Biostatistics (2018) **19**, 2, *pp.* 137–152
doi:10.1093/biostatistics/kxx026
Advance Access publication on June 25, 2017

Simple fixed-effects inference for complex functional models

SO YOUNG PARK*, ANA-MARIA STAICU, LUO XIAO

Department of Statistics, North Carolina State University, Raleigh, NC, USA
spark13@ncsu.edu

CIPRIAN M. CRAINICEANU

Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA

Summary

We propose simple inferential approaches for the fixed effects in complex functional mixed effects models. We estimate the fixed effects under the independence of functional residuals assumption and then bootstrap independent units (e.g. subjects) to conduct inference on the fixed effects parameters. Simulations show excellent coverage probability of the confidence intervals and size of tests for the fixed effects model parameters. Methods are motivated by and applied to the Baltimore Longitudinal Study of Aging, though they are applicable to other studies that collect correlated functional data.

Keywords: Bootstrap/resampling; Functional data; Measurement error; Smoothing and nonparametric regression.

1. Introduction

Rapid advancement in technology and computation has led to an increasing number of studies that collect complex-correlated functional data. In response to these studies research in structured functional data analysis (FDA) has witnessed rapid development. A major characteristic of these data is that they are strongly correlated, as multiple functions are observed on the same observational unit. Many new studies have functional structures including multilevel (Morris *and others*, 2003; Morris and Carroll, 2006; Di *and others*, 2009; Crainiceanu *and others*, 2009), longitudinal (Greven *and others*, 2010; Chen and Müller, 2012; Scheipl *and others*, 2015), spatially aligned (Baladandayuthapani *and others*, 2008; Staicu *and others*, 2010; Serban *and others*, 2013), or crossed (Aston *and others*, 2010; Shou *and others*, 2015).

While these types of data can have highly complex dependence structures, one is often interested in simple, population-level, questions for which the multi-layered structure of the correlation is just an infinite-dimensional nuisance parameter. For example, the Baltimore Study of Aging (BLSA), which motivated this article, collected physical activity levels from each of many participants at the minute level for multiple consecutive days. Thus, the BLSA activity data exhibit complex within-day and between-day correlations. However, the most important questions in the BLSA tend to be simple; in particular, one may be interested in how age affects the daily patterns of activity or whether the effect is different by gender. In this context, the high complexity and size of the data are just technical inconveniences.

*To whom correspondence should be addressed.

Such simple questions are typically answered by estimating fixed effects in complex functional mixed effects models. Our proposed approach avoids complex modeling and implementation by: (i) estimating the fixed (population-level) effects under the assumption of independence of functional residuals; and (ii) using a nonparametric bootstrap of independent units (e.g. subjects) to construct confidence intervals and conduct tests. A natural question is whether efficiency is lost by ignoring the correlation. While the loss of efficiency is well documented in longitudinal studies with few observations per subject and small dimensional within-subject correlation, little is known about inference when there are many observations per subject with an unknown large dimensional within-subject correlation matrix. An important contribution of this article is to evaluate the performance of bootstrap-based inferential approaches in this particular context. Our view is that estimating large dimensional covariance matrices of functional data may hurt fixed effects estimation by wasting degrees of freedom. Indeed, a covariance matrix for an n by p matrix of functional data (n = number of subjects and p = number of subject-specific observations) would require estimation of $p(p+1)/2$ matrix covariance entries when the covariance matrix is unstructured. When p is moderate or large this is a difficult problem. Moreover, the resulting matrix has an unknown low rank and is not invertible.

We will consider cases when multiple functional observations are observed for the same subject. This structure is inspired by many current observational studies, but we will focus on the BLSA, where activity data are recorded at the minute level over multiple consecutive days, resulting in daily activity profiles (each as a function of time of the day) observed for each participant over multiple days. Specifically, we will focus on data from 332 female BLSA participants with age varying between 50 and 90. A total of 1580 daily activity profiles were collected (an average of 4.7 monitoring days per person), where each daily profile consists of 1440 activity counts measured at the minute level. Thus, the activity data considered in this paper is stored in a 1580 × 1440 dimensional matrix. Our primary interest is to conduct inference on the fixed effects of covariates, such as age and body mass index (BMI), on daily activity profiles. Because data from each participant were collected on consecutive days in a short period (about a week on average), age and BMI in the BLSA data are subject-specific but visit-invariant.

While our covariates are time-invariant, we propose methods that can accommodate both time-invariant and time-dependent covariates. Assume that the observed data is of the form $\{Y_{ij}(\cdot), \mathbf{X}_{ij}\}$, where $Y_{ij}(\cdot)$ is the jth unit functional response (e.g. jth visit) for the ith subject, and $\mathbf{X}_{ij}$ is the corresponding vector of covariates. This general form applies to all types of functional data discussed above: multilevel, longitudinal, spatially correlated, crossed, etc. The main objective is to make statistical inference for the population-level effects of interest using repeatedly observed functional response data.

A naïve approach to analyze data with such a complex structure is to ignore the dependence over the functional argument t, but to account for the dependence across the repeated visits. That is, by assuming that the responses $Y_{ij}(t)$ are correlated over j and independent over t. Longitudinal data analysis literature offers a wide variety of models and methods for estimating the fixed effects and their uncertainty, and for conducting tests (see for example Laird and Ware (1982); Liang and Zeger (1986); Fitzmaurice *and others* (2012)). These methods allow to account for within-subject correlation, incorporate additional covariates, and make inference about the fixed effects. However, extending these estimation and inferential procedures to functional data is difficult. In the literature this has been addressed by modeling the within- and between-curve dependence structure using functional random effects. These approaches are highly computationally intensive, require inverting high dimensional covariances matrices, and make implicit assumptions about the correlation structures that may not be easy to transport across applications.

Another possible approach is to completely ignore the dependence across the repeated visits j, but account for the functional dependence. That is, assume $Y_{ij}(t)$ are dependent over t, but independent over j. Function on scalar/vector regression models can be used to estimate the fixed effects of interest; see for example Faraway (1997); Jiang *and others* (2011); Ivanescu *and others* (2015). In this context, testing procedures for hypotheses on fixed effects are available. For example, Shen and Faraway (2004) proposed the functional F statistic for testing hypotheses related to nested functional linear models. Zhang *and others* (2007) proposed L_2 norm based test for testing the effect of a linear combination of time-varying coefficients, and approximate the null sampling distribution using resampling methods. However, failing to account for dependence across visits results in tests with inflated type I error.

In contrast, development of statistical inferential methods for correlated functional data has received less attention. Fully Bayesian inference has been previously considered in the literature for complex designs; see, for example, Morris and Carroll (2006), Morris *and others* (2006), Morris *and others* (2011), Zhu *and others* (2011), and Zhang *and others* (2016). These approaches take into account both between- and within-function correlations using MCMC simulations of the posterior distribution. In contrast, we focus on a frequentist approach to inference that avoids modeling of the complex correlation structures. In the frequentist framework, Crainiceanu *and others* (2012) discussed bootstrap-based inferential methods for the difference in the mean profiles of correlated functional data. Staicu *and others* (2014) proposed a likelihood-ratio type testing procedure, while Staicu *and others* (2015) considered L_2 norm-based testing procedures for testing that multiple group mean functions are equal. Horváth *and others* (2013) developed inference for the mean function of a functional time series. However, these approaches focus on testing the effect of a categorical variable, and do not handle inference on fixed effects in full generality.

Here we consider a modeling framework that is a direct generalization of the linear mixed model framework from longitudinal data analysis, where scalar responses are replaced with functional ones. We propose to model the fixed effect of a scalar covariate either parametrically or nonparametrically while the error covariance is left unspecified to avoid model complexity. We estimate the fixed effects under the working independence and account for all known sources of data dependence by bootstrapping over subjects. Based on this procedure, we propose confidence bands and L_2 norm-based testing for fixed effects parameters. An important contribution of this article is to investigate and confirm the performance of the bootstrap-based inferential approaches when data have a complex functional dependence structure.

2. Modeling framework and estimation

Consider the case when each subject is observed at m_i visits, and data at each visit consist of a functional outcome $\{Y_{ij\ell} = Y_{ij}(t_{ij\ell}) : \ell = 1, \ldots, L_{ij}\}$ and a vector of covariates including a scalar covariate of interest, X_{ij}, and additional p-dimensional vector of covariates, $\mathbf{Z}_{ij}$. We assume that $t_{ij\ell} \in \mathcal{T}$, where $\mathcal{T}$ is a compact and closed domain; take $\mathcal{T} = [0, 1]$ for simplicity. For convenience, we assume a balanced regular sampling design, i.e. $t_{ij\ell} = t_\ell$ and $L_{ij} = L$, though all methods apply to general sampling designs. Furthermore, we assume that $\{X_{ij} : \forall\, i, j\}$ is a dense set in the closed domain $\mathcal{X}$; this assumption is needed when the fixed effect of X_{ij} is modeled nonparametrically (Ruppert *and others*, 2003; Fitzmaurice *and others*, 2012). A common approach for the study of the effect of the covariates on $Y_{ij}(\cdot)$ is to posit a model of the type

$$Y_{ij}(t) = \mu(t, X_{ij}) + \mathbf{Z}_{ij}^T \boldsymbol{\tau} + \epsilon_{ij}(t), \tag{2.1}$$

where $\mu(t, X_{ij})$ is a time-varying smooth fixed effect of the covariate of interest, X_{ij}, and $\boldsymbol{\tau}$ is a p-dimensional parameter quantifying the linear additive fixed effect of the covariate vector, $\mathbf{Z}_{ij}$. $\epsilon_{ij}(t)$ is a zero-mean random deviation that incorporates both the within- and between-subject variability. $\mu(t, X_{ij})$ can be modeled either parametrically or nonparametrically; see Section 6 (F2.) for some possible mean structures. While technically more difficult to implement, nonparametric smoothing is useful when limited information about the mean structure is available.

Here we present the most complex case where the mean structure for $\mu(t, X)$ is an unknown bivariate smooth function. We construct a bivariate basis using the tensor product of two univariate B-spline bases, $\{B_1^t(t), \cdots, B_{d_t}^t(t)\}$, and $\{B_1^x(x), \cdots, B_{d_x}^x(x)\}$, defined on $\mathcal{T}$ and $\mathcal{X}$ respectively. The unspecified mean is then expressed as $\mu(t,x) = \mathbf{B}(t,x)^T\boldsymbol{\beta}$, where $\mathbf{B}(t,x)$ is the $d_t d_x$-dimensional vector of $B_l^t(t)B_r^x(x)$'s and $\boldsymbol{\beta}$ is the vector of parameters β_{lr}. Typically, the number of basis functions is chosen sufficiently large to capture the maximum complexity of the mean function and smoothness is induced by a quadratic penalty on the coefficients. There are several penalties for bivariate smoothing; see, for example, Marx and Eilers (2005), Wood (2006), and Xiao *and others* (2013, 2016). In this article we used the following estimation criterion

$$\underset{\boldsymbol{\beta},\,\boldsymbol{\tau},\,\lambda}{\text{argmin}} \sum_{i,j,\ell}[Y_{ij\ell} - \{\mathbf{B}(t_\ell, X_{ij})^T\boldsymbol{\beta} + \mathbf{Z}_{ij}^T\boldsymbol{\tau}\}]^2 + \boldsymbol{\beta}^T P_\lambda \boldsymbol{\beta} \tag{2.2}$$

with a penalty matrix P_λ described in Wood (2006) and a vector of smoothing parameters, λ. Specifically, we used $P_\lambda = \lambda_t P_t \otimes I_{d_x} + \lambda_x I_{d_t} \otimes P_x$ and $\lambda = (\lambda_t, \lambda_x)^T$, where $\otimes$ denotes the tensor product, and P_t and λ_t are the marginal second order difference matrix and the smoothing parameter for the t direction, respectively; P_x and λ_x are defined similarly for the x direction. Here I_{d_t} and I_{d_x} are the identity matrices of dimensions d_t and d_x, respectively. For a fixed smoothing parameter, λ, the minimizer of (2.2) has the form $[\widehat{\boldsymbol{\beta}}_\lambda^T, \widehat{\boldsymbol{\tau}}_\lambda^T]^T = (\mathbf{M}^T\mathbf{M} + P_\lambda)^{-1}\mathbf{M}^T\mathbf{Y}$ where $\mathbf{M} = [\mathbf{M_1} \vdots \mathbf{M_2}]$ with $\mathbf{M_1}$ the matrix with rows $\mathbf{B}(t_\ell, X_{ij})$ and $\mathbf{M_2}$ the matrix obtained by row-stacking of $\mathbf{Z}_{ij}^T$, while the estimated mean is $\widehat{\mu}(t,x) + \mathbf{Z}_{ij}^T\widehat{\boldsymbol{\tau}} = \mathbf{B}(t,x)^T\widehat{\boldsymbol{\beta}}_\lambda + \mathbf{Z}_{ij}^T\widehat{\boldsymbol{\tau}}_\lambda$. In this article, the generalized cross validation (GCV) is used to select the optimal smoothing parameters, while other criteria such as the restricted maximum likelihood can be used; relevant literatures on selection of the smoothing parameter include Wahba (1990) and Ruppert *and others* (2003).

Estimation of the fixed effects in model (2.1) under the working independence assumption is not new; see for example Scheipl *and others* (2015) and Chen and Müller (2012). However, our approach to inference for the population level fixed effects in the context of structured functional data has not been studied. The novelty of this article consists precisely in filling this gap in the literature. We consider an estimation approach of fixed effects under working independence and a bootstrap of independent units approach to appropriately account for complex correlation.

3. CONFIDENCE BANDS FOR $\mu(t,x)$

We now discuss inference for $\mu(t,x)$ using confidence bands and formal hypothesis testing. Without loss of generality, assume that the mean structure is $\mu(t,x) = \mathbf{B}(t,x)^T\boldsymbol{\beta}$, where $\mathbf{B}(t,x)^T$ can be as simple as $(1,t,x)$ or as complex as a vector of prespecified basis functions. The mean estimator of interest is $\widehat{\mu}(t,x) = \mathbf{B}(t,x)^T\widehat{\boldsymbol{\beta}}$. One could study pointwise variability for every pair (t,x), that is var$\{\widehat{\mu}(\mathrm{t},\mathrm{x})\}$, or the joint variability for the entire domain $\mathcal{T} \times \mathcal{X}$, that is cov$\{\widehat{\mu}(t,x) : t \in \mathcal{T}, x \in \mathcal{X}\}$. Irrespective of the choice, the variability is fully described by the variability of the parameter estimator $\widehat{\boldsymbol{\beta}}$.

3.1. *Bootstrap algorithms*

We consider a flexible dependence structure for $\epsilon_{ij}(t)$ that describes both within- and between-subject variability. We make minimal assumption that $\epsilon_{ij}(t)$ is independent over i but correlated over j and t, though *we do not specify the form of this correlation*. Deriving the analytical expression for the sampling variability of the estimator $\widehat{\boldsymbol{\beta}}$ in such contexts is challenging. Instead, we propose to use two bootstrap algorithms: bootstrap of subject-level data and bootstrap of subject-level residuals. These approaches have already been studied and used in nonparametric regression for independent measurements; see, for example, Härdle and Bowman (1988), Efron and Tibshirani (1994), and Hall *and others* (2013) among many others. Bootstrap of functional data for fixed effects has also been considered, including by Politis and Romano (1994) for weakly dependent processes in Hilbert space, by Cuevas *and others* (2006) for independent functional data, and by Crainiceanu *and others* (2012) for paired samples of functional data. However, studying these bootstrap algorithms for functional data with complex correlation is new.

The subject-level bootstrap algorithm for correlated functional data is provided below.

Algorithm 1 Bootstrap of the subject-level data [uncertainty estimation]

```
for $b \in \{1, \dots, B\}$ do
    Re-sample the subject indexes from the index set $\{1, \dots, n\}$ with replacement.
    Let $I^{(b)}$ be the resulting sample of $n$ subjects.
    Define the $b$th bootstrap data by:
    $\text{data}^{(b)} = [\{Y_{i^*j}(t_\ell), X_{i^*j}, \mathbf{Z}_{i^*j}, t_\ell\} : i^* \in I^{(b)},\ j = 1, \dots, m_{i^*},$ and $\ell = 1, \dots, L]$.
    Using $\text{data}^{(b)}$ fit the model (2.1) with the mean structure of interest modeled by
    $\mu(t,x) = \mathbf{B}(t,x)^T\boldsymbol{\beta}$, by employing criterion (2.2). Let $\widehat{\boldsymbol{\beta}}^{(b)}_{\lambda(b)}$ be the corresponding estimate of
    the parameter of interest; similarly define $\widehat{\mu}^{(b)}(t,x) = \mathbf{B}(t,x)^T\widehat{\boldsymbol{\beta}}^{(b)}_{\lambda(b)}$. end for
Calculate the sample covariance of $\{\widehat{\boldsymbol{\beta}}^{(1)}_{\lambda(1)}, \dots, \widehat{\boldsymbol{\beta}}^{(B)}_{\lambda(B)}\}$; denote it by $V_{\widehat{\boldsymbol{\beta}}}$.
```

The bootstrap of subject-level data is more generally applicable, while the bootstrap of subject-level residuals approach relies on two important assumptions: (i) the covariates do not depend on visit, that is $X_{ij} = X_i$ and $\mathbf{Z}_{ij} = \mathbf{Z}_i$; and (ii) both the correlation and the error variance are independent of covariates. These assumptions ensure that sets of subject-level errors, i.e. $\{\epsilon_{ij}(t) : j = 1, \dots, m_i\}$ for $i = 1, \dots, n$, can be resampled over subjects without affecting the sampling distribution. These assumptions are reasonable when covariates are independent of the visit, as is the case in the BLSA application. Indeed, in BLSA we consider age and BMI, which are time-invariant because repeated measures per subject were collected within a week.

Similarly, we introduce the algorithm for bootstrapping residuals. We start by fitting the model (2.1) with the mean structure of interest modeled by $\mu(t,x) = \mathbf{B}(t,x)^T\boldsymbol{\beta}$, using the estimation criterion described in (2.2), and calculating the residuals $e_{ij}(t_\ell) = Y_{ij}(t_\ell) - \mathbf{B}(t_\ell, X_i)^T\widehat{\boldsymbol{\beta}}_\lambda - \mathbf{Z}_i^T\widehat{\boldsymbol{\tau}}_\lambda$.

Algorithm 2 Bootstrap of the subject-level residuals [uncertainty estimation]

```
for $b \in \{1, \dots, B\}$ do
    Re-sample the subject indexes from the index set $\{1, \dots, n\}$ with replacement. Let $I^{(b)}$ be
    the resulting sample of subjects. For each $i = 1, \dots, n$ denote by $m_i^*$ the number of repeated
    time-visits for the $i$th subject selected in $I^{(b)}$.
    Define the $b$th bootstrap sample of residuals
    $\{e^*_{ij}(t_\ell) : i = 1, \dots, n,\ j = 1, \dots, m_i^*,$ and $\ell = 1, \dots, L\}$.
    Define the $b$th bootstrap data by:
    $\text{data}^{(b)} = [\{Y^*_{ij}(t_\ell), X_i, Z_i, t_\ell\} : i = 1, \dots, n, j = 1, \dots, m_i^*, \ell = 1, \dots, L]$, where $Y^*_{ij}(t_\ell) =$
    $\mathbf{B}(t_\ell, X_i)^T\widehat{\boldsymbol{\beta}}_\lambda + \mathbf{Z}_i^T\widehat{\boldsymbol{\tau}}_\lambda + e^*_{ij}(t_\ell)$.
    Using $\text{data}^{(b)}$ fit the model (2.1) with the mean structure of interest modeled by
    $\mu(t,x) = \mathbf{B}(t,x)^T\boldsymbol{\beta}$, by employing criterion (2.2). Let $\widehat{\boldsymbol{\beta}}^{(b)}$ be the corresponding estimate of
    the parameter of interest; similarly define $\widehat{\mu}^{(b)}(t,x) = \mathbf{B}(t,x)^T\widehat{\boldsymbol{\beta}}^{(b)}_{\lambda(b)}$. end for
Calculate the sample covariance of $\{\widehat{\boldsymbol{\beta}}^{(1)}_{\lambda(1)}, \dots, \widehat{\boldsymbol{\beta}}^{(B)}_{\lambda(B)}\}$; denote it by $V_{\widehat{\boldsymbol{\beta}}}$.
```

Based on our numerical investigation (see Section 6) the bootstrap of subject-level residuals has excellent performance and is recommended when the necessary assumptions are satisfied, though the bootstrap of subjects is a good alternative.

3.2. *Bootstrap-based inference*

For fixed (t,x), the variance of the estimator $\widehat{\mu}(t,x) = \mathbf{B}(t,x)^T\widehat{\boldsymbol{\beta}}$ can be estimated as $\text{var}\{\widehat{\mu}(t,x)\} = \mathbf{B}(t,x)^T\, V_{\widehat{\boldsymbol{\beta}}}\, \mathbf{B}(t,x)$, by using the bootstrap-based estimate of the covariance of $\widehat{\boldsymbol{\beta}}$. A $100(1-\alpha)\%$ pointwise confidence interval for $\mu(t,x)$ can be calculated as $\widehat{\mu}(t,x) \pm z^*_{\alpha/2}\sqrt{\text{var}\{\widehat{\mu}(t,x)\}}$, using normal distributional assumption for the estimator $\widehat{\mu}(t,x)$, where $z^*_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal. An alternative is to use the pointwise $100(\alpha/2)\%$ and $100(1-\alpha/2)\%$ quantiles of the bootstrap estimates $\{\widehat{\mu}^{(b)}(t,x) : b = 1, ..., B\}$.

In most cases, it makes more sense to study the variability of $\widehat{\mu}(t,x)$, and draw inference about the entire true mean function $\{\mu(t,x) : (t,x) \in \mathcal{D}_t \times \mathcal{D}_x\}$. Thus, we focus our study on constructing a joint (or simultaneous) confidence band for $\mu(t,x)$. Constructing simultaneous confidence bands for univariate smooths has already been discussed in the nonparametric literature. For example, Degras (2009), Ma *and others* (2012), and Cao *and others* (2012) proposed asymptotically correct simultaneous confidence bands for different estimators, when data are independently sampled curves; Crainiceanu *and others* (2012) proposed bootstrap-based joint confidence bands for univariate smooths in the case of functional data with complex error processes. Here, we present an extension of the approach considered by Crainiceanu *and others* (2012) to bivariate smooth function estimation for general functional correlation structures.

Let $\mathbf{T}^* = \{t_{g_t} : g_t = 1, \ldots, G_t\}$ and $\mathbf{X}^* = \{x_{g_x} : g_x = 1, ..., G_x\}$ be the evaluation points that are equally spaced in the domains $\mathcal{D}_t$ and $\mathcal{D}_x$, respectively. We evaluate the bootstrap estimate $\widehat{\mu}^{(b)}(t,x)$ of one bootstrap sample at all pairs $(t,x) \in \mathbf{T}^* \times \mathbf{X}^*$, and denote by $\widehat{\boldsymbol{\mu}}^{(\boldsymbol{b})}$ the G_tG_x-dimensional vector with components $\widehat{\mu}^{(b)}(t,x)$. Let $\mathbf{B}$ be the $\dim(\boldsymbol{\beta}) \times G_tG_x$-dimensional matrix obtained by column-stacking $\mathbf{B}(t_{g_t}, x_{g_x})$ for all g_t and g_x. Let $s(t_{g_t}, x_{g_x}) = \sqrt{\text{var}\{\widehat{\mu}(t_{g_t}, x_{g_x})\}}$ as defined above. After adjusting for the bivariate structure of the problem, the main steps of the construction of the joint confidence bands for $\mu(t,x)$ follow similarly to the ones used in Crainiceanu *and others* (2012) for univariate smooth parameter functions.

Step 1. Generate a random variable $\mathbf{u}$ from the multivariate normal with mean $\mathbf{0}_{dim(\boldsymbol{\beta})}$ and covariance matrix $V_{\widehat{\boldsymbol{\beta}}}$; let $q(t_{g_t}, x_{g_x}) = \mathbf{B}(t_{g_t}, x_{g_x})^T\mathbf{u}$ for $g_t = 1, \ldots, G_t$ and $g_x = 1, \ldots, G_x$.
Step 2. Calculate $q^*_{max} = \max_{(t_{g_t}, x_{g_x})} \{|q(t_{g_t}, x_{g_x})|/\sqrt{s(t_{g_t}, x_{g_x})} : (t_{g_t}, x_{g_x}) \in \mathbf{T}^* \times \mathbf{X}^*\}$.
Step 3. Repeat *Step 1.* and *Step 2.* for $r = 1, \ldots, R$, and obtain $\{q^*_{max,r} : r = 1, \ldots, R\}$. Determine the $100(1-\alpha)\%$ empirical quantile of $\{q^*_{max,r} : r = 1, \ldots, R\}$, say $\widehat{q}_{1-\alpha}$.
Step 4. Construct the $100(1-\alpha)\%$ joint confidence band by: $\{\bar{\mu}(t_{g_t}, x_{g_x}) \pm \widehat{q}_{1-\alpha}\ \sqrt{s(t_{g_t}, x_{g_x})} : (t_{g_t}, x_{g_x}) \in \mathbf{T}^* \times \mathbf{X}^*\}$. Here $\bar{\mu}(t,x) = B^{-1}\sum_{b=1}^{B}\widehat{\mu}^{(b)}(t,x)$.
The joint confidence band, in contrast to the pointwise confidence band, can be used as an inferential tool for formal global tests about the mean function, $\mu(t,x)$. For example, one can use the joint confidence band for testing the null hypothesis, $\text{H}_0 : \mu(t,x) = f_0(t,x)$ for all pairs $(t,x) \in \mathcal{D}_t \times \mathcal{D}_x$ and for some prespecified function $f_0(t,x)$, by checking whether the confidence band $\bar{\mu}(t_{g_t}, x_{g_x}) \pm \widehat{q}_{1-\alpha}\ \sqrt{s(t_{g_t}, x_{g_x})}$ contains $\{f_0(t_{g_t}, x_{g_x})\}$ for all $(t_{g_t}, x_{g_x}) \in \mathcal{D}_t \times \mathcal{D}_x$. If the confidence band does not contain $\{f_0(t_{g_t}, x_{g_x})\}$ for some (t_{g_t}, x_{g_x}), then we conclude that there is significant evidence that the true mean function is the prespecified function f_0.

4. HYPOTHESIS TESTING FOR$\mu(t,x)$

Next, we focus on assessing the effect of the covariate of interest X on the mean function. Consider the general case when the model is (2.1) and the average effect is an unspecified bivariate smooth function, $\mu(t,x)$. Our goal is to test if the true mean function depends on x, that is testing:

$$\text{H}_0 : \quad \mu(t,x) = \mu_0(t) \text{ for all } t, x, \tag{4.1}$$

for some *unknown* smooth function $\mu_0 : \mathcal{D}_t \to R$ against $\text{H}_A : \mu(t,x)$ varies over x for some t.

To the best of our knowledge, this type of hypothesis, where the mean function is nonparametric both under the null and alternative hypotheses, has not been studied in FDA. The problem was extensively studied in nonparametric smoothing, where the primary interest centered on significance testing of a subset of covariates in a nonparametric regression model; see, for example, Fan and Li (1996), Lavergne and Vuong (2000), Delgado and Manteiga (2001), Gu *and others* (2007), and Hall *and others* (2007). However, all these methods are based on the assumption that observations are independent across sampling units; in our context requiring independence of $Y_{ij}(t_{ijk})$ over j and k is unrealistic and failing to account for this dependence leads to inflated type I error rates.

To test hypothesis (4.1) we propose a test statistic based on the L^2 distance between the mean estimators under the null and alternative hypotheses. Specifically we define it as:

$$T = \int_{\mathcal{X}} \int_{\mathcal{T}} \{\widehat{\mu}_A(t,x) - \widehat{\mu}_0(t)\}^2 \text{d}t\text{d}x, \tag{4.2}$$

where $\widehat{\mu}_0(t)$ and $\widehat{\mu}_A(t,x)$ are the estimates of $\mu(t,x)$ under the null and alternative hypotheses, respectively. In particular, $\widehat{\mu}_A(t,x)$ is estimated as in Section 2. The estimator $\widehat{\mu}_0(t)$ is obtained by modeling $\mu(t) = \sum_{l=1}^{d_t} B_l^t(t)\beta_l = \mathbf{B}(t)^T\boldsymbol{\beta}$ for the d_t-dimensional vector $\boldsymbol{\beta}$ and by estimating the mean parameters $\boldsymbol{\beta}$ based on a criterion similar to (2.2).

Deriving the finite sample distribution of the test statistic T under the null hypothesis is challenging and we propose to approximate it using the bootstrap. As in Section 3, the smoothing parameter selection is repeated for each bootstrap sample and model, $\mu_0(t)$ and $\mu_A(t,x)$.

Algorithm 3 Bootstrap approximation of the null distribution of the test statistic, T

```
for $b \in \{1,\dots,B\}$ do
    Re-sample the subject indexes from the index set $\{1,\dots,n\}$ with replacement. Let $I^{(b)}$ be
    the obtained sample of subjects. For each $i = 1,\dots,n$ denote by $m_i^*$ the number of repeated
    time-visits for the $i$th subject selected in $I^{(b)}$.
    Define the $b$th bootstrap sample of pseudo-residuals
    $\{e_{ij}^*(t_\ell) : i = 1,\dots,n,\ j = 1,\dots,m_i^*,$ and $\ell = 1,\dots,L\}$. For each $i = 1,\dots,n$
    let $\{\mathbf{Z}_{ij}^* : j = 1,\dots,m_i^*\}$ the corresponding sample of the nuisance covariates for the
    $i$th subject selected in $I^{(b)}$. Similarly define $X_{ij}^*$.
    Define the $b$th bootstrap data by: $\text{data}^{(b)} = [\{Y_{ij}^*(t_\ell), X_{ij}^*, \mathbf{Z}_{ij}^*] : i = 1,\dots,n, j = 1,\dots,m_i^*,$
    $\ell = 1,\dots,L\}$, where $Y_{ij}^*(t_\ell) = \widehat{\mu}_0(t_\ell) + \mathbf{Z}_{ij}^*\widehat{\tau}_A + e_{ij}^*(t_\ell)$
    Using $\text{data}^{(b)}$ fit two models. First, fit model (2.1) with the mean structure modeled by
    $\mu(t,x) = \mathbf{B}(t,x)^T\boldsymbol{\beta}$ and estimate $\widehat{\mu}_A^{(b)}(t,x)$. Second, fit model (2.1) with the mean model
    $\mu(t) = \mathbf{B}(t)^T\boldsymbol{\beta}$ and estimate $\widehat{\mu}_0^{(b)}(t)$. Calculate the test statistic $T^{(b)}$ using (4.2). end for
Approximate the tail probability $P(T > T_{obs})$ by the $p$-value $= B^{-1}\sum_{b=1}^{B} I(T^{(b)} > T_{obs})$,
where $T_{obs}$ is obtained using the original data and $I$ is the indicator function.
```

When the covariates X_{ij} and $\mathbf{Z}_{ij}$ do not depend on visit, i.e. $X_{ij} = X_i$ and $\mathbf{Z}_{ij} = \mathbf{Z}_i$, the algorithm can be modified along the lines of the 'bootstrap of the subject-level residuals' algorithm.

5. Application to physical activity data

Physical activity measured by wearable devices such as accelerometers provides new insights into the association between activity and health outcomes (Schrack *and others*, 2014); the complexity of the data also poses serious challenges to current statistical analysis. For example, accelerometers can record activity at the minute level for many days and for hundreds of individuals. Here we consider the physical activity data from the BLSA (Stone and Norris, 1966). Each female participant in the study wore the Actiheart portable physical activity monitor (Brage *and others* 2006) for 24 h a day for a number of consecutive days; visit duration varied among participants with an average of 4.7 days. Activity counts were measured in 1-min epochs and each daily activity profile has 1440 minute-by-minute activity counts measurements. Activity counts are proxies of activity intensity. Activity counts were log-transformed (more precisely, $x \to \log(1 + x)$) because they are highly skewed and then averaged in 30-min intervals. For simplicity, hereafter we refer to the log-transformed counts as log counts. Here we focus on 1580 daily activity profiles from a single visit of 332 female participants who have at least two days of data. Women in the study are aged between 50 and 90 years. Further details on the BLSA activity data can be found in Schrack *and others* (2014) and Xiao *and others* (2015).

Our objective is to conduct inference on the marginal effect of age on women's daily activity after adjusting for BMI. We model the mean log counts as $\mu(t, X_i) + Z_i\beta(t)$, where X_i and Z_i are the age and BMI of the ith woman during the visit, $\mu(t, x)$ is the baseline mean log counts for time t within the day for a woman who is x-years old, and $\beta(t)$ is the association of BMI with mean log counts for time t within the day. We test whether $\mu(t, x)$ varies solely with t. We use the proposed testing statistic, $T = \int\int\{\widehat{\mu}_A(t, x) - \widehat{\mu}_0(t)\}^2 \mathrm{d}t\mathrm{d}x$ as detailed in Section 4. The estimate $\widehat{\mu}_A(t, x)$ is based on the tensor product of 15 cubic basis functions in t and 7 cubic basis functions in x and the estimate $\widehat{\mu}_0(t)$ is based on 15 cubic basis functions. Goodness of fit is studied by comparing the observed data with simulated data from the fitted model; see Figure S6 of the supplementary materials available at *Biostatistics* online. Figure S1 of supplementary material available at *Biostatistics* online shows the null distribution of the statistic T. The observed test statistic is $T = 0.041$ and the corresponding p-value is less than 0.001 based on 1000 MC samples. This indicates that there is strong evidence that daily activity profiles in women vary with age.

Figure 1 displays the estimated baseline activity profile as a function of age, $\widehat{\mu}(t, x)$, using the average of all bootstrap estimates. The plot indicates that the average log counts is a decreasing function of age for most times during the day. Furthermore, it depicts two activity peaks, one around 12 pm and the other around 6 pm. The 6-pm peak seems to decrease faster with age, indicating that afternoon activity is more affected by age than morning activity. We use joint confidence band to evaluate the sampling variability of $\widehat{\mu}(t, x)$. The joint lower and upper 95% confidence limits based on methods described in Section 3 are displayed in the bottom plots of Figure 1; the plots show that across all ages, the estimated low average activity at night has relatively small variability while the estimated high-average activity during the day has relatively high variability. To visualize the results, we display the estimated activity profile for 60-years-old women, $\widehat{\mu}(t, 60)$, and the corresponding 95% joint confidence band in Figure 2. Figure S2 of supplementary material available at *Biostatistics* online displays the estimated association of BMI with mean log counts as a function of time of day; it suggests that women with higher BMI have less activity during the day and evening, albeit more activity at late night and in early morning.

5.1. *Validating the testing results via simulation study*

We conducted a simulation study designed to closely mimic the BLSA data structure. Specifically, we generated data from model (2.1) with $\mu(t, x) = \cos(2\pi t) + \delta\{\widehat{\mu}(t, x) - \cos(2\pi t)\}$, where $\widehat{\mu}(t, x)$ is the estimated mean log counts, and δ is a parameter quantifying the distance from the null and alternative hypotheses. When $\delta = 0$ the true mean profile $\mu(t, x) = \cos(2\pi t)$, whereas when $\delta = 1$ then $\mu(t, x) = \widehat{\mu}(t, x)$. The errors $\epsilon_{ij}(t)$ are generated with a covariance structure that closely mimics that of the residuals from the BLSA data. Specifically we use the model $\epsilon_{ij}(t) = u_i(t, x_i) + v_{ij}(t, x_i) + w_{ij}(t)$ and the associated model estimates from Xiao *and others* (2015), where $u_i(\cdot, x_i)$ and $v_{ij}(\cdot, x_i)$ are subject-specific and subject-

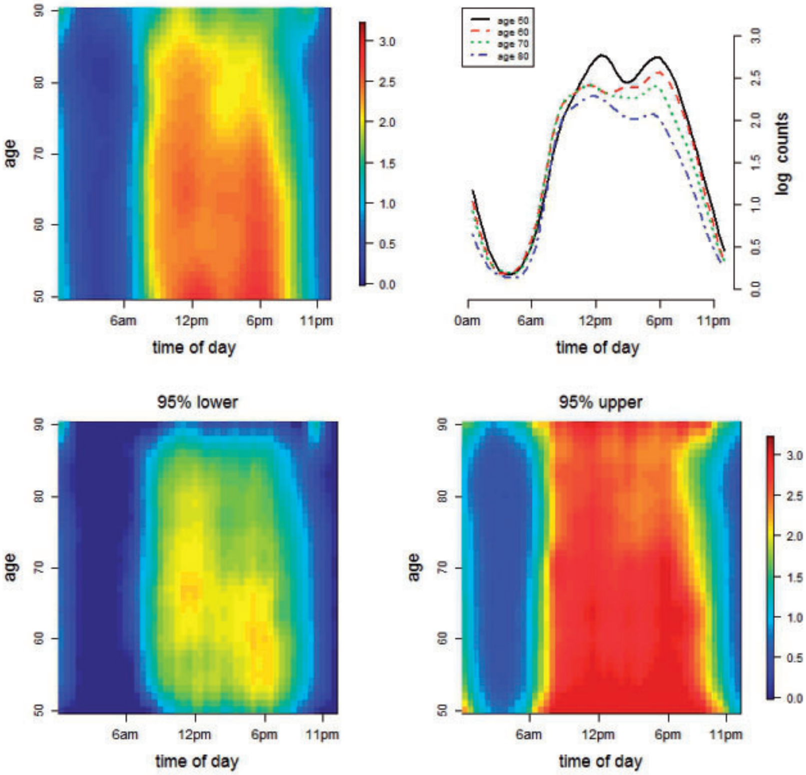

Fig. 1. Heat map of average of bootstrap estimates of log counts as a bivariate function of time of day and age (top left panel); average of bootstrap estimates of log counts for five different age groups (top right panel); and heat maps of joint confidence bands for the estimate in the top left panel (bottom panels). The legend on the right applies to both of the bottom plots.

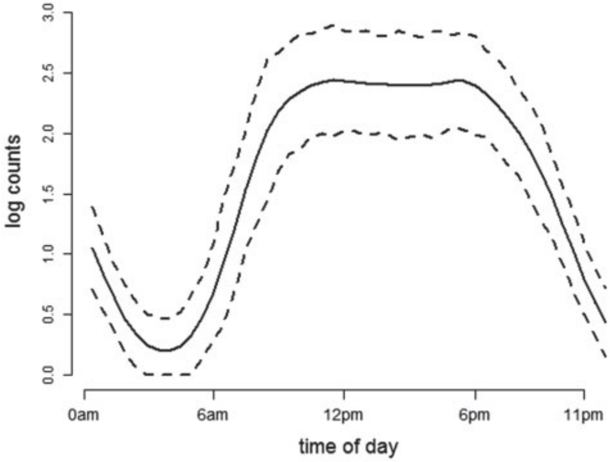

Fig. 2. Average of bootstrap estimates of log counts as a function of time of day at age 60 and the associated joint confidence bands.

Table 1. *Empirical type I error of the test statistic T based on the N_{sim} = 1000 MC samples; Mean function is $\mu(t,x) = cos(2\pi t)$, $\tau = 0$*

$\mu(t,x) = \cos(2\pi t)$, $\tau = 0$		
$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
0.06 (0.01)	0.11 (0.01)	0.16 (0.01)

Standard errors are presented in parentheses.

and visit-specific random processes with mean zero and $w_{ij}(t)$ is white noise. X_i and m_i are generated uniformly from $\{30,\ldots,90\}$ and $\{5,\ldots,9\}$, respectively. Sample size is set to be the number of female participants in the BLSA. Estimation is done exactly the same as in our data analysis. Table 1 shows the rejection probabilities in 1000 simulations when $\delta = 0$ and indicates that the empirical size is close to the nominal levels. Figure S3 of supplementary material available at *Biostatistics* online displays the power in 500 simulations, when $\delta > 0$. When the true $\mu(t,x)$ is the estimated bivariate mean log counts of the BLSA data, i.e. $\delta = 1$, the rejection probability reaches 1.

6. SIMULATION STUDY

We evaluate the performance of the proposed inferential methods. Data are simulated using the model (2.1) where $X_{ij} = X_i$, $Z_{ij} = Z_i$. The errors $\epsilon_{ij}(t)$ are generated from the model $\epsilon_{ij}(t) = \sum_{l=1}^{3} \xi_{ijl}\phi_l(t) + w_{ij}(t)$, where for each i and l the basis coefficients $\{\xi_{ijl}\}_j$ are generated from a multivariate normal distribution with mean zero and covariance $\text{cov}(\xi_{ijl}, \xi_{ij'l}) = \lambda_l \rho^{|V_{ij}-V_{ij'}|}$, where ρ is a correlation parameter and V_{ij} is the actual time of visit at which $Y_{ij}(\cdot)$ is observed; a similar dependence structure has been considered in simulation studies by Park and Staicu (2015) and Islam *and others* (2016). The residuals $w_{ij}(\cdot)$ are mutually independent with zero mean and variance σ^2. The number of repeated measures is fixed at $m_i = 5$, $(\lambda_1, \lambda_2, \lambda_3) = (3, 2, 1/3)$, and the functions $[\phi_1(t), \phi_2(t), \phi_3(t)] = [\sqrt{2}\cos(2\pi t), \sqrt{2}\sin(2\pi t), \sqrt{2}\cos(4\pi t)]$. The subject-specific covariates X_i and Z_i are generated from a Uniform[0, 1]. The grid of points $\{t_\ell : \ell = 1,\ldots,L\}$ is set as 101 equally spaced points in [0, 1]. The variance of the white noise process σ^2 is set to 5.33, which provides a signal to noise ratio SNR= $\sum_{l=1}^{3} \lambda_l/\sigma^2$ equal to 1.

We consider different combinations of the following factors: F1. number of subjects: (a) $n = 100$, (b) $n = 200$, and (c) $n = 300$; F2. bivariate mean function: (a) $\mu_1(t,x) = \beta_0 + \beta_t t + \beta_x x$ for $(\beta_0, \beta_t, \beta_x) = (5, 2, 3)$, (b) $\mu_2(t,x) = \beta_0 + \beta_t t + \beta_x x + \beta_{tx} tx$ for $(\beta_0, \beta_t, \beta_x, \beta_{tx}) = (5, 2, 3, 7)$, (c) $\mu_3(t,x) = \cos(2\pi t) + \beta_x x$ for $\beta_x = 3$, and (d) $\mu_4(t,x) = \cos(2\pi t) + \delta((x/4) - t)^3$ for $\delta = 0, 2, 4,$ and 6, *with/without* the addition of linear effect of nuisance covariate Z_i, i.e. $\tau = 0$ (no effect) and $\tau = 8$; lastly, F3. between-curves correlations: (a) $\rho = 0.2$ (weak) and (b) $\rho = 0.9$ (strong).

Confidence bands for model parameters are evaluated in two ways. First, we model the data by assuming the correct model and by evaluating the accuracy of the inferential procedures. Second, we model the data using a bivariate mean, $\mu(t,x)$, and evaluate the performance of the confidence bands of $\mu(t,x)$ for covering the true mean even when the true mean has a simpler structure, i.e. F2 i.(a)–(c). The results for the first case are included in Section B of the supplementary material available at *Biostatistics* online, whereas those for the second case are presented below, because in the BLSA we used bivariate nonparametric fitting. Estimation is done as detailed in Section 2. We use $d_t = d_x = 7$ cubic B-spline basis functions, and select the smoothing parameters via GCV; specifically, for the bivariate smooth, $d_t d_x = 49$ basis functions are used.

The performance of the pointwise and joint confidence bands is evaluated in terms of average coverage probability (ACP), and average length (AL) of the confidence intervals. Specifically, let $(\widehat{\mu}^{i_{sim},\, l}(t,x),\ \widehat{\mu}^{i_{sim},\, u}(t,x))$ be the $100(1-\alpha)\%$ pointwise confidence interval of $\mu(t,x)$ obtained at the i_{sim} Monte Carlo generation of the data, then

$$\mathrm{ACP}^{\text{point}} = \frac{1}{N_{sim} G_t G_x} \sum_{i_{sim}=1}^{N_{sim}} \sum_{g_t=1}^{G_t} \sum_{g_x=1}^{G_x} 1 \left\{ \mu(t_{g_t}, x_{g_x}) \in (\widehat{\mu}^{i_{sim},\, l}(t_{g_t}, x_{g_x}),\ \widehat{\mu}^{i_{sim},\, u}(t_{g_t}, x_{g_x})) \right\}$$

$$\mathrm{AL}^{\text{point}} = \frac{1}{N_{sim} G_t G_x} \sum_{i_{sim}=1}^{N_{sim}} \sum_{g_t=1}^{G_t} \sum_{g_x=1}^{G_x} |\widehat{\mu}^{i_{sim},\, l}(t_{g_t}, x_{g_x}) -\ \widehat{\mu}^{i_{sim},\, u}(t_{g_t}, x_{g_x}))|,$$

where $\{t_{g_t} : g_t = 1, \ldots, G_t\}$ and $\{x_{g_x} : g_x = 1, ..., G_x\}$ are equi-distanced grid points in the domains $\mathcal{D}_t$, and $\mathcal{D}_x$, respectively. Next, let $(\widehat{\mu}^{i_{sim},\, l}(t,x),\ \widehat{\mu}^{i_{sim},\, u}(t,x))$ be $100(1-\alpha)\%$ joint confidence interval. The AL is calculated as above, while the ACP is calculated as:

$$\mathrm{ACP}^{\text{joint}}_{\mu(t,x)} = \frac{1}{N_{sim}} \sum_{i_{sim}=1}^{N_{sim}} 1 \left\{ \mu(t_{g_t}, x_{g_x}) \in (\widehat{\mu}^{i_{sim},\, l}(t_{g_t}, x_{g_x}),\ \widehat{\mu}^{i_{sim},\, u}(t_{g_t}, x_{g_x})) :\ \text{for all } g_t, g_x \right\}.$$

The performance of the test statistic T is evaluated in terms of its size for the nominal levels $\alpha = 0.05$, 0.10, and 0.15, and power at $\alpha = 0.05$. The results for the size are based on $N_{sim} = 1000$ MC samples, while the results for ACP and AL of the confidence bands, and power of the test are based on $N_{sim} = 500$ MC samples. For each MC simulation we use $B = 300$ bootstrap samples.

Table 2 shows the ACP and AL for the 95% confidence bands based on the bootstrap of subject-level residuals when the sample size $n = 100$ and when $\mu(t,x)$ is modeled nonparametrically regardless of the true mean structure; the results for other nominal coverages (85% and 90%) are included in Section A of the supplementary material available at *Biostatistics* online. Overall, the pointwise/joint confidence bands achieve the nominal coverage for all of the mean structures considered. The confidence bands tend to be wider when the between-curves correlation is strong ($\rho = 0.9$).

We also investigate the performance of the confidence band when the correct structure of $\mu(t,x)$ is used; the corresponding results for the bootstrap of subject-level residuals and observations are included in Section B and Section C of the supplementary material available at *Biostatistics* online. The results show the good coverage of the pointwise/joint confidence bands based on the bootstrap of residuals by subjects for all of the mean structures considered. The bootstrap of observations by subjects leads to equally good coverage when the true effect of the covariate X is linear (cases F2 i.(a)–(c)), whereas it leads to slight under–coverage when the true effect of X is nonlinear (case F2 i.(d)). However, in the case of a visit-varying covariate X_{ij} the joint confidence band maintains nominal coverage even when the effect of X is nonlinear; see Table S9 of the supplementary material available at *Biostatistics* online. These results indicate that for a time-invariant covariate, X_i, the bootstrap of subject-level residuals is narrower and has better coverage. In terms of computational cost, fitting a nonparametric model is much slower than fitting a parametric model. For example when the true mean F2 i. (c) is used to generate the data, fitting a nonparametric model for $B = 300$ bootstrap samples takes 337 s whereas the same procedure for a parametric model takes 50 s; the results are based on 100 MC samples on a computer with a 3.60 Hz processor.

Table 2. *Simulation results for* 95% *confidence bands based on the bootstrap of subject-level residuals when a nonparametric bivariate function is fitted for* $\mu(t,x)$*; results are based on* 500 *MC samples*

Case	True mean function	ρ	$\text{ACP}^{\text{point}}$		AL^{point}		$\text{ACP}^{\text{joint}}_{\mu(t,x)}$		$\text{AL}^{\text{joint}}_{\mu(t,x)}$	
(a)	$\mu(t,X) = 5 + 2t + 3X$	0.20	0.94	(< 0.01)	1.65	(0.01)	0.94	(0.01)	3.22	(0.01)
		0.90	0.94	(< 0.01)	2.17	(0.02)	0.93	(0.01)	4.24	(0.01)
	$\tau = 8$	0.20	0.93	(0.01)	0.14	(< 0.01)				
		0.90	0.93	(0.01)	0.14	(< 0.01)				
(b)	$\mu(t,X) = 5 + 2t + 3X + 7tX$	0.20	0.94	(< 0.01)	1.65	(0.01)	0.94	(0.01)	3.22	(0.01)
		0.90	0.94	(< 0.01)	2.17	(0.02)	0.93	(0.01)	4.24	(0.01)
	$\tau = 8$	0.20	0.93	(0.01)	0.14	(< 0.01)				
		0.90	0.93	(0.01)	0.14	(< 0.01)				
(c)	$\mu(t,X) = \cos(2\pi t) + 3X$	0.20	0.94	(< 0.01)	1.65	(0.01)	0.93	(0.01)	3.23	(0.01)
		0.90	0.94	(< 0.01)	2.18	(0.02)	0.93	(0.01)	4.25	(0.01)
	$\tau = 8$	0.20	0.93	(0.01)	0.14	(< 0.01)				
		0.90	0.93	(0.01)	0.14	(< 0.01)				
(d)	$\mu_4(t,X)$	0.61	0.94	(< 0.01)	1.65	(0.01)	0.93	(0.01)	3.23	(0.01)
		0.90	0.94	(< 0.01)	2.18	(0.02)	0.93	(0.01)	4.26	(0.01)
	$\tau = 8$	0.20	0.93	(0.01)	0.14	(< 0.01)				
		0.90	0.94	(0.01)	0.14	(< 0.01)				

Standard errors are presented in parentheses.

Table 3. *Empirical Type I error of the test statistic* T *based on the* $N_{sim} = 1000$ *MC samples*

		$\mu(t,x) = \cos(2\pi t)$, $\tau = 0$					
		$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.15$	
$n = 100$	$\rho = 0.2$	0.08	(0.01)	0.14	(0.01)	0.21	(0.01)
	$\rho = 0.9$	0.09	(0.01)	0.14	(0.01)	0.20	(0.01)
$n = 200$	$\rho = 0.2$	0.07	(0.01)	0.13	(0.01)	0.17	(0.01)
	$\rho = 0.9$	0.08	(0.01)	0.12	(0.01)	0.18	(0.01)
$n = 300$	$\rho = 0.2$	0.06	(0.01)	0.11	(0.01)	0.16	(0.01)
	$\rho = 0.9$	0.06	(0.01)	0.12	(0.01)	0.16	(0.01)
		$\mu(t,x) = \cos(2\pi t)$, $\tau = 8$					
		$\alpha = 0.05$		$\alpha = 0.10$		$\alpha = 0.15$	
$n = 100$	$\rho = 0.2$	0.07	(0.01)	0.15	(0.01)	0.20	(0.01)
	$\rho = 0.9$	0.08	(0.01)	0.15	(0.01)	0.21	(0.01)
$n = 200$	$\rho = 0.2$	0.07	(0.01)	0.13	(0.01)	0.17	(0.01)
	$\rho = 0.9$	0.08	(0.01)	0.12	(0.01)	0.18	(0.01)
$n = 300$	$\rho = 0.2$	0.06	(0.01)	0.11	(0.01)	0.16	(0.01)
	$\rho = 0.9$	0.06	(0.01)	0.12	(0.01)	0.16	(0.01)

Standard errors are presented in parentheses.

Table 3 shows the empirical size of the proposed testing procedure for testing $H_0 : \mu(t,x) = \mu_0(t)$, where $\mu_0(\cdot)$ is a smooth effect depending on t only. Results indicate that, as sample size increases, the size of the test gets closer to the corresponding nominal levels. In the simulation settings considered, the test attains the correct sizes with sample size $n \geq 300$, which is the case in our motivating BLSA data application. Including an additional covariate in the model seems to have no effect on the performance of the testing procedure. Figure S4 of supplementary material available at *Biostatistics* online illustrates the power curves, when the true mean structure deviates from the null hypothesis. It presents the power as a function of the deviation from the null that involves both t and x, $\mu(t,x) = 2\cos(2\pi t) + \delta(x/4 - t)^3$. Here δ quantifies the departure from the null hypothesis. As expected, for $\delta > 0$ rejection probabilities increase as the departure from the null hypothesis increases, irrespective of the direction in which it deviates. As expected, rejection probabilities increase with the sample size. Our investigation indicates that the strength of the correlation between the functional observations corresponding to the same subject affect the rejection probability: the weaker the correlation, the larger the power. There is no competitive testing method available for this null hypothesis. Lastly we conducted a simulation study to evaluate the robustness of the proposed methods to non-Gaussian error distributions and obtained similar results with those from the Gaussian case; see Section D of the supplementary material available at *Biostatistics* online.

Supplementary Material

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

Acknowledgments

Conflict of Interest: None declared.

Funding

NSF (DMS 1007466 and DMS 1454942 to A.M.S.); NIH (R01 NS085211 and R01 MH086633 to A.M.S); NIH (R01 NS085211, R01 NS060910, R01 HL123407), NIA contracts (HHSN27121400603P and HHSN27120400775P) to C.M.C Data for these analyses were obtained from the Baltimore Longitudinal Study of Aging, performed by the National Institute on Aging.

References

Aston, J. A., Chiou, J.-M. and Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**, 297–317.

Baladandayuthapani, V., Mallick, B. K, Young H., Mee, L., Joanne R. T., Nancy D. and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics* **64**, 64–73.

Brage, S., Brage, N., Ekelund, U., Luan, J., Franks, P. W., Froberg, K. and Wareham, N. J. (2006). Effect of combined movement and heart rate monitor placement on physical activity estimates during treadmill locomotion and free-living. *European Journal of Applied Physiology* **96**, 517–524.

Cao, G., Yang, L. and Todem, D. (2012). Simultaneous inference for the mean function based on dense functional data. *Journal of Nonparametric Statistics* **24**, 359–377.

Chen, K. and Müller, H.-G. (2012). Modeling repeated functional observations. *Journal of the American Statistical Association* **107**, 1599–1609.

CRAINICEANU, C. M., STAICU, A.-M. AND DI, C.-Z. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association* **104**, 1550–1561.

CRAINICEANU, C. M., STAICU, A.-M., RAY, S. AND PUNJABI, N. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine* **31**, 3223–3240.

CUEVAS, A., FEBRERO, M. AND FRAIMAN, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis* **51**, 1063–1074.

DEGRAS, D. A. (2011). Simultaneous confidence bands for nonparametric regression with functional data. *Statistica Sinica*, **21**, 1735–1765.

Delgado M. A. and Manteiga, W. G. (2001). Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics*, **29**, 1469–1507.

DI, C.-Z., CRAINICEANU, C. M., CAFFO, B. S AND PUNJABI, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics* **3**, 458.

EFRON, B. AND TIBSHIRANI, R. J. (1994). *An Introduction to the Bootstrap*. Boca Raton, Florida: CRC press.

FAN, Y. and LI, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms, *Econometrica* **64**, 865–890.

FARAWAY, J. J. (1997). Regression analysis for a functional response. *Technometrics* **39**, 254–261.

FITZMAURICE, G. M., LAIRD, N. M. AND WARE, J. H. (2012). *Applied Longitudinal Analysis*, Volume 998. Hoboken, NJ: John Wiley & Sons.

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

GU, J., LI, D. AND LIU, D. (2007). Bootstrap non-parametric significance test. *Journal of Nonparametric Statistics* **19**, 215–230.

HALL, P. and HOROWITZ, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions. *The Annals of Statistics* **41**, 1892–1921.

HALL, P., LI, Q. AND RACINE, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *The Review of Economics and Statistics* **89**, 784–789.

HÄRDLE, W. AND BOWMAN, A. W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association* **83**, 102–110.

HORVÁTH, L., KOKOSZKA, P. AND REEDER, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 103–122.

ISLAM, M. N., STAICU, A.-M. AND VAN HEUGTEN, E. (2016). Longitudinal dynamic functional regression. *arXiv preprint arXiv:1611.01831*.

IVANESCU, A. E., STAICU, A.-M., SCHEIPL, F. AND GREVEN, S. (2015). Penalized function-on-function regression. *Computational Statistics* **30**, 539–568.

JIANG, C.-R. and WANG, J.-L. (2011). Functional single index models for longitudinal data. *The Annals of Statistics* **39**, 362–388.

LAIRD, N. M. AND WARE, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

LAVERGNE, P. AND VUONG, Q. (2000). Nonparametric significance testing. *Econometric Theory* **16**, 576–601.

LIANG, K.-Y. AND ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

MA, S., YANG, L. AND CARROLL, R. J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistica Sinica* **22**, 95.

MARX, B. D. AND EILERS, P. H. (2005). Multidimensional penalized signal regression. *Technometrics* **47**, 13–22.

MORRIS, J. S., ARROYO, C., COULL, B. A., RYAN, L. M., HERRICK, R. and GORTMAKER, S. L. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *Journal of the American Statistical Association* **101**, 1352–1364.

MORRIS, J. S, BALADANDAYUTHAPANI, V., HERRICK, R. C, SANNA, P. AND GUTSTEIN, H. (2011). Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *The Annals of Applied Statistics* **5**, 894.

MORRIS, J. S. AND CARROLL, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 179–199.

MORRIS, J. S., VANNUCCI, 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.

PARK, S. Y. AND STAICU, A.-M. (2015). Longitudinal functional data analysis. *Stat* **4**, 212–226.

POLITIS, D. N. AND ROMANO, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical association* **89**, 1303–1313.

RUPPERT, D., WAND, M. P AND CARROLL, R. J. (2003). *Semiparametric regression*, Number 12. Cambridge, UK: Cambridge university press.

SCHEIPL, F., STAICU, A.-M. AND GREVEN, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics* **24**, 477–501.

SCHRACK, J. A., ZIPUNNIKOV, V., GOLDSMITH, J., BAI, J., SIMONSICK, E. M., CRAINICEANU, C. M. AND FERRUCCI, L. (2014). Assessing the "physical cliff": detailed quantification of aging and patterns of physical activity. *The Journals of Gerontogoly Series A: Biological Sciences and Medical Sciences.* **69**, 973–979.

SERBAN, N., STAICU, A.-M. AND CARROLL, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics* **69**, 903–913.

SHEN, Q. AND FARAWAY, J. (2004). An f test for linear models with functional responses. *Statistica Sinica* **14**, 1239–1258.

SHOU, H., ZIPUNNIKOV, V., CRAINICEANU, C. M. AND GREVEN, S. (2015). Structured functional principal component analysis. *Biometrics* **71**, 247–257.

STAICU, A.-M., CRAINICEANU, C. M. AND CARROLL, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics* **11**, 177–194.

STAICU, A.-M., LAHIRI, S. N. AND CARROLL, R. J. (2015). Significance tests for functional data with complex dependence structure. *Journal of Statistical Planning and Inference* **156**, 1–13.

STAICU, A.-M., LI, Y., CRAINICEANU, C. M AND RUPPERT, D. (2014). Likelihood ratio tests for dependent data with applications to longitudinal and functional data analysis. *Scandinavian Journal of Statistics* **41**, 932–949.

STONE, J. L. AND NORRIS, A. H. (1966). Activities and attitudes of participants in the Baltimore Longitudinal Study. *Journal of Gerontology* **21**, 575–580.

WAHBA, G. (1990). *Spline Models for Observational Data*, Volume 59. Philadelphia, PA: Siam.

WOOD, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62**, 1025–1036.

XIAO, L., HUANG, L., SCHRACK, J. A, FERRUCCI, L., ZIPUNNIKOV, V. AND CRAINICEANU, C. M. (2015). Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics* **16**, 352–367.

XIAO, L., LI, Y. AND RUPPERT, D. (2013). Fast bivariate p-splines: the sandwich smoother. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**, 577–599.

XIAO, L., ZIPUNNIKOV, V., RUPPERT, D. AND CRAINICEANU, C. (2016). Fast covariance estimation for high-dimensional functional data. *Statistics and Computing* **26**, 409–421.

ZHANG, J.-T. and CHEN, J. (2007). Statistical inferences for functional data. *The Annals of Statistics* **35**, 1052–1079.

ZHANG, L., BALADANDAYUTHAPANI, V., ZHU, H., BAGGERLY, K. A, MAJEWSKI, T., CZERNIAK, B. A. AND MORRIS, J. S. (2016). Functional car models for large spatially correlated functional datasets. *Journal of the American Statistical Association* **111**, 772–786.

ZHU, H., BROWN, P. J AND MORRIS, J. S. (2011). Robust, adaptive functional regression in functional mixed model framework. *Journal of the American Statistical Association* **106**, 1167–1179.

[*Received June 16, 2016; revised April 9, 2017; accepted for publication May 7, 2017*]