ABSTRACT

WANG, WENYI. Expectation-Maximization (EM) Based Functional Data Methods for Longitudinal Data and Survival Data. (Under the direction of Luo Xiao).

Longitudinal data, where repeated measurements are collected over time, and survival data, where the focus is on time-to-event outcomes, are fundamental in biomedical and epidemiological research. Joint modeling of longitudinal and survival data is essential when these two types of data are correlated, allowing for improved inference and prediction. In many modern studies, the longitudinal outcomes may not be continuous or normally distributed but instead arise from exponential family distributions, such as binary or count data. Challenges arise from the irregular and sparse observation of longitudinal measurements, the non-Gaussian distributional structure of the data, and the joint modeling of longitudinal and survival outcomes. This dissertation addresses these challenges by developing statistical methods based on functional data analysis frameworks, including functional principal component analysis (FPCA), exponential family functional principal component analysis (EF-FPCA), Expectation-Maximization (EM) algorithms, and penalized splines. These approaches enable flexible and efficient modeling and estimation for complex joint longitudinal and survival data structures.

In Chapter 2, we develop an integrative joint modeling framework for multivariate sparse functional outcomes and survival data arising from multiple Alzheimer's disease cohorts. The proposed model captures shared disease variation patterns across cohorts while allowing cohort-specific disease progression trajectories and survival associations. Estimation is performed via an EM algorithm incorporating penalized splines to ensure smooth estimation of functional components. Application to three Alzheimer's cohorts demonstrates that the method successfully recovers meaningful disease progression patterns while accounting for missing-by-design outcomes and inter-cohort variability. Extensive simulation studies validate the robustness and estimation accuracy of the proposed approach.

In Chapter 3, we study longitudinal head impact exposure data collected from college football athletes, where the outcomes are count-valued functional observations. To model these data, we propose a tensor decomposition approach combined with exponential family functional principal component techniques to capture variations across schools, player positions and season time. An efficient EM-based estimation procedure using penalized splines is developed. Analysis of the head acceleration event (HAE) dataset uncovers interpretable differences in exposure trajectories by position and school, and simulation results demonstrate the method's effectiveness.

In Chapter 4, we propose a functional joint modeling strategy for longitudinal and survival outcomes with a baseline functional covariate, motivated by the Objective Physical Activity and Cardiovascular Health (OPACH) study. The longitudinal data is modeled by functional data methods and links to the survival data via functional principal component scores. Daily sitting bouts profiles derived from accelerometry data are treated as functional covariates, and their associations with longitudinal physical function measures and survival outcomes are modeled through linear scalar-on-function regression forms. The model is estimated via an EM algorithm combined with penalized spline smoothing. Application to the OPACH data reveals scientifically meaningful relationships between sitting bout patterns, physical function, and mortality risk, with simulations confirming the proposed method's robust performance.

Expectation-Maximization (EM) Based Functional Data Methods for Longitudinal Data and Survival Data

by
Wenyi Wang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2025

APPROVED BY:

Marie Davidian	Sujit Ghosh
Jaroslaw Harezlak External Member	Wenbin Lu

Luo Xiao
Chair of Advisory Committee

DEDICATION

To my parents.

BIOGRAPHY

The author was born in Hunan, China, in 2000. She received the Bachelor of Science degree in Statistics from the University of Science and Technology of China in 2020 and the Master of Science degree in Data Science–Statistics from the University of Wisconsin–Madison in 2021. Following her graduation, she pursued doctoral studies in the Department of Statistics at North Carolina State University. Under the advisement of Dr. Luo Xiao, she will complete her Ph.D. in Statistics in July 2025.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Luo Xiao, for his unwavering support, guidance, and encouragement throughout my Ph.D. journey. His mentorship has shaped me not only as a researcher but also as a thinker. I am especially thankful for his patience and insight during both academic and personal challenges. His passion for research, dedication to excellence, and thoughtful approach to problem-solving have deeply inspired me. Through his guidance, I have grown to become more independent, thoughtful, and confident in my work. I am truly fortunate to have had him as my advisor.

I would also like to thank Dr. Jaroslaw Harezlak, whose collaboration and guidance have been invaluable to my projects. I am especially grateful for the financial support he provided during a key part of my doctoral studies.

I am also sincerely thankful to my dissertation committee members: Dr. Marie Davidian, Dr. Sujit Ghosh, Dr. Jaroslaw Harezlak, and Dr. Wenbin Lu. Their thoughtful feedback and invaluable suggestions have strengthened this dissertation and broadened my perspective on statistical research. Each of them has provided guidance and support in unique ways, and I am grateful for the time and expertise they have shared with me. I also want to thank all the collaborators I have had the privilege to work with during my Ph.D., including Dr. Chongzhi Di, Dr. Jaroslaw Harezlak, Dr. Ilsuk Kang, Dr. Ruonan Li, Dr. Sheng Luo, Dr. Alok S. Shah, and Dr. Brian D. Stemper. Their insights and contributions have greatly enriched my research and broadened my academic experience.

Special thanks go to Dr. Kimberly Sellers, as the Department Head, whose kind messages and support during a difficult time in my life meant more than words can express. I would also like to thank Dr. Wenbin Lu, as the Director of the Statistics Graduate Program, for his continued support throughout my graduate studies. I am also grateful to Ms. Isabel Lopez Ayala for her generous help and administrative support during my time at NC State.

I am profoundly grateful to my boyfriend, Shengjie Liu, for his constant encouragement, understanding, and companionship. His belief in me and his warmth have helped me persevere through many ups and downs.

Finally, I would like to thank my parents, Zexi Wang and Qiaoyun Su, for their unconditional love and unwavering support. Their strength and belief in me have been a foundation through every stage of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Outline and Contributions

Longitudinal and survival data are two fundamental types of data structures commonly encountered in biomedical, epidemiological, and social science studies (Li et al. 2022; Wulfsohn and Tsiatis 1997; Tsiatis and Davidian 2004). Longitudinal data involve repeated measurements collected from the same observation unit over time, often at irregular and subject-specific time points, allowing for the study of within-subject changes and temporal patterns. Traditionally, longitudinal data are assumed to be continuous and normally distributed; however, in many real-world applications, longitudinal outcomes belong to the exponential family of distributions. Examples include binary outcomes (e.g., presence or absence of a condition) and count data (e.g., number of incidents) (Serban et al. 2013; Huang et al. 2014). Such exponential family longitudinal data require specialized modeling techniques that account for their exponential family nature.

Survival data involve the time until an event of interest occurs, often subject to right-censoring when the event is not observed within the study period. The joint analysis of longitudinal and survival data is important when longitudinal trajectories are predictive of survival outcomes. However, several challenges arise in this context, including handling irregular and sparse longitudinal measurements, linking longitudinal and survival processes flexibly, and developing computationally efficient estimation algorithms. This dissertation focuses on developing statistical methods that address these challenges by leveraging functional data analysis techniques, the Expectation-Maximization (EM) algorithm (Dempster et al. 1977), and nonparametric smoothing methods such as penalized splines. The primary goals are to model complex functional structures, handle longitudinal and survival data efficiently, and apply these methods to real-world problems.

In chapter 2, we develop an integrative joint model for multivariate sparse functional and survival data to analyze Alzheimer's disease (AD) across multiple studies. To address missing-by-design outcomes in multi-cohort studies, our approach extends the multivariate functional mixed model (MFMM)(Li et al. 2022), which integrates longitudinal outcomes to extract shared disease progression trajectories and links these outcomes to time-to-event data through a parsimonious survival model. This framework balances flexibility and interpretability by modeling shared progression trajectories while accommodating cohort-specific mean functions and survival parameters. For efficient estimation, we incorporate penalized splines into an EM algorithm. Application to three AD cohorts demonstrates the model's ability to capture disease trajectories and account for inter-cohort variability. Simulation studies confirm its robustness and accuracy, highlighting its value in advancing the understanding of AD progression and supporting clinical decision-making in multi-cohort settings.

In chapter 3, we apply functional data analysis methods to count-valued longitudinal data arising from head impact exposure in college football athletes. College football players often sustain impacts that might lead to concussion. Helmet-attached accelerometers track instantaneous head acceleration and are capable of detecting impacts. The dataset analyzed in this work records longitudinally collected Head Acceleration Events (HAEs) from college football players playing at different positions and from several schools. We model the number of HAEs as count data and use functional data methods to quantify the varying patterns of HAEs in a season and evaluate how they might differ across positions and from school to school. In particular, we decompose a matrix of functions via the Tucker decomposition of tensors, which leads to an interpretable and stable model estimation of school and position-specific mean patterns of HAEs. For the proposed model, we develop an efficient model estimation algorithm that combines expectation-maximization (EM) and nonparametric function estimation via penalized splines. The proposed model, when applied to the HAEs dataset, gives sensible results. Simulation studies also demonstrate good performance of the proposed model estimation algorithm.

In chapter 4, we propose a joint model for longitudinal and survival data with a baseline functional covariate, motivated by the Objective Physical Activity and Cardiovascular Health (OPACH) study. Daily profiles of sitting bouts of different lengths, derived from accelerometry data in elderly women, are treated as functional covariates. The longitudinal data (physical function via survey) is modeled by functional data methods and links to the survival data (all-cause mortality) via functional principal component scores. In both sub-models, linear scalar-on-function regression forms are used to account for the effect of the baseline functional covariate. For model estimation, an efficient EM algorithm coupled with penalized splines is proposed. When applied to the OPACH study, the proposed model gives scientifically interpretable associations of sitting-bout-length profiles with physical function and mortality. A simulation study is also conducted to confirm the good performance of the proposed model.

To set the stage for the methods developed in this dissertation, the remainder of this chapter summarizes the key statistical methodologies employed in the subsequent chapters. The aim is not to provide an exhaustive review, but rather to present the essential background necessary to support the technical developments that follow. In particular, this chapter highlights the Expectation-Maximization (EM) algorithm for parameter estimation with incomplete data, functional data analysis techniques for modeling complex trajectories, and nonparametric smoothing methods via penalized splines.

1.2 Key Statistical Methods and Techniques

1.2.1 Expectation-Maximization (EM) Algorithm

The Expectation-Maximization (EM) algorithm (Dempster et al. 1977) is a widely used iterative method for obtaining maximum likelihood estimates when data are incomplete, contain latent variables, or have missing observations. Many statistical models involve hidden structures that complicate direct maximization of the observed-data likelihood.

Let $\boldsymbol{Y}$ denote the observed data and $\boldsymbol{Z}$ denote the unobserved latent data or missing values. The complete data is $(\boldsymbol{Y}, \boldsymbol{Z})$, and let $\boldsymbol{\theta}$ represent the vector of parameters to be estimated. The maximum likelihood estimate (MLE) of the unknown parameters $\boldsymbol{\theta}$ is determined by maximizing the marginal likelihood of the observed data:

$$L(\boldsymbol{\theta};\boldsymbol{Y}) = p(\boldsymbol{Y}|\boldsymbol{\theta}) = \int p(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})d\boldsymbol{Z} = \int p(\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{\theta})p(\boldsymbol{Z}|\boldsymbol{\theta})d\boldsymbol{Z},$$

where $p(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})$ is the complete-data joint density. Direct maximization of $L(\boldsymbol{\theta};\boldsymbol{Y})$ is often infeasible due to the intractability of the integral over $\boldsymbol{Z}$.

The EM algorithm addresses this difficulty by iteratively alternating between two steps:

- **E-step (Expectation step)**: Calculate the expected value of the complete-data log-likelihood with respect to the conditional distribution of the unobserved data $\boldsymbol{Z}$ given the observed data $\boldsymbol{Y}$ and the current parameter estimates $\boldsymbol{\theta}^{(t)}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{Z}\sim p(\cdot|\boldsymbol{Y},\boldsymbol{\theta}^{(t)})}\big[\log p(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{\theta})\big].$$

- **M-step (Maximization step)**: Update the parameter estimates by maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

The EM algorithm guarantees that the marginal log-likelihood of the observed data does not decrease at each iteration, i.e.,

$$\ell(\boldsymbol{\theta}^{(t+1)};\boldsymbol{Y}) \geq \ell(\boldsymbol{\theta}^{(t)};\boldsymbol{Y}),$$

where $\ell(\boldsymbol{\theta};\boldsymbol{Y}) = \log L(\boldsymbol{\theta};\boldsymbol{Y})$ is the observed-data (i.e., marginal) log-likelihood. Under mild regularity conditions, the sequence $\boldsymbol{\theta}^{(t)}$ converges to a stationary point of the observed-data likelihood function (Wu 1983). Although an EM iteration increases the observed-data likelihood, there is no guarantee that the algorithm converges to the global maximum; it may converge to a local maximum depending on the starting values.

The EM framework is particularly advantageous when the complete-data log-likelihood belongs to a convenient family (e.g., Gaussian distribution), making the E-step analytically tractable and the M-step a standard optimization problem. However, in situations where the E-step is computationally intractable, such as when survival outcomes are involved, variants such as the Monte Carlo EM (MCEM) algorithm (Levine and Casella 2001) are employed. In MCEM, the expectation in the E-step is approximated using Monte Carlo simulation by generating samples from the conditional distribution of the missing data given the observed data and current parameters.

1.2.2 Functional Principal Component Analysis (FPCA)

Functional data analysis (FDA) is a statistical framework for analyzing data observed as functions, curves, or surfaces over a continuous domain such as time or space (Ramsay and Silverman 2005). Unlike traditional data, which are recorded as scalar measurements, functional data represent a smooth and continuous process. For example, in medicine, a patient's heart rate may be recorded as a continuous curve over time. In environmental science, temperature and rainfall are measured as functions across days or months. Functional data are widely used in fields such as healthcare (Ohalete et al. 2024), neuroscience (Rügamer et al. 2018), chemistry (Brockhaus et al. 2015), and biophysics (Li et al. 2014). Although FDA assumes an underlying continuous process, real-world data are observed discretely and are often limited by practical constraints such as sampling frequency or observation time. These data are inherently infinite-dimensional, presenting unique challenges for analysis. FDA provides tools that treat observation as an entire function rather than a finite set of values, allowing researchers to retain all essential information without reducing the data to summary measures. However, the infinite/high dimensionality of functional data requires specialized methods to manage their complexity and ensure computational efficiency.

One of the most widely used techniques in FDA is Functional Principal Component Analysis (FPCA) (Silverman 1996; Yao et al. 2005), which is an extension of classical Principal Component Analysis (PCA) to functional data. FPCA reduces the dimensionality of functional data by identifying the dominant patterns of variation and summarizing the data using a finite number of principal functional components. These components explain the primary sources of variability while preserving the smoothness and continuity of the original data. This ability to simplify complex functional data while retaining essential features has made FPCA a foundational tool in FDA.

Consider a set of observations $\{(Y_{ij}, T_{ij}) : i = 1, \ldots, n; j = 1, \ldots, m_i\}$, where Y_{ij} denotes the response of subject i observed at time point $T_{ij} \in \mathscr{T}$, with $\mathscr{T}$ a closed interval. The conceptual framework for FPCA assumes that the observed data Y_{ij} are noisy realizations of a smooth underlying random function $X(\cdot)$:

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent, identically distributed random variables. The functions $X_1(t), \ldots, X_n(t)$ are assumed to be independent and identically distributed (i.i.d.) copies of a stochastic process $X(t)$ with mean function $\mu(t) = \mathbb{E}\{X(t)\}$ and covariance function $\mathscr{C}(s,t) = \mathrm{Cov}\{X(s), X(t)\}$.

By Mercer's theorem (Mercer 1909), the covariance function can be expressed as:

$$\mathscr{C}(s,t) = \sum_{\ell \geq 1} \lambda_\ell \phi_\ell(s) \phi_\ell(t),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are eigenvalues with associated orthonormal eigenfunctions $\phi_\ell(t)$, i.e., $\int_{\mathscr{T}} \phi_\ell(t)\phi_{\ell'}(t)dt = \delta_{\ell\ell'}$. By the Karhunen–Loève representation (Karhunen 1947; Loeve 1945), each individual function $X_i(t)$ can be expressed as:

$$X_i(T_{ij}) = \mu(T_{ij}) + \sum_{\ell \geq 1} \xi_{i\ell} \phi_\ell(T_{ij}),$$

where $\xi_{i\ell}$, for $\ell \geq 1$, are uncorrelated random Functional Principal Component (FPC) scores with mean zero and variance λ_ℓ.

In practice, the infinite sum is truncated to retain a finite number L of components:

$$X_i(T_{ij}) = \mu(T_{ij}) + \sum_{\ell=1}^{L} \xi_{i\ell}\phi_\ell(T_{ij}),$$

where L is selected to capture a desired proportion of the total variation, often based on criteria such as the cumulative percentage of variance explained or information criteria (such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)).

Functional data arise in two typical observational settings. In one case, measurements are collected over a dense and regular grid of time points, often resembling high-frequency multivariate data. In the other, data are recorded at irregular and individual-specific time points, leading to what is commonly referred to as sparse functional data, similar in structure to traditional longitudinal data.

In the case of dense and regular observations, smooth estimates of the mean function can be obtained via smoothing over $\mathscr{T}$ using, for example, penalized splines as we will discuss in Subsection 1.2.4. Smooth estimates of the covariance function $\mathscr{C}(s,t)$ can be obtained via smoothing over $\mathscr{T} \times \mathscr{T}$ using methods of bivariate smoothing, such as tensor-product bivariate P-splines (Eilers and Marx 2003) and thin plate regression splines (Wood 2003). Alternatively, Xiao et al. (2016) proposed the fast covariance estimation (FACE) method based on penalized splines to directly estimate the covariance function. However, their work was developed for high-dimensional dense functional data and the extension to sparse data is far from obvious. For a study of FDA for sparse functional data, see Yao et al. (2003) and Yao et al. (2005). Several alternative methods for sparse functional data also exist in the literature: James et al. (2000a) used reduced rank spline mixed effects models, Peng and Paul (2009) proposed a geometric approach under the framework of marginal maximum likelihood estimation and Xiao et al. (2018) proposed the FACE method for directly estimating the covariance function of sparse functional data and longitudinal data. What's more, theoretical properties of functional data analysis have been well studied in several papers (Li and Hsing 2010; Cai and Yuan 2011; Zhang and Wang 2016; Xiao 2020).

FPCA provides a parsimonious low-dimensional representation of functional data by reducing infinite-dimensional functions to finite-dimensional functional principal components. This dimension reduction is essential for subsequent modeling tasks, including regression, clustering, and joint modeling of longitudinal and survival outcomes.

1.2.3 Exponential Family Functional Principal Component Analysis (EF-FPCA)

Despite the broad applicability of FPCA, it relies on the assumption that the functional data arise from a Gaussian process, which is often unrealistic in many real-world applications. In practice, functional data can arise from exponential family distributions. For example, in bioinformatics, gene expression levels are often modeled as count data using Poisson or negative binomial distributions. In social sciences, survey responses typically involve binary outcomes, such as yes/no answers. Other examples include spectral backscatter data from infrared light detection and longitudinal studies of drug use (Serban et al. 2013; Huang et al. 2014). Applying traditional FPCA to such data can result in biased estimates and fail to capture important features, such as distributional properties like skewness and specific response patterns.

Exponential Family Functional Principal Component Analysis (EF-FPCA) was developed to address these limitations by extending FPCA to handle data from exponential family distributions. The exponential family includes many commonly encountered distributions, such as Poisson for counts and Bernoulli for binary outcomes. Instead of modeling the observed data directly, EF-FPCA relates the expected value of observed data to a smooth latent process through a link function. This approach ensures that the analysis respects the distributional properties of the data while retaining the ability to model smooth functional trends, making EF-FPCA a powerful tool for analyzing functional data in scenarios where traditional FPCA is unsuitable.

EF-FPCA Framework

Using similar notation to FPCA, let Y_{ij} denote the observed response for subject i at time T_{ij}. In EF-FPCA, Y_{ij} is assumed to follow a distribution in the canonical exponential family, with probability density function:

$$f(Y_{ij}|\eta_{ij},\sigma) = \exp\left\{\frac{Y_{ij}\eta_{ij} - b(\eta_{ij})}{a(\sigma)} + c(Y_{ij},\sigma)\right\},$$

where η_{ij} is the canonical parameter, σ is a dispersion parameter, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions (McCullagh and Nelder 1989). The first order derivative of $b(\eta_{ij})$, denoted by $b'(\eta_{ij})$, equals the expectation $\mathbb{E}(Y_{ij})$, and the canonical link function $g(\mathbb{E}(Y_{ij})) = b'^{-1}(\mathbb{E}(Y_{ij}))$ transforms the observed mean into the canonical parameter space. This transformation is fundamental in EF-FPCA, as it enables variability in the data to be analyzed in the canonical parameter space.

EF-FPCA extends traditional FPCA by modeling the canonical parameter η_{ij} instead of the observed functional data Y_{ij} directly. Specifically, it assumes:

$$g(\mathbb{E}(Y_{ij})) = \eta_{ij} = X_i(T_{ij}),$$

where $\{X_i(t), t \in \mathscr{T}, i = 1, \ldots, n\}$ are independent realizations of $X(t)$, a latent process driving the behavior of the observed functional process. A key modeling assumption in many EF-FPCA literatures (e.g. Hall et al. 2008; Gertheiss et al. 2017; Huang et al. 2014) is that $X(t)$ is a latent Gaussian process, implying that any finite set of random variables $\{X_i(T_{i1}), \ldots, X_i(T_{im_i})\}$ follows a multivariate Gaussian distribution. Recent works, such as Li et al. (2018) and Zhong et al. (2023), have relaxed this Gaussian assumption, broadening the applicability of EF-FPCA.

Analogous to FPCA, the latent process $X_i(t)$ can be represented through a finite-dimensional Karhunen–Loève expansion: $X_i(T_{ij}) = \mu(T_{ij}) + \sum_{\ell=1}^{L} \xi_{i\ell}\phi_\ell(T_{ij})$. EF-FPCA thus provides a flexible framework for analyzing functional data arising from exponential family distributions while preserving the advantages of principal component decomposition.

Overview of EF-FPCA Methods

FPCA methods for Gaussian functional data typically involve either smoothing the covariance function (Staniswalis and Lee 1998; Xiao et al. 2016; Yao et al. 2003) or using probabilistic model-based approaches with explicit likelihood assumptions (Van Der Linde 2008; James et al. 2000b). In the covariance-based approach, FPCA decomposes the variance-covariance matrix of the demeaned functional observations to identify the main modes of variation (Yao et al. 2005; Goldsmith et al. 2013). Extending this idea to exponential family functional data, Hall et al. (2008) adapted the methods of Yao et al. (2005) by introducing a latent Gaussian process (LGP) model. Building on this foundation, Serban et al. (2013) refined and extended this approach for binary functional data in rare events. However, as demonstrated in Gertheiss et al. (2017), the LGP model by Hall et al. (2008) has an inherent bias due to reliance on a marginal rather than conditional mean estimate (Wrobel et al. 2019).

Probabilistic FPCA provides an alternative to covariance smoothing by modeling PCA within a likelihood-based framework, making it suitable for handling sparse or irregular data. Many Bayesian methods have been developed for FPCA (Van Der Linde 2009; Goldsmith et al. 2015; Nolan et al. 2025). And among frequentist approaches, Gertheiss et al. (2017) proposed a two-step method that performs FPCA through a generalized additive mixed model (GAMM) to accommodate exponential family functional data, while (Scheipl et al. 2015) and (Leroux et al. 2023) applied this idea through functional additive mixed models (FAMMs) and fast generalized FPCA (fast-GFPCA), respectively. Some other EF-FPCA methods estimate parameters within a probabilistic framework using the Expectation-Maximization (EM) algorithm. Huang et al.

(2014) proposed a Monte Carlo EM (MCEM) algorithm, while Li et al. (2014) improved computational efficiency by employing a “pseudo" normal approximation. Additionally, Wrobel et al. (2019) developed a fast binary FPCA method that employs a variational approximation to the binomial likelihood (Jaakkola and Jordan 1997). All EF-FPCA methods mentioned above assume a latent Gaussian process, following the foundation set by Hall et al. (2008). However, recent works relax this assumption for broader applicability. Li et al. (2018) introduced Exponential Family FPCA (EFPCA) for dense, regularly spaced data, while Zhong et al. (2023) developed Sparse Logistic FPCA (SLFPCA) for binary functional data. These developments highlight EF-FPCA's versatility in analyzing exponential family functional data across diverse applications.

Challenges with EF-FPCA

Implementing EF-FPCA comes with several challenges. The high dimensionality of functional data requires efficient methods to estimate the mean function and eigenfunctions while avoiding overfitting. Smoothness constraints, often implemented using regularization techniques like penalized splines or roughness penalties, are essential for producing interpretable results but also add complexity to the estimation process.

Sparse or irregular functional data pose additional difficulties. Sparse functional data, where each individual has limited observations, make it challenging to accurately reconstruct the underlying function. Similarly, irregular functional data, where observation points vary between individuals, also complicate the estimation process.

Computational complexity is another major concern. Unlike Gaussian FPCA, which benefits from closed-form solutions, most EF-FPCA methods rely on iterative optimization of a penalized likelihood function (e.g., Li et al. (2014) and Zhong et al. (2023)). This optimization process can be computationally expensive, especially for large datasets or high-dimensional functional domains. Additionally, establishing theoretical guarantees, such as consistency and convergence of estimated components, is more challenging in EF-FPCA due to the increased complexity introduced by the exponential family framework and regularization terms.

1.2.4 Nonparametric Function Estimation via Penalized Splines

Nonparametric function estimation aims to recover an unknown smooth function without imposing a strict parametric form. In functional data analysis and smoothing problems, it is often necessary to estimate mean functions and eigenfunctions in a flexible and data-driven manner. Penalized splines provide a widely used and effective approach for nonparametric smoothing, balancing flexibility with control against overfitting through regularization (Eilers and Marx 1996a; Ruppert et al. 2003a).

Suppose we wish to estimate a smooth function $f(t)$ based on observations (T_i, Y_i) for $i = 1, \ldots, n$, where T_i lies in a compact domain $\mathscr{T}$. The model is:

$$Y_i = f(T_i) + \epsilon_i,$$

with ϵ_i being independent random errors with mean zero and constant variance.

Regression splines represent $f(t)$ using a basis expansion:

$$f(t) = \sum_{k=1}^{K} \beta_k B_k(t),$$

where $\{B_k(t)\}_{k=1}^{K}$ are pre-specified spline basis functions (e.g., B-splines), and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\top$ are the corresponding coefficients to be estimated. Various choices of basis functions are available, including Fourier basis and wavelet basis (Morris and Carroll 2006), but B-spline bases are the most commonly employed for regression splines. A spline of order $r+1$ is a piecewise polynomial function of degree r that is smoothly connected at points called knots. B-splines, in particular, are nonzero only over at most $r+1$ adjacent intervals between knots, which makes them computationally faster and more stable (De Boor and De Boor 1978).

Rather than fitting $f(t)$ purely by minimizing the residual sum of squares, penalized splines introduce a roughness penalty to discourage overfitting. Specifically, the penalized loss function is:

$$\sum_{i=1}^{n} \left(Y_i - \sum_{k=1}^{K} \beta_k B_k(T_i) \right)^2 + \lambda \int_{\mathscr{T}} \left(f^{(2)}(t) \right)^2 dt,$$

where $f^{(2)}(t)$ denotes the second derivative of $f(t)$, and $\lambda > 0$ is a smoothing parameter that controls the trade-off between goodness-of-fit and smoothness. A larger λ imposes more smoothing, while a smaller λ allows more flexibility. While the second derivative penalty is the most common, other penalty forms are possible. Penalized splines can be categorized based on their penalty term and basis choice. A notable example is P-splines (Eilers and Marx 1996a), which use B-spline bases and a discrete difference penalty.

In practice, the roughness penalty can be expressed as a quadratic form:

$$\int_{\mathscr{T}} \left(f^{(2)}(t) \right)^2 dt = \boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta},$$

where $\boldsymbol{\Omega}$ is a positive semi-definite penalty matrix determined by the basis and derivative order.

Thus, the penalized least squares problem becomes:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \| \boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\beta} \|^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \right\},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is the vector of responses and $\boldsymbol{B}$ is the $n \times K$ matrix of basis function evaluations with entries $B_{ik} = B_k(T_i)$. The solution has a closed-form expression:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{B}^\top \boldsymbol{B} + \lambda \boldsymbol{\Omega} \right)^{-1} \boldsymbol{B}^\top \boldsymbol{Y}.$$

The penalized sum of squares objective can be generalized to penalized likelihood formulations, where the fidelity-to-data term is based on a likelihood rather than squared errors.

Penalized splines offer several advantages. They allow a relatively large number of basis functions without overfitting due to the penalty. Empirical and theoretical studies of penalized splines show that the number of knots does not matter as long as a relatively large number of knots are used (Ruppert 2002; Xiao 2019).They are computationally efficient due to the low dimensionality of $\boldsymbol{\beta}$ and are highly flexible, with smoothness easily tuned via λ (Wood 2017). The smoothing parameter λ can be selected using methods such as generalized cross-validation (GCV) and restricted marginal likelihood (REML).

In this dissertation, penalized splines are employed to estimate smooth functions such as mean functions and eigenfunctions in Chapters 2, 3, and 4. One of the main contributions is the incorporation of penalized splines within the EM algorithm, enabling flexible, stable, and interpretable modeling of complex functional data structures.

CHAPTER

2

A FUNCTIONAL JOINT MODEL FOR SURVIVAL AND MULTIVARIATE SPARSE FUNCTIONAL DATA IN MULTI-COHORT ALZHEIMER'S DISEASE STUDY

2.1 Introduction

[1] Alzheimer's disease (AD) is a progressive brain disorder that significantly impairs cognitive and behavioral functions. In the United States, AD is the fifth leading cause of death among individuals aged 65 and older, with an estimated 6.9 million people affected in 2024 (Alzheimer's Association 2024). The increasing prevalence of AD has lead to the establishment of large-scale longitudinal studies, such as the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Weiner et al. 2015), the National Alzheimer's Coordinating Center (NACC) (Besser et al. 2018), and the Religious Orders Study and Rush Memory and Aging Project (ROSMAP) (Bennett et al. 2018). These studies collect diverse longitudinal data, including neuropsychological and behavioral

[1] This chapter is joint work with Luo Xiao, Ruonan Li and Sheng Luo.

measurements, to monitor disease trajectories and assess risk factors over time (Table 2.1). Such measurements are critical for understanding disease progression and its associations with survival outcomes (e.g., dementia onset), ultimately informing clinical interventions and policy.

Analyzing data from multiple AD cohorts presents significant opportunities and challenges. Pooling data from multiple cohorts increases sample size and statistical power, enabling the investigation of complex risk and protective factors, including nonlinear and time-varying effects. Additionally, multi-cohort analysis improves the robustness and generalizability of predictive models across diverse subgroups, settings, and countries, making findings more clinically applicable. However, heterogeneity in baseline characteristics, study designs, and data collection protocols complicates integration, particularly when some longitudinal outcomes are systematically missing by design (Table 2.1). Modeling longitudinal outcomes in this context requires methods capable of capturing nonlinear subject-specific trajectories while accounting for cohort-specific differences. Traditional parametric models often fail to address these complexities, especially in the presence of missing-by-design outcomes.

Table 2.1: List of longitudinal outcomes and number of Mild Cognitive Impairment (MCD) subjects in three AD cohorts. Abbreviations: 1.MMSE, Mini-Mental State Exam; 2.WMSLM, Wechsler Memory Scale Logical Memory; 3.RAVLT, Rey Auditory Verbal Learning Test immediate recall; 4.SDMT, Symbol Digit Modalities Test; 5.CDR-SB, Clinical Dementia Rating; 6.ADAS, Alzheimer Disease Assessment Scale-Cognitive; 7.FAQ, Functional Assessment Questionnaire; 8.TRAILA, Trail Making Test Part A. Symbols: ✓, data available; ✗, data unavailable.

Order	Longitudinal outcome	Scaling Coefficient (β_j)	ADNI	NACC	ROSMAP
1	MMSE	1.00	✓	✓	✓
2	WMSLM	1.54	✓	✓	✓
3	RAVLT	3.68	✓	✗	✗
4	SDMT	3.65	✗	✗	✓
5	CDR-SB	-0.60	✓	✓	✗
6	ADAS	-2.70	✓	✗	✗
7	FAQ	-1.84	✓	✗	✗
8	TRAILA	-1.08	✗	✓	✗
	Number of MCI subjects		715	3707	522

Joint models (JMs) are widely used for analyzing longitudinal and survival data (Wulfsohn and Tsiatis 1997; Tsiatis and Davidian 2004) and have been extended to handle multiple longitudinal outcomes (Lin et al. 2002; Murray and Philipson 2023; He et al. 2024). However, traditional JMs often rely on parametric assumptions for the longitudinal sub-model, which can be overly restrictive and subject to model misspecification, in the presence of intricate nonlinear trajectories. To address this, longitudinal outcomes can alternatively be modeled as sparse functional data (James et al. 2000a; Yao et al. 2005), leveraging flexible nonparametric methods (Ramsay and Silverman 2005), to capture complex subject-specific patterns over time. The joint modeling of survival and sparse functional data, termed functional joint models (FJMs), was pioneered by works such as Yao (2007) and Yan et al. (2017). Recent advances have extended FJMs to incorporate multivariate sparse functional data (Li and Luo 2019; Li et al. 2022; Zou et al. 2023; He et al. 2024), multi-dimensional sparse functional data (Shi et al. 2024), recurrent event data (Hong et al. 2021), and imaging data (Zou et al. 2024). Despite successes of FJMs, their model estimation remains a major challenge and many existing works adopt a computationally more feasible two-stage estimation, often at the expense of substantial model bias. This challenge becomes more severe when developing FJM to address systematic missingness and inter-cohort heterogeneity in multi-cohort studies.

Motivated by the challenges, this paper develops an integrative FJM for multivariate sparse functional data and survival data from multiple cohorts, with an accompanied efficient and feasible model estimation. First, to accommodate missing-by-design outcomes while leveraging shared information across cohorts in the longitudinal sub-model, we decompose longitudinal outcomes into shared and outcome-specific disease trajectories as in the multivariate functional mixed model (MFMM) Li et al. (2022), which effectively balances model flexibility and interpretability, thereby enabling comprehensive insights into disease processes. Crucially, we assume consistent variation patterns across cohorts in modeling the shared and outcome-specific disease trajectories, which enables the integration of longitudinal data from diverse sources while maintaining a parsimonious survival model. Indeed, our approach integrates longitudinal and survival data through shared latent trajectories, offering a unified framework that leverages all available information to improve predictive accuracy. Then to address cohort-specific variability, we incorporate cohort-specific mean functions in the longitudinal sub-model and cohort-specific regression coefficients in the survival sub-model. As shall be shown in the application to AD studies, the integrative model is capable of revealing insights that cannot be identified by single-cohort analysis. Those model consideration and assumptions balance leveraging shared information across cohorts with accommodating their inherent heterogeneity.

To estimate the proposed FJM, we develop an efficient and also computationally feasible expectation-maximization (EM) algorithm, another major contribution of the paper. FJMs often have multiple nonparametric smooth functions (more than 10 in our model) to estimate and their estimation makes EM algorithms computationally very challenging. Indeed, regression splines was used earlier Li et al. (2022), for which overfitting may not only affect estimates but also result in slow convergence of EM. Notice that overfitting is almost inevitable in joint models due to the right truncation of longitudinal outcomes by the survival outcome. Therefore, we propose to use penalized splines (Eilers and Marx 1996b) to estimate the non-parametric smooth functions in the longitudinal sub-model, allowing for the modeling of complex nonlinear relationships while reducing overfitting risks associated with regression splines. Penalized splines also allow for varying smoothness across functions, accommodating diverse data structures and trajectories within and across cohorts. For example, Figure 2.3 in the data application shows that penalized splines provide not only highly nonlinear estimates but also essentially linear estimates. However, selecting smoothing parameters for multiple nonparametric functions presents great computational challenges, particularly in iterative EM algorithms. A fast algorithm for the global selection of smoothing parameters in nonparametric regression such as in generalized additive models Wood et al. (2016) is inapplicable for FJMs because it can only deal with regression functions while for FJMs, nonparametric smooth functions (eigenfunctions) are also used to model variances in functional data models. To address this, we propose to use a local selection of smoothing parameters, exploiting the iterative nature of EM algorithms. Specifically, at each iterative step of EM, we reformulate the estimation of each smooth function as a weighted least squares problem for nonparametric regression. This reformulation enables efficient simultaneous estimation of smooth functions and smoothing parameters using standard software, such as the well-developed *mgcv* R package. This approach stabilizes parameter estimation and substantially accelerates EM convergence, making it feasible for multi-cohort studies such as AD studies.

The remainder of this paper is organized as follows. Section 2.2 introduces the proposed multi-cohort functional joint model. Section 2.3 details the model estimation using the Monte Carlo Expectation-Maximization (EM) algorithm. Section 2.4 outlines the selection of key model parameters, such as principal components and smoothing parameters. In Section 2.5, we apply the model to three AD cohorts, demonstrating its ability to capture shared and cohort-specific disease progression patterns. Section 2.6 evaluates the model's performance through simulation studies, and Section 2.7 concludes with a discussion. The R code for implementing the proposed method is available at https://github.com/wenyiwang2000/Multi-Cohort-FJM.

2.2 The Multi-cohort Functional Joint Model

2.2.1 Data Structure and Notation

As shown in Table 2.1, the longitudinal outcomes measured in each cohort differ, with some outcomes observed across all cohorts and others specific to one or two cohorts. Let n be the total number of subjects across all cohorts, indexed i, and let n_c be the number of cohorts. The mapping $c(i)$ indicates the cohort to which subject i belongs. Let J denote the total number of longitudinal outcomes and $\mathscr{J}_c$ the set of outcomes observed in cohort c. For the multi-cohort AD studies and their outcomes listed in Table 2.1, $J = 8$, with $\mathscr{J}_1 = \{1,2,3,4,5,6\}$ in ADNI, $\mathscr{J}_2 = \{1,2,3,7\}$ in NACC, and $\mathscr{J}_3 = \{1,2,8\}$ in ROSMAP. For subject i in cohort $c(i)$ and outcome $j \in \mathscr{J}_{c(i)}$, let Y_{ijk} denote the kth observation of the jth longitudinal outcome at time t_{ijk}, with $1 \le k \le m_{ij}$ and m_{ij} the number of observations. Observational times t_{ijk} are assumed to lie within a compact interval $\mathscr{T}$, representing the study follow-up period.

The survival outcome, representing the time from baseline to AD dementia onset, is denoted by S_i for subject i. When S_i is right-censored, we observe $T_i = \min(S_i, C_i)$, where C_i is the censoring time, assumed independent of both the event time S_i and the longitudinal outcomes. The binary event indicator $\Delta_i = 1_{\{S_i \le C_i\}}$ specifies whether S_i is observed. Longitudinal observations t_{ijk} are restricted to $t_{ijk} \in [0, T_i] \subset \mathscr{T}$, implying no observations of longitudinal outcomes after T_i. Finally, let $\boldsymbol{z}_i = (Z_{i1}, \ldots, Z_{iP}) \in \mathbb{R}^P$ denote a common set of baseline covariates collected across all cohorts, including demographic and clinical characteristics such as age, sex, and APOE genotype, which are relevant to AD progression.

2.2.2 Multivariate Functional Mixed Model for Multi-cohort Longitudinal Data

We model the multiple longitudinal outcomes as multivariate functional data and extend the multivariate functional mixed model (MFMM) in Li et al. (2022) to multiple cohorts:

$$\begin{aligned} Y_{ijk} &= X_{ij}(t_{ijk}) + \epsilon_{ijk}, \\ X_{ij}(t) &= \mu_j^{c(i)}(t) + \beta_j\{U_i(t) + V_{ij}(t)\}, \quad t \in \mathscr{T}, \end{aligned} \tag{2.1}$$

where $X_{ij}(t)$ is the unobserved but smooth latent stochastic process, $\mu_j^{c(i)}(t)$ is the fixed mean function for the jth outcome in cohort $c(i)$, the random processes $U_i(t)$ capture shared variation patterns and induce correlations among multiple outcomes for the ith subject, $V_{ij}(t)$ is the jth subject- and outcome-specific random deviation from $\mu_j^{c(i)}(t)$, β_j is the outcome-specific scaling parameter for the jth outcome, and ϵ_{ijk} is measurement error with variance σ_j^2. These components jointly capture both shared disease progression and individual variations in AD longitudinal outcomes. It is assumed that $U_i(t)$, $V_{ij}(t)$, and ϵ_{ijk} are mutually independent both within and across subjects.

The shared latent trajectory $U_i(t)$ captures overall AD progression, reflecting variation patterns shared among outcomes. Modeled as a Gaussian process with zero mean and covariance $\mathscr{C}(s,t) = \text{Cov}\{U_i(s), U_i(t)\}$, it quantifies the overall trajectory of disease progression. Outcome-specific variations are captured by $V_{ij}(t)$, another Gaussian process with covariance $\mathscr{G}(s,t) = \text{Cov}\{V_{ij}(s), V_{ij}(t)\}$. The scaling coefficients β_js adjust for differences in outcome magnitudes, ensuring comparability. Identifiability is ensured by fixing $\beta_1 = 1$ for a sentinel outcome, such as MMSE or WMS. Note that different scaling parameters might be used for $U_i(t)$ and $V_{ij}(t)$, but we found such flexibility is unnecessary for our data application.

The covariance function of the shared latent trajectory $U_i(t)$ is decomposed using eigen-decomposition as $\mathscr{C}(s,t) = \sum_{d\geq 1} \lambda_{1d}\phi_d(s)\phi_d(t)$, where $\lambda_{11} \geq \lambda_{12} \geq \ldots \geq 0$ are the eigenvalues, and $\phi_d(t)$ are the associated orthonormal eigenfunctions, satisfying $\int_{\mathscr{T}} \phi_d(t)\phi_{d'}(t)dt = 1_{\{d=d'\}}$. Similarly, the covariance of outcome-specific process $V_{ij}(t)$ is decomposed as $\mathscr{G}(s,t) = \sum_{d\geq 1} \lambda_{2d}\psi_d(s)\psi_d(t)$, where $\lambda_{21} \geq \lambda_{22} \geq \ldots \geq 0$ are the eigenvalues and $\psi_d(t)$ are orthonormal eigenfunctions satisfying $\int_{\mathscr{T}} \psi_d(t)\psi_{d'}(t)dt = 1_{\{d=d\}}$.

To facilitate estimation and interpretation, we assume that both $U_i(t)$ and $V_{ij}(t)$ can be represented by a finite number of principal components. Specifically, $U_i(t) = \sum_{d=1}^{D_1} \xi_{id}\phi_d(t)$ with FPCA scores $\xi_{id} \sim N(0, \lambda_{1d})$, and $V_{ij}(t) = \sum_{d=1}^{D_2} \zeta_{ijd}\psi_d(t)$ with FPCA scores $\zeta_{ijd} \sim N(0, \lambda_{2d})$. Here, the parameters D_1 and D_2 represent the number of principal components retained for shared and outcome-specific processes, respectively, and are selected adaptively using data-driven criteria such as cross-validation or AIC/BIC. Finally, the measurement errors ϵ_{ijk} are assumed to be independently distributed as $N(0, \sigma_j^2)$. Given the eigenfunctions $\phi_d(t)$ and $\psi_d(t)$, Model (2.1) can be expressed in the mixed model representation:

$$Y_{ijk} = \mu_j^{c(i)}(t_{ijk}) + \beta_j \Big\{ \sum_{d_1=1}^{D_1} \xi_{id_1}\phi_{d_1}(t_{ijk}) + \sum_{d_2=1}^{D_2} \zeta_{ijd_2}\psi_{d_2}(t_{ijk}) \Big\} + \epsilon_{ijk}.$$

We make a few remarks on Model (2.1). First, the heterogeneity across cohorts is modeled via the outcome- and cohort-specific fixed mean functions, ensuring flexibility in modeling cohort-specific baseline differences. Second, the proposed model integrates the cohorts by imposing shared variation patterns across outcomes, enabling the scores ξ_{id_1} and ζ_{ijd_2} to be directly comparable among subjects across cohorts. This comparability facilitates a parsimonious survival model that integrates information across cohorts, even when different sets of longitudinal outcomes are collected by each cohort. By leveraging shared latent trajectories and cohort-specific mean functions, the MFMM naturally accommodates missing data by design, allowing outcomes observed in only a subset of cohorts to contribute to the model without loss of generalizability. Indeed, Proposition 1 below shows that the MFMM is identifiable under loose conditions. Figure 2.1 illustrates this decomposition by plotting the estimated components for two outcomes, MMSE and CDR-SB, for one subject in the ADNI study and another subject in the NACC study, highlighting the model's ability to separate shared and outcome-specific variations.

Proposition 1. *Under model* (2.1), *suppose that* $\int_{\mathcal{T}} \mathscr{C}(t,t)dt > 0$. *Assume that* $\beta_1 = 1$ *and* $1 \in \cap_{c=1}^{n_c} \mathscr{J}_c$. *Then, Model* (2.1) *is identifiable. Specifically, the model parameters* $\{\mu_j^c(t), j \in \mathscr{J}_c, 1 \le c \le n_c\}, \{\beta_j\}, \mathscr{C}(s,t), \mathscr{G}(s,t)$ *and* $\{\sigma_j^2, 1 \le j \le J\}$ *are identifiable.*

The proof of Proposition 1 is similar to that of identifiability for the MFMM in Li et al. (2022) and hence is omitted. The condition $\int_{\mathcal{T}} \mathscr{C}(t,t)dt > 0$ ensures the presence of shared variation among outcomes, while fixing $\beta_1 = 1$ anchors the scaling of the shared latent trajectories. These conditions are easily satisfied in AD studies because all cohorts collect MMSE and WMS outcomes, one of which can serve as the sentinel outcome with $\beta_1 = 1$. Moreover, the conditions in Proposition 1 can be relaxed. For example, it is not a necessary condition to have a single or multiple outcomes common to all cohorts, demonstrating the model's flexibility in accommodating different outcome collection across cohorts.

2.2.3 Joint Model for Longitudinal and Survival Data

Let $h_i(t)$ be the hazard function for the ith subject. Also, let $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iD_1})^T$ be the vector of scores corresponding to the shared latent trajectory among outcomes. We consider the following cohort-specific proportional hazards model,

$$h_i(t) = h_0(t)\exp\left\{\boldsymbol{z}_i^T \boldsymbol{\gamma}_{1c(i)} + \boldsymbol{\xi}_i^T \boldsymbol{\gamma}_{2c(i)}\right\}, \tag{2.2}$$

where $h_0(t)$ is the baseline hazard function, $\boldsymbol{\gamma}_{1c(i)}$ is the coefficient vector for the baseline covariates $\boldsymbol{z}_i$ in cohort $c(i)$, and $\boldsymbol{\gamma}_{2c(i)} = (\gamma_{2c(i),1}, \ldots, \gamma_{2c(i),D_1})^T$ is the coefficient vector associated with the shared latent score $\boldsymbol{\xi}_i$ for cohort $c(i)$. Here, $h_0(t)$ represents the baseline hazard function, which may be specified parametrically or estimated non-parametrically. The cohort-specific coefficients $\boldsymbol{\gamma}_{1c(i)}$ account for baseline differences in covariate effects among cohorts, while $\boldsymbol{\gamma}_{2c(i)}$ captures the influence of the shared latent trajectory on survival outcomes. This hazard model is parsimonious and applicable across cohorts because it incorporates the longitudinal outcomes via the shared random scores $\boldsymbol{\xi}_i$, enabling integration across cohorts without dependence on the specific set of outcomes being collected. By leveraging shared latent scores, the model avoids the complexity of cohort-specific dependency on collected outcomes, reducing dimensionality while maintaining interpretability and statistical power.

While the proportional hazards assumption is used here, the framework can be extended to accommodate time-varying effects or more complex survival models, providing additional flexibility for diverse applications. Moreover, although we employ the random effects model as the linking function between longitudinal outcomes and survival data, other functional forms can also be utilized. For example, derivative forms incorporating the rate of change in the shared latent trajectory, cumulative forms accounting for historical influence, or lag models emphasizing the weighted impact of recent history, could offer alternative approaches Zou et al. (2023). These extensions provide further flexibility to capture nuanced relationships between longitudinal and survival data, enhancing the applicability of the model across various contexts.

2.2.4 Likelihood of Joint Model

Let $\boldsymbol{y}_{ij} = (Y_{ij1}, \cdots, Y_{ijm_{ij}})^T$ be the column vector of observations for the j-th outcome of the i-th subject. Let $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iD_1})^T$ be the vector of FPCA scores corresponding to the shared latent trajectory, with a diagonal covariance matrix $\boldsymbol{\Lambda}_1 = \mathrm{diag}(\lambda_{11}, \ldots, \lambda_{1D_1})$. Similarly, let $\boldsymbol{\zeta}_{ij} = (\zeta_{ij1}, \cdots, \zeta_{ijD_2})^T$ be the vector of FPCA scores corresponding to the outcome-specific trajectory, with a diagonal covariance matrix $\boldsymbol{\Lambda}_2 = \mathrm{diag}(\lambda_{21}, \ldots, \lambda_{2D_2})$. Define $\boldsymbol{\phi}_{d,ij} = \left(\phi_d(t_{ij1}), \ldots, \phi_d(t_{ijm_{ij}})\right)^T$ and $\boldsymbol{\Phi}_{ij} = [\boldsymbol{\phi}_{1,ij}, \cdots, \boldsymbol{\phi}_{D_1,ij}]$ as the eigenfunctions for the shared latent trajectory. Similarly, let $\boldsymbol{\psi}_{d,ij} = \left(\psi_d(t_{ij1}), \ldots, \psi_d(t_{ijm_{ij}})\right)^T$ and $\boldsymbol{\Psi}_{ij} = [\boldsymbol{\psi}_{1,ij}, \cdots, \boldsymbol{\psi}_{D_2,ij}]$ represent the eigenfunctions for the outcome-specific trajectory.

The predicted values for the j-th outcome of the i-th subject are given by $\boldsymbol{x}_{ij} = \beta_j(\boldsymbol{\Phi}_{ij}\boldsymbol{\xi}_i) + \beta_j\boldsymbol{\Psi}_{ij}\boldsymbol{\zeta}_{ij}$. The predicted values $\boldsymbol{x}_{ij}$ provide the functional link between latent FPCA scores and observed longitudinal data, capturing the modeled relationship between shared and outcome-specific trajectories. Concatenate the vectors $\boldsymbol{y}_{ij}$ ($j \in \mathscr{J}_{c(i)}$) into a long column vector $\boldsymbol{y}_i$ and similarly define $\boldsymbol{x}_i$. Let $\boldsymbol{\eta}_i = (\boldsymbol{\xi}_i^T, \boldsymbol{\zeta}_{i1}^T, \ldots, \boldsymbol{\zeta}_{iJ_{c(i)}}^T)^T$ be the collection of all FPCA scores (treated as random effects in the likelihood framework). Finally, denote the covariance matrix of measurement errors by $\boldsymbol{\Sigma}_i = \mathrm{blockdiag}(\sigma_1^2 \boldsymbol{I}_{m_{i1}}, \ldots, \sigma_{J_{c(i)}}^2 \boldsymbol{I}_{m_{iJ_{c(i)}}})$. The conditional likelihood of the multivariate longitudinal data given $\boldsymbol{\eta}_i$ is,

$$f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\Sigma}_i) = (|2\pi\boldsymbol{\Sigma}_i|)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{x}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{x}_i)\right\}.$$

The FPCA scores (random effects) $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_{ij}$ are assumed to follow multivariate normal distributions: $f(\boldsymbol{\xi}_i|\boldsymbol{\Lambda}_1) = \sqrt{|2\pi\boldsymbol{\Lambda}_1|}\exp(-\frac{1}{2}\boldsymbol{\xi}_i^T\boldsymbol{\Lambda}_1^{-1}\boldsymbol{\xi}_i)$, and $f(\boldsymbol{\zeta}_{ij}|\boldsymbol{\Lambda}_2) = \sqrt{|2\pi\boldsymbol{\Lambda}_2|}\exp(-\frac{1}{2}\boldsymbol{\zeta}_{ij}^T\boldsymbol{\Lambda}_2^{-1}\boldsymbol{\zeta}_{ij})$.

The conditional partial likelihood of the time-to-event data is given by

$$
\begin{aligned}
f(T_i, \Delta_i | h_0, \boldsymbol{z}_i, \boldsymbol{\xi}_i, \boldsymbol{\gamma}_{1c(i)}, \boldsymbol{\gamma}_{2c(i)}) = &\left\{ h_0(T_i) \exp(\boldsymbol{z}_i^T \boldsymbol{\gamma}_{1c(i)} + \boldsymbol{\xi}_i^T \boldsymbol{\gamma}_{2c(i)}) \right\}^{\Delta_i} \\
&\times \exp\left\{ -\Lambda_0(T_i) \exp(\boldsymbol{z}_i^T \boldsymbol{\gamma}_{1c(i)} + \boldsymbol{\xi}_i^T \boldsymbol{\gamma}_{2c(i)}) \right\},
\end{aligned}
\tag{2.3}
$$

where $\Lambda_0(t) = \int_0^t h_0(u) du$ is the baseline cumulative hazard function.

The multivariate longitudinal data and the time-to-event data are assumed to be conditionally independent given the random effects $\boldsymbol{\eta}_i$. This assumption reflects the idea that the shared latent trajectory $\boldsymbol{\xi}_i$ and outcome-specific deviations $\boldsymbol{\zeta}_{ij}$ sufficiently capture the dependence structure between the longitudinal and survival components. To account for differences in cohort sizes, a weighted log-likelihood is adopted. The smallest cohort (cohort 1) is designated as the reference cohort, ensuring that variability in sample sizes does not disproportionately influence parameter estimation. The weights for subjects in other cohorts are given by $w_c = \frac{N_1}{N_c}$, where N_c is the total number of longitudinal observations in cohort c.

The overall marginal likelihood is obtained by integrating the product of the longitudinal and survival likelihoods over the distributions of the random effects $\boldsymbol{\eta}_i$. The weighted log marginal likelihood is

$$
\sum_{i=1}^{n} \omega_{c(i)} \log \left\{ \int f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\Sigma}_i) f(T_i, \Delta_i | h_0, \boldsymbol{z}_i, \boldsymbol{\xi}_i, \boldsymbol{\gamma}_{1c(i)}, \boldsymbol{\gamma}_{2c(i)}) f(\boldsymbol{\xi}_i | \boldsymbol{\Lambda}_1) \prod_{j \in \mathscr{J}_{c(i)}} f(\boldsymbol{\zeta}_{ij} | \boldsymbol{\Lambda}_2) d\boldsymbol{\eta}_i \right\}. \tag{2.4}
$$

2.3 Model Estimation via Monte Carlo EM

2.3.1 Spline Approximation of Smooth Functions

Estimating the smooth mean and covariance functions in the MFMM model requires a flexible yet computationally efficient approach. We employ spline approximation, which provides these properties and is particularly suited to complex functional data. Let $\boldsymbol{b}(t) = \{B_1(t), \cdots, B_K(t)\}^T$ be the vector of B-spline basis functions, where K is the number of equally-spaced interior knots plus the order (degree + 1) of the B-splines. The mean function $\mu_j^c(t)$ for the j-th outcome in cohort c is modeled as $\mu_j^c(t) = \boldsymbol{b}(t)^T \boldsymbol{\alpha}_j^c$, where $\boldsymbol{\alpha}_j^c$ is the corresponding coefficient vector.

To ensure numerical stability, we orthonormalize the B-spline bases using the Gram matrix: $\boldsymbol{G} = \int \boldsymbol{b}(t)\boldsymbol{b}(t)^T dt \in \mathbb{R}^{K\times K}$. The covariance functions $\mathscr{C}(s,t)$ (shared latent trajectory) and $\mathscr{G}(s,t)$ (outcome-specific trajectory) are decomposed into their dth eigenfunctions, which are approximated: $\phi_d(t) = \widetilde{\boldsymbol{b}}(t)^T \boldsymbol{\theta}_{1d}$ and $\psi_d(t) = \widetilde{\boldsymbol{b}}(t)^T \boldsymbol{\theta}_{2d}$, where $\boldsymbol{\theta}_{1d}$ and $\boldsymbol{\theta}_{2d}$ are coefficient vectors for the d-th eigenfunctions of $\mathscr{C}$ and $\mathscr{G}$, respectively. The orthonormality of the eigenfunctions imposes the constraints: $\boldsymbol{\theta}_{1d_1}^T \boldsymbol{\theta}_{1d_2} = 1_{\{d_1=d_2\}}$ and $\boldsymbol{\theta}_{2d_1}^T \boldsymbol{\theta}_{2d_2} = 1_{\{d_1=d_2\}}$.

2.3.2 E-Step

The spline approximation of the nonparametric smooth functions enables the use of parametric estimation methods for the MFMM model. The random score vector $\boldsymbol{\eta}_i$, which captures both shared and outcome-specific variations, introduces significant dimensionality to the model. In practice, the dimension of $\boldsymbol{\eta}_i$ often exceeds four, making direct maximization of the log marginal likelihood in Model (2.4) computationally expensive or infeasible. To address this, we employ the EM algorithm (Dempster et al. 1977), which treats $\boldsymbol{\eta}_i$ as latent (missing) data. The EM algorithm iteratively alternates between two steps until convergence: (1) computing the expectation of the conditional log-likelihood of the observed data given the current parameter estimates (E-step), and (2) maximizing the expected log-likelihood to update the parameter estimates (M-step).

Let the observed data for the ith subject be $\mathbb{Y}_i = \{\boldsymbol{y}_i, \boldsymbol{z}_i, T_i, \Delta_i\}$, where $\boldsymbol{y}_i$ represents the collection of longitudinal data. Let $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_j^c : j \in \mathscr{J}_c, 1 \le c \le n_c\}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_J^2)^T$, $\boldsymbol{\Theta}_1 = [\boldsymbol{\theta}_{11}, \cdots, \boldsymbol{\theta}_{1D_1}]$ and $\boldsymbol{\Theta}_2 = [\boldsymbol{\theta}_{21}, \cdots, \boldsymbol{\theta}_{2D_2}]$. Let $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_{1c}, \boldsymbol{\gamma}_{2c} : 1 \le c \le n_c\}$. Denote the complete collection of parameters by $\boldsymbol{\Omega} = \{h_0, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\gamma}, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$. The E-step calculates the expected value of the weighted conditional log-likelihood given the observed data and current parameter estimates, $\widehat{\boldsymbol{\Omega}}$. This is expressed as:

$$
\begin{aligned}
Q(\boldsymbol{\Omega}) = &\sum_{i=1}^{n} \omega_{c(i)} \mathbb{E}_i \Big\{ \log\big(f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\Sigma}_i) \big) \Big\} + \sum_{i=1}^{n} \omega_{c(i)} \mathbb{E}_i \Big\{ \log\big(f(T_i, \Delta_i | h_0, \boldsymbol{z}_i, \boldsymbol{\xi}_i, \boldsymbol{\gamma}) \big) \Big\} \\
&+ \sum_{i=1}^{n} \omega_{c(i)} \mathbb{E}_i \Big\{ \log\big(f(\boldsymbol{\xi}_i | \boldsymbol{\Lambda}_1) \big) \Big\} + \sum_{i=1}^{n} \omega_{c(i)} \sum_{j \in \mathscr{J}_{c(i)}} \mathbb{E}_i \Big\{ \log\big(f(\boldsymbol{\zeta}_{ij} | \boldsymbol{\Lambda}_2) \big) \Big\},
\end{aligned} \tag{2.5}
$$

where $\mathbb{E}_i(\cdot)$ is the expectation with respect to the conditional distribution of $\boldsymbol{\eta}_i$ given the observed data $\mathbb{Y}_i$ and current parameter estimates $\widehat{\boldsymbol{\Omega}}$.

Let $g(\cdot)$ be any smooth function of $\boldsymbol{\eta}_i$. The conditional expectation $\mathbb{E}_i\{g(\boldsymbol{\eta}_i)\}$ is given by:

$$
\frac{\int g(\boldsymbol{\eta}_i) f(T_i, \Delta_i | \boldsymbol{z}_i, \boldsymbol{\eta}_i, \widehat{\boldsymbol{\Omega}}) f(\boldsymbol{\eta}_i | \boldsymbol{y}_i, \widehat{\boldsymbol{\Omega}}) d\boldsymbol{\eta}_i}{\int f(T_i, \Delta_i | \boldsymbol{z}_i, \boldsymbol{\eta}_i, \widehat{\boldsymbol{\Omega}}) f(\boldsymbol{\eta}_i | \boldsymbol{y}_i, \widehat{\boldsymbol{\Omega}}) d\boldsymbol{\eta}_i},
$$

where $f(T_i, \Delta_i | \boldsymbol{z}_i, \boldsymbol{\eta}_i, \widehat{\boldsymbol{\Omega}})$ corresponds to the survival likelihood defined in Model (2.3), and

$$
f(\boldsymbol{\eta}_i | \boldsymbol{y}_i, \widehat{\boldsymbol{\Omega}}) \propto f(\boldsymbol{y}_i | \boldsymbol{x}_i, \widehat{\boldsymbol{\Omega}}) f(\boldsymbol{\xi}_i | \widehat{\boldsymbol{\Lambda}}_1) \prod_{j \in \mathscr{J}_{c(i)}} f(\boldsymbol{\zeta}_{ij} | \widehat{\boldsymbol{\Lambda}}_2).
$$

The conditional distribution $f(\boldsymbol{\eta}_i | \boldsymbol{y}_i, \widehat{\boldsymbol{\Omega}})$ is multivariate normal (see Appendix A.1).

To efficiently compute $\mathbb{E}_i\{g(\boldsymbol{\eta}_i)\}$, we use Monte Carlo integration. Specifically, we draw R

samples $\boldsymbol{\eta}_i^{(r)}$ from the multivariate normal distribution $f(\boldsymbol{\eta}_i|\boldsymbol{y}_i,\widehat{\boldsymbol{\Omega}})$ and compute the approximation:

$$\mathbb{E}_i\{g(\boldsymbol{\eta}_i)\} \approx \frac{\sum_{r=1}^R g(\boldsymbol{\eta}_i^{(r)}) f(T_i,\Delta_i|\boldsymbol{z}_i,\boldsymbol{\eta}_i^{(r)},\widehat{\boldsymbol{\Omega}})}{\sum_{r=1}^R f(T_i,\Delta_i|\boldsymbol{z}_i,\boldsymbol{\eta}_i^{(r)},\widehat{\boldsymbol{\Omega}})}.$$

The Monte Carlo approximation enables efficient implementation of the E-step, ensuring computational feasibility even in high-dimensional parameter spaces.

2.3.3 M-Step

In the M-step, updated estimates of $\boldsymbol{\Omega}$ are obtained by maximizing the function $Q(\boldsymbol{\Omega})$ in (2.5). Specifically, estimates of the longitudinal parameters $\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2\}$, the survival parameters $\{h_0,\boldsymbol{\gamma}\}$, and the eigenvalue parameters $\{\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2\}$ are updated iteratively. Below, we provide details on the estimation process for each set of parameters.

First, the longitudinal parameters $\{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\Theta}_1,\boldsymbol{\Theta}_2\}$ are estimated iteratively by minimizing $-\sum_i^n \omega_{c(i)}\mathbb{E}_i\{\log(f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Sigma}_i))\}$, expressed as:

$$\sum_{i=1}^n \omega_{c(i)} \sum_{j\in\mathscr{J}_{c(i)}} \left\{\frac{m_{ij}}{2}\log(2\pi\sigma_j^2) + \frac{1}{2\sigma_j^2}\mathbb{E}_i\left\|\boldsymbol{y}_{ij} - \boldsymbol{B}_{ij}\boldsymbol{\alpha}_j^{c(i)} - \left(\beta_j\widetilde{\boldsymbol{B}}_{ij}\boldsymbol{\Theta}_1\boldsymbol{\xi}_i + \beta_j\widetilde{\boldsymbol{B}}_{ij}\boldsymbol{\Theta}_2\boldsymbol{\zeta}_{ij}\right)\right\|^2\right\}, \quad (2.6)$$

where $\boldsymbol{B}_{ij} = [\boldsymbol{b}(t_{ij1}),\ldots,\boldsymbol{b}(t_{ijm_{ij}})]^T$ is the matrix of B-spline basis values evaluated at time points $t_{ij1},\ldots,t_{ijm_{ij}}$, and $\widetilde{\boldsymbol{B}}_{ij} = \boldsymbol{G}^{-1/2}\boldsymbol{B}_{ij}$ is the orthonormalized basis obtained by multiplying $\boldsymbol{B}_{ij}$ by the inverse square root of the Gram matrix $\boldsymbol{G}$. The outcome-specific scaling parameter β_j is obtained though weighted least squares of linear regression, as detailed in Appendix A.3. To enforce smoothness in the estimates of the mean functions and eigenfunctions, quadratic penalties are added to (2.6). Compared to the regression spline estimation used in Li et al. (2022), penalized spline estimation stabilizes the iterative updates, reduces the number of iterations required for EM convergence, and produces smoother, more interpretable estimates. Empirical and theoretical studies of penalized splines show that the number of knots does not matter as long as a relatively large number of knots are used (Ruppert 2002; Xiao 2019). In this study, seven spline bases constructed at equally spaced knots are employed for both simulations and real data. The orthonormality of eigenfunctions is also ensured by post-processing the updates.

Second, the baseline hazard function $h_0(t)$ in the Cox regression is estimated by minimizing $-\sum_i^n \omega_{c(i)}\mathbb{E}_i\{\log(f(T_i,\Delta_i|h_0,\boldsymbol{z}_i,\boldsymbol{\xi}_i,\boldsymbol{\gamma}_{1c(i)},\boldsymbol{\gamma}_{2c(i)}))\}$, expressed as:

$$\sum_{i=1}^n \omega_{c(i)}\left\{-\Delta_i\left\{\log(h_0(T_i)) + \boldsymbol{z}_i^T\boldsymbol{\gamma}_{1c(i)} + \mathbb{E}_i\left(\boldsymbol{\xi}_i^T\boldsymbol{\gamma}_{2c(i)}\right)\right\} + \Lambda_0(T_i)\mathbb{E}_i\left\{\exp\left(\boldsymbol{z}_i^T\boldsymbol{\gamma}_{1c(i)} + \boldsymbol{\xi}_i^T\boldsymbol{\gamma}_{2c(i)}\right)\right\}\right\}.$$

The baseline hazard $h_0(t)$ is then estimated by using a weighted ratio of observed events to expected contributions at each time t:

$$\widehat{h}_0(t) = \frac{\sum_{i=1}^n \omega_{c(i)} \Delta_i 1_{\{T_i = t\}}}{\sum_{i=1}^n \omega_{c(i)} \mathbb{E}_i \left\{ \exp\left(\boldsymbol{z}_i^T \widehat{\boldsymbol{\gamma}}_{1c(i)} + \boldsymbol{\xi}_i^T \widehat{\boldsymbol{\gamma}}_{2c(i)} \right) \right\} 1_{\{T_i \geq t\}}}.$$

The parameter vector $\boldsymbol{\gamma}^c = (\boldsymbol{\gamma}_{1c}^T, \boldsymbol{\gamma}_{2c}^T)^T$, representing the effects of baseline covariates and latent trajectories, is updated iteratively using the Newton-Raphson algorithm. The ℓ-th iteration is given by: $\widehat{\boldsymbol{\gamma}}_\ell^c = \widehat{\boldsymbol{\gamma}}_{\ell-1}^c + \boldsymbol{I}_{\widehat{\boldsymbol{\gamma}}_{\ell-1}^c}^{-1} \boldsymbol{S}_{\widehat{\boldsymbol{\gamma}}_{\ell-1}^c}$, where $\boldsymbol{S}_{\widehat{\boldsymbol{\gamma}}_{\ell-1}^c}$ and $\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}}_{\ell-1}^c}$ are the score vector and observed information matrix, respectively. By differentiating the survival likelihood (2.3) with respect to $\boldsymbol{\gamma}^{c(i)}$, the score for subject i is:

$$\boldsymbol{s}_i(\boldsymbol{\gamma}^{c(i)}) = \Delta_i (\boldsymbol{z}_i^T, \mathbb{E}_i \boldsymbol{\xi}_i^T)^T - \sum_{v=1}^V h_0(t_v) \mathbb{E}_i \left\{ (\boldsymbol{z}_i^T, \boldsymbol{\xi}_i^T)^T \exp\left(\boldsymbol{z}_i^T \boldsymbol{\gamma}_{1c(i)} + \boldsymbol{\xi}_i^T \boldsymbol{\gamma}_{2c(i)} \right) \right\} 1_{\{T_i \geq t_v\}},$$

where t_v are the distinct observed event times across all cohorts. The cohort-level score $\boldsymbol{S}_{\boldsymbol{\gamma}^c}$ and information matrix $\boldsymbol{I}_{\boldsymbol{\gamma}^c}$ are computed as: $\boldsymbol{S}_{\boldsymbol{\gamma}^c} = \sum_{i \in \mathscr{I}_c} \boldsymbol{s}_i(\boldsymbol{\gamma}^c)$ and $\boldsymbol{I}_{\boldsymbol{\gamma}^c} = \sum_{i \in \mathscr{I}_c} \boldsymbol{s}_i(\boldsymbol{\gamma}^c) \boldsymbol{s}_i^T(\boldsymbol{\gamma}^c) - \boldsymbol{S}_{\boldsymbol{\gamma}^c} \boldsymbol{S}_{\boldsymbol{\gamma}^c}^T / |\mathscr{I}_c|$, where $|\mathscr{I}_c|$ is the number of subjects in cohort c.

Third, the diagonal matrices of eigenvalues $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are estimated by minimizing the weighted negative log-likelihoods $-\sum_{i=1}^n \omega_{c(i)} \mathbb{E}_i \{\log(f(\boldsymbol{\xi}_i | \boldsymbol{\Lambda}_1))\}$ and $-\sum_{i=1}^n \omega_{c(i)} \mathbb{E}_i \{\log(f(\boldsymbol{\zeta}_{ij} | \boldsymbol{\Lambda}_2))\}$, respectively. For $\boldsymbol{\Lambda}_1$, this minimization simplifies to: $\sum_i \left\{ \log|\boldsymbol{\Lambda}_1| + \mathbb{E}_i \left\{ \boldsymbol{\xi}_i^T \boldsymbol{\Lambda}_1^{-1} \boldsymbol{\xi}_i \right\} \right\}$. The estimator for the eigenvalues is: $\widehat{\lambda}_{1d} = \frac{1}{\sum_{i=1}^n \omega_{c(i)}} \sum_{i=1}^n \omega_{c(i)} \mathbb{E}_i(\xi_{id}^2)$, where $\mathbb{E}_i(\xi_{id}^2)$ represents the expected squared random scores for the shared latent trajectory. Similarly, for $\boldsymbol{\Lambda}_2$, minimizing the negative log-likelihood leads to the estimator: $\widehat{\lambda}_{2d} = \frac{1}{\sum_{i=1}^n \omega_{c(i)}} \sum_{i=1}^n \omega_{c(i)} \frac{1}{J_{c(i)}} \sum_{j \in \mathscr{I}_{c(i)}} \mathbb{E}_i(\zeta_{ijd}^2)$, where $\mathbb{E}_i(\zeta_{ijd}^2)$ represents the expected squared random scores for the outcome-specific deviations. The inclusion of cohort-specific weights $\omega_{c(i)}$ ensures that the estimates appropriately account for differences in cohort sizes and data distributions. This weighting enhances model flexibility and accuracy across diverse cohort distributions.

2.4 Model Selection

Penalized splines (Eilers and Marx 1996b) are used to estimate the mean functions and eigenfunctions in the longitudinal data model, requiring the selection of appropriate smoothing parameters. These parameters are determined using generalized cross-validation (GCV) at each iteration of the EM algorithm. By reformulating function estimation as a weighted least squares problem in nonparametric regression, we enable efficient implementation using the *gam* function from the R package *mgcv* (Wood 2011; Wood et al. 2016) (see Appendix A.3 for details). Although the smoothing parameters may vary across iterations, they stabilize rapidly as the algorithm converges, ensuring consistent estimation of smooth functions and improving computational efficiency.

The number of principal components D_1 for the shared latent process and D_2 for the outcome-specific latent process are critical tuning parameters in the model. To select these parameters, we use the Bayesian information criteria (BIC), defined as $\mathrm{BIC} = -2\ell(\widehat{\boldsymbol{\Omega}}) + \left(\sum_{c=1}^{n_c} \omega_c \log(N_c) \cdot \mathrm{df_c}\right)$, where $\ell(\widehat{\boldsymbol{\Omega}})$ is the weighted marginal log-likelihood of the data, approximated as:

$$\sum_{i=1}^{n} \omega_{c(i)} \left[\log\left\{ f(\boldsymbol{y}_i | \widehat{\boldsymbol{\Omega}}) \right\} + \log\left\{ R^{-1} \sum_{r=1}^{R} f\left(T_i, \Delta_i \mid \widehat{h}_0, \boldsymbol{z}_i, \xi_i^{(r)}, \widehat{\boldsymbol{\gamma}}_{c(i)} \right) \right\} \right].$$

Here, $f(\boldsymbol{y}_i|\widehat{\boldsymbol{\Omega}})$ is the marginal density, which follows a normal distribution (see Appendix A.1). The samples $\xi_i^{(r)}$ are drawn from $f(\boldsymbol{\eta}_i \mid \boldsymbol{y}_i, \widehat{\boldsymbol{\Omega}})$; see Appendix A.2 for derivation. To ensure an accurate approximation of the marginal log-likelihood, the number of samples R must be much larger than in the EM algorithm.

The degrees of freedom (DOF) for cohort c, denoted as $\mathrm{df_c}$, are calculated as:

$$\sum_{j \in \mathcal{J}_c} \mathrm{df}_{\mu_j^c} + \frac{1}{n_c} \left\{ \sum_{d=1}^{D_1} \mathrm{df}_{\phi_d} - \frac{D_1(D_1-1)}{2} \right\} + \frac{1}{n_c} \left\{ \sum_{d=1}^{D_2} \mathrm{df}_{\psi_d} - \frac{D_2(D_2-1)}{2} \right\} + \frac{(2J-1)}{n_c} + (p + D_1).$$

Each term in this formula corresponds to a specific model component:

- $\sum_{j \in \mathcal{J}_c} \mathrm{df}_{\mu_j^c}$: DOF for estimating the mean functions in cohort c.
- $\sum_{d=1}^{D_1} \mathrm{df}_{\phi_d} - D_1(D_1-1)/2$: DOF for the shared covariance function's eigenpairs.
- $\sum_{d=1}^{D_2} \mathrm{df}_{\psi_d} - D_2(D_2-1)/2$: DOF for the outcome-specific covariance function's eigenpairs.
- $(2J-1)/n_c$: DOF for estimating error variances and scaling coefficient vector $\boldsymbol{\beta}$.
- p: Number of baseline covariates in the Cox regression.
- D_1: Number of cohort-specific coefficients for shared latent scores in the Cox regression.

Cohort c's DOF has a division factor n_c to appropriately reflect cohort-level contributions when parameters are shared across cohorts. A two-dimensional grid is employed to identify the optimal values of D_1 and D_2, balancing model complexity with goodness-of-fit.

2.5 Data Application

We apply the proposed multi-cohort functional joint model (FJM) to data from three AD cohorts, integrating longitudinal outcomes via the extended MFMM and linking them to survival outcomes through a Cox regression model. This application jointly characterize the progression of multivariate longitudinal outcomes and their association with time to AD dementia diagnosis. By leveraging cohort-specific variations while incorporating shared latent processes, the multi-cohort FJM allows for a unified yet flexible representation of disease trajectories across diverse datasets. Following Li et al. (2022), we include the following baseline covariates: age, sex, years of education, and the number of apolipoprotein E $\epsilon 4$ alleles (APOE4), in the hazard model, reflecting known demographic and genetic risk factors for AD progression (see Table A.1 in Appendix A.4 for details). To assess performance, we compare the multi-cohort FJM with the single-cohort FJM-MFMM Li et al. (2022) and the parametric multivariate joint model (MJM) in Henderson et al. (2000) (MJM), each applied separately to the cohorts.

Using the BIC criteria described in Section 2.4, we selected four principal components for the shared covariance structure ($D_1 = 4$) and three for the outcome-specific covariance structure ($D_2 = 3$). The estimated scaling coefficients $\boldsymbol{\beta}$, displayed in Table 2.1, with signs aligning with expectation: lower values of MMSE, WMSLM, RAVLT, and SDMT and higher values of CDR-SB, ADAS, FAQ, and TRAILA indicate AD progression. Figure 2.2 illustrates the first two eigenfunctions estimated using the multi-cohort FJM and separate FJM models. Consistent trends in eigenfunctions across models support the utility of the multi-cohort FJM in capturing subject-specific random trajectories. For instance, the first eigenfunction of the shared latent process ($\phi_1(t)$) is negative throughout, and its loading coefficients in $\boldsymbol{\beta}$ (Table 2.1) suggest that positive scores for the shared trajectory correspond worsening outcomes (lower MMSE, WMSLM, RAVLT, and SDMT values, and higher CDR-SB, ADAS, FAQ, and TRAILA values). This pattern indicates accelerating AD progression as time progresses. A similar interpretation holds for the first outcome-specific eigenfunction ($\psi_1(t)$), highlighting its role in capturing deviations specific to individual outcomes.

Figure 2.3 compares the estimated mean trajectories of longitudinal outcomes across three models: the multi-cohort FJM, the single-cohort FJM, and the MJM. Columns 1-3 display results for the FJM and multi-cohort FJM, while columns two and four depict results for the MJM. Unlike the MJM, which assumes linear trends, both FJM and multi-cohort FJM effectively capture nonlinear trajectories. Three important observations emerge. First, while the FJM and multi-cohort FJM produce similar mean functions, the multi-cohort FJM captures steeper declines in WMSLM and RAVLT and sharper increases in TRAILA, reflecting more rapid deterioration in some AD-related measures. Second, the flexibility of penalized splines in the FJM models ennables them to capture nonlinear trends in some outcomes (e.g., WMSLM, RAVLT, and TRAILA), which indicate accelerating deterioration in AD-related measures. In contrast, the linear MJM fails to capture such dynamics, underscoring its limitations in modeling complex trends in AD studies. Third, clear cohort-specific differences in progression are evident. For example, patients in the ROSMAP cohort demonstrate faster progression, as reflected in lower MMSE and WMSLM scores, compared to ADNI and NACC, where ADNI patients show the slowest progression. These differences highlight the importance of multi-cohort modeling in uncovering population-specific patterns.

Table 2.2 presents the estimated Cox regression coefficients from the multi-cohort FJM. APOE4 consistently emerges as a significant risk factor across all cohorts, while sex shows no significant association. Age is significant in ADNI and NACC but not in ROSMAP, and years of education have a significant protective effect in NACC and ROSMAP but not in ADNI, potentially reflecting differences in population characteristics. The shared progression scores $(\xi_{i1}, \xi_{i2}, \xi_{i3}, \xi_{i4})$ are critical predictors of AD risk. Most scores show significant associations with time to AD diagnosis, emphasizing the utility of the shared latent trajectory in capturing meaningful variability in disease progression. For instance, positive values for ξ_{i1} are associated with higher AD risk, consistent with the interpretation of the shared trajectory reflecting overall deterioration. These findings underscore the importance of incorporating longitudinal outcomes into survival models to better understand the progression and prediction of AD.

In the proposed multi-cohorts model, distinct mean functions for longitudinal outcomes and cohort-specific Cox regression coefficients are employed to account for inter-cohort variability. To investigate the impact of simplifying assumptions, we evaluate alternative models with either shared mean functions or shared Cox coefficients across cohorts. Using AIC/BIC for model selection, we assess the trade-offs between model flexibility and parsimony. When both the mean functions and Cox coefficients are shared across cohorts, the degreess of freedom $\mathrm{df_c}$ for cohort c are defined as:

$$\sum_{j\in\mathscr{J}_c}\frac{1}{n_{y_j}}\mathrm{df}_{\mu_j^c}+\frac{1}{n_c}\left\{\sum_{d=1}^{D_1}\mathrm{df}_{\phi_d}-\frac{D_1(D_1-1)}{2}\right\}+\frac{1}{n_c}\left\{\sum_{d=1}^{D_2}\mathrm{df}_{\psi_d}-\frac{D_2(D_2-1)}{2}\right\}+\frac{(2J-1)}{n_c}+\frac{(p+D_1)}{n_c},$$

where n_{y_j} represents the number of cohorts in which the j-th outcome is collected.

For each model, we apply the multi-cohort FJM with different values of the tuning parameter D_0 and D_1 (D_0, $D_1 \in \{1,2,3,4,5\}$) to the three AD cohorts; see Appendix A.4 for details. The results show that cohort-specific Cox coefficients are consistently required, as determined by both AIC and BIC. However, the two metrics diverge regarding mean functions: BIC favors shared mean functions across cohorts, reflecting its preference for simpler models, while AIC supports cohort-specific mean functions, prioritizing flexibility. This divergence likely arises from the limited overlap of outcomes across cohorts, with only three of eight outcomes shared by at least two cohorts. Allowing cohort-specific mean functions in such a scenario does not significantly increase model flexibility. Results for the simpler model selected by BIC are provided in Appendix A.4 and align with the interpretations of the full model. These findings underscore the critical role of cohort-specific Cox coefficients in capturing inter-cohort heterogeneity, while the choice to use cohort-specific mean functions depends on the extent of outcome overlap across cohorts.

Table 2.2: Estimated Cox regression coefficients (with standard errors) from the multi-cohort functional joint model, evaluating the association between baseline covariates, shared latent scores, and time to AD diagnosis. Covariates include baseline age, sex, years of education, and APOE4 status. Shared latent scores reflect subject-specific progression in longitudinal outcomes. An asterisks indicates significance at the 0.05 level.

	ADNI		NACC		ROSMAP	
Covariates	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value
Age	0.02 (0.01)*	0.01	0.01 (0.00)*	0.00	0.00 (0.01)	0.96
sex (female)	0.10 (0.17)	0.55	0.09 (0.08)	0.21	-0.31 (0.24)	0.20
Education	0.05 (0.03)	0.08	0.09 (0.01)*	0.00	0.11 (0.03)*	0.00
APOE4	0.28 (0.13)*	0.03	0.37 (0.06)*	0.00	0.40 (0.19)*	0.03
Shared score (ξ_{i1})	0.63 (0.04)*	0.00	0.59 (0.02)*	0.00	0.81 (0.06)*	0.00
Shared score (ξ_{i2})	0.68 (0.22)*	0.00	0.48 (0.07)*	0.00	0.10 (0.30)	0.74
Shared score (ξ_{i3})	-1.23 (0.26)*	0.00	-1.71 (0.12)*	0.00	-0.66 (0.39)	0.09
Shared score (ξ_{i4})	2.86 (0.68)*	0.00	3.90 (0.31)*	0.00	-3.88 (0.71)*	0.00

2.6 Simulation Study

2.6.1 Simulation Settings

We evaluate the performance of the multi-cohort FJM method by generating data similar to the AD cohorts. The longitudinal data of 3 cohorts are generated according to Model (2.1) with $J = 8$ outcomes, following the structure of AD cohorts. Each cohort includes the same set of outcomes as in the real data, with cohort-specific mean functions, scaling parameters, and error variances derived from the estimates obtained in the application study. To simplify the analysis, the multi-cohort FJM is fitted with two principal components for both the shared covariance function and the outcome-specific covariance functions. The eigenfunctions used

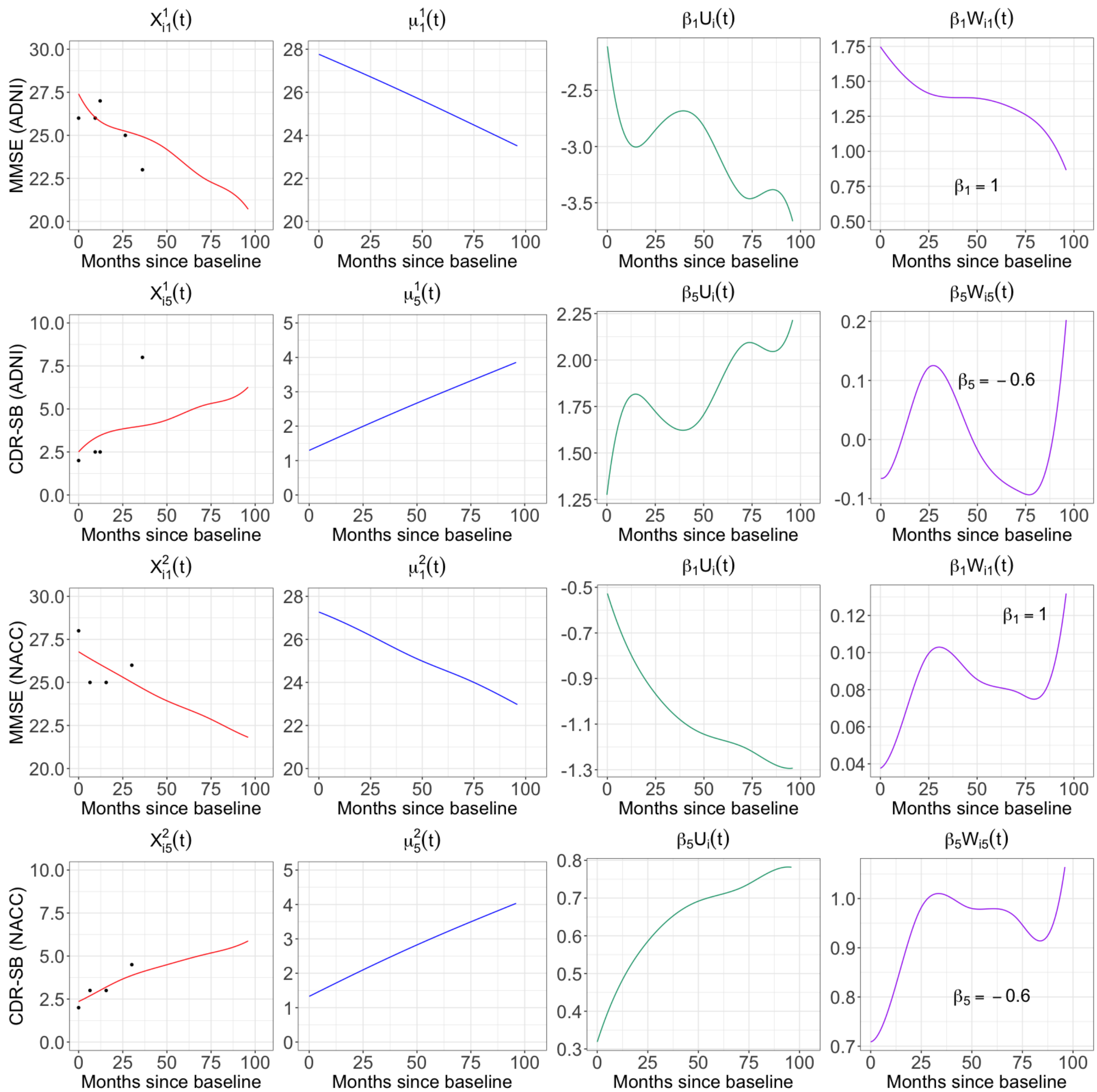


Figure 2.1: Estimated components of the multi-cohort MFMM in Model (2.1) for two outcomes: MMSE and CDR-SB for one subject from the ADNI study (top two rows) and one subject from the NACC study (bottom two rows). Columns represent the following: (1) observed MMSE/CDR-SB values (black dots) and estimated latent processes $X_{ij}(t)$; (2) cohort-specific mean functions $\mu_j^{c(i)}(t)$; (3) shared latent disease trajectory $U_i(t)$ scaled by coefficient β_j; (4)outcome-specific deviations $W_{ij}(t)$ scaled by β_j.

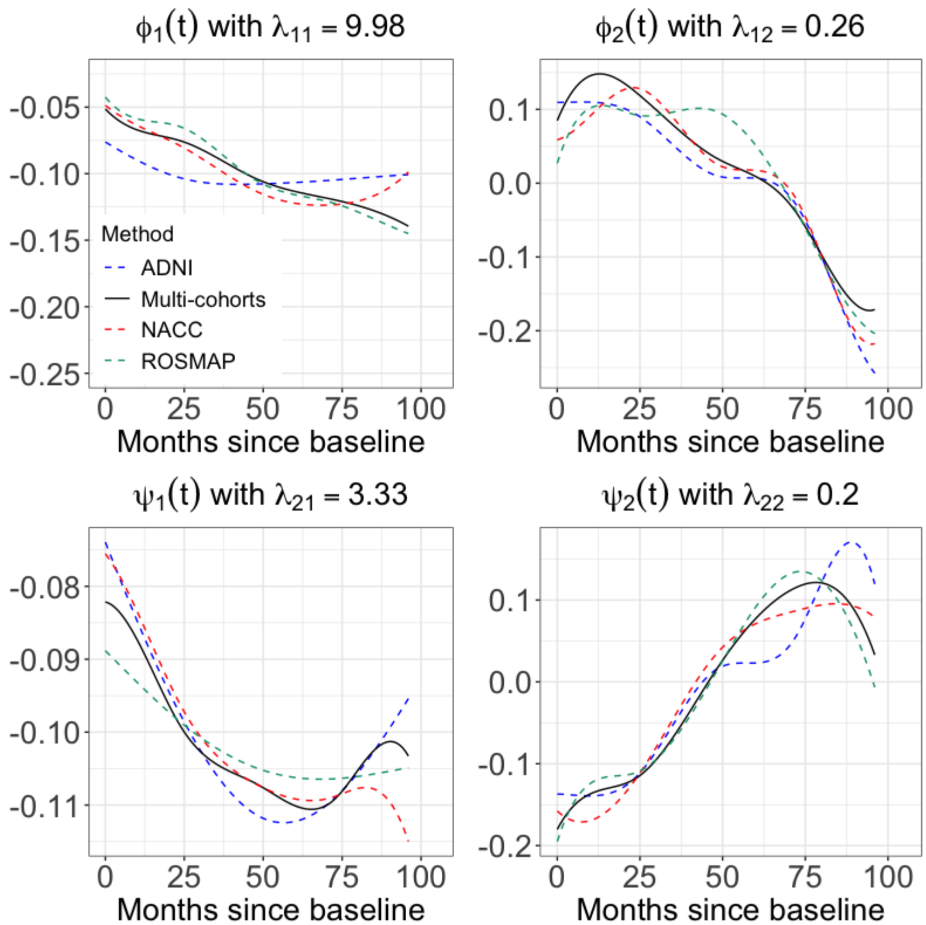


Figure 2.2: Top two estimated eigenfunctions with associated eigenvalues for the shared and outcome-specific latent processes, comparing multi-cohort FJM (solid lines) and separate FJM (dashed lines). These eigenfunctions characterize the primary patterns of variability in the longitudinal outcomes across cohorts.

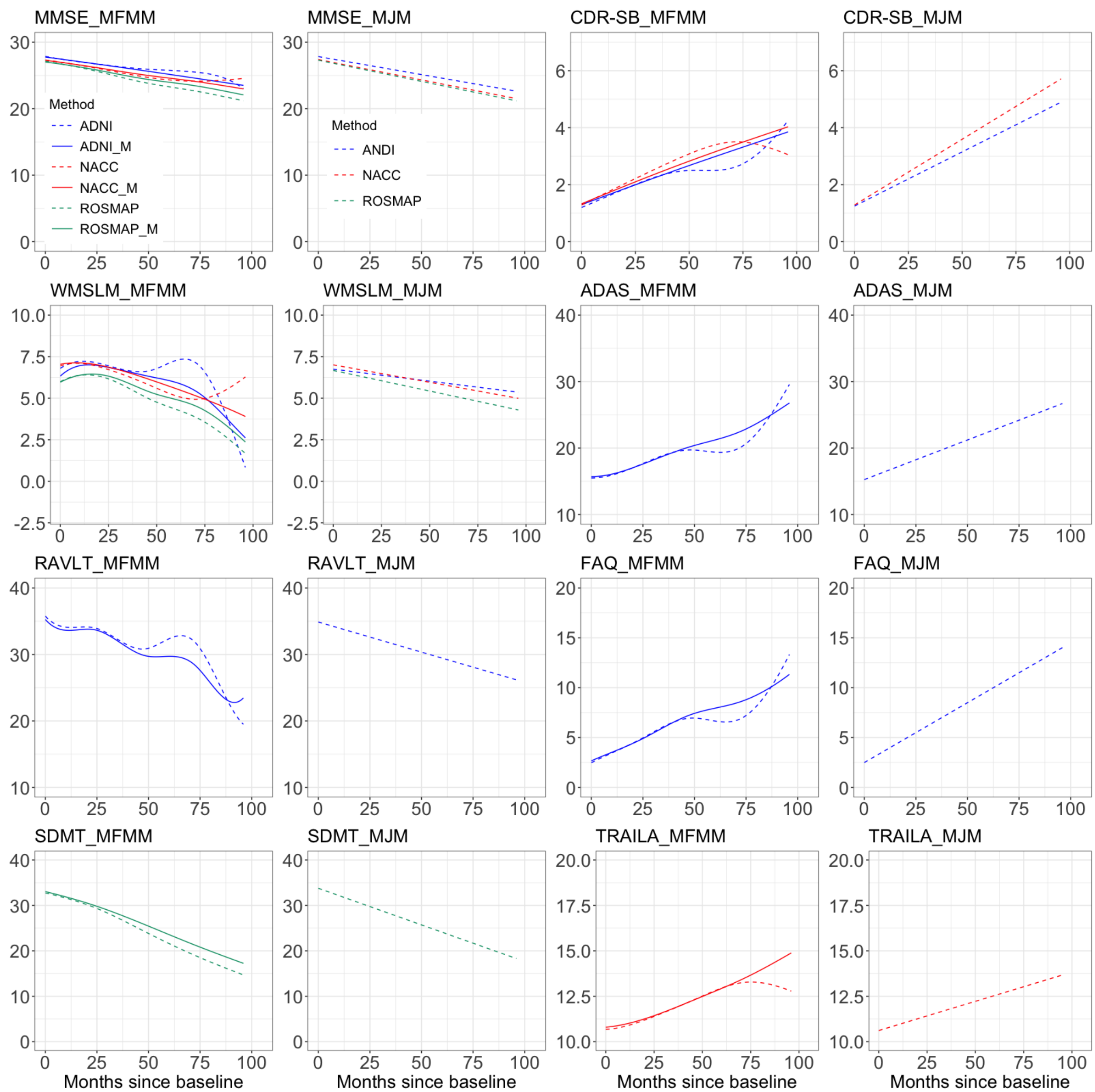


Figure 2.3: Estimated cohort-specific mean trajectories of longitudinal outcomes for ADNI (blue lines), NACC (red lines) and ROSMAP (green lines). Columns 1 and 3 compare multi-cohort FJM (solid lines) with separate FJM (dashed lines), while columns 2 and 4 show results from the parametric MJM.

in the simulations are the estimated eigenfunctions from the three AD cohorts, corresponding to two shared principal components and two outcome-specific principal components.

The shared random scores ξ_{i1} and ξ_{i2} are generated from a normal distribution $N(0, \lambda_{1d})$, where $\lambda_{11} = 12.02$ and $\lambda_{12} = 0.38$. Similarly, the outcome-specific random scores ζ_{ijd} are generated $N(0, \lambda_{2d})$, with $\lambda_{21} = 3.49$ and $\lambda_{22} = 0.16$. Measurement errors, denoted as ϵ_{ijk}, are sampled from a normal distribution $N(0, \sigma_j^2)$. The observed time points $t_{ijk} = t_{ik}$ correspond to 11 fixed time points mapped to the interval $[0, 1]$ and are subject to truncation based on censoring and event times, as described below.

The events data are generated using the Cox regression model in (2.2), incorporating the shared random scores and four baseline covariates from the data application. The linear hazard rate function is specified as $\boldsymbol{z}_i^T \boldsymbol{\gamma}_{1c(i)} + \xi_{i1}\gamma_{2c(i)1} + \xi_{i2}\gamma_{2c(i)2}$, where the Cox coefficients are based on estimates from the data application. Specifically, $\boldsymbol{\gamma}_{11} = (0.02, 0.14, 0.05, 0.30)^T$, $\boldsymbol{\gamma}_{12} = (0.02, 0.13, 0.09, 0.35)^T$, and $\boldsymbol{\gamma}_{13} = (0.02, -0.19, 0.07, 0.40)^T$, with shared coefficients $\gamma_{211} = 0.53$, $\gamma_{212} = -1.04$, $\gamma_{221} = 0.45$, $\gamma_{222} = -1.11$, $\gamma_{231} = 0.59$, and $\gamma_{232} = -0.09$. A Weibull baseline hazard $h_0(t) = \rho t^{\rho - 1}$ with $\rho = 20$ is used, and event times S_i are generated via the inverse probability method (Bender et al. 2005). Censoring time $C_i s$ are sampled independently from a Beta distribution, with $\alpha_0^{c(i)}$ and $\beta_0^{c(i)}$ chosen to approximate censoring rates of 66%, 66% and 59%, matching the three AD cohorts. For each subject, only measurements at $t_{ik} \leq T_i = \min(S_i, C_i)$ are retained. The cohort sample sizes are 715, 3707, and 522 for cohorts 1, 2, and 3, respectively, with average number of observations per subject of 7.8, 6.7, and 7.3. These settings closely mimic the AD study, ensuring realistic simulation conditions. We simulate data 100 times.

2.6.2 Simulation Results

To assess model performance, we first fit the multi-cohort FJM using the true number of shared and outcome-specific principal components. Figure 2.4 shows the estimated mean functions across 100 replications. Gray lines represent estimates from individual replicates, dashed blue lines show the average of these estimates, and solid red lines represent the true mean functions. The close alignment of the dashed blue and solid red lines indicates accurate estimation. The estimated eigenfunctions and a comprehensive assessment of estimated model parameters are presented in Appendix A.5. These parameters include the eigenvalues $\boldsymbol{\Lambda}$, the outcome-specific scaling parameter $\boldsymbol{\beta}$, the Cox regression coefficients $\boldsymbol{\gamma}$, and the white noise variance σ_j^2. The estimated eigenfunctions demonstrate strong agreement with the ground truth, and the estimated model parameters align closely with their true values, confirming the good performance of the proposed multi-cohort FJM.

Finally, we evaluate model selection using AIC and BIC to determine the number of eigenfunctions in the covariance structures. The proposed selection method demonstrates high accuracy, achieving correct selection rates of 0.98 for AIC and 1.00 for BIC, respectively. These results highlight the robustness of BIC in correctly identifying model complexity, making it the preferred criterion for model selection.

In summary, these simulation results substantiate the reliability and effectiveness of the proposed multi-cohorts FJM, demostrating its ability to accurately estimate model parameters and select the appropriate number of eigenfunctions.

2.7 Discussion

We proposed a novel extension of the multivariate functional mixed model (MFMM) (Li et al. 2022) to jointly analyze longitudinal outcomes from multiple cohorts. Our approach offers a principled solution to the challenges posed by disparate outcome collections across cohorts. By leveraging shared variation patterns extracted from the MFMM, we achieved a parsimonious yet flexible framework for linking longitudinal and survival data. This approach enables the robust characterization of disease trajectories and their associations with survival outcomes across heterogeneous cohorts.

The application to the three AD cohorts underscores the model's utility in uncovering both shared and cohort-specific disease progression patterns. By identifying differences in disease trajectories across cohorts, such as the faster progression in ROSMAP compared to ADNI and NACC, the model revealed insights that would be unavailable from single-cohort analyses. Additionally, the inclusion of cohort-specific survival model coefficients allowed us to capture inter-cohort heterogeneity in baseline covariate effects, offering a more comprehensive understanding of AD risk factors, including the differential impact of age, education, and APOE4 status across cohorts.

For the proposed functional joint model, we developed a computationally feasible algorithm that combines effectively EM and penalized splines. The iterative nature of EM matches well with the local selection of smoothing parameters for penalized splines, overcoming the computational complexity of functional joint models. The success of our algorithm will no doubt encourage further uses of penalized splines in iterative algorithms. And it is worth noting that the strategy of local selection of smoothing parameters is generally applicable to any iterative algorithm with nonparametric smooth functions to estimate.

Despite its strengths, the proposed model has limitations that warrant further investigation. Currently, the longitudinal and survival sub-models are linked parametrically via the principal scores. While this approach ensures model simplicity, it may not fully capture more complex associations between longitudinal outcomes and survival data. Future work could incorporate

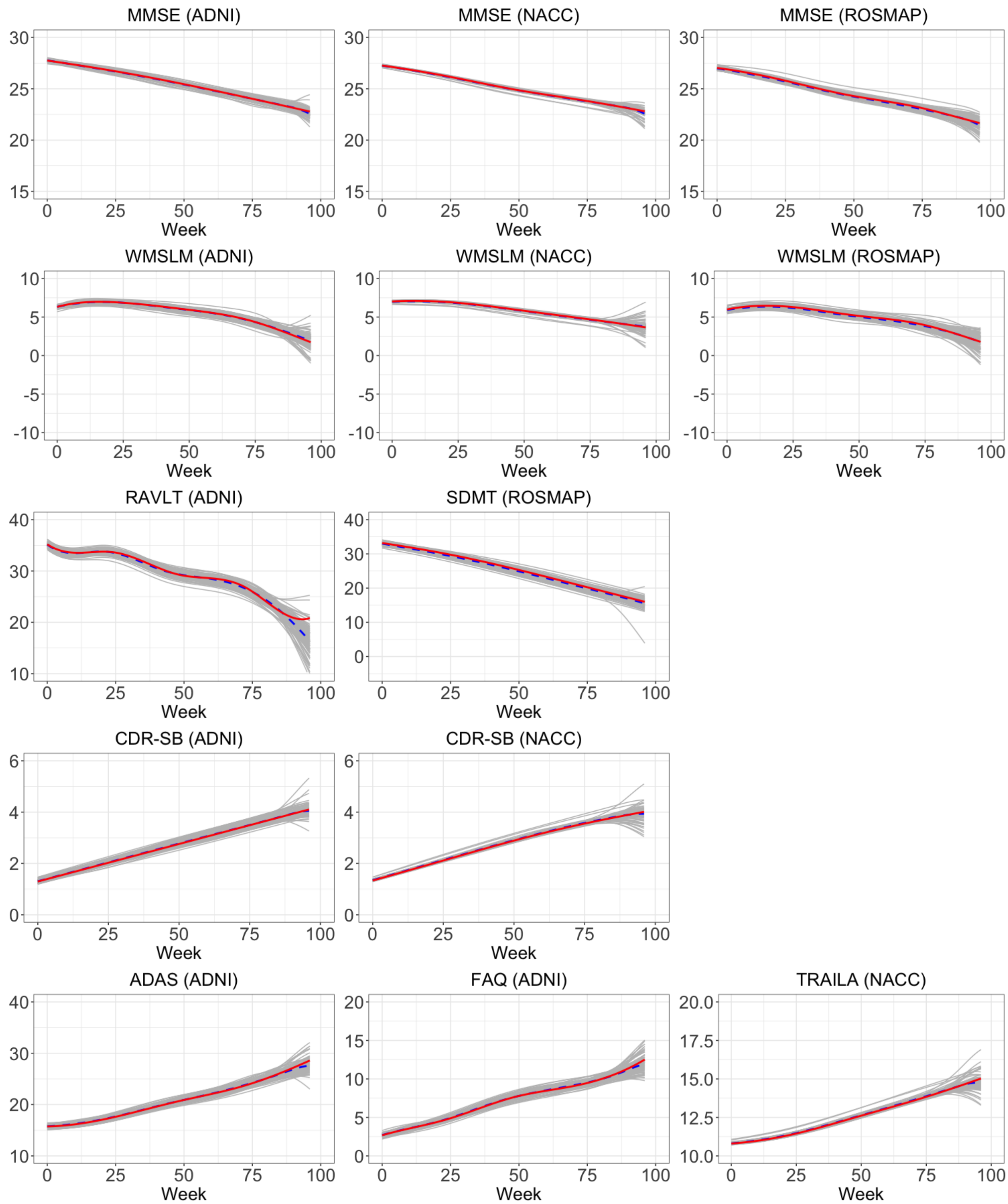


Figure 2.4: Estimated mean functions from 100 simulation replications. Gray lines: individual estimates; dashed blue lines: the average of these estimates; solid red lines: true mean functions.

nonparametric linking mechanisms Zou et al. (2023), to enhance flexibility and accommodate complex relationships between the two sub-models.

Another limitation is the reliance on a single shared variation pattern to represent commonalities across longitudinal outcomes. While sufficient for moderate numbers of outcomes, this may become restrictive when analyzing a large set of outcomes with diverse dependencies. The recent latent functional factor model Li and Xiao (2023) allows for multiple shared variation patterns and offers a promising direction for extending MFMM to handle such complexity. Incorporating multiple shared variation patterns into the functional joint model could further improve its ability to capture the rich structure of longitudinal data and enhance its applicability to more complex multi-cohort studies.

CHAPTER

3

FUNCTIONAL DATA ANALYSIS OF HEAD IMPACT EXPOSURE OF COLLEGE FOOTBALL ATHLETES

3.1 Introduction

3.1.1 Data

[1] Our project uses a subset of the data from the CARE (Concussion Assessment Research and Education) Consortium ARC (Advanced Research Core) head impact measurements (HIM) study. In total there were 717 NCAA Division I college football athletes enrolled in CARE-ARC who had head impact data collected over 5 football seasons using the Head Impact Telemetry (HIT) system resulting in data from 1,021 player-seasons. Details of the study are summarized in Broglio et al. (2017). The CARE-ARC HIM cohort provides a unique opportunity to investigate the effects of repetitive HAEs on concussion risk and its effects as the data were collected during the whole season over multiple years. All football players also had comprehensive longitudinal assessments, including neurocognitive and behavioral testing, blood biomarker

[1]This chapter is joint work with Luo Xiao, Alok S. Shah, Brian D. Stemper and Jaroslaw Harezlak.

and DNA collection (Broglio et al. 2017). We analyze here data collected from four civilian and military colleges. The seven positions of those athletes include wide receiver (WR), offensive linemen (OL), defensive linemen (DL), line backer (LB), running back (RB), defensive back (DB), and tight end (TE). A unit of analysis is defined as a combination of a player's id and season. Table 3.1 provides the number of player-season subjects for each combination of play position and school in our analyzed data.

Table 3.1: Number of player-season subjects per school and position. OL: offensive lineman; DL: defensive lineman; TE: tight end; LB: line backer; RB: running back; DB: defensive back; WR: wide receiver.

	School 1	School 2	School 3	School 4	Total
OL	38	49	33	33	153
DL	34	26	29	37	126
TE	9	10	10	14	43
LB	22	38	35	14	109
RB	9	24	18	16	67
DB	45	29	35	24	133
WR	28	22	14	18	82
Total	185	198	174	156	713

In the CARE-ARC HIM study, the enrolled athletes wore the HIT System for every game and practice throughout that period and a head acceleration event (HAE) was recorded when the Head Impact Telemetry (HIT) system detected a linear acceleration beyond the 9.6-g threshold. The enrollment in the CARE-ARC HIM core adopted a design where each enrolled athlete participated during the preseason, regular season and spring training. For this data analysis, we consider HAEs data collected during preseaon (4 weeks) and regular season (13 weeks). For each practice or game, the number of HAEs is calculated and as the season progressed, the number of HAEs for each athlete also varied. Figure 3.1 gives the longitudinal trajectories of numbers of HAEs for a number of athletes with one athlete chosen from each position at each school. The primary objectives of this study are to model the varying patterns of HAEs during the season, and to explore how they differ across the athletes' school and position. Toward this goal, we model the trajectories of numbers of HAEs as count-valued functional data and extend existing nonparametric functional data methods (Ramsay and Silverman 2005). For simplicity, we count the number of HAEs of each player by week and then add the counts together to get cumulative counts for each week.

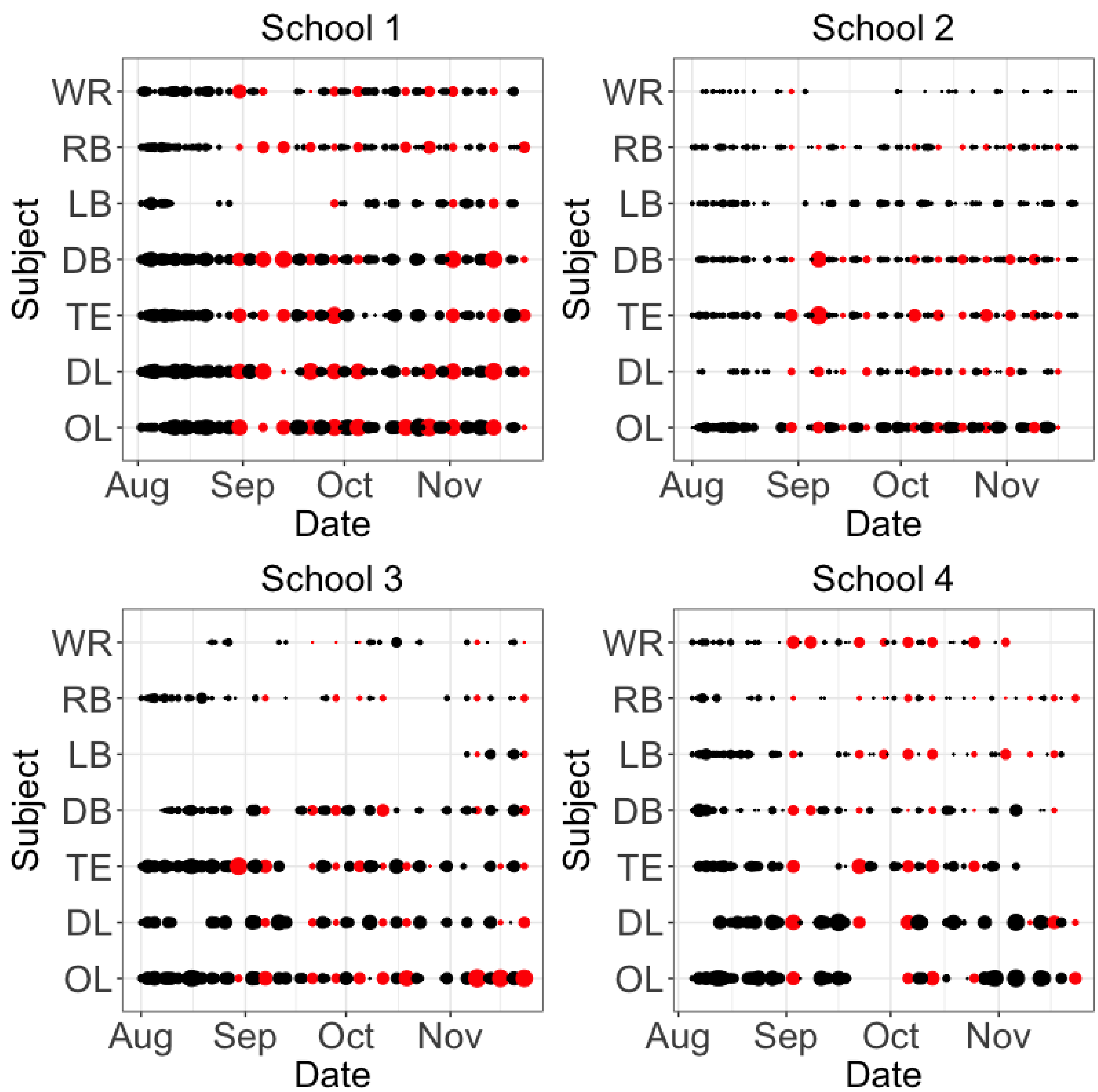


Figure 3.1: Plots of counts of HAEs for one subject from each of the 7 positions from 4 schools. Red points: counts in game; black points: counts in practice; size of points: value of counts. The average number of HAEs per practice or game for one subject is 12.01 with standard deviation 7.39.

3.1.2 Existing Literature

The rapid development and deployment of modern technology such as sensors for monitoring human health have led to remarkable success of functional data methods in the last two decades (Ramsay and Silverman 2005; Crainiceanu et al. 2024). While functional data methods have been mostly developed for Gaussian processes, non-Gaussian functional data have also become common in scientific studies. Therefore, many efforts have been devoted to extending functional principal component analysis (FPCA) to repeatedly observed data that can be modeled by distributions from the exponential family and the key idea is to introduce a latent Gaussian process. Hall et al. (2008) and Serban et al. (2013) are two earlier works that conducted model estimation using method of moments and approximations. Based on reduced-rank spline approximation of FPCs, Huang et al. (2014) developed expectation-maximization (EM) method for marginal likelihood estimation while Li et al. (2014) proposed a quasi-likelihood approach by transforming observations into “pseudo" normal variables before employing EM. To allow for covariate-dependent mean functions, Scheipl et al. (2015) proposed functional additive mixed models (FAMM) that used splines to approximate each individual latent Gaussian process and conducted inference under the generalized linear mixed model framework. To reduce the computational difficulty of FAMM due to the estimation of FPCs, Gertheiss et al. (2017) utilized a two-stage approach in which FPCs were first estimated using methods such as in Hall et al. (2008) and then were fixed for inference of other parameters. Similarly and for a computationally convenient estimation of FPCs, Leroux et al. (2023) estimated FPCs by smoothing of local mixed effects. The above literature treated FPC scores as random effects and when treating FPC scores as fixed parameters, Li et al. (2018) developed an iterative reweighted least squares (IRLS) algorithm while Zhong et al. (2023) used a majorization-minimization (MM) algorithm focusing on binary functional data. A number of Bayesian methods for exponential family functional data have also been developed; see, e.g., Van Der Linde (2009), Wrobel et al. (2019) and Kowal (2019).

Varying coefficient models (VCMs) (Hastie and Tibshirani 1993) are often an integral part of the more complex models with functional responses, for example, the aforementioned FAMM in Scheipl et al. (2015). When the number of varying coefficients is more than just a few, overfitting tends to be an issue for VCMs and dimension reduction methods have been employed for more efficient model estimation. Indeed, Jiang et al. (2013) proposed the seminal principal varying coefficient model in which all coefficient functions are linear combinations of a much smaller set of so called principal functions. This dimension reduction approach has also been successfully extended to quantile regression (Zhao et al. 2018), multivariate response models (He et al. 2018), and the Tucker decomposition of a matrix of coefficient functions for multivariate response models (Zhang et al. 2022). We shall adopt this dimension reduction strategy in our model.

3.1.3 Outline of Proposed Model and Method

In this paper, we apply non-Gaussian functional data analysis with distributions from the exponential family to analyze the data of head impact exposures. With 4 schools and 7 positions as shown in Table 3.1, there are 28 mean functions to estimate, thus susceptible or even inevitable to overfitting using existing standard methods. Moreover, for several combinations of schools and positions (TE from School 1 for example), the dataset has only very few number of players, which presents substantial variability of mean function estimation for those combinations. To stabilize or even improve model estimation, we adopt the dimension reduction strategy in the literature of varying coefficient models and in particular, we exploit the array structure of mean functions and make use of its Tucker decomposition. To carry out model estimation, we develop an effective EM algorithm and use the quasi-likelihood approach in Li et al. (2014), with the latter providing a faster convergence than standard EM algorithm as in Huang et al. (2014). The Tucker decomposition requires an iterative estimation, which fits well into the iterative EM algorithm. Moreover, we use penalized splines to estimate nonparametrically the principal functions in the Tucker decomposition, which requires selection of smoothing parameters. We adopt a local selection of smoothing parameters, which can be done effectively in each step of the proposed EM algorithm.

The rest of the paper is organized as follows. Section 3.2 details the proposed model. Section 3.3 considers model estimation. The application of our proposed method to analyze the HAEs dataset is given in Section 3.4. To assess the performance of our method, Section 3.5 presents the results of a simulation study. We conclude the paper with discussions in Section 3.6.

3.2 Model

Let Y_{ij} be the detected number of Head Acceleration Events (HAEs) during practices or games for athlete $i (1 \le i \le n)$ during week $T_{ij} (1 \le j \le m_i)$. Here $T_{ij} \in \mathscr{T} = [-3, 13]$, where 1 indicates the first week since the regular seasons started while negative values denote pre-season weeks. Let $N_{ij} = N_i(T_{ij})$, where $N_i(t)$ is the number of practices or games athlete i participated during week t. We use the grouping function $s(i) \in \{1,2,3,4\}$ to denote the athlete's school and $p(i) \in \{1,2,3,4,5,6,7\}$ to denote the athlete's position on the football field. There are $n_s = 4$ schools and $n_p = 7$ positions.

We model the count data Y_{ij} as Poisson random variables with mean $N_{ij}\mu_{ij}$, where μ_{ij} is the rate of HAEs per practice/game for athlete i during week T_{ij}. To model the dependence between the repeatedly observed counts,

we use the functional data model for $\eta_{ij} = \eta_i(T_{ij}) = \log(N_{ij}\mu_{ij})$,

$$\begin{aligned} \eta_i(t) &= \log\{N_i(t)\} + \eta_{0s(i)p(i)}(t) + X_i(t), \\ \eta_{0s(i)p(i)}(t) &= \sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3} a_{r_1 r_2 r_3} U^{(1)}_{s(i)r_1} U^{(2)}_{p(i)r_2} g_{r_3}(t), \end{aligned} \tag{3.1}$$

where $\eta_{0s(i)p(i)}(t)$ is the mean function for athletes in school $s(i)$ and position $p(i)$ and $X_i(t)$ is a Gaussian process that models the subject-specific deviation from mean function. In model (3.1), there is a 4×7 matrix of mean functions, with each mean function corresponding to one combination of school and position. As shown in Table 3.1, some positions (e.g., TE) have data from very few athletes in all schools while some school has data from very few athletes in some position (e.g., school 1 at RB); as a result, the estimation of those mean functions may have large variations and can be difficult to interpret. To enhance interpretation and achieve a more stable model estimation, we integrate the mean function estimation through dimension reduction of tensors. Indeed, the collection of mean functions can be treated as a 3rd-order tensor, with the time t as the third (infinite) dimension (Zhang et al. 2022). Then the second equation of model (3.1) gives the Tucker decomposition (Tucker 1966) for a third-order tensor. In the decomposition, $a_{r_1 r_2 r_3}$ is the (r_1, r_2, r_3)th element of a 3rd-order core tensor $\boldsymbol{A} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$, $U^{(1)}_{s(i)r_1}$ is the $(s(i), r_1)$th element of an orthogonal matrix $\boldsymbol{U}^{(1)} \in \mathbb{R}^{n_s \times R_1}$, $U^{(2)}_{p(i)r_2}$ is the $(p(i), r_2)$th element of an orthogonal matrix $\boldsymbol{U}^{(2)} \in \mathbb{R}^{n_p \times R_2}$, and $\left\{g_1(t), \cdots, g_{R_3}(t)\right\}$ has R_3 orthonormal smooth functions. We call $g_{r_3}(t)$s the principal functions as in principal varying coefficient models (Jiang et al. 2013) in which a vector of functions is decomposed via the Tucker decomposition of an order-2 tensor. The 3-dimensional core tensor $\boldsymbol{A}$, and two projection matrices $\boldsymbol{U}^{(1)}$ and $\boldsymbol{U}^{(2)}$, as well as the principal functions $g_{r_3}(t)$s are unknown and need to be estimated. The triplet (R_1, R_2, R_3) is the multilinear rank, determining the complexity of the decomposition. For simplicity, we assume $R_1 = R_2 = R_3 = R$ in our model and it is required that $R \le \min(n_s, n_p)$. The Tucker decomposition is an extension of matrix singular value decomposition to tensors (called higher-order SVD, or HOSVD), and when R is small, the number of parameters to be estimated is much smaller than estimating a full matrix of functions, thus enabling a more efficient and computationally simpler model estimation as will be given later.

Let $\mathscr{C}(s,t) = \mathrm{Cov}\{X_i(s), X_i(t)\}$ be the covariance function with the eigendecomposition $\mathscr{C}(s,t) = \sum_{\ell \ge 1} \lambda_\ell \phi_\ell(s)\phi_\ell(t)$, where $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$ are eigenvalues with associated orthonormal eigenfunctions $\phi_\ell(t)$, i.e., $\int_{\mathscr{T}} \phi_\ell(t)\phi_{\ell'}(t)dt = \delta_{\ell\ell'}$. By the Karhunen–Loève representation, $X_i(t) = \sum_{\ell \ge 1} \xi_{i\ell}\phi_\ell(t)$, where $\xi_{i\ell} \sim N(0, \lambda_\ell)$ are random eigen scores that are independent across

i and ℓ. We assume there are only a finite number of variation patterns modeled by $\phi_\ell(t)$s so that $X_i(t) = \sum_{\ell=1}^{L} \xi_{i\ell}\phi_\ell(t)$, where L is a tuning parameter and will be selected through data adaptive methods.

3.3 Model Estimation

3.3.1 Splines

We use splines for approximating the principal functions in the Tucker decomposition of mean functions and also eigenfunctions of the subject-specific smooth deviations. Let $\boldsymbol{b}(t) = \big(B_1(t), \cdots, B_K(t)\big)^\top \in \mathbb{R}^K$ be the vector of B-spline basis functions in $\mathscr{T}$ (De Boor and De Boor 1978), where K is the number of equally-spaced interior knots plus the order (degree plus 1) of the B-splines. Let $\boldsymbol{G} = \int \boldsymbol{b}(t)\boldsymbol{b}(t)^\top dt \in \mathbb{R}^{K\times K}$, which is positive definite (Shen et al. 1998). Then $\widetilde{\boldsymbol{b}}(t) = \big(\widetilde{B}_1(t), \cdots, \widetilde{B}_K(t)\big)^\top = \boldsymbol{G}^{-1/2}\boldsymbol{b}(t)$ are orthonormal B-spline bases. We approximate the function $g_{r_3}(t)$ in the Tucker decomposition by $\widetilde{\boldsymbol{b}}(t)^\top \boldsymbol{u}^{(3)}_{r_3}$, where $\boldsymbol{u}^{(3)}_{r_3} = \big(U^{(3)}_{1r_3}, \cdots, U^{(3)}_{Kr_3}\big)^\top \in \mathbb{R}^K$ are coefficient vectors and then $\boldsymbol{U}^{(3)} = [\boldsymbol{u}^{(3)}_1, \cdots, \boldsymbol{u}^{(3)}_{R_3}] \in \mathbb{R}^{K\times R_3}$ is an orthogonal matrix by the orthonormality of $\big\{g_1(t), \cdots, g_{R_3}(t)\big\}$. We approximate the ℓth eigenfunction $\phi_\ell(t)$ by $\widetilde{\boldsymbol{b}}(t)^\top \boldsymbol{\theta}_\ell$, where $\boldsymbol{\theta}_\ell$ are coefficient vectors. Let $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_L]$, then the orthonormality of eigenfunctions leads to the constraints $\boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{I}_{L\times L}$, which are equivalent to $\boldsymbol{\theta}_\ell^\top \boldsymbol{\theta}_{\ell'} = \delta_{\ell\ell'}$. Then we approximate model (3.1) as

$$
\begin{aligned}
\eta_i(t) &= \log\{N_i(t)\} + \eta_{0s(i)p(i)}(t) + \widetilde{\boldsymbol{b}}(t)^\top \boldsymbol{\Theta}\boldsymbol{\xi}_i \\
\eta_{0s(i)p(i)}(t) &= \sum_{k=1}^{K} \widetilde{B}_k(t)\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3} a_{r_1r_2r_3} U^{(1)}_{s(i)r_1} U^{(2)}_{p(i)r_2} U^{(3)}_{kr_3}\right\},
\end{aligned} \tag{3.2}
$$

where $\boldsymbol{\xi}_i = (\xi_{i1}, \cdots, \xi_{iL})^\top \sim \mathrm{MVN}(\boldsymbol{0}, \boldsymbol{\Lambda})$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_L)$. Let $\boldsymbol{T} \in \mathbb{R}^{n_s\times n_p\times K}$ be a 3rd-order tensor with $\boldsymbol{T}(s,p,k) = \sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3} a_{r_1r_2r_3} U^{(1)}_{sr_1} U^{(2)}_{pr_2} U^{(3)}_{kr_3}$. Then, $\boldsymbol{A}, \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}$ together actually gives the HOSVD of the tensor $\boldsymbol{T}$ and the second equation of (3.2) shows the matrix of mean functions is the 3–mode multiplication of $\boldsymbol{T}$ with the row vector $\widetilde{\boldsymbol{b}}^\top(t)$. Denote by $\boldsymbol{\Omega} = \{\boldsymbol{A}, \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}\}$ the collection of all parameters.

3.3.2 “Pseudo" Normal Distribution

Likelihood-based model estimation of exponential family distributions often involves iteratively reweighed least squares, which can be become computationally intensive and has slow convergence when random effects need to be integrated out from the joint likelihood. Alternatively, following Li et al. (2014), we employ a penalized quasilikelihood approach to transform the count variable Y_{ij} into a pseudo normal random variable, incorporating a second order approximation term. This method, first introduced by Goldstein and Rasbash (1996), is an improvement over the methods, e.g., standard Laplacian approximation, discussed in Breslow and Clayton (1993).

Let $H(\cdot)$ be the inverse log function, and let the first and second derivatives of $H(\cdot)$ be $H'(\cdot)$ and $H''(\cdot)$, respectively. Given current parameter estimates $\widehat{\mathbf{\Omega}} = \{\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{U}}^{(1)}, \widehat{\boldsymbol{U}}^{(2)}, \widehat{\boldsymbol{U}}^{(3)}, \widehat{\mathbf{\Theta}}, \widehat{\mathbf{\Lambda}}\}$ and prediction $\widehat{\boldsymbol{\xi}}_i$ of the score vector $\boldsymbol{\xi}_i$, let

$$\widehat{\eta}_{0sp}(t) = \sum_{k=1}^{K} \widetilde{B}_k(t) \left\{ \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} \widehat{a}_{r_1 r_2 r_3} \widehat{U}^{(1)}_{s r_1} \widehat{U}^{(2)}_{p r_2} \widehat{U}^{(3)}_{k r_3} \right\}$$

and $\widehat{\eta}_i(t) = \log\{N_i(t)\} + \widehat{\eta}_{0s(i)p(i)}(t) + \widetilde{\boldsymbol{b}}(t)^\top \widehat{\mathbf{\Theta}} \widehat{\boldsymbol{\xi}}_i$. Then we let

$$Y^*_{ij} = \widehat{\eta}_i(T_{ij}) + \frac{1}{H'\{\widehat{\eta}_i(T_{ij})\}} \left\{ Y_{ij} - H\{\widehat{\eta}_i(T_{ij})\} \right\} - \frac{1}{2H'\{\widehat{\eta}_i(T_{ij})\}} H''\{\widehat{\eta}_i(T_{ij})\} \widetilde{\boldsymbol{b}}(T_{ij})^\top \widehat{\mathbf{\Theta}} \widehat{\boldsymbol{V}}_i \widehat{\mathbf{\Theta}}^\top \widetilde{\boldsymbol{b}}(T_{ij}),$$

where $\widehat{\boldsymbol{V}}_i$ is an estimate of $\boldsymbol{V}_i = \mathrm{Var}\left(\boldsymbol{\xi}_i - \widehat{\boldsymbol{\xi}}_i\right)$. The last term on the right hand side of the above equation is the a second order approximation that accounts for the variation of the prediction error by $\widehat{\boldsymbol{\xi}}$. Appendix B.2 gives the details of $\boldsymbol{V}_i$ and its approximation $\widehat{\boldsymbol{V}}_i$.

The above approximation leads to an approximate model of model (3.2) for the observed data,

$$\begin{aligned} Y^*_{ij} &= \log(N_{ij}) + \eta_{0s(i)p(i)}(T_{ij}) + \widetilde{\boldsymbol{b}}(T_{ij})^\top \mathbf{\Theta} \boldsymbol{\xi}_i + \epsilon_{ij} \\ \eta_{0s(i)p(i)}(t) &= \sum_{k=1}^{K} \widetilde{B}_k(t) \left\{ \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \sum_{r_3=1}^{R_3} a_{r_1 r_2 r_3} U^{(1)}_{s(i) r_1} U^{(2)}_{p(i) r_2} U^{(3)}_{k r_3} \right\}, \end{aligned} \tag{3.3}$$

where ϵ_{ij} is random error with distribution $N(0, \frac{1}{H'\{\widehat{\eta}_i(T_{ij})\}})$ and independent across i and j.

3.3.3 Penalized Likelihood of Model

We now consider the likelihood of the "pseudo" normal outcomes in equation (3.3) for model estimation. Let $\boldsymbol{Y}^*_i = \left(Y^*_{i1}, \ldots, Y^*_{im_i}\right)^\top$ be the vector of "pseudo" normal observations without offset, i.e., for simplicity, we assume that the offset term $\log(N_{ij})$ has been subtracted from Y^*_{ij}. Let $\boldsymbol{\eta}_{0s(i)p(i)} = \left\{\eta_{0s(i)p(i)}(T_{i1}), \ldots, \eta_{0s(i)p(i)}(T_{im_i})\right\}^\top$, $\widehat{\boldsymbol{\eta}}_i = \left\{\widehat{\eta}_i(T_{i1}), \ldots, \widehat{\eta}_i(T_{im_i})\right\}^\top$, and $\mathbf{\Sigma}_i = \mathrm{diag}(e^{-\widehat{\boldsymbol{\eta}}_i})$ the diagonal covariance matrix. Finally, let $\boldsymbol{B}_i = \left\{\widetilde{\boldsymbol{b}}(T_{i1}), \ldots, \widetilde{\boldsymbol{b}}\left(T_{im_i}\right)\right\}^\top$ be the $m_i \times K$ B-spline design matrix. Then, for the “pseudo" normal data, we have the conditional likelihood,

$$f\left(\boldsymbol{Y}_i^* \mid \boldsymbol{\xi}_i\right) = (|2\pi\boldsymbol{\Sigma}_i|)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{Y}_i^* - \boldsymbol{\eta}_{0s(i)p(i)} - \boldsymbol{B}_i\boldsymbol{\Theta}\boldsymbol{\xi}_i\right)^\top \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{Y}_i^* - \boldsymbol{\eta}_{0s(i)p(i)} - \boldsymbol{B}_i\boldsymbol{\Theta}\boldsymbol{\xi}_i\right)\right\}, \quad (3.4)$$

and $f\left(\boldsymbol{\xi}_i \mid \boldsymbol{\Lambda}\right) = (|2\pi\boldsymbol{\Lambda}|)^{-1/2} \exp\left(-\frac{1}{2}\boldsymbol{\xi}_i^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}_i\right)$. Therefore, the joint log-likelihood for the “pseudo" normal data has the form

$$\mathscr{L} = \sum_{i=1}^{n} \log f\left(\boldsymbol{Y}_i^* \mid \boldsymbol{\xi}_i\right) + \sum_{i=1}^{n} \log f\left(\boldsymbol{\xi}_i \mid \boldsymbol{\Lambda}\right).$$

For estimation of the smooth functions, we incorporate roughness penalties to regularize the fits of the functions (Eilers and Marx 1996b). This penalized spline approach provides stable estimators and facilitates the convergence of the EM algorithm to be detailed later. Thus, we will use the penalized log-likelihood,

$$\mathscr{L}_{pen} = \mathscr{L} - \sum_{r_3=1}^{R_3} \tau_{0r_3} \boldsymbol{u}_{r_3}^{(3)\top} \boldsymbol{P} \boldsymbol{u}_{r_3}^{(3)} - \sum_{\ell=1}^{L} \tau_\ell \boldsymbol{\theta}_\ell^\top P \boldsymbol{\theta}_\ell,$$

where τ_{0r_3}, τ_ℓ are smoothness parameters, the penalty matrix is $P = \int \widetilde{\boldsymbol{b}}''(t)\widetilde{\boldsymbol{b}}''(t)^\top dt$, and the first summation has the smoothness penalty for the principal functions $g_{r_3}(t)$s while the second summation has the smoothness penalty for the eigenfunctions $\phi_\ell(t)$s.

3.3.4 Parameter Estimation Using the EM algorithm

The spline approximation of smooth functions leads to a parametric model, and we shall employ the expectation-maximization (EM) algorithm, which is an effective strategy commonly utilized in random effect models by treating random effects as missing values. The EM algorithm works by iteratively performing two steps until convergence: computing expectation of joint log-likelihood with respect to random effects conditional on the observed data and the current estimate of parameters (E-step), and maximizing the expected log-likelihood to update estimates of parameters (M-step). For our model estimation, we follow iterations involving updates of "pseudo" normal data using the current parameter values, E-steps and M-steps.

In the E-step, the conditional distribution of the random scores $\boldsymbol{\xi}_i$ is approximated via a normal distribution based on the pseudo normal distributions in equation (3.3); see Appendix B.1. The expectation of the penalized log-likelihood conditional on the observed data and the current parameter values $\widehat{\boldsymbol{\Omega}}$ is

$$Q(\boldsymbol{\Omega}) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\xi}_i | \boldsymbol{Y}_i^*, \widehat{\boldsymbol{\Omega}}} \left\{ \log f\left(\boldsymbol{Y}_i^* \mid \boldsymbol{\xi}_i\right) + \log f\left(\boldsymbol{\xi}_i \mid \boldsymbol{\Lambda}\right) \right\} - \sum_{r_3=1}^{R_3} \tau_{0r_3} \boldsymbol{u}_{r_3}^{(3)\top} \boldsymbol{P} \boldsymbol{u}_{r_3}^{(3)} - \sum_{\ell=1}^{L} \tau_\ell \boldsymbol{\theta}_\ell^\top P \boldsymbol{\theta}_\ell.$$

In the M-step, current iteration estimates are obtained by maximizing the functions $Q(\boldsymbol{\Omega})$ with respect to $\boldsymbol{\Omega}$.

First, $\boldsymbol{A}$, $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$, $\boldsymbol{U}^{(3)}$, $\boldsymbol{\Theta}$ can be estimated iteratively by maximizing the terms in $Q(\boldsymbol{\Omega})$ that involves those parameters, i.e.,

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{j=1}^{m_i} \exp\{\widehat{\eta}_i(T_{ij})\} \mathbb{E}_{\boldsymbol{\xi}_i | \boldsymbol{Y}_i^*, \widehat{\boldsymbol{\Omega}}} \left[\left\{ Y_{ij}^* - \eta_{0s(i)p(i)}(T_{ij}) - \widetilde{\boldsymbol{b}}(T_{ij})^\top \boldsymbol{\Theta} \boldsymbol{\xi}_i \right\}^2 \right] \\ &- \sum_{r_3=1}^{R_3} \tau_{0r_3} \boldsymbol{u}_{r_3}^{(3)\top} \boldsymbol{P} \boldsymbol{u}_{r_3}^{(3)} - \sum_{\ell=1}^{L} \tau_\ell \boldsymbol{\theta}_\ell^\top P \boldsymbol{\theta}_\ell. \end{aligned} \tag{3.5}$$

Appendix B.3 provides the technical details of each estimation step. Second, $\boldsymbol{\Lambda}$ are estimated by maximizing $\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\xi}_i | \boldsymbol{Y}_i^*, \widehat{\boldsymbol{\Omega}}} \log f\left(\boldsymbol{\xi}_i \mid \boldsymbol{\Lambda}\right)$, from which we obtain the estimator $\widehat{\lambda}_\ell = n^{-1} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\xi}_i | \boldsymbol{Y}_i^*, \widehat{\boldsymbol{\Omega}}} \left(\xi_{i\ell}^2\right)$.

3.3.5 Estimation Algorithm

The proposed iterative algorithm requires initial values of the model parameters $\boldsymbol{A}$, $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$, $\boldsymbol{U}^{(3)}$, $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$ and predictions of scores $\boldsymbol{\xi}_i$. We begin by simply transforming the observed count data Y_{ij} to $Y_{ij}^{(0)} = \log(Y_{ij}) - \log(N_{ij})$. Using $Y_{ij}^{(0)}$, we apply the `gam` function from the `mgcv` (Wood 2011; Wood et al. 2016) package in `R` to compute the smoothed mean functions and corresponding coefficient vectors $\boldsymbol{T}(s, p, \cdot)$ for each combination of school and position, under the working independence assumption for the observations from each subject. The resulting $\boldsymbol{T}(s, p, k)$ values are then processed using the higher-order orthogonal iteration (HOOI) algorithm (De Lathauwer et al. 2000) to obtain the initial estimates of $\boldsymbol{A}$, $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$, $\boldsymbol{U}^{(3)}$. After subtracting the estimated mean functions from $Y_{ij}^{(0)}$, we compute the initial estimates of $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\xi}_i$ using the method of fast covariance estimation for sparse functional data (FACEs) (Xiao et al. 2018).

Assume current estimated values of parameters at the dth step are

$$\boldsymbol{A}^{(d)}, \boldsymbol{U}^{(1)^{(d)}}, \boldsymbol{U}^{(2)^{(d)}}, \boldsymbol{U}^{(3)^{(d)}}, \boldsymbol{\theta}_1^{(d)}, \cdots, \boldsymbol{\theta}_L^{(d)}, \lambda_1^{(d)}, \cdots, \lambda_L^{(d)}, \boldsymbol{\xi}_i^{(d)}.$$

Then we can calculate $\eta_i^{(d)}(T_{ij})$, $\boldsymbol{\Psi}^{(d)}$, $\boldsymbol{\Sigma}_i^{(d)}$ and $\boldsymbol{R}_i^{(d)}$. Based on these values, we carry out the $(d+1)$th iteration as in Algorithm 1.

Algorithm 1 Estimation Algorithm

1: Calculate Y_{ij}^* by equation (3.3).

2: Update $\boldsymbol{A}^{(d+1)}, \boldsymbol{U}^{(1)^{(d+1)}}, \boldsymbol{U}^{(2)^{(d+1)}}, \boldsymbol{U}^{(3)^{(d+1)}}, \boldsymbol{\theta}_1^{(d+1)}, \cdots, \boldsymbol{\theta}_L^{(d+1)}, \lambda_1^{(d+1)}, \cdots, \lambda_L^{(d+1)}$ using the EM algorithm. Specifically, in the M step:

   2.1 Update $\boldsymbol{A}^{(d+1)}, \boldsymbol{U}^{(1)^{(d+1)}}, \boldsymbol{U}^{(2)^{(d+1)}}, \boldsymbol{U}^{(3)^{(d+1)}}$ by minimizing the objective function in (3.5).

   2.2 Calculate $\boldsymbol{T}(s,p,k) = \sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3} a_{r_1 r_2 r_3}^{(d+1)} U_{s r_1}^{(1)^{(d+1)}} U_{p r_2}^{(2)^{(d+1)}} U_{k r_3}^{(3)^{(d+1)}}$. Apply the higher-order orthogonal iteration (HOOI) algorithm (De Lathauwer et al. 2000) to $\boldsymbol{T}$ to obtain updated $\boldsymbol{A}^{(d+1)}$ and orthogonal $\boldsymbol{U}^{(1)^{(d+1)}}, \boldsymbol{U}^{(2)^{(d+1)}}, \boldsymbol{U}^{(3)^{(d+1)}}$.

   2.3 Update $\boldsymbol{\theta}_1^{(d+1)}, \cdots, \boldsymbol{\theta}_L^{(d+1)}, \lambda_1^{(d+1)}, \cdots, \lambda_L^{(d+1)}$.

   2.4 Calculate $\boldsymbol{\Theta}^\top \boldsymbol{\Lambda} \boldsymbol{\Theta}$ and apply eigendecomposition to obtain eigenvectors and eigenvalues, yielding updated orthonormal $\boldsymbol{\theta}_1^{(d+1)}, \cdots, \boldsymbol{\theta}_L^{(d+1)}$ and $\lambda_1^{(d+1)}, \cdots, \lambda_L^{(d+1)}$.

3: Update $\eta_i^{(d+1)}(T_{ij}), \boldsymbol{\Psi}^{(d+1)}, \boldsymbol{\Sigma}_i^{(d+1)}, \boldsymbol{R}_i^{(d+1)}$.

4: Update $\xi_i^{(d+1)} = \boldsymbol{\Lambda}^{(d+1)} \boldsymbol{\Theta}^{(d+1)^\top} \boldsymbol{B}_i^\top \boldsymbol{R}_i^{(d+1)^{-1}} (\boldsymbol{Y}_i^* - \boldsymbol{\eta}_{0s(i)p(i)}^{(d+1)})$ (BLUP).

Repeat the above procedure to the $(d+1)$th step until convergence.

3.3.6 Model Selection

There are a few model parameters that need to be selected. The smoothness parameters τ_{0r_3} and τ_ℓ are selected by using the method of restricted maximum likelihood (REML) of the pseudo normal distributions locally in each iteration. As shown in Appendix B.3, in each iteration, the estimation of principal functions and eigenfunctions can all be formulated as a weighted least square problem for a varying coefficient nonparametric regression model. We exploit this formulation and employ the well-developed R package *mgcv* to simultaneously conduct model estimation and selection of the smoothness parameter.

We next select the number of eigenfunctions, L, as well as the dimension of the core tensor, R. We shall employ information criteria (Akaike and Bayesian Information Criteria) for model selection, for which we need to evaluate the marginal log-likelihood function of the data and calculate the degrees of freedom of the model. Let $\boldsymbol{Y}_i = \{Y_{i1}, \ldots, Y_{im_i}\}^\top$, the marginal log-likelihood of the Poisson data can be approximated by Monte Carlo approximation:

$$\begin{aligned}
\ell(\boldsymbol{\Omega}) &= \sum_{i=1}^n \log \int \left\{ \prod_{j=1}^{m_i} f\left(Y_{ij} \mid \xi_i\right) \right\} f\left(\xi_i\right) d\xi_i \\
&\approx \sum_{i=1}^n \log \left\{ \frac{1}{N} \sum_{c=1}^N \left\{ \prod_{j=1}^{m_i} f\left(Y_{ij} \mid \xi_i^{(c)}\right) \right\} \right\}
\end{aligned}$$

where $\xi_i^{(c)}$ is the cth sample from the normal distribution $f\left(\xi_i | \boldsymbol{\Lambda}\right)$ and N random samples are drawn. We use a sufficiently large N to ensure the Monte Carlo error is small.

We approximate the degrees of freedom of the model by

$$\mathrm{df} := R^3 + n_s R - R(R+1)/2 + n_p R - R(R+1)/2 + \sum_{r_3=1}^{R_3} \mathrm{df}_{g_{r_3}} - R(R+1)/2 + \sum_{\ell=1}^{L} \mathrm{df}_{\phi_\ell} + L - L(L+1)/2,$$

where R^3 is the degrees of freedom for estimating $\boldsymbol{A}$, $n_s R - R(R+1)/2$ is the degrees of freedom for estimating $\boldsymbol{U}^{(1)}$, $n_p R - R(R+1)/2$ is the degrees of freedom for estimating $\boldsymbol{U}^{(2)}$, $\sum_{r_3=1}^{R_3} \mathrm{df}_{g^{(3)}_{r_3}} - R(R+1)/2$ is the degrees of freedom for estimating the principal functions, the $-R(R+1)/2$ term is due to orthonormality constraints of $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$ and $\boldsymbol{U}^{(3)}$, $\sum_{\ell=1}^{L} \mathrm{df}_{\phi_\ell}$ correspond to the degrees of freedom for estimating eigenfunctions, L corresponds to the number of eigenvalues and $-L(L+1)/2$ is due to orthonormality constraints of $\boldsymbol{\Theta}$, the matrix of spline coefficients for eigenfunctions. Here the degrees of freedom of penalized splines for estimating smooth functions can be easily calculated; see, e.g., the textbook Wood (2017). The final model can be either selected by $\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\Omega}}) + 2 \cdot \mathrm{df}$ or $\mathrm{BIC} = -2\ell(\widehat{\boldsymbol{\Omega}}) + \log(N_{\mathrm{obs}}) \cdot \mathrm{df}$, where N_{obs} is the total number of observations.

3.4 Data Application

We apply our method to the Head Acceleration Events (HAEs) data collected in the CARE-ARC study to capture the varying patterns of counts of HAEs during a season and their differences among player's school and position. The original dataset includes 1,021 player-season subjects over 5 football seasons with 5 colleges and 9 positions. We exclude one college due to poor data quality and select a representative sample of seven positions. Additionally, we remove a few seasons from the remaining four colleges where the data are incomplete. As a result, our final dataset consists of 713 player-season subjects, capturing HAEs from four colleges and seven positions during different pre-seasons and regular seasons, as summarized in Table 3.1. To get the estimated mean function of cumulative counts, we first model the counts in each week and then add up these counts together. Referring back to the notation of Section 3.2, $Y_i(t)$ is the count of HAEs for subject i at week t and $N_i(t)$ is the number of days during which the subject i takes part in practice or competition in week t. We analyzed up to 17 weeks in one season, including four pre-season weeks and 13 regular season weeks.

To fit the model, we used cubic B-splines with seven equally spaced interior knots. The number of Monte Carlo samples used in model selection and the stopping criteria are set as described in Appendix B.4. We select the core-tensor dimension R and the number of eigenfunctions, L, according to BIC, which is found to work well in the simulations. The BIC gives $R = 2$ and $L = 7$.

To better display results and make comparisons between schools and positions and also to visually check the fit of the model, we calculate the estimated mean functions of the accumulated HAEs given the observed N_{ij} at T_{ij}. Specifically, for subject i, we estimate the subject's mean count of HAEs at T_{ij} as $\widehat{Y}_{ij} = N_{ij}\exp\{\widehat{\eta}_{0s(i)p(i)}(T_{ij}) + 2^{-1}\widetilde{\boldsymbol{b}}(T_{ij})^\top \widehat{\boldsymbol{\Theta}}\widehat{\Lambda}\widehat{\boldsymbol{\Theta}}^\top \widetilde{\boldsymbol{b}}(T_{ij})\}$. Note that this estimation is at the population level as it is not conditional on the observed Y_{ij}s. Then the estimated mean function of the accumulated HAEs for school s and position p is given by

$$\frac{\sum_{i:s(i)=s,p(i)=p}\sum_j \delta(T_{ij} \leq t)\widehat{Y}_{ij}}{\sum_{i:s(i)=s,p(i)=p} 1}.$$

In Figure 3.2, the estimated mean functions are grouped by positions and School 3 consistently exhibits higher cumulative counts than other schools at most positions, which seems to suggest the physical playing styles of the school. Figure 3.3 organizes the estimated mean functions by schools. The plots show that athletes playing OL or DL have higher HAE counts in all 4 schools, while athletes playing WR and DB tend to have fewer HAE counts in all 4 schools, revealing the different level of exposure to head impacts by positions. In addition, the estimated mean functions exhibit a consistent profile of higher slope (rate of HAE accumulation) during pre-season (weeks -3 to 0) followed by a fairly consistent and linear profile of HAE accumulation during the regular season (weeks 1-13).

Figure 3.4 presents the first three estimated eigenfunctions and eigenvalues. The first eigenfunction is nearly a constant line, the second eigenfunction exhibits a contrast between before week 5 and after, while the third eigenfunction shows a contrast between the first 7 weeks of regular season and the other weeks. As the second and third eigenvalues are much smaller than the first one, it appears that the subject-specific variation of HAE counts does not vary much over time. Additional data application results are presented in Appendix B.5.

3.5 Simulation Studies

3.5.1 Simulation Settings

In this section, we conduct simulation studies to evaluate the performance of our method on simulated data that mimic the HAEs data. The simulated dataset consists of 713 subjects, matching the real data. The observed time points, T_{ij}, are 17 fixed time points equally distanced in the interval $[0,1]$, with the same observed time points as in the HAEs data. The varying linear predictor, $\eta_i(t)$, for each subject is generated according to model (3.1), incorporating the same school, position, and the number of days, $N_i(t)$, during which the subject participates

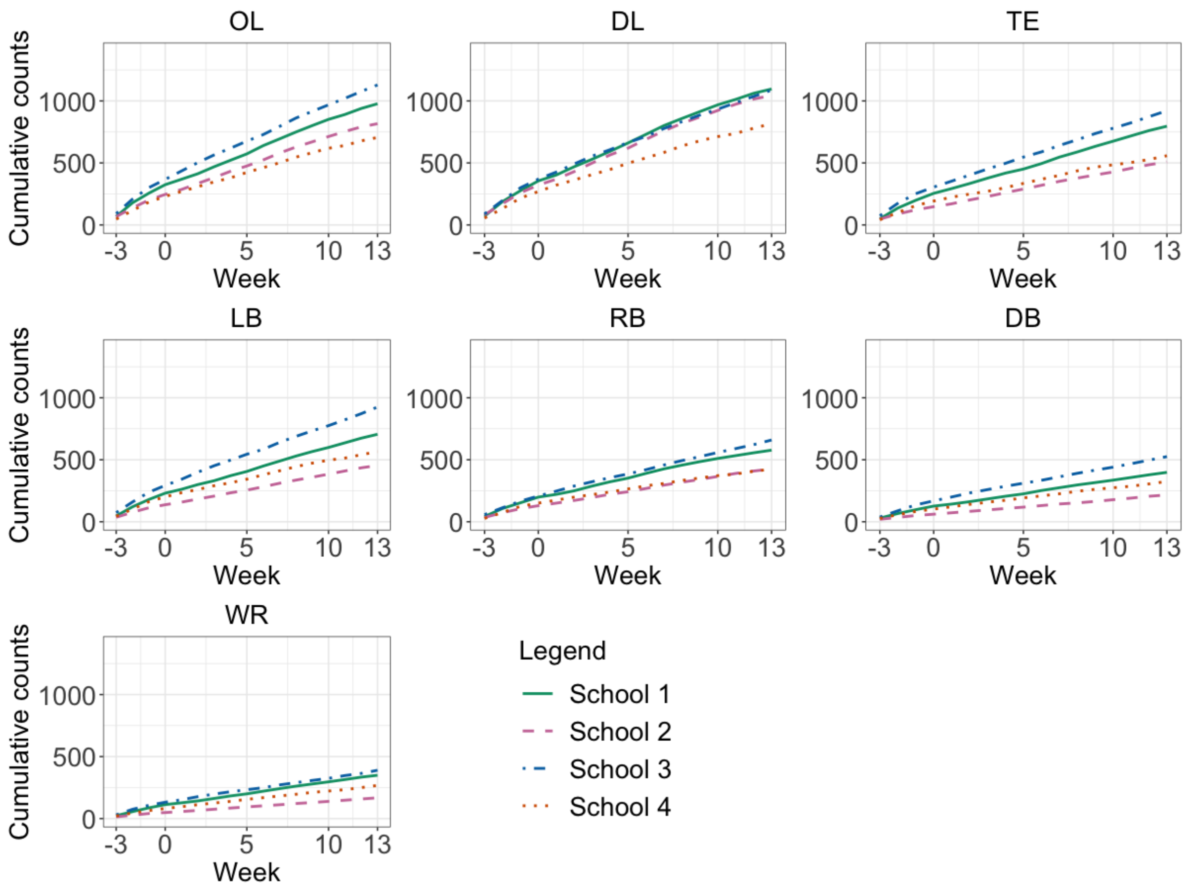


Figure 3.2: Estimated mean functions of cumulative counts of HAEs grouped by positions.

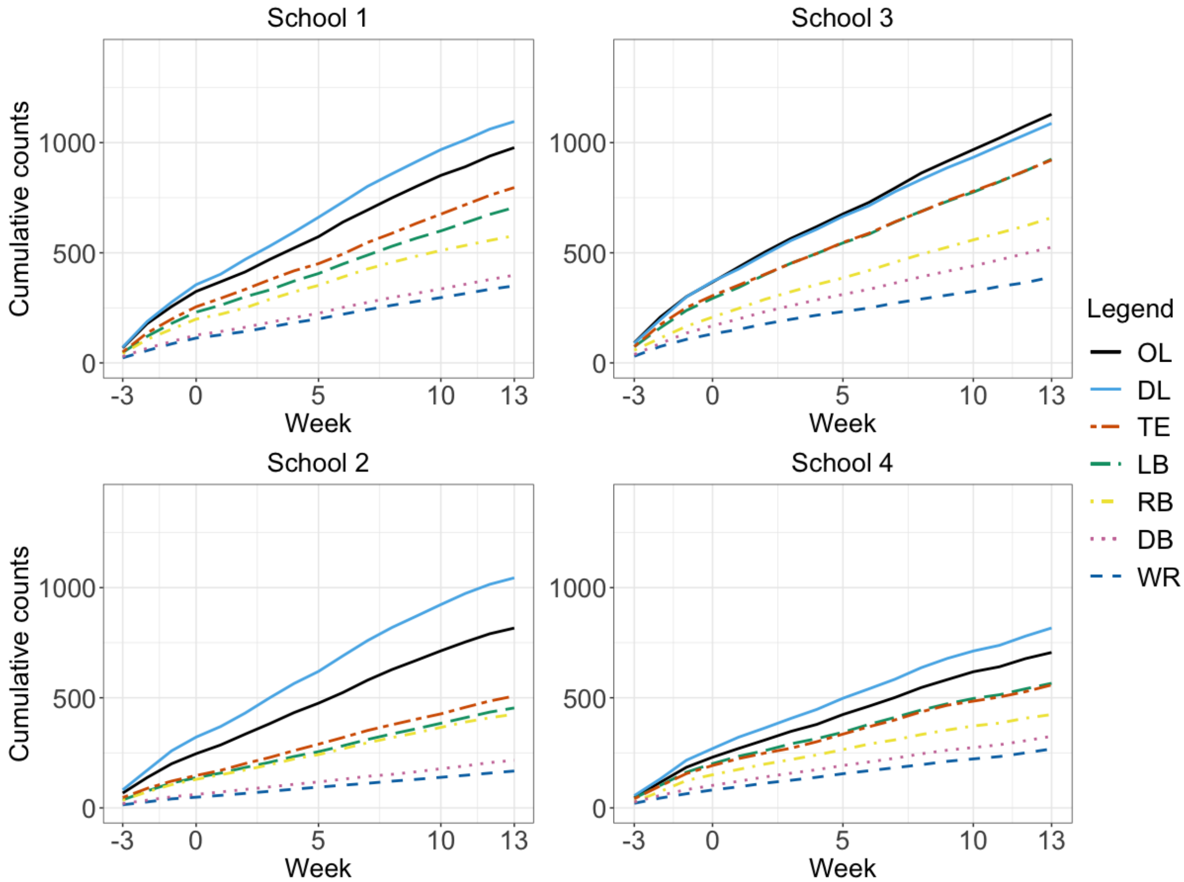


Figure 3.3: Estimated mean functions of cumulative counts of HAEs grouped by schools.

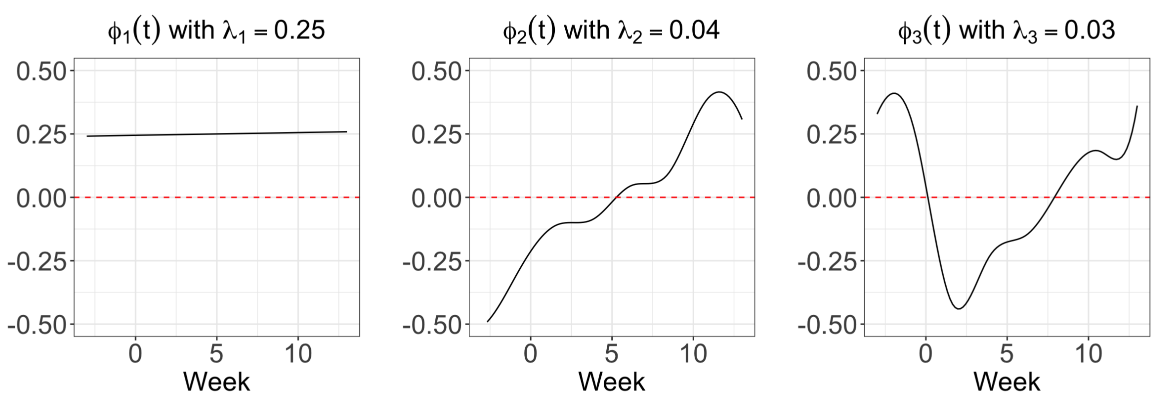


Figure 3.4: Estimated eigenfunctions with associated top 3 eigenvalues from HAEs data.

in practice or competition, as in the HAEs data. The mean functions are derived from the estimated mean functions. In the data application, the covariance function includes 7 principal components; here, we only use the first three principal components with the largest eigenvalues. The three eigenfunctions are the estimates presented in Figure 3.4. The random scores $\xi_{i\ell}$ are generated from a normal distribution $N(0, \lambda_\ell)$, where $\lambda_1 = 0.25$, $\lambda_2 = 0.04$ and $\lambda_3 = 0.03$. Once $\eta_i(T_{ij})$ is obtained, we generate one Poisson observation with mean $\eta_i(T_{ij})$ as our longitudinal outcome. We simulate 100 replications.

To quantify the estimation accuracy of the mean and covariance functions, we use the relative integrated squared error (RISE) as follows:

$$\mathrm{RISE}_{\eta_0} = \frac{\int_0^1 \left\{\eta_0(t) - \widehat{\eta}_0(t)\right\}^2 dt}{\int_0^1 \left\{\eta_0(t)\right\}^2 dt},$$

$$\mathrm{RISE}_{\mathscr{C}} = \frac{\int_0^1 \int_0^1 \left\{\mathscr{C}(s,t) - \widehat{\mathscr{C}}(s,t)\right\}^2 ds\,dt}{\int_0^1 \int_0^1 \{\mathscr{C}(s,t)\}^2 ds\,dt},$$

where η_0 and $\mathscr{C}$ represent the true mean and covariance functions, and $\widehat{\eta}_0$ and $\widehat{\mathscr{C}}$ denote their estimates, respectively.

3.5.2 Simulation Results

First, we fix the the number of principal functions and the number of eigenfunctions at the truth values ($R = 2$, and $L = 3$) and estimate the model components. Figure 3.5 displays boxplots of the relative integrated squared error (RISE) for all 28 mean functions, $\eta_{0sp}(t)$, across different schools and positions, as well as for the covariance function, $\mathscr{C}(s,t)$. Figure 3.6a presents boxplots of estimated eigenvalues, while Figure 3.6b shows the estimated eigenfunctions. The RISEs for estimating both the mean and covariance function are low, and the estimated eigenvalues seem close to the true values, and the estimated eigenfunctions are overall similar to the truth ones, all demonstrating the estimation accuracy of the proposed model estimation.

Next, we apply AIC and BIC to select the number of principal functions and the number of eigenfunctions in the covariance function, given that the true values $R = 2$ and $L = 3$. We consider candidate values for R in {1,2,3,4} and for L in {1,2,3,4,5}. The correct selection rates are 0.66 for AIC and 1.00 for BIC. We find that AIC tends to overselect the number of principal functions and the number of eigenfunctions when it selects incorrectly the values of R and L. Therefore, BIC is recommended for model selection.

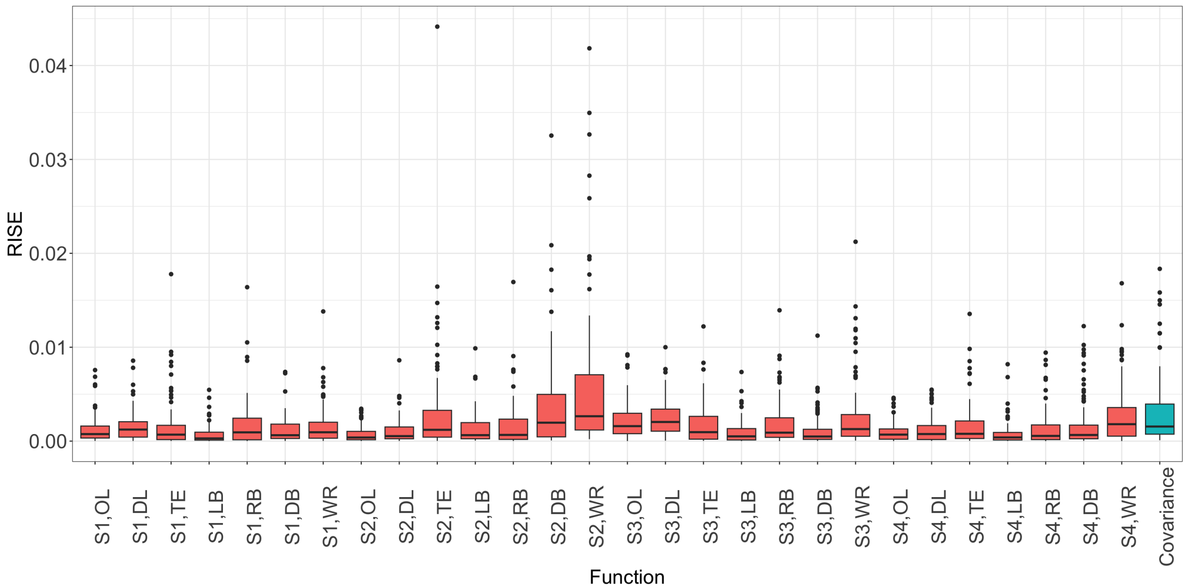


Figure 3.5: Boxplots of RISEs of estimates of 28 mean functions $\eta_{0sp}(t)$ for each combination of school and position and covariance function $\mathscr{C}(s,t)$ across 100 replicates.

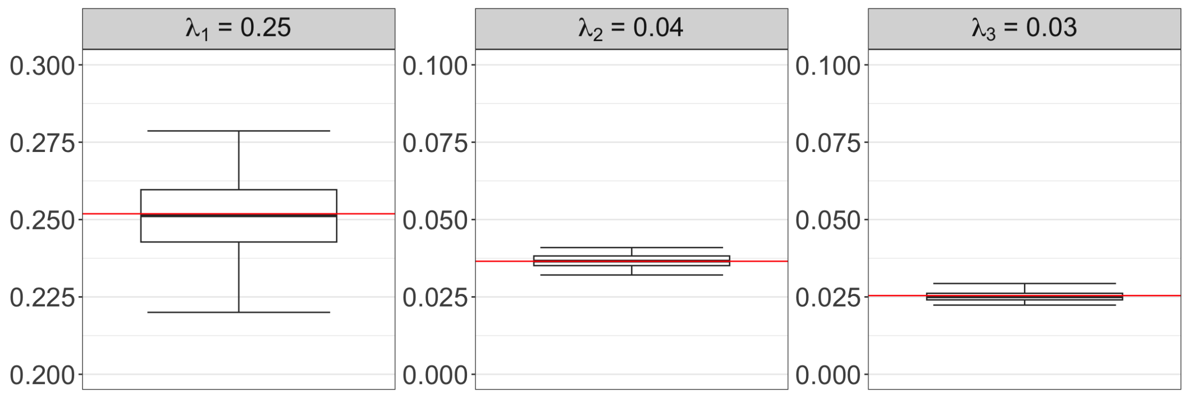
3.6 Discussion

In this paper, we applied non-Gaussian functional data methods to analyze the head impact exposure data collected from college football athletes. The Tucker decomposition of the collection of mean functions led to an interpretable and stable estimation of mean functions. We developed an effective estimation algorithm that integrates with EM the iterative estimation of Tucker decomposition and fast estimation of nonparametric principal functions via penalized splines.

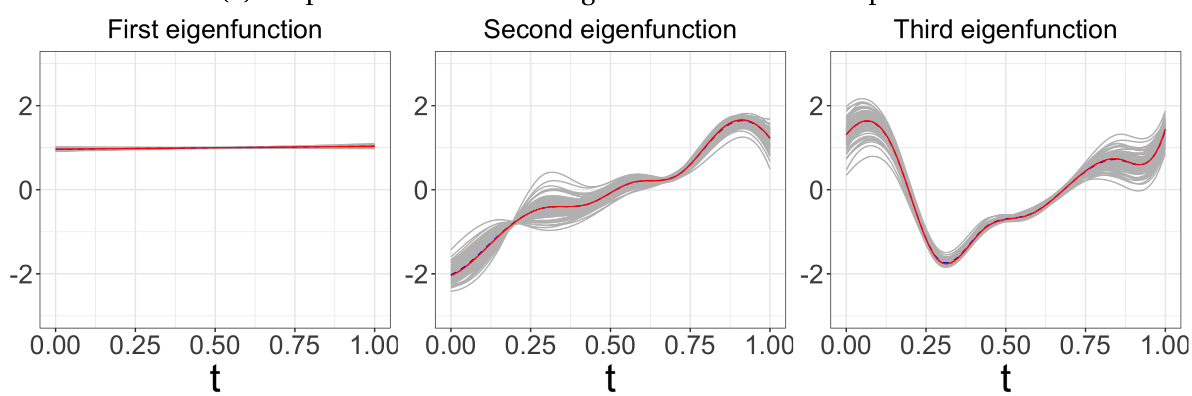
The application of the proposed method to head impact exposure data expands our previous findings, indicating a relatively higher exposure during the preseason compared to the exposure during the fall seasons (see (McCrea et al. 2021)). Furthermore, we also address our claim that cumulative exposure to head impacts increases the risk of concussion in college football athletes that we have studied previously (see (Stemper et al. 2019)).

In this analysis, we only considered the number of HAEs without taking into account the magnitude of each HAE defined as maximum linear acceleration, which varied substantially from 10-g to beyond 100-g and could potentially have different effects. Thus, it would be of interest to incorporate the magnitude of HAE into the analysis. For such a more refined analysis, a more complex method such as a joint modeling approach is likely useful and is left for a further research direction.

Some athletes in the data experienced concussion during practices or games and hence



(a) Boxplots of estimates of eigenvalues across 100 replicates.



(b) Estimated eigenfunctions across 100 replications. Gray lines: estimates from 100 replications; dashed blue lines: mean functions of gray lines; solid red lines: true eigenfunctions.

Figure 3.6: Plots of estimated eigenvalues and eigenfunctions across 100 replications.

might be absent from later games for treatment, which resulted in missing data. While only a small proportion of athletes experienced concussions which likely won't change the results substantially, nevertheless it is of future interest to model concussion as event data and extend the current model.

CHAPTER

4

FUNCTIONAL JOINT MODEL FOR LONGITUDINAL AND SURVIVAL DATA WITH A FUNCTIONAL COVARIATE WITH APPLICATION TO A LONGITUDINAL STUDY OF OBJECTIVE PHYSICAL ACTIVITY

4.1 Introduction

4.1.1 OPACH data

[1] The Women's Health Initiative (WHI) program investigated the most common causes of morbidity and mortality among 161,808 postmenopausal women, ages 50-79 years who were initially enrolled at 40 clinical sites in the United States between 1993-1998 (Group et al. 1998). The WHI Long Life Study was a part of the second WHI extension study (2010-2015) in which

[1]This chapter is joint work with Luo Xiao, Ilsuk Kang and Chongzhi Di.

in-home examinations were completed among a sub-cohort of 7,875. The Objective Physical Activity and Cardiovascular Health (OPACH) Study was an accelerometry ancillary cross-sectional study to the WHI Long Life Study conducted between 2012 and 2014 (LaCroix et al. 2017) that recorded hip-worn accelerometry on 6,489 older women (63-99 years old).

4.1.2 Accelerometry-based Sitting Bout Profiles

Accelerometers have been widely utilized in physical activity epidemiological studies (Yang and Hsu 2010). Device manufacturers often provide software (e.g., ActiLife) that transform raw acceleration signals into summarized metrics called counts per pre-specified time intervals or epochs using their proprietary algorithm. Specifically for identifying sedentary behavior, a common approach is to apply a cutpoint to epoch level counts data (e.g., < 100 counts per minute) and calculate daily or weekly summary metrics (e.g., total sedentary time). A sedentary bout is identified as a period of time in continuous, uninterrupted sedentary behavior. However, it has been recognized that the standard cutpoint-based method described above might overestimate sitting time. Recently, a deep learning algorithm was designed to focus specifically on sitting, called the convolutional neural network hip accelerometer posture (CHAP) classification algorithm (Greenwood-Hickman et al. 2021; Eanes 2018), which has been shown to substantially improve accuracy of predicting sitting. The CHAP is unique in that it utilizes two competitive architectures in the deep learning community, convolutional neural network (CNN) and bi-directional long short-term memory network (BiLSTM), for predicting sitting in 10-second epochs.

Once sitting/non-sitting is predicted for each epoch, a natural next step is to explore sitting bout accumulation patterns and investigate their relationships with health outcomes. Reporting results based on simple summary metrics of sedentary behavior accumulation patterns (e.g., mean/median bout duration) might omit features in bout accumulation profiles that are predictive of health outcomes. To address these challenges, we adopt a different approach that view the entire distribution of sitting bout durations from a participant as a whole entity (profile), which is functional data. We applied the CHAP algorithm to raw accelerometry data to predict sitting behaviors in each 10-second epochs and obtain binary sequence of 10s epochs. We then transformed this sequence into a sitting bout accumulation profile by computing total times of sitting bouts of different lengths per day averaged across multiple days. Our objective is to examine the association of daily sitting profiles with physical function and mortality.

Figure 4.1 shows the longitudinal trajectories of physical function for two subjects over time. The trajectories highlight substantial differences in baseline levels and patterns across individuals. Figure 4.2 displays the corresponding daily sitting profiles for the same two subjects, illustrating the distribution of total sitting time across various sitting bout lengths. Noticeably, subject 2 has not only overall longer total sitting time but also longer sitting bouts compared to subject 1, which agrees with the subject's much lower values of physical functioning compared with those of subject 1.

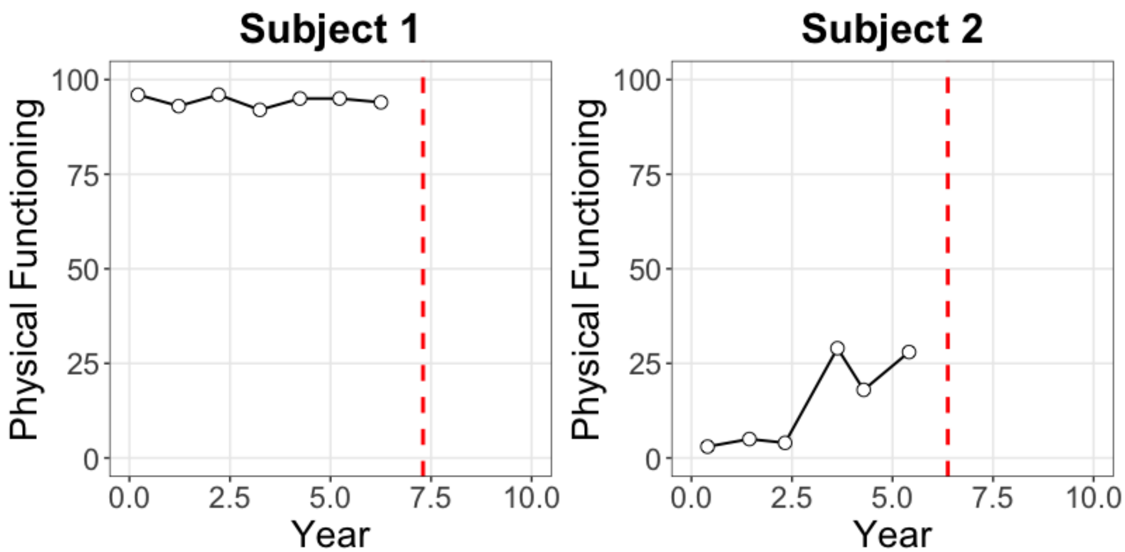


Figure 4.1: Observed SF-36 physical function over time for two subjects. Each line connects repeated measures of physical function, with the red dashed vertical line indicating the time of all-cause mortality.

4.1.3 Existing Literature

Joint models (JMs) have been widely used for the joint analysis of longitudinal and survival data (Wulfsohn and Tsiatis 1997; Tsiatis and Davidian 2004; Lin et al. 2002; Murray and Philipson 2023; He et al. 2024). A major limitation of traditional JMs lies in their reliance on parametric assumptions for the longitudinal sub-model, which can be inadequate for capturing complex and nonlinear subject-specific trajectories. To address this, an alternative framework models longitudinal trajectories as sparse functional data (James et al. 2000a; Yao et al. 2005), utilizing flexible nonparametric techniques (Ramsay and Silverman 2005) to better accommodate individual variation. This approach has led to the development of functional joint models (FJMs), initially introduced by Yao (2007) and Yan et al. (2017). FJMs have since been extended to a wide range of applications, including multivariate sparse functional data (Li and Luo 2019; Li et al. 2022; Zou et al. 2023; He et al. 2024), multi-dimensional functional data (Shi et al. 2024),

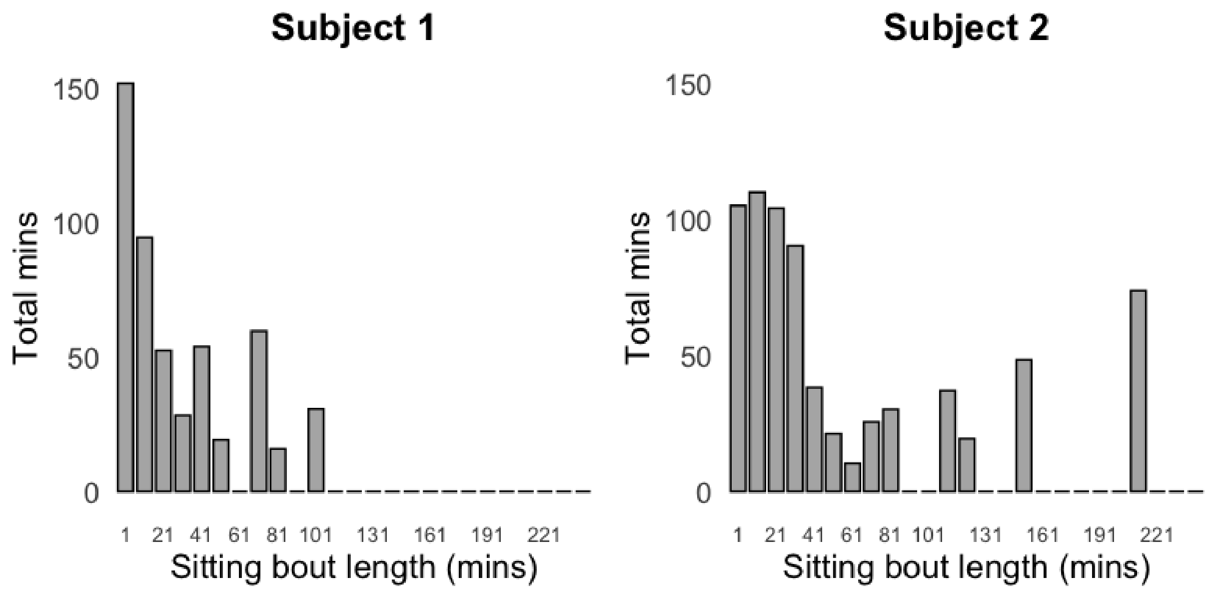


Figure 4.2: Distribution of total sitting time across different bout lengths for two subjects.

recurrent event data (Hong et al. 2021), and imaging data (Zou et al. 2024).

Despite these methodological advances, estimation in FJMs remains a core challenge. Many existing methods rely on two-stage estimation procedures for computational feasibility, often resulting in substantial model bias. To address this, Li et al. (2022) estimated FJM via an expectation-maximization (EM) algorithm, which helps reduce bias by jointly modeling and estimating all components in longitudinal sub-model and survival sub-model. However, their use of regression splines to estimate smooth functions can lead to overfitting, which in turn may degrade estimation accuracy and slow the convergence of the EM algorithm. Wang et al. (2025) propose an EM-based estimation procedure that incorporates penalized splines (Eilers and Marx 1996b) for modeling smooth functions. This approach offers greater flexibility in capturing nonlinear effects while effectively mitigating overfitting, thereby improving both estimation stability and computational efficiency.

To clarify, joint modeling here refers to the situation where the domain of function is longitudinal time. In contrast, a large body of work has focused on Cox regression models where functional data are incorporated as baseline covariates. For instance, Gellar et al. (2015), Qu et al. (2016), and Kong et al. (2018) studied Cox models to characterize the association between time-to-event outcomes and sets of functional and scalar predictors, developing likelihood-based estimation methods. Similarly, Lee et al. (2015) proposed a Bayesian functional Cox regression framework with both functional and scalar covariates, introducing alternative regularization techniques. Beyond Cox models, Li and Luo (2017) developed a joint functional regression model to simultaneously model longitudinal and survival outcomes, adjusting for high-dimensional images and low-dimensional covariates using functional principal component analysis (FPCA). Building on this, Wang et al. (2020) introduced a functional partial least squares (FPLS) approach for more efficient estimation within this joint modeling framework, demonstrating improved accuracy and robustness compared to FPCA in various scenarios. More recently, Cui et al. (2021) proposed an additive functional Cox model to flexibly capture the effects of functional covariates on survival outcomes. In the context of high-dimensional imaging data, Jiang et al. (2023) proposed a supervised FPCA (sFPCA) method for feature extraction prioritized by association with survival, while Jiang et al. (2024) further developed a Cox model using FPLS for such predictors. Beyond the Cox framework, Ghosal et al. (2025b) introduced a more general functional time transformation model for survival analysis with both scalar and functional covariates, and Ghosal et al. (2025a) proposed distributional outcome regression via quantile functions (DORQF) to accommodate scalar and distributional predictors.

4.1.4 Proposed Method

Although both functional joint models (FJMs) and Cox models with functional covariates have been extensively studied, relatively few works have addressed FJMs that incorporate baseline functional covariates. In this paper, we propose a FJM framework for jointly analyzing longitudinal outcomes (physical function measured via survey) and time-to-event outcomes (all-cause mortality), while incorporating baseline scalar covariates and a baseline functional covariate (daily sitting profiles). Both the longitudinal and survival sub-models include linear scalar-on-function regression terms to capture the effect of the baseline functional covariate, allowing us to simultaneously evaluate its influence on both outcome types.

To estimate the proposed model efficiently and robustly, we develop a computationally feasible expectation-maximization (EM) algorithm. We use penalized splines to estimate the nonparametric smooth functions in the longitudinal sub-model as well as the coefficient functions for the baseline functional covariate in both sub-models. This approach flexibly models nonlinear relationships while mitigating the overfitting risks commonly associated with regression splines. A key computational challenge lies in selecting smoothing parameters for multiple nonparametric components. To address this, we adopt a local selection strategy that exploits the EM algorithm's iterative framework. Specifically, for the longitudinal sub-model, smooth function estimation is reformulated as a weighted least squares problem for nonparametric regression and solved using the *mgcv* (Wood 2011; Wood et al. 2016) R package at each EM iteration. In contrast, smoothing parameter selection in the survival sub-model is more delicate.

We address this by applying a local model selection procedure based on the Akaike Information Criterion (AIC), which evaluates the expected survival log-likelihood and incorporates an effective degrees of freedom penalty derived from the observed Fisher information matrix. A predefined grid search identifies the optimal smoothing level at each iteration. This unified and adaptive approach enhances stability and accelerates EM convergence, while maintaining the model's flexibility to accommodate complex data structures.

The remainder of the paper is structured as follows. Section 4.2 introduces the proposed functional joint model that incorporates a baseline functional covariate. Section 4.3 describes the model estimation procedure based on the Monte Carlo Expectation-Maximization (EM) algorithm. Section 4.4 discusses the selection of key model parameters, including the number of principal components and smoothing parameters. In Section 4.5, we apply the proposed method to the OPACH dataset to illustrate its ability to jointly model longitudinal outcomes, survival outcomes, and baseline scalar and functional covariates. Section 4.6 assesses the model's performance through simulation studies. Finally, Section 4.7 concludes with a discussion of the findings and potential future directions.

4.2 Functional Joint Model with Baseline Functional Covariate

4.2.1 Data Structure and Notation

Let n be the number of subjects, indexed by i. For subject i, let Y_{ij} denote the jth observation at longitudinal time t_{ij}, with $1 \le j \le m_i$ and m_i the number of observations. Observational times t_{ij} are assumed to lie within a compact interval $\mathscr{T}$, representing the study follow-up period. The survival outcome, representing the time from baseline to the event of interest, is denoted by S_i for subject i. When S_i is right-censored, we observe $T_i = \min(S_i, C_i)$, where C_i is the censoring time, assumed independent of both the event time S_i and the longitudinal outcomes. The binary event indicator $\Delta_i = 1_{\{S_i \le C_i\}}$ specifies whether S_i is observed. Longitudinal observations t_{ij} are restricted to $t_{ij} \in [0, T_i] \subset \mathscr{T}$, implying no observations of longitudinal outcomes after T_i. Finally, let $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iP}) \in \mathbb{R}^P$ denote a common set of baseline scalar covariates and let $W_i(s), s \in \mathscr{S}$ denote the baseline functional covariate, where $\mathscr{S}$ is a compact interval.

4.2.2 Longitudinal Model with a Baseline Functional Covariate

We model the longitudinal outcome as functional data and consider the following functional linear regression model:

$$Y_{ij} = \mu(t_{ij}) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_1 + \int_{s\in\mathscr{S}} W_i(s)\beta_1(s)\mathrm{d}s + X_i(t_{ij}) + \epsilon_{ij}\,, \tag{4.1}$$

where $\mu(t)$ is the fixed mean function, $\boldsymbol{\gamma}_1$ is the coefficient vector for the baseline scalar covariates $\boldsymbol{Z}_i$, $\beta_1(s)$ is the smooth coefficient function for the baseline functional covariate $W_i(s)$, $X_i(t)$ is the random subject-specific smooth deviation, and ϵ_{ij} is measurement error. It is assumed that $X_i(t)$ and ϵ_{ij} are mutually independent.

The subject-specific variations are captured by $X_i(t)$, modeled as a Gaussian process with zero mean and covariance $\mathscr{C}(t,t') = \mathrm{Cov}\{X_i(t), X_i(t')\}$. The covariance function is decomposed using eigendecomposition as $\mathscr{C}(t,t') = \sum_{\ell\geq 1}\lambda_\ell\phi_\ell(t)\phi_\ell(t')$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the eigenvalues, and $\phi_\ell(t)$ are the associated orthonormal eigenfunctions, satisfying $\int_{\mathscr{T}}\phi_\ell(t)\phi_{\ell'}(t)dt = 1_{\{\ell=\ell'\}}$. To facilitate estimation and interpretation, we assume that $X_i(t)$ can be represented by a finite number of principal components. Specifically, $X_i(t) = \sum_{\ell=1}^{L}\xi_{i\ell}\phi_\ell(t)$ with FPCA scores $\xi_{i\ell} \sim N(0,\lambda_\ell)$. Let $\boldsymbol{\xi}_i = (\xi_{i1},\ldots,\xi_{iL})^\top$ be the vector of FPCA scores corresponding to subject i, with a diagonal covariance matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1,\ldots,\lambda_L)$. Here, the parameter L represents the number of principal components retained for the subject-specific process and is selected adaptively using data-driven criteria such as cross-validation or AIC/BIC. Finally, the measurement errors ϵ_{ij} are assumed to be independently distributed as $N(0,\sigma^2)$. Given the eigenfunctions $\phi_\ell(t)$, Model (4.1) can be expressed in the mixed model representation:

$$Y_{ij} = \mu(t_{ij}) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_1 + \int_{s\in\mathscr{S}} W_i(s)\beta_1(s)ds + \sum_{\ell=1}^{L}\xi_{i\ell}\phi_\ell(t_{ij}) + \epsilon_{ij}\,.$$

4.2.3 Joint Model for Longitudinal and Survival Data

Let $h_i(t)$ be the hazard function for the ith subject. We consider the following proportional hazard model,

$$h_i(t) = h_0(t)\exp\left\{\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2 + \int_{s\in\mathscr{S}} W_i(s)\beta_2(s)\mathrm{d}s + \boldsymbol{\xi}_i^\top\boldsymbol{\gamma}_3\right\}, \tag{4.2}$$

where $h_0(t)$ represents the baseline hazard function, which may be specified parametrically or estimated non-parametrically, $\boldsymbol{\gamma}_2$ is the coefficient vector for the baseline scalar covariates $\boldsymbol{Z}_i$, $\beta_2(s)$ is the smooth coefficient function for the baseline functional covariate $W_i(s)$, $\boldsymbol{\gamma}_3 = (\gamma_{31},\ldots,\gamma_{3L})^\top$ is the coefficient vector associated with the FPCA scores $\boldsymbol{\xi}_i$ linking longitudinal and survival outcomes.

4.2.4 Likelihood of Joint Model

Let $\boldsymbol{y}_i = (Y_{i1},\ldots,Y_{im_i})^\top$, $\boldsymbol{\mu}_i = (\mu(t_{i1}),\ldots,\mu(t_{im_i}))^\top$, and $\boldsymbol{F}_i = \left(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_1 + \int_{s\in\mathscr{S}} W_i(s)\beta_1(s)\mathrm{d}s\right)\mathbf{1}_{m_i}$ be vectors of observations, mean function at the observed time points, and baseline covariates for the ith subject, respectively. Define $\boldsymbol{\phi}_{\ell,i} = \left(\phi_\ell(t_{i1}),\ldots,\phi_\ell(t_{im_i})\right)^\top$ and $\boldsymbol{\Phi}_i = [\boldsymbol{\phi}_{1,i},\cdots,\boldsymbol{\phi}_{L,i}]$ as the eigenfunctions for the i-th subject. The noiseless observations for the ith subject are given by $\boldsymbol{x}_i = \boldsymbol{\mu}_i + \boldsymbol{F}_i + \boldsymbol{\Phi}_i\xi_i$. Finally, denote the covariance matrix of measurement errors by $\boldsymbol{\Sigma}_i = \sigma^2\boldsymbol{I}_{m_i}$. The conditional likelihood of the longitudinal data given ξ_i is,

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Sigma}_i) = (|2\pi\boldsymbol{\Sigma}_i|)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{x}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{y}_i-\boldsymbol{x}_i)\right\}.$$

The FPCA random scores ξ_i are assumed to follow multivariate normal distributions: $f(\xi_i|\boldsymbol{\Lambda}) = (|2\pi\boldsymbol{\Lambda}|)^{-\frac{1}{2}}\exp(-\frac{1}{2}\xi_i^\top\boldsymbol{\Lambda}^{-1}\xi_i)$.

The conditional partial likelihood of the time-to-event data is given by

$$\begin{aligned} f(T_i,\Delta_i|h_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\boldsymbol{\gamma}_2,\beta_2(s),\boldsymbol{\gamma}_3) = &\left\{h_0(T_i)\exp(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2 + \int_{s\in\mathscr{S}} W_i(s)\beta_2(s)\mathrm{d}s + \xi_i^\top\boldsymbol{\gamma}_3)\right\}^{\Delta_i} \\ &\times\exp\left\{-\Lambda_0(T_i)\exp(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2 + \int_{s\in\mathscr{S}} W_i(s)\beta_2(s)\mathrm{d}s + \xi_i^\top\boldsymbol{\gamma}_3)\right\}, \end{aligned} \tag{4.3}$$

where $\Lambda_0(t) = \int_0^t h_0(u)\mathrm{d}u$ is the baseline cumulative hazard function.

The overall marginal likelihood is obtained by integrating the product of the longitudinal and survival likelihoods over the distributions of the random score vector ξ_i. The log marginal likelihood is

$$\sum_{i=1}^n \log\left\{\int f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Sigma}_i)f(T_i,\Delta_i|h_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\boldsymbol{\gamma}_2,\beta_2(s),\boldsymbol{\gamma}_3)f(\xi_i|\boldsymbol{\Lambda})\mathrm{d}\xi_i\right\}. \tag{4.4}$$

4.3 Model Estimation via Monte Carlo EM

4.3.1 Spline Approximation of Smooth Functions

Estimating the smooth mean function, covariance functions and coefficient functions for the baseline functional covariate in the longitudinal model and survival model requires a flexible yet computationally efficient approach. We employ spline approximation, which provides these properties. Let $\boldsymbol{b}(t) = \{B_1(t),\cdots,B_K(t)\}^\top$ be the vector of B-spline basis functions on $\mathscr{T}$ and $\boldsymbol{b}^\beta(s) = \{B_1^\beta(s),\cdots,B_{K^\beta}^\beta(s)\}^\top$ be the vector of B-spline basis functions on $\mathscr{S}$, where K and

K^{β} are the numbers of equally-spaced interior knots plus the order (degree + 1) of the B-splines on $\mathscr{T}$ and $\mathscr{S}$, respectively. The mean function $\mu(t)$ is modeled as $\mu(t) = \boldsymbol{b}(t)^{\top}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the corresponding coefficient vector. The smooth coefficient functions for the baseline functional covariate $W_i(s)$ in longitudinal model and survival model are approximated as $\beta_1(s) = \boldsymbol{b}^{\beta}(s)^{\top}\boldsymbol{\omega}_1$ and $\beta_2(s) = \boldsymbol{b}^{\beta}(s)^{\top}\boldsymbol{\omega}_2$, where $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$ are the corresponding coefficient vectors.

To ensure numerical stability, we orthonormalize the B-spline bases on $\mathscr{T}$ using the Gram matrix: $\boldsymbol{G} = \int \boldsymbol{b}(t)\boldsymbol{b}(t)^{\top}dt \in \mathbb{R}^{K\times K}$. The orthonormal B-spline bases are given by: $\widetilde{\boldsymbol{b}}(t) = \boldsymbol{G}^{-1/2}\boldsymbol{b}(t)$. The covariance function $\mathscr{C}(t,t')$ is decomposed into its ℓth eigenfunction, which is approximated: $\phi_{\ell}(t) = \widetilde{\boldsymbol{b}}(t)^{\top}\boldsymbol{\theta}_{\ell}$, where $\boldsymbol{\theta}_{\ell}$ is coefficient vector for the ℓ-th eigenfunction of $\mathscr{C}$. The orthonormality of the eigenfunctions imposes the constraints: $\boldsymbol{\theta}_{\ell_1}^{\top}\boldsymbol{\theta}_{\ell_2} = 1_{\{\ell_1=\ell_2\}}$. These orthonormality constraints ensure the eigenfunctions are linearly independent, and form a valid basis for the decomposition of the covariance structures. This finite-dimensional approximation reduces model complexity while retaining flexibility, facilitating efficient parameter estimation in the subsequent EM algorithm steps.

4.3.2 E-step

The spline approximation of the nonparametric smooth functions enables the use of parametric estimation methods for the functional joint model. Since direct maximization of the log marginal likelihood in Model (4.4) is computationally expensive, we employ the EM algorithm (Dempster et al. 1977), which treats $\boldsymbol{\xi}_i$ as latent (missing) data. The EM algorithm iteratively alternates between two steps until convergence: (1) computing the expectation of the conditional log-likelihood of the observed data given the current parameter estimates (E-step), and (2) maximizing the expected log-likelihood to update the parameter estimates (M-step).

Let the observed data for the ith subject be $\mathbb{Y}_i = \{\boldsymbol{y}_i, \boldsymbol{Z}_i, W_i(s), T_i, \Delta_i\}$. Let $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_L]$. Denote the collection of parameters by $\boldsymbol{\Omega} = \{h_0, \boldsymbol{\alpha}, \boldsymbol{\gamma}_1, \boldsymbol{\omega}_1, \sigma^2, \boldsymbol{\gamma}_2, \boldsymbol{\omega}_2, \boldsymbol{\gamma}_3, \boldsymbol{\Lambda}, \boldsymbol{\Theta}\}$. The E-step calculates the expected value of the conditional log-likelihood given the observed data and current parameter estimates, $\widehat{\boldsymbol{\Omega}}$. This is expressed as:

$$
\begin{aligned}
Q(\boldsymbol{\Omega}) = &\sum_{i=1}^{n}\mathbb{E}_i\Big\{\log\big(f(\boldsymbol{y}_i|\boldsymbol{x}_i,\boldsymbol{\Sigma}_i)\big)\Big\} + \sum_{i=1}^{n}\mathbb{E}_i\Big\{\log\big(f(T_i,\Delta_i|h_0,\boldsymbol{Z}_i,W_i(s),\boldsymbol{\xi}_i,\boldsymbol{\gamma}_2,\boldsymbol{\omega}_2,\boldsymbol{\gamma}_3)\big)\Big\} \\
&+ \sum_{i=1}^{n}\mathbb{E}_i\Big\{\log\big(f(\boldsymbol{\xi}_i|\boldsymbol{\Lambda})\big)\Big\},
\end{aligned} \tag{4.5}
$$

where $\mathbb{E}_i(\cdot)$ is the expectation with respect to the conditional distribution of $\boldsymbol{\xi}_i$ given the observed data $\mathbb{Y}_i$ and current parameter estimates $\widehat{\boldsymbol{\Omega}}$.

Let $g(\cdot)$ be any smooth function of $\boldsymbol{\xi}_i$. The conditional expectation $\mathbb{E}_i\{g(\boldsymbol{\xi}_i)\}$ is given by:

$$\frac{\int g(\boldsymbol{\xi}_i) f(T_i, \Delta_i | \boldsymbol{Z}_i, W_i(s), \boldsymbol{\xi}_i, \widehat{\boldsymbol{\Omega}}) f(\boldsymbol{\xi}_i | \boldsymbol{y}_i, \boldsymbol{Z}_i, W_i(s), \widehat{\boldsymbol{\Omega}}) \mathrm{d}\boldsymbol{\xi}_i}{\int f(T_i, \Delta_i | \boldsymbol{Z}_i, W_i(s), \boldsymbol{\xi}_i, \widehat{\boldsymbol{\Omega}}) f(\boldsymbol{\xi}_i | \boldsymbol{y}_i, \boldsymbol{Z}_i, W_i(s), \widehat{\boldsymbol{\Omega}}) \mathrm{d}\boldsymbol{\xi}_i},$$

where $f(T_i, \Delta_i | \boldsymbol{Z}_i, W_i(s), \boldsymbol{\xi}_i, \widehat{\boldsymbol{\Omega}})$ corresponds to the survival likelihood defined in Model (4.3), and

$$f(\boldsymbol{\xi}_i | \boldsymbol{y}_i, \boldsymbol{Z}_i, W_i(s), \widehat{\boldsymbol{\Omega}}) \propto f(\boldsymbol{y}_i | \boldsymbol{x}_i, \widehat{\boldsymbol{\Omega}}) f(\boldsymbol{\xi}_i | \widehat{\boldsymbol{\Lambda}}).$$

The conditional distribution $f(\boldsymbol{\xi}_i | \boldsymbol{y}_i, \boldsymbol{Z}_i, W_i(s), \widehat{\boldsymbol{\Omega}})$ is multivariate normal (see Appendix C.1).

To efficiently compute $\mathbb{E}_i\{g(\boldsymbol{\eta}_i)\}$, we use Monte Carlo integration. Specifically, we draw R samples $\xi_i^{(r)}$ from the multivariate normal distribution $f(\boldsymbol{\xi}_i | \boldsymbol{y}_i, \boldsymbol{Z}_i, W_i(s), \widehat{\boldsymbol{\Omega}})$ and compute the approximation:

$$\mathbb{E}_i\left\{g(\boldsymbol{\xi}_i)\right\} \approx \frac{\sum_{r=1}^R g(\xi_i^{(r)}) f(T_i, \Delta_i | \boldsymbol{Z}_i, W_i(s), \xi_i^{(r)}, \widehat{\boldsymbol{\Omega}})}{\sum_{r=1}^R f(T_i, \Delta_i | \boldsymbol{Z}_i, W_i(s), \xi_i^{(r)}, \widehat{\boldsymbol{\Omega}})}.$$

The Monte Carlo approximation enables efficient implementation of the E-step, ensuring computational feasibility even in high-dimensional parameter spaces.

4.3.3 M-step

In the M-step, updated estimates of $\boldsymbol{\Omega}$ are obtained by maximizing the function $Q(\boldsymbol{\Omega})$ in (4.5). Specifically, estimates of the longitudinal parameters $\left\{\boldsymbol{\alpha}, \boldsymbol{\gamma}_1, \boldsymbol{\omega}_1, \boldsymbol{\sigma}^2, \boldsymbol{\Theta}\right\}$, the survival parameters $\left\{h_0, \boldsymbol{\gamma}_2, \boldsymbol{\omega}_2, \boldsymbol{\gamma}_3\right\}$, and the eigenvalue parameter $\boldsymbol{\Lambda}$ are updated iteratively. Below, we provide details on the estimation process for each set of parameters.

First, the longitudinal parameters $\left\{\boldsymbol{\alpha}, \boldsymbol{\gamma}_1, \boldsymbol{\omega}_1, \boldsymbol{\sigma}^2, \boldsymbol{\Theta}\right\}$ are estimated iteratively by minimizing $-\sum_{i=1}^n \mathbb{E}_i\{\log(f(\boldsymbol{y}_i | \boldsymbol{x}_i, \boldsymbol{\Sigma}_i))\}$, expressed as:

$$\sum_{i=1}^n \left\{ \frac{m_i}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \mathbb{E}_i \left\| \boldsymbol{y}_i - \boldsymbol{B}_i \boldsymbol{\alpha} - \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_1 \mathbf{1}_{m_i} - \Big(\int_{s \in \mathscr{S}} W_i(s) \boldsymbol{b}^\beta(s)^\top \boldsymbol{\omega}_1 \mathrm{d}s \Big) \mathbf{1}_{m_i} - \widetilde{\boldsymbol{B}}_i \boldsymbol{\Theta} \xi_i \right\|^2 \right\}, \quad (4.6)$$

where $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{B}_i = [\boldsymbol{b}(t_{i1}), \ldots, \boldsymbol{b}(t_{im_i})]^\top$ is the matrix of B-spline basis values evaluated at time points $t_{i1}, \ldots, t_{im_i}$, and $\widetilde{\boldsymbol{B}}_i = \boldsymbol{G}^{-1/2} \boldsymbol{B}_i$ is the orthonormalized basis obtained by multiplying $\boldsymbol{B}_i$ by the inverse square root of the Gram matrix $\boldsymbol{G}$. To enforce smoothness in the estimates of the mean function, coefficient function and eigenfunctions, quadratic penalties are added to (4.6). Compared to the regression spline estimation, penalized spline estimation stabilizes the iterative updates, reduces the number of iterations required for EM convergence, and produces smoother, more interpretable estimates. However, selecting smoothing parameters for these nonparametric functions presents great computational challenges, particularly in iterative EM algorithms. To address this, we propose to use a local selection of smoothing parameters, exploiting the iterative nature of EM algorithms. Specifically, we reformulate the estimation of these smooth functions as weighted least squares problems for nonparametric regression. This reformulation enables efficient simultaneous estimation of smooth functions and smoothing parameters using the well-developed *mgcv* (Wood 2011; Wood et al. 2016) R package. Moreover, (4.6) involves integrals of the baseline functional covariate and coefficient function, adding more challenges to our estimation. Fortunately, *mgcv* package allows the response to depend on linear functionals of smooth terms, i.e., the response can depend on the weighted sum of the same smooth function evaluated at different values, which allows for the response to depend on integrals of smooth function by approximating the integrals. Without loss of generality, we assume $\mathscr{S} = [0,1]$. We approximate $\int_{s\in\mathscr{S}} W_i(s)\boldsymbol{b}^{\beta}(s)^\top \boldsymbol{\omega}_1 ds$ by $\frac{1}{J}\sum_{j_s=1}^{J} W_i(s_{j_s})\boldsymbol{b}^{\beta}(s_{j_s})^\top \boldsymbol{\omega}_1$, where $\{s_1,\ldots,s_J\}$ is regular and dense grid on $[0,1]$. Now all longitudinal parameters and smoothing parameters can be estimated iteratively using *mgcv* package after approximation and reformulation, as detailed in Appendix C.3. We also present the derivative of the observed Fisher information matrix for $(\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}_1^\top, \boldsymbol{\omega}_1^\top)^\top$ in Appendix C.3. This approach stabilizes parameter estimation and substantially accelerates EM convergence. In practice, we employ seven B-spline bases constructed at equally spaced knots on both $\mathscr{T}$ and $\mathscr{S}$ $(K = K^{\beta} = 7)$ for simulations and real data. The orthonormality of eigenfunctions is also ensured by post-processing the updates.

Second, the survival parameters $\{h_0, \boldsymbol{\gamma}_2, \boldsymbol{\omega}_2, \boldsymbol{\gamma}_3\}$ in the Cox regression are estimated by minimizing $-\sum_{i=1}^n \mathbb{E}_i\{\log\big(f(T_i,\Delta_i|h_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\boldsymbol{\gamma}_2,\boldsymbol{\omega}_2,\boldsymbol{\gamma}_3)\big)\}$, expressed as:

$$\sum_{i=1}^{n}\Bigg\{-\Delta_i\Big\{\log(h_0(T_i)) + \boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2 + \int_{s\in\mathscr{S}} W_i(s)\boldsymbol{b}^{\beta}(s)^\top\boldsymbol{\omega}_2 ds + \mathbb{E}_i\big(\boldsymbol{\xi}_i^\top\boldsymbol{\gamma}_3\big)\Big\} + \Lambda_0(T_i)\mathbb{E}_i\Big\{\exp\big(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2 + \int_{s\in\mathscr{S}} W_i(s)\boldsymbol{b}^{\beta}(s)^\top\boldsymbol{\omega}_2 ds + \boldsymbol{\xi}_i^\top\boldsymbol{\gamma}_3\big)\Big\}\Bigg\}.$$

The baseline hazard $h_0(t)$ is then estimated by using a ratio of observed events to expected contributions at each time t:

$$\widehat{h}_0(t) = \frac{\sum_{i=1}^n \Delta_i 1_{\{T_i=t\}}}{\sum_{i=1}^n \mathbb{E}_i\Big\{\exp\big(\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\gamma}}_2 + \int_{s\in\mathscr{S}} W_i(s)\boldsymbol{b}^{\beta}(s)^\top\widehat{\boldsymbol{\omega}}_2 ds + \boldsymbol{\xi}_i^\top\widehat{\boldsymbol{\gamma}}_3\big)\Big\}1_{\{T_i\geq t\}}}.$$

To enforce smoothness in the estimate of the coefficient function $\beta_2(s)$ in the survival model, we also add quadratic penalties to the expected survival log-likelihood:

$$\sum_{i=1}^{n}\Big\{\Delta_i\Big\{\log(h_0(T_i))+\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2+\Big(\int_{s\in\mathscr{S}}W_i(s)\boldsymbol{b}^\beta(s)ds\Big)^\top\boldsymbol{\omega}_2+\mathbb{E}_i\big(\boldsymbol{\xi}_i^\top\boldsymbol{\gamma}_3\big)\Big\}-$$
$$\Lambda_0(T_i)\mathbb{E}_i\Big\{\exp\Big(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2+\Big(\int_{s\in\mathscr{S}}W_i(s)\boldsymbol{b}^\beta(s)ds\Big)^\top\boldsymbol{\omega}_2+\boldsymbol{\xi}_i^\top\boldsymbol{\gamma}_3\Big)\Big\}\Big\}-\tau_2^\beta\boldsymbol{\omega}_2^\top\boldsymbol{P}^\beta\boldsymbol{\omega}_2\,,$$

where $\int_{s\in\mathscr{S}}W_i(s)\boldsymbol{b}^\beta(s)ds=(\int_{s\in\mathscr{S}}W_i(s)B_1^\beta(s)ds,\ldots,\int_{s\in\mathscr{S}}W_i(s)B_{K^\beta}^\beta(s)ds)^\top$ is known, $\boldsymbol{P}^\beta$ is a penalty matrix and τ_2^β is the smoothing parameter. The parameter vector $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_2^\top,\boldsymbol{\omega}_2^\top,\boldsymbol{\gamma}_3^\top)^\top$, representing the effects of baseline scalar covariates, baseline functional covariate and latent trajectories, is updated iteratively using the Newton-Raphson algorithm. The d-th iteration is given by: $\widehat{\boldsymbol{\gamma}}_d=\widehat{\boldsymbol{\gamma}}_{d-1}+\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}}_{d-1}}^{-1}\boldsymbol{S}_{\widehat{\boldsymbol{\gamma}}_{d-1}}$, where $\boldsymbol{S}_{\widehat{\boldsymbol{\gamma}}_{d-1}}$ and $\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}}_{d-1}}$ are the score vector and observed information matrix with penalized term, respectively. By differentiating the expected survival log-likelihood with respect to $\boldsymbol{\gamma}$, the score for subject i is:

$$\boldsymbol{s}_i(\boldsymbol{\gamma})=\Delta_i\Big(\boldsymbol{Z}_i^\top,(\int_{s\in\mathscr{S}}W_i(s)\boldsymbol{b}^\beta(s)ds)^\top,\mathbb{E}_i\boldsymbol{\xi}_i^\top\Big)^\top-$$
$$\sum_{v=1}^{V}h_0(t_v)\mathbb{E}_i\Big\{\Big(\boldsymbol{Z}_i^\top,(\int_{s\in\mathscr{S}}W_i(s)\boldsymbol{b}^\beta(s)ds)^\top,\boldsymbol{\xi}_i^\top\Big)^\top\exp\Big(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_2+\Big(\int_{s\in\mathscr{S}}W_i(s)\boldsymbol{b}^\beta(s)ds\Big)^\top\boldsymbol{\omega}_2+\boldsymbol{\xi}_i^\top\boldsymbol{\gamma}_3\Big)\Big\}1_{\{T_i\geq t_v\}},$$

where t_v are the distinct observed event times. The score $\boldsymbol{S}_{\boldsymbol{\gamma}}$ and information matrix $\boldsymbol{I}_{\boldsymbol{\gamma}}$ are computed as: $\boldsymbol{S}_{\boldsymbol{\gamma}}=\sum_{i=1}^n\boldsymbol{s}_i(\boldsymbol{\gamma})-(0\cdot\mathbf{1}_P^\top,(2\tau_2^\beta\boldsymbol{P}^\beta\boldsymbol{\omega}_2)^\top,0\cdot\mathbf{1}_L^\top)^\top$ and $\boldsymbol{I}_{\boldsymbol{\gamma}}=\sum_{i=1}^n\boldsymbol{s}_i(\boldsymbol{\gamma})\boldsymbol{s}_i^\top(\boldsymbol{\gamma})-(\sum_{i=1}^n\boldsymbol{s}_i(\boldsymbol{\gamma}))(\sum_{i=1}^n\boldsymbol{s}_i^\top(\boldsymbol{\gamma}))/n+\text{Blockdiag}\{0\cdot\boldsymbol{I}_P,2\tau_2^\beta\boldsymbol{P}^\beta,0\cdot\boldsymbol{I}_L\}$.

Third, the diagonal matrix of eigenvalues $\boldsymbol{\Lambda}$ is estimated by minimizing the negative log-likelihoods $-\sum_{i=1}^n\mathbb{E}_i\{\log(f(\boldsymbol{\xi}_i|\boldsymbol{\Lambda}))\}$. This minimization simplifies to: $\sum_{i=1}^n\Big\{\log|\boldsymbol{\Lambda}|+\mathbb{E}_i\big\{\boldsymbol{\xi}_i^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\xi}_i\big\}\Big\}$. The estimator for the eigenvalues is: $\widehat{\lambda}_\ell=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_i(\xi_{i\ell}^2)$, where $\mathbb{E}_i(\xi_{i\ell}^2)$ represents the expected squared random scores of subject-specific trajectories.

4.4 Model Selection

Penalized splines (Eilers and Marx 1996b) are used to estimate the mean function and eigenfunctions in the longitudinal data model and the coefficient functions in the longitudinal and survival model, requiring the selection of appropriate smoothing parameters. These smoothing parameters in the longitudinal model are determined using Restricted Maximum Likelihood (REML) at each iteration of the EM algorithm. By reformulating function estimation as a weighted least squares problem in nonparametric regression, we enable efficient implementation using the *gam* function from the R package *mgcv* (Wood 2011; Wood et al. 2016) (see

Appendix C.3 for details). The smoothing parameter for the coefficient function in the survival model is selected locally at each iteration of the EM algorithm using Akaike information criterion (AIC), defined as $\mathrm{AIC} = -2\sum_{i=1}^{n}\mathbb{E}_i\{\log\big(f(T_i,\Delta_i|\widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3)\big)\} + 2\cdot\mathrm{tr}(\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}}}^{-1}\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}},0})$, where $\sum_{i=1}^{n}\mathbb{E}_i\{\log\big(f(T_i,\Delta_i|\widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3)\big)\}$ is the expected survival log-likelihood evaluated at the current estimates, $\mathrm{tr}(\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}}}^{-1}\boldsymbol{I}_{\widehat{\boldsymbol{\gamma}},0})$ is the effective degrees of freedom for the coefficient $\boldsymbol{\gamma}$ in the survival model, and $\boldsymbol{I}_{\boldsymbol{\gamma},0} = \sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\gamma})\boldsymbol{s}_i^{\top}(\boldsymbol{\gamma}) - (\sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\gamma}))(\sum_{i=1}^{n}\boldsymbol{s}_i^{\top}(\boldsymbol{\gamma}))/n$ is the information matrix without penalized term. A grid $\{e^{-10}, e^{-9},\ldots,e^{0},\ldots,e^{9},e^{10}\}$ multiplying the mean of the diagonal elements of $\boldsymbol{I}_{\boldsymbol{\gamma},0}$ evaluated at estimates of $\boldsymbol{\gamma}$ in previous iteration is used to identify the optimal value of the smoothing parameter τ_2^{β}. Although the smoothing parameters may vary across iterations, they stabilize rapidly as the algorithm converges, ensuring consistent estimation of smooth functions and improving computational efficiency.

The number of principal components L is also a critical tuning parameter in the model. To select it, we use the Bayesian information criteria (BIC), defined as $\mathrm{BIC} = -2\ell(\widehat{\boldsymbol{\Omega}}) + \log(N)\cdot \mathrm{df}$, where N is the number of observations, $\ell(\widehat{\boldsymbol{\Omega}})$ is the marginal log-likelihood of the data, approximated as:

$$\sum_{i=1}^{n}\left[\log\left\{f(\boldsymbol{y}_i|\boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}})\right\} + \log\left\{R^{-1}\sum_{r=1}^{R} f\left(T_i,\Delta_i \mid \widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i^{(r)},\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3\right)\right\}\right].$$

Here, $f(\boldsymbol{y}_i|\boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}})$ is the marginal density, which follows a normal distribution (see Appendix C.1). The samples $\xi_i^{(r)}$ are drawn from $f(\xi_i|\boldsymbol{y}_i,\boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}})$; see Appendix C.2 for derivation. To ensure an accurate approximation of the marginal log-likelihood, the number of samples R must be much larger than in the EM algorithm. The degrees of freedom (DOF), denoted as df, is calculated as:

$$\mathrm{df}_{\mu} + \mathrm{df}_{\beta_1} + \mathrm{df}_{\beta_2} + \left\{\sum_{\ell=1}^{L}\mathrm{df}_{\phi_\ell} - \frac{L(L-1)}{2}\right\} + 2P + 1 + L.$$

Each term in this formula corresponds to a specific model component:

- df_{μ}: DOF for estimating the mean function in longitudinal model.
- $\mathrm{df}_{\beta_1} + \mathrm{df}_{\beta_2}$: DOF for estimating the coefficient function for baseline functional covariate $W_i(s)$ in longitudinal model and Cox regression.
- $\sum_{\ell=1}^{L}\mathrm{df}_{\phi_\ell} - \frac{L(L-1)}{2}$: DOF for the covariance function's eigenpairs.
- $2P+1$: DOF for estimating coefficients for the baseline scalar covariates $\boldsymbol{Z}_i$ in longitudinal model and Cox regression and error variance in longitudinal model.

- L: Number of coefficients for FPCA scores in the Cox regression.

A grid is employed to identify the optimal values of L, balancing model complexity with goodness-of-fit.

4.5 Data Application

We apply the proposed functional joint model (FJM) to a dataset of 5,706 women that closely resembles the original OPACH data. Due to data access restrictions, we use a modified version that retains the key characteristics of the original dataset for illustrative purposes. The longitudinal outcome is a summary measure from the 36-item short form survey (SF-36) with higher values indicating stronger physical function. The event outcome is all-cause mortality. We include in the analysis the following baseline scalar covariates: race (white, black and Hispanic), age and awake time (see Table C.1 in Appendix C.4 for details). The baseline functional covariate is the daily profiles of sitting bout lengths. In particular, we can consider the daily sitting profile as point process. Assume we just look at one day of data for subject i, and we let $s_{ik}, 1 \le k \le K_i$ denote duration of kth sitting bout. Let $N_i(s) = \sum_{k=1}^{K_i} \mathbb{I}(s_{ik} \le s), s \in \mathscr{S}$ represent the number of sitting bouts whose duration is less or equal to s. Then $s\mathrm{d}N_i(s)$ is the total sitting time of bouts of duration s. We can explore the effect of sitting bout accumulation profile using the association $\int_{s\in\mathscr{S}} \beta_1(s)W_i(s)\mathrm{d}s = \int_{s\in\mathscr{S}} \beta_1(s)s\mathrm{d}N_i(s)$. If $\beta_1(s) = c$, which means the effect of different sitting bouts lengths are same, $\int_{s\in\mathscr{S}} \beta_1(s)s\mathrm{d}N_i(s) = c(\sum_{k=1}^{K_i} s_{ik})$.

The proposed FJM with the BIC criteria described in Section 4.4 selected four functional principal components for modeling the longitudinal outcome. For the longitudinal outcome, Figure 4.3 and Figure 4.4 show the estimated mean function and two estimated eigenfunctions associated with the top two eigenvalues. The decreasing trend of the estimated mean function is apparent, demonstrating the continuing decline of physical function of the population as subjects age. The first eigenfunction captures overall difference across subjects, while the second eigenfunction characterizes subjects with accelerating decline (with negative scores) or slowing decline (with positive scores) of physical function. Figure 4.5 plots the estimated coefficient functions with their 95% point-wise confidence bands and 95% simultaneous confidence bands (Ruppert et al. 2003b) for the baseline functional covariate in longitudinal model and survival model, respectively. The left plot implies that sitting bouts are negatively associated with physical function significant at the 0.05 level and in particular, longer sitting bouts are associated with more physical function decline. Scientifically sensible interpretation can also be made for the survival outcome. For example, longer sitting bouts are associated with higher risk of mortality. Table 4.1 provides the estimated coefficients of baseline scalar covariates in longitudinal model and the Cox regression coefficients, which are consistent with scientific knowledge in aging. The table also gives the estimated coefficient of FPC scores of the longitudinal outcome, linking physical function to mortality. The first two FPC scores are significant and the signs are interpretable. For example, for the first FPC score, the estimated coefficient is positive and the associated eigenfunction is always negative (Figure 4.4 left panel), which means if a subject has a negative score, then this subject has higher values of physical function and also lower risk of mortality.

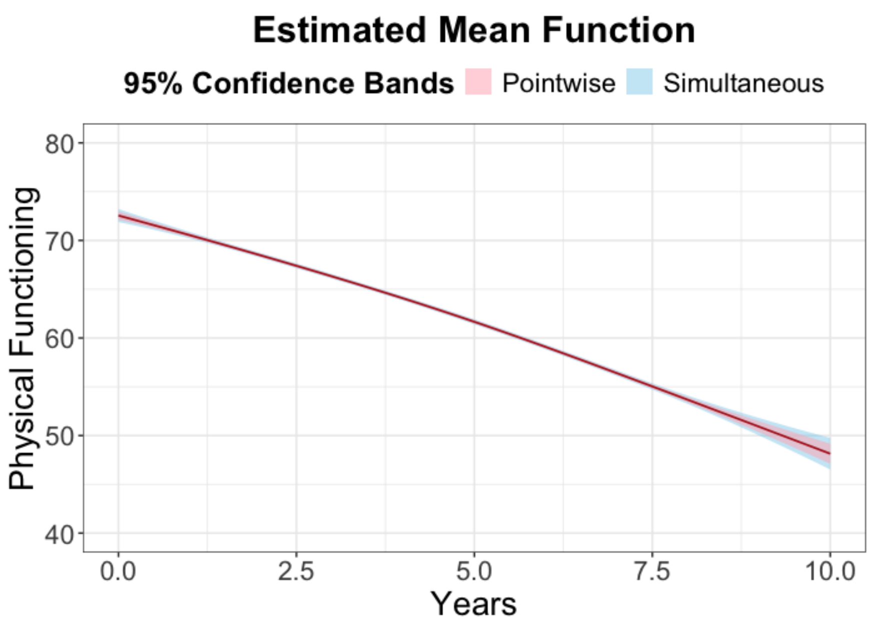


Figure 4.3: Estimated mean trajectories of SF-36 physical function.

4.6 Simulation Study

4.6.1 Simulation Settings

We evaluate the performance of the FJM method by generating data similar to the real data. The longitudinal data are generated according to Model (4.1), following the structure of the real data. The baseline coefficient vector, mean function, coefficient function, eigenfunctions and error variance are derived from the estimates obtained in the application study. To simplify the analysis, we use two functional principal components for the longitudinal outcome and the eigenfunctions used in the simulations are those two associated with largest eigenvalues from

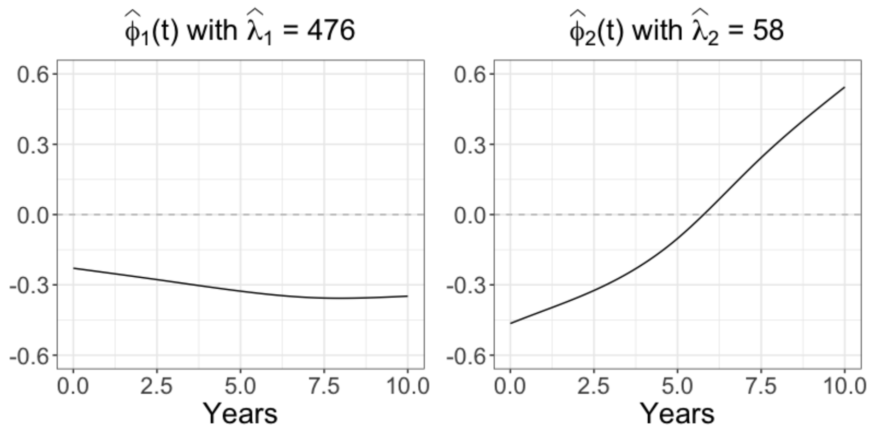


Figure 4.4: Top two estimated eigenfunctions with associated eigenvalues.

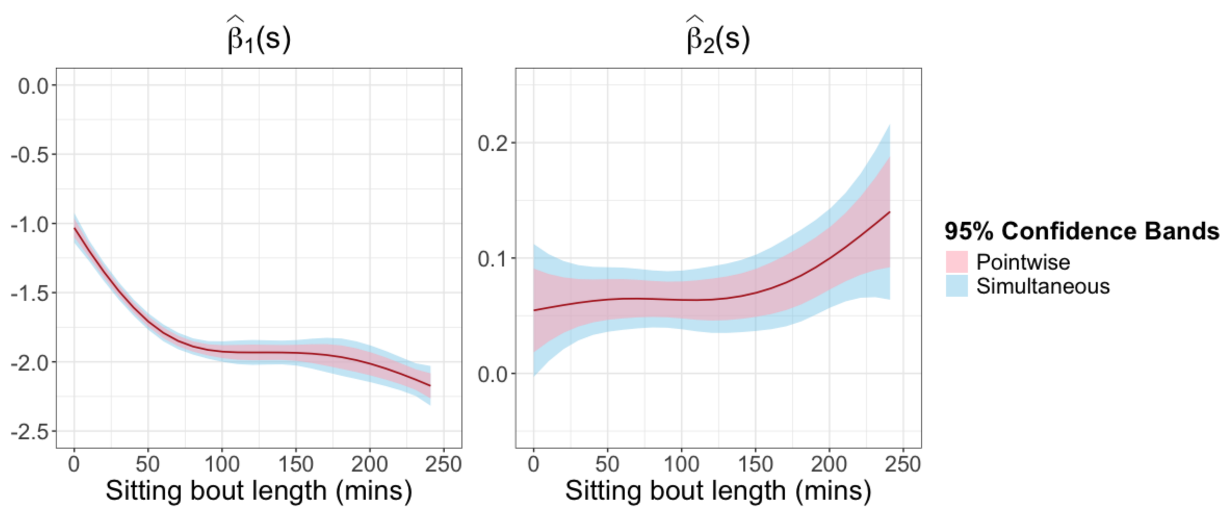


Figure 4.5: Estimated coefficient functions (red line) with 95% point-wise confidence bands and 95% simultaneous confidence bands in longitudinal model (left panel) and survival model (right panel).

Table 4.1: Estimated coefficients of baseline scalar covariates in longitudinal model and estimated Cox regression coefficients (with standard errors), evaluating the association between baseline scalar covariates, FPCA scores, and time to event. Baseline covariates include race, age at baseline, and awake time. An asterisks indicates significance at the 0.05 level.

	Longitudinal		Survival	
Covariates	Estimate (s.e.)	P-value	Estimate (s.e.)	P-value
Race: black	-0.67 (0.06)*	0.00	-0.09 (0.03)*	0.01
Race: hispanic	0.00 (0.06)	0.97	-0.15 (0.04)*	0.00
Age	-7.05 (0.06)*	0.00	0.67 (0.03)*	0.00
Awake time (hours per day)	6.46 (0.06)*	0.00	-0.30 (0.03)*	0.00
SF-36 physical function score (ξ_{i1})	-	-	0.02 (0.00)*	0.00
SF-36 physical function score (ξ_{i2})	-	-	-0.03 (0.01)*	0.00
SF-36 physical function score (ξ_{i3})	-	-	0.00 (0.02)	0.94
SF-36 physical function score (ξ_{i4})	-	-	0.03 (0.03)	0.24

the real data.

The shared random scores ξ_{i1} and ξ_{i2} are generated from a normal distribution $\mathcal{N}(0, \lambda_{1\ell})$, where $\lambda_{11} = 477.39$ and $\lambda_{12} = 58.22$, as estimated from the real data. Measurement error, denoted as ϵ_{ij}, are sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 = 111.85$. The observed time points t_{ij} correspond to 11 fixed time points mapped to the interval $[0, 1]$ and are subject to right truncation due to censoring and event times, as described below.

The events data are generated using the Cox regression model in (4.2), incorporating the random scores, four baseline scalar covariates and a baseline functional covariate from the data application. Speicially, the linear hazard rate function is specified as $\boldsymbol{Z}_i^T\boldsymbol{\gamma}_2 + \left(\int_{s\in\mathcal{S}} W_i(s)\boldsymbol{b}^{\beta}(s)ds\right)^{\top}\boldsymbol{\omega}_2 + \xi_{i1}\gamma_{31} + \xi_{i2}\gamma_{32}$, where the Cox coefficients and coefficient function are estimates from the data application. A Weibull baseline hazard $h_0(t) = \rho t^{\rho-1}$ with $\rho = 20$ is used, and event times S_i are generated via the inverse probability method (Bender et al. 2005). Censoring time $C_i s$ are sampled independently from a Beta distribution, with α_0 and β_0 chosen to approximate censoring rates of 77%, matching the real data. For each subject, only measurements at $t_{ij} \le T_i = \min(S_i, C_i)$ are retained. The sample size is 5708, with average number of observations per subject of 8.9. These settings closely mimic the real data, ensuring realistic simulation conditions. We simulate data 100 times.

4.6.2 Simulation Results

To assess model performance, we first fit the FJM using the true number (two) of functional principal components. Figure 4.6 shows the estimated mean functions and eigenfunctions across 100 replications. Figure 4.7 shows the two estimated coefficient functions in longitudinal model and survival model across 100 replications. Gray lines represent estimates from individual replicates, dashed blue lines show the average of these estimates, and solid red lines represent the true function. The dashed blue and solid red lines align closely, except for estimates of $\widehat{\beta}_1(t)$ with some noticeable bias. A comprehensive assessment of estimated model parameters are presented in Appendix C.5. These parameters include the eigenvalues λ_1 and λ_2, coefficient vector $\boldsymbol{\gamma}_1$ in longitudinal model, the Cox regression coefficients $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$, and the white noise variance σ^2. The estimated model parameters align closely with their true values, confirming the good performance of the proposed FJM.

Finally, we evaluate model selection using AIC and BIC to determine the number of functional principal component for modeling the longitudinal outcome. The proposed selection method demonstrates high accuracy, achieving correct selection rates of 0.93 for AIC and 1.00 for BIC, respectively. These results highlight the robustness of BIC in correctly identifying model complexity, making it the preferred criterion for model selection.

In summary, these simulation results substantiate the reliability and effectiveness of the proposed FJM, demonstrating its ability to accurately estimate model parameters and select the appropriate number of eigenfunctions.

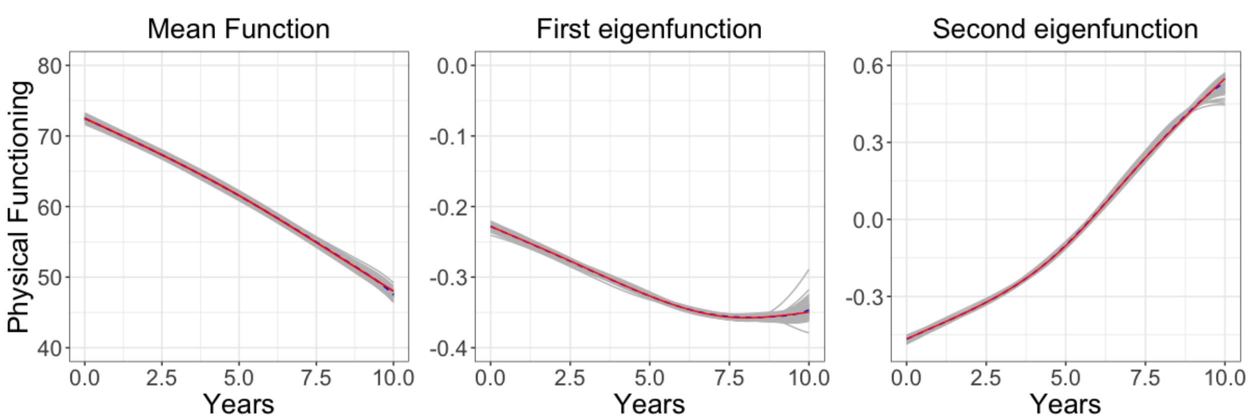


Figure 4.6: Estimated mean functions and eigenfunctions from 100 simulation replications. Gray lines: estimates from 100 replicates; dashed blue lines: the average of these estimates; solid red lines: true mean function and eigenfunctions.

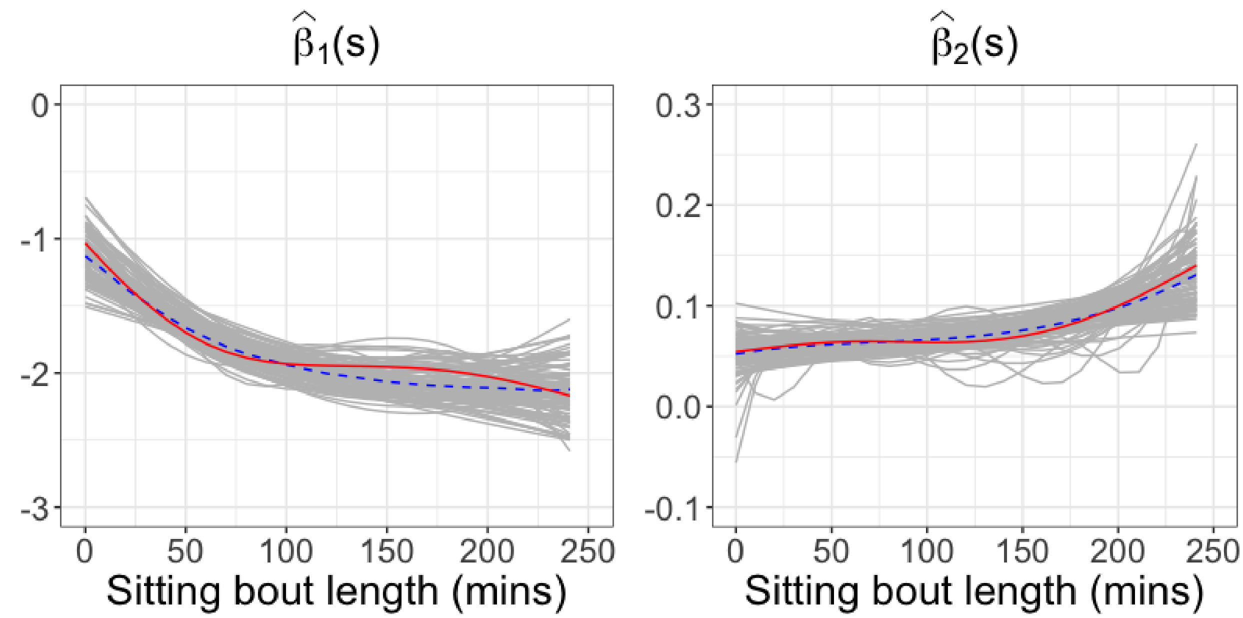


Figure 4.7: Estimated coefficient functions from 100 simulation replications. Gray lines: individual estimates; dashed blue lines: the average of these estimates; solid red lines: true coefficient functions.

4.7 Discussion

In this paper, we proposed a functional joint model (FJM) that integrates a baseline functional covariate into both the longitudinal and survival sub-models. Our model is motivated by the need to better understand the relationship between daily sitting bout profiles and physical functions and all-cause mortality in older women from the OPACH study. By incorporating the entire sitting profile as a functional predictor, our approach goes beyond traditional summary measures and enables a more nuanced characterization of sedentary behavior and its health impacts. Through application to the OPACH data, we demonstrated that our model can successfully detect interpretable associations between baseline sitting patterns and both physical function trajectories and survival outcomes.

To estimate the proposed model, we developed an efficient and stable EM based algorithm that leverages penalized spline approximations for nonparametric smooth functions including the coefficient functions for the baseline functional covariate. Our estimation framework includes an adaptive, local smoothing parameter selection strategy within each EM iteration, improving both flexibility and computational feasibility. The use of penalized splines, as opposed to traditional regression splines, helps mitigate overfitting and ensures smoother, interpretable estimates.

Our work contributes to a growing body of literature on functional joint modeling, particularly by filling a gap in models that incorporate baseline functional covariates. Nonetheless, several limitations warrant future research. First, current framework uses a parametric structure to link the longitudinal and survival sub-models, relying specifically on principal scores. Although this strategy offers simplicity and computational efficiency, it may be insufficient for capturing more complex relationships. Future extensions could consider more flexible alternatives such as nonparametric association structures (Zou et al. 2023) to better reflect the complexity of real-world relationships. Second, although we focus on a single functional covariate in this work, extending the framework to accommodate multiple or time-varying functional covariates would increase its applicability and potential in broader longitudinal and survival study contexts.

REFERENCES

Alzheimer's Association (2024). 2024 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 20(5):3708–3821.

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine*, 24(11):1713–1723.

Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., Schneider, J. A., Perry, G., Avila, J., Moreira, P., Sorensen, A., and Tabaton, M. (2018). Religious orders study and rush memory and aging project. *Journal of Alzheimer's Disease*, 64(s1):S161–S189. PMID: 29865057.

Besser, L., Kukull, W., Knopman, D. S., Chui, H., Galasko, D., Weintraub, S., Jicha, G., Carlsson, C., Burns, J., Quinn, J., Sweet, R. A., Rascovsky, K., Teylan, M., Beekly, D., Thomas, G., Bollenbeck, M., Monsell, S., Mock, C., Zhou, X. H., Thomas, N., Robichaud, E., Dean, M., Hubbard, J., Jacka, M., Schwabe-Fry, K., Wu, J., Phelps, C., and Morris, J. C. (2018). Version 3 of the national alzheimer's coordinating center's uniform data set. *Alzheimer Disease & Associated Disorders*, 32(4):351–358.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(1):265–285.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, pages 9–25.

Brockhaus, S., Scheipl, F., Hothorn, T., and Greven, S. (2015). The functional linear array model. *Statistical Modelling*, 15(3):279–300.

Broglio, S. P., McCrea, M., McAllister, T., Harezlak, J., Katz, B., Hack, D., and Hainline, B. (2017). A national study on the effects of concussion in collegiate athletes and us military service academy members: the ncaa–dod concussion assessment, research and education (care) consortium structure and methods. *Sports medicine*, 47:1437–1451.

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Annals of Statistics*.

Crainiceanu, C. M., Goldsmith, J., Leroux, A., and Cui, E. (2024). *Functional Data Analysis with R*. Chapman and Hall/CRC, New York, 1st edition. First published 2024, eBook published 10 March 2024.

Cui, E., Crainiceanu, C. M., and Leroux, A. (2021). Additive functional cox model. *Journal of Computational and Graphical Statistics*, 30(3):780–793.

De Boor, C. and De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-verlag, New York.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). On the best rank-1 and rank-(r 1, r 2,..., rn) approximation of higher-order tensors. *SIAM journal on Matrix Analysis and Applications*, 21(4):1324–1342.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Eanes, L. (2018). Ce: Too much sitting: A newly recognized health risk. *AJN The American journal of nursing*, 118(9):26–34.

Eilers, P. H. and Marx, B. D. (1996a). Flexible smoothing with b-splines and penalties. *Statistical science*, 11(2):89–121.

Eilers, P. H. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, 66(2):159–174.

Eilers, P. H. C. and Marx, B. D. (1996b). Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–102.

Gellar, J. E., Colantuoni, E., Needham, D. M., and Crainiceanu, C. M. (2015). Cox regression models with functional covariates for survival data. *Statistical modelling*, 15(3):256–278.

Gertheiss, J., Goldsmith, J., and Staicu, A.-M. (2017). A note on modeling sparse exponential-family functional response curves. *Computational statistics & data analysis*, 105:46–52.

Ghosal, R., Ghosh, S. K., Schrack, J. A., and Zipunnikov, V. (2025a). Distributional outcome regression via quantile functions and its application to modelling continuously monitored heart rate and physical activity. *Journal of the American Statistical Association*, (just-accepted):1–20.

Ghosal, R., Matabuena, M., and Ghosh, S. K. (2025b). Functional time transformation model with applications to digital health. *Computational Statistics & Data Analysis*, page 108131.

Goldsmith, J., Greven, S., and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics*, 69(1):41–51.

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71(2):344–353.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 159(3):505–513.

Greenwood-Hickman, M. A., Nakandala, S., Jankowska, M. M., Rosenberg, D. E., Tuz-Zahra, F., Bellettiere, J., Carlson, J., Hibbing, P. R., Zou, J., Lacroix, A. Z., et al. (2021). The cnn hip accelerometer posture (chap) method for classifying sitting patterns from hip accelerometers: A validation study. *Medicine and science in sports and exercise*, 53(11):2445.

Group, W. H. I. S. et al. (1998). Design of the women's health initiative clinical trial and observational study. *Controlled clinical trials*, 19(1):61–109.

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 70(4):703–723.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779.

He, K., Lian, H., Ma, S., and Huang, J. Z. (2018). Dimensionality reduction and variable selection in multivariate varying-coefficient models with a large number of covariates. *Journal of the American Statistical Association*, 113(522):746–754.

He, Y., Song, X., and Kang, K. (2024). Joint mixed membership modeling of multivariate longitudinal and survival data for learning the individualized disease progression. *The Annals of Applied Statistics*, 18(3):1924–1946.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hickey, G. L., Philipson, P., Jorgensen, A., and Kolamunnage-Dona, R. (2018). joinerml: a joint model and software package for time-to-event and multivariate longitudinal outcomes. *BMC medical research methodology*, 18:1–14.

Hong, Y., Su, L., Song, S., and Yan, F. (2021). Dynamic prediction of disease processes based on recurrent history and functional principal component analysis of longitudinal biomarkers: Application for ovarian epithelial cancer. *Statistics in Medicine*, 40(8):2006–2023.

Huang, H., Li, Y., and Guan, Y. (2014). Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data. *Journal of the American Statistical Association*, 109(508):1412–1424.

Jaakkola, T. S. and Jordan, M. I. (1997). A variational approach to bayesian logistic regression models and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, pages 283–294. PMLR.

James, G., Hastie, T., and Sugar, C. (2000a). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

James, G. M., Hastie, T. J., and Sugar, C. A. (2000b). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

Jiang, Q., Wang, H., Xia, Y., and Jiang, G. (2013). On a principal varying coefficient model. *Journal of the American Statistical Association*, 108(501):228–236.

Jiang, S., Cao, J., and Colditz, G. A. (2024). Functional partial least squares with censored outcomes: Prediction of breast cancer risk with mammogram images. *The Annals of Applied Statistics*, 18(2):1051–1063.

Jiang, S., Cao, J., Colditz, G. A., and Rosner, B. (2023). Predicting the onset of breast cancer using mammogram imaging data with irregular boundary. *Biostatistics*, 24(2):358–371.

Karhunen, K. (1947). Under lineare methoden in der wahr scheinlichkeitsrechnung. *Annales Academiae Scientiarun Fennicae Series A1: Mathematia Physica*, 47.

Kong, D., Ibrahim, J. G., Lee, E., and Zhu, H. (2018). Flcrm: Functional linear cox regression model. *Biometrics*, 74(1):109–117.

Kowal, D. R. (2019). Integer-valued functional data analysis for measles forecasting. *Biometrics*, 75(4):1321–1333.

LaCroix, A. Z., Rillamas-Sun, E., Buchner, D., Evenson, K. R., Di, C., Lee, I.-M., Marshall, S., LaMonte, M. J., Hunt, J., Tinker, L. F., et al. (2017). The objective physical activity and cardiovascular disease health in older women (opach) study. *BMC public health*, 17:1–12.

Lee, E., Zhu, H., Kong, D., Wang, Y., Giovanello, K. S., and Ibrahim, J. G. (2015). Bflcrm: A bayesian functional linear cox regression model for predicting time to conversion to alzheimer's disease. *The annals of applied statistics*, 9(4):2153.

Leroux, A., Crainiceanu, C., and Wrobel, J. (2023). Fast generalized functional principal components analysis. *arXiv preprint arXiv:2305.02389*.

Levine, R. A. and Casella, G. (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439.

Li, C., Xiao, L., and Luo, S. (2022). Joint model for survival and multivariate sparse functional data with application to a study of alzheimer's disease. *Biometrics*, 78(2):435–447.

Li, G., Huang, J. Z., and Shen, H. (2018). Exponential family functional data analysis via a low-rank model. *Biometrics*, 74(4):1301–1310.

Li, H., Staudenmayer, J., and Carroll, R. J. (2014). Hierarchical functional data with mixed continuous and binary measurements. *Biometrics*, 70(4):802–811.

Li, K. and Luo, S. (2017). Functional joint model for longitudinal and time-to-event data: an application to alzheimer's disease. *Statistics in medicine*, 36(22):3560–3572.

Li, K. and Luo, S. (2019). Dynamic prediction of alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Statistics in medicine*, 38(24):4804–4818.

Li, R. and Xiao, L. (2023). Latent factor model for multivariate functional data. *Biometrics*, 79(4):3307–3318.

Li, Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*.

Lin, H., McCulloch, C. E., and Mayne, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. *Statistics in medicine*, 21(16):2369–2382.

Loeve, M. (1945). *Fonctions aléatoires du second ordre*, volume 220.

McCrea, M. A., Shah, A., Duma, S., Rowson, S., Harezlak, J., McAllister, T. W., Broglio, S. P., Giza, C. C., Goldman, J., Cameron, K. L., et al. (2021). Opportunities for prevention of concussion and repetitive head impact exposure in college football players: a concussion assessment, research, and education (care) consortium study. *JAMA neurology*, 78(3):346–350.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition.* Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Mercer, J. (1909). Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2):179–199.

Murray, J. and Philipson, P. (2023). Fast estimation for generalised multivariate joint models using an approximate em algorithm. *Computational Statistics & Data Analysis*, 187:107819.

Nolan, T. H., Goldsmith, J., and Ruppert, D. (2025). Bayesian functional principal components analysis via variational message passing with multilevel extensions. *Bayesian Analysis*, 20(1):1459–1485.

Ohalete, N. C., Ayo-Farai, O., Onwumere, C., Maduka, C. P., and Olorunsogo, T. O. (2024). Functional data analysis in health informatics: A comparative review of developments and applications in the usa and africa. *World Journal of Advanced Research and Reviews*, 21(1):1097–1114.

Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics*, 18(4):995–1015.

Qu, S., Wang, J.-L., and Wang, X. (2016). Optimal estimation for the functional cox model.

Ramsay, J. and Silverman, B. (2005). *Functional data analysis.* Springer, New York.

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of statistical software*, 35:1–33.

Rügamer, D., Brockhaus, S., Gentsch, K., Scherer, K., and Greven, S. (2018). Boosting factor-specific functional historical models for the detection of synchronization in bioelectrical signals. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(3):621–642.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11(4):735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003a). *Semiparametric regression*. Number 12. Cambridge university press.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003b). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Scheipl, F., Staicu, A.-M., and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24(2):477–501.

Serban, N., Staicu, A.-M., and Carroll, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics*, 69(4):903–913.

Shen, X., Wolfe, D., and Zhou, S. (1998). Local asymptotics for regression splines and confidence regions. *The annals of statistics*, 26(5):1760–1782.

Shi, H., Jiang, S., Ma, D., Beg, M. F., and Cao, J. (2024). Dynamic survival prediction using sparse longitudinal images via multi-dimensional functional principal component analysis. *Journal of Computational and Graphical Statistics*, (just-accepted):1–18.

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1):1–24.

Staniswalis, J. G. and Lee, J. J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418.

Stemper, B. D., Shah, A. S., Harezlak, J., Rowson, S., Mihalik, J. P., Duma, S. M., Riggen, L. D., Brooks, A., Cameron, K. L., Campbell, D., et al. (2019). Comparison of head impact exposure between concussed football athletes and matched controls: evidence for a possible second mechanism of sport-related concussion. *Annals of biomedical engineering*, 47:2057–2072.

Tsiatis, A. A. and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311.

Van Der Linde, A. (2008). Variational bayesian functional pca. *Computational statistics & data analysis*, 53(2):517–533.

Van Der Linde, A. (2009). A bayesian latent variable approach to functional principal components analysis with binary and count data. *AStA Advances in Statistical Analysis*, 93:307–333.

Wang, W., Xiao, L., Li, R., and Luo, S. (2025). A functional joint model for survival and multivariate sparse functional data in multi-cohort alzheimer's disease study. Manuscript in preparation.

Wang, Y., Ibrahim, J. G., and Zhu, H. (2020). Partial least squares for functional joint models with applications to the alzheimer's disease neuroimaging initiative study. *Biometrics*, 76(4):1109–1119.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Luthman, J., Morris, J. C., Petersen, R. C., Saykin, A. J., Shaw, L., Shen, L., Schwarz, A., Toga, A. W., Trojanowski, J. Q., and Initiative, A. D. N. (2015). 2014 update of the alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 11(6):e1–e120.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):95–114.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York, 2nd edition. First published 2017, eBook published 17 May 2017.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Wrobel, J., Zipunnikov, V., Schrack, J., and Goldsmith, J. (2019). Registration for exponential family functional data. *Biometrics*, 75(1):48–57.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, pages 330–339.

Xiao, L. (2019). Asymptotic theory of penalized splines. *Electronic Journal of Statistics*, 13(1):747–794.

Xiao, L. (2020). Asymptotic properties of penalized splines for functional data.

Xiao, L., Li, C., Checkley, W., and Crainiceanu, C. (2018). Fast covariance estimation for sparse functional data. *Statistics and computing*, 28:511–522.

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

Yan, F., Lin, X., and Huang, X. (2017). Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *The Annals of Applied Statistics*, 11(3):1649 – 1670.

Yang, C.-C. and Hsu, Y.-L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, 10(8):7772–7788.

Yao, F. (2007). Functional principal component analysis for longitudinal and survival data. *Statistica Sinica*, pages 965–983.

Yao, F., Müller, H.-G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590.

Zhang, F., Zhou, Y., He, K., and Wong, R. K. W. (2022). Multivariate varying-coefficient models via tensor decomposition. *Statistica Sinica*. Preprint No: SS-2022-0103.

Zhang, X. and Wang, J.-L. (2016). From sparse to dense functional data and beyond.

Zhao, W., Jiang, X., and Lian, H. (2018). A principal varying-coefficient model for quantile regression: Joint variable selection and dimension reduction. *Computational Statistics & Data Analysis*, 127:269–280.

Zhong, R., Liu, S., Li, H., and Zhang, J. (2023). Sparse logistic functional principal component analysis for binary data. *Statistics and Computing*, 33(1):15.

Zou, H., Xiao, L., Zeng, D., and Luo, S. (2024). Dynamic prediction with multivariate longitudinal outcomes and longitudinal magnetic resonance imaging data. *The Annals of Applied Statistics*. accepted.

Zou, H., Zeng, D., Xiao, L., and Luo, S. (2023). Bayesian inference and dynamic prediction for multivariate longitudinal and survival data. *The Annals of Applied Statistics*, 17(3):2574–2595.

APPENDICES

APPENDIX

A

SUPPLEMENT TO CHAPTER 2

A.1 Joint Distribution of Longitudinal Outcomes and FPC Scores

We derive the joint distribution of $f(\boldsymbol{y}_i, \boldsymbol{\eta}_i | \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2)$, which characterizes the relationship between the observed longitudinal outcomes and the latent principal scores. The observed outcomes follow the equation: $\boldsymbol{y}_{ij} = \boldsymbol{\mu}_{ij} + \beta_j \boldsymbol{\Phi}_{ij} \boldsymbol{\xi}_i + \beta_j \boldsymbol{\Psi}_{ij} \boldsymbol{\zeta}_{ij} + \boldsymbol{\epsilon}_{ij}$, where $\boldsymbol{\mu}_{ij} = (\mu_j^{c(i)}(t_{ij1}), \ldots, \mu_j^{c(i)}(t_{ijm_{ij}}))^T$ is the vector of the jth mean function for cohort $c(i)$ at the observed time points and $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijm_{ij}})^T$ is the vector of measurement errors. Let $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^T, \ldots, \boldsymbol{\mu}_{iJ_{c(i)}}^T)^T$. We have

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{\eta}_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}(\boldsymbol{y}_i) & \mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{\eta}_i) \\ \mathrm{Cov}(\boldsymbol{\eta}_i, \boldsymbol{y}_i) & \mathrm{Cov}(\boldsymbol{\eta}_i) \end{pmatrix}\right],$$

where $\mathrm{Cov}(\boldsymbol{y}_i) =$

$$\begin{pmatrix} \beta_1^2 \boldsymbol{\Phi}_{i1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_{i1}^T + \beta_1^2 \boldsymbol{\Psi}_{i1} \boldsymbol{\Lambda}_2 \boldsymbol{\Psi}_{i1}^T + \sigma_1^2 \boldsymbol{I}_{m_{i1}}, & \cdots & \beta_1 \beta_{J_{c(i)}} \boldsymbol{\Phi}_{i1} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_{iJ_{c(i)}}^T \\ \vdots & \cdots & \vdots \\ \beta_{J_{c(i)}} \beta_1 \boldsymbol{\Phi}_{iJ_{c(i)}} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_{i1}^T, & \cdots & \beta_{J_{c(i)}}^2 \boldsymbol{\Phi}_{iJ_{c(i)}} \boldsymbol{\Lambda}_1 \boldsymbol{\Phi}_{iJ_{c(i)}}^T + \beta_{J_{c(i)}}^2 \boldsymbol{\Psi}_{iJ_{c(i)}} \boldsymbol{\Lambda}_2 \boldsymbol{\Psi}_{iJ_{c(i)}}^T + \sigma_{J_{c(i)}}^2 \boldsymbol{I}_{m_{iJ_{c(i)}}} \end{pmatrix},$$

$$
\mathrm{Cov}(\boldsymbol{y}_i,\boldsymbol{\eta}_i)=\begin{pmatrix}
\beta_1\boldsymbol{\Phi}_{i1}\boldsymbol{\Lambda}_1, & \beta_1\boldsymbol{\Psi}_{i1}\boldsymbol{\Lambda}_2 & 0 & \dots & 0\\
\beta_2\boldsymbol{\Phi}_{i2}\boldsymbol{\Lambda}_1, & 0 & \beta_2\boldsymbol{\Psi}_{i2}\boldsymbol{\Lambda}_2 & \dots & 0\\
\vdots & \vdots & \vdots & \dots & \vdots\\
\beta_{J_{c(i)}}\boldsymbol{\Phi}_{iJ_{c(i)}}\boldsymbol{\Lambda}_1, & 0 & 0 & \dots & \beta_{J_{c(i)}}\boldsymbol{\Psi}_{iJ_{c(i)}}\boldsymbol{\Lambda}_2
\end{pmatrix},
$$

and $\mathrm{Cov}(\boldsymbol{\eta}_i)=\mathrm{blockdiag}\{\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2,\dots,\boldsymbol{\Lambda}_2\}$. Using conditional expectation, the posterior mean and covariance of $\boldsymbol{\eta}_i$ given $\boldsymbol{y}_i$ are derived as: $\mathbb{E}_i(\boldsymbol{\eta}_i|\boldsymbol{y}_i)=\mathrm{Cov}(\boldsymbol{\eta}_i,\boldsymbol{y}_i)\mathrm{Cov}(\boldsymbol{y}_i)^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$, $\mathrm{Cov}(\boldsymbol{\eta}_i|\boldsymbol{y}_i)=\mathrm{Cov}(\boldsymbol{\eta}_i)-\mathrm{Cov}(\boldsymbol{\eta}_i,\boldsymbol{y}_i)\mathrm{Cov}^{-1}(\boldsymbol{y}_i)\mathrm{Cov}(\boldsymbol{y}_i,\boldsymbol{\eta}_i)$.

A.2 Marginal Log-likelihood

The marginal log-likelihood is a key quantity for model evaluation and comparison, providing a measure of how well the proposed model fits the observed data. Since the exact computation of this likelihood involves intractable integrals, we employ an approximation strategy as detailed below.

The marginal log-likelihood of the data is given by

$$
\begin{aligned}
\ell(\widehat{\boldsymbol{\Omega}})&=\sum_{i=1}^{n}\omega_{c(i)}\log\left\{\int f\left(\boldsymbol{y}_i\mid\boldsymbol{x}_i,\widehat{\boldsymbol{\Sigma}}_i\right)f\left(T_i,\Delta_i\mid\widehat{h}_0,\boldsymbol{z}_i,\boldsymbol{\xi}_i,\widehat{\boldsymbol{\gamma}}_{c(i)}\right)f\left(\boldsymbol{\xi}_i\mid\widehat{\boldsymbol{\Lambda}}_1\right)\sum_{j\in\mathscr{J}_{c(i)}}f\left(\boldsymbol{\zeta}_{ij}\mid\widehat{\boldsymbol{\Lambda}}_2\right)\mathrm{d}\boldsymbol{\eta}_i\right\}\\
&=\sum_{i=1}^{n}\omega_{c(i)}\log\left\{f\left(\boldsymbol{y}_i\mid\widehat{\boldsymbol{\Omega}}\right)\int f\left(\boldsymbol{\eta}_i\mid\boldsymbol{y}_i,\widehat{\boldsymbol{\Omega}}\right)f\left(T_i,\Delta_i\mid\widehat{h}_0,\boldsymbol{z}_i,\boldsymbol{\xi}_i,\widehat{\boldsymbol{\gamma}}_{c(i)}\right)\mathrm{d}\boldsymbol{\eta}_i\right\}.
\end{aligned}
$$

Since this integral does not have a closed-form solution, we approximate it using Monte Carlo integration. Specifically, the integral is approximated as:

$$
\left\{R^{-1}\sum_{r=1}^{R}f\left(T_i,\Delta_i\mid\widehat{h}_0,\boldsymbol{z}_i,\xi_i^{(r)},\widehat{\boldsymbol{\gamma}}_{c(i)}\right)\right\},
$$

where $\xi_i^{(r)}$ are samples drawn from the posterior distribution $f\left(\boldsymbol{\eta}_i\mid\boldsymbol{y}_i,\widehat{\boldsymbol{\Omega}}\right)$. Substituting this approximation, the marginal log-likelihood becomes:

$$
\ell(\widehat{\boldsymbol{\Omega}})\approx\sum_{i=1}^{n}\omega_{c(i)}\left[\log\left\{f\left(\boldsymbol{y}_i\mid\widehat{\boldsymbol{\Omega}}\right)\right\}+\log\left\{R^{-1}\sum_{r=1}^{R}f\left(T_i,\Delta_i\mid\widehat{h}_0,\boldsymbol{z}_i,\xi_i^{(r)},\widehat{\boldsymbol{\gamma}}_{c(i)}\right)\right\}\right].
$$

A.3 Details on M-step

Unlike the B-splines used in Li et al. (2022), we estimate $\boldsymbol{\alpha}$, $\boldsymbol{\Theta}_1$, and $\boldsymbol{\Theta}_2$ using P-splines (Eilers and Marx 1996b) with smoothing parameters. P-splines provide stable estimates that enhance the convergence speed of the iterative algorithm, contributing to computational efficiency.

First, we estimate $\boldsymbol{\alpha}$ by minimizing the following penalized least squares objective function:

$$\sum_{i\in\mathcal{I}_c}\left\|\boldsymbol{y}_{ij}-\widehat{\beta}_j\widehat{\boldsymbol{\Phi}}_{ij}\mathbb{E}_i(\boldsymbol{\xi}_i)-\widehat{\beta}_j\widehat{\boldsymbol{\Psi}}_{ij}\mathbb{E}_i(\boldsymbol{\zeta}_{ij})-\boldsymbol{B}_{ij}\boldsymbol{\alpha}_j^{c(i)}\right\|^2+\tau_j^{c(i)}\boldsymbol{\alpha}_j^{c(i)^T}\boldsymbol{P}\boldsymbol{\alpha}_j^{c(i)}, \tag{A.1}$$

where $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{P}$ is a penalty matrix, and $\tau_j^{c(i)}$ is the smoothing parameter.

Second, given $\boldsymbol{\theta}_{1d'}$, $d'\neq d$, $\boldsymbol{\theta}_{1d}$, we estimate $\boldsymbol{\Theta}_1$ by minimizing:

$$\sum_{i=1}^n\omega_{c(i)}\sum_{j\in\mathcal{I}_{c(i)}}\mathbb{E}_i\left\|\boldsymbol{y}_{ij}-\widehat{\boldsymbol{\mu}}_{ij}-\widehat{\beta}_j\widehat{\boldsymbol{\Psi}}_{ij}\boldsymbol{\zeta}_{ij}-\widehat{\beta}_j\sum_{d'\neq d}\widetilde{\boldsymbol{B}}_{ij}\widehat{\boldsymbol{\theta}}_{1d'}\xi_{id'}-\widehat{\beta}_j\widetilde{\boldsymbol{B}}_{ij}\boldsymbol{\theta}_{1d}\xi_{id}\right\|^2+\tau_{1d}\boldsymbol{\theta}_{1d}^T\boldsymbol{P}\boldsymbol{\theta}_{1d}.$$

This is equivalent to minimizing:

$$\begin{aligned}\sum_{i=1}^n\omega_{c(i)}\sum_{j\in\mathcal{I}_{c(i)}}\Bigg\|&\frac{\mathbb{E}_i(\xi_{id})}{\sqrt{\mathbb{E}_i(\xi_{id}^2)}}(\boldsymbol{y}_{ij}-\widehat{\boldsymbol{\mu}}_{ij})-\frac{\widehat{\beta}_j}{\sqrt{\mathbb{E}_i(\xi_{id}^2)}}\widehat{\boldsymbol{\Psi}}_{ij}\mathbb{E}_i(\boldsymbol{\zeta}_{ij}\xi_{id})\\&-\frac{\widehat{\beta}_j}{\sqrt{\mathbb{E}_i(\xi_{id}^2)}}\sum_{d'\neq d}\widetilde{\boldsymbol{B}}_{ij}\widehat{\boldsymbol{\theta}}_{1d'}\mathbb{E}_i(\xi_{id'}\xi_{id})-\widehat{\beta}_j\widetilde{\boldsymbol{B}}_{ij}\boldsymbol{\theta}_{1d}\sqrt{\mathbb{E}_i(\xi_{id}^2)}\Bigg\|^2+\tau_{1d}\boldsymbol{\theta}_{1d}^T\boldsymbol{P}\boldsymbol{\theta}_{1d}.\end{aligned} \tag{A.2}$$

Similarly, $\boldsymbol{\theta}_{2d}$ is estimated by minimizing:

$$\begin{aligned}\sum_{i=1}^n\omega_{c(i)}\sum_{j\in\mathcal{I}_{c(i)}}\Bigg\|&\frac{\mathbb{E}_i(\zeta_{ijd})}{\sqrt{\mathbb{E}_i(\zeta_{ijd}^2)}}(\boldsymbol{y}_{ij}-\widehat{\boldsymbol{\mu}}_{ij})-\frac{\widehat{\beta}_j}{\sqrt{\mathbb{E}_i(\zeta_{ijd}^2)}}\widehat{\boldsymbol{\Phi}}_{ij}\mathbb{E}_i(\boldsymbol{\xi}_i\zeta_{ijd})\\&-\frac{\widehat{\beta}_j}{\sqrt{\mathbb{E}_i(\zeta_{ijd}^2)}}\sum_{d'\neq d}\widetilde{\boldsymbol{B}}_{ij}\widehat{\boldsymbol{\theta}}_{2d'}\mathbb{E}_i(\zeta_{ijd'}\zeta_{ijd})-\widehat{\beta}_j\widetilde{\boldsymbol{B}}_{ij}\boldsymbol{\theta}_{2d}\sqrt{\mathbb{E}_i(\zeta_{ijd}^2)}\Bigg\|^2+\tau_{2d}\boldsymbol{\theta}_{2d}^T\boldsymbol{P}\boldsymbol{\theta}_{2d}.\end{aligned} \tag{A.3}$$

Minimizing (A.1), (A.2), and (A.3) can be solved directly via the *gam* function in the R package *mgcv* (Wood 2011; Wood et al. 2016). The smoothing parameters τ are selected locally through generalized cross-validation during each iteration. The procedure is repeated for each column of $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ until convergence, ensuring robust estimation.

Third, the scaling parameter $\boldsymbol{\beta}$ is updated using a linear model by minimizing the weighted least square: $\sum_{i=1}^n\omega_{c(i)}\mathbb{E}_i\left\|\boldsymbol{y}_{ij}-\widehat{\boldsymbol{\mu}}_{ij}-\beta_j(\widehat{\boldsymbol{\Psi}}_{ij}\boldsymbol{\zeta}_{ij}+\widehat{\boldsymbol{\Phi}}_{ij}\boldsymbol{\xi}_i)\right\|^2$.

The optimal β_j is then:

$$\widehat{\beta}_j = \frac{\sum_{i=1}^n \omega_{c(i)} \Big\{ (\boldsymbol{y}_{ij} - \widehat{\boldsymbol{\mu}}_{ij})^T (\widehat{\boldsymbol{\Phi}}_{ij} \mathbb{E}_i \boldsymbol{\xi}_i + \widehat{\boldsymbol{\Psi}}_{ij} \mathbb{E}_i \boldsymbol{\zeta}_{ij}) \Big\}}{\sum_{i=1}^n \omega_{c(i)} \Big[\mathrm{tr}\big\{ \widehat{\boldsymbol{\Phi}}_{ij}^T \widehat{\boldsymbol{\Phi}}_{ij} \mathbb{E}_i(\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T) \big\} + \mathrm{tr}\big\{ \widehat{\boldsymbol{\Psi}}_{ij}^T \widehat{\boldsymbol{\Psi}}_{ij} \mathbb{E}_i(\boldsymbol{\zeta}_{ij} \boldsymbol{\zeta}_{ij}^T) \big\} + 2\mathrm{tr}\big\{ \widehat{\boldsymbol{\Phi}}_{ij}^T \widehat{\boldsymbol{\Psi}}_{ij} \mathbb{E}_i(\boldsymbol{\zeta}_{ij} \boldsymbol{\xi}_i^T) \big\} \Big]}.$$

Finally, the variance of random noise σ_j^2 is updated as:

$$\begin{aligned}
\widehat{\sigma}_j^2 &= \frac{1}{\sum_{i=1}^n \omega_{c(i)} m_{ij}} \sum_{i=1}^n \omega_{c(i)} \mathbb{E}_i \Big\| \boldsymbol{y}_{ij} - \widehat{\boldsymbol{\mu}}_{ij} - \widehat{\beta}_j \widehat{\boldsymbol{\Phi}}_{ij} \boldsymbol{\xi}_i - \widehat{\beta}_j \widehat{\boldsymbol{\Psi}}_{ij} \boldsymbol{\zeta}_{ij} \Big\|^2 \\
&= \frac{1}{\sum_{i=1}^n \omega_{c(i)} m_{ij}} \sum_{i=1}^n \omega_{c(i)} \Big\{ \|\boldsymbol{y}_{ij} - \widehat{\boldsymbol{\mu}}_{ij}\|^2 - 2(\boldsymbol{y}_{ij} - \widehat{\boldsymbol{\mu}}_{ij})^T \big(\widehat{\beta}_j \widehat{\boldsymbol{\Phi}}_{ij} \mathbb{E}_i(\boldsymbol{\xi}_i) + \widehat{\beta}_j \widehat{\boldsymbol{\Psi}}_{ij} \mathbb{E}_i(\boldsymbol{\zeta}_{ij}) \big) \\
&+ \widehat{\beta}_j^2 \mathrm{tr}\big\{ \widehat{\boldsymbol{\Phi}}_{ij}^T \widehat{\boldsymbol{\Phi}}_{ij} \mathbb{E}_i(\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T) \big\} + \widehat{\beta}_j^2 \mathrm{tr}\big\{ \widehat{\boldsymbol{\Psi}}_{ij}^T \widehat{\boldsymbol{\Psi}}_{ij} \mathbb{E}_i(\boldsymbol{\zeta}_{ij} \boldsymbol{\zeta}_{ij}^T) \big\} \\
&+ 2\widehat{\beta}_j^2 \mathrm{tr}\big\{ \widehat{\boldsymbol{\Phi}}_{ij}^T \widehat{\boldsymbol{\Psi}}_{ij} \mathbb{E}_i(\boldsymbol{\zeta}_{ij} \boldsymbol{\xi}_i^T) \big\} \Big\}.
\end{aligned}$$

A.4 Additional Data Application Results

Table A.1: Baseline covariates and censoring rates in three AD cohorts. APOE4 indicates the number of alleles (0/1/2, %). Censoring rates are calculated as the proportion of censored observations among the total.

	ADNI	NACC	ROSMAP
Covariates	Median (range) / %	Median (range) / %	Median (range) / %
Age in years	73.5 (55.0 - 91.4)	74.0 (28.0 - 99.0)	82.3 (56.2 - 102.1)
Gender (female, %)	42.8	48.0	32.0
Education in years	16.0 (7.0 - 20.0)	16.0 (0.0 - 30.0)	16.0 (5.0 - 30.0)
APOE4 (0/1/2, %)	50.9/38.7/10.4	57.2/35.3/7.5	68.2/29.3/2.5
censoring rate	65.59%	65.58%	58.81%

Table A.1 provides baseline covariates and censoring rates in three AD cohorts, highlighting differences in age, education, and APOE4 allele distributions across cohorts.

Figure A.1 shows the estimated baseline hazard function in real data application.

Table A.2 summarizes AIC and BIC values for different models. The model with common mean functions and cohort-specific Cox coefficients achieves the lowest BIC, indicating a favorable balance of parsimony and fit.

To provide additional insights, we also display the result of multi-cohort FJM with consistent

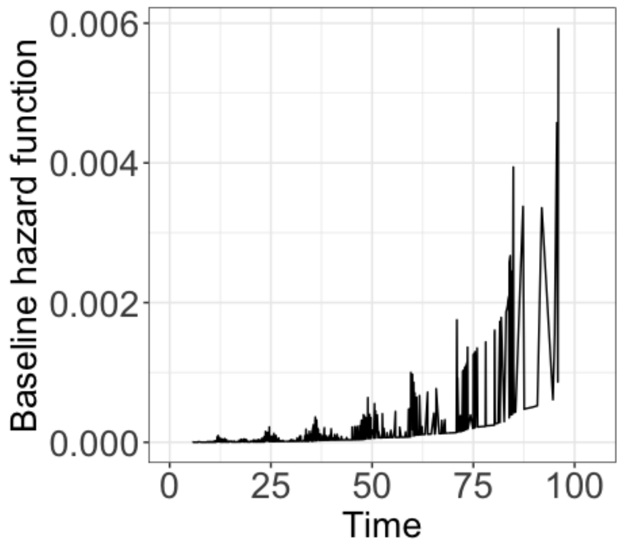


Figure A.1: Estimated baseline hazard function ($h_0(t)$) from the multi-cohort FJM applied to real data.

Table A.2: AIC and BIC values (with selected principle components) for different models. Common MFMM: same mean functions across cohorts; Multi-cohorts MFMM: cohort-specific mean functions; Common Cox regression: same Cox coefficients across cohorts; Multi-cohorts Cox regression: cohort-specific Cox coefficients. Values in parentheses indicate the number of principal components for shared and outcome-specific covariance structures.

Model	AIC	BIC
Common MFMM + Common Cox regression	130270 (5,5)	130658 (5,2)
Common MFMM + Multi-cohorts Cox regression	130153 (5,4)	130602 (4,3)
Multi-cohorts MFMM + Common Cox regression	130216 (5,4)	130649 (3,3)
Multi-cohorts MFMM + Multi-cohorts Cox regression	130114 (5,5)	130620 (4,3)

mean functions for longitudinal outcomes and cohort-specific Cox coefficients for time-to-event data. Using the BIC criteria, this model selects four principal components for the shared covariance structure and three principal components for the outcome-specific covariance structure. The estimated scale coefficients are given by $\boldsymbol{\beta} = (1.00, 1.55, 3.69, 3.67, -0.61, -2.72, -1.84, -1.08)^{\top}$. Figure A.2 shows the first two estimated eigenfunctions obtained from multi-cohort FJM and separate FJM, highlighting the shared and outcome-specific variation patterns captured by the model. Figure A.3 compares the estimated mean trajectories of longitudinal outcomes across the three models: columns one and three depict results for the multi-cohort and separate FJM models, while columns two and four present results for the MJM. These figures underscore the advantages of the multi-cohort FJM in capturing nonlinear trends and cohort-specific differences effectively. Table A.3 displays the estimated Cox regression coefficients from the multi-cohort FJM, demonstrating significant associations between shared latent scores and survival outcomes, along with the effects of baseline covariates.

Table A.3: Cox regression coefficient estimates (standard errors) from the multi-cohort FJM with common MFMM. Covariates include age, gender, education, and APOE4 alleles. Shared scores (ξ_{i1} to ξ_{i4}) quantify latent disease progression. An asterisk indicates significance at the 0.05 level.

	ADNI		NACC		ROSMAP	
Covariates	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value	Estimate (s.e.)	p-value
Age	0.02 (0.01)*	0.00	0.01 (0.00)*	0.00	0.00 (0.01)	0.61
Gender (female)	0.09 (0.17)	0.61	0.09 (0.08)	0.21	-0.32 (0.24)	0.18
Education	0.04 (0.03)	0.09	0.08 (0.01)*	0.00	0.11 (0.03)*	0.00
APOE4	0.28 (0.13)*	0.03	0.36 (0.06)*	0.00	0.40 (0.19)*	0.03
Shared score (ξ_{i1})	0.62 (0.04)*	0.00	0.58 (0.02)*	0.00	0.81 (0.06)*	0.00
Shared score (ξ_{i2})	0.59 (0.24)*	0.01	0.36 (0.07)*	0.00	0.18 (0.31)	0.56
Shared score (ξ_{i3})	-1.18 (0.24)*	0.00	-1.82 (0.12)*	0.00	-0.58 (0.38)	0.12
Shared score (ξ_{i4})	2.54 (0.65)*	0.00	3.58 (0.29)*	0.00	-3.73 (0.62)*	0.00

A.5 Additional Simulation Results

This section evaluates the model's performance across 100 simulation replicates, with key findings illustrated in several figures to highlight its accuracy and robustness in estimating critical parameters. Figure A.4 shows the estimated eigenfunctions across replicates. The solid red lines represent the true eigenfunctions, the dashed blue lines depict their mean

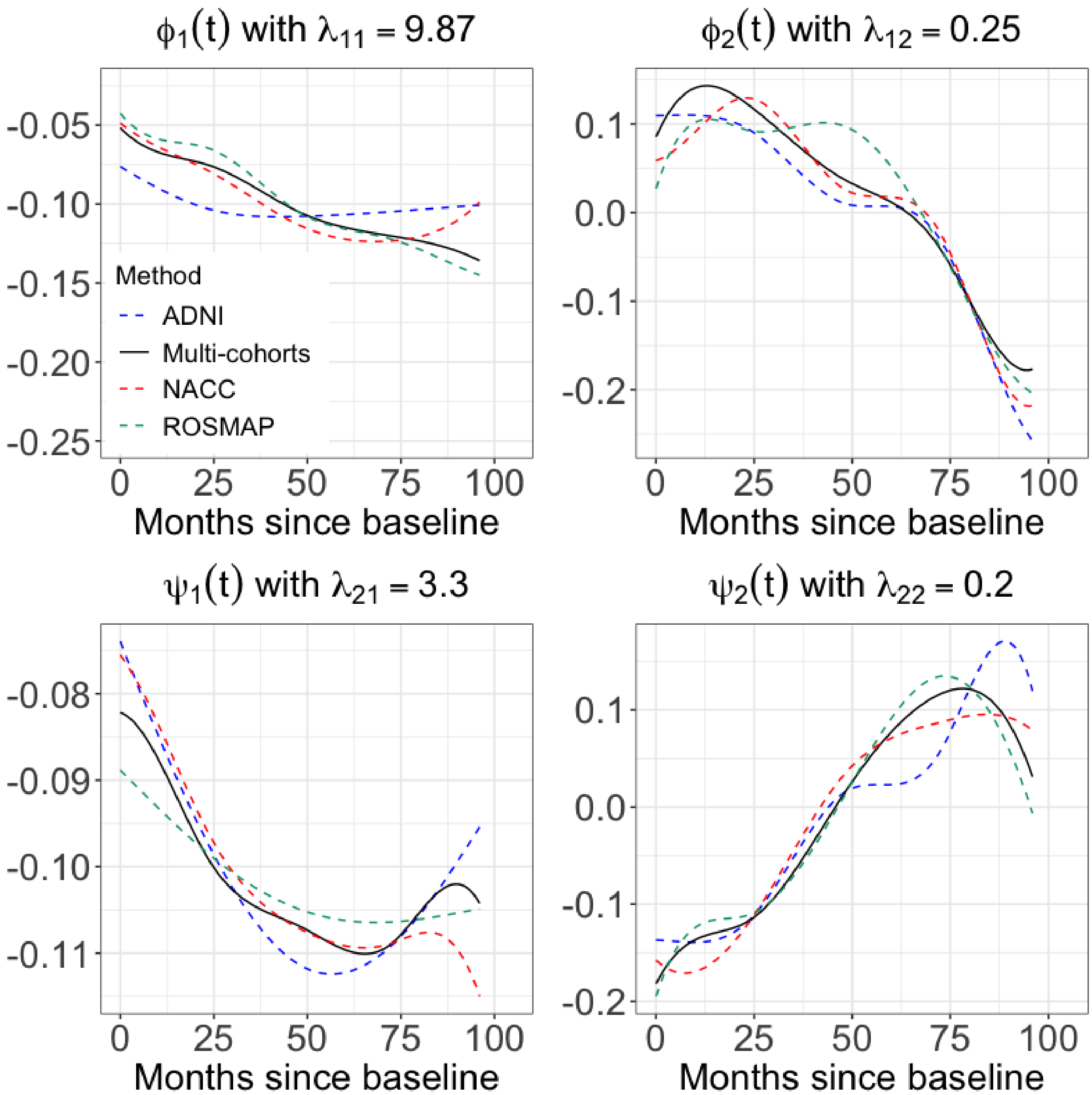


Figure A.2: First two eigenfunctions ($\phi_1(t)$ and $\phi_2(t)$) with associated eigenvalues (λ_{11} and λ_{12}) estimated from the multi-cohort FJM (solid lines) and separate FJM (dashed lines). These eigenfunctions represent shared patterns of variability across longitudinal outcomes.

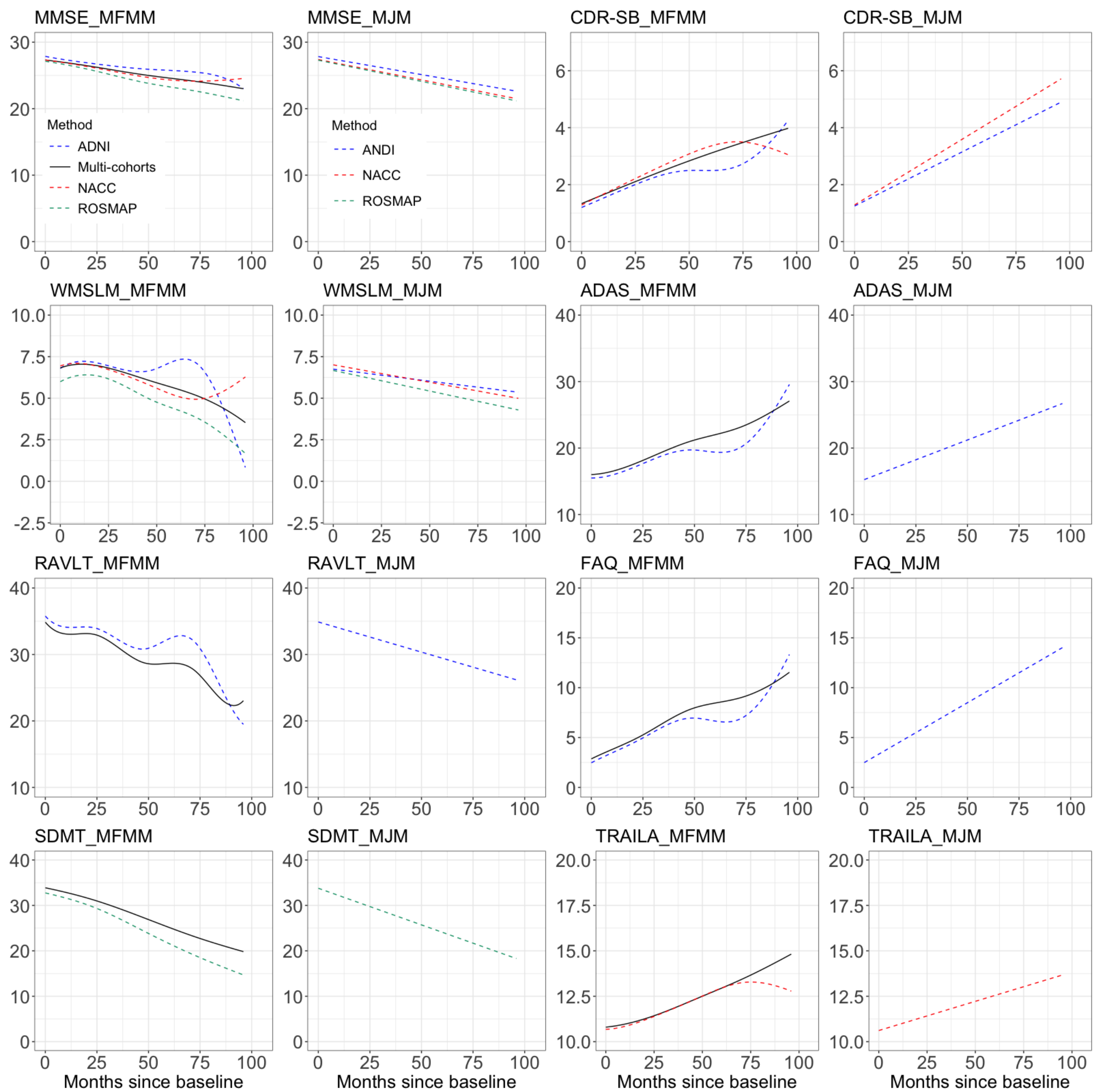


Figure A.3: Estimated mean trajectories of longitudinal outcomes from multi-cohort FJM with common MFMM. Solid lines represent multi-cohort FJM estimates, and dashed lines represent separate FJM estimates. Colored lines correspond to cohorts: ADNI (blue), NACC (red), and ROSMAP (green). Columns 1 and 3 compare multi-cohort and separate FJM, while columns 2 and 4 compare MJM results.

estimates, and the gray lines correspond to individual replicates. The close alignment of the mean estimates with the true eigenfunctions confirms the model's robustness in capturing shared variation patterns across longitudinal outcomes.

Figure A.5 provides boxplots of the estimated eigenvalues. The close agreement between the red lines (true values) and the boxplot medians validates the model's stability and accuracy in capturing variance structures. Figure A.6 displays the estimated scaling parameters for different outcomes, with β_1 fixed at 1. The narrow variability in the estimates suggests the model effectively captures outcome-specific scaling properties, supporting its capability to handle multiple longitudinal outcomes. Figure A.7 presents boxplots of the estimated Cox regression coefficients. The close alignment of the estimates with true values underscores the model's ability to link longitudinal trajectories to survival outcomes while addressing cohort-specific variations. Figure A.8 illustrates the estimated white noise variances for each outcome. The tight clustering of estimates around true values demonstrates the model's precision in capturing measurement noise.

These findings collectively affirm the proposed model's effectiveness in handling complex multi-cohort longitudinal and survival data, ensuring accurate parameter estimation while maintaining interpretability and statistical power.

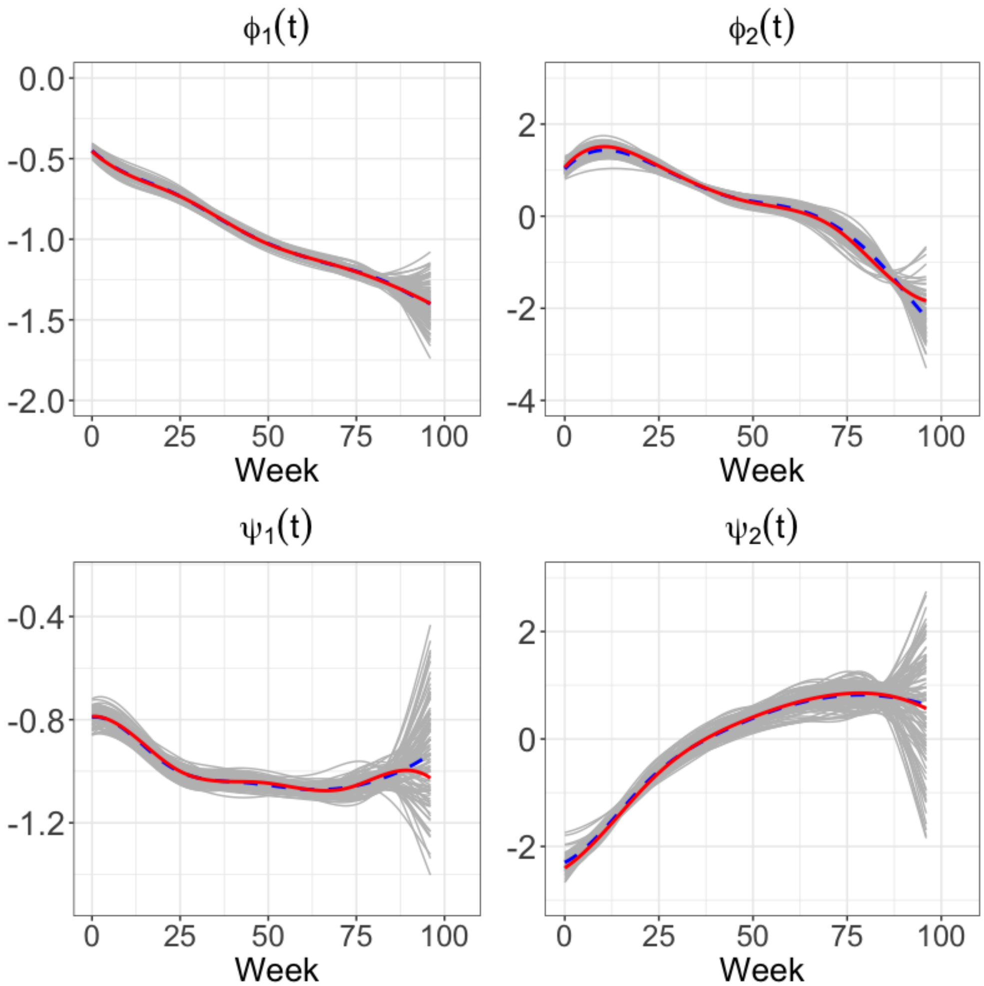


Figure A.4: Estimated eigenfunctions across 100 replications. Gray lines: estimates from 100 replicates; dashed blue lines: the average estimated eigenfunctions; and solid red lines: the true eigenfunctions.

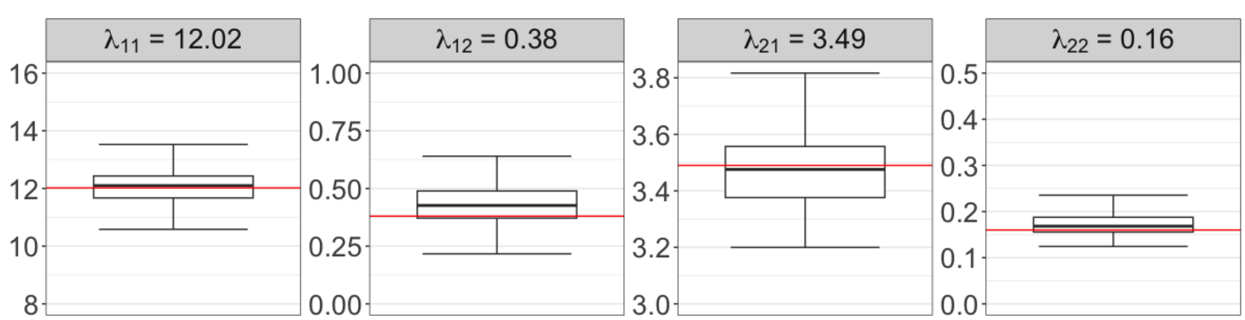


Figure A.5: Boxplots of estimated eigenvalues across 100 replicates.

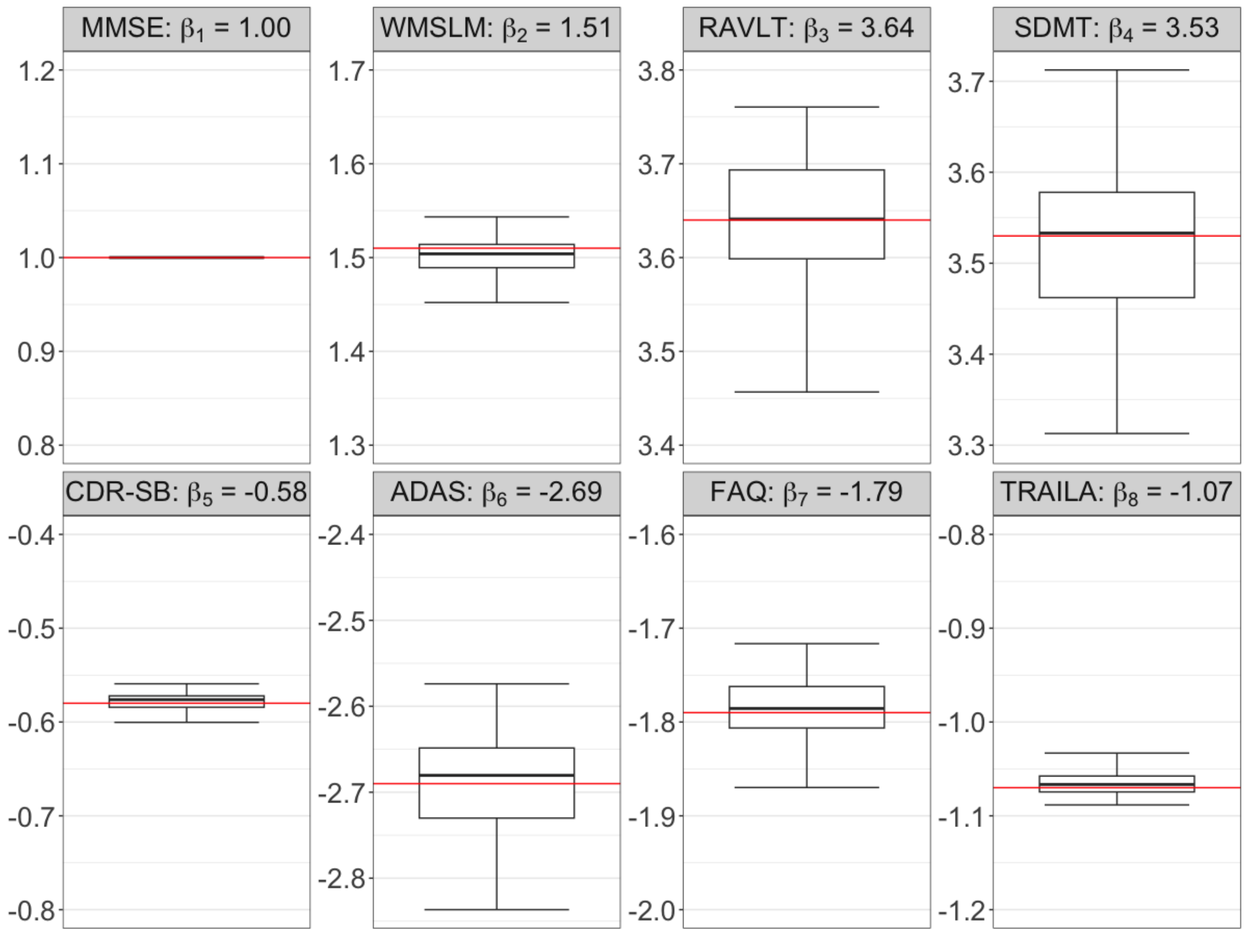


Figure A.6: Boxplots of estimated outcome-specific scaling parameter ($\boldsymbol{\beta}$) across 100 replicates. Note that β_1 is fixed at 1 and thus exhibits no variability.

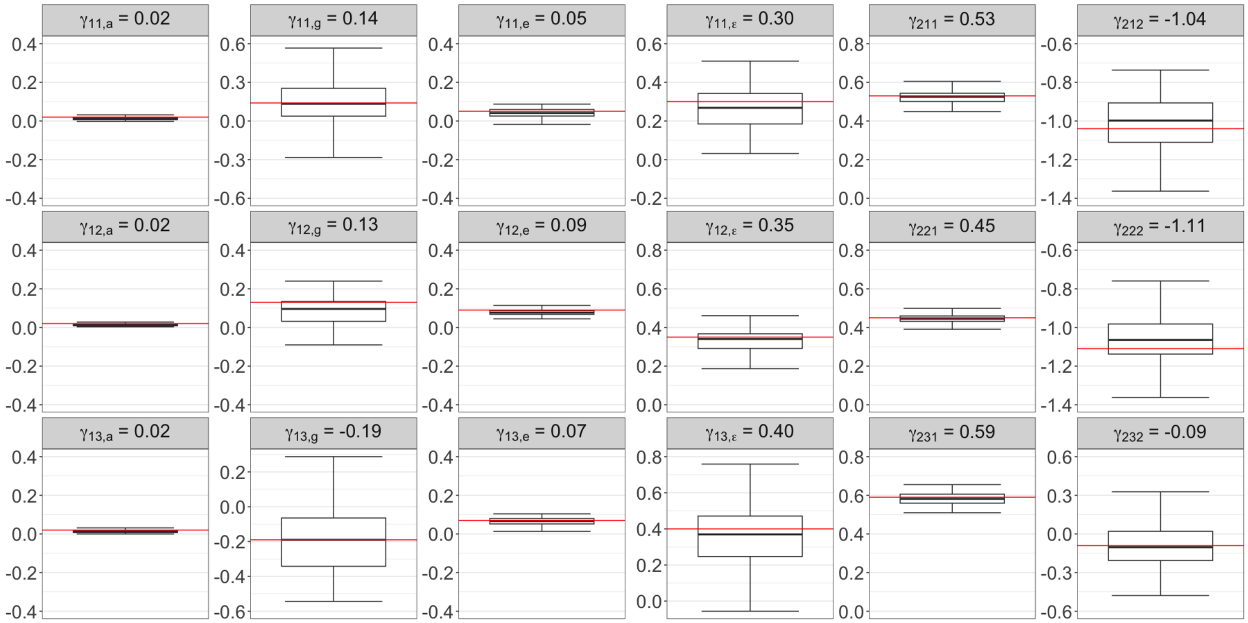


Figure A.7: Boxplots of estimated Cox regression coefficients ($\boldsymbol{\gamma}$) across 100 replicates.

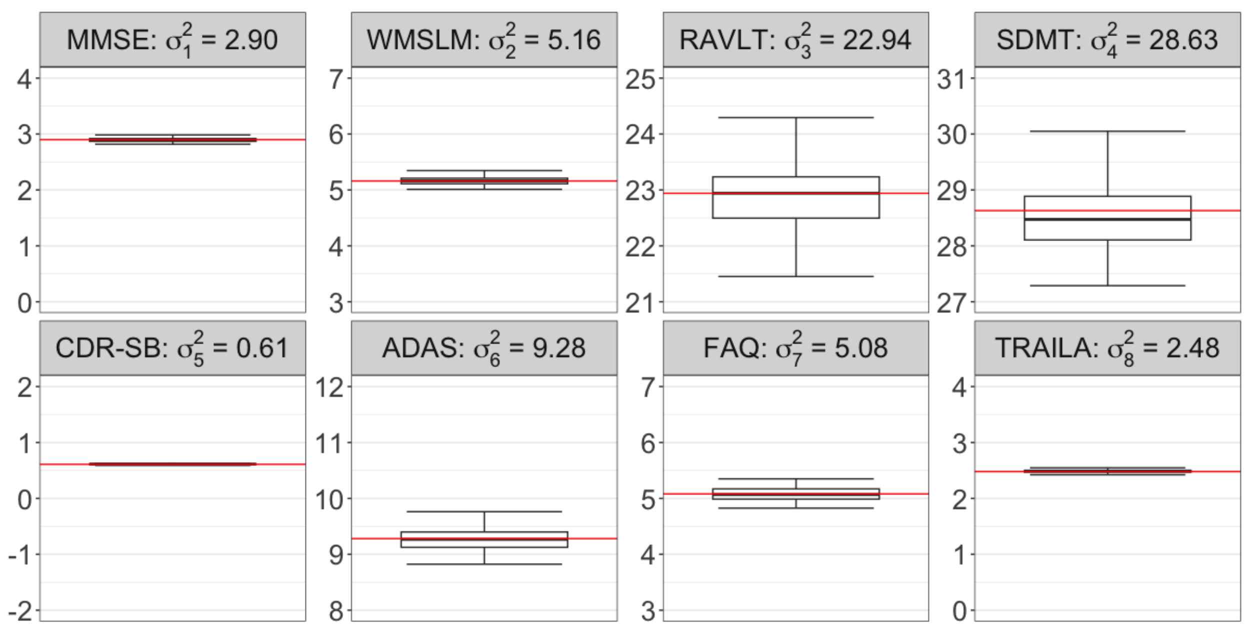


Figure A.8: Boxplots of estimated white noise variance (σ_j^2) across 100 replicates.

APPENDIX

B

SUPPLEMENT TO CHAPTER 3

B.1 Joint Distribution of “Pseudo" Normal Outcomes and Principal Scores

The vector of “pseudo" normal observations $\boldsymbol{Y}_i^* = \boldsymbol{\eta}_{0s(i)p(i)} + \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i = \{\epsilon_i(T_{i1}), \ldots, \epsilon_i(T_{im_i})\}^\top$ is the vector of errors. The joint distribution of $\boldsymbol{Y}_i^*$ and $\boldsymbol{\xi}_i$ is

$$\begin{pmatrix} \boldsymbol{Y}_i^* \\ \boldsymbol{\xi}_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\eta}_{0s(i)p(i)} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \operatorname{Cov}\left(\boldsymbol{Y}_i^*\right) & \operatorname{Cov}\left(\boldsymbol{Y}_i^*, \boldsymbol{\xi}_i\right) \\ \operatorname{Cov}\left(\boldsymbol{\xi}_i, \boldsymbol{Y}_i^*\right) & \operatorname{Cov}\left(\boldsymbol{\xi}_i\right) \end{pmatrix}\right],$$

where

$$\operatorname{Cov}\left(\boldsymbol{Y}_i^*\right) = \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\Lambda} \boldsymbol{\Theta}^\top \boldsymbol{B}_i^\top + \boldsymbol{\Sigma}_i$$

$$\operatorname{Cov}\left(\boldsymbol{\xi}_i\right) = \boldsymbol{\Lambda}$$

$$\operatorname{Cov}\left(\boldsymbol{Y}_i^*, \boldsymbol{\xi}_i\right) = \operatorname{Cov}\left(\boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \boldsymbol{\xi}_i\right) = \boldsymbol{B}_i \boldsymbol{\Theta} \operatorname{Cov}\left(\boldsymbol{\xi}_i, \boldsymbol{\xi}_i\right) = \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\Lambda}.$$

Therefore, $\boldsymbol{\xi}_i \mid \boldsymbol{Y}_i^*$ follows normal distribution with mean $E_{\boldsymbol{\xi}_i}(\boldsymbol{\xi}_i \mid \boldsymbol{Y}_i^*)$ and covariance $\operatorname{Cov}_{\boldsymbol{\xi}_i}(\boldsymbol{\xi}_i \mid \boldsymbol{Y}_i^*)$. Let $\boldsymbol{R}_i = \boldsymbol{B}_i \boldsymbol{\Theta} \boldsymbol{\Lambda} \boldsymbol{\Theta}^\top \boldsymbol{B}_i^\top + \boldsymbol{\Sigma}_i$. By conditional expectation, we have

$$E_{\boldsymbol{\xi}_i}(\boldsymbol{\xi}_i \mid \boldsymbol{Y}_i^*) = \operatorname{Cov}\left(\boldsymbol{\xi}_i, \boldsymbol{Y}_i^*\right) \operatorname{Cov}\left(\boldsymbol{Y}_i^*\right)^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\eta}_{0s(i)p(i)}) = \boldsymbol{\Lambda} \boldsymbol{\Theta}^\top \boldsymbol{B}_i^\top \boldsymbol{R}_i^{-1} (\boldsymbol{Y}_i^* - \boldsymbol{\eta}_{0s(i)p(i)})$$

$$\text{Cov}_{\xi_i}(\xi_i \mid \boldsymbol{Y}_i^*) = \text{Cov}\big(\xi_i\big) - \text{Cov}\big(\xi_i, \boldsymbol{Y}_i^*\big)\text{Cov}\big(\boldsymbol{Y}_i^*\big)^{-1}\text{Cov}\big(\boldsymbol{Y}_i^*, \xi_i\big) = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Theta}^\top \boldsymbol{B}_i^\top \boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Theta}\boldsymbol{\Lambda}.$$

B.2 Details of Updating “Pseudo" Normal Outcomes

In the “pseudo" normal approximation, we need $\boldsymbol{\Theta}\text{Var}\big(\xi_i - \widehat{\xi}_i\big)\boldsymbol{\Theta}^\top$ to update the “pseudo" normal data. Let $\widehat{\boldsymbol{\eta}}_i = \big\{\widehat{\eta}_i(T_{i1}),\ldots,\widehat{\eta}_i(T_{im_i})\big\}^{\text{T}}$. Denote by $\boldsymbol{\Sigma}_i = \text{diag}\{e^{-\widehat{\eta}_i}\}$ the covariance matrix of errors. Let $\boldsymbol{B}_i$ be the $m_i \times q$ matrix with rows $\big\{\widetilde{\boldsymbol{b}}(T_{i1}),\ldots,\widetilde{\boldsymbol{b}}\big(T_{im_i}\big)\big\}^{\text{T}}$. Define $\boldsymbol{\Psi} = \text{Cov}(\boldsymbol{\Theta}\xi_i, \boldsymbol{\Theta}\xi_i) = \boldsymbol{\Theta}\text{Cov}(\xi_i,\xi_i)\boldsymbol{\Theta}^\top = \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top$, and $\boldsymbol{R}_i = \boldsymbol{B}_i\boldsymbol{\Psi}\boldsymbol{B}_i^\top + \boldsymbol{\Sigma}_i$. Let $\boldsymbol{u}_i = \boldsymbol{\Theta}\xi_i$, then $\boldsymbol{\Theta}\text{Var}\big(\xi_i - \widehat{\xi}_i\big)\boldsymbol{\Theta}^\top = \text{Cov}(\boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i, \boldsymbol{u}_i - \widehat{\boldsymbol{u}}_i)$.

For all subjects with school s and position p, since the mean function $\eta_{0sp}(t)$ can be approximated by $\widetilde{\boldsymbol{b}}(t)^\top\boldsymbol{\beta}_{sp}$, where $\boldsymbol{\beta}_{sp}$ are coefficient vectors, we can directly look at the linear mixed model

$$\boldsymbol{Y}_i = \boldsymbol{B}_i\boldsymbol{\beta}_{sp} + \boldsymbol{B}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\epsilon}_i = \big\{\epsilon_1,\cdots,\epsilon_{m_i}\big\}^\top$ and follow $N(\boldsymbol{0},\boldsymbol{\Sigma}_i)$, and $\boldsymbol{u}_i$ follow $N(\boldsymbol{0},\boldsymbol{\Psi})$.

By results of linear mixed model, we have $\widehat{\boldsymbol{\beta}}_{sp} = (\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i)^{-1}(\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{Y}_i)$ and $\widehat{\boldsymbol{u}}_i = \boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{B}_i\widehat{\boldsymbol{\beta}}_{sp})$. Let $\boldsymbol{S} = (\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i)^{-1}$ and $\boldsymbol{\gamma}_i = \boldsymbol{B}_i\boldsymbol{u}_i + \boldsymbol{\epsilon}_i$, then $\widehat{\boldsymbol{\beta}}_{sp} = \boldsymbol{S}(\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{Y}_i) = \boldsymbol{S}(\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{B}_i\boldsymbol{\beta}_{sp} + \boldsymbol{\gamma}_i)) = \boldsymbol{\beta}_{sp} + \boldsymbol{S}(\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{\gamma}_i)$.

Let $\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{B}_i\widehat{\boldsymbol{\beta}}_{sp} = \boldsymbol{B}_i\boldsymbol{\beta}_{sp} + \boldsymbol{\gamma}_i - \boldsymbol{B}_i\boldsymbol{\beta}_{sp} - \boldsymbol{B}_i\boldsymbol{S}(\sum_i \boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{\gamma}_i) = (\boldsymbol{I} - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1})\boldsymbol{\gamma}_i - \boldsymbol{B}_i\boldsymbol{S}\sum_{j\neq i}\boldsymbol{B}_j^\top\boldsymbol{R}_j^{-1}\boldsymbol{\gamma}_j$, then $\widehat{\boldsymbol{u}}_i = \boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{e}_i$.

$$\text{E}(\boldsymbol{e}_i) = \boldsymbol{0},$$

$$\begin{aligned}
\text{Var}(\boldsymbol{e}_i) &= (\boldsymbol{I} - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1})\boldsymbol{R}_i(\boldsymbol{I} - \boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top) + \boldsymbol{B}_i\boldsymbol{S}(\sum_{j\neq i}\boldsymbol{B}_j^\top\boldsymbol{R}_j^{-1}\boldsymbol{R}_j\boldsymbol{R}_j^{-1}\boldsymbol{B}_j)\boldsymbol{S}\boldsymbol{B}_i^\top \\
&= \boldsymbol{R}_i - 2\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top + \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top + \boldsymbol{B}_i\boldsymbol{S}(\sum_{j\neq i}\boldsymbol{B}_j^\top\boldsymbol{R}_j^{-1}\boldsymbol{B}_j)\boldsymbol{S}\boldsymbol{B}_i^\top \\
&= \boldsymbol{R}_i - 2\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top + \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{S}^{-1}\boldsymbol{S}\boldsymbol{B}_i^\top \\
&= \boldsymbol{R}_i - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top.
\end{aligned}$$

Then

$$\text{Var}(\widehat{\boldsymbol{u}}_i) = \boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top)\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi},$$

$$\text{Cov}(\boldsymbol{e}_i, \boldsymbol{u}_i) = (\boldsymbol{I} - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1})\text{Cov}(\boldsymbol{\gamma}_i, \boldsymbol{u}_i) = (\boldsymbol{I} - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1})\boldsymbol{B}_i\boldsymbol{\Psi},$$

$$\text{Cov}(\widehat{\boldsymbol{u}}_i, \boldsymbol{u}_i) = \boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\text{Cov}(\boldsymbol{e}_i, \boldsymbol{u}_i) = \boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{I} - \boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1})\boldsymbol{B}_i\boldsymbol{\Psi}.$$

So

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i,\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i) &= \mathrm{Var}(\boldsymbol{u}_i)+\mathrm{Var}(\widehat{\boldsymbol{u}}_i)-2\mathrm{Cov}(\widehat{\boldsymbol{u}}_i,\boldsymbol{u}_i)\\ &= \boldsymbol{\Psi}+\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i-\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top)\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}-2\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{I}-\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1})\boldsymbol{B}_i\boldsymbol{\Psi}\\ &= \boldsymbol{\Psi}+\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}-\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}\\ &\quad -2\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}+2\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}\\ &= \boldsymbol{\Psi}-\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}+\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{S}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}.\end{aligned}$$

Additionally, if we directly ignore the impact of fixed effects variability and assume $\boldsymbol{Y}_i = \boldsymbol{B}_i\boldsymbol{u}_i+\boldsymbol{\epsilon}_i$, the resulting estimate for $\widehat{\boldsymbol{u}}_i$ is $\widehat{\boldsymbol{u}}_i=\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{Y}_i$. Then

$$\mathrm{Var}(\widehat{\boldsymbol{u}}_i)=\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}(\boldsymbol{R}_i)\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}=\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi},$$

$$\mathrm{Cov}(\widehat{\boldsymbol{u}}_i,\boldsymbol{u}_i)=\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\mathrm{Cov}(\boldsymbol{B}_i\boldsymbol{u}_i+\boldsymbol{\epsilon}_i,\boldsymbol{u}_i)=\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}.$$

So we have

$$\begin{aligned}\boldsymbol{\Theta}\mathrm{Var}\left(\xi_i-\widehat{\xi}_i\right)\boldsymbol{\Theta}^\top &= \mathrm{Cov}(\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i,\boldsymbol{u}_i-\widehat{\boldsymbol{u}}_i)\\ &= \mathrm{Var}(\boldsymbol{u}_i)+\mathrm{Var}(\widehat{\boldsymbol{u}}_i)-2\mathrm{Cov}(\widehat{\boldsymbol{u}}_i,\boldsymbol{u}_i)\\ &= \boldsymbol{\Psi}+\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}-2\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}\\ &= \boldsymbol{\Psi}-\boldsymbol{\Psi}\boldsymbol{B}_i^\top\boldsymbol{R}_i^{-1}\boldsymbol{B}_i\boldsymbol{\Psi}.\end{aligned}$$

B.3 Details on M-step

In the M step of the expectation-maximization (EM) algorithm, $\boldsymbol{A},\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\boldsymbol{U}^{(3)},\boldsymbol{\Theta}$ can be estimated iteratively by minimizing

$$\begin{aligned}&-\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\mathbb{E}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}\left[\left(\boldsymbol{Y}_i^*(T_{ij})-\eta_{0s(i)p(i)}(T_{ij})-\sum_{\ell=1}^{L}\xi_{i\ell}\widetilde{\boldsymbol{b}}^\top(T_{ij})\boldsymbol{\theta}_\ell\right)^2\right]\\ &+\sum_{r_3=1}^{R_3}\tau_{0r_3}\boldsymbol{u}_{r_3}^{(3)\top}\boldsymbol{P}\boldsymbol{u}_{r_3}^{(3)}+\sum_{\ell=1}^{L}\tau_\ell\boldsymbol{\theta}_\ell^\top P\boldsymbol{\theta}_\ell,\end{aligned}$$

where

$$\eta_{0s(i)p(i)}(T_{ij})=\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3}a_{r_1r_2r_3}U_{s(i)r_1}^{(1)}U_{p(i)r_2}^{(2)}U_{kr_3}^{(3)}\right\}.$$

We estimate $\boldsymbol{U}^{(3)}$ and $\boldsymbol{\Theta}$ using P-splines (Eilers and Marx 1996b) with the smoothing parameters. P-splines can provide stable estimators which facilitate the convergence of the iteration algorithm.

For mean functions, given $\boldsymbol{\Theta}$, we estimate $\boldsymbol{A}, \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \boldsymbol{U}^{(3)}$ by minimizing

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{\boldsymbol{Y}_i^*(T_{ij})-\sum_{\ell=1}^{L}\mathbb{E}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}(\xi_{i\ell})\widetilde{\boldsymbol{b}}^\top(T_{ij})\widehat{\boldsymbol{\theta}}_\ell-\eta_{0s(i)p(i)}(T_{ij})\right\}^2\\&+\sum_{r_3=1}^{R_3}\tau_{0r_3}\boldsymbol{u}_{r_3}^{(3)^\top}\boldsymbol{P}\boldsymbol{u}_{r_3}^{(3)},\end{aligned}\tag{B.1}$$

where $\boldsymbol{P}$ is a penalty matrix and τ_{0r_3} are smoothing parameters.

Let $w_{ij}=\boldsymbol{Y}_i^*(T_{ij})-\sum_{\ell=1}^{L}\mathbb{E}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}(\xi_{i\ell})\widetilde{\boldsymbol{b}}^\top(T_{ij})\widehat{\boldsymbol{\theta}}_\ell$, then model (B.1) becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3}a_{r_1r_2r_3}U_{s(i)r_1}^{(1)}U_{p(i)r_2}^{(2)}U_{kr_3}^{(3)}\right\}\right\}^2+\sum_{r_3=1}^{R_3}\tau_{0r_3}\boldsymbol{u}_{r_3}^{(3)^\top}\boldsymbol{P}\boldsymbol{u}_{r_3}^{(3)}.$$

First, given $\boldsymbol{A}, \boldsymbol{U}^{(1)}$ and $\boldsymbol{U}^{(2)}$, we estimate $\boldsymbol{U}^{(3)}$ by minimizing

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_3=1}^{R_3}\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\widehat{a}_{r_1r_2r_3}\widehat{U}_{s(i)r_1}^{(1)}\widehat{U}_{p(i)r_2}^{(2)}\right\}\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})U_{kr_3}^{(3)}\right\}^2\\&+\sum_{r_3=1}^{R_3}\tau_{0r_3}\boldsymbol{u}_{r_3}^{(3)^\top}\boldsymbol{P}\boldsymbol{u}_{r_3}^{(3)}\\&=\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_3=1}^{R_3}\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\widehat{a}_{r_1r_2r_3}\widehat{U}_{s(i)r_1}^{(1)}\widehat{U}_{p(i)r_2}^{(2)}\right\}\widetilde{\boldsymbol{b}}^\top(T_{ij})\boldsymbol{u}_{r_3}\right\}^2+\sum_{r_3=1}^{R_3}\tau_{0r_3}\boldsymbol{u}_{r_3}^{(3)^\top}\boldsymbol{P}\boldsymbol{u}_{r_3}^{(3)}.\end{aligned}$$

We update the columns of $\boldsymbol{U}^{(3)}$ one at a time, while holding all other columns fixed. So given $\boldsymbol{A}, \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}$ and $\boldsymbol{u}_{r_3'}^{(3)}$, $r_3'\neq r_3$, r_3 is estimated by minimizing

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\Bigg\{w_{ij}-\sum_{r_3'\neq r_3}&\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\widehat{a}_{r_1r_2r_3'}\widehat{U}_{s(i)r_1}^{(1)}\widehat{U}_{p(i)r_2}^{(2)}\right\}\widetilde{\boldsymbol{b}}^\top(T_{ij})\boldsymbol{u}_{r_3'}\\&-\left\{\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\widehat{a}_{r_1r_2r_3}\widehat{U}_{s(i)r_1}^{(1)}\widehat{U}_{p(i)r_2}^{(2)}\right\}\widetilde{\boldsymbol{b}}^\top(T_{ij})\boldsymbol{u}_{r_3}\Bigg\}^2+\tau_{0r_3}\boldsymbol{u}_{r_3}^{(3)^\top}\boldsymbol{P}\boldsymbol{u}_{r_3}^{(3)}.\end{aligned}\tag{B.2}$$

Second, given $\boldsymbol{A}, \boldsymbol{U}^{(2)}$ and $\boldsymbol{U}^{(3)}$, we estimate $\boldsymbol{U}^{(1)}$ by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_1=1}^{R_1}\left\{\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3}\widehat{a}_{r_1r_2r_3}\widehat{U}_{p(i)r_2}^{(2)}\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\widehat{U}_{kr_3}^{(3)}\right\}U_{s(i)r_1}^{(1)}\right\}^2.$$

We update the sth row of $\boldsymbol{U}^{(1)}$ using all subject from school s, that is minimizing

$$\sum_{i:s(i)=s}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_1=1}^{R_1}\Big\{\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3}\widehat{a}_{r_1r_2r_3}\widehat{U}^{(2)}_{p(i)r_2}\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\widehat{U}^{(3)}_{kr_3}\Big\}U^{(1)}_{s(i)r_1}\right\}^2. \tag{B.3}$$

Third, similarly, given $\boldsymbol{A}$, $\boldsymbol{U}^{(1)}$ and $\boldsymbol{U}^{(3)}$, we estimate $\boldsymbol{U}^{(2)}$ by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_2=1}^{R_2}\Big\{\sum_{r_1=1}^{R_1}\sum_{r_3=1}^{R_3}\widehat{a}_{r_1r_2r_3}\widehat{U}^{(1)}_{s(i)r_1}\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\widehat{U}^{(3)}_{kr_3}\Big\}U^{(2)}_{p(i)r_2}\right\}^2.$$

We update the pth row of $\boldsymbol{U}^{(2)}$ using all subject from position p, that is minimizing

$$\sum_{i:p(i)=p}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_2=1}^{R_2}\Big\{\sum_{r_1=1}^{R_1}\sum_{r_3=1}^{R_3}\widehat{a}_{r_1r_2r_3}\widehat{U}^{(1)}_{s(i)r_1}\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\widehat{U}^{(3)}_{kr_3}\Big\}U^{(2)}_{p(i)r_2}\right\}^2. \tag{B.4}$$

Fourth, given $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$ and $\boldsymbol{U}^{(3)}$, we estimate $\boldsymbol{A}$ by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\left\{w_{ij}-\sum_{r_1=1}^{R_1}\sum_{r_2=1}^{R_2}\sum_{r_3=1}^{R_3}\Big\{\widehat{U}^{(1)}_{s(i)r_1}\widehat{U}^{(2)}_{p(i)r_2}\sum_{k=1}^{K}\widetilde{B}_k(T_{ij})\widehat{U}^{(3)}_{kr_3}\Big\}a_{r_1r_2r_3}\right\}^2. \tag{B.5}$$

For eigenfunctions, given $\boldsymbol{A}$, $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$, $\boldsymbol{U}^{(3)}$ and $\boldsymbol{\theta}_{\ell'}$, $\ell'\neq\ell$, $\boldsymbol{\theta}_\ell$ is estimated by minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\mathbb{E}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}\left\{\boldsymbol{Y}_i^*(T_{ij})-\widehat{\eta}_{0s(i)p(i)}(T_{ij})-\sum_{\ell'\neq\ell}\xi_{i\ell'}\widetilde{\boldsymbol{b}}^\top(T_{ij})\widehat{\boldsymbol{\theta}}_{\ell'}-\xi_{i\ell}\widetilde{\boldsymbol{b}}^\top(T_{ij})\boldsymbol{\theta}_\ell\right\}^2$$
$$+\tau_\ell\boldsymbol{\theta}_\ell^\top P\boldsymbol{\theta}_\ell.$$

Minimizing the above weighted least square is equivalent to minimizing

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\exp\{\widehat{\eta}_i(T_{ij})\}\bigg[\mathbb{E}^{-\frac{1}{2}}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}(\xi_{i\ell}^2)\Big\{\big(\boldsymbol{Y}_i^*(T_{ij})-\widehat{\eta}_{0s(i)p(i)}(T_{ij})\big)\mathbb{E}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}(\xi_{i\ell})\\&-\sum_{\ell'\neq\ell}\mathbb{E}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}(\xi_{i\ell'}\xi_{i\ell})\widetilde{\boldsymbol{b}}^\top(T_{ij})\widehat{\boldsymbol{\theta}}^{(d)}_{\ell'}\Big\}-\mathbb{E}^{\frac{1}{2}}_{\xi_i|\boldsymbol{Y}_i^*,\widehat{\boldsymbol{\Omega}}}(\xi_{i\ell}^2)\widetilde{\boldsymbol{b}}^\top\big(T_{ij}\big)\boldsymbol{\theta}_\ell\bigg]^2+\tau_\ell\boldsymbol{\theta}_\ell^\top\boldsymbol{P}\boldsymbol{\theta}_\ell.\end{aligned} \tag{B.6}$$

Minimizing (B.3), (B.4) and (B.5) can by solved by weighted linear regression via the *lm* function in R. Minimizing (B.2) and (B.6) can be solved directly via the *gam* function in the R package *mgcv* (Wood 2011; Wood et al. 2016). The smoothing parameters τ are selected by using Restricted Maximum Likelihood in each iteration. We repeat the procedure for $\boldsymbol{U}^{(3)}$, $\boldsymbol{U}^{(1)}$, $\boldsymbol{U}^{(2)}$, $\boldsymbol{A}$ and each column of $\boldsymbol{\Theta}$ until convergence.

B.4 EM Implementation Details

The convergence criteria of the EM algorithm is specified following Li et al. (2022). We employ three widely used criteria: relative and absolute changes in parameters, and relative changes in the marginal likelihood. The first two are utilized in the R package *JoineRML* (Hickey et al. 2018), and the third is adopted in the R package *JM* (Rizopoulos 2010). The algorithm stops when any one of the three criteria falls below a predetermined threshold, specified as follows:

1. $\max\left\{\frac{\left|\widehat{\mathbf{\Omega}}^{(l+1)}-\widehat{\mathbf{\Omega}}^{(l)}\right|}{\left|\widehat{\mathbf{\Omega}}^{(l)}\right|+\delta_0}\right\} < \delta_1$
2. $\max\left\{\left|\widehat{\mathbf{\Omega}}^{(l+1)}-\widehat{\mathbf{\Omega}}^{(l)}\right|\right\} < \delta_2$
3. $\frac{\left|\ell\left(\widehat{\mathbf{\Omega}}^{(l+1)}\right)-\ell\left(\widehat{\mathbf{\Omega}}^{(l)}\right)\right|}{\left|\ell\left(\widehat{\mathbf{\Omega}}^{(l)}\right)\right|+\delta_0} < \delta_3$

Here, $\widehat{\mathbf{\Omega}}^{(l)}$ represents the estimate of $\mathbf{\Omega}$ at the ℓ th iteration, the maximum is taken over the components of $\mathbf{\Omega}, \ell\left(\widehat{\mathbf{\Omega}}^{(l)}\right)$ is the marginal likelihood at the ℓ th iteration. We set $\delta_0 = 0.001, \delta_1 = 0.005, \delta_2 = 0.001$, $\delta_3 = 10^{-7}$ following the suggestions in Booth and Hobert (1999).

In the Monte Carlo approximation, a higher number of samples may help alleviate numerical errors. However, in practice, a large MC sample in the early iterations is unnecessary as the parameters are still far from the optimal values. Therefore, we incrementally adjust the number of samples to approximate the marginal log-likelihood of our Poisson data as the estimates converge towards the truth. In our implementation, we use Monte Carlo samples with N=2000 during the burn-in iterations 1-20, and subsequently, we set N=10000. The maximum number of iterations is set at 1000.

B.5 Additional Data Application Results

Figure B.1 displays the estimated mean functions of cumulative HIE counts, along with trajectories of subject-specific cumulative HIE counts in each group. Most group mean functions seem to capture the average trend of subject-specific trajectories, indicating reasonable model fit to the data.

Figure B.2 and Figure B.3 present the estimated mean rate of HIEs per practice/game, organized by positions and schools respectively. The estimated mean rate is given by $\widehat{\mu}_{ij} = \exp\{\widehat{\eta}_{0s(i)p(i)}(T_{ij}) + 2^{-1}\widetilde{\boldsymbol{b}}(T_{ij})^\top \widehat{\mathbf{\Theta}}\widehat{\Lambda}\widehat{\mathbf{\Theta}}^\top \widetilde{\boldsymbol{b}}(T_{ij})\}$. These plots convey information consistent with the corresponding plots of the estimated mean functions for cumulative HIE counts, highlighting similar trends and insights.

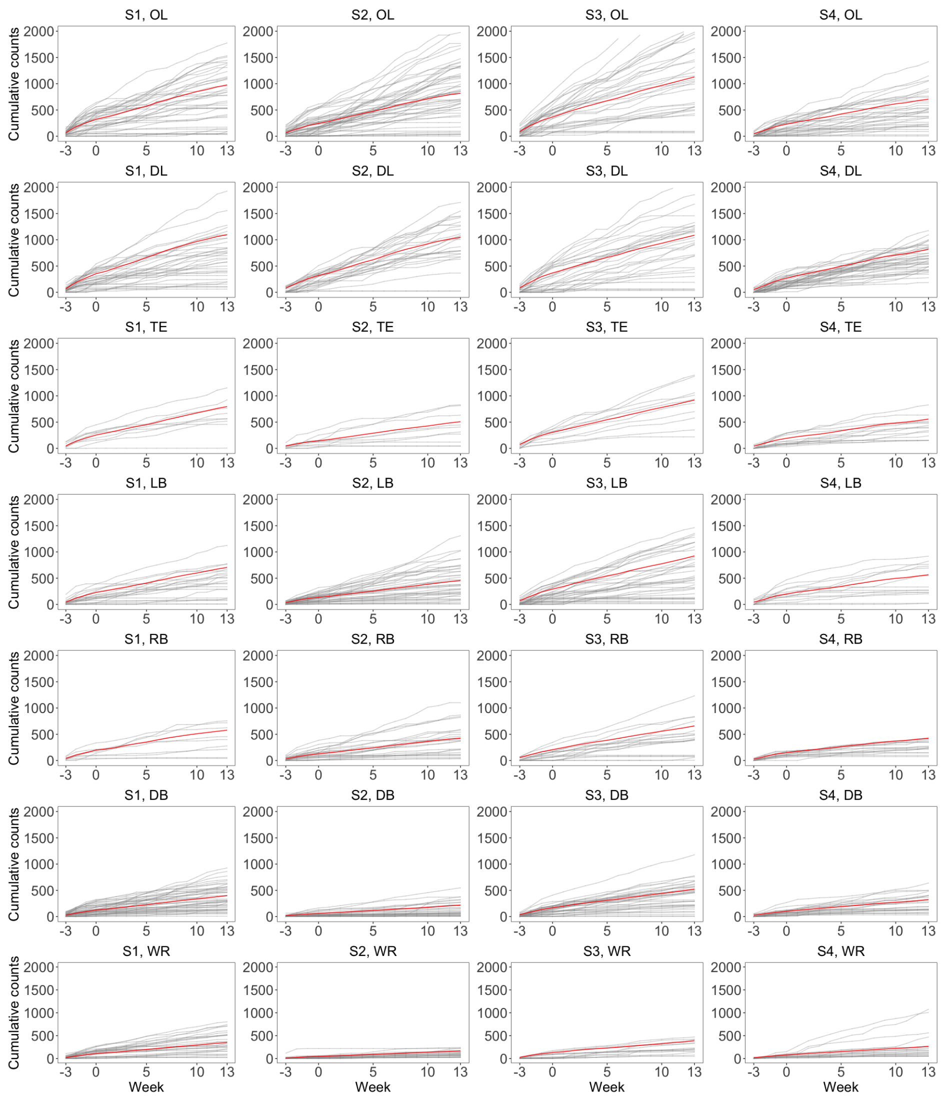


Figure B.1: Estimated mean functions (red) of cumulative counts of HIEs for 7 positions (rows) at 4 schools (columns). Gray lines: observed cumulative counts outcomes of individual subjects.

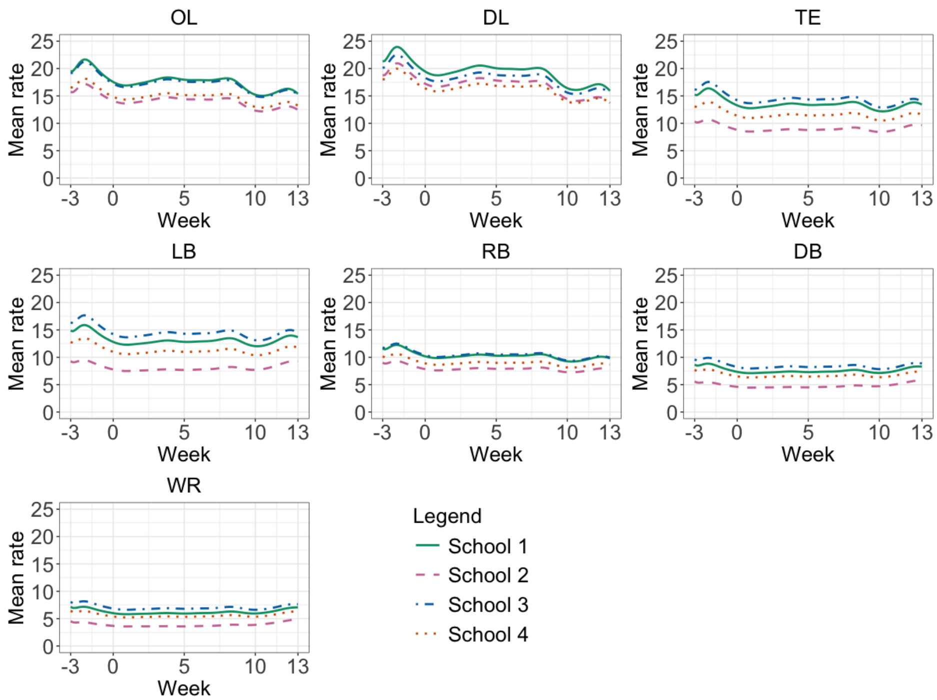


Figure B.2: Estimated mean rate of HIEs per practice/game grouped by positions.

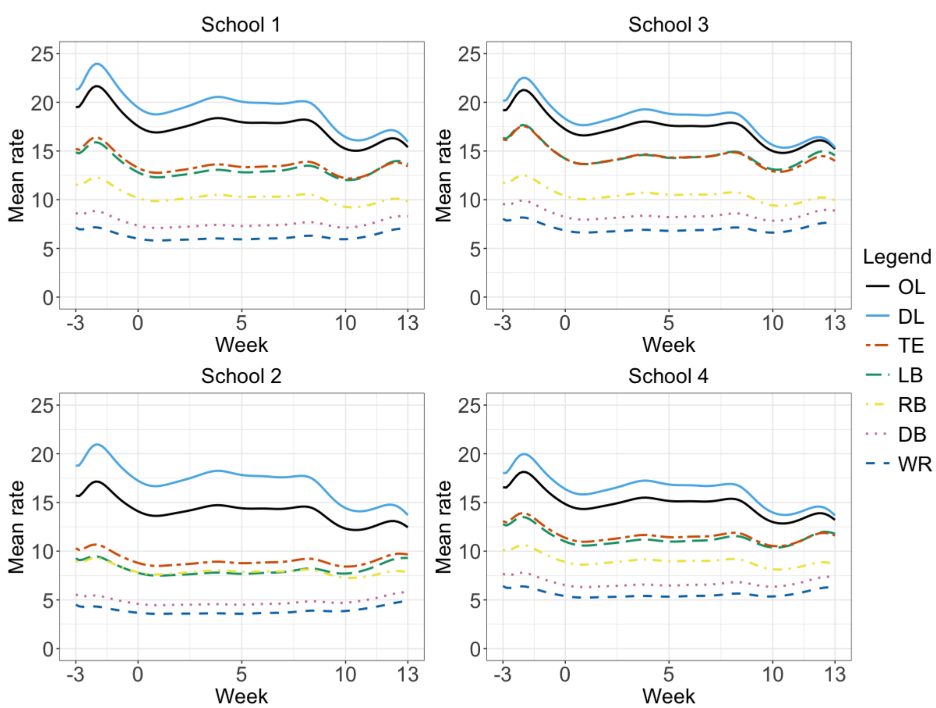


Figure B.3: Estimated mean rate of HIEs per practice/game grouped by schools.

APPENDIX

C

SUPPLEMENT TO CHAPTER 4

C.1 Joint Distribution of Longitudinal Outcomes and FPC Scores

We derive the joint distribution of $f(\boldsymbol{y}_i, \boldsymbol{\xi}_i|\boldsymbol{\Lambda})$, which characterizes the relationship between the observed longitudinal outcomes and the latent principal scores. The observed outcomes follow the equation: $\boldsymbol{y}_i = \boldsymbol{\mu}_i + \boldsymbol{F}_i + \boldsymbol{\Phi}_i \boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^\top$ is the vector of measurement errors. We have

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{\xi}_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu}_i + \boldsymbol{F}_i \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \mathrm{Cov}(\boldsymbol{y}_i) & \mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{\xi}_i) \\ \mathrm{Cov}(\boldsymbol{\xi}_i, \boldsymbol{y}_i) & \mathrm{Cov}(\boldsymbol{\xi}_i) \end{pmatrix}\right],$$

where $\mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{\Phi}_i \boldsymbol{\Lambda} \boldsymbol{\Phi}_i^\top + \sigma^2 \boldsymbol{I}_{m_i} = \boldsymbol{\Phi}_i \boldsymbol{\Lambda} \boldsymbol{\Phi}_i^\top + \boldsymbol{\Sigma}_i$, $\mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{\xi}_i) = \boldsymbol{\Phi}_i \boldsymbol{\Lambda}$ and $\mathrm{Cov}(\boldsymbol{\xi}_i) = \boldsymbol{\Lambda}$. Using conditional expectation, the posterior mean and covariance of $\boldsymbol{\xi}_i$ given $\boldsymbol{y}_i$ are derived as: $\mathbb{E}_i(\boldsymbol{\xi}_i|\boldsymbol{y}_i) = \mathrm{Cov}(\boldsymbol{\xi}_i, \boldsymbol{y}_i)\mathrm{Cov}(\boldsymbol{y}_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i - \boldsymbol{F}_i)$, $\mathrm{Cov}(\boldsymbol{\xi}_i|\boldsymbol{y}_i) = \mathrm{Cov}(\boldsymbol{\xi}_i) - \mathrm{Cov}(\boldsymbol{\xi}_i, \boldsymbol{y}_i)\mathrm{Cov}^{-1}(\boldsymbol{y}_i)\mathrm{Cov}(\boldsymbol{y}_i, \boldsymbol{\xi}_i)$.

C.2 Marginal Log-likelihood

The marginal log-likelihood is a key quantity for model evaluation and comparison, providing a measure of how well the proposed model fits the observed data. Since the exact computation of this likelihood involves intractable integrals, we employ an approximation strategy as detailed below.

The marginal log-likelihood of the data is given by

$$\ell(\widehat{\boldsymbol{\Omega}})=\sum_{i=1}^{n}\log\left\{\int f\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i,\widehat{\boldsymbol{\Sigma}}_i\right) f\left(T_i,\Delta_i \mid \widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3\right) f\left(\boldsymbol{\xi}_i \mid \widehat{\boldsymbol{\Lambda}}\right)\mathrm{d}\boldsymbol{\xi}_i\right\}$$

$$=\sum_{i=1}^{n}\log\left\{f\left(\boldsymbol{y}_i \mid \boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}}\right)\int f\left(\xi_i \mid \boldsymbol{y}_i,\boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}}\right) f\left(T_i,\Delta_i \mid \widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i,\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3\right)\mathrm{d}\xi_i\right\}.$$

Since this integral does not have a closed-form solution, we approximate it using Monte Carlo integration. Specifically, the integral is approximated as:

$$\left\{R^{-1}\sum_{r=1}^{R} f\left(T_i,\Delta_i \mid \widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i^{(r)},\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3\right)\right\},$$

where $\xi_i^{(r)}$ are samples drawn from the posterior distribution $f\left(\xi_i \mid \boldsymbol{y}_i,\boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}}\right)$. Substituting this approximation, the marginal log-likelihood becomes:

$$\ell(\widehat{\boldsymbol{\Omega}})\approx\sum_{i=1}^{n}\left[\log\left\{f\left(\boldsymbol{y}_i \mid \boldsymbol{Z}_i,W_i(s),\widehat{\boldsymbol{\Omega}}\right)\right\}+\log\left\{R^{-1}\sum_{r=1}^{R} f\left(T_i,\Delta_i \mid \widehat{h}_0,\boldsymbol{Z}_i,W_i(s),\xi_i^{(r)},\widehat{\boldsymbol{\gamma}}_2,\widehat{\boldsymbol{\omega}}_2,\widehat{\boldsymbol{\gamma}}_3\right)\right\}\right].$$

C.3 Details on M-step

We estimate $\boldsymbol{\alpha}$, $\boldsymbol{\omega}_1$ and $\boldsymbol{\Theta}$ using P-splines (Eilers and Marx 1996b) with smoothing parameters. P-splines provide stable estimates that enhance the convergence speed of the iterative algorithm, contributing to computational efficiency.

First, we estimate $\boldsymbol{\alpha},\boldsymbol{\gamma}_1$ and $\boldsymbol{\omega}_1$ by minimizing the following penalized least squares objective function:

$$\sum_{i=1}^{n}\left\|\boldsymbol{y}_i-\widehat{\boldsymbol{\Phi}}_i\mathbb{E}_i(\xi_i)-\boldsymbol{B}_i\boldsymbol{\alpha}-\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\boldsymbol{\omega}_1\Big)\mathbf{1}_{m_i}\right\|^2+\tau_\alpha\boldsymbol{\alpha}^\top\boldsymbol{P}\boldsymbol{\alpha}+\tau_1^\beta\boldsymbol{\omega}_1^\top\boldsymbol{P}^\beta\boldsymbol{\omega}_1, \tag{C.1}$$

where $\|\cdot\|$ is the Euclidean norm, $\boldsymbol{P}$ and $\boldsymbol{P}^\beta$ are two penalty matrices, τ_α and τ_1^β are smoothing parameters.

Second, given $\boldsymbol{\theta}_{1\ell'}$, $\ell'\neq\ell$, we estimate $\boldsymbol{\theta}_\ell$ by minimizing:

$$\sum_{i=1}^{n}\mathbb{E}_i\left\|\boldsymbol{y}_i-\widehat{\boldsymbol{\mu}}_i-\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\gamma}}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\widehat{\boldsymbol{\omega}}_1\Big)\mathbf{1}_{m_i}-\sum_{\ell'\neq\ell}\widetilde{\boldsymbol{B}}_i\widehat{\boldsymbol{\theta}}_{\ell'}\xi_{i\ell'}-\widetilde{\boldsymbol{B}}_i\boldsymbol{\theta}_\ell\xi_{i\ell}\right\|^2+\tau_{\theta,\ell}\boldsymbol{\theta}_\ell^\top\boldsymbol{P}\boldsymbol{\theta}_\ell.$$

This is equivalent to minimizing:

$$\sum_{i=1}^{n}\Bigg\| \frac{\mathbb{E}_i(\xi_{i\ell})}{\sqrt{\mathbb{E}_i(\xi_{i\ell}^2)}}\Big(\boldsymbol{y}_i-\widehat{\boldsymbol{\mu}}_i-\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\gamma}}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\widehat{\boldsymbol{\omega}}_1\Big)\mathbf{1}_{m_i}\Big) \\ -\frac{1}{\sqrt{\mathbb{E}_i(\xi_{i\ell}^2)}}\sum_{\ell'\neq\ell}\widetilde{\boldsymbol{B}}_i\widehat{\boldsymbol{\theta}}_{\ell'}\mathbb{E}_i(\xi_{i\ell'}\xi_{i\ell})-\widetilde{\boldsymbol{B}}_i\boldsymbol{\theta}_\ell\sqrt{\mathbb{E}_i(\xi_{i\ell}^2)}\Bigg\|^2+\tau_{\theta,\ell}\boldsymbol{\theta}_\ell^\top\boldsymbol{P}\boldsymbol{\theta}_\ell\,. \tag{C.2}$$

Minimizing (C.1) and (C.2) can be solved directly via the *gam* function in the R package *mgcv* (Wood 2011; Wood et al. 2016). The smoothing parameters are selected locally through Restricted Maximum Likelihood (REML) during each iteration. The procedure is repeated for each column of $\boldsymbol{\Theta}$ until convergence, ensuring robust estimation.

Third, the variance of random noise σ^2 is updated as:

$$\begin{aligned}\widehat{\sigma}^2&=\frac{1}{\sum_{i=1}^n m_i}\sum_{i=1}^{n}\mathbb{E}_i\Big\|\boldsymbol{y}_i-\widehat{\boldsymbol{\mu}}_i-\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\gamma}}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\widehat{\boldsymbol{\omega}}_1\Big)\mathbf{1}_{m_i}-\widehat{\boldsymbol{\Phi}}_i\boldsymbol{\xi}_i\Big\|^2\\ &=\frac{1}{\sum_{i=1}^n m_i}\sum_{i=1}^{n}\Bigg\{\Big\|\boldsymbol{y}_i-\widehat{\boldsymbol{\mu}}_i-\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\gamma}}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\widehat{\boldsymbol{\omega}}_1\Big)\mathbf{1}_{m_i}\Big\|^2-\\ &2\Big(\boldsymbol{y}_i-\widehat{\boldsymbol{\mu}}_i-\boldsymbol{Z}_i^\top\widehat{\boldsymbol{\gamma}}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\widehat{\boldsymbol{\omega}}_1\Big)\mathbf{1}_{m_i}\Big)^\top\big(\widehat{\boldsymbol{\Phi}}_i\mathbb{E}_i(\boldsymbol{\xi}_i)\big)+\mathrm{tr}\big\{\widehat{\boldsymbol{\Phi}}_i^\top\widehat{\boldsymbol{\Phi}}_i\mathbb{E}_i(\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top)\big\}\Bigg\}\,.\end{aligned}$$

Finally, we derive the observed information matrix for $(\boldsymbol{\alpha}^\top,\boldsymbol{\gamma}_1^\top,\boldsymbol{\omega}_1^\top)^\top$. The penalized expected longitudinal log-likelihood is expressed as:

$$\begin{aligned}\ell_1&=-\sum_{i=1}^{n}\Bigg\{\frac{m_i}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\mathbb{E}_i\Big\|\boldsymbol{y}_i-\boldsymbol{B}_i\boldsymbol{\alpha}-\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_1\mathbf{1}_{m_i}-\Big(\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\boldsymbol{\omega}_1\Big)\mathbf{1}_{m_i}-\widetilde{\boldsymbol{B}}_i\boldsymbol{\Theta}\boldsymbol{\xi}_i\Big\|^2\Bigg\}\\ &-\tau_\alpha\boldsymbol{\alpha}^\top\boldsymbol{P}\boldsymbol{\alpha}-\tau_1^\beta\boldsymbol{\omega}_1^\top\boldsymbol{P}^\beta\boldsymbol{\omega}_1-\sum_{\ell=1}^{L}\tau_{\theta,\ell}\boldsymbol{\theta}_\ell^\top\boldsymbol{P}\boldsymbol{\theta}_\ell\\ &=-\sum_{i=1}^{n}\sum_{j=1}^{m_i}\Bigg\{\frac{1}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\mathbb{E}_i\Big(y_{ij}-\boldsymbol{b}(t_{ij})^\top\boldsymbol{\alpha}-\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_1-\frac{1}{J}\sum_{j_s=1}^{J}W_i(s_{j_s})\boldsymbol{b}^\beta(s_{j_s})^\top\boldsymbol{\omega}_1-\widetilde{\boldsymbol{b}}(t_{ij})^\top\boldsymbol{\Theta}\boldsymbol{\xi}_i\Big)^2\Bigg\}\\ &-\tau_\alpha\boldsymbol{\alpha}^\top\boldsymbol{P}\boldsymbol{\alpha}-\tau_1^\beta\boldsymbol{\omega}_1^\top\boldsymbol{P}^\beta\boldsymbol{\omega}_1-\sum_{\ell=1}^{L}\tau_{\theta,\ell}\boldsymbol{\theta}_\ell^\top\boldsymbol{P}\boldsymbol{\theta}_\ell.\end{aligned}$$

The score vector is:

$$
\begin{aligned}
\boldsymbol{S}_1 &= \frac{\partial \ell_1}{\partial(\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}_1^\top, \boldsymbol{\omega}_1^\top)^\top} \\
&= -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Bigg\{ 2\Big(y_{ij} - \boldsymbol{b}(t_{ij})^\top \boldsymbol{\alpha} - \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_1 - \frac{1}{J} \sum_{j_s=1}^{J} W_i(s_{j_s}) \boldsymbol{b}^\beta(s_{j_s})^\top \boldsymbol{\omega}_1 - \widetilde{\boldsymbol{b}}(t_{ij})^\top \boldsymbol{\Theta} \mathbb{E}_i(\boldsymbol{\xi}_i) \Big) \\
&\Big(-\boldsymbol{b}(t_{ij})^\top, -\boldsymbol{Z}_i^\top, -\frac{1}{J} \sum_{j_s=1}^{J} W_i(s_{j_s}) \boldsymbol{b}^\beta(s_{j_s})^\top \Big)^\top \Bigg\} - 2\Big(\tau_\alpha (\boldsymbol{P}\boldsymbol{\alpha})^\top, 0 \cdot \mathbf{1}_P^\top, \tau_1^\beta (\boldsymbol{P}^\beta \boldsymbol{\omega}_1)^\top \Big)^\top.
\end{aligned}
$$

Then the observed information matrix is:

$$
\begin{aligned}
\boldsymbol{I}_1 &= -\frac{\partial S_1}{\partial(\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}_1^\top, \boldsymbol{\omega}_1^\top)} \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Big\{ \Big(-\boldsymbol{b}(t_{ij})^\top, -\boldsymbol{Z}_i^\top, -\frac{1}{J} \sum_{j_s=1}^{J} W_i(s_{j_s}) \boldsymbol{b}^\beta(s_{j_s})^\top \Big)^\top \Big(-\boldsymbol{b}(t_{ij})^\top, -\boldsymbol{Z}_i^\top, -\frac{1}{J} \sum_{j_s=1}^{J} W_i(s_{j_s}) \boldsymbol{b}^\beta(s_{j_s})^\top \Big) \Big\} \\
&+ \text{Blockdiag}\{2\tau_\alpha \boldsymbol{P}, 0 \cdot \boldsymbol{I}_P, 2\tau_1^\beta \boldsymbol{P}^\beta\} \\
&= \frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \Big\{ \Big(\boldsymbol{b}(t_{ij})^\top, \boldsymbol{Z}_i^\top, \frac{1}{J} \sum_{j_s=1}^{J} W_i(s_{j_s}) \boldsymbol{b}^\beta(s_{j_s})^\top \Big)^\top \Big(\boldsymbol{b}(t_{ij})^\top, \boldsymbol{Z}_i^\top, \frac{1}{J} \sum_{j_s=1}^{J} W_i(s_{j_s}) \boldsymbol{b}^\beta(s_{j_s})^\top \Big) \Big\} \\
&+ \text{Blockdiag}\{2\tau_\alpha \boldsymbol{P}, 0 \cdot \boldsymbol{I}_P, 2\tau_1^\beta \boldsymbol{P}^\beta\}.
\end{aligned}
$$

C.4 Additional Data Application Results

Table C.1: Baseline scalar covariates and censoring rates in OPACH data. Race includes white, black and hispanic (w/b/h, %). Censoring rates are calculated as the proportion of censored observations among the total.

Covariates	Median (range) / %
Race (w/b/h, %)	49.9/33.1/17.0
Age	81.0 (64.0 - 98.0)
Awake time (hours per day)	15.0 (10.4 - 20.6)
censoring rate	76.56%

Table C.1 provides baseline scalar covariates and censoring rate in OPACH data.

Figure C.1 shows the estimated baseline hazard function in real data application.

We further investigate whether the baseline functional covariate influences the rate of change (slope) of the longitudinal outcome over time. To this end, we extend the functional

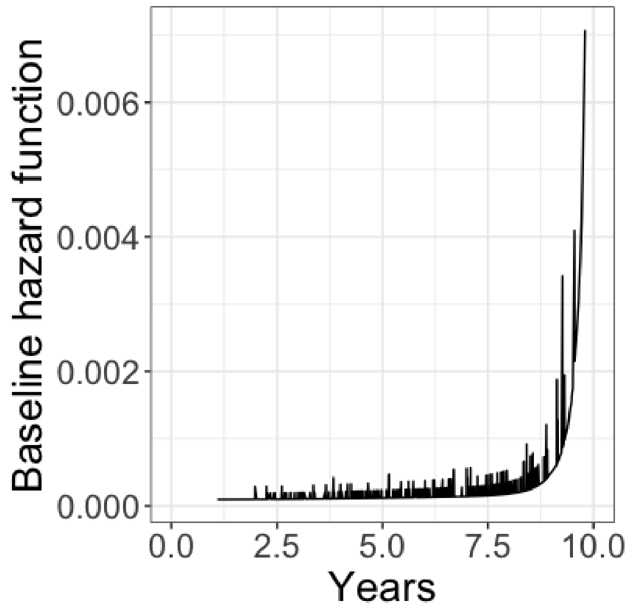


Figure C.1: Estimated baseline hazard function ($h_0(t)$) from the FJM applied to real data.

linear regression model for the longitudinal outcome as follows:

$$Y_{ij} = \mu(t_{ij}) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}_1 + \int_{s \in \mathcal{S}} W_i(s)\beta_1(s)\mathrm{d}s + \Big(\int_{s \in \mathcal{S}} W_i(s)\beta_{11}(s)\mathrm{d}s\Big) t_{ij} + X_i(t_{ij}) + \epsilon_{ij}\,,$$

where $\beta_{11}(s)$ denotes a smooth coefficient function representing the effect of the baseline functional covariate $W_i(s)$ on the slope of the longitudinal trajectory. All other components of the functional joint model remain unchanged. The extended model selects four functional principal components for the longitudinal outcome, yielding a Bayesian Information Criterion (BIC) value of 277,832.9, compared to a BIC of 277,835 for the model without the slope term. This modest improvement in BIC suggests that incorporating a slope effect for the baseline functional covariate enhances model fit and provides a more comprehensive understanding of the longitudinal dynamics, without introducing substantial additional model complexity.

For the longitudinal outcome, Figure C.2 and Figure C.3 display the estimated mean function and the two leading eigenfunctions corresponding to the top two eigenvalues for the new FJM that incorporates the slope term. Table C.2 presents the estimated coefficients for the baseline scalar covariates in the longitudinal submodel with the slope term, as well as the estimated Cox regression coefficients. The results and their interpretations are largely consistent with those from the original FJM without the slope term.

Figure C.4 shows the estimated coefficient functions for the baseline functional covariate, along with their 95% pointwise and 95% simultaneous confidence bands, for both the longitudinal model with the slope term and the survival model. The overall patterns in the left and right panels remain similar to the earlier model. The middle panel depicts the estimated coefficient function for the slope term in the longitudinal model.

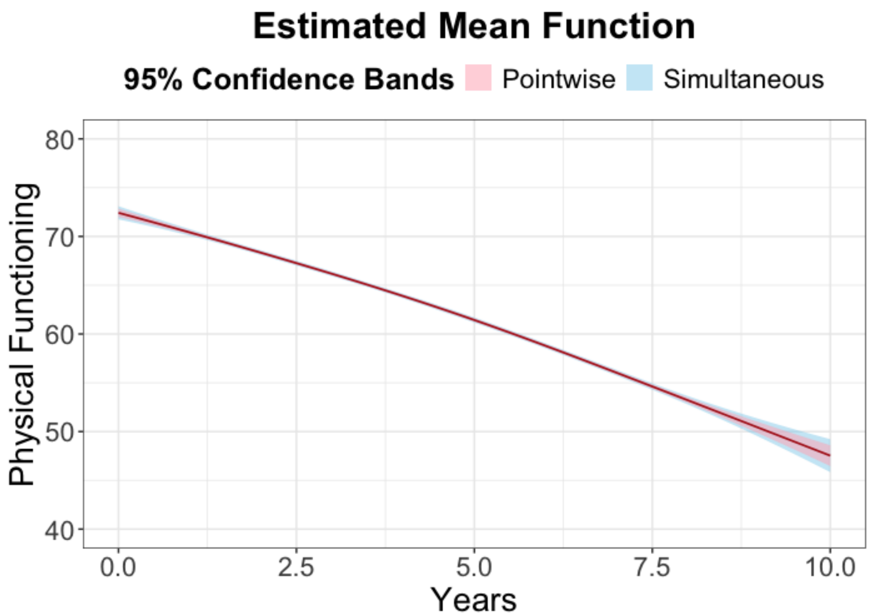


Figure C.2: Estimated mean trajectories of SF-36 physical function for the longitudinal model with slope term.

C.5 Additional Simulation Results

This section shows additional simulation results to evaluate the model's performance across 100 simulation replicates, with key findings illustrated in several figures to highlight its accuracy and robustness in estimating critical parameters. Figure C.5 provides boxplots of the estimated eigenvalues. The close agreement between the red lines (true values) and the boxplot medians validates the model's stability and accuracy in capturing variance structures. Figure C.6 displays the estimated model parameters, including the coefficient vector $\boldsymbol{\gamma}_1$ in longitudinal model, the Cox regression coefficients $\boldsymbol{\gamma}_2$ and $\boldsymbol{\gamma}_3$, and the white noise variance σ^2. The close alignment of the estimates with true values underscores the model's ability to link longitudinal trajectories to survival outcomes.

These findings collectively affirm the proposed model's effectiveness in handling baseline scalar and functional covariates in longitudinal and survival data, ensuring accurate parameter

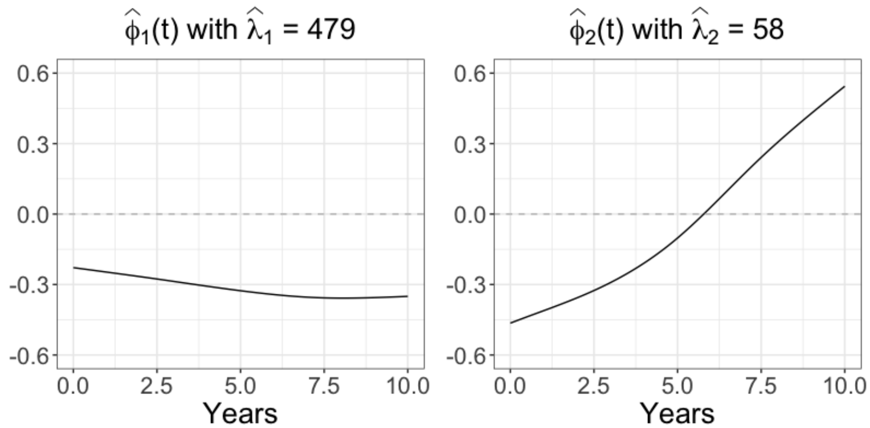


Figure C.3: Top two estimated eigenfunctions with associated eigenvalues for the longitudinal model with slope term.

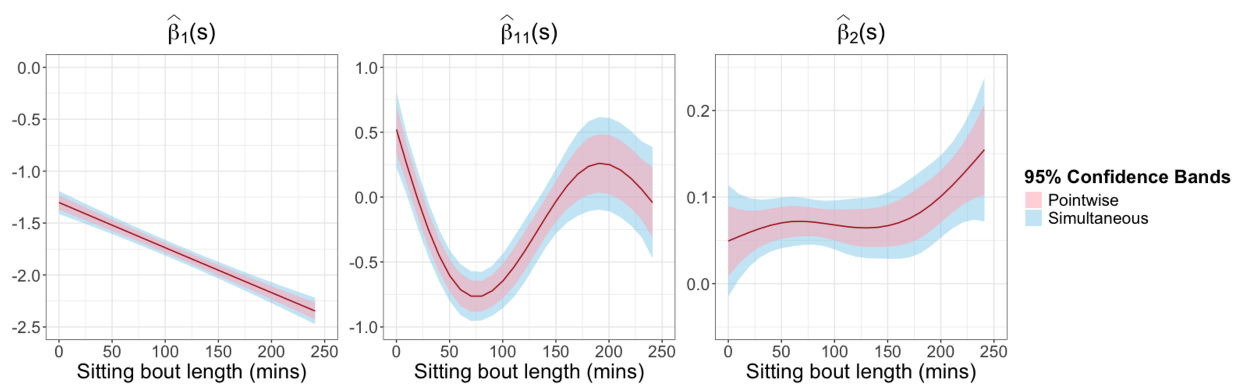


Figure C.4: Estimated coefficient functions (red line) with 95% point-wise confidence bands and 95% simultaneous confidence bands in the longitudinal model with slope term (left two panels) and survival model (right panel).

Table C.2: Estimated coefficients of baseline scalar covariates in the longitudinal model with slope term and estimated Cox regression coefficients (with standard errors), evaluating the association between baseline scalar covariates, FPCA scores, and time to event. Baseline covariates include race, age at baseline, and awake time. An asterisks indicates significance at the 0.05 level.

	Longitudinal		Survival	
Covariates	Estimate (s.e.)	P-value	Estimate (s.e.)	P-value
Race: black	-0.71 (0.06)*	0.00	-0.09 (0.04)*	0.01
Race: hispanic	-0.01 (0.06)	0.84	-0.15 (0.04)*	0.00
Age	-6.85 (0.06)*	0.00	0.66 (0.03)*	0.00
Awake time (hours per day)	6.57 (0.06)*	0.00	-0.30 (0.03)*	0.00
SF-36 physical function score (ξ_{i1})	-	-	0.03 (0.00)*	0.00
SF-36 physical function score (ξ_{i2})	-	-	-0.03 (0.01)*	0.00
SF-36 physical function score (ξ_{i3})	-	-	0.00 (0.02)	0.85
SF-36 physical function score (ξ_{i4})	-	-	0.03 (0.03)	0.27

estimation while maintaining interpretability.

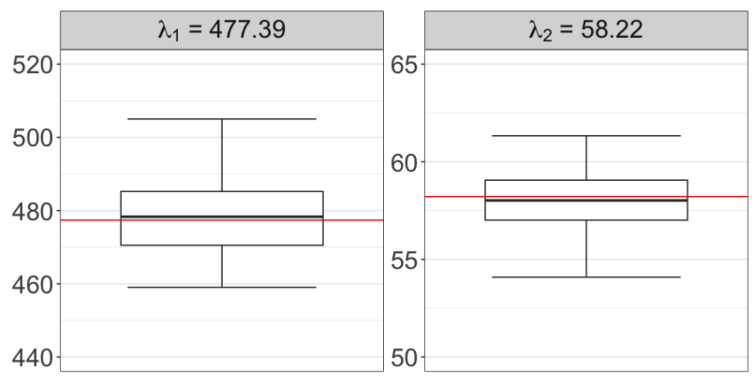


Figure C.5: Boxplots of estimated eigenvalues across 100 replicates.

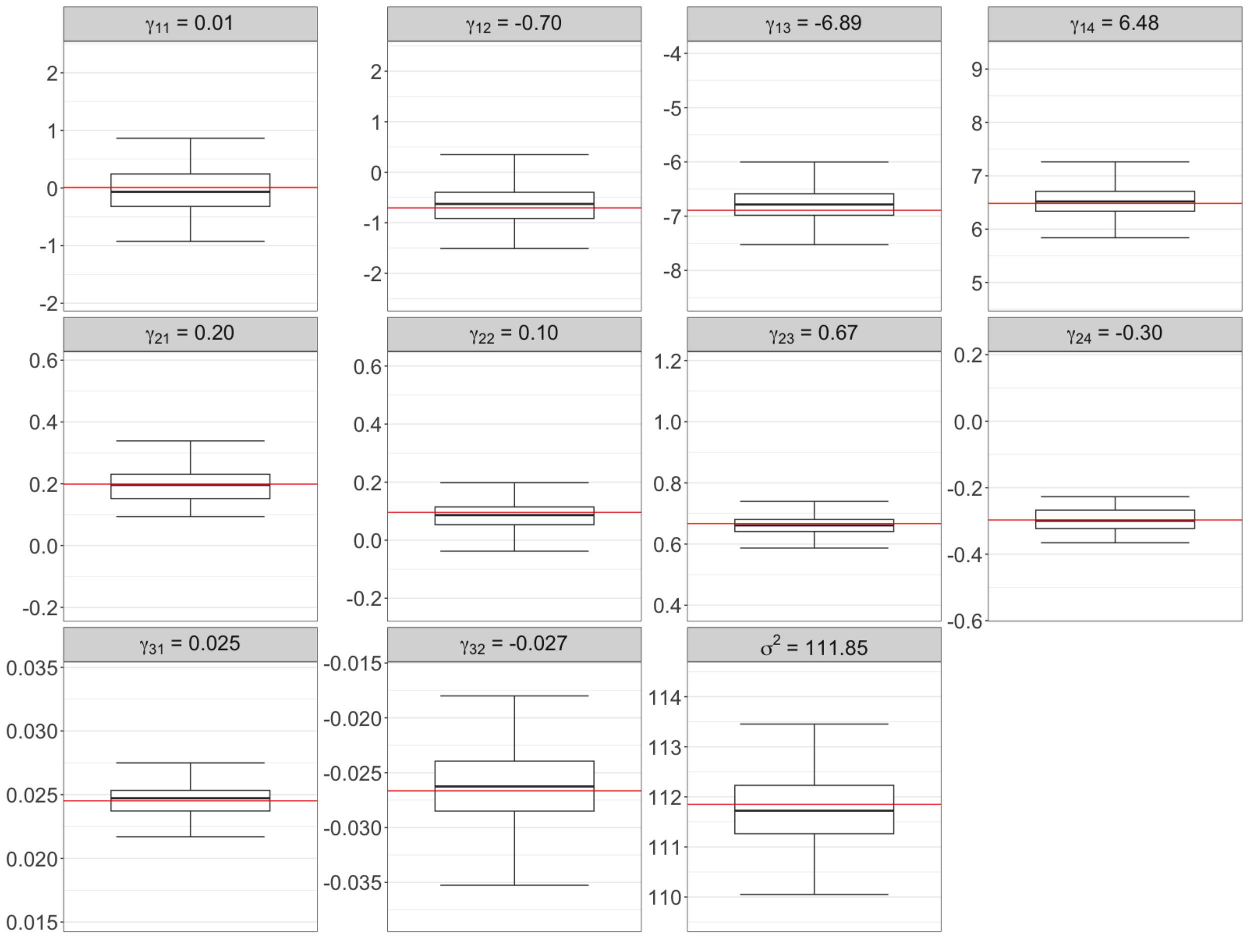


Figure C.6: Boxplots of estimated coefficents ($\boldsymbol{\gamma}_1$) for baseline scalar covariates and estimated white noise variance (σ^2) in longitudinal model and boxplots of Cox regression coefficients ($\boldsymbol{\gamma}_2$, $\boldsymbol{\gamma}_3$) across 100 replicates.