, the relative size is more complicated to explain and interpret.

3 Bayesian inference

Because of the potential problems associated with two-stage procedures we propose to use joint modeling. Bayesian inference using Markov Chain Monte Carlo (MCMC) simulations of the posterior distribution provides a reasonable, robust, and well tested computational approach for this type of problems. Moreover, Bayesian inference can easily be extended to the more general models described in Sections 4 and 5. Possible reasons for the current lack of Bayesian methodology in functional regression analysis could be: 1) the connection between functional regression models and joint mixed effects models was not known; and 2) the Bayesian inferential tools were perceived as unnecessarily complex and hard to implement. We clarified the connection to mixed effects models in Section 2 and we now show that 2) is not true, thanks to intense methodological and computational research conducted over the last 10-20 years. See, for example, the monographs [1, 4, 11, 13] and the citations therein for a good overview of recent developments.

To be specific, we focus on a Bernoulli/logit outcome model with functional regressors. Other outcome models would be treated similarly. Consider the joint model with the outcome $Y_i \sim \text{Bernoulli}(p_i)$, linear predictor $\text{logit}(p_i) = \boldsymbol{\xi}_i^t \boldsymbol{\beta} + \boldsymbol{Z}_i^t \boldsymbol{\gamma}$ and functional exposure model $W_i(t_{im}) = \boldsymbol{\psi}_{im}^t \boldsymbol{\xi}_i + \epsilon_i(t_{im})$. We assume that $\xi_{ik} \sim \text{N}(0, \lambda_k)$ and $\epsilon_i(t_{im}) \sim \text{N}(0, \sigma_\epsilon^2)$ are a-priori mutually independent for $i = 1, \ldots, n$ and $m = 1, \ldots, M_i$. The parameters of the model are $\boldsymbol{\Omega} = \{(\boldsymbol{\xi}_i : i = 1, \ldots, n), \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}, \sigma_\epsilon^2\}$. While $\epsilon_i(t_{ij})$ are also unknown, we do not incorporate them in the set of parameters because they are automatically updated by $\epsilon_i(t_{im}) = W_i(t_{im}) - \boldsymbol{\psi}_{im}^t \boldsymbol{\xi}_i$. The prior for $\boldsymbol{\xi}_i$ was already defined and it is standard to assume that the fixed effects parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, are apriori independent, with $\boldsymbol{\beta} \sim \text{Normal}(0, \sigma_\beta^2 \boldsymbol{I}_K)$ and $\boldsymbol{\gamma} \sim \text{Normal}(0, \sigma_\gamma^2 \boldsymbol{I}_P)$

where σ_β^2 and σ_γ^2 are very large and P is the number of Z covariates. In our applications we used $\sigma_\beta^2 = \sigma_\gamma^2 = 10^6$, which we recommend when there is no reason to expect that the components of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ could be outside of the interval $[-1000, 1000]$. In some applications this priors might be inconsistent with the true value of the parameter. In this situations we recommend re-scaling $W_i(t_{im})$ and normalizing, or re-scaling, the Z covariates.

While standard choices of priors for fixed effects parameters exist and are typically non-controversial, the same is not true for priors of variance components. Indeed, the estimates of the variance components are known to be sensitive to the prior specification, see, for example, [6, 10]. In particular, the popular inverse-gamma priors may induce bias when their parameters are not tuned to the scale of the problem. This is dangerous in the shrinkage context where the variance components control the amount of smoothing. However, we find that with reasonable care, the conjugate gamma priors can be used in practice. Alternatives to gamma priors are discussed by, for example, [10, 25], and have the advantage of requiring less care in the choice of the hyperparameters. Nonetheless, exploration of other prior families for functional regression would be well worthwhile, though beyond the scope of this paper.

We propose to use the following independent inverse gamma priors $\lambda_k \sim \text{IG}(A_k, B_k)$, $k = 1, \ldots, K$, and $\sigma_\epsilon^2 \sim \text{IG}(A_\epsilon, B_\epsilon)$, where $\text{IG}(A, B)$ is the inverse of a gamma prior with mean A/B and variance A/B^2. We first write the full conditional distributions for all the parameters and than discuss choices of non-informative inverse gamma parameters. Here we treat λ_k as parameters to be estimated, but a simpler Empirical Bayes (EB) method proved to be a reasonable alternative in practice. More precisely, the EB method estimates λ_k by diagonalizing the functional covariance operator as described in Section 2.1. These estimators are than fixed in the joint model. In the following we present the inferential procedure for the case when λ_ks are estimated with obvious simplifications for the EB procedure where they would be fixed.

We use Gibbs sampling [12] to simulate $[\boldsymbol{\Omega}|\boldsymbol{D}]$, where $\boldsymbol{D}$ denotes the observed data. A particularly convenient partition of the parameter space and the associated full conditional distributions are described below.

$$\begin{array}{lcl}
[\boldsymbol{\beta}, \boldsymbol{\gamma}|\text{others}] & \propto & \exp[\sum_{i=1}^{n} Y_i(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma}) - \sum_{i=1}^{n} \log\{1 + \exp(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma})\}] \\
 & & \times \exp(-0.5\boldsymbol{\beta}^t\boldsymbol{\beta}/\sigma_\beta^2 - 0.5\boldsymbol{\gamma}^t\boldsymbol{\gamma}/\sigma_\gamma^2); \\
[\boldsymbol{\xi}_i|\text{others}] & \propto & \exp[Y_i(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma}) - \log\{1 + \exp(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \boldsymbol{Z}_i^t\boldsymbol{\gamma})\}] \\
 & & \times \exp\{-0.5||\boldsymbol{W}_i - \boldsymbol{\Psi}_i\boldsymbol{\xi}_i||^2/\sigma_\epsilon^2 - 0.5\boldsymbol{\xi}_i\boldsymbol{\Lambda}\boldsymbol{\xi}_i\}; \\
[\lambda_k|\text{others}] & \propto & \text{IG}\left\{n/2 + A_k, \sum_{i=1}^{n} \xi_{ik}^2/2 + B_k\right\}; \\
[\sigma_\epsilon^2|\text{others}] & \propto & \text{IG}\{\sum_{i=1}^{n} T_i/2 + A_\epsilon, \sum_{i=1}^{n} ||\boldsymbol{W}_i - \boldsymbol{\Psi}_i\boldsymbol{\xi}_i||^2/2 + B_\epsilon\}.
\end{array}$$

The first two full-conditionals do not have an explicit form, but can be sampled using Markov Chain Monte Carlo (MCMC). For Bernoulli outcomes the MCMC methodology is routine. We use the Metropolis-Hastings algorithm with a normal proposal distribution centered at the current value and small variance tuned to provide an acceptance rate around 30-40%. The last two conditionals are explicit and can be easily sampled. However, understanding the various components of these distributions will provide insights into rational choices of inverse gamma prior parameters. Indeed, the first parameter of the full conditional for λ_k is $n/2+A_k$, where n is the number of subjects. Thus, it is safe to choose $A_k \leq 0.01$. The second parameter is $\sum_{i=1}^{n} \xi_{ik}^2/2 + B_k$, where $\sum_{i=1}^{n} \xi_{ik}^2$ is an estimator of $n\lambda_k$. Thus, it is safe to choose $B_k \leq 0.01\lambda_k$. This discussion is especially relevant for those variance components or, equivalently, eigenvalues of the covariance operator, that are small but estimable. A similar discussion holds for σ_ϵ^2 and we recommend to choose $A_\epsilon \leq 0.01$ and $B_\epsilon \leq 0.01\sigma_\epsilon^2$. Note that MOM estimators for λ_k and σ_ϵ^2 are available and reasonable choices of B_k and B_ϵ are easy to propose. While we find these rules of thumb useful in practice, they should be used as any other rule of thumb, cautiously. Moreover, for every application we do not recommend to rigidly use these prior parameters but rather tune them according to the general principles described here.

4 Multi-level functional regression models

Multilevel functional data occurs naturally in scientific studies where subject-level functional data observed at multiple visits are becoming increasingly common. For example, our research was motivated by the largest collection of sleep EEG data on an epidemiologic cohort, which contains at each of two visits, quasi-continuous EEG signals for each subject. We provide two other examples inspired by our current research, but otherwise not covered in this paper. First, magnetic resonance imaging (MRI) has become commonly used in epidemiological studies and our applications contain images (e.g., of the brain or heart) at multiple visits. Second, the daily trajectory of blood glucose concentration may provide more information than simple summaries, such as fasting glucose. These are examples of what we refer to as Multilevel Functional Data (MFD), where functional data are observed at multiple visits. MFD should not be mistaken for ”functional longitudinal data”, which typically refers to data containing one function per subject.

This section expands the methodology described in Section 2 to account for the natural multilevel structure of functional data. Most results in Section 2 generalize directly to the multilevel case and require mainly notational and computational effort. However, there are important differences that we are noting here before providing the technical details below. The most important difference is that the subject-specific scores have higher variability due to the additional within-level/between-visit variability. This, in turn, leads to larger bias in a two-stage procedure when the outcome model is not linear. We will address this problem below, after introducing the framework for multilevel functional regression.

4.1 Joint mixed effects models

The observed data for the ith subject in a Generalized Multilevel Functional Model (GMFM) is $[Y_i, \boldsymbol{Z}_i, \{W_{ij}(t_{ijm}), t_{ijm} \in [0,1]\}]$, where Y_i is the continuous or discrete outcome, $\boldsymbol{Z}_i$ is a vector of covariates, and $W_{ij}(t_{ijm})$ is a random curve in $L_2[0,1]$ observed at time t_{ijm}, which is the mth observation, $m = 1, \ldots, M_{ij}$, for the jth visit, $j = 1, \ldots, J_i$ of the ith subject. We assume that $W_{ij}(t)$ is a proxy observation of the true underlying subject-specific functional signal $X_i(t)$, and that $W_{ij}(t) = \mu(t) + \eta_j(t) + X_i(t) + U_{ij}(t) + \epsilon_{ij}(t)$. Here $\mu(t)$ is the overall mean function, $\eta_j(t)$ is the visit j specific shift from the overall mean function, $X_i(t)$ is the subject i specific deviation from the visit specific mean function, and U_{ij} is the residual subject/visit specific deviation from the subject specific mean. To ensure identifiability we assume that $X_i(t)$, $U_{ij}(t)$, and $\epsilon_{ij}(t)$ are uncorrelated and that $\epsilon_{ij}(t)$ is a white noise process with variance σ_ϵ^2. Given the large sample size of the SHHS data, we can assume that $\mu(t)$ and $\eta_j(t)$ are estimated with negligible error by $\bar{W}_{\cdot\cdot}(t)$ and $\bar{W}_{\cdot j}(t) - \bar{W}_{\cdot\cdot}$, respectively. Here $\bar{W}_{\cdot\cdot}(t)$ is the average over all subjects, i, and visits, j, of $W_{ij}(t)$ and $\bar{W}_{\cdot j}(t)$ is the average over all subjects, i, of observation at visit j of $W_{ij}(t)$. We can assume that these estimates have been subtracted from $W_{ij}(t)$, so that $W_{ij}(t) = X_i(t) + U_{ij}(t) + \epsilon_{ij}(t)$.

We also assume that the distribution of Y_i is in the exponential family with linear predictor η_i and dispersion parameter α, denoted here by $\mathrm{EF}(\eta_i, \alpha)$. The linear predictor is assumed to have the following form $\eta_i = \int_0^1 X_i(t)\beta(t)dt + \boldsymbol{Z}_i^t \boldsymbol{\gamma}$, where $\beta(\cdot) \in L_2[0,1]$ is a functional parameter and the main target of inference. If $\psi_k^{(1)}(t)$ and $\psi_l^{(2)}(t)$ are two orthonormal basis in $L_2[0,1]$ then $X_i(\cdot)$, $U_{ij}(\cdot)$ have unique representations

$$X_i(t) = \sum_{k\geq 1} \xi_{ik}\psi_k^{(1)}(t), \;\; U_{ij}(t) = \sum_{l\geq 1} \zeta_{ijl}\psi_l^{(2)}(t); \;\; \beta(t) = \sum_{k\geq 1} \beta_k\psi_k^{(1)}(t). \qquad (10)$$

Using the same arguments as in Section 2, we use the truncated versions of these equalities. If K and L are the truncation lags, the multilevel outcome model can be written as

$$\begin{cases} Y_i & \sim & \text{EF}(\eta_i^K, \alpha); \\ \eta_i^K & = & \sum_{k=1}^{K} \xi_{ik}\beta_k + \boldsymbol{Z}_i^t\boldsymbol{\gamma}, \end{cases} \qquad (11)$$

which is identical to the single-level outcome model (3). Other multilevel outcome models could be considered by including regression terms for the $U_{ij}(t)$ process or, implicitly, for ζ_{ijl}. However, we restrict our discussion to models of the type (11).

In this paper we use Multilevel Functional Principal Component Analysis (MFPCA) [7] to obtain the bases that capture most of the functional variability of the space spanned by $X_i(t)$ and $U_{ij}(t)$, respectively, with the the first few components. MFPCA is based on the spectral decomposition of the within- and between-visit functional variability covariance operators. We summarize here the main components of this methodology. Denote by $K_T^W(s,t) = \text{cov}\{\,W_{ij}(s), W_{ij}(t)\,\}$ and $K_B^W(s,t) = \text{cov}\{\,W_{ij}(s), W_{ik}(t)\,\}$ for $j \neq k$ the total and the between covariance operator corresponding to the observed process, $W_{ij}(\cdot)$, respectively. Denote by $K^X(t,s) = \text{cov}\{X_i(t), X_i(s)\}$ the covariance operator of the $X_i(\cdot)$ process and by $K_T^U(t,s) = \text{cov}\{\,U_{ij}(s), U_{ij}(t)\,\}$ the total covariance covariance operator of the $U_{ij}(\cdot)$ process. Note that, by definition, $K_B^U(s,t) = \text{cov}\{\,U_{ij}(s), U_{ik}(t)\,\} = 0$ for $j \neq k$. Moreover, $K_B^W(s,t) = K^X(s,t)$ and $K_T^W(s,t) = K^X(s,t) + K_T^U(s,t) + \sigma_\epsilon^2\delta_{ts}$, where δ_{ts} is equal to 1 when $t = s$ and 0 otherwise. Thus, $K^X(s,t)$ can be estimated using a method of moments estimator of $K_B^W(s,t)$, say $\widehat{K}_B^W(s,t)$. For $t \neq s$ a method of moment estimator of $K_T^W(s,t) - K_B^W(s,t)$, say $\widehat{K}_T^U(s,t)$, can be used to estimate $K_T^U(s,t)$. To estimate $\widehat{K}_T^U(t,t)$ it was proposed [7] to predict $K_T^U(t,t)$ using a bivariate thin-plate spline smoother of $\widehat{K}_T^U(s,t)$ for $s \neq t$. This method was suggested by [31, 34] for single-level FPCA and shown to work well in the MFPCA context [7].

Once consistent estimators of $K^X(s,t)$ and $K_T^U(s,t)$ are available, the spectral decomposition and functional regression proceed as in the single-level case. More precisely, Mercer's theorem (see [15], Chapter 4) provides the following convenient spectral decompositions $K^X(t,s) = \sum_{k=1}^{\infty} \lambda_k^{(1)} \psi_k^{(1)}(t)\psi_k^{(1)}(s)$, where $\lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \ldots$ are the ordered eigenvalues and $\psi_k^{(1)}(\cdot)$ are the associated orthonormal eigenfunctions of $K^X(\cdot,\cdot)$ in the L^2 norm. Similarly, $K_T^U(t,s) = \sum_{l=1}^{\infty} \lambda_l^{(2)} \psi_l^{(2)}(t)\psi_l^{(2)}(s)$, where $\lambda_1^{(2)} \geq \lambda_2^{(2)} \geq \ldots$ are the ordered eigenvalues and $\psi_l^{(2)}(\cdot)$ are the associated orthonormal eigenfunctions of $K_T^U(\cdot,\cdot)$ in the L^2 norm. The Karhunen-Loève (KL) decomposition [18, 21] provides the following infinite decompositions $X_i(t) = \sum_{k=1}^{\infty} \xi_{ik}\psi_k^{(1)}(t)$ and $U_{ij}(t) = \sum_{l=1}^{\infty} \zeta_{ijl}\psi_l^{(2)}(t)$ where $\xi_{ik} = \int_0^1 X_i(t)\psi_k^{(1)}(t)dt$, $\zeta_{ijl} = \int_0^1 U_{ij}(t)\psi_l^{(2)}(t)dt$ are the principal component scores with $E(\xi_{ik}) = E(\zeta_{ijl}) = 0$, $\text{Var}(\xi_{ik}) = \lambda_k^{(1)}$, $\text{Var}(\zeta_{ijl}) = \lambda_l^{(2)}$. The zero-correlation assumption between the $X_i(\cdot)$ and $U_{ij}(\cdot)$ processes is ensured by the assumption that $\text{cov}(\xi_i, \zeta_{ijl}) = 0$. These properties hold for every i, j, k, and l.

Once the eigenfunctions and the truncation lags K and L are fixed, the model for observed functional data can be written as a linear mixed model. Indeed, by assuming a normal shrinkage distribution for scores and errors, the model can be rewritten as

$$\begin{cases} W_{ij}(t) & = & \sum_{k=1}^{K} \xi_{ik}\psi_k^{(1)}(t) + \sum_{l=1}^{L} \zeta_{ijl}\psi_k^{(2)}(t) + \epsilon_{ij}(t); \\ \xi_{ik} & \sim & \text{N}(0, \lambda_k^{(1)});\ \zeta_{ijl} \sim \text{N}(0, \lambda_l^{(2)});\ \epsilon_{ij}(t) \sim \text{N}(0, \sigma_\epsilon^2). \end{cases} \qquad (12)$$

For simplicity we will refer to $\psi_k^{(1)}(\cdot)$, $\psi_l^{(2)}(\cdot)$ and $\lambda_k^{(1)}$, $\lambda_l^{(2)}$ as the level 1 and 2 eigenfunctions and eigenvalues, respectively.

We propose to jointly fit the outcome model (11) and the exposure model (12). Because the joint model is a generalized linear mixed effects model the inferential arsenal for mixed effects models can be used. In particular, we propose to use a Bayesian analysis via posterior MCMC simulations. An alternative would be to use a two-stage analysis by first predicting the scores from model (12) using, for example, BLUP and then plug-in these estimates into model (11).

While the parallels between single-level and multilevel functional regression are obvious our presentation is far from being unnecessarily repetitive. Indeed, close inspection of exposure models (4) and (12) will reveal differences with important consequences. The most important difference is that model (12) contains the term $\sum_{l=1}^{L} \zeta_{ijl}\psi_l^{(2)}(t)$ which quantifies the visit/subject-specific deviations from the subject specific mean. This variability is typically large and makes estimation of the subject-specific scores, $\boldsymbol{\xi}_i$, difficult even when the functions are perfectly observed, that is $\sigma_\epsilon^2 = 0$. Thus, the effects of variability on bias in a two-stage procedure will typically be more severe in a multilevel context, especially when the within-subject variability is large compared to the between-subject variability. In the next section we provide the technical details associated with a two stage procedure and provide a simple example to build up the intuition.

4.2 Posterior distribution of subject-specific functional scores

We now turn our attention to calculating the posterior distribution of subject-specific scores for the MFPCA model (12). While this section is more technical and contains some pretty heavy notation, the results are important because they form the basis of any reasonable inferential procedure in this context, be it two-stage or joint modeling. We first introduce some notation for a subject i. Let $\mathbf{W}_{ij}$ =

$\{W_{ij}(t_{ij1}),\ldots,W_{ij}(t_{ijM_{ij}})\}^t$ be the $M_{ij} \times 1$ vector of observations at visit j, $\mathbf{W}_i = (\mathbf{W}_{i1}^t,\ldots,\mathbf{W}_{iJ_i}^t)^t$ be the $(\sum_{j=1}^{J_i} M_{ij}) \times 1$ vector of observations obtained by stacking $\boldsymbol{W}_{ij}$, $\boldsymbol{\psi}_{ij,k}^{(1)} = \{\psi_k^{(1)}(t_{ij1}),\ldots,\psi_k^{(1)}(t_{ijM_{ij}})\}^t$ be the $M_{ij} \times 1$ dimensional vector corresponding to the kth level 1 eigenfunction at visit j, and $\boldsymbol{\psi}_{ik}^{(1)} = \{\boldsymbol{\psi}_{i1,k}^{(1)t},\ldots,\boldsymbol{\psi}_{iJ_i,k}^{(1)t}\}^t$ be the $(\sum_{j=1}^{J_i} M_{ij}) \times 1$ dimensional vector corresponding to the kth level 1 eigenfunction at all visits. Also, let $\boldsymbol{\Psi}_{ij}^{(1)} = \{\boldsymbol{\psi}_{ij,1}^{(1)},\ldots,\boldsymbol{\psi}_{ij,K}^{(1)}\}$ be the $M_{ij} \times K$ dimensional matrix of level 1 eigenvectors obtained by binding the column vectors $\boldsymbol{\psi}_{ij,k}^{(1)}$ corresponding to the jth visit and $\boldsymbol{\Psi}_i^{(1)} = (\boldsymbol{\psi}_{i1}^{(1)},\ldots,\boldsymbol{\psi}_{iK}^{(1)})$ be the $(\sum_{j=1}^{J_i} M_{ij}) \times K$ dimensional matrix of level 1 eigenfunctions obtained by binding the column vectors $\boldsymbol{\psi}_{i1}^{(1)}$. Similarly, we define the vectors $\boldsymbol{\psi}_{ijl}^{(2)}$, $\boldsymbol{\psi}_{il}^{(2)}$, $\boldsymbol{\Psi}_{ij}^{(2)}$ and $\boldsymbol{\Psi}_i^{(2)}$. Finally, let $\boldsymbol{\Lambda}^{(1)} = \text{diag}\{\lambda_1^{(1)},\ldots,\lambda_K^{(1)}\}$ and $\boldsymbol{\Lambda}^{(2)} = \text{diag}\{\lambda_1^{(2)},\ldots,\lambda_L^{(2)}\}$ be the $K \times K$ and $L \times L$ dimensional diagonal matrices of level 1 and level 2 eigenvalues, respectively.

As in the single level case, $[\boldsymbol{\xi}_i|\boldsymbol{W}_i] = \text{Normal}\{m(\boldsymbol{W}_i),\boldsymbol{\Sigma}_i\}$, where $m(\boldsymbol{W}_i)$ and $\boldsymbol{\Sigma}_i$ have a more complex structure. Indeed, if $\Sigma_{\mathbf{W}_i}$ denotes the covariance matrix of $\boldsymbol{W}_i$ then $m(\boldsymbol{W}_i) = \boldsymbol{\Lambda}^{(1)}\ \boldsymbol{\Psi}_i^{(1)t}\ \boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1}\mathbf{W}_i$ and $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}^{(1)} - \boldsymbol{\Lambda}^{(1)}\ \boldsymbol{\Psi}_i^{(1)t}\ \boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1}\ \boldsymbol{\Psi}_i^{(1)}\boldsymbol{\Lambda}^{(1)}$. It can be shown that $\boldsymbol{\Sigma}_{\mathbf{W}_i}$ is a matrix with the (j,j')th block matrix equal to $B_{i,jj'}$ where $B_{i,jj'} = B_{i,j'j}^t = \boldsymbol{\Psi}_{ij}^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{ij'}^{(1)t}$ if $j \neq j'$ and $B_{i,jj} = \sigma_\epsilon^2\boldsymbol{I}_{M_{ij}} + \boldsymbol{\Psi}_{ij}^{(2)}\boldsymbol{\Lambda}^{(2)}\boldsymbol{\Psi}_{ij}^{(2)t} + \boldsymbol{\Psi}_{ij}^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{ij}^{(1)t}$ for $1 \leq j,j' \leq J_i$.

Theorem 1 *Consider the multilevel functional exposure model (12) with a fixed number of observations per visit, i.e.* $M_{ij} = M_i$, *at the same subject-specific times for each visit, i.e.* $t_{ijm} = t_{im}$ *for all* $j = 1,\ldots,J_i$. *Denote by* $K^X = \boldsymbol{\Psi}_{i1}^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{i1}^{(1)t}$, *by* $K_T^U = \boldsymbol{\Psi}_{i1}^{(2)}\boldsymbol{\Lambda}^{(2)}\boldsymbol{\Psi}_{i1}^{(2)t}$, *by* $\mathbf{1}_{J_i \times J_i}$ *the* $J_i \times J_i$ *dimensional matrix of ones, and by* $\otimes$ *the Kronecker product of matrices. Then* $\boldsymbol{\Sigma}_{\mathbf{W}_i} = \mathbf{1}_{J_i \times J_i} \otimes K^X + \boldsymbol{I}_{J_i} \otimes (\sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U)$ *and* $\boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1} = \boldsymbol{I}_{J_i} \otimes (\sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U)^{-1} - \mathbf{1}_{J_i \times J_i} \otimes \{(\sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U)^{-1}\ K^X\ (J_iK^X + \sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U)^{-1}\}$.

Theorem 2 *Assume the balanced design considered in Theorem 1 and denote by*

$\bar{\boldsymbol{W}}_i = \sum_{j=1}^{J_i} \boldsymbol{W}_{ij}/J_i$. *Then* $m(\boldsymbol{W}_i) = \boldsymbol{\Lambda}^{(1)}\ \boldsymbol{\Psi}_{i1}^{(1)t}\{K^X + \frac{1}{J_i}(\sigma_\epsilon^2 \boldsymbol{I}_{M_i} + K_T^U)\}^{-1}\ \bar{\boldsymbol{W}}_i$ *and* $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}^{(1)} - \boldsymbol{\Lambda}^{(1)}\ \boldsymbol{\Psi}_{i1}^{(1)t}\{K^X + \frac{1}{J_i}(\sigma_\epsilon^2 \boldsymbol{I}_{M_i} + K_T^U)\}^{-1}\ \boldsymbol{\Psi}_{i1}^{(1)}\boldsymbol{\Lambda}^{(1)}$.

See the Appendix for proofs.

Theorem 2 provides a particularly simple description of the conditional distribution $\boldsymbol{\xi}_i|\boldsymbol{W}_i$. Moreover, it shows that, conditional on the smoothing matrices $\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{\Lambda}^{(2)}$, the conditional distribution $\boldsymbol{\xi}_i|\boldsymbol{W}_i$ is the same as the conditional distribution $\boldsymbol{\xi}_i|\bar{\boldsymbol{W}}_i$. We now provide a simple example where all calculations can be done explicitly to illustrate the contribution of each individual source of variability to the variability of the posterior distribution $\boldsymbol{\xi}_i|\boldsymbol{W}_i$, $\boldsymbol{\Sigma}_i$. As described in section 2.2, this variability affects the size of the estimation bias in a two-stage procedure. Thus, it is important to understand in what applications this might be a problem.

Consider a balanced design model with $K = L = 1$ and $\psi^{(1)}(t) = 1$, $\psi^{(2)}(t) = 1$ for all t. The exposure model becomes a balanced mixed two-way ANOVA model

$$\begin{cases} W_{ij}(t) & = & \xi_i + \zeta_{ij} + \epsilon_{ij}(t); \\ \xi_i & \sim & \mathrm{N}(0, \lambda_1);\ \zeta_{ij} \sim \mathrm{N}(0, \lambda_2);\ \epsilon_{ij}(t) \sim \mathrm{N}(0, \sigma_\epsilon^2), \end{cases} \tag{13}$$

where, for simplicity, we denoted by $\xi_i = \xi_{i1}$, $\zeta_{ij} = \zeta_{ij1}$, $\lambda_1 = \lambda_1^{(1)}$ and by $\lambda_2 = \lambda_1^{(2)}$. In this case the conditional variance Σ_i is a scalar and, using the results from Theorem 2, we obtain $\Sigma_i = \frac{\lambda_1\{\lambda_2/J_i + \sigma_\epsilon^2/(M_i J_i)\}}{\lambda_1 + \{\lambda_2/J_i + \sigma_\epsilon^2/(M_i J_i)\}} \leq \min\{\lambda_1,\ \lambda_2/J_i + \sigma_\epsilon^2/(M_i J_i)\}$. Several important characteristics of this formula have direct practical consequences. First, the within-subject/between-visit variability, λ_2, is divided by the number of visits, J_i. In many applications λ_2 is large compared to λ_1 and J_i is small, leading to a large variance Σ_i. For example, in the SHHS study $J_i = 2$ and the functional analog of λ_2 is roughly 4 times larger than the functional analog of λ_1. Second, in contrast to the single-level case, even when functions are perfectly observed, that is $\sigma_\epsilon^2 = 0$, the variance $\boldsymbol{\Sigma}_i$ is not zero. Third, in many applications $\sigma_\epsilon^2/(M_iJ_i)$ is negligeable because the total number of observations for subject i, M_iJ_i, is large. For example, in the SHHS, $M_iJ_i \approx 1600$.

5 Generalizing the functional regression model

We have shown that single- and multilevel functional regression models can be viewed as mixed effects models. Thus, the inferential machinery developed for mixed effects models can be applied to complex functional regression settings with only minimal changes. Another important consequence is that the mixed effects framework provides a natural and modular framework for generalization. Indeed, mixed effects regression modules developed for other problems can easily be incorporated with the methodology described here. For example, consider the case when an additional covariate, say d, in the linear predictor equation (3) has a smooth effect on the outcome. More precisely, the linear predictor has the form $\eta_i^K = \sum_{m=1}^K \xi_{ik}\beta_k + \boldsymbol{Z}_i^t\boldsymbol{\gamma} + f(d_i)$, where $f(\cdot)$ is unspecified. Using, for example, penalized splines regression [29] the function $f(\cdot)$ can be parameterized as $f(d_{1i}) = \alpha_0 + \alpha_1 d_i + \ldots + \alpha_p d_i^p + \sum_{l=1}^L a_l(d_i - \kappa_l)_+^p$, where $a_l \sim \mathrm{N}(0, \sigma_a^2)$, p is the degree of the spline, κ_l, $l = 1, \ldots, L$ are fixed knots, and $x_+^p = x^p$ if $x > 0$ and 0 otherwise. Thus, the outcome model remains a mixed effects model by simply viewing $\alpha_0, \ldots, \alpha_p$ as fixed effects and a_l, $l = 1, \ldots, L$ as random effects parameters. Extensions to multiple uni- or multivariate smooth functions is similar and will not be described here in detail. Such extensions are neither exotic nor rare. For example, in the SHHS the outcome, Y_i, could be the Chronic Heart Disease (CHD) indicator, the functional regressors, $W_{ij}(\cdot)$, could be the normalized sleep EEG δ-power, and d could be age or body mass index (BMI) or both.

A different type of generalization occurs when the outcome observations are clustered. For example, consider the case when one observes the outcome Y_{ij} for a subject i at visit j. A standard approach to account for correlation is to add a visit-specific random intercept. More precisely, the linear predictor of the outcome model is $\eta_{ij}^K = r_i + \eta_{ij}^K$, where $r_i \sim \text{Normal}(0, \sigma_v^2)$ are visit-specific random intercepts and η_{ij}^K is the linear predictor with a structure as in (3). Thus, as with adding smooth functions, adding random intercepts is equivalent to adding a layer of random effects that control the shrinkage of the cluster means or, equivalently, the correlation among same-cluster observations. Other, more complex, random effects structures may also be added using similar constructions.

6 Simulation studies

In this section, we compare the performance of the joint analysis procedure with the two-stage procedure through simulation studies. We examine the Bernoulli model with probit link when the functional exposure model is single-level, as in Section 2, and multilevel, as in Section 4.

The outcome data was simulated from a Bernoulli/probit model with linear predictor $\Phi^{-1}(p_i) = \beta_0 + \int_0^1 X_i(t)\beta(t)\,dt + z_i\gamma$, for $i = 1, \ldots, n$, where $n = 1000$ is the number of subjects. We used the functional predictor $X_i(t) = \xi_i\psi_1(t)$, where $\xi_i \sim N(0, \lambda_1)$ and $\psi_1(t) \equiv 1$, evaluated at $M = 15$ equidistant time points in $[0, 1]$. We set $\beta_0 = 1$, $\gamma = 1$ and a constant functional parameter $\beta(t) \equiv \beta$. The z_is are taken equally spaced between $[-1, 1]$ with $z_1 = -1$ and $z_n = 1$. Note that the linear predictor can be re-written as $\Phi^{-1}(p_i) = \beta_0 + \beta\xi_i + z_i\gamma$. In the following subsections we conduct simulations with different choices of β and type of functional exposure model. All models are fit using joint Bayesian inference via MCMC posterior simulations and a two-stage approach using either BLUP or numerical integration [23]. We simulated $N = 100$ data sets from each model.

6.1 Single-level functional exposure model

Consider the case when for each subject, i, instead of observing $X_i(t)$, one observes the noisy predictors $W_i(t)$, where $W_i(t) = X_i(t) + \epsilon_i(t)$, $i = 1, \ldots, n$ and $\epsilon_i(t_m) \sim \text{Normal}(0, \sigma_\epsilon^2)$ is the measurement error. We set $\lambda_1 = 1$, consider three values of the signal $\beta = 0.5$, 1.0, 1.5 and three different magnitudes of noise $\sigma_\epsilon = 0$ (no noise), $\sigma_\epsilon = 1$ (moderate) and $\sigma_\epsilon = 3$ (very large). Figure 8 shows the boxplots of the parameter estimates $\hat{\beta}$ and $\hat{\gamma}$. The top and bottom panels provide results for the joint Bayesian analysis and the two-stage analysis with BLUP, respectively. The left and middle panels display the parameter estimates for different magnitudes of noise and the right panel presents the bias of the estimates of β for several true values of β. For the two-stage procedure when the amount of noise, σ_ϵ, or the absolute value of the true parameter, $|\beta|$, increases, the bias increases. These results confirm our theoretical discussion in Section 2 and indicate that bias is a problem both for the parameters of the functional variables measured with error and of the perfectly observed covariates. Moreover, bias increases when the true functional effect increases as well as when measurement error increases.

For the case $\sigma_\epsilon = 3$, Table 1 displays the root mean squared error (RMSE) and coverage probability of confidence intervals for β and γ. The two-stage approach with scores estimated by numerical integration has a much higher RMSE than the other two methods, which have a practically equal RMSE. However, it would be misleading to simply compare the RMSE for the joint Bayesian inference and the two-stage procedure based on BLUP estimation. Indeed, the coverage probability for the latter procedure is far from the nominal level and can even drop to zero. This is an example of good RMSE obtained by a combination of two wrong reasons: the point estimate is biased and the variance is underestimated.

Method	β	$\hat{\beta}$			$\hat{\gamma}$		
		RMSE	80%CI cov.	50%CI cov.	RMSE	80%CI cov.	50%CI cov.
Numerical integration	0.5	0.20	0.00	0.00	0.10	0.79	0.46
	1.0	0.46	0.00	0.00	0.17	0.41	0.09
	1.5	0.81	0.00	0.00	0.27	0.03	0.00
BLUP	0.5	0.06	0.84	0.56	0.10	0.79	0.46
	1.0	0.16	0.26	0.11	0.17	0.41	0.09
	1.5	0.40	0.01	0.00	0.27	0.03	0.00
Bayesian	0.5	0.07	0.85	0.58	0.11	0.77	0.54
	1.0	0.14	0.83	0.48	0.14	0.80	0.52
	1.5	0.39	0.85	0.51	0.23	0.86	0.49

Table 1: The comparison between the two-stage estimates (with numerical integration or BLUP) and Bayesian estimates of β and γ with respect to root mean squared error (RMSE), and coverage probability of the 80% and 50% confidence intervals (80%CI cov. and 50%CI cov.) for $\sigma_\epsilon = 3$.

6.2 Multilevel functional exposure model

Consider now the situation when the predictors are measured through a hierarchical functional design, as in SHHS. To mimic the design of the SHHS, we assume $J = 2$ visits per subject and that the observed noisy predictors $W_{ij}(t)$ are generated from the model $W_{ij}(t) = X_i(t) + U_{ij}(t) + \epsilon_{ij}(t)$, for each subject $i = 1, \ldots, n$ and visit $j = 1, \ldots, J$, where $\epsilon_{ij}(t) \sim \text{Normal}(0, \sigma_\epsilon^2)$ and $U_{ij}(t) = \zeta_{ij}\psi_2(t)$ with $\zeta_{ij} \sim \text{Normal}(0, \lambda_2)$, $\psi_2(t) \equiv 1$. We used various choices of λ_1, λ_2 and σ_ϵ^2, and compared the two-stage analysis with the scores estimated by BLUP with a joint Bayesian analysis. As in the single-level case, the bias depends on the factor $1 + \beta^2\Sigma_i$ and the only technical difference is the calculation of Σ_i. Thus, we limit our analyses to the case $\beta = 1$ and examine the effects of the other factors that may influence estimation.

Figure 2 presents the boxplots of the estimates of β using the joint Bayesian analysis (top panels) and the two-stage method with BLUP estimation of scores (bottom panels). The left panels correspond to $\lambda_1 = 1$, $\lambda_2 = 1$ and three values of σ_ϵ, 0.5, 1 and 3. The joint Bayesian inference produces unbiased estimates, while the two-stage procedure produces biased estimates with the bias increasing only slightly with the measurement error variance. This confirms our theoretical results that indicated that, typically, in the hierarchical setting the noise magnitude is not the main source of bias. The middle and right panels display results when the measurement error variance is fixed, $\sigma_\epsilon = 1$. The middle panels show results for the case when the between-subject variance is small, $\lambda_1 = 0.1$, and three values of the within-subject variance, $\lambda_2 = 0.1$, 0.4 and 0.8. The right panels show results for the case when the between-subject variance is large, $\lambda_1 = 3$, and three values of the within-subject variance, $\lambda_2 = 1$, 3 and 5. These results confirm our theoretical analyses in Section 4. Indeed, bias is small when the between-subject variability, λ_1, even when the within subject variability, λ_2, is much larger relative to λ_1. When λ_1 is large then bias is much larger and increases with λ_2. In contrast, the joint Bayesian analysis produces unbiased estimators with variability increasing with λ_2. The RMSE and coverage probability results were similar to the ones for the single-level case. We have also obtained similar results for γ. While these results are not reported here they are available upon request and can be reproduced using the attached simulation software.

In spite of the obvious advantages of the joint Bayesian analysis, the message is more nuanced than simply recommending this method. In practice, the two-stage method with BLUP estimation of scores is a robust alternative that often produces similar results to the joint analysis with less computational effort. Our recommendation is to apply both methods and compare their results. We also provided insight into why and when inferential differences may be observed, and, especially, how to address such differences.

7 The analysis of sleep data from the SHHS

We now apply our proposed methods to the SHHS data. We considered $3,201$ subjects with complete baseline and visit 2 data with sleep duration that exceeds 4 hours at both visits and we analyzed data for the first 4 hours of sleep. We focus on the association between hypertension (HTN) and sleep EEG δ-power spectrum. Complete descriptions of the SHHS data set and of this functional regression problem can be found in [5, 7]. We provide here a short summary.

A quasi-continuous EEG signal was recorded during sleep for each subject at two visits, roughly 5 years apart. This signal was processed using the Discrete Fourier Transform (DFT). More precisely, if $x_0, \ldots, x_{N-1}$ are the N measurements from a raw EEG signal then the DFT is $F_{x,k} = \sum_{n=0}^{N-1} x_n e^{-2\pi ink/N}, k = 0, \ldots, N-1$, where $i = \sqrt{-1}$. If W denotes a range of frequencies, then the power of the signal in that frequency range is defined as $P_W = \sum_{k \in W} F_{x,k}^2$. Four frequency bands were of particular interest: 1) δ [0.8-4.0Hz]; 2) θ [4.1-8.0Hz]; 3) α [8.1-13.0Hz]; 4) β [13.1-20.0Hz]. These bands are standard representations of low (δ) to high (β) frequency neuronal activity. The normalized power in the δ band is $\mathrm{NP}_\delta = P_\delta/(P_\delta+P_\theta+P_\alpha+P_\beta)$. Because of the nonstationary nature of the EEG signal the DTF and normalization are applied in adjacent 30 second intervals resulting in the function of time $t \rightarrow \mathrm{NP}_\delta(t)$, where t indicates the time corresponding to a particular 30 second interval. To illustrate this, the dots in Figure 3 are the pairs $\{t, \mathrm{NP}_\delta(t)\}$ while the solid lines represent the estimated mean function using penalized splines. Our goal is to regress HTN on the subject-specific functional characteristics that do not depend on random or visit-specific fluctuations.

The first step was to subtract from each observed normalized function the corresponding visit-specific population average. Following notations in Section 4.2, $W_{ij}(t)$

denotes these "centered" functional data for subject i at visit j during the tth 30-second interval. We used model (12) as the exposure model where the subject-level function, $\sum_{k=1}^{K} \xi_{ik}\psi_k^{(1)}(t)$, is the actual functional predictor used for HTN.

To obtain the subject- and visit-level eigenfunctions and eigenvalues we used the MFPCA methodology introduced by [7] and summarized in section 4.1. Table 2 provides the estimated eigenvalues at both levels indicating that Level 1 variability, associated with subject-level variability, is practically explained by the first three dimensions. For example, the first eigenvalue explains 80.6% of the variation, while the second and third eigenvalues explain 7.7% and 3.7% of variation, respectively. Together, they explain more than 91% of the subject level variation. Figure 4 provides the graphical representation of subject-level variability. The top-left plot displays the average over all subjects and visits as well as the visit specific averages of the normalized sleep δ-power. The other three panels display the first three subject-level eigenfunctions. The first eigenfunction is positive and roughly constant, indicating that subjects with positive scores on this component will tend to get a consistently larger proportion of sleep EEG δ-power than the population average.

Table 2 indicates that there are more directions of variation in the Level 2 functional space, associated with visit deviations from the subject-specific means. Indeed, 90% of the variability is explained by the first 14 principal components with 50% of the variability being explained by the first 4 components. The proportion of variability explained by subject-level functional clustering was $\hat{\rho}_W = 0.213$ (95% confidence interval: $0.210, 0.236$), i.e, 21.3% of variability in the sleep EEG δ-power is attributable to the subject-level variability.

We considered the following model $\text{logit}\{P(Y_i = 1)\} = \beta_0 + \sum_{k=1}^{K} \beta_k \xi_{ik} + Z_i^T$, where Y_i is the HTN indicator variable, $K = 3$, ξ_{ik} is the score of subject i on the kth subject specific eigenfunction, $\psi_k^{(1)}(t)$, and Z_i is a vector of other covariates. Table 3

Level 1 eigenvalues							
Component	1	2	3				
eigenvalue ($\times 10^{-3}$)	13.00	1.24	0.55				
% var	80.59	7.68	3.38				
cumm. % var	80.59	88.27	91.66				
Level 2 eigenvalues							
Component	1	2	3	4	5	6	7
eigenvalue ($\times 10^{-3}$)	12.98	7.60	7.46	6.45	5.70	4.47	3.07
% var	21.84	12.79	12.55	10.85	9.58	7.52	5.17
cumm. % var	21.84	34.63	47.17	58.02	67.61	75.13	80.30

Table 2: Estimated eigenvalues on both levels for SHHS data. We showed the first 3 components for level 1 (subject level), and 7 components for level 2.

provides results for two models, one without confounding adjustment (labeled Model 1) and one with confounding adjustment (labeled Model 2). The confounders in Model 2 are sex, smoking status (with three categories: never smokers, former smokers, and current smokers), age, body mass index (BMI) and respiratory disturbance index (RDI). Each model was fitted using a two-stage analysis with BLUP estimates of scores from the exposure model and a joint Bayesian analysis.

Using a two-stage analysis, both models indicated that the first principal component score is strongly and negatively associated with hypertension. The magnitude of association varies with the amount of confounding adjustment. For example, Model 1 estimates that a subject with one unit increase in the first principal component has $e^{-1.58} = 0.204$ (p value: < 0.001) times the odds of HTN. Considering the scale of the principal component scores (the first principal component scores have mean zero and standard deviation 0.11), standardized coefficients would be easier to interpret. After standardizing, one standard deviation increase in the first principal component score is associated with an odds ratio $e^{-0.205} = 0.815$ (p value: < 0.001). Model 2, which adjusts for all the confounders, estimated an odds ratio of $e^{-0.86} = 0.423$ per unit increase in the first principal component score, or an odds ratio $e^{-0.11} = 0.895$ per one standard deviation increase in the first principal component score. The second and third principal components were not found to be associated with hypertension. The negative relationship between smoking and hypertension may seem counterintuitive. However, in this study smokers are younger, have a lower body mass index and many other smokers with severe disease were not included in the study [35].

The point estimators and scientific findings are similar for the two methods, while the joint Bayesian analysis produces wider confidence intervals. Given the results in this paper, these results are easy to explain. As in our simulation example in Section 6.2, the between-subject variability is relatively small and bias is minimal in spite of the much larger within-subject variability. However, the joint Bayesian analysis correctly incorporates the variability that the two-stage procedure ignores and produces wider confidence intervals.

It is important to note that the joint Bayesian analysis is simple, robust and requires only minimal tunning. This is possible because of our MFPCA method, which produces a parsimonious decomposition of the functional variability using orthonormal bases. The effect of using orthonormal bases is to reduce the posterior correlation of corresponding parameters, which leads to excellent mixing properties and stable inference. For example, Figure 5 displays chains for regression coefficients of the principal component scores for Model 1 and the corresponding autocorrelation functions. The lack of correlation and fast convergence to the target distribution is very encouraging. Thus, we recommend the joint Bayesian analysis as a practical, not only theoretically appealing, approach to inference in this context.

8 Discussion

The methodology introduced in this paper was motivated by many current studies where exposure or covariates are functional data collected at multiple time points.

	Two-stage analysis		Joint analysis	
	Model 1	Model 2	Model 1	Model 2
score 1	-1.58 (0.28)*	-0.86 (0.30)*	-1.56 (0.36)*	-0.77 (0.35)*
score 2	0.66 (0.97)	-0.26 (1.04)	0.54 (1.74)	-2.94 (1.35)*
score 3	1.74 (1.56)	-0.26 (1.67)	3.82 (2.03)	-0.20 (3.39)
sex		0.10 (0.08)		0.12 (0.08)
smk:former		-0.19 (0.08)*		-0.19 (0.08)*
smk:current		-0.11 (0.13)		-0.10 (0.13)
age		0.06 (0.00)*		0.07 (0.00)*
BMI		0.06 (0.01)*		0.06 (0.01)*
RDI		0.01 (0.00)*		0.01 (0.00)*

Table 3: Models for association between hypertension and sleep EEG δ-power. Smoking status has three categories: never smokers (reference), former smokers (smk:former) and current smokers (smk.current). For the variable sex, female is the reference group and an asterisks indicates significance at level 0.05.

The SHHS is just one example of such studies. The GMFLM methodology provides a self contained set of statistical tools that is robust, fast and reasonable for such studies. These properties are due to: 1) the connection between GMFLMs and mixed effects models; 2) the parsimonious decomposition of functional variability in principal directions of variation; 3) the modular way mixed effects models can incorporate desirable generalizations; and 4) the good properties of Bayesian posterior simulations due to the orthogonality of the directions of variation.

The methods described in this paper have a few limitations. First, they require a large initial investment in developing and understanding the multilevel functional structure. Second, they require many choices including number and type of basis functions, distribution of random effects, method of inference, etc. The choices we made are reasonable, but other choices may be more appropriate in other applications. Third, the computational problems may seem daunting, especially when we propose a joint Bayesian analysis of a data set with thousands of subjects, multiple visits and thousands of random effects. However, we do not think that this is a real problem, and to address this issue we posted the software we developed for our simulations in Section 6 at `www.biostat.jhsph.edu/~ccrainic/webpage/software/GFR.zip`.

Appendix

Proof of Theorem 1. In a balanced design, the M_i-dimensional vector $\boldsymbol{\psi}_{ij,k}^{(1)}$ does not depend on j. Thus, the M_iJ_i-dimensional vector corresponding to the level 1 eigenfunction $\psi_k^{(1)}$ can be written as $\boldsymbol{\psi}_{ik}^{(i)} = \mathbf{1}_{J_i} \otimes \boldsymbol{\psi}_{i1,k}^{(1)}$, where $\mathbf{1}_{J_i}$ is the J_i-dimensional vector of ones. It follows that the matrix $\boldsymbol{\Psi}_{ij}^{(1)} = \boldsymbol{\Psi}_{i1}^{(1)}$ for all $1 \leq j \leq J_i$. Moreover, $B_{i,jj'} = \boldsymbol{\Psi}_{i1}^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{i1}^{(1)t} = K^X$ for $j \neq j'$ and $B_{i,jj} = \sigma_\epsilon^2\boldsymbol{I}_{M_i} + \boldsymbol{\Psi}_{i1}^{(2)}\Lambda^{(2)}\Psi_{i1}^{(2)t} + \boldsymbol{\Psi}_{i1}^{(1)}\boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{i1}^{(1)t} = \sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U + K^X$. For simplicity, the dependence of K_T^U and K^X on i has been suppressed. Thus, $\boldsymbol{\Sigma}_{\mathbf{W}_i} = \mathbf{1}_{J_i \times J_i} \otimes K_T^X + \boldsymbol{I}_{J_i} \otimes (\sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U)$. It is enough to show that $\boldsymbol{\Sigma}_{\mathbf{W}_i}\boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1} = \boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1}\boldsymbol{\Sigma}_{\mathbf{W}_i} = \boldsymbol{I}_{M_iJ_i}$, where $\boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1}$ is given in Theorem 1. We only prove that $\boldsymbol{\Sigma}_{\mathbf{W}_i}\boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1} = I_{M_iJ_i}$ since the proof of the second equality is analogous.

For simplicity of presentation, denote by $D = \sigma_\epsilon^2\boldsymbol{I}_{M_i} + K_T^U$ and $C = K^X$. Also, $\boldsymbol{\Sigma}_{\mathbf{W}_i} = \mathbf{1}_{J_i \times J_i} \otimes C + \boldsymbol{I}_{J_i} \otimes D$. Thus, $\boldsymbol{\Sigma}_{\mathbf{W}_i}\boldsymbol{\Sigma}_{\mathbf{W}_i}^{-1}$ is equal to

$$\boldsymbol{I}_{M_iJ_i} + \mathbf{1}_{J_i \times J_i} \otimes \{CD^{-1} - J_iCD^{-1}C\ (D + J_iC)^{-1} - C(D + J_iC)^{-1}\}.$$

It is enough to show that $CD^{-1} - J_iCD^{-1}C\ (D + J_iC)^{-1} - C(D + J_iC)^{-1} = 0$. This equality holds because $J_iCD^{-1}C\ (D + J_iC)^{-1} = CD^{-1}(D + J_iC - D)\ (D + J_iC)^{-1} = CD^{-1} - C(D + J_iC)^{-1}$.

Proof of Theorem 2. We use the same notations as in Theorem 1. For balanced designs $\boldsymbol{\psi}_{ij,k}^{(1)} = \boldsymbol{\psi}_{i1,k}^{(1)}$ for $1 \leq j \leq J_i$. Hence, $\boldsymbol{\Psi}_i^{(1)} = \mathbf{1}_{J_i} \otimes \boldsymbol{\Psi}_{i1}^{(1)}$ as $\boldsymbol{\psi}_{ik}^{(i)} = \mathbf{1}_{J_i} \otimes \boldsymbol{\psi}_{i1,k}^{(1)}$ and $\boldsymbol{\Psi}_{i1}^{(1)} = (\boldsymbol{\psi}_{i1,1}^{(1)}, \ldots, \boldsymbol{\psi}_{i1,K}^{(1)})$. Thus, $m(\boldsymbol{W}_i)$ is

$$\begin{aligned}
&= \boldsymbol{\Lambda}^{(1)}\left\{\mathbf{1}_{J_i}^T \otimes \boldsymbol{\Psi}_{i1}^{(1)t}\right\}\ \left[\boldsymbol{I}_{J_i} \otimes D^{-1} - \mathbf{1}_{J_i \times J_i} \otimes \left\{D^{-1}C(D + J_iC)^{-1}\right\}\right]\boldsymbol{W}_i\\
&= \boldsymbol{\Lambda}^{(1)}\left[\mathbf{1}_{J_i}^T \otimes \{\boldsymbol{\Psi}_{i1}^{(1)t}D^{-1}\} - J_i\mathbf{1}_{J_i}^T \otimes \{\boldsymbol{\Psi}_{i1}^{(1)t}D^{-1}C(D + J_iC)^{-1}\}\right]\boldsymbol{W}_i\\
&= \boldsymbol{\Lambda}^{(1)}\left[\mathbf{1}_{J_i}^T \otimes \{\boldsymbol{\Psi}_{i1}^{(1)t}(D + J_iC)^{-1}\}\right]\boldsymbol{W}_i\\
&= \boldsymbol{\Lambda}^{(1)}\textstyle\sum_{j=1}^{J_i}\{\boldsymbol{\Psi}_{i1}^{(1)t}(D + J_iC)^{-1}\}\boldsymbol{W}_{ij}\\
&= \boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{i1}^{(1)t}(J_i^{-1}D + C)^{-1}\bar{\boldsymbol{W}}_i,
\end{aligned}$$

where $\bar{\boldsymbol{W}}_i = \sum_{j=1}^{J_i}\boldsymbol{W}_{ij}/J_i$ is the mean vector of $\boldsymbol{W}_{ij}$'s. We used the fact that $\boldsymbol{\Psi}_{i1}^{(1)t}(J_i^{-1}D + C)^{-1}$ does not vary with j. Using a similar technique, $\boldsymbol{\Sigma}_i$ is

$$\begin{aligned}
&= \boldsymbol{\Lambda}^{(1)} - \boldsymbol{\Lambda}^{(1)}\{\mathbf{1}_{J_i}^T \otimes \boldsymbol{\Psi}_{i1}^{(1)t}\}\ [\boldsymbol{I}_{J_i} \otimes D^{-1} - \mathbf{1}_{J_i \times J_i} \otimes \{D^{-1}C(D + J_iC)^{-1}\}]\ \{\mathbf{1}_{J_i} \otimes \boldsymbol{\Psi}_{i1}^{(1)}\}\boldsymbol{\Lambda}^{(1)}\\
&= \boldsymbol{\Lambda}^{(1)} - \boldsymbol{\Lambda}^{(1)}\left[\mathbf{1}_{J_i}^T \otimes \{\boldsymbol{\Psi}_{i1}^{(1)t}(D + J_iC)^{-1}\}\right]\ \left\{\mathbf{1}_{J_i} \otimes \boldsymbol{\Psi}_{i1}^{(1)}\right\}\boldsymbol{\Lambda}^{(1)}\\
&= \boldsymbol{\Lambda}^{(1)} - J_i\boldsymbol{\Lambda}^{(1)}\boldsymbol{\Psi}_{i1}^{(1)t}(D + J_iC)^{-1}\boldsymbol{\Psi}_{i1}^{(1)}\boldsymbol{\Lambda}^{(1)}.
\end{aligned}$$

References

[1] B.P. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition.* Chapman & Hall/CRC, 2000.

[2] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C.M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman & Hall/CRC, New York, 2006.

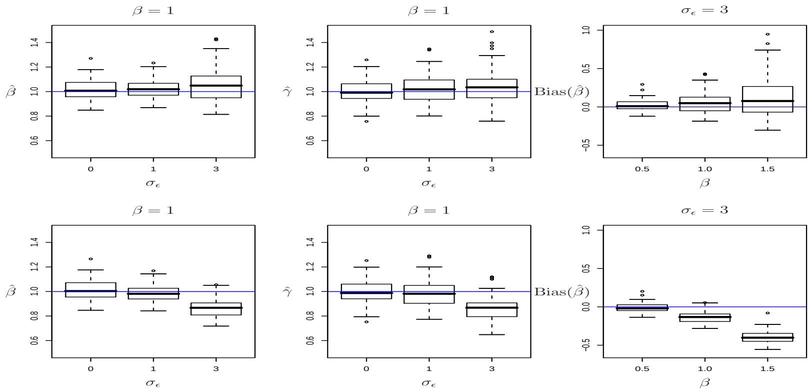

Figure 1: Joint Bayesian analysis (upper panel) versus two-stage analysis with BLUP (bottom panel): box plots of $\hat{\beta}$ and $\hat{\gamma}$ for different values of β and σ_ϵ.

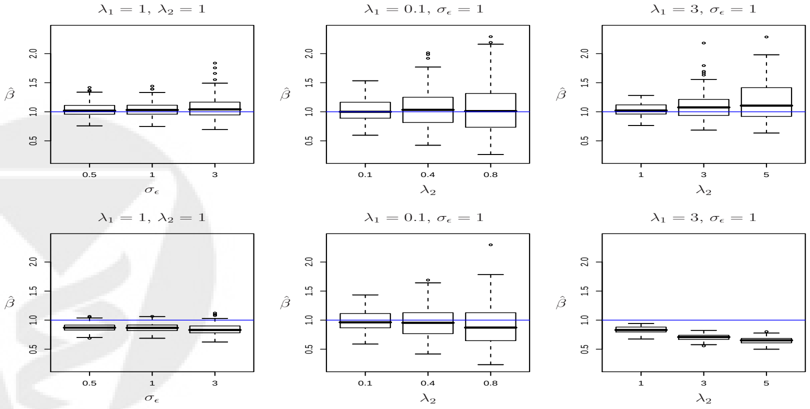

Figure 2: Joint Bayesian analysis (upper panel) versus two-stage analysis with BLUP (bottom panel): box plots of $\hat{\beta}$ for $\beta = 1$ and various values of σ_ϵ and λ's.

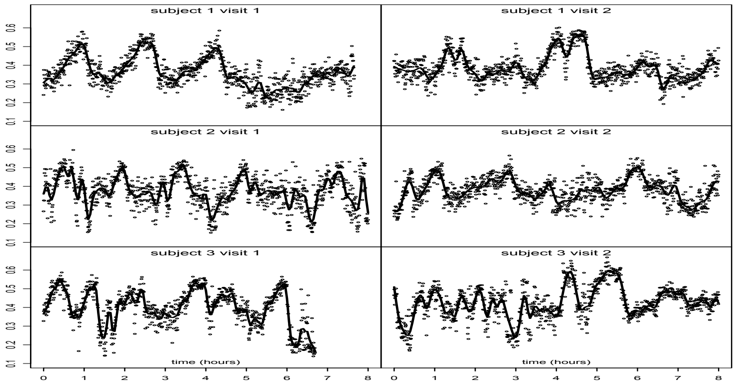

Figure 3: The EEG signal series for three subjects at both visits. The horizonal axis is time in hours, and the vertical axis is the percentage of delta power sleep in 30 seconds windows. Each subject was measured at both visit 1 and visit 2. The solid lines are smooth estimates of the mean functions.

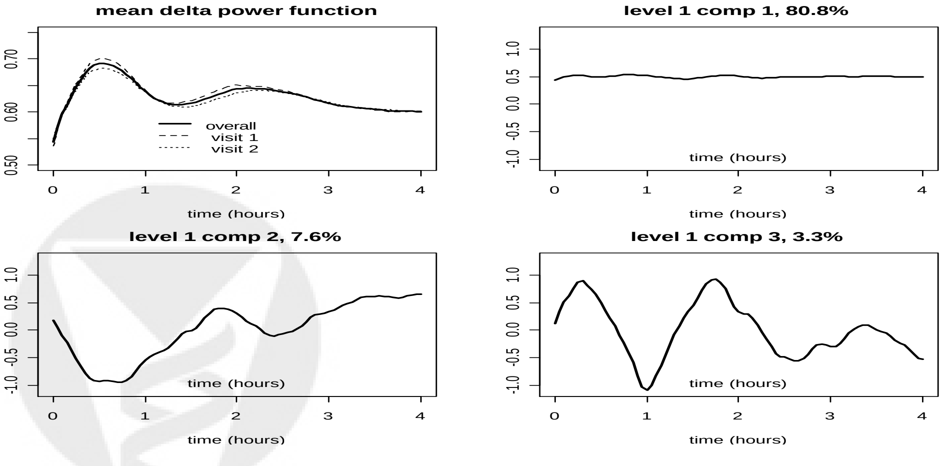

Figure 4: Characteristics of normalized sleep EEG δ-power. Top-left panel: overall mean (solid line), baseline mean (dashed line) and visit 2 mean (dotted line). Other three panels: First three eigenvalues of the subject specific deviations from the visit mean function.

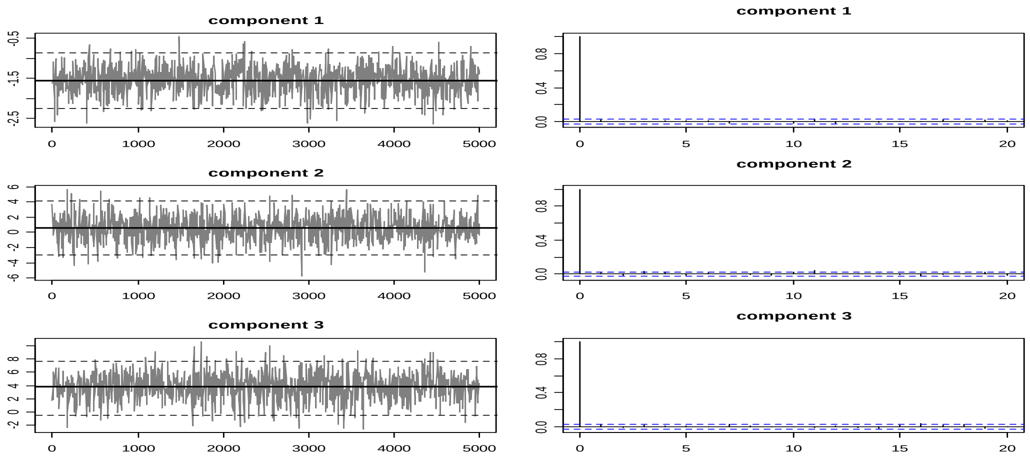

Figure 5: Markov chains (left panels) and autocorrelation functions (right panels) for regression coefficients of 3 principal component scores from 5000 iterations using Model 1 from Section 7.

[3] J.-M. Chiou, H.-G. Müller, and J.-L. Wang. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, Series B*, 65:405–423, 2003.

[4] P. Congdon. *Applied Bayesian Modelling.* Wiley, 2003.

[5] C.M. Crainiceanu, B. Caffo, and P. Naresh. Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *under review*, 2008.

[6] C.M. Crainiceanu, D. Ruppert, R.J. Carroll, J. Adarsh, and B. Goodner. Spatially adaptive Penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16(2), 2007.

[7] C. Di, C.M. Crainiceanu, B. Caffo, and P. Naresh. Multilevel functional principal component analysis. *under review.*

[8] J. Fan and J. T. Zhang. Functional linear models for functional data. *Journal of the Royal Statistical Society, Series B*, 39:254–261, 1998.

[9] J. Fan and J.-T. Zhang. Functional quasi-likelihood regression models with smooth random effects. *Journal of the Royal Statistical Society, Series B*, 62:303–322, 2000.

[10] A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[11] A. Gelman, J.B. Carlin, H.A. Stern, and D.B Rubin. *Bayesian Data Analysis, Second Edition.* Chapman & Hall/CRC, 2003.

[12] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[13] W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman & Hall/CRC, 1996.

[14] P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34:1493–1517, 2006.

[15] J. Indritz. *Methods in analysis.* Macmillan & Colier-Macmillan, 1963.

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

[17] G.M. James. Generalized Linear Models with Functional Predictors. *Journal of the Royal Statistical Society, Series B*, 64:411–432, 2002.

[18] K. Karhunen. *Über lineare Methoden in der Wahrscheinlichkeitsrechnung.* Suomalainen Tiedeakatemia, 1947.

[19] NM Laird and JH Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–74, 1982.

[20] X. Lin and R.J. Carroll. Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured Without/With Error. *Journal of the American Statistical Association*, 95(450), 2000.

[21] M. Loève. Functions aleatoire de second ordre. *Comptes Rendus Acad. Sci*, 220, 1945.

[22] H.-G. Müller. Functional modelling and classification of longitudinal data. *Scandivanian Journal of Statistics*, 32:223–240, 2005.

[23] H.-G. Müller and U. Stadtmüller. Generalized Functional Linear Models. *The Annals of Statististics*, 33(2):774–805, 2005.

[24] H.-G. Müller and Y. Zhang. Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. *Biometrics*, 61:1064–1075, 2005.

[25] R. Natarajan and R.E. Kass. Reference bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95:227–237, 2000.

[26] J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis.* Springer-Verlag, New York, 2005.

[27] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis.* Springer-Verlag, New York, 2006.

[28] J. Rice. Functional and longitudinal data analysis. *Statistica Sinica*, 14:631–647, 2004.

[29] D. Ruppert, R.J. Carroll, M.P. Wand, and MP Wand. *Semiparametric Regression.* Cambridge University Press, 2003.

[30] N. Wang, R.J. Carroll, and X. Lin. Efficient semiparametric marginal estimation for longitudinal/clustered data. *Journal of the American Statistical Association*, 100:147–157, 1998.

[31] F. Yao and T.C.M. Lee. Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society. Series B, statistical methodology*, 68:3–25, 2006.

[32] F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590, 2005.

[33] F. Yao, H.-G. Müller, and J.-L. Wang. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33:2873–2903, 2005.

[34] F. Yao, H.G. Muller, A.J. Clifford, S.R. Dueker, J. Follett, Y. Lin, B.A. Buchholz, and J.S. Vogel. Shrinkage Estimation for Functional Principal Component Scores with Application to the Population Kinetics of Plasma Folate. *Biometrics*, 59(3):676–685, 2003.

[35] L. Zhang, J. Samet, B. Caffo, and N.M. Punjabi. Cigarette smoking and nocturnal sleep architecture. *American Journal of Epidemiology*, 164(6):529, 2006.

[36] X. Zhao, J.S. Marrron, and M.T. Wells. The functional data analysis view of longitudinal data. *Statistica Sinica*, 14:789–808, 2004.