Supplementary materials for this article are available at http://pubs.amstat.org/toc/jasa/104/488.

Generalized Multilevel Functional Regression

Ciprian M. CRAINICEANU, Ana-Maria STAICU, and Chong-Zhi DI

We introduce Generalized Multilevel Functional Linear Models (GMFLMs), a novel statistical framework for regression models where exposure has a multilevel functional structure. We show that GMFLMs are, in fact, generalized multilevel mixed models. Thus, GMFLMs can be analyzed using the mixed effects inferential machinery and can be generalized within a well-researched statistical framework. We propose and compare two methods for inference: (1) a two-stage frequentist approach; and (2) a joint Bayesian analysis. Our methods are motivated by and applied to the Sleep Heart Health Study, the largest community cohort study of sleep. However, our methods are general and easy to apply to a wide spectrum of emerging biological and medical datasets. Supplemental materials for this article are available online.

KEY WORDS: Functional principal components; Sleep EEG; Smoothing.

1. INTRODUCTION

Recording and processing of functional data has become routine due to advancements in technology and computation. Many current studies contain observations of functional data on the same subject at multiple visits. For example, the Sleep Heart Health Study (SHHS) described in Section 7 contains, for each subject, quasi-continuous electroencephalogram (EEG) signals at two visits. In this paper we introduce a class of models and inferential methods for association studies between functional data observed at multiple levels/visits, such as sleep EEG or functional magnetic resonance imaging (fMRI), and continuous or discrete outcomes, such as systolic blood pressure (SBP) or Coronary Heart Disease (CHD). As most of these datasets are very large, feasibility of methods is a primary concern.

Functional regression is a generalization of regression to the case when outcomes or regressors or both are functions instead of scalars. Functional Regression Analysis is currently under intense methodological research (James 2002; Chiou, Müller, and Wang 2003; Müller and Stadtmüller 2005; Yao, Müller, and Wang 2005; Ramsay and Silverman 2006) and is a particular case of Functional Data Analysis (FDA) (James, Hastie, and Sugar 2000; Wang, Carroll, and Lin 2005; Hall, Müller, and Wang 2006; Yao and Lee 2006). Two comprehensive monographs of FDA with applications to curve and image analysis are Ramsay and Silverman (2005, 2006). There has been considerable recent effort to apply FDA to longitudinal data, for example, Fan and Zhang (2000); Rice (2004); Zhao, Marrron, and Wells (2004); see Müller (2005) for a thorough review. However, in all current FDA research, the term "longitudinal" represents single-level time series.

FDA was extended to multilevel functional data; see, for example, Guo (2002); Morris et al. (2003); Morris and Carroll (2006); Baladandayuthapani et al. (2008); Di et al. (2009); Staicu, Crainiceanu, and Carroll (2009). However, all these papers have focused on models for functional data and not on functional regression. The multilevel functional principal component analysis (MFPCA) approach in Di et al. (2009) uses functional principal component bases to reduce data dimensionality and accelerate the associated algorithms, which is especially useful in moderate and large datasets. Thus, MFPCA provides an excellent platform for methodological extensions to the multilevel regression case.

We introduce Generalized Multilevel Functional Linear Models (GMFLMs), a novel statistical framework for regression models where exposure has a multilevel functional structure. This framework extends MFPCA in several ways. First, GMFLMs are designed for studies of association between outcome and functional exposures, whereas MFPCA is designed to describe functional exposure only; this extension is needed to answer most common scientific questions related to longitudinal collection of functional/image data. Second, we show that GMFLMs are the functional analog of measurement error regression models; in this context MFPCA is the functional analog of the exposure measurement error models (Carroll et al. 2006). Third, we show that all regression models with functional predictors contain two mixed effects submodels: an outcome and an exposure model. Fourth, we propose and compare two methods for inference: (1) a two-stage frequentist approach; and (2) a joint Bayesian analysis. Using the analogy with measurement error models we provide insight into when using a two-stage method is a reasonable alternative to the joint analysis and when it is expected to fail. Our methods are an evolutionary development in a growth area of research. They build on and borrow strength from multiple methodological frameworks: functional regression, measurement error, and multilevel modeling. Given the range of applications and methodological flexibility of our methods, we anticipate that they will become one of the standard approaches in functional regression.

The paper is organized as follows. Section 2 introduces the functional multilevel regression framework. Section 3 describes estimation methods based on best linear prediction. Section 4 presents our approach to model selection. Section 5 discusses the specific challenges of a Bayesian analysis of the joint mixed effects model corresponding to functional regression. Section 6 provides simulations. Section 7 describes an application to sleep EEG data from the SHHS. Section 8 summarizes our conclusions.

Ciprian M. Crainiceanu is Associate Professor, Department of Biostatistics, Johns Hopkins University, Baltimore, MD 21205 (E-mail: *ccrainic@jhsph.edu*). Ana-Maria Staicu is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *staicu@stat.ncsu.edu*). Chong-Zhi Di is Assistant Professor, Biostatistics Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109 (E-mail: *cdi@jhsph.edu*). Crainiceanu's and Di's research was supported by award 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.

Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Theory and Methods
DOI: 10.1198/jasa.2009.tm08564

2. MULTILEVEL FUNCTIONAL REGRESSION MODELS

In this section we introduce the GMFLM framework and inferential methods.

2.1 Joint Mixed Effects Models

The observed data for the ith subject in a GMFLM is $[Y_i, \mathbf{Z}_i, \{W_{ij}(t_{ijm}), t_{ijm} \in [0,1]\}]$, where Y_i is the continuous or discrete outcome, $\mathbf{Z}_i$ 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, $i = 1, \ldots, I$. For presentation simplicity we only discuss the case of equally spaced t_{ijm}, but our methods can be applied with only minor changes to unequally/random spaced t_{ijm}; see Di et al. (2009) for more details.

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 $U_{ij}(t)$ is the residual subject/visit specific deviation from the subject specific mean. Note that multilevel functional models are a generalization of: (1) the classical measurement error models for replication studies when there is no t variable and $\eta_j(t) = 0$; (2) the standard functional models when $J_i = 1$ for all i, $\eta_j(\cdot) = 0$ and $U_{ij}(\cdot) = 0$; and (3) the two-way ANOVA models when $X_i(\cdot)$ and $U_{ij}(\cdot)$ do not depend on t. This is not important because "a more general model is better," but because it allows us to borrow and adapt methods from seemingly unrelated areas of statistics. We contend that this synthesis is both necessary and timely to address the increasing challenges raised by ever larger and more complex datasets.

To ensure identifiability we assume that $X_i(t)$, $U_{ij}(t)$, and $\epsilon_{ij}(t)$ are uncorrelated, that $\sum_j \eta_j(t) = 0$ 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 $\overline{W}_{\cdot\cdot}(t)$ and $\overline{W}_{\cdot j}(t) - \overline{W}_{\cdot\cdot}$, respectively. Here $\overline{W}_{\cdot\cdot}(t)$ is the average over all subjects, i, and visits, j, of $W_{ij}(t)$ and $\overline{W}_{\cdot j}(t)$ 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 $W_{ij}(t) = X_i(t) + U_{ij}(t) + \epsilon_{ij}(t)$. Note that consistent estimators of $\widetilde{W}_{ij}(t) = X_i(t) + U_{ij}(t)$ can be obtained by smoothing $\{t, W_{ij}(t)\}$. Moreover, consistent estimators for $X_i(t)$ and $U_{ij}(t)$ can be constructed as estimators of $\sum_{k=1}^{J} \widetilde{W}_{ik}(t)/J$ and $\widetilde{W}_{ij}(t) - \sum_{k=1}^{J} \widetilde{W}_{ik}(t)/J$, respectively.

We assume that the distribution of the outcome, Y_i, is in the exponential family with linear predictor ϑ_i and dispersion parameter α, denoted here by $\mathrm{EF}(\vartheta_i, \alpha)$. The linear predictor is assumed to have the following form $\vartheta_i = \int_0^1 X_i(t)\beta(t)\,dt + \mathbf{Z}_i^t\boldsymbol{\gamma}$, where $\mathbf{X}_i(t)$ is the subject-specific deviation from the visit-specific mean, $\beta(\cdot) \in L_2[0,1]$ is a functional parameter and the main target of inference, $\mathbf{Z}_i^t$ is a vector of covariates and $\boldsymbol{\gamma}$ are fixed effects parameters. If $\{\psi_k^{(1)}(t)\}$ and $\{\psi_l^{(2)}(t)\}$ are two orthonormal bases in $L_2[0,1]$ then $X_i(\cdot)$, $U_{ij}(\cdot)$ and $\beta(\cdot)$ have unique representations

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

This form of the model is impractical because it involves three infinite sums. Instead, we will approximate model (1) with a series of models where the number of predictors is truncated at $K = K_{I,J}$ and $L = L_{I,J}$ and the dimensions K and L increase asymptotically with the total number of subjects, I, and visits per subject, J. A good heuristic motivation for this truncation strategy can be found, for example, in Müller and Stadtmüller (2005). In Section 4 we provide a theoretical and practical discussion of alternatives for estimating K and L. For fixed K and L the multilevel outcome model becomes

$$\begin{cases} Y_i \sim \mathrm{EF}(\vartheta_i^K, \alpha); \\ \vartheta_i^K = \sum_{k=1}^{K} \xi_{ik}\beta_k + \mathbf{Z}_i^t\boldsymbol{\gamma}. \end{cases} \tag{2}$$

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 (2).

We use MFPCA (Di et al. 2009) to obtain the parsimonious bases that capture most of the functional variability of the space spanned by $X_i(t)$ and $U_{ij}(t)$, respectively. MFPCA is based on the spectral decomposition of the within-visit and between-visit functional variability covariance operators. We summarize here the main components of this methodology. Denote by $K_T^W(s,t) = \mathrm{cov}\{W_{ij}(s), W_{ij}(t)\}$ and $K_B^W(s,t) = \mathrm{cov}\{W_{ij}(s), W_{ik}(t)\}$ for $j \neq k$ the total and between covariance operator corresponding to the observed process, $W_{ij}(\cdot)$, respectively. Denote by $K^X(t,s) = \mathrm{cov}\{X_i(t), X_i(s)\}$ the covariance operator of the $X_i(\cdot)$ process and by $K_T^U(t,s) = \mathrm{cov}\{U_{ij}(s), U_{ij}(t)\}$ the total covariance operator of the $U_{ij}(\cdot)$ process. By definition, $K_B^U(s,t) = \mathrm{cov}\{U_{ij}(s), U_{ik}(t)\} = 0$ 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 $\widehat{K}_T^U(t,t)$ one predicts $K_T^U(t,t)$ using a bivariate thin-plate spline smoother of $\widehat{K}_T^U(s,t)$ for $s \neq t$. This method was proposed for single-level FPCA (Yao and Lee 2006) and shown to work well in the MFPCA context (Di et al. 2009).

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 Indritz 1963, chapter 4) provides the following convenient spectral decompositions: $K^X(t,s) = \sum_{k=1}^{\infty} \lambda_k^{(1)}\psi_k^{(1)}(t) \times \psi_k^{(1)}(s)$, where $\lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \cdots$ 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, $K_T^U(t,s) = \sum_{l=1}^{\infty} \lambda_l^{(2)} \times$

$\psi_l^{(2)}(t)\psi_l^{(2)}(s)$, where $\lambda_1^{(2)} \geq \lambda_2^{(2)} \geq \cdots$ 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 (Loève 1945; Karhunen 1947) 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 $E(\xi_{ik}) = E(\zeta_{ijl}) = 0$, $\mathrm{Var}(\xi_{ik}) = \lambda_k^{(1)}$, $\mathrm{Var}(\zeta_{ijl}) = \lambda_l^{(2)}$. The zero-correlation assumption between the $X_i(\cdot)$ and $U_{ij}(\cdot)$ processes is ensured by the assumption that $\mathrm{cov}(\xi_i, \zeta_{ijl}) = 0$. These properties hold for every i, j, k, and l.

Conditional on the eigenfunctions and truncation lags K and L, 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

$$\begin{cases} W_{ij}(t) = \displaystyle\sum_{k=1}^{K} \xi_{ik}\psi_k^{(1)}(t) + \sum_{l=1}^{L} \zeta_{ijl}\psi_l^{(2)}(t) + \epsilon_{ij}(t); \\ \xi_{ik} \sim \mathrm{N}\{0, \lambda_k^{(1)}\}; \zeta_{ijl} \sim \mathrm{N}\{0, \lambda_l^{(2)}\}; \epsilon_{ij}(t) \sim \mathrm{N}(0, \sigma_\epsilon^2). \end{cases} \tag{3}$$

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 (2) and the exposure model (3). 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 Markov chain Monte Carlo (MCMC) simulations as described in Section 5. An alternative would be to use a two-stage analysis by first predicting the scores from model (3) and then plug-in these estimates into model (2).

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 $\mathbf{W}_i = \{W_{i1}(t_{i11}), \ldots, W_{i1}(t_{i1M_{i1}}), \ldots, W_{iJ_i}(t_{iJ_iM_{iJ_i}})\}^t$. With a slight abuse of notation $[Y_i|\mathbf{W}_i, \mathbf{Z}_i] = \int [Y_i, \boldsymbol{\xi}_i|\mathbf{W}_i, \mathbf{Z}_i]\,d\boldsymbol{\xi}_i$, where $[\cdot|\cdot]$ denotes the probability density function of the conditional distribution. The assumptions in models (2) and (3) imply that $[Y_i, \boldsymbol{\xi}_i|\mathbf{W}_i, \mathbf{Z}_i] = [Y_i|\boldsymbol{\xi}_i, \mathbf{Z}_i] \times [\boldsymbol{\xi}_i|\mathbf{W}_i]$, which, in turn, implies that

$$[Y_i|\mathbf{W}_i, \mathbf{Z}_i] = \int [Y_i|\boldsymbol{\xi}_i, \mathbf{Z}_i][\boldsymbol{\xi}_i|\mathbf{W}_i]\,d\boldsymbol{\xi}_i. \tag{4}$$

Under normality assumptions it is easy to prove that $[\boldsymbol{\xi}_i|\mathbf{W}_i] = \mathrm{N}\{m(\mathbf{W}_i), \boldsymbol{\Sigma}_i\}$, where $m(\mathbf{W}_i)$ and $\boldsymbol{\Sigma}_i$ are the mean and covariance matrix of the conditional distribution of $\boldsymbol{\xi}_i$ given the observed functional data and model (3). In Section 3 we provide the derivation of $m(\mathbf{W}_i)$ and $\boldsymbol{\Sigma}_i$ and additional insight into their effect on inference.

For most nonlinear models the induced model for observed data (4) 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 (3); and (2) estimate the parameters of the outcome model (2) by replacing $\boldsymbol{\xi}_i$ with $\widehat{\boldsymbol{\xi}}_i$. It is reasonable to use the best linear unbiased predictor (BLUP) of $\boldsymbol{\xi}_i$, $\widehat{\boldsymbol{\xi}}_i = m(\mathbf{W}_i)$, but other predictors could also be used. For example, for the single-level functional model Müller and Stadtmüller (2005) used $\widehat{\xi}_{ik} = \int_0^1 W_i(t)\psi_k(t)\,dt$, which are unbiased predictors of ξ_{ik}. Such estimators have even higher variance than $\boldsymbol{\Sigma}_i$ because they do not borrow strength across subjects. This may lead to estimation bias and misspecified variability. The problem is especially serious in multilevel functional models as we discuss below.

Consider, for example, the outcome model $Y_i|\boldsymbol{\xi}_i, \mathbf{Z}_i \sim \mathrm{Bernoulli}(p_i)$, where $\Phi^{-1}(p_i) = \boldsymbol{\xi}_i^t\boldsymbol{\beta} + \mathbf{Z}_i^t\gamma$, 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 $Y_i|\mathbf{W}_i, \mathbf{Z}_i \sim \mathrm{Bernoulli}(q_i)$, where

$$\Phi^{-1}(q_i) = \{m^t(\mathbf{W}_i)\boldsymbol{\beta} + \mathbf{Z}_i^t\gamma\}/(1 + \boldsymbol{\beta}^t\boldsymbol{\Sigma}_i\boldsymbol{\beta})^{1/2}. \tag{5}$$

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

There are important potential differences between joint and two-stage analyses in a multilevel functional regression context. Indeed, the term $\sum_{l=1}^{L} \zeta_{ijl}\psi_l^{(2)}(t)$ in Equation (3) 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 when $\sigma_\epsilon^2 = 0$. Thus, the effects of variability on bias in a two-stage procedure can be severe, 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.

3. POSTERIOR DISTRIBUTION OF SUBJECT–SPECIFIC SCORES

We now turn our attention to calculating the posterior distribution of subject-specific scores for the MFPCA model (3). 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 $\mathbf{W}_{ij}$, $\boldsymbol{\psi}_{ij,k}^{(1)} = \{\psi_k^{(1)}(t_{ij1}), \ldots, \psi_k^{(1)}(t_{ijM_{ij}})\}^t$ 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 $\boldsymbol{\psi}_{ij,l}^{(2)}$, $\boldsymbol{\psi}_{il}^{(2)}$, $\boldsymbol{\Psi}_{ij}^{(2)}$, and $\boldsymbol{\Psi}_i^{(2)}$. Finally, let $\boldsymbol{\Lambda}^{(1)} = \mathrm{diag}\{\lambda_1^{(1)}, \ldots, \lambda_K^{(1)}\}$

and $\mathbf{\Lambda}^{(2)} = \text{diag}\{\lambda_1^{(2)}, \ldots, \lambda_L^{(2)}\}$ be the $K \times K$ and $L \times L$ dimensional diagonal matrices of level 1 and level 2 eigenvalues, respectively.

If $\mathbf{\Sigma}_{\mathbf{W}_i}$ denotes the covariance matrix of $\mathbf{W}_i$ then its (j, j')th block matrix is equal to $\mathbf{B}_{i,jj'}$ where $\mathbf{B}_{i,jj'} = \mathbf{B}_{i,j'j}^t = \mathbf{\Psi}_{ij}^{(1)} \mathbf{\Lambda}^{(1)} \times \mathbf{\Psi}_{ij'}^{(1)t}$ if $j \neq j'$ and $\mathbf{B}_{i,jj} = \sigma_\epsilon^2 \mathbf{I}_{M_{ij}} + \mathbf{\Psi}_{ij}^{(2)} \mathbf{\Lambda}^{(2)} \mathbf{\Psi}_{ij}^{(2)t} + \mathbf{\Psi}_{ij}^{(1)} \mathbf{\Lambda}^{(1)} \times \mathbf{\Psi}_{ij}^{(1)t}$ for $1 \leq j, j' \leq J_i$. Moreover, under normality assumptions $[\boldsymbol{\xi}_i | \mathbf{W}_i] = \text{N}\{m(\mathbf{W}_i), \mathbf{\Sigma}_i\}$, where $m(\mathbf{W}_i) = \mathbf{\Lambda}^{(1)} \mathbf{\Psi}_i^{(1)t} \Sigma_{\mathbf{W}_i}^{-1} \mathbf{W}_i$ and $\mathbf{\Sigma}_i = \mathbf{\Lambda}^{(1)} - \mathbf{\Lambda}^{(1)} \mathbf{\Psi}_i^{(1)t} \Sigma_{\mathbf{W}_i}^{-1} \mathbf{\Psi}_i^{(1)} \mathbf{\Lambda}^{(1)}$. The following results provide simplified expressions for $\Sigma_{\mathbf{W}_i}$, $m(\mathbf{W}_i)$, and $\mathbf{\Sigma}_i$ that greatly reduce computational burden of algorithms.

Theorem 1. Consider the exposure model (3) with a fixed number of observations per visit, that is, visit, that is, $t_{ijm} = t_{im}$ for all $j = 1, \ldots, J_i$. Denote by $\mathbf{K}^X = \mathbf{\Psi}_{i1}^{(1)} \mathbf{\Lambda}^{(1)} \mathbf{\Psi}_{i1}^{(1)t}$, by $\mathbf{K}_T^U = \mathbf{\Psi}_{i1}^{(2)} \mathbf{\Lambda}^{(2)} \mathbf{\Psi}_{i1}^{(2)t}$, 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 $\mathbf{\Sigma}_{\mathbf{W}_i} = \mathbf{1}_{J_i \times J_i} \otimes \mathbf{K}^X + \mathbf{I}_{J_i} \otimes (\sigma_\epsilon^2 \mathbf{I}_{M_i} + \mathbf{K}_T^U)$ and $\Sigma_{\mathbf{W}_i}^{-1} = \mathbf{I}_{J_i} \otimes (\sigma_\epsilon^2 \mathbf{I}_{M_i} + \mathbf{K}_T^U)^{-1} - \mathbf{1}_{J_i \times J_i} \otimes \{(\sigma_\epsilon^2 \mathbf{I}_{M_i} + \mathbf{K}_T^U)^{-1} \mathbf{K}^X (J_i \mathbf{K}^X + \sigma_\epsilon^2 \mathbf{I}_{M_i} + \mathbf{K}_T^U)^{-1}\}$.

Theorem 2. Assume the balanced design considered in Theorem 1 and denote by $\overline{\mathbf{W}}_i = \sum_{j=1}^{J_i} \mathbf{W}_{ij} / J_i$. Then $m(\mathbf{W}_i) = \mathbf{\Lambda}^{(1)} \mathbf{\Psi}_{i1}^{(1)t} \{\mathbf{K}^X + \frac{1}{J_i}(\sigma_\epsilon^2 \mathbf{I}_{M_i} + \mathbf{K}_T^U)\}^{-1} \overline{\mathbf{W}}_i$ and $\Sigma_i = \mathbf{\Lambda}^{(1)} - \mathbf{\Lambda}^{(1)} \times \mathbf{\Psi}_{i1}^{(1)t} \{\mathbf{K}^X + \frac{1}{J_i}(\sigma_\epsilon^2 \mathbf{I}_{M_i} + \mathbf{K}_T^U)\}^{-1} \mathbf{\Psi}_{i1}^{(1)} \mathbf{\Lambda}^{(1)}$.

Proofs can be found in the accompanying web supplement. Theorem 2 provides a particularly simple description of the conditional distribution $\boldsymbol{\xi}_i | \mathbf{W}_i$. Moreover, it shows that, conditional on the smoothing matrices $\mathbf{\Lambda}^{(1)}$ and $\mathbf{\Lambda}^{(2)}$, the conditional distribution $\boldsymbol{\xi}_i | \mathbf{W}_i$ is the same as the conditional distribution $\boldsymbol{\xi}_i | \overline{\mathbf{W}}_i$. 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 $\boldsymbol{\xi}_i | \mathbf{W}_i$, $\mathbf{\Sigma}_i$. 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

$$\begin{cases} W_{ij}(t) = \xi_i + \zeta_{ij} + \epsilon_{ij}(t); \\ \xi_i \sim \text{N}(0, \lambda_1); \zeta_{ij} \sim \text{N}(0, \lambda_2); \epsilon_{ij}(t) \sim \text{N}(0, \sigma_\epsilon^2), \end{cases} \tag{6}$$

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 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, $\Sigma_i \leq \lambda_1$ indicating that Σ_i is small when the variability at first level, λ_1, is small. In this situation one could expect the two-stage procedure to work well. Second, 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. Third, even when functions are perfectly observed, that is, $\sigma_\epsilon^2 = 0$, the variance Σ_i is not zero. Fourth, in many applications $\sigma_\epsilon^2 / (M_i J_i)$ is negligible because the total number of observations for subject i, $M_i J_i$, is large. For example, in the SHHS, $M_i J_i \approx 1600$.

4. MODEL UNCERTAINTY

Our framework is faced with two distinct types of model uncertainty related to: (1) the choice of K and L, the dimensions of the two functional spaces in the exposure model (3); and (2) estimating $\beta(t)$, the functional effect parameter, conditional on K and L, in the outcome model (2).

To address the first problem we focus on estimating K, as estimating L is similar. Note that, as K increases, the models described in (3) form a nested sequence of mixed effects models. Moreover, testing for the dimension of the functional space being equal to K versus $K + 1$ is equivalent to testing $H_{0,K} : \lambda_{K+1}^{(1)} = 0$ versus $H_{A,K} : \lambda_{K+1}^{(1)} > 0$, which is testing for the null hypothesis that a particular variance component is *equal to zero*. This connection provides a paradigm shift for estimating the dimension of the functional space or, more generally, the number of nonzero eigenvalues in PCA. Current methods are based on random matrix theory and require that eigenvalues be *bounded away from zero*; see, for example, Bilodeau and Brenner (1999); Hall and Hosseini-Nasab (2006). This is not the correct approach when the null hypothesis is that the eigenvalue is zero.

In this context Staicu, Crainiceanu, and Carroll (2009) proposed a sequence of Restricted Likelihood Ratio Tests (RLRTs) for zero variance components (Stram and Lee 1994; Crainiceanu and Ruppert 2004; Crainiceanu et al. 2005) to estimate K. Müller and Stadtmüller (2005) proposed to use either the Akaike's Information Criterion (AIC) (Akaike 1973) or the Bayesian Information Criterion (BIC) (Schwarz 1978). Moreover, they found these criteria to be more stable and less computationally intensive than methods based on cross-validation (Rice and Silverman 1991) or relative difference between the Pearson criterion and deviance (Chiou and Müller 1998). Staicu, Crainiceanu, and Carroll (2009) show that both AIC and BIC are particular cases of sequential RLRT with nonstandard α levels. They also explain that AIC performs well because its associated α level is 0.079, which is different from the standard $\alpha = 0.05$, but might be reasonable in many applications. In contrast, they recommend against using the BIC in very large datasets, such as in our application, because the corresponding α level becomes extremely small.

In practice we actually prefer an even simpler method for estimating the number of components based on the estimated explained variance. More precisely, let P_1 and P_2 be two thresholds and define $N_1 = \min\{k : \rho_k^{(1)} \geq P_1, \lambda_k < P_2\}$, where $\rho_k^{(1)} = (\lambda_1^{(1)} + \cdots + \lambda_k^{(1)}) / (\lambda_1^{(1)} + \cdots + \lambda_T^{(1)})$. For the cumulative explained variance threshold we used $P_1 = 0.9$ and for the individual explained variance we used $P_2 = 1/T$, where T is the number of grid points. We used a similar method for choosing the number of components at level 2. These choices were

slightly conservative, but worked well in our simulations and application. However, the two thresholds should be carefully tuned in any other particular application using simulations.

To address the second problem we note that it can be reduced to a standard model selection problem. Forward, backward, single-variable or all subset selection can be used to identify statistically significant predictors in the outcome model (2). Typical pitfalls reported for these methods are avoided because predictors are mutually orthogonal by construction. In practice, we prefer to do a backward selection combined with sensitivity analysis around the chosen model. More precisely, we obtain an optimal model and the two next best models. For all these models we provide the functional estimates and the log-likelihood differences.

A powerful alternative to estimating $\beta(t)$ was proposed in a series of papers by Reiss and Ogden (2007, 2009) for the single-level functional regression case. In short, they project the original (un-smooth) matrix of functional predictors onto a B-spline basis and use the P-spline basis penalty to induce shrinkage directly on the functional parameter. Another alternative is to adapt the forward selection method using pseudo-variables (Luo, Stefanski, and Boos 2006; Wu, Boos, and Stefanski 2007), which could work especially well because the estimated eigenvalues are sorted. Both methods could easily be used in our framework. However, they would need to be adapted to a joint analysis context to overcome the bias problem induced by the two-stage analysis described in Section 2.

5. BAYESIAN INFERENCE

Because of the potential problems associated with two-stage procedures, we propose to use joint modeling. Bayesian inference using MCMC simulations of the posterior distribution provides a reasonable, robust, and well-tested computational approach for this type of problems. 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 2.1 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 of Gilks, Richardson, and Spiegelhalter (1996); Carlin and Louis (2000); Congdon (2003); Gelman et al. (2003), and the citations therein for a good overview.

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 $\text{logit}(p_i) = \boldsymbol{\xi}_i^t\boldsymbol{\beta} + \mathbf{Z}_i^t\boldsymbol{\gamma}$, and functional exposure model (3). The parameters of the model are $\boldsymbol{\Omega} = \{(\boldsymbol{\xi}_i : i = 1, \ldots, I), (\boldsymbol{\zeta}_{ij} : i = 1, \ldots, I; j = 1, \ldots, J_i), \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\Lambda}, \sigma_\epsilon^2\}$, where $\boldsymbol{\xi}_i$ was defined in Section 2.2 and $\boldsymbol{\zeta}_{ij} = (\zeta_{ij1}, \ldots, \zeta_{ijL})^T$. While $\epsilon_i(t_{ijm})$ are also unknown, we do not incorporate them in the set of parameters because they are automatically updated by $\epsilon_i(t_{ijm}) = W_{ij}(t_{ijm}) - \sum_{k=1}^K \xi_{ik}\psi_k^{(1)}(t_{ijm}) - \sum_{l=1}^L \zeta_{ijl}\psi_l^{(2)}(t_{ijm})$.

The priors for $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_{ij}$ were already defined and it is standard to assume that the fixed effects parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, are apriori independent, with $\boldsymbol{\beta} \sim \text{N}(0, \sigma_\beta^2\mathbf{I}_K)$ and $\boldsymbol{\gamma} \sim \text{N}(0, \sigma_\gamma^2\mathbf{I}_P)$ 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 prior might be inconsistent with the true value of the parameter. In this situations we recommend rescaling $W_{ij}(t_{ijm})$ and normalizing, or rescaling, the Z covariates.

While standard choices of priors for fixed effects parameters exist and are typically noncontroversial, 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, Crainiceanu et al. (2007) and Gelman (2006). 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, Gelman (2006) and Natarajan and Kass (2000), 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 $\lambda_k^{(1)} \sim \text{IG}(A_k^{(1)}, B_k^{(1)}), k = 1, \ldots, K$, $\lambda_l^{(2)} \sim \text{IG}(A_l^{(2)}, B_l^{(2)})$, $l = 1, \ldots, L$, 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 then discuss choices of noninformative inverse gamma parameters. Here we treat $\lambda_k^{(1)}$ and $\lambda_l^{(2)}$ 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 $\lambda_k^{(1)}$ and $\lambda_l^{(2)}$ by diagonalizing the functional covariance operators as described in Section 2.1. These estimators are then fixed in the joint model. In the following we present the inferential procedure for the case when $\lambda_k^{(1)}$ and $\lambda_l^{(2)}$ are estimated with obvious simplifications for the EB procedure where they would be fixed.

We use Gibbs sampling (Geman and Geman 1984) to simulate $[\boldsymbol{\Omega}|\mathbf{D}]$, where $\mathbf{D}$ denotes the observed data. A particularly convenient partition of the parameter space and the associated full conditional distributions are described below:

$$[\boldsymbol{\beta}, \boldsymbol{\gamma}|\text{others}] \propto \exp\left[\sum_{i=1}^n Y_i(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \mathbf{Z}_i^t\boldsymbol{\gamma}) - \sum_{i=1}^n \log\{1 + \exp(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \mathbf{Z}_i^t\boldsymbol{\gamma})\}\right] \times \exp(-0.5\boldsymbol{\beta}^t\boldsymbol{\beta}/\sigma_\beta^2 - 0.5\boldsymbol{\gamma}^t\boldsymbol{\gamma}/\sigma_\gamma^2);$$

$$[\boldsymbol{\xi}_i|\text{others}] \propto \exp\big[Y_i(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \mathbf{Z}_i^t\boldsymbol{\gamma}) - \log\{1 + \exp(\boldsymbol{\xi}_i^t\boldsymbol{\beta} + \mathbf{Z}_i^t\boldsymbol{\gamma})\}\big] \times \exp\left[-0.5\sum_{j=1}^{J_i} \big\|\mathbf{W}_{ij} - \boldsymbol{\Psi}_{ij}^{(1)}\boldsymbol{\xi}_i - \boldsymbol{\Psi}_{ij}^{(2)}\boldsymbol{\zeta}_{ij}\big\|^2/\sigma_\epsilon^2 - 0.5\boldsymbol{\xi}_i\{\boldsymbol{\Lambda}^{(1)}\}^{-1}\boldsymbol{\xi}_i\right];$$

$$[\boldsymbol{\zeta}_{ij}|\text{others}] = \mathrm{N}\big[\mathbf{A}_{ij}^{-1}\big\{\mathbf{W}_{ij} - \boldsymbol{\Psi}_{ij}^{(1)}\boldsymbol{\xi}_i\big\}, \mathbf{A}_{ij}^{-1}\big];$$

$$\big[\lambda_k^{(1)}|\text{others}\big] = \mathrm{IG}\left\{I/2 + A_k^{(1)}, \sum_{i=1}^{n}\xi_{ik}^2/2 + B_k^{(1)}\right\};$$

$$\big[\lambda_l^{(2)}|\text{others}\big] = \mathrm{IG}\left\{\sum_{i=1}^{I}J_i/2 + A_k^{(2)}, \sum_{i=1}^{I}\sum_{i=1}^{J_i}\zeta_{ijl}^2/2 + B_k^{(2)}\right\};$$

$$[\sigma_\epsilon^2|\text{others}] = \mathrm{IG}\left\{\sum_{i=1}^{I}J_i/2 + A_\epsilon, \sum_{i=1}^{n}\big\|\mathbf{W}_{ij} - \boldsymbol{\Psi}_{ij}^{(1)}\boldsymbol{\xi}_i - \boldsymbol{\Psi}_{ij}^{(2)}\boldsymbol{\zeta}_{ij}\big\|^2/2 + B_\epsilon\right\},$$

where $\mathbf{A}_{ij} = \{\boldsymbol{\Psi}_{ij}^{(2)}\}^T\boldsymbol{\Psi}_{ij}^{(2)} + \{\boldsymbol{\Lambda}^{(2)}\}^{-1}$. The first two full conditionals do not have an explicit form, but can be sampled using 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 four 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. The first parameter of the full conditional for $\lambda_k^{(1)}$ is $I/2 + A_k^{(1)}$, where I is the number of subjects and it is safe to choose $A_k^{(1)} \le 0.01$. The second parameter is $\sum_{i=1}^{n}\xi_{ik}^2/2 + B_k^{(1)}$, where $\sum_{i=1}^{n}\xi_{ik}^2$ is an estimator of $n\lambda_k^{(1)}$ and it is safe to choose $B_k^{(1)} \le 0.01\lambda_k^{(1)}$. This is especially relevant for those variance components or, equivalently, eigenvalues of the covariance operator, that are small, but estimable. A similar discussion holds for $\lambda_l^{(2)}$. For σ_ϵ^2 we recommend to choose $A_\epsilon \le 0.01$ and $B_\epsilon \le 0.01\sigma_\epsilon^2$. Note that method of moments estimators for $\lambda_k^{(1)}$, $\lambda_l^{(2)}$, and σ_ϵ^2 are available and reasonable choices of $B_k^{(1)}$, $B_l^{(2)}$, and B_ϵ are easy to propose. These rules of thumb are useful in practice, but 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.

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 and multilevel.

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 $\xi_i \sim \mathrm{N}(0, \lambda_1)$ 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 rewritten 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 analysis via MCMC posterior simulations and a two-stage approach using either BLUP or numerical integration (Müller and Stadtmüller 2005). We simulated $N = 100$ datasets 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 $\epsilon_i(t_m) \sim \mathrm{N}(0, \sigma_\epsilon^2)$ 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 1 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 2.2 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 analysis 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.

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 $\epsilon_{ij}(t) \sim \mathrm{N}(0, \sigma_\epsilon^2)$ and $U_{ij}(t) = \zeta_{ij}\psi_2(t)$ with $\zeta_{ij} \sim \mathrm{N}(0, \lambda_2)$, $\psi_2(t) \equiv 1$. 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,

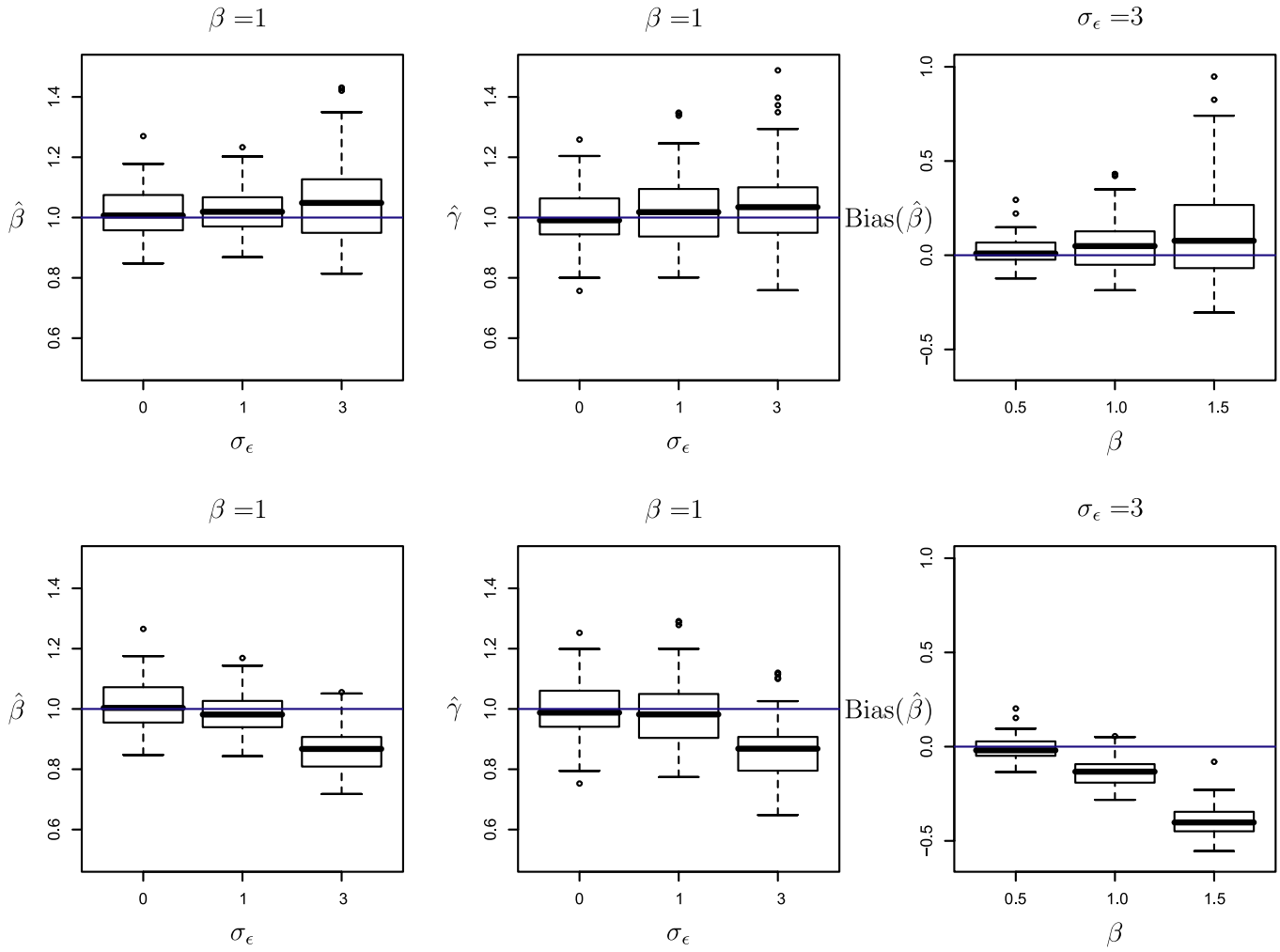

Figure 1. Joint Bayesian analysis (upper panel) versus two-stage analysis with BLUP (bottom panel): box plots of $\hat{\beta}$ and $\hat{\gamma}$ for different values of β and σ_ϵ.

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, 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 pan-

Table 1. 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 $\sigma_\epsilon = 3$. The Monte Carlo standard error for the Bayesian analysis was small compared to the RMSE; for $\beta = 0.5$ it ranged between $(0.002, 0.005)$ for β and $(0.002, 0.004)$ for γ; for $\beta = 1.5$ it ranged between $(0.009, 0.043)$ for β and $(0.005, 0.024)$ for γ

		$\hat{\beta}$			$\hat{\gamma}$		
Method	β	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

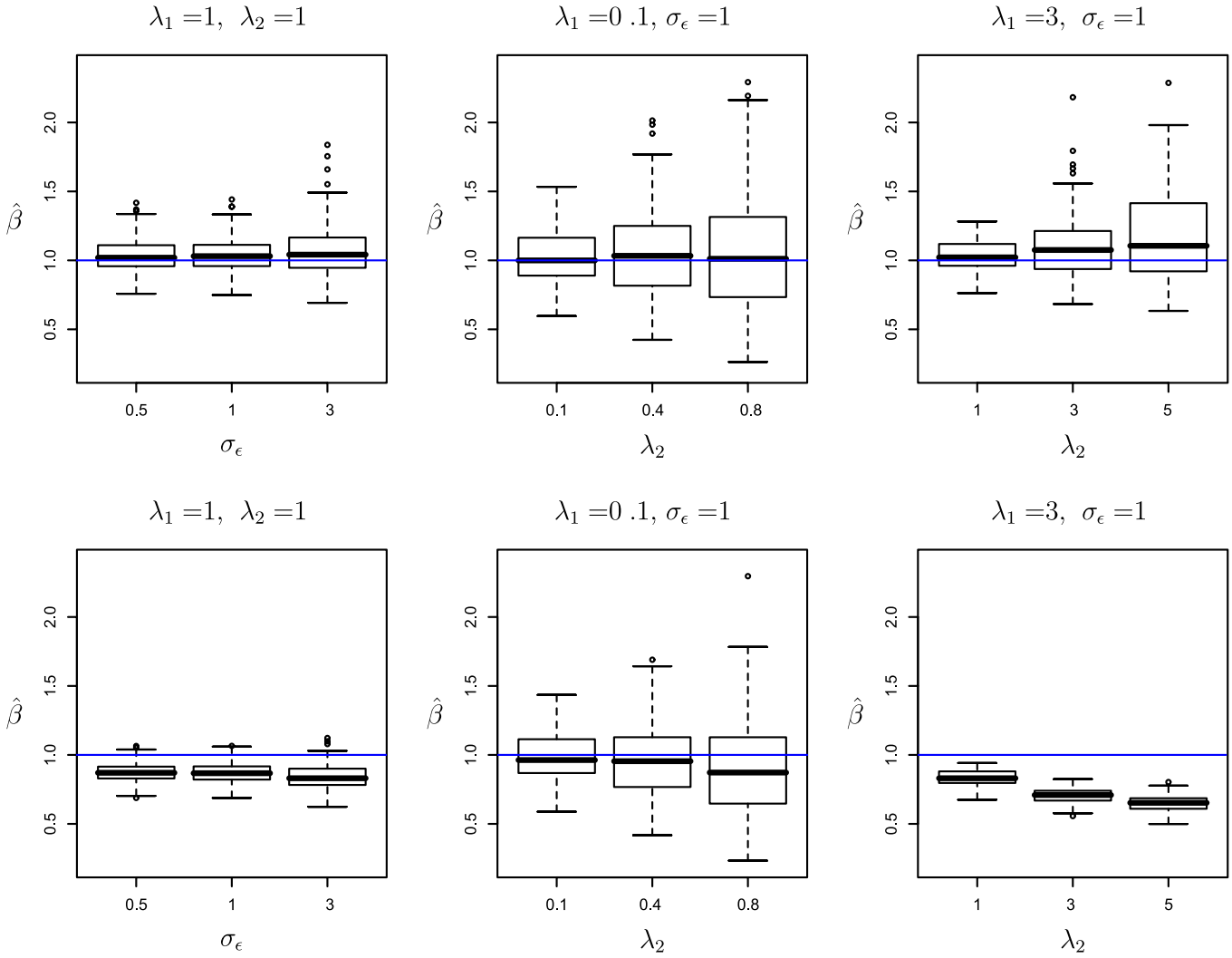

Figure 2. Joint Bayesian analysis (upper panel) versus two-stage analysis with BLUP (bottom panel): box plots of $\hat{\beta}$ for $\beta = 1$ and various values of σ_ϵ and λ's.

els 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. We conclude that bias is small when the between-subject variability, λ_1, is small even when the within subject variability, λ_2, is much larger than λ_1. If 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; results are not reported here, but 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 3201 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 dataset and of this functional regression problem can be found in Crainiceanu et al. (2009) and Di et al. (2009). 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 $F_{x,k} = \sum_{n=0}^{N-1} x_n e^{-2\pi i n k/N}$, $k = 0, \ldots, N-1$, 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 [0.8–4.0 Hz]; (2) θ [4.1–8.0 Hz]; (3) α [8.1–13.0 Hz]; (4) β [13.1–20.0 Hz]. These bands are standard representations of low (δ) to high (β) frequency neu-

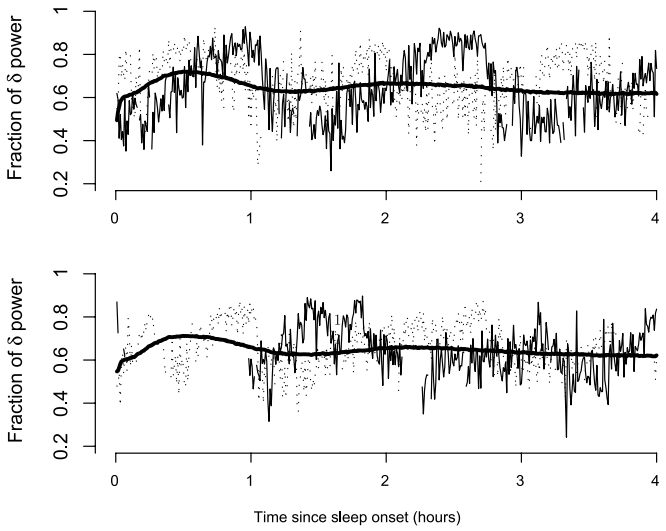

Figure 3. Gray solid and dashed lines display percent δ-power in 30 seconds intervals for the same 2 subjects at baseline (top panel) and visit 2 (bottom panel). Missing data correspond to wake periods. Solid black line displays visit-specific average δ power over all subjects.

ronal 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 $t \to \mathrm{NP}_\delta(t)$, where t indicates the time corresponding to a particular 30 second interval. For illustration, Figure 3 displays the pairs $\{t, \mathrm{NP}_\delta(t)\}$ for two subjects (gray solid and dashed lines) at baseline and visit 2. Time $t = 1$ corresponds to the first 30 second interval after sleep onset. Figure 3 also displays the visit-specific average percent δ power across all subjects (solid black line). 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 3, $W_{ij}(t)$ denotes these "centered" functional data for subject i at visit j during the tth 30-second interval. We used model (3) 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-level and visit-level eigenfunctions and eigenvalues we used the MFPCA methodology introduced by Di et al. (2009) and summarized in Section 2.1. Table 2 provides the estimated eigenvalues at both levels indicating that 95% of level 1 (subject) variability is explained by the first five eigenfunctions and 80% is explained by the first eigenfunction. Table 2 indicates that there are more directions of variation in the level 2 (visit) space. Indeed, 80% of the variability is explained by the first 7 eigenfunctions and 90% of the variability is explained by the first 14 components (results not shown). The proportion of variability explained by subject-level functional clustering was $\hat{\rho}_W = 0.213$ with a 95% confidence interval: $(0.210, 0.236)$, that is, 21.3% of variability in the sleep EEG δ-power is attributable to the subject-level variability.

We started with $K = 5$ and performed a backward selection starting with the full outcome model $\mathrm{logit}\{P(Y_i = 1)\} = \beta_0 + \sum_{k=1}^{K} \beta_k \xi_{ik}$, where Y_i is the HTN indicator variable and no additional covariates were included into the model. Three principal components were eliminated in the following order: PC4 (p-value $= 0.49$), PC2 (p-value $= 0.46$), PC3 (p-value $= 0.23$). The other two principal components (PCs) were retained in the model: PC1 (p-value < 0.001) and PC5 (p-value $= 0.0012$). For illustration, Figure 4 displays principal components 1, 2, 3, and 5. PC1 is, basically, a vertical shift. Thus, subjects who are positively loaded on it have a higher long-term δ-power than the population average. PC5 is roughly centered around 0 and it has a more interesting behavior: a subject who is positively loaded on PC5 will have a lower percent δ-power (faster brain activity) in the first 45 minutes. This difference is more pronounced in the first 10, 15 minutes of sleep, with the subject "catching-up" to the population average between minute 45 and 60. After 1 hour of sleep the subject will have a higher percent δ-power (slower brain activity) than the average population. After 2 hours, the behavior along this component returns to the population average. Both PC1 and PC5 are very strong predictors of HTN, even though they explain very different proportions of subject-level variability: PC1 (80%) and PC5 (2%). As will be seen below, the parameter of PC5 is negative indicating that subjects who are positively loaded on this component are less likely to have HTN.

Table 3 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

Table 2. Estimated eigenvalues on both levels for SHHS data. We showed the first 5 components for level 1 (subject level), and 7 components for level 2

Level 1 eigenvalues							
Component	1	2	3	4	5		
Eigenvalue ($\times 10^{-3}$)	12.97	1.22	0.53	0.45	0.33		
% var	80.81	7.60	3.29	2.79	2.05		
cum. % var	80.81	88.40	91.70	94.48	96.53		
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
cum. % var	21.84	34.63	47.17	58.02	67.61	75.13	80.30

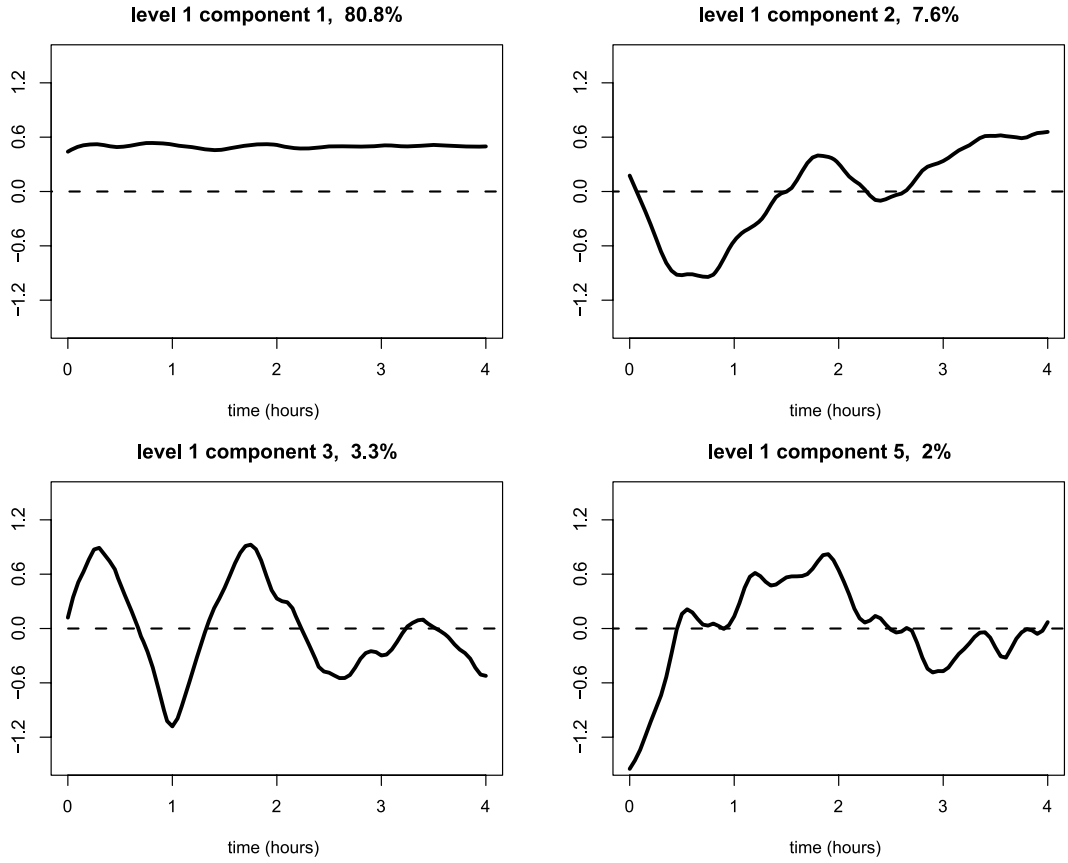

Figure 4. Characteristics of normalized sleep EEG δ-power. Principal components 1, 2, 3, and 5 of the subject-level functional space.

was fitted using a two-stage analysis with BLUP estimates of scores from the exposure model and a joint Bayesian analysis. We note that there is good agreement between the two methods with the exception of the statistical significance of PC5: the two-stage analysis finds it highly significant whereas the Bayesian analysis does not. As expected, the magnitude of association varies with the amount of confounding adjustment. For example, Model 1 estimates that a one standard deviation increase in PC1 scores corresponds to a relative risk $e^{-1.55*0.11} = 0.84$ (Table 2 provides the variance of PC1 scores). Model 2, which adjusts for confounders, estimated that a one standard deviation increase in PC1 scores corresponds to a relative risk $e^{-0.85*0.11} = 0.91$.

These results are now easy to explain. The bias of point estimators is likely due to the variability of PC scores. The wider credible intervals obtained from the Bayesian analysis are likely due to the appropriate incorporation of the sources of variability. The negative relationship between smoking and hypertension may seem counterintuitive. However, in this study smokers are younger, have a lower BMI and many other smokers with severe disease were not included in the study Zhang et al. (2006).

Figure 5 displays results for $\beta(t)$, the functional association effect between subject-specific deviations, $X_i(t)$, from the visit-specific mean, $\mu(t) + \eta_j(t)$, and HTN without accounting for confounders. The top panel shows results for the optimal model

Table 3. Mean and standard error estimates (within brackets) for parameters of models of 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

	Two-stage analysis		Joint analysis	
	Model 1	Model 2	Model 1	Model 2
Score 1	−1.55 (0.28)*	−0.85 (0.30)*	−1.75 (0.33)*	−1.08 (0.40)*
Score 5	−7.03 (2.18)*	−4.67 (2.34)*	−7.68 (3.90)	−1.97 (3.80)
Sex		0.10 (0.08)		0.09 (0.08)
smk:former		−0.18 (0.08)*		−0.19 (0.08)*
smk:current		−0.10 (0.13)		−0.10 (0.13)
Age		0.06 (0.00)*		0.06 (0.00)*
BMI		0.06 (0.01)*		0.06 (0.01)*
RDI		0.01 (0.00)*		0.01 (0.00)*

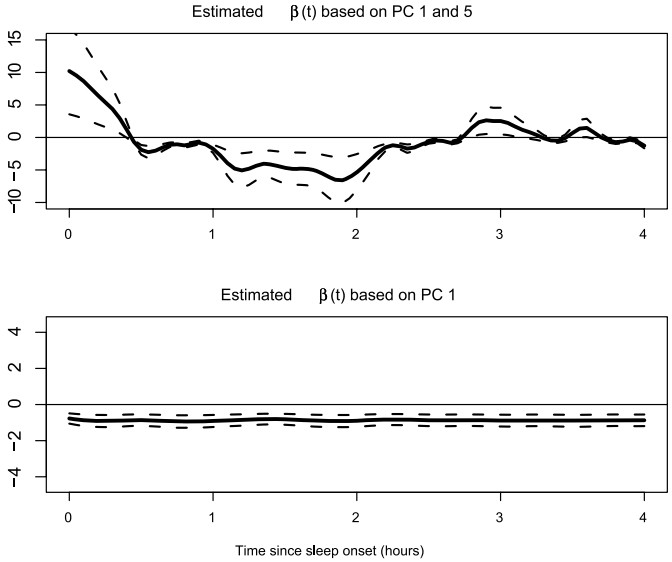

Figure 5. Results for $\beta(t)$, the functional association effect between subject-specific deviations, $X_i(t)$, from the visit-specific mean, $\mu(t) + \eta_j(t)$, and HTN in the model without confounders. Two-stage (top panel); joint Bayesian (bottom panel).

using a two-stage frequentist analysis. This model includes PCs 1 and 5. The bottom panel shows results for the optimal model using a joint Bayesian analysis. This model includes only PC1, because PC5 was not found to be statistically significant using a joint approach. The differences are visually striking, but they are due to the special shape of PC5 and to the fact that the methods disagree on its importance. Indeed, point estimators of the PC5 component are very close, but Bayesian analysis estimates an 80% larger standard error.

Joint Bayesian analysis is simple, robust and requires minimal tunning. This is possible because MFPCA produces a parsimonious decomposition of the functional variability using orthonormal bases. The use of orthonormal bases leads to reduction in the number of parameters and of posterior correlation among parameters, which lead to excellent mixing properties. For example, the web supplement displays chains for the regression coefficients indicating independence-like behavior.

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. 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, our framework opened many new theoretical problems; addressing all these problems exceeds the scope of the current paper and will be addressed in subsequent papers. Fourth, the computational problems may seem daunting, especially when we propose a joint Bayesian analysis of a dataset with thousands of subjects, multiple visits and thousands of random effects. However, we do not think that they are insurmountable; see the software we posted at *www.biostat.jhsph.edu/~ccrainic/webpage/software/GFR.zip*.

SUPPLEMENTAL MATERIALS

Theorem Proofs: Proofs of the two theorems. (web_supplement.pdf)

Chain Histories: Plot of chain histories for a functional regression model. (sleep-GFLR-chain.pdf)

[Received October 2008. Revised May 2009.]

REFERENCES

Akaike, H. (1973), "Maximum Likelihood Identification of Gaussian Autoregressive Moving Average Models," *Biometrika*, 60, 255–265.

Baladandayuthapani, V., Mallick, B. K., Hong, M. Y., Lupton, J. R., Turner, N. D., and Carroll, R. J. (2008), "Bayesian Hierarchical Spatially Correlated Functional Data Analysis With Application to Colon Carcinogenesis," *Biometrics*, 64, 64–73.

Bilodeau, M., and Brenner, D. (1999), *Theory of Multivariate Statistics*, New York: Springer-Verlag.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

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

Chiou, J.-M., and Müller, H.-G. (1998), "Quasi-Likelihood Regression With Unknown Link and Variance Functions," *Journal of the American Statistical Association*, 93, 1376–1387.

Chiou, J.-M., Müller, H.-G., and Wang, J.-L. (2003), "Functional Quasi-Likelihood Regression Models With Smooth Random Effects," *Journal of the Royal Statistical Society, Ser. B*, 65, 405–423.

Congdon, P. (2003), *Applied Bayesian Modelling*, Chichester, England: Wiley.

Crainiceanu, C. M., and Ruppert, D. (2004), "Likelihood Ratio Tests in Linear Mixed Models With One Variance Component," *Journal of the Royal Statistical Society, Ser. B*, 66, 165–185.

Crainiceanu, C. M., Caffo, B., Di, C., and Punjabi, N. (2009), "Nonparametric Signal Extraction and Measurement Error in the Analysis of Electroencephalographic Activity During Sleep," *Journal of the American Statistical Association*, 104, 541–555.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Adarsh, J., and Goodner, B. (2007), "Spatially Adaptive Penalized Splines With Heteroscedastic Errors," *Journal of Computational and Graphical Statistics*, 16 (2), 265–288.

Crainiceanu, C. M., Ruppert, D., Claeskens, G., and Wand, M. P. (2005), "Exact Likelihood Ratio Tests for Penalized Splines," *Biometrika*, 92 (1), 91–103.

Di, C., Crainiceanu, C. M., Caffo, B., and Naresh, P. (2009), "Multilevel Functional Principal Component Analysis," *The Annals of Applied Statistics*, 3 (1), 458–488.

Fan, J., and Zhang, J.-T. (2000), "Two-Step Estimation of Functional Linear Models With Application to Longitudinal Data," *Journal of the Royal Statistical Society, Ser. B*, 62, 303–322.

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models," *Bayesian Analysis*, 1 (3), 515–533.

Gelman, A., Carlin, J. B., Stern, H. A., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.), Boca Raton, FL: Chapman & Hall/CRC.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice*, Boca Raton, FL: Chapman & Hall/CRC.

Guo, W. (2002), "Functional Mixed Effects Models," *Biometrics*, 58, 121–128.

Hall, P., and Hosseini-Nasab, M. (2006), "On Properties of Functional Principal Components Analysis," *Journal of the Royal Statistical Society, Ser. B*, 68, 109–126.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006), "Properties of Principal Component Methods for Functional and Longitudinal Data Analysis," *The Annals of Statistics*, 34, 1493–1517.

Indritz, J. (1963), *Methods in Analysis*, New York: Macmillan & Colier-Macmillan.

James, G. M. (2002), "Generalized Linear Models With Functional Predictors," *Journal of the Royal Statistical Society, Ser. B*, 64, 411–432.

James, G. M., Hastie, T. G., and Sugar, C. A. (2000), "Principal Component Models for Sparse Functional Data," *Biometrika*, 87, 587–602.

Karhunen, K. (1947), "Über lineare Methoden in der Wahrscheinlichkeitsrechnung," *Annales AcadeniæScientiarum Fennicæ, Series A1: Mathematica-Physica*, 37, 3–79.

Loève, M. (1945), "Functions Aleatoire de Second Ordre," *Comptes Rendus des Séances de l'Academie des Sciences*, 220.

Luo, X., Stefanski, L. A., and Boos, D. D. (2006), "Tuning Variable Selection Procedures by Adding Noise," *Technometrics*, 48, 165–175.

Morris, J. S., and Carroll, R. J. (2006), "Wavelet-Based Functional Mixed Models," *Journal of the Royal Statistical Society, Ser. B*, 68, 179–199.

Morris, J. S., Vanucci, M., Brown, P. J., and Carroll, R. J. (2003), "Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis," *Journal of the American Statistical Association*, 98, 573–583.

Müller, H.-G. (2005), "Functional Modelling and Classification of Longitudinal Data," *Scandivanian Journal of Statistics*, 32, 223–240.

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

Natarajan, R., and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 95, 227–237.

Ramsay, J. O., and Silverman, B. W. (2005), *Applied Functional Data Analysis*, New York: Springer-Verlag.

——— (2006), *Functional Data Analysis*, New York: Springer-Verlag.

Reiss, P. T., and Ogden, R. T. (2007), "Functional Principal Component Regression and Functional Partial Least Squares," *Journal of the American Statistical Association*, 102, 984–996.

——— (2009), "Functional Generalized Linear Models With Images as Predictors," *Biometrics*, to appear.

Rice, J. A. (2004), "Functional and Longitudinal Data Analysis," *Statistica Sinica*, 14, 631–647.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society, Ser. B*, 53, 233–243.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2009), "Fast Methods for Spatially Correlated Multilevel Functional Data," manuscript, North Carolina State University.

Stram, D. O., and Lee, J. W. (1994), "Variance Components Testing in the Longitudinal Mixed Effects Model," *Biometrics*, 50, 1171–1177.

Wang, N., Carroll, R. J., and Lin, X. (2005), "Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data," *Journal of the American Statistical Association*, 100, 147–157.

Wu, Y., Boos, D. D., and Stefanski, L. A. (2007), "Controlling Variable Selection by the Addition of Pseudovariables," *Journal of the American Statistical Association*, 102, 235–243.

Yao, F., and Lee, T. C. M. (2006), "Penalized Spline Models for Functional Principal Component Analysis," *Journal of the Royal Statistical Society, Ser. B*, 68, 3–25.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903.

Zhang, L., Samet, J., Caffo, B., and Punjabi, N. M. (2006), "Cigarette Smoking and Nocturnal Sleep Architecture," *American Journal of Epidemiology*, 164 (6), 529–537.

Zhao, X., Marrron, J. S., and Wells, M. T. (2004), "The Functional Data Analysis View of Longitudinal Data," *Statistica Sinica*, 14, 789–808.