FUNCTIONAL DATA ANALYSIS METHODS FOR LARGE SCALE PHYSICAL ACTIVITY STUDIES

by

Erjia Cui

A dissertation submitted to Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

March 2023

Abstract

Wearable devices have been increasingly deployed in large epidemiological and clinical studies to provide objective measures of human activity in the free-living environment. The statistical analysis of these wearable device data collected in large cohort studies is challenging due to its size, dimension, and complexity. This thesis presents three novel functional data analysis methods, each of which addresses an important problem in the large cohort physical activity studies. An additive functional Cox model is proposed to flexibly quantify the association between functional predictors and survival outcomes. A fast multilevel functional principal component analysis method is proposed to perform variability decomposition for functional data measured at multiple visits. A fast univariate inferential approach is proposed to model the association between predictors and longitudinal functional data.

Thesis Committee

Primary Readers

Ciprian M. Crainiceanu (Primary Advisor)
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Lingxin Hao (Chair)
Benjamin H. Griswold III Professor
Department of Sociology
Johns Hopkins Krieger School of Arts & Sciences

Mei-Cheng Wang
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Jennifer Schrack
Associate Professor
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

Vadim Zipunnikov
Associate Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Adam Spira
Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor, Ciprian Crainiceanu, for his unwavering support and guidance throughout my doctoral studies. While he is undoubtedly a world-class researcher, he is also a kind and compassionate individual. His invaluable insights, constructive feedback, and constant encouragement, together with his passion for wearable device research, played a major role in shaping the direction of my research and helping me stay motivated through my PhD journey. I really enjoy the WIT seminar and am proud to be a part of our group. I am also grateful to Andrew Leroux for his invaluable guidance and support when I embarked on this journey. I really learned a lot from him on how to do reproducible innovative research.

I wish to thank the other members of my thesis committee, Mei-Cheng Wang, Lingxin Hao, Jennifer Schrack, Vadim Zipunnikov, and Adam Spira, for their valuable feedback that helped me refine my research over the years. Special thanks to Mei-Cheng for her kindness, patience, and deep insights, and for organizing the SLAM seminar. I really enjoyed these presentations and hope to attend more in the future.

Furthermore, I wish to express my appreciation to the mentors and collaborators who supported me outside the Johns Hopkins University. My sincere thanks go to Raymond J. Carroll at Texas A&M University and David Ruppert at Cornell University for their insights, guidance and support. I am also deeply grateful to work with Luo Xiao at North Carolina State University, Irina Gaynanova at Texas A&M University, and Philip Tzvi Reiss at the University of Haifa. I have learned a lot from all these amazing researchers, and I look forward to learning even more in the future.

I would like to acknowledge the contributions of all my friends who made my time at Johns Hopkins University a memorable experience. It has been a pleasure to have crossed paths with you at various stages of this journey, and I wish you all the best in your future endeavors.

Finally, I would like to extend my sincere gratitude to my family for their unconditional love. I want to express my deepest appreciation to my mother, Shuguang Jiao, who has been a constant source of inspiration and support in my life. Her unwavering encouragement and unwavering belief in me have helped me to overcome challenges, pursue my dreams, and discover the beauty of this world. Heartfelt thanks to Han Wang, whose admirable personality and kindness not only inspire me to work harder, but also let me realize the true meaning of life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past two decades, wearable devices have been increasingly deployed in large epidemiological and clinical studies to provide objective measures of human activity in the free-living environment. Specifically, each participant of the National Health and Nutrition Examination Survey (NHANES) 2003-2004 and 2005-2006 waves was asked to wear a hip-worn physical activity monitor (Actigraph AM-7164) for seven consecutive days. The accelerometry data were collected, processed, and released as minute-level activity counts (AC), a measure of the physical activity intensity. NHANES also collected demographic, socioeconomic, and health-related information as well as examination data, including medical, dental, and physiological measurements.

The statistical analysis of this data set is challenging because: (i) it contains many study participants and many covariates (large n large p); (ii) each study participant has multiple days of data, which induces a multilevel structure; (iii) for each day physical activity data are high-dimensional because it contains 1440 observations; (iv) the activity counts exhibit non-stationary behavior across time of the day. In the thesis, we propose three novel functional

data analysis (FDA) methods to model this NHANES data set and answer important scientific questions.

In Chapter 2, the research question is "what is the association between the high dimensional baseline objective measurements of physical activity (activity counts for every minute of the day) and time to death?". To answer this question, we propose an Additional Functional Cox Model (AFCM) (Cui, Crainiceanu, and Leroux, 2021). The model extends the linear functional Cox model to the nonparametric functional Cox model by replacing the functional linear predictor $\int_0^1 X_i(t)\beta(t)dt$ with the functional nonparametric predictor $\int_0^1 F\{X_i(t), t\}dt$. An important finding is that high levels of physical activity at night and low levels of physical activity during the day are associated with higher risk of all-cause mortality. We have validated these findings in two other large data sets that contain objectively measured physical activity data: NHANES 2011-2014 and UK Biobank. The paper is published in the *Journal of Computational and Graphical Statistics*.

In Chapter 3, the research question is "given the multilevel structure of the data (multiple days of physical activity), what is the structure of the within- and between- study participants variability of physical activity?". To answer this question, we propose Fast Multilevel Functional Principal Component Analysis (Fast MFPCA) (Cui et al., 2022b), which is two orders of magnitude faster than MFPCA and achieves similar estimation accuracy. To facilitate the use of this method, we have contributed the function `mfpca.face()` to the R package `refund` available in CRAN. A theoretical study of the fast MFPCA approach is also provided. The paper is published in the *Journal of*

Computational and Graphical Statistics.

In Chapter 4, the research question is "what is the association between the patterns of daily activity and covariates (e.g., age, sex, day of the week)?". To answer this question, we propose a Fast Univariate Inference (FUI) method for longitudinal functional models (Cui et al., 2022a). We introduce a novel method for model estimation and inference by first fitting massively univariate mixed models and then smoothing along the functional domain. Despite the rich literature on functional mixed models, the new approach is the only one that is computationally feasible for the NHANES dataset, which has 1440 observations per day for more than 8000 days. The paper is published in the *Journal of Computational and Graphical Statistics*.

References

Cui, Erjia, Ciprian M Crainiceanu, and Andrew Leroux (2021). "Additive functional Cox model". In: *Journal of Computational and Graphical Statistics* 30.3, pp. 780–793.

Cui, Erjia, Ruonan Li, Ciprian M Crainiceanu, and Luo Xiao (2022b). "Fast Multilevel Functional Principal Component Analysis". In: *Journal of Computational and Graphical Statistics*, pp. 1–12.

Cui, Erjia, Andrew Leroux, Ekaterina Smirnova, and Ciprian M Crainiceanu (2022a). "Fast univariate inference for longitudinal functional models". In: *Journal of Computational and Graphical Statistics* 31.1, pp. 219–230.

Chapter 2

Additive Functional Cox Model

2.1 Introduction

We introduce a class of nonparametric additive functional Cox regression models for quantifying the association between a time to event outcome and functional covariates. This expands the rich literature on survival analysis by allowing for one or multiple functional covariates. It also expands the sparser literature on functional data analysis with survival outcomes by allowing a more flexible association between the functional covariate and time-to-event outcome. The approach is fully reproducible, fast, is implemented in `R` (R Core Team, 2019), and can be used with minimal effort on personal laptops. Our work is motivated by the study of the association between time to death and physical activity (PA). PA has long been known to confer health benefits (Cooper et al., 2017) and has been associated with reduced risk of mortality (Schmid, Ricci, and Leitzmann, 2015; Matthews et al., 2016). However, until the relatively recent development and adoption of wearable accelerometers, researchers relied on crude, inaccurate, and biased measures obtained from

self-report questionnaires (Sallis and Saelens, 2000; Silsbury, Goldsmith, and Rushton, 2015). In contrast, accelerometers offer an unintrusive, continuous, and unbiased alternative to objectively measure PA over the course of several days, weeks, or even months. For these reasons they have been deployed in many large epidemiologic studies; see, for example, Troiano et al., 2008; Schrack et al., 2014; Bai et al., 2016; Doherty et al., 2017.

Here we are interested in quantifying the effect of timing and volume of PA on all-cause mortality in the National Health and Nutrition Examination Survey (NHANES). NHANES is a nationally representative study conducted by the Centers for Disease Control (CDC) to assess the health and nutritional status of adults and children in the United States. Participants were selected for inclusion according to the CDC sample design (Mirel et al., 2013) and assigned a survey weight based on the proportion the individual represents in the US population. Broadly, the NHANES data can be divided into three main categories: (1) questionnaire data, including responses to demographic, socioeconomic, dietary and health-related questions; (2) examination and laboratory data, including results of medical, dental, physiological measurements and laboratory tests; (3) accelerometer-measured PA. The processed NHANES 2003-2004 and 2005-2006 data are available in the `R` package `rnhanesdata` (Leroux et al., 2019a).

Specifically, the high resolution PA was measured by hip-worn accelerometers in the NHANES 2003-2004 and 2005-2006 waves. Each eligible participant was asked to wear the device for 7 consecutive days, and data were summarized in minute-level activity counts (a proprietary measure of PA

intensity level). The minute-level activity counts are then transformed as $\text{AC} \rightarrow \log(1 + \text{AC})$ resulting in the log-transformed activity counts (LAC), which reduces the severe skewness of original data and is an appropriate measure of PA volume of lower levels of physical activity which have been adopted in the physical activity research literature (Varma et al., 2017; Varma et al., 2018). A sample of recorded minute-level LAC for one individual in NHANES is shown in the upper-left of Figure 2.1. Data are displayed on rows, where each row corresponds to a day of the week, where higher values correspond to more intense PA.

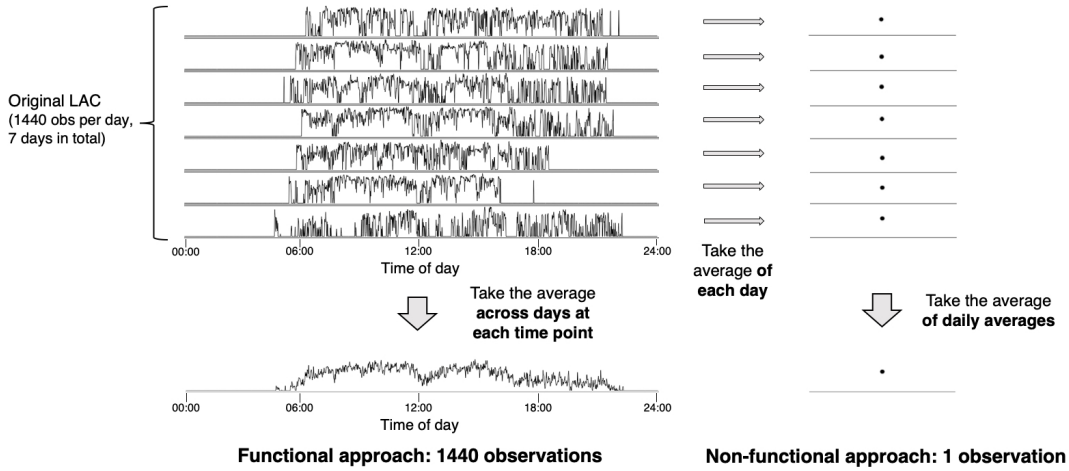


Figure 2.1: A sample of minute-level LAC for one individual in NHANES shown in the upper left and two summarising approaches. The right panel illustrates the traditional summarising approach, which calculates a daily average (or sum) of LAC and then averages these means (or totals) across days. The bottom panel illustrates a less aggressive summarising approach, where LAC are averaged at each time point across days.

There are many different approaches for compressing and using these high dimensional accelerometry data. The most popular is to calculate a daily average (or sum) of LAC and then average these means (or totals) across days.

This is illustrated in the right panel of Figure 2.1 by the horizontal arrows labeled "Take the average of each day" and by the arrow labeled "Take the average of daily averages". The bottom panel illustrates a less aggressive summarising approach, where LAC are averaged at each time point across days. We will use this approach to create our functional covariates, which loses information on day-to-day variability in PA but retains substantially more information than traditional averaging over time of days and day. The pre-processing steps used to create the functional covariates is described in detail in Section 2.3.1.

Because NHANES can be linked to the National Death Index (NDI) released by National Center for Health Statistics (NCHI), it provides an unique opportunity to investigate the association between accelerometer-based PA measurements and time to death in a nationally representative sample. Figure 2.2 provides the intuition behind the problem and describes the data structure. PA data measured as minute-level LAC averaged over available days are shown for six study participants as a function of time of day. For each individual the data contain sociodemographic factors (age, race, employment status, education attainment, poverty-income ratio), health factors (self-reported overall health, smoking status, alcohol consumption, body mass index, mobility difficulty), and disease indicators (diabetes, coronary heart disease, congestive heart failure, stroke, cancer, systolic blood pressure, total cholesterol). For each study participant we display only their age, though much more additional information is available. The right panel in Figure 2.2 displays the mortality information. For example, the first study participant,

who was 83 years old at the time the PA data were collected was deceased 2 years later (red horizontal line with a $\times$ symbol at the end to indicate a death event). The fourth study participant was 70 years old when the PA data were collected and was still alive 9.08 years later, the last time data were available for this individual (black horizontal line with a $\bullet$ symbol at the end).

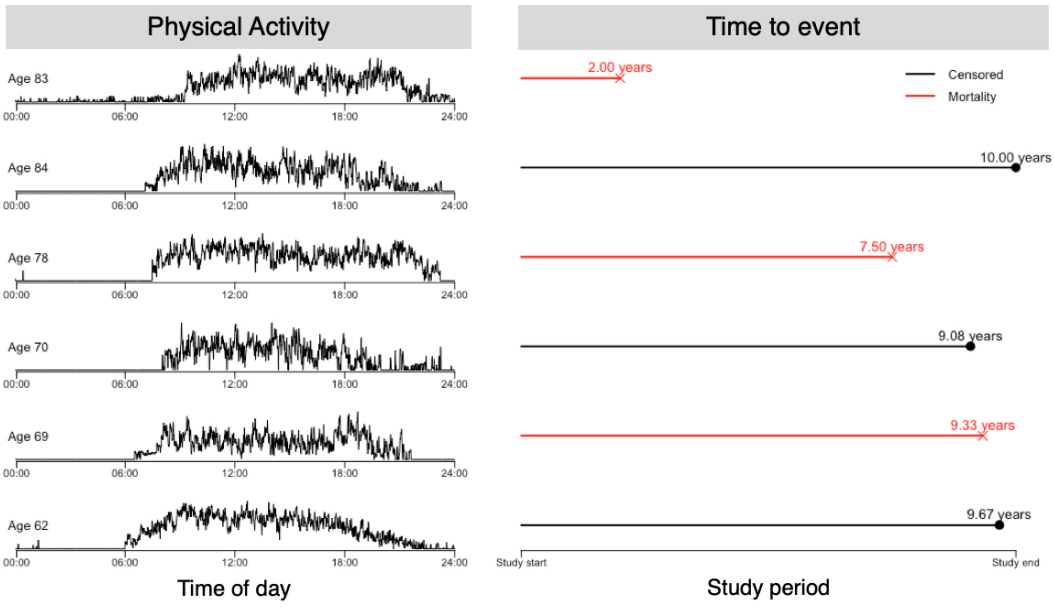


Figure 2.2: Physical activity and survival data of six study participants in NHANES. Each function represents the minute-level average LAC over the available days of valid data for that study participant. The age of the study participant is shown together with their mortality status (red for dead, black for alive) and the follow up time.

In the NHANES 2003-2004 and 2005-2006 study, accelerometry data were collected from a total of 14631 study participants. For the purpose of this analysis, we exclude participants who: (1) were younger than 50 years of age, or 85 and older at the time they wore the accelerometer (10859 participants); (2) had fewer than 3 days of data with at least 10 hours of estimated wear time

or were deemed by NHANES to have poor quality data (517 participants); (3) had missing covariates of interest, including age, employment status, educational attainment, poverty-income ratio, body mass index, self-reported overall health, coronary heart disease, congestive heart failure, stroke, cancer, diabetes, smoking status and alcohol consumption (436 participants); or (4) had missing mortality information (3 participants). The final data contained 2816 participants with 659 deaths in the first 10 years after the time PA data were collected. Individuals with observed mortality beyond 10 years are administratively censored at 10 years in our application.

Surprisingly, there are few published methods for analyzing this type of data. In particular, Gellar et al., 2015, Qu, Wang, Wang, et al., 2016 and Kong et al., 2018 proposed different versions of the "linear functional Cox model", which included a linear functional term of the form $\int_{\mathcal{S}} X_i(s)\beta(s)ds$ in the log-hazard expression to capture the effect of the functional covariate $\{X_i(s) : s \in \mathcal{S}\}$. In practice we only observe $X_i(s)$ at a finite number of points. In our example, $X_i(s)$ is the smoothed minute-level average log-transformed activity count (smoothed LAC) for study participant i at time s of day, and the domain $\mathcal{S}$ is midnight to midnight. We introduce three important methodological innovations: (1) extending the linear functional form to $\int_{\mathcal{S}} F\{s, X_i(s)\}ds$, where $F(\cdot, \cdot)$ is an unspecified smooth function, as done by McLean et al., 2014 for generalized linear models; (2) introducing a flexible class of transformation functions for $X_i(\cdot)$ to account for the complexity of the NHANES accelerometry data; and (3) providing necessary assumptions and constraints to ensure the estimability and identifiability of the functional coefficient. We implement

our method using easy-to-use software and provide a vignette which provides a detailed introduction of our model estimation procedure.

The remainder of the paper is organized as follows. Section 2.2 introduces the model and functional data transformations. Section 2.3 provides the results of the model applied to NHANES and interpretations. Section 2.4 proposes a simulation framework for both functional covariates and survival data. Section 2.5 summarizes the major findings and provides conclusions.

2.2 Methods

2.2.1 Model Setup

Motivated by the data structure illustrated in Figure 2.2, we model the log hazard function for $i = 1, ..., N$ study participants in the presence of independent right censoring. Denote the mortality event time as T_i and censoring time as C_i. We observe $Y_i = \min(T_i, C_i)$ and the event indicator $\Delta_i = I(T_i \leq C_i)$ for each study participant, where $I(\cdot)$ is the indicator function. The censoring time, C_i, is assumed to be independent of the event time, T_i, conditional on covariates. Suppose that at baseline we observe for each study participant p scalar covariates $\boldsymbol{Z}_i \in \mathbb{R}^p$, and a functional covariate $\boldsymbol{X}_i = \{X_i(s) : s \in \mathcal{S}\}$. The framework extends to multiple functional predictors, but we use single functional predictor for presentation purposes. We assume that $\boldsymbol{X}_i$ takes values on a compact interval, and denote the partial information in the functional covariate up to s as $X_i^{\mathcal{P}}(s) = \{X_i(u) : u \leq s\}$. Hereafter we refer to this partial information as the "history" of the functional covariate, though this "history" is distinct from the notion of time as it relates to the survival process.

Although the functional domain in our application is time of day, in other applications it may be, for example, space or some other argument. Using this notation we propose the following additive functional Cox model

$$\log \lambda_i(t|\boldsymbol{Z}_i, \boldsymbol{X}_i) = \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \int_{\mathcal{S}} F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\} ds \,, \tag{2.1}$$

where $F(\cdot, \cdot)$ is an unspecified bivariate twice differentiable function; see McLean et al., 2014 for a similar approach in the context of outcomes from the exponential family. We discuss the identifiability of $F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\}$ in Section 2.2.3.

This formulation allows the hazard function to vary smoothly with respect to both the functional domain and the value of the functional covariate, relaxing the assumption of linearity in the linear functional Cox model. We will show that this is important in our application, where activity during the day and night have opposite effects on the hazard of mortality. Another innovation is to allow for a known subject- and domain-specific transformation, $h_{is}[X_i^{\mathcal{P}}(s)]$, of the partial history of the functional covariate up to time s of the functional domain, $X_i^{\mathcal{P}}(s)$. The main reason for considering transformations in the NHANES accelerometry data is that its structure is highly complex and exhibits substantial skewness, missingness, and heterogeneity within- and between-study participants. In addition, transformations of the functional data can be used to improve the estimability of $F(\cdot, \cdot)$; see our discussion in Section 2.2.3.

2.2.2 Transformations of the Functional Covariate

We consider two types of transformations, one that is domain-specific, $h_s(\cdot)$, and one that is subject/domain-specific, $h_{is}(\cdot)$. The difference is that the second type of transformation depends on the subject, i, in addition to the domain, s.

2.2.2.1 Domain-specific Transformations

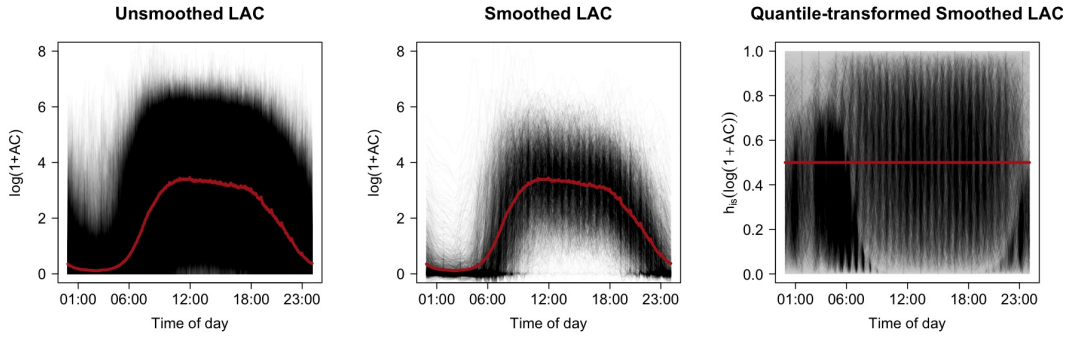


Figure 2.3: Distribution of the transformed minute-level LAC of all selected participants in the NHANES study, including unsmoothed (left), smoothed (middle), and smoothed + quantile transformation (right). The white top-left regions in the two left panels indicate the lack of high activity counts during the night.

The NHANES study activity data (minute-level LAC) shown in the left panel of Figure 2.3 indicates that during the night (1AM-4AM), PA measurements are much smaller than during the day. Therefore, estimating the function $F(\cdot,\cdot)$ on the entire rectangular domain $[0,24] \times [0,8]$ is nearly impossible. Here 24 stands for the number of hours in a day and 8 stands for an upper bound on the LAC. Indeed, there is basically no data in the $[1,6] \times [5,8]$ sub-domain. Therefore, estimates will be entirely driven by extrapolation of the smooth function $F(\cdot,\cdot)$ that borrows information from regions that are too

far away to provide meaningful information. Fundamentally, the problem is that the function $F(\cdot,\cdot)$ cannot be well estimated in areas where there is little or no data. This is a limitation of the model and the primary motivation for our emphasis on transformation functions. The middle panel in Figure 2.3 displays the same data after smoothing each individual curve. Results indicate that the data sparsity becomes even more serious in certain parts of the domain of $F(\cdot,\cdot)$. Below we propose two classes of domain-specific transformations to address this issue.

2.2.2.1.1 Quantile transformation. The first domain-specific transformation is the "quantile transformation", which takes the form

$$h_{is}[X_i^{\mathcal{P}}(s)] = h_s[X_i(s)] = \mathrm{P}(X(s) \leq X_i(s)|s)\ . \tag{2.2}$$

Here $\{X(s) : s \in \mathcal{S}\}$ is the stochastic process on the functional domain $\mathcal{S}$ and $X_i(s)$ is the observed functional realization for the ith study participant. As a result, the functional covariate at each $s \in \mathcal{S}$ is transformed to the cumulative distribution function (cdf) conditional on s. The right panel in Figure 2.3 displays the NHANES data after being smoothed and quantile-transformed. In contrast to the original data, these transformed data cover well its range, $[0,24] \times [0,1]$. The difference is that the interpretation of $\hat{F}(\cdot,\cdot)$ changes because the first argument is the relative, not absolute, size of the intensity of physical activity for an individual at a given time of day. More precisely, $\hat{F}(s,\tau)$ is the effect of being in the τth quantile of the functional covariate (physical activity intensity) at time $s \in \mathcal{S}$. Results can be interpreted on the original scale of physical activity intensity $h_{is}^{-1}(\tau)$, but interpretation

of results should only be conducted in regions of the domain with sufficient data density.

A similar approach was proposed by McLean et al., 2014, who used the empirical cdf separately for each observed $s \in \mathcal{S}$, $\hat{\mathrm{P}}(X(s) \leq X_i(s)|s) = N^{-1}\sum_{j=1}^{N} I(X_j(s) \leq X_i(s))$ to estimate the marginal cdf. When the functional covariate is irregularly sampled or measured with error we propose a complementary approach using additive quantile regression, which assumes smoothness of the quantiles of $X(s)$ across the functional domain. Specifically, consider the model

$$\mu_\tau(s) = f_{0,\tau}(s)\,, \tag{2.3}$$

where τ is the quantile to be estimated, $\mu_\tau(s) = \inf[X_i(s) : \mathrm{P}(X(s) \leq X_i(s)|s) \geq \tau]$ is the τth quantile of $X(\cdot)$ given s, and $f_{0,\tau}(s)$ is a smooth function of s. Computationally stable estimation of $f_{0,\tau}(s)$ can be done via penalized splines (Fasiolo et al., 2017; Fasiolo et al., 2019). Quantile regression estimates the inverse cdf and requires a separate model fit for each quantile of interest. Therefore, our estimator for (2.2) involves separate regression models for $\tau \in \boldsymbol{\tau}_0$ where $\boldsymbol{\tau}_0$ is a fine grid on $(0, 1)$. Given these model fits, the quantile transformation is obtained by $\hat{\mathrm{P}}(X(s) \leq X_i(s)|s) = \sup\{\tau : X_i(s) \leq \hat{\mu}_\tau(s), \tau \in \boldsymbol{\tau}_0\}$. While the empirical cdf approach may suffice in many applications, the proposed estimator can be extended to conditioning on subject-specific features, as discussed in Section 2.2.2.2.

2.2.2.1.2 Domain-specific standardization. The second domain-specific transformation subtracts the domain specific mean and divides by the domain specific standard deviation:

$$h_{is}[X_i^{\mathcal{P}}(s)] = h_s[X_i(s)] = \frac{X_i(s) - E[X(s)|s]}{\sqrt{\text{Var}[X(s)|s]}}\,, \tag{2.4}$$

where $E[X(s)|s]$ and $\text{Var}[X(s)|s]$ can be estimated using their empirical estimators. After this transformation the interpretation of $\hat{F}(s,x)$ is the effect of being x standard deviations from the population mean at each $s \in \mathcal{S}$. Unlike quantile transformation, the domain standardization approach is more sensitive to skewness and may not cover the domain of $F(\cdot,\cdot)$ well.

2.2.2.2 Subject-specific Transformations

We also consider transformations that depend on subject-specific characteristics. In our application we will use such transformations to conduct age-specific standardization of PA profiles. This will allow to assess the predictive power of PA on mortality independent of the natural decline of PA with age. Suppose that $\boldsymbol{U}_i$ is a q-dimensional vector of subject-specific characteristics and we want to extend the quantile transformation introduced in Section 2.2.2.1 to account for $\boldsymbol{U}_i$. In this extended setting, $X(s)$ is defined as the stochastic process on the functional domain $\mathcal{S}$ that also depends on $\boldsymbol{U}_i$.

2.2.2.2.1 Subject-specific quantile transformation. Consider the subject/domain-specific transformation

$$h_{is}[X_i^{\mathcal{P}}(s)] = h_{is}[X_i(s)] = \text{P}(X(s) \leq X_i(s)|s, \boldsymbol{U}_i)\,. \tag{2.5}$$

We propose to extend model (2.3) to the more general additive quantile regression model

$$\mu_\tau(s|\boldsymbol{U}_i) = \sum_{j=1}^{q} f_{j,\tau}(U_{ij}, s) \; . \tag{2.6}$$

The functions $f_{j,\tau}(\cdot, s)$ are smooth functions of each covariate and the functional domain $s \in \mathcal{S}$. While the model may seem involved, it can be easily estimated by existing software; see, for example, the `qgam` package (Fasiolo et al., 2019) in `R`. Estimating (2.5) follows the same procedure described for the domain-specific quantile transformation. First, we estimate separate models for $\tau \in \boldsymbol{\tau}_0$ where $\boldsymbol{\tau}_0$ is a fine grid in $(0, 1)$. Then, given these model fits, we estimate $\hat{\mathrm{P}}(X(s) \leq X_i(s)|s, \boldsymbol{U}_i) = \sup\{\tau : X_i(s|\boldsymbol{U}_i) \leq \hat{\mu}_\tau(s|\boldsymbol{U}_i), \tau \in \boldsymbol{\tau}_0\}$. Note that extending the empirical cdf ideas to account for subject-specific covariates, $\boldsymbol{U}_i$, would be difficult, especially if the number of covariates is large.

2.2.2.2.2 Subject-specific standardization. The second subject/domain-specific transformation is

$$h_{is}[X_i^{\mathcal{P}}(s)] = h_{is}[X_i(s)] = \frac{X_i(s) - E[X(s)|s, \boldsymbol{U}_i]}{\sqrt{\mathrm{Var}[X(s)|s, \boldsymbol{U}_i]}} \; . \tag{2.7}$$

As with the subject- and domain-specific quantile transformation, this transformation will likely involve some modeling of the first and second moments of $X(s)$ conditional on s and $\boldsymbol{U}_i$. Separate additive regression models for $E[X(s)|s, \boldsymbol{U}_i]$ and $E[X^2(s)|s, \boldsymbol{U}_i]$ with linear predictors of the same form as Model (2.6) could be used.

2.2.2.2.3 History of the functional domain. The third subject/domain-specific transformation is

$$h_{is}[X_i^{\mathcal{P}}(s)] = \int_0^s X_i(u)du \ . \tag{2.8}$$

Just as with the other transformations, the interpretation of $F(\cdot,\cdot)$ changes compared to using the original functional covariates. For example, in the NHANES study $F(\cdot,\cdot)$ becomes "the effect of volume and timing of cumulative PA".

2.2.2.3 Choosing a transformation function

Choosing a transformation function for any given application is an open and important problem. We propose to choose the transformation function based on interpretability of results, ability to cover the domain of interest, and predictive performance. In our application predictive performance was roughly comparable for models with or without transformations, so the first two criteria took precedence. We also strongly suggest to display density plots and identify regions of the space where there with sparse or no data. Model coefficients should not be interpreted in these areas, as little is known about extrapolation of complex nonparametric smoothers.

It could be tempting to jointly model the transformation function $h_{is}(\cdot)$ and $F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\}$, though the exact procedure for doing so is not currently available. Such an approach would require the building of custom software and could substantially increase the computational complexity of the associated algorithms. To preserve computational efficiency and interpretability

we do not pursue this idea, though this could be an important area for future research.

2.2.3 Identifiability

Wood, 2017 discussed the necessity of adding constraints on the smooth functions to ensure the identifiability of additive models. Specifically, the constraint

$$\sum_{i=1}^{N} \int_{\mathcal{S}} F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\} ds = 0 \,, \tag{2.9}$$

is imposed by default when fitting an additive model using the `R mgcv` package. However, this constraint is not sufficient to ensure identifiability of the additive functional Cox model. For example, for any bivariate smooth function $F(s,x)$ let $g(s)$ be a function such that $\int_{\mathcal{S}} g(s)ds = 0$. If we define the function $F^*(s,x) = F(s,x) + g(s)$ then

$$\int_{\mathcal{S}} F^*(s,x)ds = \int_{\mathcal{S}} F(s,x) + g(s)ds = \int_{\mathcal{S}} F(s,x)ds + \int_{\mathcal{S}} g(s)ds = \int_{\mathcal{S}} F(s,x)ds \,. \tag{2.10}$$

Therefore, the integrals are the same, but $F^*(s,x) \neq F(s,x)$ as long as $g(s) \neq 0$. Müller, Wu, and Yao, 2013 proved that $F(s,x)$ is identifiable up to a function that does not depending on x. However, this result applies only in regions of the domain covered by $\{s, X_i(s)\}$. Hence, the identifiability condition is not sufficient to ensure that the model is estimable in areas of the domain sparsely covered or not covered by $\{s, X_i(s)\}$. The domain covered by $\{s, X_i(s)\}$ is often different from and much smaller than the rectangular domain defined by the minimum and maximum of s and $X_i(s)$ for all s and i. We refer to this

as the "rectangular domain".

We will show that this distinction is crucial in our application, where the functional coefficient is estimable only in a sub-region of the rectangular domain. This suggests that, when possible, transformations of the functional covariate should be considered to improve the coverage of the rectangular domain. This is particularly important as automatic nonparametric smoothers tend to work well on rectangular domains; see the supplementary materials for a more detailed discussion.

Suppose enough observations are available in the functional parametric region of interest. To address identifiability over the estimable domain we impose the additional identifiability constraints

$$\sum_{i=1}^{N} F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\} = 0, \text{ for each } s \in \mathcal{S} . \tag{2.11}$$

These constraints restrict $F(s, x)$ at each $s \in \mathcal{S}$ to have a unique form within the range of $h_{is}[X_i^{\mathcal{P}}(s)]$, thus ensuring identifiability over the area of interest. This restriction can be implemented directly in our software. The simulation results in Section 2.4 confirm that this approach provides a reasonable solution; see implementation details in Section 2.2.4.

2.2.4 Estimation and Inference

2.2.4.1 Penalized Spline Smoothing

Penalized splines smoothing (Ruppert, Wand, and Carroll, 2003; Wood, Pya, and Säfken, 2016; Wood, 2017) and its connection with mixed effects modeling provide a powerful inferential platform for nonparametric regression

modeling. Thus, pairing penalized spline smoothing and functional modeling (Goldsmith et al., 2011; Goldsmith et al., 2012; Scheipl, Staicu, and Greven, 2015; Greven and Scheipl, 2017) provides a modern, easy to implement, extendable framework for data analysis. Here we follow this principle and provide only the essential modeling details, as we consider penalized splines to be a mainstream inferential approach. Other methods include regressing on the functional principal component scores; see, for example, Müller and Yao, 2008. While this approach leads to comparable predictive performance, the estimation of the functional parameter is highly sensitive to the choice of the number of principal components. For the bivariate case McLean et al., 2014 suggested using the tensor products of two univariate P-splines to model $F(\cdot, \cdot)$

$$F(s,x) = \sum_{j=1}^{K_s} \sum_{k=1}^{K_x} \theta_{j,k} B_j(s) B_k(x) \,, \tag{2.12}$$

where $B_j(\cdot)$ and $B_k(\cdot)$ are two univariate splines on the domains of s and x, respectively. The parameters $\{\theta_{j,k} : j = 1, 2, ..., K_s; k = 1, 2, ..., K_x\}$ are the spline coefficients. We use cyclic cubic regression splines for the functional domain s, and cubic regression splines for the functional covariate domain x. Given the excellent `mgcv` software in `R` this can be implemented as (users of `mgcv` will find this easy to understand):

```
ti(x, s, bs=c("cr","cc"), k=c(Kx,Ks), mc=c(TRUE,FALSE))
```

The `mc` parameter specifies marginal centering constraints to the functional covariate domain, which coincides with the identifiability constraints (2.11) discussed in Section 2.2.3. The cyclic cubic regression splines are used for the

functional domain, s, to account for the periodicity of PA as both $s = 0$ and $s = 24$ hours indicate midnight in our notation.

2.2.4.2 Estimation

Using the tensor product notation, the additive functional Cox model can be rewritten as

$$\begin{aligned}
\log \lambda_i(t|\boldsymbol{Z}_i, \boldsymbol{X}_i) &= \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \int_{\mathcal{S}} F\{s, h_{is}[X_i^{\mathcal{P}}(s)]\} ds \\
&= \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \sum_{j=1}^{K_s} \sum_{k=1}^{K_x} \theta_{j,k} \int_{\mathcal{S}} B_j(s) B_k\{h_{is}[X_i^{\mathcal{P}}(s)]\} ds \\
&= \log \lambda_0(t) + \boldsymbol{Z}_i^T \boldsymbol{\beta} + \boldsymbol{V}_i^T \boldsymbol{\theta} \\
&= \log \lambda_0(t) + \boldsymbol{W}_i^T \boldsymbol{\gamma} \, .
\end{aligned} \tag{2.13}$$

Here $\boldsymbol{W}_i^T = (\boldsymbol{Z}_i^T, \boldsymbol{V}_i^T)$ and $\boldsymbol{\gamma}^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)$, where $\boldsymbol{\theta}$ is the vector with entries $\theta_{j,k}$ and $\boldsymbol{V}_i$ is the vector with entries $\int_{\mathcal{S}} B_j(s) B_k\{h_{is}[X_i^{\mathcal{P}}(s)]\} ds$, and both vectors $\boldsymbol{\theta}$ and $\boldsymbol{V}_i$ are organized in the same order of the indices $j = 1, 2, ..., K_s; k = 1, 2, ..., K_x$. The parameters $\boldsymbol{\gamma}$ are estimated by maximizing the penalized partial log likelihood, where the penalty is induced on the $\boldsymbol{\theta}$ parameters (the vector of parameters of the bivariate spline function) using standard quadratic penalties that depend on the vector of smoothing parameters λ. Selection of λ is discussed in Section 2.2.4.3. The penalized partial log likelihood has the following form

$$l_p(\boldsymbol{\gamma}|\lambda) = l(\boldsymbol{\gamma}) - \lambda J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \delta_i [\boldsymbol{W}_i^T \boldsymbol{\gamma} - \log \sum_{Y_j \geq Y_i} e^{\boldsymbol{W}_i^T \boldsymbol{\gamma}}] - \lambda J(\boldsymbol{\theta}) \, . \tag{2.14}$$

For every fixed smoothing parameter λ, the estimator of the regression coefficients is obtained by $\hat{\gamma}(\lambda) = \arg\min_{\gamma} -l_p(\gamma|\lambda)$ using the Newton-Raphson algorithm. Detailed information on this approach can be found in Wood, Pya, and Säfken, 2016 supplementary materials G. Following ideas in Wood, Pya, and Säfken, 2016 we use cubic spline penalties. The practical implication of this approach is that it is easy to implement in the `gam` function of the `mgcv` package. For example, suppose that the functional covariates are observed on an equally-spaced grid $\{s_1, ..., s_m\}$ of the functional domain. The integral in equation (2.13) is approximated through weighted numerical summation of functional observations, where the weights are the increments between each neighboring pair and are stored in the vector `l`. In the case with only one scalar covariate, `z`, if the event indicator δ_i and observed survival time Y_i are stored in the variables `delta` and `Y`, respectively, the code is simply

```
fit <- gam(Y ~ z + ti(x, s, by = l, bs=c("cr","cc"), k=c(Kx,Ks),
mc=c(TRUE,FALSE)), weights = delta, data, family = cox.ph())
```

The detailed procedure of fitting the model, extracting estimates on a fine grid, and visualizing the results is provided in the vignette in the supplementary materials. We would like to underline the simplicity of the code. This was possible because of the careful and novel methodological work and is an important contribution. Indeed, it is only through the use of powerful, reproducible, inferential code that functional methods can become popular after publication in highly specialized journals.

An alternative approach to Cox regression is to use estimation of the nonparametric proportional hazard model; see, for example, Lin, He, and

Huang, 2016 and Hiabu et al., 2017. However, here we focus on generalizing the Cox proportional hazard model.

2.2.4.3 Smoothing Parameter Selection

An important problem is the selection of the smoothing parameter λ. Several selection criteria have been proposed, including GCV (Gu, 2013), AIC (Hurvich, Simonoff, and Tsai, 1998), EPIC (Shinohara et al., 2011) and REML (Ruppert, Wand, and Carroll, 2003). In the context of functional Cox regression, Gellar et al., 2015 proposed using a criteria based on AIC. Here we follow the estimation procedure described in Wood, Pya, and Säfken, 2016, which involves maximizing the Laplace approximation of the marginal likelihood of the smoothing parameter.

2.2.4.4 Statistical Inference

In addition to estimating the model parameters, γ, the corresponding Hessian matrix $\boldsymbol{H}$ is also estimated; see Wood, 2017 supplementary material G for details. Several estimators of the covariance matrix have been proposed in the literature, including a "sandwich estimator" $\boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}$ proposed by Gray, 1992, and a "pseudo standard error" $\boldsymbol{H}^{-1}$ proposed by Verweij and Van Houwelingen, 1994. Here $\boldsymbol{G}$ denotes the corresponding Hessian matrix without a penalty term. Therneau, Grambsch, and Pankratz, 2003 recommended to perform significance tests on the estimator $\boldsymbol{H}^{-1}$ instead of $\boldsymbol{H}^{-1}\boldsymbol{G}\boldsymbol{H}^{-1}$. Because the structure of the problem does not change fundamentally, the inference for our model follows a similar procedure with that introduced by McLean et al., 2014 for the functional generalized additive model.

2.3 Applications

The additive functional Cox model was motivated by studying the association between the high-resolution physical activity measures and time to death. We present results using different transformations of the functional covariate, and compare their interpretation and predictive performance with those of traditional approaches.

2.3.1 NHANES

As discussed in Section 4.1, NHANES contains a large number of individual characteristics together with physical activity data measured by hip-worn accelerometry. Data are linked to mortality outcomes and are available, for example, through the `rnhanesdata` package in `R`. For more details on organizing and analyzing NHANES physical activity data see Leroux et al., 2019b. In our application, the functional covariate is the smoothed minute-level average LAC over available days, referred to as "smoothed LAC" below. We now describe the pre-processing procedure for creating the functional covariate (smoothed LAC). Denote the minute level activity counts $\text{AC}_{ij}(s)$ for subject $i = 1, \ldots, N$, and day $j = 1, \ldots, 7$, for minute $s = 1, \ldots, 1440$. To account for subject non-compliance with study wear-time protocols, we use the default estimated wear/non-wear at every minute available in the `rnhanesdata` package, which were created using established algorithms (Troiano et al., 2008). Denote wear/non-wear indicators by $W_{ij}(s)$, where 0 and 1 correspond to estimated non-wear and wear, respectively. Next step introduces an indicator variable G_{ij}, which encodes a "good" day of accelerometry data as 1 and is

defined as a day with at least 10 hours of estimated wear time. More precisely, $G_{ij} = 1(\sum_{s=1}^{1440} W_{ij}(s) \geq 600)$, where $1(\cdot)$ is the indicator function. The index set for all good days for subject i is denoted by $J_i^* = \{j : G_{ij} = 1\}$. To create the functional predictor the daily activity counts are transformed as $\text{LAC}_{ij}(s) = g(\text{AC}_{ij}(s))$ where $g(y) = \log(1 + y)$. This $g(\cdot)$ is introduced when building our functional predictors and is conceptually completely separated from the transformation function $h_{is}(\cdot)$ in the model. These log activity profiles are averaged across all the "good" accelerometry days for study participant i: $\text{LAC}_i(s) = |J_i^*|^{-1} \sum_{j \in J_i^*} \text{LAC}_{ij}(s)$. These individual profiles are then smoothed using FPCA. Therefore, we start our NHANES application with these smoothed LAC, denoted by $\text{LAC}_i'(s) = X_i(s) = \sum_{k=1}^{K} \tilde{\xi}_{ik} \hat{\phi}_k(s)$ where $\tilde{\xi}_{ik}$ are the predicted scores and $\hat{\phi}_k(s)$ are the estimated eigenfunctions obtained from functional principal component analysis (FPCA) (Xiao et al., 2016b). These steps are all pre-processing steps and are conceptually distinct from the subject-domain transformation function $h_{is}(\cdot)$.

Survival time is measured in months from accelerometer wear and, for the purpose of this study, all survival times are censored at 10 years. Among the 2816 study participants who met the inclusion criteria, 2157 (76.6%) survived for more than 10 years from the time when accelerometry data were collected. We adjust for sociodemographic factors (age, race, employment status, education, poverty-income ratio), health factors (self-reported overall health, smoking status, alcohol consumption, body mass index, mobility difficulty), and disease indicators (diabetes, coronary heart disease, congestive heart failure, stroke, cancer, systolic blood pressure, total cholesterol).

2.3.2 Application Results

2.3.2.1 Estimated Functional Surface $\hat{F}(\cdot, \cdot)$

All models are fit using the `R` code described in Section 2.2. A vignette to reproduce the analysis is provided in the supplementary materials. Different transformations were used on the functional covariate, including identity transformation, quantile transformation, and subject-specific quantile transformation. We focus on the density of different transformed functional covariates and its connection with model estimation and interpretation.

To illustrate the complexity of the problem, Figure 2.4 displays density plots for the observations $\{s, X_i(s)\}$, where $X_i(s)$ is a generic notation for the LAC before or after transformation. First and second row correspond to individuals who were deceased within and alive for 10 years, respectively. First column: unsmoothed LAC. Second column: smoothed LAC. Third column: quantile-transformed smoothed LAC. The rectangular domain was partitioned into small sub-rectangles and the number of points $\{s, X_i(s)\}$ was counted in each sub-rectangle and plotted. For example, for unsmoothed and smoothed LAC the $[0, 24] \times [0, 8]$ rectangle domain was partitioned into $24 \times 20 = 480$ equal size rectangles, where each rectangle corresponds to one hour and an increment of 0.4 on the $\log(1 + \text{AC})$ scale. A similar partition (into 480 equal size sub-rectangles) was done for the quantile-transformed data, though the domain in this case was $[0, 24] \times [0, 1]$, because the quantile transformed data spans the $[0, 1]$ domain, whereas the original LAC data spans the $[0, 8]$ domain. The number in each block decreases from red (largest) to blue (smallest). Color scales are different across plots.

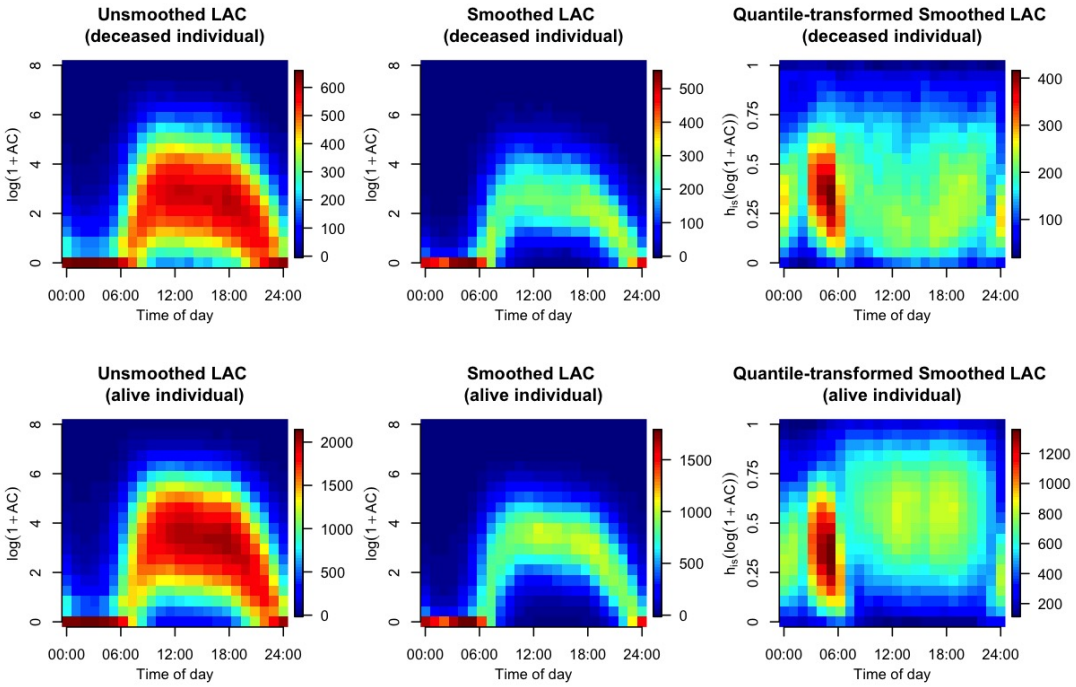


Figure 2.4: Density plots for the log-transformed activity counts (LAC) before or after transformation. First and second row correspond to individuals who were deceased within and alive for at least 10 years, respectively. First column: unsmoothed LAC. Second column: smoothed LAC. Third column: quantile-transformed smoothed LAC. The rectangular domain was partitioned into small sub-rectangles and the number of $\{s, X_i(s)\}$ was counted in each sub-rectangle and plotted. The number in each block decreases from red (largest) to blue (smallest). Color scales are different across plots.

The panels for unsmoothed LAC (left panels) show that data are extremely sparse in the sub-domain corresponding to high activity counts during the night; see the dark blue in the top-left region of the grid. This illustrates the estimability principle that we have discussed in this paper. Indeed, the regions $[0, 6] \times [3, 8]$ and $[6, 24] \times [7, 8]$ contain little or no data, despite the fact that we have a relatively large sample size (2816 study participants). An additional concern is that between 12AM to 6AM, when most people sleep, the density of observations is highly concentrated around zero counts. Thus, in

this case, imposing the identifiability condition in Müller, Wu, and Yao, 2013 is necessary but insufficient to ensure that we obtain meaningful estimates in these regions. In fact, we expect similar results even if the sample size were 100 times larger. The panels for smoothed LAC (second column) show that the problem is further exacerbated by smoothing. In contrast, the panels for quantile-transformed smoothed LAC (right panels in Figure 2.4) show a much better coverage of the rectangle domain $[0, 24] \times [0, 1]$. This suggests that the quantile transformation could be an effective approach for addressing the estimability problem over the entire domain.

The estimates using smoothed LAC before and after transformations are shown in Figure 2.5, where each plot is visualized as a function of both the functional domain and the value of the functional covariate. The value of $\hat{F}(\cdot, \cdot)$ decreases from red (highest) to white to blue (lowest), where a higher value corresponds to a higher hazard of death. The top-left panel in Figure 2.5 provides the functional surface estimates for the smoothed LAC ($h_{is}(x) = x$). A superficial look at the results could indicate that low activity intensity is associated with a higher hazard of mortality at any time of a day. This seems unreasonable, as a vast scientific literature exists on the benefits of restful sleep. We believe that this result is due to spurious extrapolation in regions of the functional domain with sparse or no data; compare these results with the data density panels in the second column of Figure 2.4.

To further explore whether this is, indeed the case, we conducted a stratified analysis by separating the time of day into night (12am to 8am) and day (8am to 12am). Results of this analysis are shown in the two side-by-side

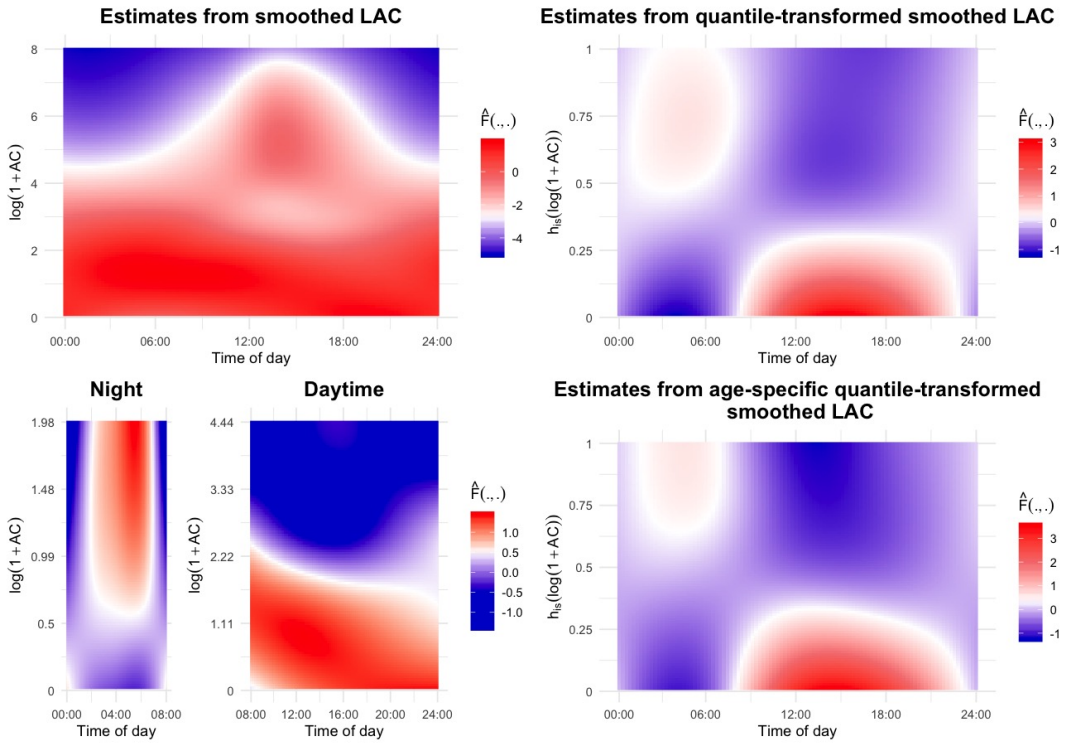


Figure 2.5: Estimated surfaces using the additive functional Cox model from untransformed (top-left), quantile-transformed (top-right), and age-specific quantile-transformed (bottom-right) smoothed LAC. The bottom-left panels show the estimates from smoothed LAC when stratifying the analysis by night (12am to 8am) and day (8am to 12am). For each time period of the stratified analysis, the functional covariate region of interest is set at the 90th percentile of the functional covariate values to ensure good coverage of data. The value of $\hat{F}(\cdot,\cdot)$ decreases from red (highest) to blue (lowest hazard of death). Color scales are different across plots.

panels on the bottom in the first column (titled "Night" and "Daytime"). The interpretation of these results is that, higher activity during the night and lower activity during the day are associated with a higher hazard of mortality. However, even in this case the results during the night continue to be affected by data sparsity (left-upper and right-upper corners of the panel labeled "Night"). Another problem is that when conducting stratified analyses, the y-axis (which corresponds to the value of smoothed LAC) changes for each

strata ($[0, 1.98]$ for night and $[0, 4.44]$ for day), which makes interpretation of results more difficult. The boundary value of the functional covariate is set to the 90th percentile of smoothed LAC for each time period to ensure good domain coverage. The lower boundary value at night is due to the lower LAC during the night. Moreover, the choice of threshold of 8am for night/day transition is debatable and 7am could provide a better transition point. Deciding which transition threshold to use is not obvious in practice, which further reduces the appeal of the stratified analysis.

For all these reasons we considered quantile transformations of smoothed LAC. As shown on the right panels of Figure 2.4, the improved coverage on the grid indicates that the functional surface is more likely to be estimable on the $[0, 24] \times [0, 1]$ grid of the transformed data. The top-right panel of Figure 2.5 indicates that lower relative activity during the day and higher relative activity during the night are associated with a higher hazard of mortality. Specifically, being below the 30th percentile of smoothed LAC in the population during daytime (9am to 9pm) is associated with a higher hazard of mortality. In contrast with the results in the top-left panel, this plot indicates that a lower relative LAC (less than the 35th percentile of smoothed LAC in the population) during the night (12am to 8am) is associated with a lower hazard of mortality. These results agree with those obtained from the stratified analysis.

While both approaches yield similar interpretable results, we favor the use of quantile transformation because: (1) the quantile transformation automatically unifies different scales of functional observations across the domain; (2) stratified analysis requires manual choice of the threshold, while the y-axis

of the domain may be different; (3) results using quantile transformations are interpretable and translatable, whereas stratified analyses are based on quantities that are difficult to use for providing physical activity guidance; (4) the quantile transformation is easier to implement; and (5) the quantile transformation has a long and successful history in genomics analyses.

Building on the success of the quantile transformation, we have further applied the age-specific quantile transformation, where age is the subject-specific characteristic. This eliminates the effects of age on the individual quantile, as older individuals tend to have lower levels of activity. The result is illustrated on the bottom-right panel of Figure 2.5. The plot indicates that even after using age-specific quantile transformations, the pattern of the effect of diurnal and nocturnal activity intensity on the hazard of mortality remains relatively unchanged. Results indicate that individuals who are above the 60th percentile of activity during the night and below the 35th percentile during the day in their corresponding age group are at increased risk of mortality, irrespective of age.

2.3.2.2 Predictive Performance

Cross-validated Harrell's C-index (Harrell Jr et al., 1982; Harrell Jr et al., 1984; Harrell Jr, Lee, and Mark, 1996) and Brier score (Brier, 1950) are used as measures of predictive performance. Across models, the non-functional covariates are kept the same, allowing for a comparison of different approaches for modelling the association between activity and mortality while adjusting for

common confounders. The results of 10-fold cross validation are shown in Table 2.1. Two functional models, additive functional Cox model ("AFCM") and linear functional Cox model ("LFCM"), are implemented as the comparison. For each functional model, we evaluate the predictive performance using three forms of LAC including unsmoothed, smoothed, and quantile-transformed smoothed. In addition, the non-functional Cox proportional hazard model ("Cox PHM") is implemented as the baseline model, where the the average smoothed LAC over the entire day is used as a scalar predictor.

Table 2.1: The average 10-fold cross-validated Harrell's C-index and Brier score of all combinations of model and physical activity measures. "AFCM" denotes the additive functional Cox model, "LFCM" denotes the linear functional Cox model, and "Cox PHM" denotes the standard Cox proportional hazard model using the average activity as predictor.

Model	LAC	Harrell's C-index	Brier score
AFCM	unsmoothed	0.795	0.0751
	smoothed	0.795	0.0751
	smoothed + quantile	0.793	0.0753
LFCM	unsmoothed	0.791	0.0754
	smoothed	0.791	0.0754
	smoothed + quantile	0.791	0.0753
Cox PHM		0.791	0.0758

The predictive performance of our models (with or without transformed data) is better than that of the linear functional Cox model and non-functional model, though differences are small. Among the additive functional Cox models the difference in predictive performance is marginal. This may be due to the fact that the test and training datasets share the regions where the functional parameters are well estimated, irrespective of the transformation used. This indicates that using prediction measures may not be sufficient to differentiate between models that use raw or transformed data or among

different types of transformations. However, interpretation of results is substantially improved by the quantile transformation and agrees with stratified analyses by time of day, as shown in Figure 2.5.

2.4 Simulation Study

2.4.1 Simulation Framework

For simplicity, we consider the case with only one functional covariate $\boldsymbol{X}_i$ and no scalar covariate. Consider the case when $h_{is}[X_i^{\mathcal{P}}(s)] = h_{is}[X_i(s)]$ and denote by $\eta_i = \int_{\mathcal{S}} F\{s, h_{is}[X_i(s)]\}ds$. The model introduced in Section 2.2 can be simplified as

$$\log \lambda_i(t|\boldsymbol{X}_i) = \log \lambda_0(t) + \int_{\mathcal{S}} F\{s, h_{is}[X_i(s)]\}ds = \log \lambda_0(t) + \eta_i \,. \tag{2.15}$$

Functional covariates are simulated using functional principal component analysis (FPCA) (Ramsay, 2004) applied to the NHANES data. Survival data are simulated using either the estimated $F(\cdot,\cdot)$ based on the NHANES data or pre-specified forms of $F(\cdot,\cdot)$ in combination with simulated functional covariates and estimated cumulative baseline hazards.

2.4.1.1 Simulating Functional Covariates

FPCA has been widely used to smooth functional data by restricting the projection to the first M principal components of the Karhunen-Loève expansion (Karhunen, 1947; Loeve, 1978). If we denote by $\mu(s) = E[X(s)]$, then the subject-specific functional predictors can be expanded as $X_i(s) \approx \mu(s) + \sum_{j=1}^{M} \sqrt{\lambda_j}\xi_{ij}\psi_j(s)$. Here $\lambda_1 \geq ... \geq \lambda_M$ and $\psi_1(\cdot), ..., \psi_M(\cdot)$ are the first

M eigenvalues and eigenfunctions, respectively. The scores are derived by $\xi_{ij} = \frac{1}{\sqrt{\lambda_j}} \int X_i(t)\psi_j(t)dt$ and $E(\xi_{ij}) = 0$, $E(\xi_{ij}\xi_{ik}) = I(j = k)$, which is equal to 1 if $j = k$ and 0 otherwise. The functional covariates $\tilde{X}_i(s)$ are simulated as $\tilde{X}_i(s) = \hat{\mu}(s) + \sum_{j=1}^{M} \sqrt{\hat{\lambda}_j} e_{ij} \hat{\psi}_j(s)$, where e_{ij} are i.i.d. $N(0,1)$ random variables. The mean, $\hat{\mu}(s)$, eigenvalues, $\hat{\lambda}_j$, and eigenfunctions, $\hat{\psi}_j(s)$, are estimated using FPCA on the NHANES data. This was done using the `R` function `fpca.face` (Xiao et al., 2016a) in the `refund` package (Crainiceanu et al., 2012).

In our simulation, the functional covariates $\tilde{X}_i(s)$ are generated by applying FPCA to the smoothed LAC, the functional covariates $X_i(s)$ of NHANES application. We then impose quantile transformation h_{is} on simulated functional covariates to reduce data sparsity observed on the middle panels of Figure 2.4. See `R` code in the supplementary materials for implementation details.

2.4.1.2 Simulating Survival Data

Simulating survival data with non-pathological properties that mimic the NHANES data was one of the most difficult tasks addressed by this paper. We propose to use the estimated survival function, which proved to be both practical and realistic. While methods for estimating survival times under parametric assumptions on the distribution of survival times exist (Bender, Augustin, and Blettner, 2005; Austin, 2012), we have been unable to adapt these methods to NHANES. Part of the problem is that small changes on the modeling assumptions can lead to substantial changes in the distribution of survival times. Moreover, we could not find a general set of recommendations

on how to choose parameters, especially in the context of functional predictors.

Thus, we are taking a different approach and use the `gam` function in `R` package `mgcv` to estimate the cumulative baseline hazard $\tilde{\Lambda}_0(t) = \int_0^t \lambda_0(u)du$ from the fitted model based on the NHANES data, where certain constraints are imposed to ensure non-negative and non-decreasing estimates; see `R` code in the supplementary materials for details. We use two simulation approaches to derive the estimated linear predictor $\tilde{\eta}_i$ based on: (1) the surface estimated from NHANES; and (2) several pre-specified functional forms of $F(\cdot,\cdot)$. The estimated survival function is calculated as $\tilde{S}_i(t) = \exp\{-e^{\tilde{\eta}_i}\tilde{\Lambda}_0(t)\}$, and the simulated survival time $\tilde{T}_i$ is obtained using the relationship between the density and the survival function. The censoring times $\tilde{C}_i$ are simulated from the empirical distribution of censoring times in the NHANES data to control the censoring rate.

In summary, the simulation procedure has the following steps: (1) derive the estimated cumulative baseline hazard function $\tilde{\Lambda}_0(t)$; (2) derive the estimated linear predictor $\tilde{\eta}_i$; (3) derive the estimated survival function $\tilde{S}_i(t)$; (4) simulate survival time $\tilde{T}_i$ from $\tilde{S}_i(t)$; and (5) simulate censoring time $\tilde{C}_i$ from the empirical distribution of censoring times in NHANES. The `R` code for this simulation approach is provided in the supplementary materials.

2.4.2 Simulation Results

As discussed in Section 2.4.1.1, we simulate functional covariates using FPCA on the NHANES data. We use two choices of $F(\cdot,\cdot)$, one based on NHANES

and one based on pre-specified functional forms to evaluate model performance from different perspectives.

2.4.2.1 The Functional Surface Estimated from NHANES

We simulate survival and functional data that mimic real NHANES data with different sample sizes in the first simulation. The "true" $F(\cdot,\cdot)$ is set as the estimator using the quantile-transformed smoothed LAC in NHANES. In this section we show the model fitting performance using the correctly specified quantile transformation, while additional results using the misspecified identity transformation are included in the supplementary materials. The functional domain is rescaled to $[0,1]$ for notation convenience. We focus on the estimation accuracy of the surface $F(s,x)$ and cumulative baseline hazard $\Lambda_0(t)$ under different sample sizes. The surface is estimated on the grid $\mathcal{S}\times\mathcal{X}=[0,1]\times[0,1]$ with 100 equally-spaced points in each dimension. Thus, the estimated surface is a 100×100 dimensional matrix where the value in each cell represents the estimated $\hat{F}(\cdot,\cdot)$ at that point in the domain. The cumulative baseline hazard function is estimated on the interval $[0,10]$ on a 1000 dimensional equally-spaced grid of points.

The estimated surface based on quantile-transformed smoothed LAC of all $N=2816$ NHANES participants is shown in the top-left panel of Figure 2.6, serving as the baseline for comparing the estimation performance of simulated data with different sample sizes. This plot is different from the application results in Section 2.3 since no other covariates are included. In simulations we used three sample sizes $N=1000,2000,5000$. The sample size N controls the

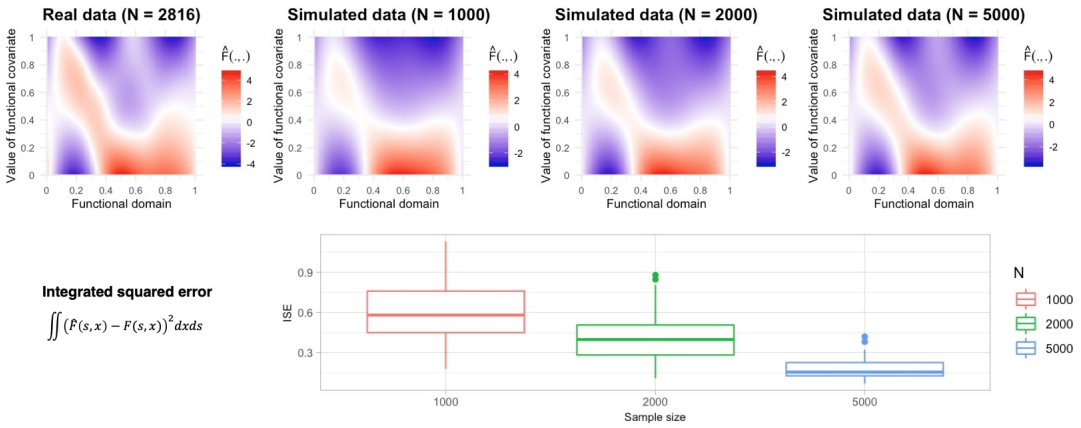


Figure 2.6: Estimated surface in NHANES, $\hat{F}(\cdot,\cdot)$, (first panel in the top row), which was used as true surface in simulations. Average estimated surfaces based on 100 simulations for $N = 1000, 2000, 5000$ (second, third, and fourth panel in the top row). Red, white, and blue correspond to highest, median, and lowest hazard of mortality. For each N, the distribution of the integrated squared error (ISE) is shown in the second row.

amount of information, in general, and the data density on the functional grid in particular. For each N, we performed 100 simulations and the average of the estimated surfaces are shown in Figure 2.6. A sample of randomly selected estimates from 100 simulations are included in the supplementary materials.

As sample size increases, the average estimated surfaces are getting closer to the baseline functional surface; see panels from left to right in Figure 2.6. These results provide a first check that the new simulation framework is reasonable and produces datasets with similar characteristics with the original NHANES. Moreover, the estimation method provides, at least on average, reasonable estimators of the target functional predictor surface. To better quantify how well surfaces and cumulative baseline hazards are estimated, the integrated square error (ISE) is calculated for each simulated data set. For surfaces ISE is defined as $\text{ISE}(\hat{F}(s,x)) = \int_{\mathcal{S}}\int_{\mathcal{X}}(\hat{F}(s,x) - F(s,x))^2 dxds$, where

$F(s,x)$ refers to the baseline functional term estimated from the real data and used in simulations. The bottom panel of Figure 2.6 displays the distribution of ISE as a function of sample size. Results illustrate a large decrease in ISE as sample size increases. More precisely, the median ISE when $N = 5000$ is less than a third the median ISE for $N = 1000$. The ISE for the cumulative baseline hazard functions is defined analogously and we show their distributions under different sample sizes in the supplementary materials. Further decompositions of ISE into integrated squared bias (denoted by "bias^2") and average variance (denoted by "variance") for both surfaces and cumulative baseline hazards are reported in Table 2.2. Results suggest that both bias and variance decrease as sample size increases. In addition, the estimation procedure is fast even for large sample sizes. Indeed, it took only ~ 2 minutes to obtain one fit with 5000 study participants on a regular laptop (2.7GHz dual-core Intel Core i5 processor), as shown in the right column of Table 2.2.

Table 2.2: The integrated squared bias and average variance for the estimated surface $\hat{F}(\cdot,\cdot)$ and cumulative baseline hazard function $\hat{\Lambda}_0(\cdot)$ based on 100 simulations with different sample sizes $N = 1000, 2000, 5000$. The average computing time per simulation is shown on the right column.

Sample size	$\hat{F}(\cdot,\cdot)$		$\hat{\Lambda}_0(\cdot)(\times 10^{-4})$		Average Comp. Time (sec.)
	bias^2	variance	bias^2	variance	
$N = 1000$	0.303	0.292	0.011	1.061	21.07
$N = 2000$	0.150	0.211	0.009	0.598	46.24
$N = 5000$	0.042	0.138	0.002	0.203	126.20

2.4.2.2 Pre-specified Functional Forms of $F(\cdot,\cdot)$

We also considered pre-specified functional forms for $F(s,x)$, while keeping the simulation of the functional covariates the same. We considered the

following functional forms for $F(s,x)$: (1) $F(s,x) = 2x$, which scales linearly with respect to x, and remains constant across s; (2) $F(s,x) = xs$, which scales linearly with respect to both x and s; (3) $F(s,x) = x^3 s$, which scales linearly with respect to s, but is nonlinear with respect to x; and (4) $F(s,x) = \sin xs$, which is nonlinear with respect to both x and s. The term $\int X_i(s)\beta(s)ds$ in the linear functional Cox model corresponds to $\beta(s) = 2$ in the first scenario and $\beta(s) = s$ in the second. However, the linear functional Cox model is misspecified for the last two scenarios. For each $F(s,x)$, we perform 100 simulations with sample size $N = 5000$ and derive the average estimated functional surfaces from each model. To reduce the linear approximation effect of nonlinear functions within small regions, for example $f(s,x) = xs$ and $g(s,x) = x^3 s$ are very close for x and s between 0 and 1, and to comply with the necessary identifiability constraints, the grid is modified to $[0,2] \times [-1,1]$ for all $F(s,x)$. The simulated functional covariates are rescaled to the same range to ensure good data coverage.

Figure 2.7 displays the true surfaces (first row) and the average estimated surfaces based on the linear (second row) and additive (third row) functional Cox model. The estimated surfaces and cumulative baseline hazard functions from a sample of randomly selected simulations are provided in the supplementary materials. The color scale is the same within each $F(s,x)$, but varies across different functions, as they have different ranges. The first two columns correspond to functions $F(s,x)$ that are linear in x. Both the linear and additive functional Cox models estimate the true surfaces well, at least when comparing the average surfaces. The ISE distributions shown in the last

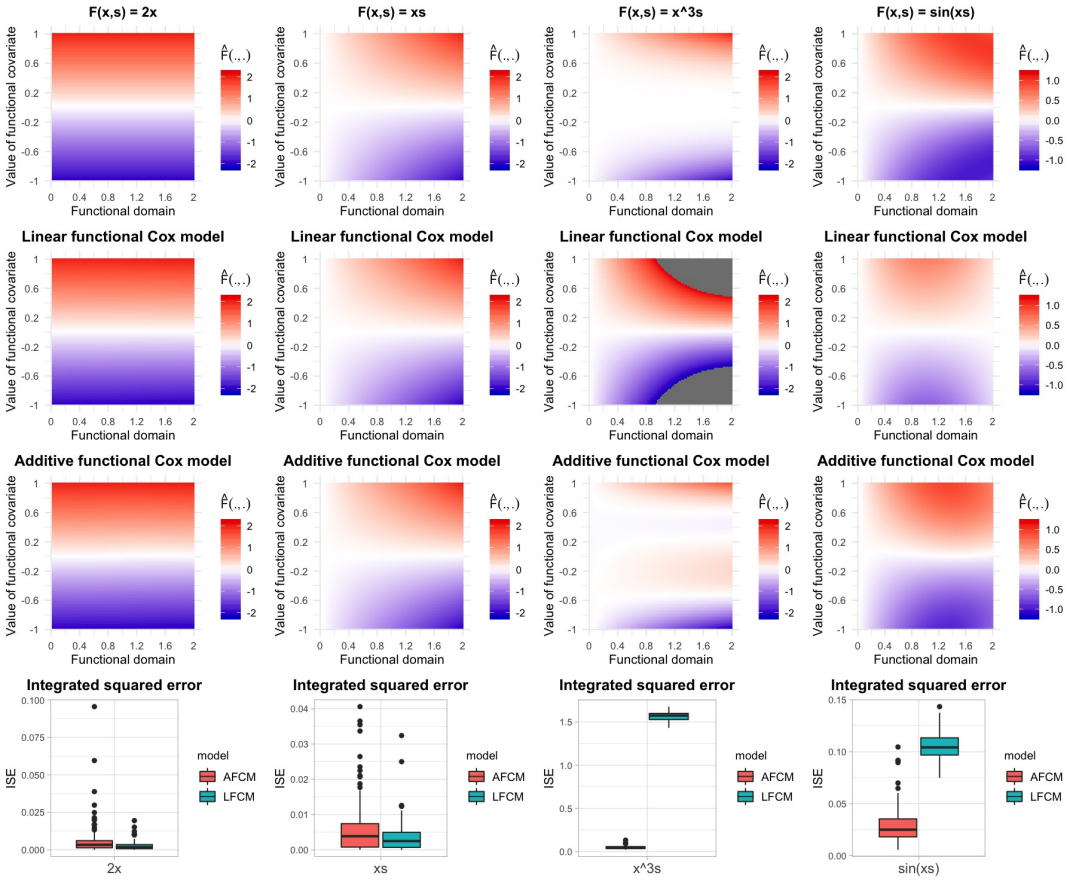


Figure 2.7: True surface (first row) and average estimated surfaces based on 100 simulations with sample size $N = 5000$ (second and third row). The second row corresponds to the linear functional Cox model and the third row corresponds to the additive functional Cox model. The fourth row displays the integrated squared error for the additive (red) and linear (blue) functional Cox models. Each column corresponds to a specific functional form of $F(\cdot, \cdot)$.

row indicate that the linear model performs slightly better, probably because of the higher complexity of the functional additive model. These results are expected and reassuring, indicating that the additive functional Cox model performs well when the true model is linear. The last two columns correspond

to functions $F(s, x)$ that are nonlinear in x. In both scenarios the additive functional Cox model substantially outperforms the linear Cox model. This can be observed both from the comparison of the average of estimated surfaces (first three rows) and from the distributions of ISE (last row).

2.5 Discussion

The major contribution of our paper is the introduction of the nonparametric additive functional Cox model. This allows to quantify complex associations between a time to event outcome and functional covariates. This approach is crucial in the NHANES application where activity intensity during the night and day has different implications for the hazard of mortality. The technical argument is to use an unspecified bivariate function $F(s, x)$ that depends on the functional domain, $\mathcal{S}$, and the transformed functional covariates $h_{is}[X_i^{\mathcal{P}}(s)]$, where necessary constraints are imposed to ensure the model identifiability.

Another important contribution is to introduce a class of transformations of functional covariates, which can alleviate problems related to data sparsity in particular areas of the domain of the $F(\cdot, \cdot)$ function and substantially improve the model estimability. We have discussed several types of domain-specific transformations and extended the idea to subject-specific transformation. While the interpretation of results changes with the transformation, this provides a flexible approach for exploring the type of association between the functional predictors and time to event.

Our model was motivated by the NHANES study, where we identified

highly interpretable patterns of association between daily trajectories of physical activity and the hazard of mortality. The prediction performance of the proposed model also improved slightly relative to the linear functional Cox model. Important advantages of the model are that it can be implemented using existing software, implementation is very fast even for large datasets, and reproducible code is provided with this paper.

We also introduced the first approach for realistic simulations of survival data for Cox models with functional predictors. Detailed `R` simulation code is provided in the supplementary materials and the associated vignette. Simulations indicate that the additive functional Cox model performs almost as well as the linear functional Cox model when the function is linear and much better when it is not. A vignette is provided in the `rnhanesdata` package introducing and implementing all our work.

Our approach shows that complex functional models can be fit quickly and efficiently using state of the art software. However, our work has also opened several exciting avenues of research including establishing the theoretical properties of the estimation approach and exploring additional functional transformations.

2.6 Supplementary Material

The supplementary material of this project is available at https://doi.org/10.1080/10618600.2020.1853550.

References

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Cooper, Rachel, Lei Huang, Rebecca Hardy, Adina Crainiceanu, Tamara Harris, Jennifer A Schrack, Ciprian Crainiceanu, and Diana Kuh (2017). "Obesity history and daily patterns of physical activity at age 60–64 years: findings from the MRC National Survey of Health and Development". In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 72.10, pp. 1424–1430.

Schmid, Daniela, Cristian Ricci, and Michael F Leitzmann (2015). "Associations of objectively assessed physical activity and sedentary time with all-cause mortality in US adults: the NHANES study". In: *PloS one* 10.3, e0119591.

Matthews, Charles E, Sarah Kozey Keadle, Richard P Troiano, Lisa Kahle, Annemarie Koster, Robert Brychta, Dane Van Domelen, Paolo Caserotti, Kong Y Chen, Tamara B Harris, et al. (2016). "Accelerometer-measured dose-response for physical activity, sedentary time, and mortality in US adults". In: *The American journal of clinical nutrition* 104.5, pp. 1424–1432.

Sallis, James F and Brian E Saelens (2000). "Assessment of physical activity by self-report: status, limitations, and future directions". In: *Research quarterly for exercise and sport* 71.sup2, pp. 1–14.

Silsbury, Zoë, Robert Goldsmith, and Alison Rushton (2015). "Systematic review of the measurement properties of self-report physical activity questionnaires in healthy adult populations". In: *BMJ open* 5.9, e008430.

Troiano, Richard P, David Berrigan, Kevin W Dodd, Louise C Masse, Timothy Tilert, and Margaret McDowell (2008). "Physical activity in the United States measured by accelerometer". In: *Medicine & Science in Sports & Exercise* 40.1, pp. 181–188.

Schrack, Jennifer A, Vadim Zipunnikov, Jeff Goldsmith, Jiawei Bai, Eleanor M Simonsick, Ciprian Crainiceanu, and Luigi Ferrucci (2014). "Assessing the "physical cliff": detailed quantification of age-related differences in daily patterns of physical activity". In: *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 69.8, pp. 973–979.

Bai, Jiawei, Chongzhi Di, Luo Xiao, Kelly R Evenson, Andrea Z LaCroix, Ciprian M Crainiceanu, and David M Buchner (2016). "An activity index for raw accelerometry data and its comparison with other activity metrics". In: *PloS one* 11.8, e0160644.

Doherty, Aiden, Dan Jackson, Nils Hammerla, Thomas Plötz, Patrick Olivier, Malcolm H Granat, Tom White, Vincent T Van Hees, Michael I Trenell, Christoper G Owen, et al. (2017). "Large scale population assessment of physical activity using wrist worn accelerometers: the UK biobank study". In: *PloS one* 12.2, e0169649.

Mirel, LB, LK Mohadjer, SM Dohrmann, J Clark, VL Burt, CL Johnson, and LR Curtin (2013). "National Health and Nutrition Examination Survey: estimation procedures, 2007-2010." In: *Vital and health statistics. Series 2, Data evaluation and methods research* 159, pp. 1–17.

Leroux, A, C Crainiceanu, E Smirnova, and Q Cao (2019a). *rnhanesdata: NHANES Accelerometry Data Pipeline*. URL: https://github.com/andrew-leroux/rnhanesdata.

Varma, Vijay R, Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov (2017). "Re-evaluating the effect of age on physical activity over the lifespan". In: *Preventive medicine* 101, pp. 102–108.

Varma, Vijay R, Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov (2018). "Total volume of physical activity: TAC, TLAC or TAC (λ)". In: *Preventive medicine* 106, p. 233.

Gellar, Jonathan E, Elizabeth Colantuoni, Dale M Needham, and Ciprian M Crainiceanu (2015). "Cox regression models with functional covariates for survival data". In: *Statistical modelling* 15.3, pp. 256–278.

Qu, Simeng, Jane-Ling Wang, Xiao Wang, et al. (2016). "Optimal estimation for the functional cox model". In: *The Annals of Statistics* 44.4, pp. 1708–1738.

Kong, Dehan, Joseph G Ibrahim, Eunjee Lee, and Hongtu Zhu (2018). "FLCRM: Functional linear cox regression model". In: *Biometrics* 74.1, pp. 109–117.

McLean, Mathew W, Giles Hooker, Ana-Maria Staicu, Fabian Scheipl, and David Ruppert (2014). "Functional generalized additive models". In: *Journal of Computational and Graphical Statistics* 23.1, pp. 249–269.

Fasiolo, Matteo, Yannig Goude, Raphael Nedellec, and Simon N. Wood (2017). *Fast calibrated additive quantile regression*. URL: https://arxiv.org/abs/1707.03307.

Fasiolo, Matteo, Yannig Goude, Raphael Nedellec, and Simon N. Wood (2019). *qgam: quantile non-parametric additive models*.

Wood, Simon N (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Müller, Hans-Georg, Yichao Wu, and Fang Yao (2013). "Continuously additive models for nonlinear functional regression". In: *Biometrika* 100.3, pp. 607–622.

Ruppert, David, M. P. Wand, and R. J. Carroll (2003). "Frontmatter". In: *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, pp. i–vi.

Wood, Simon N, Natalya Pya, and Benjamin Säfken (2016). "Smoothing parameter and model selection for general smooth models". In: *Journal of the American Statistical Association* 111.516, pp. 1548–1563.

Goldsmith, Jeff, Jennifer Bobb, Ciprian M Crainiceanu, Brian Caffo, and Daniel Reich (2011). "Penalized functional regression". In: *Journal of Computational and Graphical Statistics* 20.4, pp. 830–851.

Goldsmith, Jeff, Ciprian M Crainiceanu, Brian Caffo, and Daniel Reich (2012). "Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.3, pp. 453–469.

Scheipl, Fabian, Ana-Maria Staicu, and Sonja Greven (2015). "Functional Additive Mixed Models". In: *Journal of Computational and Graphical Statistics* 24.2, pp. 477–501.

Greven, Sonja and Fabian Scheipl (2017). "A general framework for functional regression modelling". In: *Statistical Modelling* 17.1-2, pp. 1–35.

Müller, Hans-Georg and Fang Yao (2008). "Functional additive models". In: *Journal of the American Statistical Association* 103.484, pp. 1534–1544.

Lin, Huazhen, Ye He, and Jian Huang (2016). "A global partial likelihood estimation in the additive Cox proportional hazards model". In: *Journal of Statistical Planning and Inference* 169, pp. 71–87.

Hiabu, Munir, Enno Mammen, Maria Dolores Martinez-Miranda, and Jens Perch Nielsen (2017). "Smooth backfitting of proportional hazards–A new approach projecting survival data". In: *arXiv preprint arXiv:1707.04622*.

Gu, Chong (2013). *Smoothing spline ANOVA models*. Vol. 297. Springer Science & Business Media.

Hurvich, Clifford M, Jeffrey S Simonoff, and Chih-Ling Tsai (1998). "Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.2, pp. 271–293.

Shinohara, Russell T, Ciprian M Crainiceanu, Brian S Caffo, and Daniel S Reich (2011). "Longitudinal analysis of spatiotemporal processes: a case study of dynamic contrast-enhanced magnetic resonance imaging in multiple sclerosis". In.

Gray, Robert J (1992). "Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis". In: *Journal of the American Statistical Association* 87.420, pp. 942–951.

Verweij, Pierre JM and Hans C Van Houwelingen (1994). "Penalized likelihood in Cox regression". In: *Statistics in medicine* 13.23-24, pp. 2427–2436.

Therneau, Terry M, Patricia M Grambsch, and V Shane Pankratz (2003). "Penalized survival models and frailty". In: *Journal of computational and graphical statistics* 12.1, pp. 156–175.

Leroux, Andrew, Junrui Di, Ekaterina Smirnova, Elizabeth J. Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K. Urbanek, and Ciprian Crainiceanu (2019b). "Organizing and Analyzing the Activity Data in NHANES". In: *Statistics in Biosciences* 11.2, pp. 262–287.

Xiao, Luo, Vadim Zipunnikov, David Ruppert, and Ciprian Crainiceanu (2016b). "Fast covariance estimation for high-dimensional functional data". In: *Statistics and computing* 26.1-2, pp. 409–421.

Harrell Jr, Frank E, Robert M Califf, David B Pryor, Kerry L Lee, and Robert A Rosati (1982). "Evaluating the yield of medical tests". In: *Jama* 247.18, pp. 2543–2546.

Harrell Jr, Frank E, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati (1984). "Regression modelling strategies for improved prognostic prediction". In: *Statistics in medicine* 3.2, pp. 143–152.

Harrell Jr, Frank E, Kerry L Lee, and Daniel B Mark (1996). "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Statistics in medicine* 15.4, pp. 361–387.

Brier, Glenn W (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review* 78.1, pp. 1–3.

Ramsay, James O (2004). "Functional data analysis". In: *Encyclopedia of Statistical Sciences* 4.

Karhunen, Kari (1947). *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*. Vol. 37. Sana.

Loeve, Michel (1978). "Probability theory. II, volume 46 of". In: *Graduate Texts in Mathematics*.

Xiao, L., V. Zipunnikov, D. Ruppert, and C.M. Crainiceanu (2016a). "Fast covariance estimation for high-dimensional functional data". In: *Statistics and Computing* 26, pp. 409–421.

Crainiceanu, Ciprian, Philip Reiss, Jeff Goldsmith, Lei Huang, Lan Huo, Fabian Scheipl, S Greven, J Harezlak, MG Kundu, and Y Zhao (2012). "refund: Regression with functional data". In: *R package version 0.1-6*.

Bender, Ralf, Thomas Augustin, and Maria Blettner (2005). "Generating survival times to simulate Cox proportional hazards models". In: *Statistics in medicine* 24.11, pp. 1713–1723.

Austin, Peter C (2012). "Generating survival times to simulate Cox proportional hazards models with time-varying covariates". In: *Statistics in medicine* 31.29, pp. 3946–3958.

Chapter 3

Fast Multilevel Functional Principal Component Analysis

3.1 Introduction

Functional data measured at multiple visits have become increasingly common. A standard technique for analyzing such data is multilevel functional principal component analysis (MFPCA) (Di et al., 2009), which provides a decomposition of the observed data into within- and between-subject variation. The MFPCA model generalizes traditional measurement error and multilevel models to the case when the basic measurement unit is a function.

The motivating data is a large physical activity dataset from the National Health and Nutrition Examination Survey (NHANES), a study conducted in two-year waves by the United States Centers for Disease Control and Prevention (CDC). Each study participant in the 2003-2004 and 2005-2006 waves was asked to wear a hip-worn physical activity monitor (PAM) for seven consecutive days. Acceleration data were publicly released as minute-level activity counts (AC), a proprietary measure of physical activity intensity.

For quality control purposes, some days were excluded from the analysis; see Section 3.6 for exclusion criteria. Figure 3.1 displays the physical activity profiles of three randomly selected NHANES study participants (left, middle and right panels). The number of available days varies by study participant with a maximum of 7. For example, the data for the study participant shown in the left panels contains only 6 days, while the data for the study participant shown in the middle panels contains only 5 days. Within a column, each row shows the minute-level AC of one day from midnight to midnight, where the title for each panel indicates the corresponding day of the week. The dataset has 12802 study participants and 65777 days in total, with 1440 observations per day for a total of 94718880 minute-level observations.

The NHANES dataset is an example of large-scale multilevel high dimensional functional data. For subject i on day j of the week, the physical activity intensity value at minute $s \in \mathcal{S}$ can be denoted as $Y_{ij}(s)$, where $\mathcal{S}$ is the time interval from midnight to midnight. Functional principal component analysis (FPCA) is a popular approach in functional data analysis (Ramsay and Silverman, 2005). Some early work, including Ramsay and Dalzell (1991), Silverman et al. (1996), and Yao, Müller, and Wang (2005), focused on single-level analysis. For multilevel functional data, multilevel functional principal component analysis (MFPCA) (Di et al., 2009) provides an explicit decomposition of the within- and between-subject variation in the functional space. MFPCA is designed to analyze functional data with two levels of functional variation. This is the simplest model in the rapidly expanding family of multilevel functional mixed effects models, including multilevel functional models (Brumback and

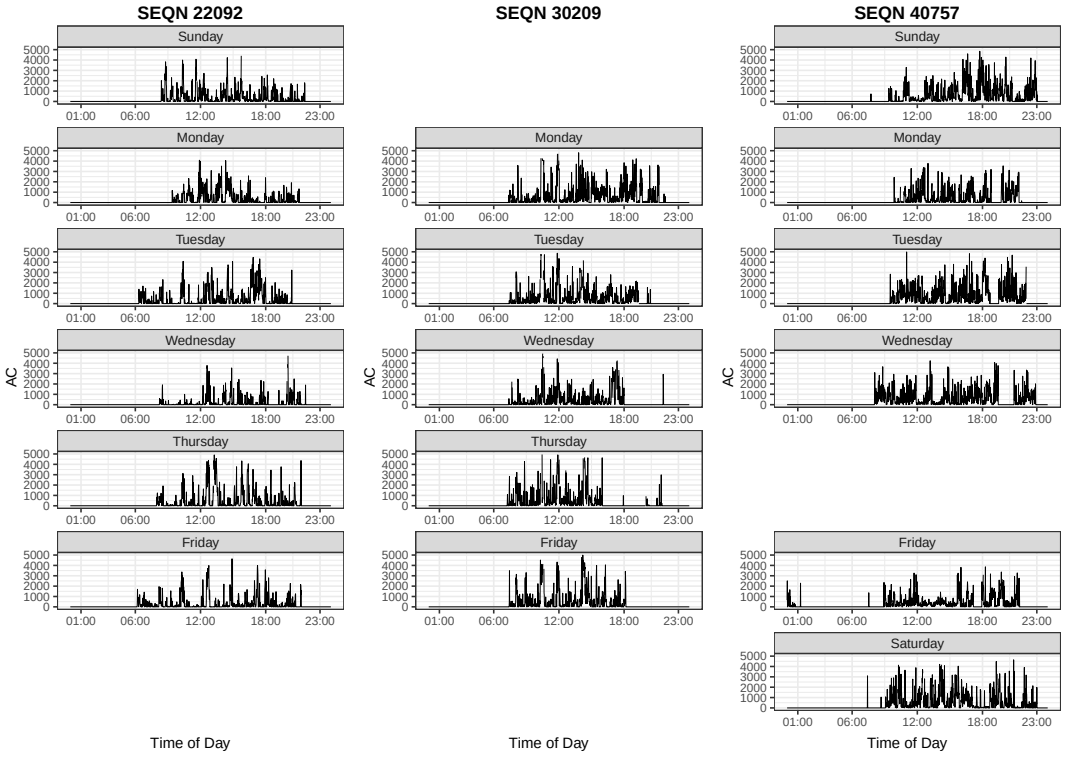


Figure 3.1: Physical activity profiles of three NHANES study participants over available days. Each study participant is uniquely identified by the SEQN number. Left column: SEQN 22092. Middle column: SEQN 30209. Right column: SEQN 40757. Within each column, each row displays the minute-level AC of one day from midnight to midnight, titled by day of the week from Sunday (top row) to Saturday (bottom row).

Rice, 1998; Morris and Carroll, 2006; Aston, Chiou, and Evans, 2010; Morris et al., 2011; Chen and Müller, 2012; Serban and Jiang, 2012; Goldsmith, Zipunnikov, and Schrack, 2015; Li et al., 2015; Xu, Li, and Nettleton, 2018; Gaynanova, Punjabi, and Crainiceanu, 2022), longitudinal functional models (Greven et al., 2010; Zipunnikov et al., 2014; Park and Staicu, 2015; Scheffler et al., 2020; Cui et al., 2021; Boland et al., 2022; Shamshoian et al., 2022; Li et al., 2022), spatial functional models (Zhang et al., 2016; Li et al., 2021), and structured functional models (Shou et al., 2015; Scheipl, Staicu, and Greven,

2015). MFPCA is also related to but distinct from multivariate functional data (Berrendero, Justel, and Svarc, 2011; Chiou, Chen, and Yang, 2014; Kowal, Matteson, and Ruppert, 2017; Happ and Greven, 2018; Wong, Li, and Zhu, 2019).

Applying MFPCA to a dataset with over 10000 study participants and over 1000 observations per function remains computationally challenging. The current implementation of MFPCA is slow for high dimensional functional data, as the number of computations is proportional to the *cube* of the number of observations per function. The problem is that the current MFPCA requires: (1) the construction, smoothing and eigendecomposition of covariance matrices with the dimension equal to the number of observations per function; and (2) the score prediction which relies on the inversion of multiple high dimensional covariance matrices. The fast covariance estimation in Xiao et al. (2016), referred to as FACE, addressed these problems for single-level functional data. Indeed, FACE, implemented in the `fpca.face()` function of the `refund R` package (Goldsmith et al., 2020), requires only minutes to smooth covariance matrices of dimension 100000. Here we provide methods that substantially accelerate MFPCA by extending methods inspired by FACE.

Therefore, we propose fast multilevel functional principal component analysis (fast MFPCA) and implement it in the `mfpca.face()` function of the `refund R` package. The fast MFPCA approach improves MFPCA by: (1) constructing transformed functional data instead of calculating the method of moments estimators of covariance matrices; (2) obtaining level-specific eigendecompositions by extending FACE to multilevel functional data, which

avoids using high dimensional covariance matrices; and (3) predicting principal component scores based on mixed model equations (MME). Using the combination of these ideas, the fast MFPCA scales up *linearly* with the number of observations per function and is orders of magnitude faster than MFPCA. For example, fast MFPCA took less than 5 minutes to fit the NHANES data compared with MFPCA, which took more than 5 days.

The rest of the paper is organized as follows. We review MFPCA in Section 3.2 and introduce the fast MFPCA approach in Section 3.3. A theoretical study of the proposed method is provided in Section 3.4. A simulation study is conducted in Section 3.5 to compare the computation time and accuracy of our new approach with existing methods. We discuss the NHANES application results in Section 3.6 and conclude with a discussion in Section 3.7.

3.2 Multilevel Functional Principal Component Analysis

We briefly review the multilevel functional data model proposed in Di et al., 2009. Denote by $Y_{ij}(s)$ the observed data for subject $i = 1, \ldots, I$ at visit $j = 1, \ldots, J_i$ and location $s \in \{s_1, \ldots, s_L\} \in \mathcal{S}$, where $\mathcal{S}$ is a compact domain. Each function has L observations at the same set of time points and we focus on the case when L is relatively large. Denote by $n = \sum_{i=1}^{I} J_i$ the total number of visits. Consider a functional ANOVA model with measurement error: $Y_{ij}(s) = X_{ij}(s) + \epsilon_{ij}(s) = \mu(s) + \eta_j(s) + Z_i(s) + W_{ij}(s) + \epsilon_{ij}(s)$, where $\mu(s)$ is the population mean function, $\eta_j(s)$ is the jth visit-specific shift from $\mu(s)$, $Z_i(s)$ is the random subject-specific mean deviation for the ith

subject, $W_{ij}(s)$ is the random jth visit-specific deviation from $Z_i(s)$, and $\epsilon_{ij}(s)$ is a white noise with variance σ^2. The random functions $Z_i(s)$ and $W_{ij}(s)$ are mutually independent zero mean processes with covariance functions $K_B(s,t) = \text{cov}\{Z_i(s), Z_i(t)\}$ and $K_W(s,t) = \text{cov}\{W_{ij}(s), W_{ij}(t)\}$, respectively.

Algorithm 1 MFPCA

1. Estimate mean functions $\mu(s)$ and $\eta_j(s)$ by applying univariate smoothers to observed data under working independence assumption and subtract them from observed data.
2. Construct method of moment (MoM) estimators of the total covariance $K_T(s,t)$ and between-subject covariance $K_B(s,t)$, denoted by $\widehat{K}_T(s,t)$ and $\widehat{K}_B(s,t)$, respectively.
3. Smooth $\widehat{K}_T(s,t)$ and $\widehat{K}_B(s,t)$ using bivariate smoothing, leading to two smooth estimates, denoted by $\widetilde{K}_T(s,t)$ and $\widetilde{K}_B(s,t)$, respectively. Let $\widetilde{K}_W(s,t) = \widetilde{K}_T(s,t) - \widetilde{K}_B(s,t)$.
4. Conduct eigenanalysis on discretized $\widetilde{K}_B(s,t)$ and $\widetilde{K}_W(s,t)$ matrices.
5. Estimate error variance σ^2 by $\hat{\sigma}^2 = \int_{\mathcal{S}} \{\hat{K}_T(s,s) - \tilde{K}_T(s,s)\} ds$.
6. Predict scores using best linear unbiased prediction (BLUP).

Let $K_T(s,t) := \text{cov}\left\{X_{ij}(s), X_{ij}(t)\right\} = K_B(s,t) + K_W(s,t)$ be the total variance of the smooth functional data. Suppose that the between-subject covariance function K_B admits the eigendecomposition $K_B(s,t) = \sum_{k\geq 1} \lambda_k^{(1)} \phi_k(s)\phi_k(t)$, where $\lambda_1^{(1)} \geq \lambda_2^{(1)} \geq \cdots \geq 0$ are eigenvalues with associated orthonormal eigenfunctions $\phi_k(s)$, that is, $\int_{\mathcal{S}} \phi_{k_1}(s)\phi_{k_2}(s)ds = \delta_{\{k_1=k_2\}}$ for any pair (k_1, k_2). Here $\delta_{\{\cdot\}}$ is an indicator function which is equal to 1 if the statement is true and 0 otherwise. Then the random function $Z_i(s)$ can be written as $Z_i(s) = \sum_{k\geq 1} \xi_{ik}\phi_k(s)$, where ξ_{ik} are scores with zero mean and variance $\lambda_k^{(1)}$ and are mutually uncorrelated. Similarly, suppose that the within-subject covariance function K_W also has an eigendecomposition $K_W(s,t) = \sum_{k\geq 1} \lambda_k^{(2)} \psi_k(s)\psi_k(t)$, where $\lambda_1^{(2)} \geq \lambda_2^{(2)} \geq \cdots \geq 0$ are the eigenvalues with associated orthonormal

eigenfunctions $\psi_k(s)$. Then $W_{ij}(s) = \sum_{k\geq 1} \zeta_{ijk}\psi_k(s)$, where ζ_{ijk} are mutually uncorrelated scores with zero mean and variance $\lambda_k^{(2)}$.

The primary goal of MFPCA is to reduce the functional data to two sets of uncorrelated scores: level-1 scores ξ_{ik_1} and level-2 scores ζ_{ijk_2}. The eigenfunctions are also useful for understanding the variation patterns in functional data. The traditional MFPCA methods in Di et al., 2009 are summarized in Algorithm 1.

3.3 Fast MFPCA

MFPCA slows down substantially when the number of observations per function increases. Indeed, constructing the $L \times L$ sample covariance matrices (step 2 of Algorithm 1) requires $O(IL^2)$ computations, where I is the total number of functions. Next, standard bivariate smoothing of $L \times L$ covariance matrices (step 3 of Algorithm 1) requires $O(L^3)$ computations. Furthermore, the eigenanalysis of $L \times L$ matrices (step 4 of Algorithm 1) also requires $O(L^3)$ computations. Finally, the score prediction (step 6 of Algorithm 1) requires inversion of covariance matrices of size $L \times L$, which require $O(L^3)$ computations.

To deal with the computational challenges of traditional MFPCA, we propose the fast MFPCA approach, which differs from traditional MFPCA in three aspects. First, we construct transformed functional data for which the underlying smooth curves have the desired covariance operators. This construction takes only $O(IL)$ computations compared to $O(IL^2)$ computations for the method of moment (MoM) sample covariance estimators. Second, we

apply the fast covariance estimation method (FACE, Xiao et al., 2016) to the transformed data to estimate the covariance operators. FACE avoids the direct calculation of MoM and eigenanalysis of empirical of $L \times L$ covariance matrices. The computational complexity of FACE is $O(ILc)$, where c is the number of B-spline bases functions used for smoothing and is much smaller than I and L. Finally, we predict the principal component scores using mixed model equations (MME), which is computationally efficient because the number of eigenfunctions is much smaller than the number of observations per curve.

To the best of our knowledge, this is the first time FACE is extended and applied to general multilevel functional data. Moreover, we are not aware of any literature using MME for score prediction in functional data analysis. This idea combination reduces the computational complexity of MFPCA from from $O(IL^2 + L^3)$ to $O(ILc)$.

3.3.1 FACE and Eigenanalysis

The fast covariance estimation method (FACE) in Xiao et al. (2016) is a bivariate smoothing method based on the tensor-product splines. FACE is computationally fast as it scales up linearly with the number of functions and the number of observations per function. In addition, the eigenanalysis via FACE avoids computationally expensive eigendecompositions of large covariance matrices. Below, we provide the technical details of FACE, which inspired the functional data transformation described in Section 3.3.2.

Let $\mathbf{Y}$ be an $L \times I$ data matrix with each column corresponding to one observed single-level function (centered and scaled) evaluated at the time

points $\{s_1, \ldots, s_L\}$. Let $\widehat{\mathbf{K}} = I^{-1}\mathbf{Y}\mathbf{Y}^T$ be the sample covariance matrix estimator. Denote by $\mathbf{B}(s) = [B_1(s), \ldots, B_c(s)]^T$ the $c \times 1$ dimensional vector of c cubic B-spline basis evaluated at s. Let $\mathbf{B} = [\mathbf{B}(s_1), \ldots, \mathbf{B}(s_L)]^T$ be the $L \times c$ design matrix, where each row corresponds to a sampling point and each column corresponds to a spline basis. An $L \times L$ smoother matrix is constructed as $\mathbf{S} = \mathbf{B}(\mathbf{B}^T\mathbf{B}/L + \lambda\mathbf{P})^{-1}\mathbf{B}^T/L$, where $\mathbf{P}$ is the qth order penalty matrix in P-splines (Eilers and Marx, 1996) and λ is the smoothing parameter; see Section S.1 in the supplementary material for more details. FACE uses $\widetilde{\mathbf{K}} = \mathbf{S}\widehat{\mathbf{K}}\mathbf{S}$ as the smooth estimator of the covariance. Thus, the (s,t) entry of the covariance estimator is $\widetilde{K}(s,t) = \mathbf{B}^T(s)\mathbf{\Theta}\mathbf{B}(t)$, where $\mathbf{\Theta} = (\mathbf{B}^T\mathbf{B}/L + \lambda\mathbf{P})^{-1}(\mathbf{B}^T\widehat{\mathbf{K}}\mathbf{B}/L^2)(\mathbf{B}^T\mathbf{B}/L + \lambda\mathbf{P})^{-1}$ is a $c \times c$ symmetric and positive semi-definite matrix. FACE does not directly calculate $\widetilde{\mathbf{K}}$ and only computes the $c \times c$ coefficient matrix $\mathbf{\Theta}$, which can be written as $\mathbf{F}\mathbf{F}^T$ with $\mathbf{F} = (L\sqrt{I})^{-1}(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{P})^{-1}\mathbf{B}^T\mathbf{Y}$. Notice that $\mathbf{F}$ is of dimension $c \times I$ and its calculation requires $O(ILc)$ computations. Because c is the number of spline functions, which is much smaller than L and I, the calculation of $\mathbf{F}$ is much faster than direct bivariate smoothing of a covariance operator. The main idea is to use the decomposition of the sample covariance $\widehat{\mathbf{K}}$ combined with the property that the smoothed covariance $\widetilde{\mathbf{K}}$ matrix is low rank. FACE selects the smoothing parameter λ by minimizing the pooled generalized cross validation (PGCV) (Xiao, Li, and Ruppert, 2013), which is fast because it relies on univariate functional smoothing to control bivariate covariance smoothing.

The tensor product spline form of $\widetilde{K}(s,t)$ can be used to reduce the computational complexity of eigenfunctions and eigenvalues estimation procedure. Indeed, let $\mathbf{G} = \int \mathbf{B}(s)\mathbf{B}^T(s)ds$, which can be constructed to be positive definite. Take the eigendecomposition $\mathbf{G}^{1/2}\mathbf{\Theta}\mathbf{G}^{1/2} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_c]$ is an orthonormal matrix and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_c)$ is a diagonal matrix with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_c \geq 0$. Then, $\mathbf{U}_k^T\mathbf{G}^{-1/2}\mathbf{B}(s)$ is the estimated kth eigenfunction corresponding to eigenvalue λ_k. In contrast, the direct approach is to conduct a spectral decomposition of the covariance matrix evaluated on a dense grid; see, for example, Yao, Müller, and Wang, 2005. This method is computationally expensive, especially in high dimensions.

3.3.2 Transformed Functional Data

To simplify the notation introduced in Section 3.2, let $\widetilde{Y}_{ij}(s) = Y_{ij}(s) - \mu(s) - \eta_j(s)$ be the demeaned observed data, where $\mu(s)$ and $\eta_j(s)$ will be replaced by their estimates in applications. Recall that the data are observed on a regular grid $\{s_1, ..., s_L\}$. Define $\widetilde{\mathbf{Y}}_{ij} = [\widetilde{Y}_{ij}(s_1), \ldots, \widetilde{Y}_{ij}(s_L)]^T$ and $\widetilde{\mathbf{Y}} = [\widetilde{\mathbf{Y}}_{11}, \ldots, \widetilde{\mathbf{Y}}_{1J_1}, \ldots, \widetilde{\mathbf{Y}}_{I1}, \ldots, \widetilde{\mathbf{Y}}_{IJ_I}] \in \mathbb{R}^{L\times n}$. Denote by $n = \sum_{i=1}^{I} J_i$, the total number of curves, and define $n_I = \sum_{i=1}^{I} J_i(J_i - 1)$.

As suggested in Shou et al. (2015), MoM estimators of covariance matrices in structured functional data often take the "sandwich" form $n^{-1}\widetilde{\mathbf{Y}}\mathbf{H}\widetilde{\mathbf{Y}}^T$, where $\mathbf{H}$ are design-specific matrices. For example, for the MoM estimator of total covariance, $\mathbf{H}$ is the $n \times n$ identity matrix. Therefore, FACE can be applied to the transformed data $\widetilde{\mathbf{Y}}\mathbf{H}^{1/2}$ to smooth the covariance whenever $\mathbf{H}$ is positive semi-definite. However, there are two potential issues with this approach.

First, the sample covariance matrix for the between-subject covariance $\widehat{\mathbf{K}}_B$ and the corresponding $\mathbf{H}$ matrix are not positive semi-definite. Therefore, FACE is not directly applicable. One solution was proposed by Xiao et al., 2016, who truncated negative eigenvalues to zero. The second issue the computation of $\mathbf{H}^{1/2}$, which is of dimension $n \times n$. When the total number of curves, n, is large, calculating $\mathbf{H}^{1/2}$ can become challenging.

To address the first problem, we smooth the total and within-subject covariance, which are positive semi-definite. The between-subject covariance is then estimated by the difference between total and within-subject covariance. To address the second problem, we provide an analytic form for the transformed functional data. We provide the technical details below.

As in Di et al., 2009, an empirical estimator of the between-subject covariance is given by the $L \times L$ matrix $\widehat{\mathbf{K}}_B = n_I^{-1}\widetilde{\mathbf{Y}}\mathbf{H}_B\widetilde{\mathbf{Y}}^T$, where $\mathbf{H}_B = \text{blockdiag}(\mathbf{1}_{J_1}\mathbf{1}_{J_1}^T - \boldsymbol{I}_{J_1}, \ldots, \mathbf{1}_{J_I}\mathbf{1}_{J_I}^T - \boldsymbol{I}_{J_I})$, $\boldsymbol{I}_J$ is the identity matrix of size J and $\mathbf{1}_J = (1, 1, \ldots, 1)_J^T$. As each block matrix $\mathbf{1}_{J_i}\mathbf{1}_{J_i}^T - \boldsymbol{I}_{J_i}$ has only two different eigenvalues, $J_i - 1$ (with geometric multiplicity 1) and -1 (with geometric multiplicity $J_i - 1$), $\mathbf{H}_B$ is not a positive semi-definite matrix. Because most eigenvalues are equal to -1, this may be the reason why the estimation performance was found to be sub-optimal when trimming negative eigenvalues of $\mathbf{H}_B$ (Xiao et al., 2016). We next focus on the total and within-subject covariance.

For the total covariance, let $\bar{J} = n^{-1}\sum_i J_i$ be the average number of visits per subject. The MoM estimator of the total covariance, $K_T(s,t)$, is $\widehat{K}_T(s,t) = \sum_{i=1}^{n}\sum_{j=1}^{J_i} w_i \widetilde{Y}_{ij}(s)\widetilde{Y}_{ij}(t)$, where $w_i > 0$ are weights that satisfy the constraint $\sum_i J_i w_i = 1$. In Di et al., 2009, the same weight is used for each visit, which

means $w_i = 1/(n\bar{J})$, though other weights could be used. For example, by setting $w_i = 1/(IJ_i)$, equal weights are assigned to study participants instead of visits. The matrix format of $\widehat{K}_T(s,t)$ is $\widehat{\mathbf{K}}_T = \sum_{i=1}^{I}\sum_{j=1}^{J_i} w_i \widetilde{\mathbf{Y}}_{ij}\widetilde{\mathbf{Y}}_{ij}^T$. The key insight is that $\widehat{\mathbf{K}}_T$ is the sample covariance of the transformed functional data $\{\sqrt{nw_i}\widetilde{\mathbf{Y}}_{ij}, 1 \le j \le J_i, 1 \le i \le I\}$. Therefore, smoothing $\widehat{\mathbf{K}}_T$ can be achieved by applying FACE to the transformed functional data.

For the within-subject covariance, notice that $K_W(s,t) = E\{\widetilde{Y}_{ij}(s) - \widetilde{Y}_{ik}(s)\}\{\widetilde{Y}_{ij}(t) - \widetilde{Y}_{ik}(t)\}^T/2$ if $j \neq k$. Let $v_i \ge 0$ be weights such that $\sum_{i=1}^{I} J_i(J_i - 1)v_i = 1$. An estimator of $K_W(s,t)$ is $\widehat{\mathbf{K}}_W = \sum_{i=1}^{I} v_i/2\sum_{j\neq k}(\widetilde{\mathbf{Y}}_{ij} - \widetilde{\mathbf{Y}}_{ik})(\widetilde{\mathbf{Y}}_{ij} - \widetilde{\mathbf{Y}}_{ik})^T$. The constraint on the weights ensures that when functional data are observed without random noise, $\sigma^2 = 0$, $\widehat{\mathbf{K}}_W$ is an unbiased estimator of K_W. If the same weight is used for each visit, then $v_i = n_I^{-1}$. If the same weight is used for each participant, then $v_i = \{(\sum_{l=1}^{I}\delta_{\{J_l \ge 2\}})J_i(J_i - 1)\}^{-1}$ if $J_i \ge 2$ and 0 otherwise. Let $\bar{\mathbf{Y}}_{i\cdot} = J_i^{-1}(\sum_{j=1}^{J_i}\widetilde{\mathbf{Y}}_{ij})$. It can be shown that $\widehat{\mathbf{K}}_W = \sum_{i=1}^{I}\sum_{j=1}^{J_i} v_i J_i(\widetilde{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i\cdot})(\widetilde{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i\cdot})^T$, which is the sample covariance of the transformed data $\{\sqrt{nv_iJ_i}(\widetilde{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i\cdot}), 1 \le j \le J_i, 1 \le i \le I\}$. Therefore, smoothing of $\widehat{\mathbf{K}}_W$ can be achieved by applying FACE to the above transformed data.

The construction of transformed functional data for the total and within-subject covariance require $O(nL)$ operations. This is a critical difference from the traditional MoM estimators, which require $O(nL^2)$ operations.

3.3.3 Multilevel FACE

We apply FACE to the transformed functional data for the total covariance and within-subject covariance, respectively, and obtain the smooth estimates. The corresponding eigenfunctions are obtained as described in the FACE approach for univarate functional data. Let $\widetilde{K}_T(s,t) = \mathbf{B}^T(s)\mathbf{\Theta}_T\mathbf{B}(t)$ be the estimate of $K_T(s,t)$ and $\widetilde{K}_W(s,t) = \mathbf{B}^T(s)\mathbf{\Theta}_W\mathbf{B}(t)$ be the estimate of $K_W(s,t)$. Here $\mathbf{\Theta}_T$ and $\mathbf{\Theta}_W$ are both $c \times c$ positive semi-definite matrices obtained from FACE. The between-subject covariance $K_B(s,t)$ is estimated by $\widetilde{K}_B(s,t) = \mathbf{B}(s)^T\mathbf{\Theta}_B\mathbf{B}(t)$, where $\mathbf{\Theta}_B = \mathbf{\Theta}_T - \mathbf{\Theta}_W$. To ensure that $\widetilde{K}_B(s,t)$ is positive semi-definite, an eigendecomposition of $\mathbf{\Theta}_B$ is taken and the eigenvectors associated with negative eigenvalues are discarded. For details, see Section S.2 of the supplementary material.

3.3.4 Score Prediction via Mixed Model Equations

Predicting the principal component scores via best linear unbiased prediction (BLUP) requires the inversion of matrices that are of dimension equal to the number of observations per curve, L. Here we propose a novel solution based on mixed model equations (MME), which further reduces the computational complexity.

Assuming the level-1 eigenfunctions $\phi_k(s)$ and level-2 eigenfunctions $\psi_k(s)$ are known, the multilevel functional model becomes the mixed effects model

$$\widetilde{Y}_{ij}(s) = \sum_{k_1 \geq 1} \xi_{ik_1}\phi_{k_1}(s) + \sum_{k_2 \geq 1} \zeta_{ijk_2}\psi_{k_2}(s) + \epsilon_{ij}(s), \tag{3.1}$$

where ξ_{ik_1}, ζ_{ijk_2} are uncorrelated scores that are uncorrelated with the $\epsilon_{ij}(s)$.

After obtaining $\phi_k(s)$ and $\psi_k(s)$ by multilevel FACE in Section 3.3.3, Equation (3.1) can be approximated by

$$\widetilde{Y}_{ij}(s) = \sum_{k_1=1}^{N_1} \xi_{ik_1}\phi_{k_1}(s) + \sum_{k_2=1}^{N_2} \zeta_{ijk_2}\psi_{k_2}(s) + \epsilon_{ij}(s),$$

where we have retained N_1 level-1 scores and N_2 level-2 scores.

Define $M_i = J_i L$. Let $\widetilde{\mathbf{Y}}_{ij} = [\widetilde{Y}_{ij}(s_1), \ldots, \widetilde{Y}_{ij}(s_L)]^T$, $\boldsymbol{\xi}_i = [\xi_{i1}, \ldots, \xi_{iN_1}]^T$, $\boldsymbol{\zeta}_{ij} = [\zeta_{ij1}, \ldots, \zeta_{ijN_2}]^T$, $\boldsymbol{\phi}_{k_1} = [\phi_{k_1}(s_1), \ldots, \phi_{k_1}(s_L)]^T$, $\boldsymbol{\psi}_{k_2} = [\psi_{k_2}(s_1), \ldots, \psi_{k_2}(s_L)]^T$, $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{N_1}] \in \mathbb{R}^{L\times N_1}$, $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_{N_2}] \in \mathbb{R}^{L\times N_2}$, and $\boldsymbol{\epsilon}_{ij} = [\epsilon_{ij1}, \ldots, \epsilon_{ijL}]^T$. Then $\widetilde{\mathbf{Y}}_{ij} = \boldsymbol{\Phi}\boldsymbol{\xi}_i + \boldsymbol{\Psi}\boldsymbol{\zeta}_{ij} + \boldsymbol{\epsilon}_{ij}$. We further define $\boldsymbol{\Phi}_i = \mathbf{1}_{J_i} \otimes \boldsymbol{\Phi} \in \mathbb{R}^{M_i\times N_1}$, $\boldsymbol{\Psi}_i = \mathbf{I}_{J_i} \otimes \boldsymbol{\Psi} \in \mathbb{R}^{M_i\times(J_iN_2)}$, $\boldsymbol{\zeta}_i = [\boldsymbol{\zeta}_{i1}^T, \ldots, \boldsymbol{\zeta}_{iJ_i}^T]^T \in \mathbb{R}^{J_iN_2}$, $\widetilde{\mathbf{Y}}_i = [\widetilde{\mathbf{Y}}_{i1}^T, \ldots, \widetilde{\mathbf{Y}}_{iJ_i}^T]^T \in \mathbb{R}^{M_i}$, and $\boldsymbol{\epsilon}_i = [\boldsymbol{\epsilon}_{i1}^T, \ldots, \boldsymbol{\epsilon}_{iJ_i}^T]^T \in \mathbb{R}^{M_i}$. The covariance matrix is $\boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_1^{(1)}, \ldots, \lambda_{N_1}^{(1)})$ for $\boldsymbol{\xi}_i$, $\boldsymbol{\Lambda}_2 = \mathrm{diag}(\lambda_1^{(2)}, \ldots, \lambda_{N_2}^{(2)})$ for $\boldsymbol{\zeta}_{ij}$, and $\sigma^2\mathbf{I}_{M_i}$ for $\boldsymbol{\epsilon}_i$. We then have the matrix form of the mixed effects model

$$\widetilde{\mathbf{Y}}_i = \boldsymbol{\Phi}_i\boldsymbol{\xi}_i + \boldsymbol{\Psi}_i\boldsymbol{\zeta}_i + \boldsymbol{\epsilon}_i,$$

$$\mathrm{E}\begin{pmatrix}\boldsymbol{\xi}_i\\ \boldsymbol{\zeta}_i\\ \boldsymbol{\epsilon}_i\end{pmatrix} = \begin{pmatrix}\mathbf{0}_{N_1}\\ \mathbf{0}_{J_iN_2}\\ \mathbf{0}_{M_i}\end{pmatrix}, \quad \mathrm{Cov}\begin{pmatrix}\boldsymbol{\xi}_i\\ \boldsymbol{\zeta}_i\\ \boldsymbol{\epsilon}_i\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Lambda}_1 & 0 & 0\\ 0 & \mathbf{I}_{J_i}\otimes\boldsymbol{\Lambda}_2 & 0\\ 0 & 0 & \sigma^2\mathbf{I}_{M_i}\end{pmatrix}. \tag{3.2}$$

It follows that the BLUP of $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ is

$$\begin{pmatrix}\widehat{\boldsymbol{\xi}}_i\\ \widehat{\boldsymbol{\zeta}}_i\end{pmatrix} = \begin{pmatrix}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_i^T\\ (\mathbf{I}_{J_i}\otimes\boldsymbol{\Lambda}_2)\boldsymbol{\Psi}_i^T\end{pmatrix}\left\{\mathrm{cov}(\widetilde{\mathbf{Y}}_i)\right\}^{-1}\widetilde{\mathbf{Y}}_i, \tag{3.3}$$

where $\mathrm{cov}(\widetilde{\mathbf{Y}}_i) = \boldsymbol{\Phi}_i\boldsymbol{\Lambda}_1\boldsymbol{\Phi}_i^T + \boldsymbol{\Psi}_i(\mathbf{I}_{J_i}\otimes\boldsymbol{\Lambda}_2)\boldsymbol{\Psi}_i^T + \sigma^2\mathbf{I}_{M_i} \in \mathbb{R}^{M_i\times M_i}$.

When M_i is large, implementing equation (3.3) is difficult as the inverse of $\mathrm{cov}(\widetilde{\mathbf{Y}}_i)$ requires $O(M_i^3)$ calculations. In practice, it takes more than a day to predict scores when $I = 1000$, $J_i = J = 3$, $L = 1000$ using existing

methods. The mixed model equations (MME) (Henderson, 1973) can be used to re-express the scores as

$$\begin{pmatrix} \widehat{\boldsymbol{\xi}}_i \\ \widehat{\boldsymbol{\zeta}}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Phi}_i^T \boldsymbol{\Phi}_i + \sigma^2 \boldsymbol{\Lambda}_1^{-1} & \boldsymbol{\Phi}_i^T \boldsymbol{\Psi}_i \\ \boldsymbol{\Psi}_i^T \boldsymbol{\Phi}_i & \boldsymbol{\Psi}_i^T \boldsymbol{\Psi}_i + \sigma^2 \mathbf{I}_{J_i} \otimes \boldsymbol{\Lambda}_2^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Phi}_i^T \widetilde{\mathbf{Y}}_i \\ \boldsymbol{\Psi}_i^T \widetilde{\mathbf{Y}}_i \end{pmatrix}, \tag{3.4}$$

where the dimension of the matrix that is inverted is $(N_1 + J_i N_2)$, which is usually much smaller than M_i. As a result, the total computational time using equation (3.4) is reduced to $O(M_i(N_1 + J_i N_2)^2 + (N_1 + J_i N_2)^3)$, which is linear in L since $M_i = J_i L$. In addition, the matrix inverse can be computed using block-wise calculations. The equivalence of BLUP and random effects solutions in MME was shown in Henderson (1963).

3.3.5 Fast MFPCA Algorithm

We summarize the steps of the fast MFPCA method in Algorithm 2. When compared with Algorithm 1, there are major differences in Steps 2-4 and 6. Steps 2-4 of fast MFPCA avoid computations that involve construction, smoothing and eigendecomposition of high dimensional covariance matrices. Step 6 of fast MFPCA uses a faster approach to the prediction of scores. The algorithm was implemented in the function `mfpca.face()` and released in the `R` package `refund`.

3.3.6 Incomplete Data

As in the physical activity data, some data might be missing. Xiao et al. (2016) proposed an iterative approach for single-level functional data which consists

Algorithm 2 fast MFPCA

1. Estimate mean functions $\mu(s)$ and $\eta_j(s)$ and subtract them from the observed data.
2. Apply FACE to the transformed functional data $\{\sqrt{nw_i}\widetilde{\mathbf{Y}}_{ij}, 1 \leq j \leq J_i, 1 \leq i \leq I\}$ for total covariance.
3. Apply FACE to the transformed functional data $\{\sqrt{nv_i J_i}(\widetilde{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i\cdot}), 1 \leq j \leq J_i, 1 \leq i \leq I\}$ for within-subject covariance and obtain within-subject eigenfunctions/eigenvalues.
4. Calculate between-subject covariance from the difference between total and within-subject covariance and extract between-subject eigenfunctions/eigenvalues.
5. Estimate error variance σ^2 by $\hat{\sigma}^2 = \int_{\mathcal{S}}\{\hat{K}_T(s,s) - \tilde{K}_T(s,s)\}ds$.
6. Estimate scores by MME in equation (3.4).

of: (1) initializing the missing data by imputation from any smoother; (2) applying FACE to the data and impute missing data by their BLUP; (3) iterating step 2 until reaching convergence. They reported convergence usually within 10 iterations. As we rely on FACE, which was designed to deal with missing data, the problem is solved automatically in our approach. The MME for score prediction in (3.4) can be easily modified to work with observed data only and hence the details are omitted. We have found that this method works well for incomplete data; see simulation results in Section 3.5.

3.4 Asymptotic Theory

We establish the L_2 convergence rate of the proposed fast MFPCA method for estimating the between- and within-subject covariance functions and show that a parametric convergence rate can be achieved when functional data are densely observed. To simplify theoretical analysis, we assume that $\mu(s)$ and $\eta_j(s)$ are known and that the study participants have the same number of

visits, $J_i = J$. In this case the weights are $w_i = (nJ)^{-1}$ for estimating the total covariance function and $v_i = \{nJ(J-1)\}^{-1}$ for estimating the within-subject covariance function. Finally, we use the same smoothing parameter λ for both $\widetilde{K}_T$ and $\widetilde{K}_W$, the smooth estimates from fast MFPCA.

We introduce some notation. The little o and big O notation are with respect to the number of study participants I and we allow the number of observations per curve L to increase with I. For two scalars a and b, let $a \wedge b = \min(a, b)$ and $a \vee b = \max(a, b)$. For a bivariate continuous function g over $\mathcal{S}^2$ let $\|g\|_{L_2}$ be its L_2 norm. For an integer $p \geq 2$, let $\mathcal{C}^p(\mathcal{S}^2)$ be the class of bivariate functions such that if $K \in \mathcal{C}^p(\mathcal{S}^2)$, then for any $0 \leq j \leq p$, $\partial^p K(s,t)/\partial s^j \partial t^{p-j}$ is continuous in $\mathcal{S}^2$.

Assumption 1 *(a). The random functions Z_i are independent across the subjects with zero-mean function and the same covariance function $K_B(s,t)$; (b). The random functions W_{ij} are independent across i and j with zero-mean function and the same covariance function $K_W(s,t)$; (c). The random errors $\epsilon_{ij\ell} = \epsilon_{ij}(s_\ell)$ are independent across i, j and ℓ with zero-mean and the same variance $\sigma_\epsilon^2 < \infty$; (d). The random functions Z_i, $\{W_{i1}, \ldots, W_{iJ_i}\}$, and the random errors $\{\epsilon_{ij1}, \ldots, \epsilon_{ijL}\}$ are mutually independent across the subjects.*

Assumption 2 $\sup_{s\in\mathcal{S}} \mathbb{E}[Z_i^4(s)] < \infty$; $\sup_{s\in\mathcal{S}} \mathbb{E}[W_{ij}^4(s)] < \infty$; $\mathbb{E}[\epsilon_{ij\ell}^4] < \infty$.

Assumption 3 *(a). $L \geq I^{\delta_1}$ for some constant $\delta_1 > 0$; (b). $c \geq I^{\delta_2}$ for some constant $\delta_2 > 0$ and $c = o(I)$; (c). There exists a sufficiently small constant δ_3 such that $c \leq \delta_3 L$; (d). $\lambda = o(I^{-2q\delta_1})$.*

Recall that m is the order of spline functions, c is the number of spline functions, and q is the order of smoothness penalty. Let $h = c^{-1}$ and $h_e = h \vee \lambda^{1/(2q)}$, and the latter is the effective bandwidth for penalized splines (Xiao, 2019). The proof for the theorem below is given in Section S.3 of the supplementary material.

Theorem 1 *Suppose that Assumptions 1 - 3 hold. If $K_B \in \mathcal{C}^p(\mathcal{S}^2)$ and $K_W \in \mathcal{C}^p(\mathcal{S}^2)$ with $q \leq p \wedge m$, then*

$$\mathbb{E}\left(\|\widetilde{K}_B - K_B\|_{L_2}^2\right) = O(L^{-2}h_e^{-1}) + O(h^{2m}) + o(h^{2p}) + O(\lambda^2 h_e^{-2q}) + O(I^{-1}),$$

$$\mathbb{E}\left(\|\widetilde{K}_W - K_W\|_{L_2}^2\right) = O(h^{2m}) + o(h^{2p}) + O(\lambda^2 h_e^{-2q}) + O(I^{-1}).$$

Except for the term $O(L^{-2}h_e^{-1})$, the derived rate in Theorem 1 is the same as those in Theorem 4.1 in Xiao, 2020, which considers covariance function estimation using penalized splines for functional data with a fixed common design. The convergence rate for estimating the between-subject covariance contains the term $O(L^{-2}h_e^{-1})$, which is due to the bias in the MoM estimator of the total covariance, which involves the extra variance, σ^2, along the main diagonal. The convergence rate for estimating the within-subject covariance does not have the term $O(L^{-2}h_e^{-1})$ because the empirical estimate $\widehat{K}_W$ is an unbiased estimate of K_W.

In both of the derived rates in Theorem 1, the term $O(h^{2m}) + o(h^{2p})$ is the approximation bias of spline functions, the term $O(\lambda^2 h_e^{-2q})$ is the shrinkage bias due to the smoothness penalty, and the term $O(I^{-1})$ is the variability of the estimate. To achieve a parametric rate of $O(I^{-1})$ for $\widetilde{K}_W$, a simple choice is to let $q = p$ and $h = O(I^{-1/(2p)})$. The second condition on h means that

the method could achieve a parametric rate as long as the number of knots (or spline functions) is sufficiently large. As for $\widetilde{K}_B$, the parametric rate can also be achieved if the condition $L^{-2}h_e^{-1} = O(I^{-1})$ also holds. This means that a sufficiently dense sampling design for each function is required. The additional condition seems reasonable for high-dimensional functional data. When L is small and the condition $L^{-2}h_e^{-1} = O(I^{-1})$ becomes stringent, one could smooth the empirical estimate $\widehat{K}_T$ without its diagonal terms or we replace the diagonal terms by other estimates with negligible bias.

While these theoretical results are for multilevel functional data, the proofs can be applied to other functional data and the convergence rate of the FACE method for single-level functional data has also been derived; see Section S.3 of the supplement for more details.

3.5 Simulation Studies

We perform simulations to: (1) assess the computational improvement and scaling behavior of fast MFPCA; and (2) evaluate the estimation accuracy of the fast MFPCA method. The `mfpca.face()` function is provided in the supplementary material and published in the `refund` package. For the implementation of traditional MFPCA we use the `mfpca.sc()` function in the `refund` package. For fast MFPCA, we use the same weight for each visit (curve), the same as traditional MFPCA. Additional simulation results for fast MFPCA with equal weight per subject are given in Section S.4 of the supplementary material.

3.5.1 Simulation Settings

We assume that the functions are observed on an equally-spaced grid $\{s_1, ..., s_L\}$ of domain $\mathcal{S} = [0, 1]$ such that $s_l = l/L$ for $l = 1, \ldots, L$. We also consider scenarios with incomplete (missing) data, where the number of observed points per function is $T_{ij} = 0.5 \times L$. Denote by J the mean number of visits per subject. For each subject i, the number of visits J_i is either balanced ($J_i = J$) or unbalanced (J_i varies by subject). We use a similar simulation setting as in Di et al. (2009). For visit j of subject i consider the following model

$$Y_{ij}(s_l) = \sum_{k_1=1}^{4} \xi_{ik_1}\phi_{k_1}(s_l) + \sum_{k_2=1}^{4} \zeta_{ijk_2}\psi_{k_2}(s_l) + \epsilon_{ij}(s_l),$$

where $\xi_{ik_1} \sim \mathcal{N}\{0, \lambda^{(1)}_{k_1}\}$, $\zeta_{ijk_2} \sim \mathcal{N}\{0, \lambda^{(2)}_{k_2}\}$, $\epsilon_{ij}(s) \sim \mathcal{N}(0, \sigma^2)$. We assume that there are $N_1 = 4$ components at the between-subject level and $N_2 = 4$ components at the within-subject level. Higher ranks on both levels with higher frequency eigenfunctions were further evaluated for both methods and the results are shown in Section S.4 of the supplementary material. The true eigenvalues are $\lambda^{(1)}_{k_1} = 0.5^{k_1-1}, k_1 = 1, 2, 3, 4$ and $\lambda^{(2)}_{k_2} = 0.5^{k_2-1}, k_2 = 1, 2, 3, 4$. The true eigenfunctions are selected as

Level 1: $\phi_{k_1}(s) = \{\sqrt{2}\sin(2\pi s), \sqrt{2}\cos(2\pi s), \sqrt{2}\sin(4\pi s), \sqrt{2}\cos(4\pi s)\}$.

Level 2: $\psi_{k_2}(s) = \{1, \sqrt{3}(2s-1), \sqrt{5}(6s^2-6s+1), \sqrt{7}(20s^3-30s^2+12s-1)\}$.

The eigenfunctions within levels 1 and 2 are mutually orthogonal, but they are not orthogonal between levels. We fix $\sigma = 1$, as the difference on computation time is marginal for different noise levels. We consider the following sample

size parameters: (1) number of subjects: $I \in \{100, 200, 1000, 5000\}$; (2) number of visits per subject: $J \in \{2, 4, 20, 100\}$. For unbalanced design, the number of visits J_i is drawn from Poisson(J) with a minimum of 1 visit for subject i; and (3) dimension of the functional domain: $L \in \{100, 200, 1000, 5000, 50000\}$. To reduce computational burden, we set the baseline as $\{I = 100, J = 2, L = 100\}$ and increase the sample size one at a time while fixing the others. For example, we fix $J = 2, L = 100$ and increase I from 100 to 5000. This gives a total of $2 \times 2 \times (4 + 3 + 4) = 44$ simulation scenarios. For each scenario we conduct 100 replications on a high performance computing cluster using 1 core per simulation. The computation time of fast MFPCA (`mfpca.face`) and MFPCA (`mfpca.sc`) is obtained under different scenarios. In addition, we derive the estimation accuracy of both methods by calculating MISE$(\mathbf{Y})$, MISE$(\boldsymbol{\phi}) = (N_1 L)^{-1}||\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}||_F^2$ and MISE$(\boldsymbol{\psi}) = (N_2 L)^{-1}||\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}||_F^2$ for each simulation.

3.5.2 Simulation Results

Table 3.1: Simulation results for different I when $J = 2$ and $L = 100$. The computation time ("Time(s)"), MISE of $\mathbf{Y}$ ("MISE($\mathbf{Y}$)") and eigenfunctions ("MISE($\boldsymbol{\phi}$)", "MISE($\boldsymbol{\psi}$)") reported in the table are median values across 100 replications.

Design	I	Method	Balanced				Unbalanced			
			Time(s)	MISE($\mathbf{Y}$)	MISE($\boldsymbol{\phi}$)	MISE($\boldsymbol{\psi}$)	Time(s)	MISE($\mathbf{Y}$)	MISE($\boldsymbol{\phi}$)	MISE($\boldsymbol{\psi}$)
Complete	100	fast MFPCA	**1.81**	0.9450	0.0781	0.0319	**1.74**	0.9450	0.1203	0.0416
		MFPCA	**14.82**	0.9451	0.0797	0.0315	**21.51**	0.9571	0.1930	0.1286
	200	fast MFPCA	**1.70**	0.9469	0.0413	0.0182	**1.88**	0.9465	0.0469	0.0229
		MFPCA	**25.42**	0.9474	0.0377	0.0171	**48.90**	0.9530	0.0780	0.0698
	1000	fast MFPCA	**2.20**	0.9505	0.0093	0.0075	**2.07**	0.9524	0.0120	0.0063
		MFPCA	**109.15**	0.9501	0.0073	0.0042	**262.76**	0.9525	0.0146	0.0176
	5000	fast MFPCA	**4.60**	0.9506	0.0034	0.0043	**5.31**	0.9516	0.0037	0.0046
		MFPCA	**540.84**	0.9503	0.0019	0.0008	**943.56**	0.9514	0.0036	0.0033
Incomplete	100	fast MFPCA	**2.16**	0.9003	0.0942	0.0348	**2.03**	0.9010	0.1570	0.0461
		MFPCA	**3.58**	0.8921	0.0999	0.0614	**5.00**	0.9089	0.2496	0.1644
	200	fast MFPCA	**2.34**	0.9028	0.0554	0.0198	**2.46**	0.9056	0.0671	0.0278
		MFPCA	**4.80**	0.8950	0.0489	0.0303	**8.34**	0.9045	0.0917	0.0857
	1000	fast MFPCA	**5.48**	0.9087	0.0230	0.0081	**5.83**	0.9074	0.0246	0.0074
		MFPCA	**13.19**	0.9037	0.0102	0.0064	**29.46**	0.9055	0.0179	0.0194
	5000	fast MFPCA	**19.99**	0.9077	0.0147	0.0048	**22.21**	0.9086	0.0150	0.0051
		MFPCA	**58.19**	0.9041	0.0022	0.0012	**134.49**	0.9066	0.0043	0.0037

Table 3.2: Simulation results for different J when $I = 100$ and $L = 100$. The computation time ("Time(s)"), MISE of $\mathbf{Y}$ ("MISE($\mathbf{Y}$)") and eigenfunctions ("MISE($\boldsymbol{\phi}$)", "MISE($\boldsymbol{\psi}$)") reported in the table are median values across 100 replications. Computation time more than 24 hours is denoted as ∞.

Design	J	Method	Balanced				Unbalanced			
			Time(s)	MISE($\mathbf{Y}$)	MISE($\boldsymbol{\phi}$)	MISE($\boldsymbol{\psi}$)	Time(s)	MISE($\mathbf{Y}$)	MISE($\boldsymbol{\phi}$)	MISE($\boldsymbol{\psi}$)
Complete	2	fast MFPCA	**1.81**	0.9450	0.0781	0.0319	**1.74**	0.9450	0.1203	0.0416
		MFPCA	**14.82**	0.9451	0.0797	0.0315	**21.51**	0.9571	0.1930	0.1286
	4	fast MFPCA	**1.98**	0.9532	0.0547	0.0126	**1.80**	0.9528	0.0634	0.0171
		MFPCA	**46.93**	0.9537	0.0483	0.0099	**75.87**	0.9597	0.0800	0.0463
	20	fast MFPCA	**2.33**	0.9614	0.0364	0.0056	**2.66**	0.9618	0.0314	0.0054
		MFPCA	**10146.82**	0.9617	0.0317	0.0019	**10334.10**	0.9626	0.0346	0.0088
	100	fast MFPCA	**6.02**	0.9626	0.0335	0.0042	**7.15**	0.9632	0.0332	0.0043
		MFPCA	∞	-	-	-	∞	-	-	-
Incomplete	2	fast MFPCA	**2.16**	0.9003	0.0942	0.0348	**2.03**	0.9010	0.1570	0.0461
		MFPCA	**3.58**	0.8921	0.0999	0.0614	**5.00**	0.9089	0.2496	0.1644
	4	fast MFPCA	**2.48**	0.9094	0.0616	0.0142	**2.48**	0.9100	0.0784	0.0200
		MFPCA	**11.07**	0.9147	0.0615	0.0161	**17.14**	0.9173	0.0842	0.0544
	20	fast MFPCA	**5.88**	0.9158	0.0432	0.0060	**6.08**	0.9162	0.0393	0.0067
		MFPCA	**733.01**	0.9254	0.0325	0.0037	**903.30**	0.9262	0.0399	0.0099
	100	fast MFPCA	**21.44**	0.9159	0.0384	0.0044	**23.72**	0.9160	0.0375	0.0046
		MFPCA	∞	-	-	-	∞	-	-	-

Table 3.3: Simulation results for different L when $I = 100$ and $J = 2$. The computation time ("Time(s)"), MISE of $\mathbf{Y}$ ("MISE($\mathbf{Y}$)") and eigenfunctions ("MISE($\boldsymbol{\phi}$)", "MISE($\boldsymbol{\psi}$)") reported in the table are median values across 100 replications. Computation time more than 24 hours is denoted as ∞.

Design	L	Method	Balanced				Unbalanced			
			Time(s)	MISE($\mathbf{Y}$)	MISE($\boldsymbol{\phi}$)	MISE($\boldsymbol{\psi}$)	Time(s)	MISE($\mathbf{Y}$)	MISE($\boldsymbol{\phi}$)	MISE($\boldsymbol{\psi}$)
Complete	100	fast MFPCA	**1.81**	0.9450	0.0781	0.0319	**1.74**	0.9450	0.1203	0.0416
		MFPCA	**14.82**	0.9451	0.0797	0.0315	**21.51**	0.9571	0.1930	0.1286
	200	fast MFPCA	**1.67**	0.9722	0.0804	0.0277	**1.78**	0.9746	0.1193	0.0395
		MFPCA	**119.92**	0.9733	0.0785	0.0272	**334.66**	0.9830	0.1933	0.1398
	1000	fast MFPCA	**2.51**	0.9953	0.0758	0.0244	**3.06**	0.9976	0.1145	0.0368
		MFPCA	**16883.52**	0.9963	0.0784	0.0237	**24395.62**	1.0055	0.1794	0.1058
	5000	fast MFPCA	**7.79**	0.9990	0.0756	0.0246	**12.40**	1.0023	0.1103	0.0369
		MFPCA	∞	-	-	-	∞	-	-	-
	50000	fast MFPCA	**29.05**	1.0000	0.0767	0.0243	**68.71**	1.0028	0.1081	0.0362
		MFPCA	∞	-	-	-	∞	-	-	-
Incomplete	100	fast MFPCA	**2.16**	0.9003	0.0942	0.0348	**2.03**	0.9010	0.1570	0.0461
		MFPCA	**3.58**	0.8921	0.0999	0.0614	**5.00**	0.9089	0.2496	0.1644
	200	fast MFPCA	**2.52**	0.9515	0.0828	0.0291	**2.41**	0.9498	0.1308	0.0408
		MFPCA	**17.01**	0.9484	0.0914	0.0368	**26.18**	0.9579	0.2138	0.1426
	1000	fast MFPCA	**3.35**	0.9923	0.0811	0.0263	**3.86**	0.9949	0.1160	0.0363
		MFPCA	**1883.75**	0.9903	0.0788	0.0274	**2697.17**	1.0023	0.1889	0.1425
	5000	fast MFPCA	**10.09**	0.9985	0.0773	0.0247	**14.56**	1.0022	0.1147	0.0374
		MFPCA	∞	-	-	-	∞	-	-	-
	50000	fast MFPCA	**60.59**	1.0006	0.0761	0.0250	**91.23**	1.0034	0.1091	0.0368
		MFPCA	∞	-	-	-	∞	-	-	-

Tables 3.1-3.3 provide the simulation results for different scenarios. For each table, we only increase one parameter in the order by I, J, L, while fixing the others at their baseline, as discussed in Section 3.5.1. Within each table, we show the computation time ("Time(s)"), MISE of $\mathbf{Y}$ ("MISE($\mathbf{Y}$)") and MISE of eigenfunctions ("MISE($\boldsymbol{\phi}$)", "MISE($\boldsymbol{\psi}$)") using both methods for complete and incomplete data for balanced and unbalanced designs. For fast MFPCA

we display the results weighted by visits. The results weighted by subjects are in Section S.4 of the supplementary material.

When I is large, fast MFPCA achieves similar accuracy for eigenfunction estimation with MFPCA for a balanced design and performs slightly better when the data are unbalanced. From a computational perspective, both methods exhibit a linear increase in computation time with I, though fast MFPCA is still much faster than MFPCA. For example, for $I = 5000$ with complete data and unbalanced design fast MFPCA takes less than 6 seconds compared to 900 seconds for traditional MFPCA (Table **??**). The computational advantage of fast MFPCA is more pronounced when J and L increase; see Table 3.2 and Table 3.3. For example, for complete data with an unbalanced design, MFPCA takes at least 3 hours when $J = 20$ and more than a day when $J = 100$. In contrast, fast MFPCA takes 2.7 seconds for $J = 20$ and less than 8 seconds for $J = 100$. For large L, MFPCA slows down substantially and takes, for example, over 6 hours for an analysis of complete unbalanced data when $L = 1000$. In contrast, fast MFPCA takes fewer than 100 seconds for $L = 50000$ (MFPCA would simply not run on such large examples).

Figure 3.2 shows the estimated eigenvalues for the complete unbalanced data when $I = 1000$. True eigenvalues are shown as gray dashed lines, while results from 100 simulations are shown in red for fast MFPCA and in blue for MFPCA. The eigenvalue estimates of both levels are close to their nominal values, while the level-2 estimates have higher precision using both methods. For level-1 there is a slight bias for the third and fourth eigenvalues using both methods, while the first eigenvalue estimates appears to be more precise for

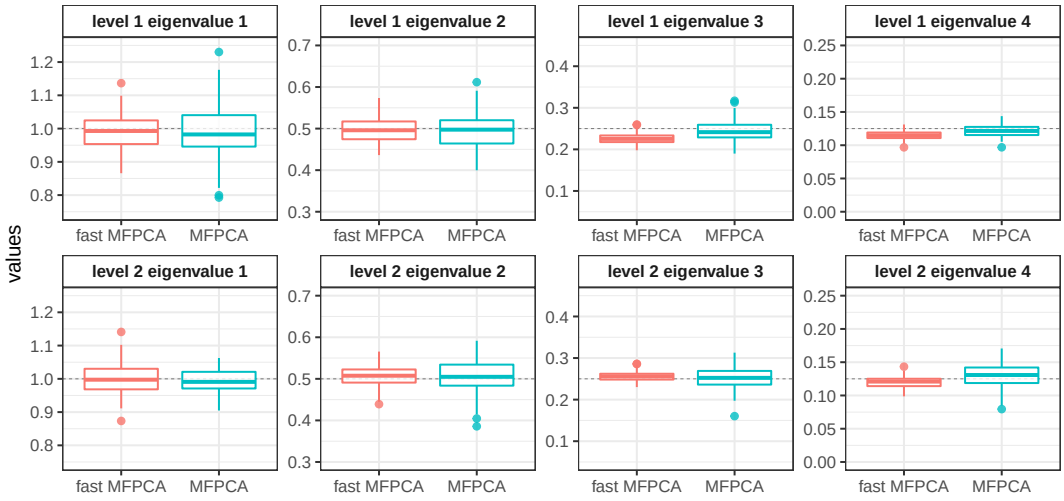


Figure 3.2: Boxplots of estimated eigenvalues from 100 replications when the data are complete with $I = 1000$, $J = 2$, $L = 100$ under unbalanced design for level-1 (first row) and level-2 (second row). True eigenvalues are shown as gray dashed lines, fast MFPCA are shown in red while MFPCA are shown in blue.

fast FPCA. In general, the eigenvalue estimates are accurate for large datasets using both approaches.

Figure 3.3 shows the estimated eigenfunctions under the same simulation setting. The top two rows display estimates from fast MFPCA in red and the bottom two rows display estimates from MFPCA in blue. Within each panel, the black solid curves indicate the true eigenfunctions at each level. For level-1 eigenfunctions, both methods exhibit similar accuracy. For level-2 eigenfunctions, we observe a larger variability for MFPCA, especially on the third and fourth eigenfunctions. This higher accuracy of fast MFPCA is also reflected by its slightly smaller MISE($\boldsymbol{\phi}$) (0.0120 vs. 0.0146) and much smaller MISE($\boldsymbol{\psi}$) (0.0063 vs. 0.0176) shown in Table 3.1.

In summary, fast MFPCA achieves similar estimation accuracy with MFPCA under different simulation settings, while the computation is at least two

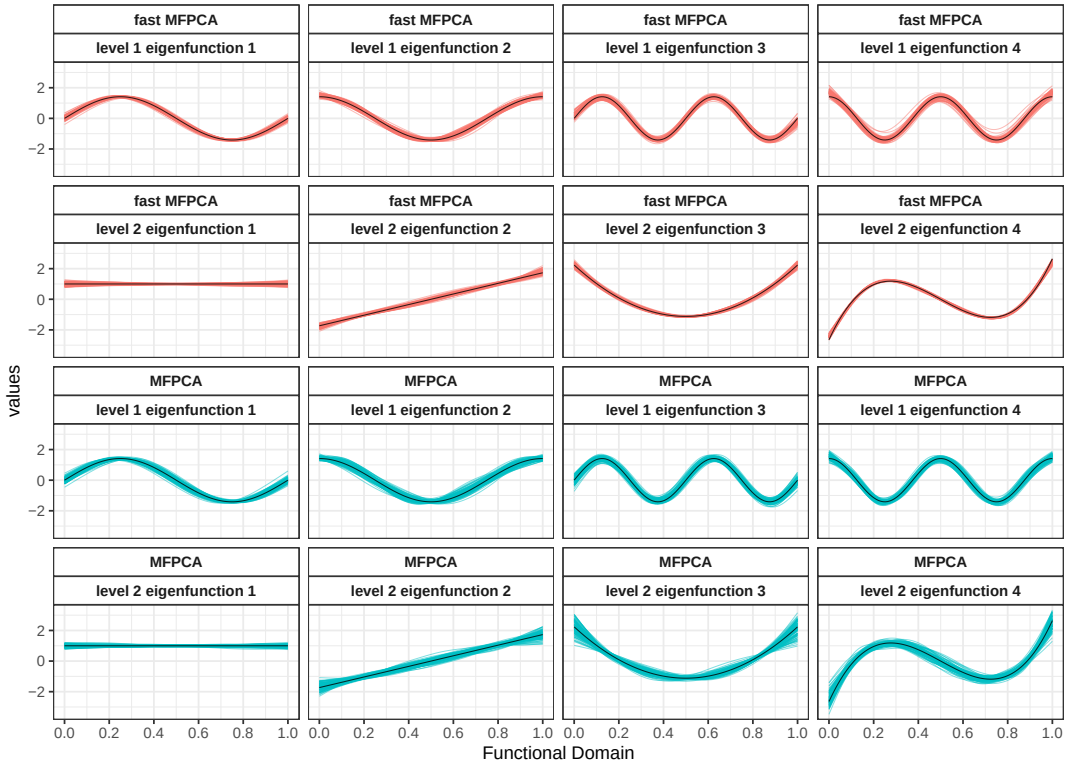


Figure 3.3: Estimated eigenfunctions for fast MFPCA (top two rows) and MFPCA (bottom two rows) when the data are complete with $I = 1000$, $J = 2$, $L = 100$ with unbalanced design. Within each model, the top row displays level-1 estimates and the bottom row displays level-2 estimates. Black lines: true eigenfunction; red lines: 100 fast MFPCA estimates; blue lines: 100 MFPCA estimates.

orders of magnitude faster. For a dataset with a large number of visits per subject ($J \sim 100$) or very high dimensions of the functional domain ($L \sim 50000$), fast MFPCA helps reduce the total computation time from several days or longer to just a few minutes.

3.6 Application

Objective physical activity measured by accelerometers and its association with health outcomes is an active area of research (Smirnova et al., 2020; Cui,

Crainiceanu, and Leroux, 2021; Cui et al., 2022). The National Health and Nutrition Examination Survey (NHANES) is a study conducted by the United States Centers for Disease Control and Prevention (CDC) with the aim of assessing the health and nutritional status of the US population. It became a continuous program conducted in two-year waves since 1999. The NHANES study collected accelerometry data using hip-worn physical activity monitors (ActiGraph model AM-7164) in the 2003-2004 and 2005-2006 waves. Both waves share the same protocol, where each study participant was asked to wear the device for 7 consecutive days. Data were released by the National Center for Health Statistics (NCHS) as minute-level activity counts (AC), a proprietary measure of physical activity intensity. We use the processed accelerometry data as described in Leroux et al. (2019). To reduce the severe skewness of the original data, for this analysis the minute-level AC were transformed into LAC $:= \log(1 + \text{AC})$, as suggested by Varma et al. (2017) and Varma et al. (2018). The 2003-2004 and 2005-2006 waves have a total of 14631 study participants with accelerometry data. Days with less than 10 hours of estimated wear time or days that were deemed by NHANES to have poor quality data were excluded. The final dataset has 12802 study participants and 65777 participant-days, with 1440 observations per day. The average number of available days per study participant is 5.14.

For this analysis, we are interested in decomposing the variability of the minute-level accelerometry data at both study participant (level-1) and day of the week (level-2) levels. While the problem is stated in simple terms, applying the existing MFPCA method to this large dataset takes at least 5 days

on a regular laptop (2.7GHz Dual-Core i5 Processor). In contrast, fast MFPCA took less than 5 minutes on the same laptop.

Figure 3.4 displays the estimated overall mean function $\mu(s)$ and the mean function for each day of the week $\mu(s) + \eta_j(s)$, $j = 1, \ldots, 7$. The weekend curves are shown as dashed lines, while the weekday curves are shown as dotted lines. The overall mean function exhibits a clear circadian rhythm. In addition, there are distinguishable weekend-weekday patterns, as the physical activity intensity is higher than average on Friday and Saturday nights and lower than average on Saturday and Sunday mornings. These results provide visual evidence of a weekend effect in the NHANES cohort.

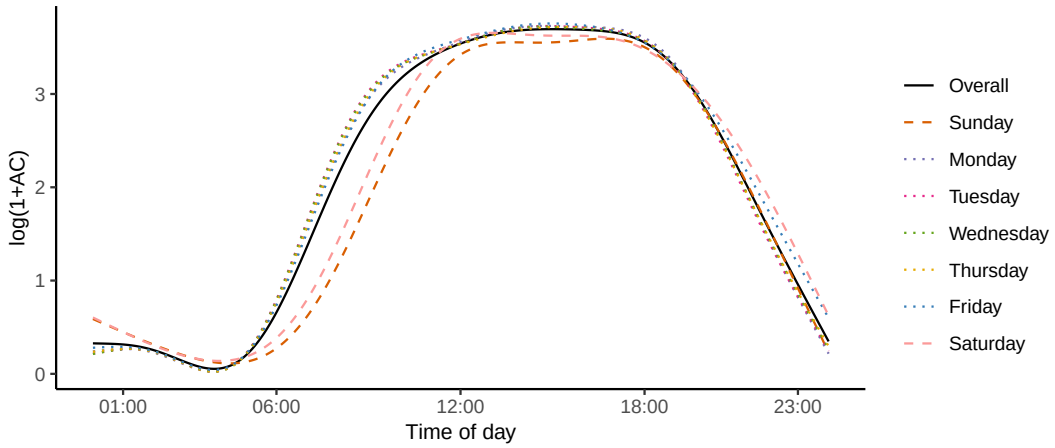
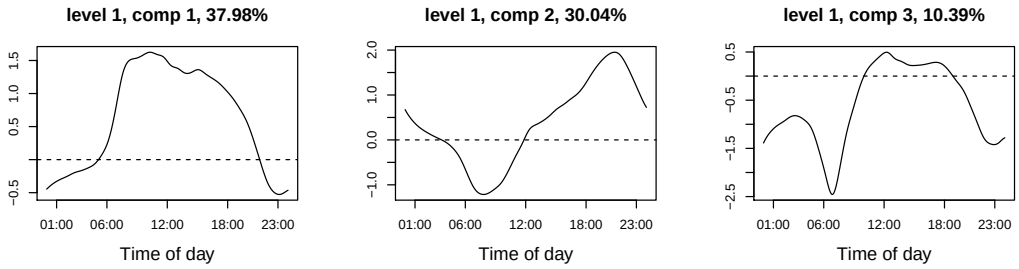


Figure 3.4: Estimated overall mean function $\mu(s)$ and day-of-the-week-specific mean function $\mu(s) + \eta_j(s)$ in the NHANES dataset using fast MFPCA. Overall mean curve: black solid line; weekend days means: dashed lines; weekday mean curves: dotted lines.

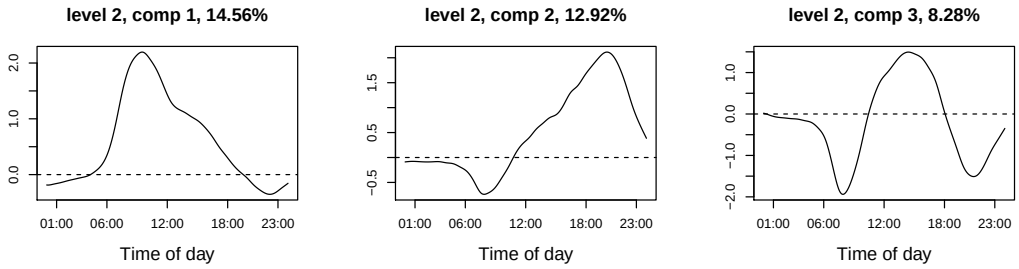
We identify 22 level-1 principal components and 31 level-2 principal components using the pre-specified percentage of variance explained (PVE) with a value of 0.99 at both levels. The total explained between-subject variance is 1.02. The total explained within-subject variance is 1.78, which is nearly

twice that of the between-subject variance. The proportion of variability explained by level 1 is 0.36, defined as $\sum_{k_1=1}^{\infty} \lambda_{k_1}^{(1)} / (\sum_{k_1=1}^{\infty} \lambda_{k_1}^{(1)} + \sum_{k_2=1}^{\infty} \lambda_{k_2}^{(2)})$ in Di et al. (2009). Figure 3.5a shows the first three estimated level-1 eigenfunctions $\phi_{k_1}^{(1)}(s)$ of the physical activity data, which explain 78.4% of the total variability. The first eigenfunction is negative at night and positive during the day, suggesting that study participants with positive scores on this component will have less activity at night and more activity during the day. The second eigenfunction is only negative during the morning (5am to 12pm), suggesting that study participants with positive scores on this component will have less activity in the morning and more activity during the rest of a day. Study participants with positive scores on the third component have less activity during working hours (10am to 6pm) and more activity at all other times of a day.

At level 2 the first 3 components explain only 35.8% of the level 2 variability. Figure 3.5b shows the first three estimated level-2 eigenfunctions $\psi_{k_2}(s)$. The interpretation is different, as level-2 characterizes within-subject behavior. Specifically, days of the week with positive scores on the first principal component correspond to lower physical activity at night and sharply higher in the morning compared to the average activity of the individual. Similarly, days of the week with positive scores on the second principal component correspond to lower activity during the morning and higher during the rest of the day compared to the average activity of the individual.



(a) The first three estimated level-1 eigenfunctions.



(b) The first three estimated level-2 eigenfunctions.

Figure 3.5: The top three estimated level-1 (first row) and level-2 (second row) eigenfunctions from the NHANES dataset using fast MFPCA. The proportion of variability explained in each principal component within each level is shown on the title of each panel.

3.7 Discussion

We propose fast MFPCA, which solves the major computational bottlenecks of the traditional MFPCA (Di et al., 2009), enabling it to be used on much larger and higher dimensional data sets. For example, the NHANES dataset contains minute-level physical activity information of more than 10000 study participants over multiple days. While applying MFPCA on such dataset takes more than 5 days on a regular laptop, the proposed fast MFPCA takes less than 5 minutes. The substantial computational improvement is due both to the development of new methods and to their careful coding. Simulation results

show that fast MFPCA achieves similar estimation accuracy with MFPCA, while the computation times are at least two orders of magnitude faster.

In this paper we only considered a dense design for functional data, the most common scenario. However, the extension of FACE (Xiao et al., 2018) to sparse designs suggests possible extensions to multilevel sparse functional data (Di, Crainiceanu, and Jank, 2014). Such extensions will be studied in future work.

3.8 Supplamentary Material

The supplementary material of this project is available at https://doi.org/10.1080/10618600.2022.2115500.

References

Di, Chongzhi, Ciprian M Crainiceanu, Brian S Caffo, and Naresh M Punjabi (2009). "Multilevel functional principal component analysis". In: *The Annals of Applied Statistics* 3.1, p. 458.

Ramsay, JO and BW Silverman (2005). *Functional data analysis*. New York: Springer.

Ramsay, James O and CJ Dalzell (1991). "Some tools for functional data analysis". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 53.3, pp. 539–561.

Silverman, Bernard W et al. (1996). "Smoothed functional principal components analysis by choice of norm". In: *The Annals of Statistics* 24.1, pp. 1–24.

Yao, Fang, Hans-Georg Müller, and Jane-Ling Wang (2005). "Functional data analysis for sparse longitudinal data". In: *Journal of the American Statistical Association* 100.470, pp. 577–590.

Brumback, Babette A and John A Rice (1998). "Smoothing spline models for the analysis of nested and crossed samples of curves". In: *Journal of the American Statistical Association* 93.443, pp. 961–976.

Morris, Jeffrey S and Raymond J Carroll (2006). "Wavelet-based functional mixed models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2, pp. 179–199.

Aston, John AD, Jeng-Min Chiou, and Jonathan P Evans (2010). "Linguistic pitch analysis using functional principal component mixed effect models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59.2, pp. 297–317.

Morris, Jeffrey S, Veerabhadran Baladandayuthapani, Richard C Herrick, Pietro Sanna, and Howard Gutstein (2011). "Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data". In: *The Annals of Applied Statistics* 5.2A, p. 894.

Chen, Kehui and Hans-Georg Müller (2012). "Modeling repeated functional observations". In: *Journal of the American Statistical Association* 107.500, pp. 1599–1609.

Serban, Nicoleta and Huijing Jiang (2012). "Multilevel functional clustering analysis". In: *Biometrics* 68.3, pp. 805–814.

Goldsmith, Jeff, Vadim Zipunnikov, and Jennifer Schrack (2015). "Generalized multilevel function-on-scalar regression and principal component analysis". In: *Biometrics* 71.2, pp. 344–353.

Li, Haocheng, Sarah Kozey Keadle, John Staudenmayer, Houssein Assaad, Jianhua Z Huang, and Raymond J Carroll (2015). "Methods to assess an exercise intervention trial based on 3-level functional data". In: *Biostatistics* 16.4, pp. 754–771.

Xu, Yuhang, Yehua Li, and Dan Nettleton (2018). "Nested hierarchical functional data modeling and inference for the analysis of functional plant phenotypes". In: *Journal of the American Statistical Association* 113.522, pp. 593–606.

Gaynanova, Irina, Naresh Punjabi, and Ciprian Crainiceanu (2022). "Modeling continuous glucose monitoring (CGM) data during sleep". In: *Biostatistics* 23.1, pp. 223–239.

Greven, Sonja, Ciprian Crainiceanu, Brian Caffo, and Daniel Salo Reich (2010). "Longitudinal functional principal component analysis". In: *Electronic Journal of Statistics* 4, pp. 1022–1054.

Zipunnikov, Vadim, Sonja Greven, Haochang Shou, Brian Caffo, Daniel S Reich, and Ciprian Crainiceanu (2014). "Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis". In: *The Annals of Applied Statistics* 8.4, pp. 2175–2202.

Park, So Young and Ana-Maria Staicu (2015). "Longitudinal functional data analysis". In: *Stat* 4.1, pp. 212–226.

Scheffler, Aaron, Donatello Telesca, Qian Li, Catherine A Sugar, Charlotte Distefano, Shafali Jeste, and Damla Şentürk (2020). "Hybrid principal components analysis for region-referenced longitudinal functional EEG data". In: *Biostatistics* 21.1, pp. 139–157.

Cui, Erjia, Andrew Leroux, Ekaterina Smirnova, and Ciprian M Crainiceanu (2021). "Fast univariate inference for longitudinal functional models". In: *Journal of Computational and Graphical Statistics* 31.1, pp. 1–12.

Boland, Joanna, Donatello Telesca, Catherine Sugar, Shafali Jeste, Cameron Goldbeck, and Damla Senturk (2022). "A study of longitudinal trends

in time-frequency transformations of EEG data during a learning experiment". In: *Computational Statistics & Data Analysis* 167, p. 107367.

Shamshoian, John, Damla Şentürk, Shafali Jeste, and Donatello Telesca (2022). "Bayesian analysis of longitudinal and multidimensional functional data". In: *Biostatistics* 23.2, pp. 558–573.

Li, Ruonan, Luo Xiao, Ekaterina Smirnova, Erjia Cui, Andrew Leroux, and Ciprian M Crainiceanu (2022). "Fixed-effects inference and tests of correlation for longitudinal functional data". In: *Statistics in Medicine*.

Zhang, Lin, Veerabhadran Baladandayuthapani, Hongxiao Zhu, Keith A Baggerly, Tadeusz Majewski, Bogdan A Czerniak, and Jeffrey S Morris (2016). "Functional CAR models for large spatially correlated functional datasets". In: *Journal of the American Statistical Association* 111.514, pp. 772–786.

Li, Yihao, Danh V Nguyen, Sudipto Banerjee, Connie M Rhee, Kamyar Kalantar-Zadeh, Esra Kürüm, and Damla Şentürk (2021). "Multilevel modeling of spatially nested functional data: Spatiotemporal patterns of hospitalization rates in the US dialysis population". In: *Statistics in Medicine* 40.17, pp. 3937–3952.

Shou, Haochang, Vadim Zipunnikov, Ciprian M Crainiceanu, and Sonja Greven (2015). "Structured functional principal component analysis". In: *Biometrics* 71.1, pp. 247–257.

Scheipl, Fabian, Ana-Maria Staicu, and Sonja Greven (2015). "Functional additive mixed models". In: *Journal of Computational and Graphical Statistics* 24.2, pp. 477–501.

Berrendero, José R, Ana Justel, and Marcela Svarc (2011). "Principal components for multivariate functional data". In: *Computational Statistics & Data Analysis* 55.9, pp. 2619–2634.

Chiou, Jeng-Min, Yu-Ting Chen, and Ya-Fang Yang (2014). "Multivariate functional principal component analysis: A normalization approach". In: *Statistica Sinica* 24.4, pp. 1571–1596.

Kowal, Daniel R, David S Matteson, and David Ruppert (2017). "A Bayesian multivariate functional dynamic linear model". In: *Journal of the American Statistical Association* 112.518, pp. 733–744.

Happ, Clara and Sonja Greven (2018). "Multivariate functional principal component analysis for data observed on different (dimensional) domains". In: *Journal of the American Statistical Association* 113.522, pp. 649–659.

Wong, Raymond KW, Yehua Li, and Zhengyuan Zhu (2019). "Partially linear functional additive models for multivariate functional data". In: *Journal of the American Statistical Association* 114.525, pp. 406–418.

Xiao, Luo, Vadim Zipunnikov, David Ruppert, and Ciprian Crainiceanu (2016). "Fast covariance estimation for high-dimensional functional data". In: *Statistics and Computing* 26.1, pp. 409–421.

Goldsmith, Jeff, Fabian Scheipl, Lei Huang, Julia Wrobel, Chongzhi Di, Jonathan Gellar, Jaroslaw Harezlak, Mathew W. McLean, Bruce Swihart, Luo Xiao, Ciprian Crainiceanu, and Philip T. Reiss (2020). *refund: Regression with functional data*. URL: https://CRAN.R-project.org/package=refund.

Eilers, P.H.C. and B.D. Marx (1996). "Flexible smoothing with B-splines and penalties (with Discussion)". In: *Statistical Science* 11, pp. 89–121.

Xiao, Luo, Yingxing Li, and David Ruppert (2013). "Fast bivariate P-splines: the sandwich smoother". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.3, pp. 577–599.

Henderson, Charles R (1973). "Sire evaluation and genetic trends". In: *Journal of Animal Science* 1973.Symposium, pp. 10–41.

Henderson, Charles R (1963). "Selection index and expected genetic advance". In: *Statisitical Genetics and Plant Breeding* 982, pp. 140–163.

Xiao, Luo (2019). "Asymptotic theory of penalized splines". In: *Electronic Journal of Statistics* 13.1, pp. 747–794. URL: https://doi.org/10.1214/19-EJS1541.

Xiao, Luo (2020). "Asymptotic properties of penalized splines for functional data". In: *Bernoulli* 26.4, pp. 2847–2875.

Smirnova, Ekaterina, Andrew Leroux, Quy Cao, Lucia Tabacu, Vadim Zipunnikov, Ciprian Crainiceanu, and Jacek K Urbanek (2020). "The predictive performance of objective measures of physical activity derived from accelerometry data for 5-year all-cause mortality in older adults: National Health and Nutritional Examination Survey 2003–2006". In: *The Journals of Gerontology: Series A* 75.9, pp. 1779–1785.

Cui, Erjia, Ciprian M Crainiceanu, and Andrew Leroux (2021). "Additive functional Cox model". In: *Journal of Computational and Graphical Statistics* 30.3, pp. 780–793.

Cui, Erjia, E Christi Thompson, Raymond J Carroll, and David Ruppert (2022). "A semiparametric risk score for physical activity". In: *Statistics in Medicine* 41.7, pp. 1191–1204.

Leroux, Andrew, Junrui Di, Ekaterina Smirnova, Elizabeth J Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu (2019). "Organizing and analyzing the activity data in NHANES". In: *Statistics in Biosciences* 11.2, pp. 262–287.

Varma, Vijay R, Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov (2017). "Re-evaluating the effect of age on physical activity over the lifespan". In: *Preventive Medicine* 101, pp. 102–108.

Varma, Vijay R, Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov (2018). "Total volume of physical activity: TAC, TLAC or TAC(λ)". In: *Preventive Medicine* 106, pp. 233–235.

Xiao, Luo, Cai Li, William Checkley, and Ciprian Crainiceanu (2018). "Fast covariance estimation for sparse functional data". In: *Statistics and Computing* 28.3, pp. 511–522.

Di, Chongzhi, Ciprian M Crainiceanu, and Wolfgang S Jank (2014). "Multilevel sparse functional principal component analysis". In: *Stat* 3.1, pp. 126–143.

Chapter 4

Fast Univariate Inference for Longitudinal Functional Data

4.1 Introduction

Longitudinal high dimensional data have become ubiquitous. For example, a Diffusion Tensor Imaging (DTI) study (Greven et al., 2010; Goldsmith et al., 2011; Goldsmith et al., 2012; Scheipl, Staicu, and Greven, 2015) collected fractional anisotropy (FA) at multiple locations along the corpus callosum at multiple visits for each study participant. FA is a measure of water diffusion direction that is thought to be associated with white matter integrity and myelination. Figure 4.1 displays the FA measures along the corpus callosum for two study participants (left and right panels). Each curve represents data collected during a particular visit (color-coded according to visit number). The visit time, covariates, and health outcomes are collected at each visit but are not displayed; for more details see Goldsmith et al. (2012). Visual inspection of the left panel does not indicate an obvious temporal trend across visits, while each FA profile exhibits substantial within-visit noise. The FA profiles are less

noisy for the second study participant (right panel), though the much lower FA values at the second visit could indicate substantial technical or biological variability. Given such data, the scientific goal is to quantify how FA at each location of the corpus callosum changes between visits and how these changes are associated with study-participant characteristics. To conceptualize the data structure, we denote by $Y_{ij}(s)$ the FA value for study participant i at visit j at time t_{ij} and location $s \in \mathcal{S}$ of the corpus callosum.

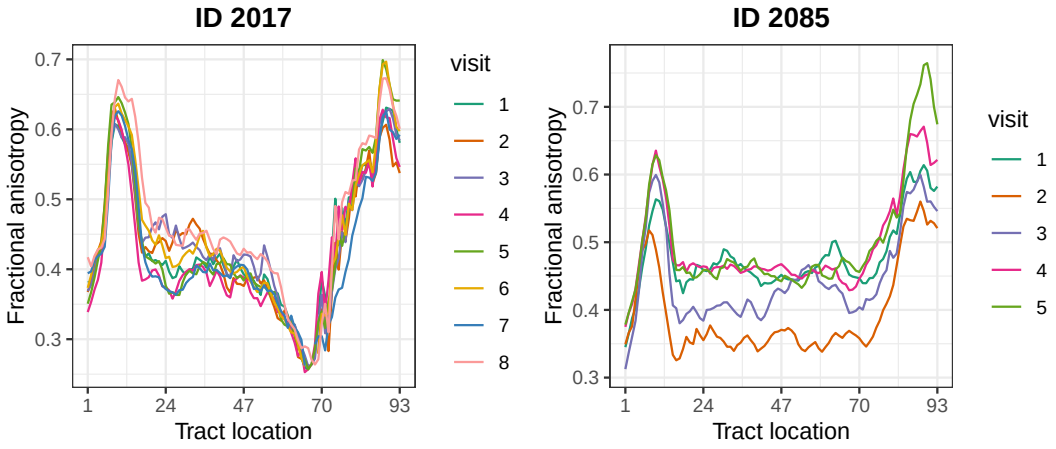


Figure 4.1: The fractional anisotropy (FA) tract profiles for the corpus callosum (functional domain) of two study participants in the DTI study. Left panel: ID 2017. Right Panel: ID 2085. For each study participant, each curve represents the tract profiles at one longitudinal visit. The visit number is color coded.

Another example is the National Health and Nutrition Examination Survey (NHANES) (Leroux et al., 2019; Smirnova et al., 2019; Cui, Crainiceanu, and Leroux, 2020), where minute-level activity counts (AC), a proprietary measure of physical activity (PA), were collected for up to seven consecutive days by hip-worn accelerometers. Data can be further transformed at the minute level to an active/inactive indicator depending on whether the observed AC is above/below a threshold. Here we use the threshold of 100, which was

proposed by Matthews et al. (2008) for the NHANES hip accelerometry data and has been used extensively in the literature (Koster et al., 2012). Therefore, the data are binary (active/inactive), functional (minute-level measurements during the day), and multilevel (multiple days). To conceptualize the data structure, $Y_{ij}(s)$ denotes the binary indicator of whether a study participant i was active on day j at minute $s \in \mathcal{S} = \{1, \ldots, 1440\}$. For the purpose of this paper, we are interested in studying the association between $Y_{ij}(s)$ and sex, age, day number from the beginning of the study, and day of the week.

Both these examples contain functional data measured at multiple visits, while the sample size in NHANES is over 10 times larger than the DTI study. Such data structures generalize standard longitudinal data, as instead of observing one scalar variable at each visit, one observes a high dimensional function. Functional data methods are sometimes used to model scalar longitudinal data (Yao, Müller, and Wang, 2005), but longitudinal functional data has a more complex structure due to the large number of observations with complex correlations *at each visit*. Methods have been proposed for functional data with complex correlation structures (Guo, 2002; Morris and Carroll, 2006; Greven et al., 2010; Zipunnikov et al., 2014; Scheipl, Staicu, and Greven, 2015; Brockhaus et al., 2015; Scheipl, Gertheiss, Greven, et al., 2016; Shou et al., 2015; Zhu et al., 2019). In particular, Scheipl, Staicu, and Greven (2015) and Scheipl, Gertheiss, Greven, et al. (2016) proposed an inferential framework and associated software for correlated functional responses based on additive mixed models. This is an important step forward, though the method cannot currently handle very large data sets. For example, it takes more than 24

hours on a regular laptop (2.7 GHz Dual-Core Intel Core i5, 8GB RAM) to fit a functional random intercept model with $I = 1000$ subjects and an average of 5 visits per subject. The approach of Goldsmith, Zipunnikov, and Schrack (2015) runs into similar scaling up problems. Indeed, they reported a total computation time of 10 days in their applications using a dataset with fewer than 600 subjects, 5 visits per subject, and 144 observations per curve. Applying either method to our NHANES example is computationally impractical.

Thus, we conclude that this area of research is still in its initial stages of statistical development. Indeed, there is an increased need for methods that: (1) are scalable both in terms of number of study participants and of the dimension of the functional data; (2) can be applied to Gaussian and non-Gaussian data; and (3) preserve the interpretability of standard mixed effects models. To address this need, we propose fast univariate inferential (FUI) approaches for longitudinal functional data of any type. The approach consists of three steps: (1) fit massively univariate pointwise mixed effects models; (2) apply a smoother along the functional domain; and (3) obtain joint confidence bands using analytic approaches for Gaussian data or a bootstrap of study participants for non-Gaussian data. The first two steps are conceptually similar with Fan and Zhang (2000) and Reiss et al. (2017) for function-on-scalar regressions. However, to the best of our knowledge, these approaches did not model correlated functional responses or provided joint inference that accounts for this correlation. The methods proposed by Park et al. (2018) are the closest to our methods, though there are important differences. First, they estimate the fixed effects under independence across-

and within-visits. We use mixed effects models across visits. Second, because Park et al. (2018) rely on fixed effects, their approach cannot be used to predict visit-specific functional effects or missing data. Third, their approach was developed for a narrower set of models.

The remainder of the paper is organized as follows. Section 4.2 introduces the longitudinal functional model. Section 4.3 describes two approaches for fixed effects inference, one analytic for Gaussian data and one based on the bootstrap of study participants for any type of data. We further discuss in Section 4.4 a simple and flexible simulation-based approach to obtain joint confidence bands. Section 4.5 presents the simulation results and comparisons with existing methods. Section 4.6 describes the applications of our model on DTI and NHANES study. We close with a discussion in Section 4.7. All code for model implementation, simulation and application is available on the supplementary material.

4.2 Massively Univariate Longitudinal Functional Model

Assume that data is of the type $Y_{ij}(s)$ on a grid $\{s_1, s_2, ..., s_L\}$ of the compact functional domain $\mathcal{S}$. Data can be Gaussian or non-Gaussian, $i = 1, 2, ..., I$ is the index of the study participant, and $j = 1, 2, ..., J_i$ is the index of the longitudinal visit at time t_{ij}. In addition to the functional outcomes, $Y_{ij}(s)$, $\boldsymbol{X}_{ij} = [X_{ij1}, X_{ij2}, ..., X_{ijp}]^T \in \mathbb{R}^p$ are the fixed and $\boldsymbol{Z}_{ij} = [Z_{ij1}, Z_{ij2}, ..., Z_{ijq}]^T \in \mathbb{R}^q$ are the random effects variables. We posit the following marginal three-step inferential approach:

First step: At each location $s_l \in \mathcal{S}$, $l = 1, 2, ..., L$, fit a separate pointwise generalized linear mixed model (GLMM) $Y_{ij}(s_l) \sim \text{EF}\{\mu_{ij}(s_l)\}$, where EF denotes the exponential family distribution, $\mu_{ij}(s_l)$ is the conditional mean, and

$$\eta_{ij}(s_l) = g\{\mu_{ij}(s_l)\} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}(s_l) + \boldsymbol{Z}_{ij}^T \boldsymbol{u}_i(s_l) \; . \tag{4.1}$$

Here $g(\cdot)$ is a link function and $\boldsymbol{u}_i(s_l)$ is a $q \times 1$ dimensional vector of random effects that depends on the location. Denote the estimates of fixed effects as $\tilde{\boldsymbol{\beta}}(s_1), ..., \tilde{\boldsymbol{\beta}}(s_L)$ and of the linear predictors as $\tilde{\eta}_{ij}(s_1), ..., \tilde{\eta}_{ij}(s_L)$ obtained from these univariate GLMMs. We refer to this as massively univariate analysis because a GLMM is fit at every location s_l, $l = 1, \ldots, L$, where L can be very large (hence, the use of the word "massive").

Second step: Smooth the estimated fixed effects $\tilde{\boldsymbol{\beta}}(s_1), ..., \tilde{\boldsymbol{\beta}}(s_L)$ and/or linear predictors $\tilde{\eta}_{ij}(s_1), ..., \tilde{\eta}_{ij}(s_L)$ along the functional domain. Denote these smooth estimators by $\{\hat{\boldsymbol{\beta}}(s), s \in \mathcal{S}\}$ and $\{\hat{\eta}_{ij}(s), s \in \mathcal{S}\}$, respectively. This can use any smoother that is or is not data adaptive, including not smoothing and taking the average over all locations.

Third step: Obtain joint confidence bands for fixed effects parameters and/or linear predictors using analytic approaches for Gaussian data or a bootstrap of study participants for Gaussian and non-Gaussian data.

The key insight of our method is to decompose the complex correlation structure into longitudinal and functional directions. The first step of the analysis is to use the familiar univariate GLMMs, a procedure that can be easily implemented using parallel computing. This substantially reduces the

computational burden of existing methods. The third step allows for joint inferences that take into account within- and between-visit correlations. This approach is not limited to estimating $\boldsymbol{\beta}(s)$ and $\eta_{ij}(s)$ and can be used for any measures of interest, including random effects, quantiles, and group means. We discuss the fixed effects inference of our approach below.

4.3 Fixed Effects Inference

Developing a principled statistical inferential framework for fixed effects in longitudinal functional models is difficult. Several approaches that account for the complex within- and between-study participant correlations exist and include different Bayesian approaches (Morris and Carroll, 2006; Morris et al., 2006; Zhu, Brown, and Morris, 2011; Goldsmith, Zipunnikov, and Schrack, 2015; Zhang et al., 2016). Unfortunately, many of these methods require specialized software, are slow, and do not scale up with the number of study participants and dimension of the functional domain. Our proposed approach is philosophically closer to the methods proposed in Crainiceanu et al. (2012) and Park et al. (2018), which use bootstrap of study participants. For Gaussian functional data we provide an analytic solution, while for all types of data we propose a bootstrap approach.

4.3.1 Analytic Inference for Gaussian Functional Data

For Gaussian functional data the pointwise linear mixed model has the form

$$Y_{ij}(s_l) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}(s_l) + \boldsymbol{Z}_{ij}^T \boldsymbol{u}_i(s_l) + \epsilon_{ij}(s_l) \ . \tag{4.2}$$

For the ith study participant at $s_l \in \mathcal{S}$ data are $\{\boldsymbol{Y}_i(s_l), \boldsymbol{X}_i, \boldsymbol{Z}_i\}$, where $\boldsymbol{Y}_i(s_l) = [Y_{i1}(s_l), ..., Y_{iJ_i}(s_l)]$, $\boldsymbol{X}_i = [\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{iJ_i}]$, and $\boldsymbol{Z}_i = [\boldsymbol{Z}_{i1}, ..., \boldsymbol{Z}_{iJ_i}]$. The observations for the entire study population are then denoted as $\{\boldsymbol{Y}(s_l), \boldsymbol{X}, \boldsymbol{Z}\}$, where $\boldsymbol{Y}(s_l) = [\boldsymbol{Y}_1(s_l), ..., \boldsymbol{Y}_I(s_l)]^T$, $\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_I]^T$, and $\boldsymbol{Z} = \text{diag}(\boldsymbol{Z}_1^T, ..., \boldsymbol{Z}_I^T)$. The matrix notation of equation (4.2) is $\boldsymbol{Y}(s_l) = \boldsymbol{X}\boldsymbol{\beta}(s_l) + \boldsymbol{Z}\boldsymbol{u}(s_l) + \boldsymbol{\epsilon}(s_l)$, where $\boldsymbol{u}(s_l) = [\boldsymbol{u}_1(s_l)^T, ..., \boldsymbol{u}_I(s_l)^T]^T$ and $\boldsymbol{\epsilon}(s) = [\epsilon_{11}(s), \epsilon_{12}(s), ..., \epsilon_{IJ_I}(s)]^T$ are mutually independent random coefficients and errors with a joint multivariate Gaussian distribution. The pointwise estimator of the fixed effects is $\tilde{\boldsymbol{\beta}}(s_l) = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(s_l)\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(s_l)\boldsymbol{Y}(s_l)$, where $\boldsymbol{V}(s_l) = \boldsymbol{Z}\boldsymbol{H}(s_l)\boldsymbol{Z}^T + \boldsymbol{R}(s_l)$ and $\boldsymbol{H}(s_l)$ and $\boldsymbol{R}(s_l)$ are covariance matrices of $\boldsymbol{u}(s_l)$ and $\boldsymbol{\epsilon}(s_l)$, respectively. These fixed effects estimators are correlated across s_l, which needs to be taken into account when conducting joint inference, including when building joint confidence bands or conducting multiple testing. This correlation is modeled intrinsically by assuming $\text{Cov}\{\boldsymbol{u}(s_{l_1}), \boldsymbol{u}(s_{l_2})\} = \boldsymbol{G}(s_{l_1}, s_{l_2})$ and $\text{Cov}\{\boldsymbol{\epsilon}(s_{l_1}), \boldsymbol{\epsilon}(s_{l_2})\} = \boldsymbol{0}$ for all $s_{l_1} \neq s_{l_2}$. Thus, the covariance of the raw estimates at $s_l \in \mathcal{S}$ is $\text{Var}\{\tilde{\boldsymbol{\beta}}(s_l)\} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(s_l)\boldsymbol{X}\}^{-1}$, and between s_{l_1} and $s_{l_2} \in \mathcal{S}$ is

$$\text{Cov}\{\tilde{\boldsymbol{\beta}}(s_{l_1}), \tilde{\boldsymbol{\beta}}(s_{l_2})\} = \{\boldsymbol{X}^T\boldsymbol{V}^{-1}(s_{l_1})\boldsymbol{X}\}^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}(s_{l_1})\boldsymbol{W}(s_{l_1}, s_{l_2})\boldsymbol{V}^{-1}(s_{l_2})\boldsymbol{X}\{\boldsymbol{X}^T\boldsymbol{V}^{-1}(s_{l_2})\boldsymbol{X}\}^{-1}. \tag{4.3}$$

Here $\boldsymbol{W}(s_{l_1}, s_{l_2}) = \boldsymbol{Z}\boldsymbol{G}(s_{l_1}, s_{l_2})\boldsymbol{Z}^T$.

Estimates of $\boldsymbol{H}(s_l)$ and $\boldsymbol{R}(s_l)$ can be obtained directly from the mixed effects model software. An additional smoothing approach, for example using penalized splines (Ruppert, Wand, and Carroll, 2003), can be applied to each entry of these matrices along the functional domain. Estimations of

the $\boldsymbol{G}(s_{l_1}, s_{l_2})$ is not as intuitive. However, the method of moments (MoM) procedure introduced in Greven et al. (2010) provides the blueprint for our procedure. More precisely, for any $s_{l_1} \neq s_{l_2}$

$$E[\{Y_{ik}(s_{l_1}) - \boldsymbol{X}_{ik}^T\boldsymbol{\beta}(s_{l_1})\}\{Y_{ij}(s_{l_2}) - \boldsymbol{X}_{ij}^T\boldsymbol{\beta}(s_{l_2})\}] = \sum_{v=1}^{q}\sum_{t=1}^{q} Z_{ijv}Z_{ikt}\text{Cov}\{u_{it}(s_{l_1}), u_{iv}(s_{l_2})\}\ , \tag{4.4}$$

for any $j, k = 1, ..., J_i$. This suggests a simple approach: regressing linearly the residual products $\{Y_{ik}(s_{l_1}) - \boldsymbol{X}_{ik}^T\boldsymbol{\beta}(s_{l_1})\}\{Y_{ij}(s_{l_2}) - \boldsymbol{X}_{ij}^T\boldsymbol{\beta}(s_{l_2})\}$ onto the covariates $\{Z_{ijv}Z_{ikt} : j, k = 1, ..., J_i\}$. A bivariate smoother, for example, the fast bivariate P-splines (Xiao, Li, and Ruppert, 2013), could be used to further reduce the variability of the covariance estimators. This approach does not guarantee that the resulting $\hat{\boldsymbol{H}}(s_l)$, $\hat{\boldsymbol{R}}(s_l)$, and $\hat{\boldsymbol{G}}(s_{l_1}, s_{l_2})$ are positive definite. This is handled by trimming the negative eigenvalues of these estimators at 0, as suggested in the literature (Yao et al., 2003; Hall, Müller, and Yao, 2008; Greven et al., 2010).

While the model allows different smoothers in the second step, we provide the closed form solution for penalized splines. For simplicity, we focus on the rth fixed effect $\beta_r(s)$. Denoted by $\tilde{\boldsymbol{\beta}}_r = [\tilde{\beta}_r(s_1), ..., \tilde{\beta}_r(s_L)]^T$ the raw estimates of $\beta_r(s)$ obtained from the first step. Let $\boldsymbol{B}_r = [\boldsymbol{b}_{r1}, ..., \boldsymbol{b}_{rK}]$ be the K-dimensional spline basis matrix, where $\boldsymbol{b}_{rk} = [b_{rk}(s_1), ..., b_{rk}(s_L)]^T$, $k = 1, ..., K$. K is usually chosen to be much smaller than L. Given a smoothing parameter λ_r, penalty matrix $\boldsymbol{P}_r$, define $\boldsymbol{S}_r = \boldsymbol{B}_r(\boldsymbol{B}_r^T\boldsymbol{B}_r + \lambda_r\boldsymbol{P}_r)^{-1}\boldsymbol{B}_r^T$. The smoothed estimator of the rth fixed effect is $\hat{\boldsymbol{\beta}}_r = \boldsymbol{S}_r\tilde{\boldsymbol{\beta}}_r$. The covariance matrix of $\hat{\boldsymbol{\beta}}_r$ is $\text{Cov}(\hat{\boldsymbol{\beta}}_r) = \boldsymbol{S}_r\text{Cov}(\tilde{\boldsymbol{\beta}}_r)\boldsymbol{S}_r$, a sandwich smoother of $\text{Cov}(\tilde{\boldsymbol{\beta}}_r)$ obtained from the first step.

This inferential approach is explicit when working with linear mixed effects models with Gaussian random effects and errors. While the inferential approach described in this section involves additional notation, the computations are straightforward. This analytic inferential solution offers the potential for substantially reducing the computational burden of performing bootstrap inference for large-scale Gaussian functional data and performs well in our simulation study (close to nominal coverage for 95% confidence bands). For outcomes with general (Gaussian and non-Gaussian) distributional assumptions we now discuss the bootstrap of study participants as a general solution.

4.3.2 Nonparametric Bootstrap Approach

Bootstrapping functional data is a practical approach for fixed effects inference (Cuevas, Febrero, and Fraiman, 2006; Crainiceanu et al., 2012). For complex correlated functional data, Park et al. (2018) proposed both study participant and residual bootstrap. Here we focus only on the study participant bootstrap approach because the residual bootstrap is not defined for generalized outcomes. The approach is described in Algorithm 3.

4.3.3 Extension to Random Effects Inference

The proposed inference procedure has a natural extension to random effects and visit-specific predictions. We provide a brief introduction for this framework. Given equation (4.2) and notations introduced in Section 4.3.1, the pointwise BLUP of the random effects is $\tilde{\boldsymbol{u}}(s_l) = \boldsymbol{H}(s_l)\boldsymbol{Z}^T\boldsymbol{V}^{-1}(s_l)\{\boldsymbol{Y}(s_l) - \boldsymbol{X}\tilde{\boldsymbol{\beta}}(s_l)\}$. Without loss of generality assume $\boldsymbol{R}(s_l) = \sigma^2(s_l)\boldsymbol{I}$. The uncertainty

Algorithm 3 Nonparametric Bootstrap for fixed effects inference

Data: $\{\boldsymbol{Y}(s_l), l = 1, ..., L\}$, $\boldsymbol{X}$, $\boldsymbol{Z}$.
Result: $\text{Var}(\hat{\boldsymbol{\beta}}(s_l)), l = 1, ..., L$.
For $b = 1, ..., B$ **do**

1. Re-sample I subject indices from $\{1, ..., I\}$ with replacement. Denote the vector of re-sampled indices as $M_{(b)}$
2. For the i'th element of $M_{(b)}$, $i' = 1, ..., I$, include all observations of the corresponding subject in the bootstrap sample. Denote the bth bootstrap sample as $\{\{\boldsymbol{Y}_{M_{(b)}}(s_l), l = 1, ..., L\}, \boldsymbol{X}_{M_{(b)}}, \boldsymbol{Z}_{M_{(b)}}\}$
3. Fit the model in Section 4.2 using the bth bootstrap sample. Derive the fixed effects estimates $\{\hat{\boldsymbol{\beta}}(s_l)_{(b)}, l = 1, ..., L\}$

4. For $l = 1, ..., L$, derive $\text{Var}(\hat{\boldsymbol{\beta}}(s_l))$ from B bootstrap estimates $\{\hat{\boldsymbol{\beta}}(s_l)_{(1)}, ..., \hat{\boldsymbol{\beta}}(s_l)_{(B)}\}$. In practice we calculate the sample variance and use it as the estimator.

of $\boldsymbol{u}(s_l)$ can be measured through the conditional variance (Morris, 1983) as $\text{Var}\{\boldsymbol{u}(s_l)|\boldsymbol{Y}(s_l)\} = \sigma^2(s_l)\{\boldsymbol{H}(s_l) - \boldsymbol{H}(s_l)\boldsymbol{Z}^T\boldsymbol{V}^{-1}(s_l)\boldsymbol{Z}\boldsymbol{H}(s_l)\}$. The uncertainty of random effects between locations can be measured similarly using the empirical Bayes estimator.

4.4 Joint Confidence Bands

Inference for functional data has a natural connection with the problem of multiple testing due to the inherent correlations in the observed data along the functional domain. The pointwise confidence bands described in Sections 4.3.1 and 4.3.2 do not provide any information about joint coverage probabilities over the entire domain and are, in fact, valid when averaged across the functional domain. The Bonferonni correction (Bonferroni, 1936) is exact for independent data, but is inappropriate for functional data, which is correlated and can have arbitrary sampling density. For example, a functional

domain could be sampled at one hundred or one million equally spaced points. In both situations the point estimators would barely change, whereas the joint Bonferonni corrected confidence intervals would depend on the choice of number of samples within the functional domain. At one extreme, if the number of points is allowed to go to infinity, the length of the joint confidence intervals diverges to infinity. This is unacceptable and cannot be addressed by changing the testing criterion to, say, the Benjamini-Hochberg false discovery rate (Benjamini and Hochberg, 1995). The reason is that both methods are overly conservative when the test statistics are highly correlated. As functional data can be sampled densely and correlations between observations increase with sampling density, the probability of failing to reject the null hypothesis using these corrections rapidly approaches one as the data are more densely sampled. An alternative was proposed by Cox and Lee (2008) for functional data based on the Westfall-Young randomization method (Westfall and Young, 1993).

The construction of joint confidence bands in the context of functional data analysis has been studied using various approaches including local linear estimators (Degras, 2011), piecewise constant splines (Ma, Yang, and Carroll, 2012), and polynomial splines (Cao, Yang, and Todem, 2012). However, most of these methods assume independence between curves, which is not the setting considered here. Crainiceanu et al. (2012) and Park et al. (2018) proposed the bootstrap of study participants as a general inferential procedure for functional data with arbitrarily complex functional correlation structures. Here we follow a similar approach.

Based on a bootstrap of study participants we obtain the following estimators $\hat{\boldsymbol{\beta}}_r = [\hat{\beta}_r(s_1), ..., \hat{\beta}_r(s_L)]^T$ and $\text{Var}(\hat{\boldsymbol{\beta}}_r)$, where $\{\beta_r(s), s \in \mathcal{S}\}$ is the rth functional fixed effect. Let N_s be the sample size of simulated data. Our approach requires simulations from the multivariate normal distribution $\mathcal{N}\{\hat{\boldsymbol{\beta}}_r, \text{Var}(\hat{\boldsymbol{\beta}}_r)\}$. When the dimension L of the functional domain is large, this can be quite slow, but the problem can be addressed using a PCA decomposition of the bootstrap samples $\hat{\boldsymbol{\beta}}_{r(1)}, \ldots, \hat{\boldsymbol{\beta}}_{r(B)}$; for details see Algorithm 4.

Algorithm 4 Level α joint confidence bands of $\hat{\beta}_r(s)$

Data: $\hat{\boldsymbol{\beta}}_r$, $\text{Var}(\hat{\boldsymbol{\beta}}_r)$, $\hat{\boldsymbol{\beta}}_{r(1)}, ..., \hat{\boldsymbol{\beta}}_{r(B)}$, N_s.
Result: Joint confidence bands of $\{\hat{\beta}_r(s), s \in \mathcal{S}\}$.
1. Perform Functional Principal Component Analysis (FPCA) on $[\hat{\boldsymbol{\beta}}_{r(1)}, ..., \hat{\boldsymbol{\beta}}_{r(B)}]^T$. Derive the mean function $\boldsymbol{\mu} = [\mu(s_1), ..., \mu(s_L)]^T$, eigenvalues $\lambda_1, ..., \lambda_L$ and eigenfunctions $\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_L$, where $\boldsymbol{\psi}_k = [\psi_k(s_1), ..., \psi_k(s_L)]^T, k = 1, ..., L$

For $n = 1, ..., N_s$ **do**

2. Simulate $\xi_{nk} \sim \mathcal{N}(0, \lambda_k)$ for $k = 1, \ldots, K_T$. Calculate $\hat{\boldsymbol{\beta}}_{r,n} = \boldsymbol{\mu} + \sum_{k=1}^{K_T} \xi_{nk} \boldsymbol{\psi}_k$

3. Calculate $u_n = \max_{s_l \in \mathcal{S}} \{|\hat{\boldsymbol{\beta}}_{r,n} - \hat{\boldsymbol{\beta}}_r| / \sqrt{\text{diag}(\text{Var}(\hat{\boldsymbol{\beta}}_r))}\}$

4. Obtain $q_{1-\alpha}$, the $(1-\alpha)$ empirical quantile of $\{u_1, ..., u_{N_s}\}$
5. The joint confidence interval at $s_l \in \mathcal{S}$ is calculated as $\hat{\beta}_r(s_l) \pm q_{1-\alpha} \sqrt{\text{Var}(\hat{\boldsymbol{\beta}}_r)_{(l,l)}}$. The upper and lower bounds of the joint confidence bands can be smoothed.

The number of functional principal component basis K_T in Step 2 is selected based on proportion of variance explained, as suggested in the FPCA literature, and usually does not exceed 100 when using a 95% variance explained threshold. As a result, this modified algorithm reduces the dimension of simulation from potentially large L to a more acceptable K_T; see supplementary material for a comparison of computing time between this method and direct simulations from multivariate normal distribution.

4.5 Simulations

An extensive simulation study is used to assess: (1) the performance of the estimators and the pointwise/joint confidence bands; (2) how methods compare to existing approaches. The `R` code for the simulation study is provided in the supplementary material.

4.5.1 Simulation Setup

We simulate functional responses on an equally-spaced grid of $\mathcal{S} = [0, 1]$ with length L. For simplicity, we fix $p = 2$, the number of fixed effects for each point s on the functional domain and $q = 1$, the number of random effects for each point on the functional domain. Therefore, $\boldsymbol{X}_{ij} = [1, X_{ij1}]^T$ and $\boldsymbol{u}_i(s) = u_i(s)$, though the approach is designed for much larger p and q. For subject i at visit j the data generating model is

$$\eta_{ij}(s) = g\{\mu_{ij}(s)\} = \beta_0(s) + X_{ij1}\beta_1(s) + u_i(s), s \in \mathcal{S} \ .$$

The fixed effects covariates are simulated as $X_{ij1} \sim \mathcal{N}(0, 4)$, while the random effects are simulated as $u_i(s) = c_{i1}\psi_1(s) + c_{i2}\psi_2(s)$. We use the scaled orthonormal functions $\psi_1(s) \propto 1.5 - \sin(2\pi s) - \cos(2\pi s)$ and $\psi_2(s) \propto \sin(4\pi s)$ to capture the individual-level fluctuations. The random coefficients are generated from $c_{i1} \sim \mathcal{N}(0, 2\sigma_B^2)$ and $c_{i2} \sim \mathcal{N}(0, \sigma_B^2)$, respectively. Here σ_B^2 is determined by the relative importance of random effects SNR_B, as described below. We consider the following simulation scenarios:

1. Simulation parameters

- Distribution of the functional responses: (a) Gaussian, (b) binary.
- Functional fixed effects $\boldsymbol{\beta}(s)$:

  S1: $\beta_0(s) = -0.15 - 0.1 * \sin(2\pi s) - 0.1 * \cos(2\pi s)$, $\beta_1(s) = \frac{1}{20}\phi(\frac{s-0.6}{0.0225})$;

  S2: $\beta_0(s) = 0.53 + 0.06\sin(3\pi s) - 0.03\sin(6.5\pi s)$, $\beta_1(s) = \frac{1}{60}\phi(\frac{s-0.2}{0.1^2}) + \frac{1}{200}\phi(\frac{s-0.35}{0.1^2}) - \frac{1}{250}\phi(\frac{s-0.65}{0.06^2}) + \frac{1}{60}\phi(\frac{s-1}{0.07^2})$.

2. Sample size parameters
   - Number of subjects: $I \in \{50, 100, 200, 400\}$.
   - Mean number of visits per subject: $J \in \{5, 10, 20, 40\}$. For subject i the number of visits J_i is drawn from Poisson(J) with a minimum of 1 visit.
   - Dimension of the functional domain: $L \in \{50, 100, 200, 400\}$.
3. Signal-noise parameters
   - Relative importance of random effects: $\text{SNR}_B \in \{0.5, 1\}$. Here SNR_B is the standard deviation of the fixed effects functions divided by the standard deviation of the random effects functions; see Scheipl, Staicu, and Greven (2015) for detailed descriptions of this parameter.
   - Signal-to-noise ratio: $\text{SNR}_\epsilon \in \{0.5, 1\}$ (Gaussian response only). Here SNR_ϵ is the standard deviation of the linear predictors divided by the standard deviation of the noise σ_ϵ.

In terms of fixed effects, the S1 functions are similar to those used by Goldsmith, Zipunnikov, and Schrack (2015), are smooth and easy-to-estimate.

The S2 functions have more complex shapes and are designed to approximate the estimated effects in our DTI application. To be specific, $\beta_0(s)$ mimics a typical fractional anisotropy (FA) trajectory on the corpus callosum in DTI while $\beta_1(s)$ represents a non-monotonic effect of scan date on the FA.

For each scenario we conducted 200 simulations. Considering all combinations of parameters would have been impossible even for our extensive computing resources. Instead, for each combination of simulation parameters (denoted by "Gaussian S1", "Gaussian S2", "Binary S1", "Binary S2" in the simulation results), we: (1) fix the sample size parameters at their baseline ($I = 50, J = 5, L = 50$), then change the signal-noise parameters; and (2) fix the signal-noise parameters at their baseline ($\text{SNR}_B = 0.5, \text{SNR}_\epsilon = 1$), then change the sample size parameter one at a time while fixing the other two sample size parameters at baseline. For example, we fix $J = 5, L = 50$ to evaluate model performance for a different number of study participants, I.

4.5.2 Comparisons to Existing Methods

Many methods have been developed to model correlated functional responses, including Functional Additive Mixed Models (FAMM) (Scheipl, Staicu, and Greven, 2015; Scheipl, Gertheiss, Greven, et al., 2016), Generalized Multilevel Function-on-Scalar Regression and Principal Component Analysis (GenMF-PCA) (Goldsmith, Zipunnikov, and Schrack, 2015), Functional Linear Array Model (FLAM) (Brockhaus et al., 2015), Wavelet-based Functional Mixed Models (WFMM) (Morris and Carroll, 2006), FMEM (Yuan et al., 2014; Zhu

et al., 2019). As described in Scheipl, Staicu, and Greven (2015), FAMM outperformed Bayesian WFMM for smooth curves. For FLAM, Brockhaus et al. (2015) reported a similar estimation accuracy with FAMM in their simulations. Although inference for FLAM could potentially be conducted via subject-level bootstrap, this implementation was not available in the `FDboost` package (Brockhaus, Rügamer, and Greven, 2017). For FMEM we could not identify general purpose software. Therefore, we compare FAMM and GenMFPCA, which have well-documented and easy-to-use implementations for inference.

Our method (FUI) is implemented in the `lfosr3s()` function in the supplementary material, which implements univariate GLMMs and then applies a penalized cubic spline smoother. Results are highly robust to the choice of smoother. For FAMM we use the `pffr()` function from `refund` package (Goldsmith et al., 2020) in `R`. We used 15 and 20 cubic B-splines bases with first order difference penalty for the population average and global functional intercept respectively; see `bs.yindex` and `bs.int` arguments in the `pffr()` function. We have increased the number of bases from the default to increase the performance of FAMM.

Because Goldsmith, Zipunnikov, and Schrack (2015) reported considerably larger computing time than FAMM, comparisons with GenMFPCA are restricted to smaller sample sizes. In addition, GenMFPCA software is only applicable to binary response. We have attempted to manually implement GenMFPCA for Gaussian responses, but, probably due to our sub-optimal implementation, our implementation was quite slow. To ensure a fair comparison of computing time, we focus on comparing with GenMFPCA only for

binary response. The results are shown in the supplementary material.

4.5.3 Model Evaluation Criteria

We compare the performance of each method (FUI, FAMM, GenMFPCA) with respect to: (1) accuracy in estimating fixed effects; (2) inference on fixed effects; and (3) computational efficiency.

Accuracy of fixed effects estimation was assessed using the integrated squared error (ISE) of fixed effects, defined as $\text{ISE}_k = \int_0^1 (\hat{\beta}_k(s) - \beta_k(s))^2 ds, k = 0, 1$. The mean integrated squared error (MISE) is calculated by averaging ISE across simulated datasets. We show ISE of $\beta_1(s)$, as similar results were obtained for $\beta_0(s)$.

Inferential performance was assessed by calculating the empirical coverage probability of 95% pointwise confidence bands at each location, then taking the average along the functional domain. For FUI, we also report the empirical coverage probability of 95% joint confidence bands proposed in Section 4.4. For FUI we use analytic inference (mean $\pm$ 2sd) for Gaussian responses and bootstrap inference for other families of distributions. In the nonparametric bootstrap, we have used the formula mean $\pm$ 2.2sd instead of mean $\pm$ 2sd. This is a simple solution that provides remarkably good results and seems to account for the extra within-subject variability that may be missed by conducting a bootstrap of study participants.

For each scenario, we do 200 simulations on the Joint High Performance Computing Exchange (JHPCE) Cluster with 1 core per simulation. The computing time of each method is obtained under different scenarios.

4.5.4 Simulation Results: Signal-to-noise parameters

The simulation results for different signal-noise parameters are shown in Figure 4.2 under the scenario with the smallest number of subjects, mean number of visits per subject, and dimension of the observed functional responses ($I = 50, J = 5, L = 50$). We only display results for the Gaussian response, as similar results were obtained for binary responses. Left two panels: fixed effect is S1. Right two panels: fixed effect is S2. The MISE decreases as signal increases, either by increasing SNR_B or SNR_ϵ. In addition, these two parameters exhibit similar scaling behavior using both estimation methods. Specifically, increasing SNR_B (or SNR_ϵ) from 0.5 to 1 decreases MISE by about 60% in S1 and about 50% in S2 using either FUI or FAMM. For this small sample size, the coverage is close to the nominal level and the computing time for both methods is almost identical (not displayed).

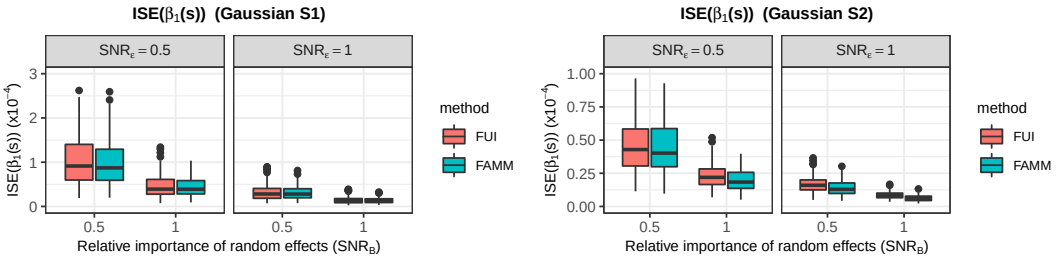


Figure 4.2: Estimation accuracy for FUI (red) and FAMM (blue) under different relative importance of random effects (SNR_B, x axis) and signal-to-noise ratios (SNR_ϵ, labels in the gray-shaded area of each panel). Functional response is Gaussian; parameters: $I = 50, J = 5, L = 50$. Left two panels: S1. Right two panels: S2.

4.5.5 Simulation Results: Sample Size Parameters

The simulation results for different sample size parameters are shown in Figure 4.3. As results tend to be quite consistent, we display the results for Gaussian outcomes and fixed effects S1 (denoted by "Gaussian S1" in the title of each panel). Results for other combinations are in the supplementary material. The baseline setting is $I = 50, J = 5, L = 50, \text{SNR}_B = 0.5, \text{SNR}_\epsilon = 1$. All other parameters are fixed at their baseline values when one sample size parameter is changed. Left column: number of subjects (I). Middle column: mean number of visits per subject (J). Right column: dimension of the functional domain (L). The ISE and computing time for 200 simulations are displayed in the top and bottom row, respectively. The inference results are shown in Table 4.1. For FUI, we report the empirical coverage probability of both joint (denoted as "Coverage (Joint)") and pointwise (denoted as "Coverage (Pointwise)") 95% confidence bands. For FAMM, we report the coverage of the pointwise confidence bands (denoted as "Coverage").

As the number of subjects increases, the MISE for both FUI and FAMM decreases. The estimation accuracy of the two methods is similar, and the coverage of the confidence bands, both joint and pointwise for FUI and pointwise for FAMM, reach their nominal level. However, FAMM computing time increases substantially when the number of subjects increases (see bottom left panel). Indeed, the median computing time exhibits a second-order polynomial shape in number of subjects, with median computing time exceeding 10 hours on the cluster when $I = 400$. In addition, the memory usage for

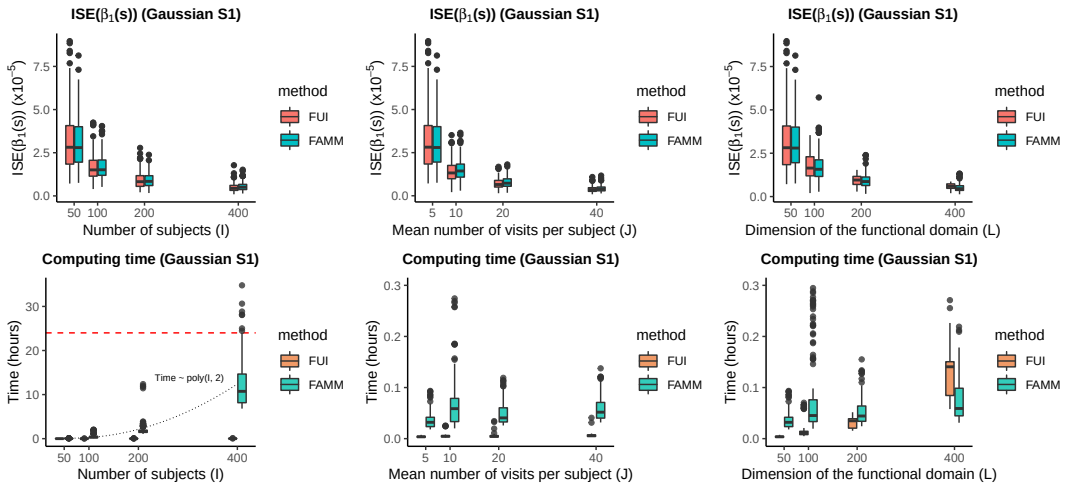


Figure 4.3: Estimation accuracy (top row) and computing time (bottom row) for FUI (red) and FAMM (blue) from 200 simulations. Response is Gaussian and the true fixed effects functions are S1. The baseline setting is $I = 50, J = 5, L = 50, \text{SNR}_B = 0.5, \text{SNR}_\epsilon = 1$. All other parameters are fixed at their baseline values when one sample size parameter is changed. Left column: number of subjects (I). Middle column: mean number of visits per subject (J). Right column: dimension of the functional domain (L).

FAMM increases substantially despite the use of the efficient `mgcv::bam` implementation. For example, we were not able to perform simulations for FAMM when $I = 1000$, as both memory (more than 40 GB RAM) and computing time (unknown) exceeded our extensive resources. In contrast, FUI required only 40 seconds for $I = 500$ study participants and there are no problems with fitting even for $I = 10000$.

As the mean number of visits per subject (second column of panels) and dimension of the functional domain (third column of panels) increase, both methods display similar estimation accuracy. The confidence bands of both methods, including joint and pointwise confidence bands for FUI and pointwise confidence bands for FAMM, also have good coverage close to the nominal level (middle and bottom blocks of Table 4.1). For $I = 50, L = 50$ FAMM

Table 4.1: Empirical coverage probability of 95% joint and pointwise confidence bands using FUI and 95% pointwise confidence bands using FAMM from 200 simulations. Response is Gaussian and the true fixed effects functions are S1. The pointwise confidence band is constructed as mean $\pm$ 2sd and the joint is mean $\pm\, q_{0.975} \times$ sd. The baseline setting is $I = 50, J = 5, L = 50, \text{SNR}_B = 0.5, \text{SNR}_\epsilon = 1$. All other parameters are fixed at their baseline values when one sample size parameter is changed.

Method	**Type**	**Number of subjects (I)**			
		50	**100**	**200**	**400**
FUI	Coverage (Joint)	0.93	0.96	0.94	0.95
	Coverage (Pointwise)	0.94	0.95	0.94	0.95
FAMM	Coverage	0.96	0.96	0.96	0.94
Method	**Type**	**Mean number of visits per subject (J)**			
		5	**10**	**20**	**40**
FUI	Coverage (Joint)	0.93	0.95	0.97	0.96
	Coverage (Pointwise)	0.94	0.95	0.95	0.95
FAMM	Coverage	0.96	0.96	0.96	0.96
Method	**Type**	**Dimension of the functional domain (L)**			
		50	**100**	**200**	**400**
FUI	Coverage (Joint)	0.93	0.94	0.94	0.94
	Coverage (Pointwise)	0.94	0.94	0.95	0.95
FAMM	Coverage	0.96	0.96	0.96	0.96

requires similar computation time (3 to 7 minutes) for 5 to 40 visits per study participant, while FUI takes on average less than 1 minute in all scenarios. For $I = 50, J = 5$ computing time of FUI increases with the dimension of the functional domain (bottom right panel) while FAMM remains unaffected. This increase is expected as we run a GLMM at every location and the time for these GLMMs simply add up. However, our method is easy to parallelize, which would reduce the fitting time to the the time it would take to fit a single GLMM. To the best of our knowledge, FAMM does not currently have a parallel implementation.

The computation time advantages of FUI should not be surprising given the way `mgcv` is used to estimate functional models. Specifically, random effects are incorporated by fully constructing the random effects design matrix

and applying ridge penalties. Therefore, for a model with a subject-specific functional random intercept, $u_{0i}(s)$, the design matrix adds k_b columns, where k_b is the number of spline bases used to represent the functional random intercept. This is not a problem when I is in the range of 50 to 100, but it becomes problematic when $I > 200$. Take our physical activity data application for example where $L = 1440$, $J = 7$ and $I = 1680$ and consider a simple functional model $\beta_0(s) + u_{i0}(s)$. Assume that the population mean function $\beta_0(s)$ uses 20 B-spline basis functions (the default for FAMM). The design matrix without random effects is $1440 \times 7 \times 1680 = 16,934,400$ rows and 20 columns. Assume that the model $u_{i0}(s)$ uses 15 B-spline basis functions ($k_b = 15$). Then the full design matrix is as large as $16,934,400 \times 25,220$, since $20 + (15 \times 1680) = 25,220$, amounting to over 400 billion elements. This explains why FAMM runs into substantial computational challenges when the number of study participants increases. The `mgcv` package does have method for handling large datasets through the `mgcv::bam` function which avoids constructing and performing computations on the full design matrix. However, even with these added efficiencies FAMM runs into substantial computational challenges. This could be addressed in the future, but our current solution provides a practical, "read-and-use", stable alternative for a moderate to large number of study participants.

4.5.6 Simulation Summary

Our method achieves similar accuracy with the state-of-the-art FAMM method for fixed effects under different simulation settings, including different signal-to-noise parameters and sample size parameters. FUI is much faster than FAMM when the number of subjects is large. To the best of our knowledge, FUI is the first inferential method that is demonstrated to work with over 1500 study participants. The reason for implementing such approaches is practical, as many datasets, including our NHANES application, contain such sample sizes. Both joint and pointwise confidence bands of FUI exhibit good coverage to the nominal level. For both FUI and FAMM, estimation accuracy is affected by the change of signal-noise parameters and sample size parameters. These results for FAMM are consistent with the simulation results reported in Scheipl, Staicu, and Greven (2015).

4.6 Applications

In this section, we apply our method to the motivating examples introduced in Section 4.1.

4.6.1 DTI Study

Multiple sclerosis (MS) is an autoimmune mediated disease that affects the central nervous system (CNS) and can lead to substantial motor and cognitive disability. While the exact cause of MS remains unknown, modern neuroimaging has played a crucial role in the diagnosis and management of

MS. A promising imaging technique is Diffusion Tensor Imaging based on Magnetic Resonance Imaging (DTI-MRI or, shorter, DTI). DTI provides measures of water diffusion in the brain, which are thought to be associated with white matter integrity. Fractional anisotropy (FA) is a measure of diffusion anisotropy derived from DTI. A zero value of FA corresponds to perfectly isotropic diffusion (water diffuses unrestricted in all directions), while a value of one of FA corresponds to perfectly anisotropic diffusion (water diffuses only in one direction). Values of FA fall somewhere within the $(0, 1)$ range with higher values corresponding to more anisotropic (more organized) water diffusion. FA can be calculated at every location in the brain.

Here we focus on the FA calculated along the corpus callosum, a nerve tract connecting the left and right cerebral hemispheres; see Goldsmith et al. (2011) and Greven et al. (2010) for in-depth descriptions of the data. For our purposes, the data set consists of 142 study participants (42 healthy individuals and 100 MS patients). For healthy individuals there is only one visit, whereas for MS patients there are multiple visits with an average of 3.4 and a maximum of 8 visits per MS patient. There were a total of 382 visits across MS patients and healthy individuals. Corpus callosum is a three-dimensional C-shaped nerve fiber bundle that connects the left and right brain hemispheres. For the purpose of this application, several landmarks were manually identified on the brain image and FA was calculated as an average FA at 93 locations along the corpus callosum. Thus, the data consist of a 382×93 dimensional matrix, where each row corresponds to a brain image visit and each column corresponds to a particular location in the corpus callosum, resulting in a total

of 35526 observations. The study participant ID, age, sex and date of scan information are also available for each study participant at each visit.

For illustration purposes, we are interested in quantifying the association between age, sex and date of scan with FA measurements along the corpus callosum. Using the notation introduced in Section 4.2, the longitudinal functional responses are denoted by $Y_{ij}(s)$, and are the observed FA values along the equally-spaced grid of $s \in \mathcal{S} = \{1, \dots 93\}$. For the ith individual at the jth visit, the fixed effects $\boldsymbol{X}_{ij} = [1, X_{ij1}, X_{ij2}, X_{ij3}, X_{ij4}]^T$ where X_{ij1} is a binary indicator of case (1 for MS patients and 0 for healthy individuals), X_{ij2} is the date of scan (converted into year unit and treated as numeric), X_{ij3} is a binary indicator of sex (there were no self-identified non-binary participants in this study) and X_{ij4} is the age at baseline scan (in years). For each location we fit a random intercept and slope model

$$Y_{ij}(s) = \beta_0(s) + X_{ij1}\beta_1(s) + X_{ij2}\beta_2(s) + X_{ij3}\beta_3(s) + X_{ij4}\beta_4(s) + u_{i0}(s) + u_{i1}(s)X_{ij2} + \epsilon_{ij}(s) \ ,$$

where $(u_{i0}(s), u_{i1}(s))$ have a joint zero-mean bivariate Normal distribution independent of $\epsilon_{ij}(s)$, which are iid $N\{0, \sigma^2(s)\}$. This standard linear mixed effects model is fit 93 times, once for each location s. After pointwise fitting, a penalized spline smoother is used to smooth each $\beta_l(\cdot)$, $l = 0, 1, 2, 3, 4$ coefficient separately. All five smoothers use a cubic basis with 15 equally spaced knots and REML estimation of the smoothing parameter. The pointwise and joint confidence intervals are obtained as described in Section 4.3 and Section 4.4.

Figure 4.4 displays the point estimators of the fixed effects parameters

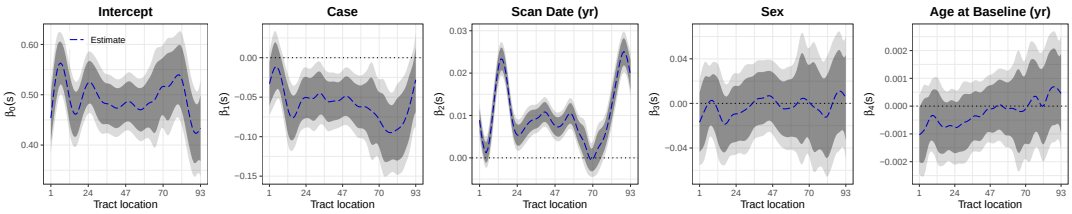


Figure 4.4: Fixed effects estimates (dashed blue line), 95% pointwise confidence intervals (dark gray shaded area), and 95% joint confidence intervals (light gray shaded area) in the DTI study. Panels from left to right: intercept, case, date of scan, sex, age at baseline.

(dashed blue lines) for the intercept, case, scan date, sex, and age (five panels from left to right), respectively. The dark gray regions correspond to the 95% pointwise confidence bands, while the light gray regions correspond to the 95% joint confidence bands. The intercept estimator is consistent with the geometry of the corpus callosum and previously published literature. Compared with healthy individuals, MS patients of the same sex and age at the same date of scan have significantly lower FA at most locations along the corpus callosum. This result may indicate lower anisotropy corresponds among MS patients, which may be consistent with brain micro-structure damage. The middle panel shows a highly significant *increase* in the FA as a function of scan date at most locations of the corpus callosum. To the best of our knowledge, there is no biological plausible reason for such an increase in anisotropy. Therefore, the result may correspond to the change in technology and software, which led to a large, deterministic, increase in measured FA. The effects of sex and age are not statistically significant at any point along the corpus callosum.

Figure 4.5 displays the data (left panels), together with the pointwise estimators (middle panels) and smoothed estimators along the functional domain (right panels). The first and second rows correspond to study participants ID

2017 and 2085, respectively. For study participant ID 2017, the point estimators are consistent with substantial reduction in the visit-to-visit variability. Unsurprisingly, after smoothing (right-top panel) visit-specific profiles are slightly smoother along the functional domain, but with a similar reduced visit-to-visit variability. Comparing the middle and right top panels indicates that the pointwise linear mixed effects models did the "heavy lifting", while the functional smoothing led to mostly cosmetic changes. This need not be the case in general when the noise and correlation structures could be quite different. Results are similar for study participant ID 2085. These results suggest that: (1) there is a statistically significant, but small fixed effect for the date of the visit; and (2) much of the observed variability is due to visit-to-visit fluctuations in FA trajectories (measurement error); and (3) the effect of scan date is largely contained in the fixed effects. The results about the decomposition of the observed residual variability after accounting for fixed effects are consistent with the literature. Indeed, Greven et al. (2010) showed that only 2 to 3% of the observed variability can be attributed to the longitudinal functional slope.

4.6.2 NHANES Study

The National Health and Nutrition Examination Survey (NHANES) is a large cohort study conducted by the US Centers for Disease Control (CDC) in two-year waves to assess the health and nutritional status of the US population. The objectively measured physical activity (PA) data were collected using hip-worn accelerometers on study participants in the 2003-2006 waves. The

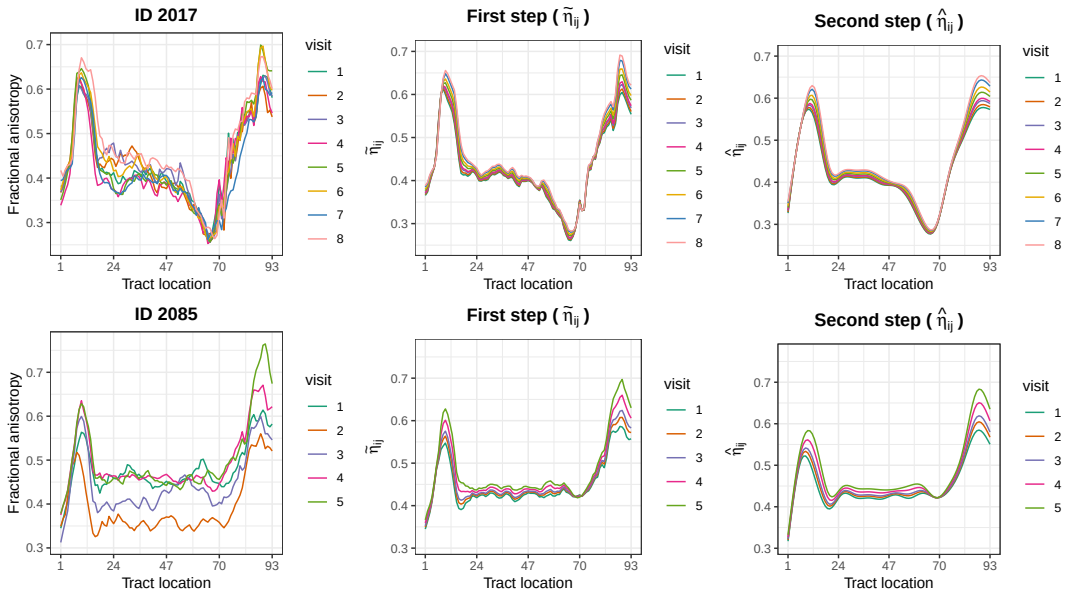


Figure 4.5: Fractional anisotropy (FA) tract profiles and estimated predictors for two study participants (first row: ID 2017, second row: ID 2085). First column: FA tract profiles for the corpus callosum over multiple visits. Second column: pointwise estimated predictor $\tilde{\eta}_{ij}$. Third column: smoothed estimated predictor $\hat{\eta}_{ij}$ of the pointwise predictors.

accelerometry data are publicly available as minute-level activity counts (AC), a proprietary measure of PA, and can be accessed in an analysis-ready format through the `R rnhanesdata` package (Leroux et al., 2019). Specifically, the accelerometry data were collected on 14631 individuals in the NHANES 2003-2004 and 2005-2006 waves. In this study, we focus on individuals with age between 18 and 30 at the time of accelerometer wear. In addition, we exclude individuals who had less than 3 days of data with at least 10 hours of estimated wear time or were labeled as poor data quality by NHANES. The number of available days vary between individuals with a maximum of 7. The final data include 1680 individuals with 8765 days, each with 1440 observations per day for a total of 12621600 minute-level observations.

We would like to investigate whether being non-sedentary is associated with gender, age, and day of the week (e.g., Monday, Tuesday). For the jth day of the ith study participant, the longitudinal functional response $Y_{ij}(s)$ is now a binary indicator, which equals to 1 if the AC at minute $s \in \{1, ..., 1440\}$ exceeds 100 and 0 if not. The fixed effect $\boldsymbol{X}_{ij} = [1, X_{ij1}, X_{ij2}, X_{ij3}, \boldsymbol{X}_{ij4}^T]^T$ where $X_{ij1} = j$ is the day number, X_{ij2} is a binary indicator of sex (female= 1), X_{ij3} is the age, and $\boldsymbol{X}_{ij4}$ is a 6×1 binary vector indicating the day of the week of day j with order $\{\text{Mon}, \text{Tue}, \ldots, \text{Sat}\}$. For example, for study participant i if day 3 is Tuesday, then the second element of $\boldsymbol{X}_{i34}$ is 1 while all others are 0; if day 3 is Sunday, all elements in $\boldsymbol{X}_{i34}$ are 0. Denote $\boldsymbol{f}(s) = [f_1(s), ..., f_6(s)]^T$. At every minute s of the day we fit a random intercept and slope model

$$\text{logit}\{\Pr(Y_{ij}(s) = 1 | \boldsymbol{X}_{ij}, \boldsymbol{u}_i)\} = \beta_0(s) + X_{ij1}\beta_1(s) + X_{ij2}\beta_2(s) + X_{ij3}\beta_3(s) + \boldsymbol{X}_{ij4}^T \boldsymbol{f}(s) + u_{i0}(s) + u_{i1}(s) \cdot j \,,$$

where $[u_{i0}(s), u_{i1}(s)]^T \sim N\{0, \Sigma_u(s)\}$. This GLMM is fit 1440 times at every location s. The same penalized spline smoother using cubic basis with 15 equally spaced knots and REML estimated smoothing parameter is applied to smooth estimated coefficients from pointwise fits separately.

Figure 4.6 displays the estimated coefficients together with the 95% pointwise (dark gray shaded area) and joint (light gray shaded area) confidence bands based on 100 bootstrap replicates. The shape of the functional intercept is consistent with the published literature and indicates less activity during the night, a sharp increase in the morning, sustained activity during the day and a reduction of activity in late evening. The effect of sex in this age group (18 to

30) is statistically significant throughout most of the day with the exception of the late afternoon/early evening period ($\approx$ 4-10PM). The sex effect is strongest during the late evening/early morning hours, when, on average, females are less active. This result could correspond to more restful sleep, more sleep, or higher compliance to study protocol for women. The effect of age is also highly significant during the night and early to late morning indicating that older study participants in this age group (18 to 30) tend to have less activity during the night and more activity in the morning and early afternoon. These findings are consistent with those reported by Varma et al. (2017).

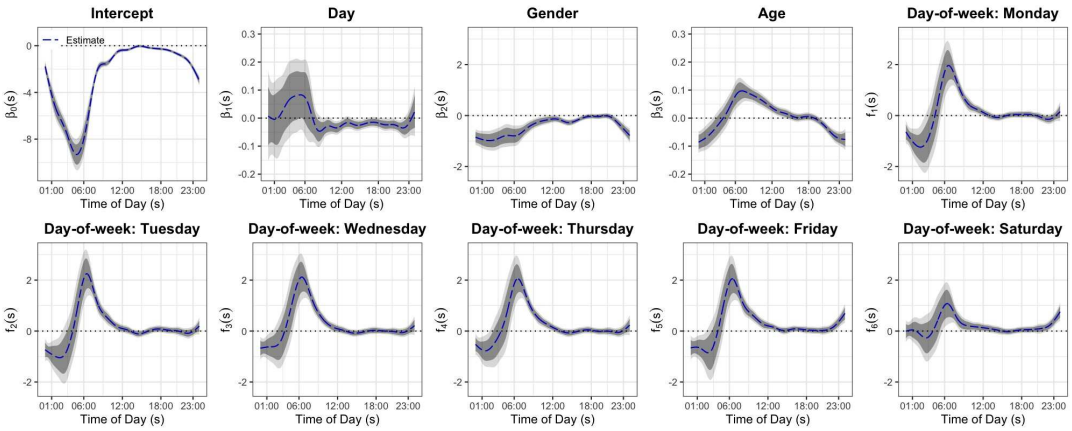


Figure 4.6: Estimated coefficients from the NHANES data application. Smoothed coefficient estimates are denoted using blue dashed lines. Pointwise and joint 95% confidence intervals are shown as the dark and light gray shaded area, respectively.

The fixed effect of day number ($\hat{\beta}_1(s)$) indicates that individuals are slightly more likely to be active during the nighttime hours and less likely to be active during normal waking hours as a function of day, though the nighttime effect is only pointwise significant during the period roughly corresponding to 4AM-6AM, but not significant when considering the joint confidence bands. This suggests that there may be a small "habituation effect", particularly during the

daytime. Habituation effects was proposed as a potential psychological effect of increasing PA at the beginning of wearing a device merely by its presence. Compared to Sundays (the reference category), weekdays correspond to lower levels of activity during the pre-dawn hours (12AM-4AM), higher levels activity in the morning (6AM-11AM), and about the same levels of activity during the afternoon and evening hours. Fridays and Saturdays correspond to more activity in the evening than Sundays. Saturdays tend to have lower activity in the morning compared to weekdays, but more activity than Sundays. These results are consistent with previous findings that individuals tend to be less active during the night and more active during the day on weekdays. These differences are likely due to social behaviors on the weekends and obligations related to school and/or work during the weekdays.

Computationally, the initial model was fit in 672 minutes, with bootstrapping requiring $\sim$ 67200 additional minutes (1120 hours); results are reported on a standard laptop. However, because of the parallel nature of our method, each location-specific fit can be estimated separately and combined at the end. This would reduce the computation time by 3 orders of magnitude, as most computational time is taken by fitting 1440 univariate GLMMs. A fully parallel implementation the entire procedure would take $\sim$ 30 seconds for one model fit and $\sim$ 1 hour for the inferential procedure. Though this may seem like a long time, we are not aware of any other methods that could fit such a model for this large longitudinal functional data set.

4.7 Discussion

We have introduced a fast univariate inferential approach for longitudinal functional models, a computationally efficient method for quantifying the association between covariates and a broad family of longitudinal functional outcomes. The model is estimated using a three-step procedure: (1) fit a series of separate standard longitudinal mixed models; (2) smooth estimators along the functional domain; and (3) construct pointwise and joint confidence bands using analytic approaches for Gaussian data or a nonparametric bootstrap of study participants for any type of data. The proposed method is highly computationally efficient because the first step can be parallelized to allow fitting large high-dimensional datasets. The second step is actually optional and one can either smooth or not smooth the resulting coefficients. Building joint confidence bands is a crucial component for conducting joint inference and performing testing multiplicity adjustment. Another major advantage of the proposed approach is its conceptual simplicity and availability in the R software. Most importantly, methods are "read-and-use", meaning that data scientists with a working knowledge of GLMMs can easily implement and apply our procedures.

The most important methodological contribution of this paper is to provide practical methods for building pointwise and joint confidence bands for very large longitudinal functional datasets. Simulation results suggest that our method achieves similar estimation accuracy and nominal coverage compared with existing methods, while the computation is much faster when the number of subjects is large (> 100).

Our work is not without limitations. First, the smoothing parameter selection assumes that the residuals of the raw estimated fixed effects around the true coefficient along the functional domain are independent. Second, changing the quantile of the confidence bands from 2 to 2.2 (lengthening the confidence bands by 10%) for bootstrap inference works well in our simulation study, but a more rigorous procedure and associated simulations may be necessary. Third, our method is only applicable to concurrent functional models, and can only take into account functional covariates that are measured on the same grid as the functional responses. Fourth, we focus on the fixed effects inference in this paper. While the inference for visit-specific predictions and other metrics falls into a similar framework, as introduced in Section 4.3.3, the extension is nontrivial and exceeds the scope of the current paper.

4.8 Supplementary Material

The supplementary material of this project is available at https://doi.org/10.1080/10618600.2021.1950006.

References

Greven, Sonja, Ciprian Crainiceanu, Brian S Caffo, and Daniel S Reich (2010). "Longitudinal functional principal component analysis". In: *Electronic Journal of Statistics*, pp. 1022–1054.

Goldsmith, Jeff, Jennifer Bobb, Ciprian M Crainiceanu, Brian Caffo, and Daniel Reich (2011). "Penalized functional regression". In: *Journal of Computational and Graphical Statistics* 20.4, pp. 830–851.

Goldsmith, Jeff, Ciprian M Crainiceanu, Brian Caffo, and Daniel Reich (2012). "Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.3, pp. 453–469.

Scheipl, Fabian, Ana-Maria Staicu, and Sonja Greven (2015). "Functional additive mixed models". In: *Journal of Computational and Graphical Statistics* 24.2, pp. 477–501.

Leroux, Andrew, Junrui Di, Ekaterina Smirnova, Elizabeth J Mcguffey, Quy Cao, Elham Bayatmokhtari, Lucia Tabacu, Vadim Zipunnikov, Jacek K Urbanek, and Ciprian Crainiceanu (2019). "Organizing and analyzing the activity data in nhanes". In: *Statistics in Biosciences* 11.2, pp. 262–287.

Smirnova, Ekaterina, Andrew Leroux, Quy Cao, Lucia Tabacu, Vadim Zipunnikov, Ciprian Crainiceanu, and Jacek K Urbanek (2019). "The Predictive Performance of Objective Measures of Physical Activity Derived From Accelerometry Data for 5-Year All-Cause Mortality in Older Adults: National Health and Nutritional Examination Survey 2003–2006". In: *The Journals of Gerontology: Series A*.

Cui, Erjia, Ciprian M Crainiceanu, and Andrew Leroux (2020). "Additive Functional Cox Model". In: *Journal of Computational and Graphical Statistics*, pp. 1–14.

Matthews, Charles E, Kong Y Chen, Patty S Freedson, Maciej S Buchowski, Bettina M Beech, Russell R Pate, and Richard P Troiano (2008). "Amount

of time spent in sedentary behaviors in the United States, 2003–2004". In: *American Journal of Epidemiology* 167.7, pp. 875–881.

Koster, Annemarie, Paolo Caserotti, Kushang V Patel, Charles E Matthews, David Berrigan, Dane R Van Domelen, Robert J Brychta, Kong Y Chen, and Tamara B Harris (2012). "Association of sedentary time with mortality independent of moderate to vigorous physical activity". In: *PloS One* 7.6, e37696.

Yao, Fang, Hans-Georg Müller, and Jane-Ling Wang (2005). "Functional data analysis for sparse longitudinal data". In: *Journal of the American Statistical Association* 100.470, pp. 577–590.

Guo, Wensheng (2002). "Functional mixed effects models". In: *Biometrics* 58.1, pp. 121–128.

Morris, Jeffrey S and Raymond J Carroll (2006). "Wavelet-based functional mixed models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2, pp. 179–199.

Zipunnikov, Vadim, Sonja Greven, Haochang Shou, Brian Caffo, Daniel S Reich, and Ciprian Crainiceanu (2014). "Longitudinal high-dimensional principal components analysis with application to diffusion tensor imaging of multiple sclerosis". In: *The Annals of Applied Statistics* 8.4, p. 2175.

Brockhaus, Sarah, Fabian Scheipl, Torsten Hothorn, and Sonja Greven (2015). "The functional linear array model". In: *Statistical Modelling* 15.3, pp. 279–300.

Scheipl, Fabian, Jan Gertheiss, Sonja Greven, et al. (2016). "Generalized functional additive mixed models". In: *Electronic Journal of Statistics* 10.1, pp. 1455–1492.

Shou, Haochang, Vadim Zipunnikov, Ciprian M Crainiceanu, and Sonja Greven (2015). "Structured functional principal component analysis". In: *Biometrics* 71.1, pp. 247–257.

Zhu, Hongtu, Kehui Chen, Xinchao Luo, Ying Yuan, and Jane-Ling Wang (2019). "Fmem: Functional mixed effects models for longitudinal functional responses". In: *Statistica Sinica* 29.4, p. 2007.

Goldsmith, Jeff, Vadim Zipunnikov, and Jennifer Schrack (2015). "Generalized multilevel function-on-scalar regression and principal component analysis". In: *Biometrics* 71.2, pp. 344–353.

Fan, Jianqing and J-T Zhang (2000). "Two-step estimation of functional linear models with applications to longitudinal data". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.2, pp. 303–322.

Reiss, Philip T, Lei Huang, Pei-Shien Wu, Huaihou Chen, and Stan Colcombe (2017). "Pointwise influence matrices for functional-response regression". In: *Biometrics* 73.4, pp. 1092–1101.

Park, So Young, Ana-Maria Staicu, Luo Xiao, and Ciprian M Crainiceanu (2018). "Simple fixed-effects inference for complex functional models". In: *Biostatistics* 19.2, pp. 137–152.

Morris, Jeffrey S, Cassandra Arroyo, Brent A Coull, Louise M Ryan, Richard Herrick, and Steven L Gortmaker (2006). "Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study". In: *Journal of the American Statistical Association* 101.476, pp. 1352–1364.

Zhu, Hongxiao, Philip J Brown, and Jeffrey S Morris (2011). "Robust, adaptive functional regression in functional mixed model framework". In: *Journal of the American Statistical Association* 106.495, pp. 1167–1179.

Zhang, Lin, Veerabhadran Baladandayuthapani, Hongxiao Zhu, Keith A Baggerly, Tadeusz Majewski, Bogdan A Czerniak, and Jeffrey S Morris (2016). "Functional CAR models for large spatially correlated functional datasets". In: *Journal of the American Statistical Association* 111.514, pp. 772–786.

Crainiceanu, Ciprian M, Ana-Maria Staicu, Shubankar Ray, and Naresh Punjabi (2012). "Bootstrap-based inference on the difference in the means of two correlated functional processes". In: *Statistics in Medicine* 31.26, pp. 3223–3240.

Ruppert, David, Matt P Wand, and Raymond J Carroll (2003). *Semiparametric regression*. 12. Cambridge university press.

Xiao, Luo, Yingxing Li, and David Ruppert (2013). "Fast bivariate P-splines: the sandwich smoother". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pp. 577–599.

Yao, Fang, Hans-Georg Müller, Andrew J Clifford, Steven R Dueker, Jennifer Follett, Yumei Lin, Bruce A Buchholz, and John S Vogel (2003). "Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate". In: *Biometrics* 59.3, pp. 676–685.

Hall, Peter, Hans-Georg Müller, and Fang Yao (2008). "Modelling sparse generalized longitudinal observations with latent Gaussian processes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, pp. 703–723.

Cuevas, Antonio, Manuel Febrero, and Ricardo Fraiman (2006). "On the use of the bootstrap for estimating functions with functional data". In: *Computational Statistics & Data Analysis* 51.2, pp. 1063–1074.

Morris, Carl N (1983). "Parametric empirical Bayes inference: theory and applications". In: *Journal of the American statistical Association* 78.381, pp. 47–55.

Bonferroni, Carlo (1936). "Teoria statistica delle classi e calcolo delle probabilita". In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8, pp. 3–62.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57.1, pp. 289–300.

Cox, Dennis D and Jong Soo Lee (2008). "Pointwise testing with functional data using the Westfall–Young randomization method". In: *Biometrika* 95.3, pp. 621–634.

Westfall, Peter H and S Stanley Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. Vol. 279. John Wiley & Sons.

Degras, David A (2011). "Simultaneous confidence bands for nonparametric regression with functional data". In: *Statistica Sinica*, pp. 1735–1765.

Ma, Shujie, Lijian Yang, and Raymond J Carroll (2012). "A simultaneous confidence band for sparse longitudinal regression". In: *Statistica Sinica* 22, p. 95.

Cao, Guanqun, Lijian Yang, and David Todem (2012). "Simultaneous inference for the mean function based on dense functional data". In: *Journal of Nonparametric Statistics* 24.2, pp. 359–377.

Yuan, Ying, John H Gilmore, Xiujuan Geng, Styner Martin, Kehui Chen, Janeling Wang, and Hongtu Zhu (2014). "Fmem: Functional mixed effects modeling for the analysis of longitudinal white matter tract data". In: *NeuroImage* 84, pp. 753–764.

Brockhaus, Sarah, David Rügamer, and Sonja Greven (2017). "Boosting functional regression models with FDboost". In: *arXiv preprint arXiv:1705.10662*.

Goldsmith, Jeff, Fabian Scheipl, Lei Huang, Julia Wrobel, Chongzhi Di, Jonathan Gellar, Jaroslaw Harezlak, Mathew W. McLean, Bruce Swihart, Luo Xiao, Ciprian Crainiceanu, and Philip T. Reiss (2020). *refund: Regression with Functional Data*. URL: https://CRAN.R-project.org/package=refund.

Varma, Vijay R., Debangan Dey, Andrew Leroux, Junrui Di, Jacek Urbanek, Luo Xiao, and Vadim Zipunnikov (2017). "Re-evaluating the effect of age on physical activity over the lifespan". In: *Preventive Medicine* 101, pp. 102–108.

Erjia Cui

EDUCATION

2018–present **Johns Hopkins University**
PhD in Biostatistics (expected to graduate in 2023)
Advisor: Ciprian M. Crainiceanu

2014-2018 **Zhejiang University, China**
BSc in Statistics
Minor in Advanced Honor Class of Engineering Education(ACEE), Chu Kochen Honors College

PROFESSIONAL EXPERIENCE

JOHNS HOPKINS UNIVERSITY

2019-present Teaching Assistant
Department of Biostatistics, Johns Hopkins University

2018-present Research Assistant for Prof. Ciprian M. Crainiceanu
Department of Biostatistics, Johns Hopkins University

Fall 2022 Visiting Student for Prof. Philip Tzvi Reiss
Department of Statistics, University of Haifa, Israel

ZHEJIANG UNIVERSITY

2017-2018 Research Assistant for Prof. Minzhi Zhao
Department of Statistics, Zhejiang University

2016-2017 Research Assistant for Prof. Xiqun (Michael) Chen
Department of Civil Engineering and Architecture, Zhejiang University

Fall 2017 Exchange Student
Department of Mathematics, University of Zurich and ETH Zurich, Switzerland

Summer 2017 GREAT Summer Research Program Participant
Department of Biostatistics, UC Davis

INDUSTRY

Summer 2022 Data Scientist Intern
Google, Mountain View, CA

Summer 2021 Data Science Intern
Liberty Mutual Insurance, Boston, MA

EDITORIAL ACTIVITIES

Reviewer *Journal of Computational and Graphical Statistics (1)*,
Journal of Nonparametric Statistics (2),
Gait & Posture (1),
Journal of Pediatric Rehabilitation Medicine (3)
PLOS ONE (1)

PUBLICATIONS (* co-first authors)

STATISTICAL METHODOLOGY

1. **Cui E***, Li R*, Crainiceanu CM, Xiao L (2022+). Fast Multilevel Functional Principal Component Analysis. *Journal of Computational and Graphical Statistics*.
2. Li R, Xiao L, Smirnova E, **Cui E**, Leroux A, Crainiceanu CM (2022). Fixed-effects Inference and Tests of Correlation for Longitudinal Functional Data. *Statistics in Medicine*, 41(17), 3349-3364.
3. **Cui E***, Thompson EC*, Carroll RJ, Ruppert D (2022). A Semiparametric Risk Score for Physical Activity. *Statistics in Medicine*, 41(7), 1191-1204.
4. **Cui E**, Leroux A, Smirnova E, Crainiceanu CM (2022). Fast Univariate Inference for Longitudinal Functional Models. *Journal of Computational and Graphical Statistics*, 31(1), 219-230.
5. **Cui E**, Crainiceanu CM, Leroux A (2021). Additive Functional Cox Model. *Journal of Computational and Graphical Statistics*, 30(3), 780-793.
6. Xin C, Zhao M, Yao Q, **Cui E** (2020). On the Distribution of the Hitting Time for the N-urn Ehrenfest Model. *Statistics & Probability Letters*, 157, 108625.

SCIENTIFIC APPLICATIONS

7. Hwang J, Blair PW, Siddarthan T, Liu G, **Cui E**, Bai J, ..., Clark DV (2022). Do Lung Ultrasound Abnormalities Change During Hospitalization for COVID-19? *Open Forum Infectious Disease*, 9(Supplement_2), ofac492.386.
8. Blair PW, Hwang J, Pearce J, Fong TC, **Cui E**, Herrera P, ... & CCPSEI Research Team. (2022). Do Worsening Lung Ultrasound Scans Identify Severe COVID-19 Trajectories? *Frontiers in Medicine*, 3255.
9. Blair PW, Siddharthan T, Liu G, Bai J, **Cui E**, East J, ..., Clark DV (2022). Point-of-care Lung Ultrasound Predicts Severe Disease and Death Due to COVID-19: A Prospective Cohort Study. *Critical Care Explorations*, 4(8): e0732.
10. Durfee AZ, Sheppard SM, Meier EL, Bunker L, **Cui E**, Crainiceanu CM, Hillis AE (2021). Explicit Training to Improve Affective Prosody Recognition in Adults with Acute Right Hemisphere Stroke. *Brain sciences*, 11(5), 667.
11. Chen X, Liu J, Hu H, **Cui E**, Zhang S (2018). Evaluation Method and Influence Factors of Network Travel Time Reliability. *Journal of Traffic and Transportation Engineering*, 18(4), 132-142.
12. Liu J, **Cui E**, Hu H, Chen X, Chen X, Chen F (2017). Short-term Forecasting of Emerging On-demand Ride Services. *4th International Conference on Transportation Information and Safety*, 489-495.

BOOKS

13. Crainiceanu CM, Goldsmith J, Leroux A, **Cui E**. Functional Data Analysis with R. *Chapman & Hall/CRC Statistics*. (expected to be published in 2023)

HONORS AND AWARDS

2021 **June B. Culley Award** (for Outstanding School-wide Examination Paper), Johns Hopkins University

2021 **Jane and Steve Dykacz Award** (for Outstanding Paper in Medical Statistics), Johns Hopkins University

2017 **Outstanding Research Performance Award**, *GREAT* program, UC Davis

2015 **First-Class Scholarship**, Zhejiang University

TEACHING

TEACHING ASSISTANT AT JOHNS HOPKINS UNIVERSITY

2022-2023 Statistical Methods in Public Health I-II
Instructor: Marie Diener-West, Karen Bandeen-Roche
Enrollment approximately 650 students

2021-2022 Methods in Biostatistics I-II (ScM core course, lead TA)
Instructor: Ciprian M. Crainiceanu
Enrollment approximately 60 students
Deliver independent one-hour lab lectures twice a week.

2020-2021 Methods in Biostatistics III-IV (ScM core course, lead TA)
Instructor: Elizabeth Colantuoni
Enrollment approximately 40 students
Deliver independent one-hour lab lectures twice a week.

2019-2020 Probability Theory I-IV (PhD core course)
Instructors: Michael A. Rosenblum, Cristian Tomasetti, Abhirup Datta
Enrollment approximately 10 students

SCM ADVISEES (* expected graduation)

2022-2023* Mu Jin, Master of Science, Epidemiology (First job: PhD student, UNC-Chapel Hill)

2021-2022 Qier Meng, Master of Science, Biostatistics (First job: Senior Statistician, Eli Lilly and Company)

RESEARCH GRANT PARTICIPATION

CURRENT OR PAST GRANT SUPPORT

2022-2027 Statistical Methods for Multilevel Multivariate Functional Studies
Studies Agency: NIH/NINDS (PI: Crainiceanu)
Grant Type: R01 NS060910
Current Effort: 50% (Research Assistant)
Potential Future Effort: 20% (Faculty)

2021-2022 Prophylaxis and Treatment of COVID19 - Adaptive Platform Trial: PROTECT-APT
Studies Agency: Henry Jackson Foundation (PI: Siddharthan)
Current Effort: 50% (Research Assistant)
Potential Future Effort: 20% (Faculty)

ACADEMIC SERVICE

PROFESSIONAL MEMBERSHIPS

- American Statistical Association
- International Biometric Society (East North America Region (ENAR))

PROGRAM DEVELOPMENT

2022 Topic Contributed Session Chair, JSM 2022

2021 Invited Session Organizer and Chair, JSM 2021

2019-2020 Departmental Journal Club Organizer

PRESENTATIONS (* invited)

2023* FDAWG, Department of Statistics, North Carolina State University, February 23, Raleigh, NC

2023* Division of Biostatistics, University of Minnesota, February 3, Minneapolis, MN

2023* Department of Information Systems and Statistics, Baruch College, January 26, New York, NY

2023* Department of Biostatistics, University of Florida, January 18, Gainesville, FL

2022* CMStatistics 2022, December 18, online

2022* Department of Mathematical Sciences, University of Texas at Dallas, December 9, Richardson, TX

2022* Department of Statistics, Ohio State University, December 6, Columbus, OH

2022* Department of Biostatistics and Epidemiology, Rutgers University, November 30, online

2022* FDAWG, Department of Biostatistics, Columbia University, November 29, New York, NY

2022* Department of Population Health Sciences, University of Utah, November 21, online

2022* Department of Statistics, University of Haifa, November 2, Haifa, Israel

2022* JSM 2022, August 10, Washington, DC

2022* The EBA Training Program Annual Joint Presentation, Johns Hopkins University, Apr 26, online

2022* Department of Biostatistics & Informatics, Colorado School of Public Health, Apr 21, online

2022 ENAR 2022, March 28, Houston, TX

2021* JSM 2021, August 9, online

2021 ENAR 2021, March 15, online

2021* Division of Biostatistics, Thomas Jefferson University, Feb 11, online

2020 JSM 2020, August 3, online

2020 ENAR 2020, March 23, online

2020 SLAM Working Group, Johns Hopkins University, Jan 31, Baltimore, MD

WORKSHOP PARTICIPATION

2020 Use of Wearable and Implantable Devices in Health Research, February 23-28, Banff, Canada

[Compiled on March 28, 2023]