ABSTRACT

WEAVER, CALEB JOSHUA. Biostatistical Methods for Longitudinal and Neuroimaging Data. (Under the direction of Luo Xiao).

Longitudinal data are essential to many biological and medical data applications as they allow for a deeper understanding of how response variables change over time. However, given the breadth of such applications, each with its own set of nuances associated with data collection and behavior, standard statistical methods of analysis can be inadequate to address questions of scientific interest appropriately. This is also true in the study of data collected via techniques of neuroimaging, which measure brain properties related to structure and function with the purpose of better understanding the human brain. The complexity of neuroimaging data leads to unique challenges associated with the application of traditional statistical models. This dissertation addresses settings and questions of scientific interest in longitudinal and neuroimaging data for which existing methodologies are insufficient.

First, we propose a methodology for modeling nonlinear relationships between response variables and functional connectivity network-valued predictors. We exploit the network structure of functional connectivity by imposing meaningful sparsity constraints, which leads to the identification of associations of brain region interactions with the response and the assessment of whether or not the functional connectivity associated with a region is related to the response variable. We demonstrate the effectiveness of the proposed model in simulation studies and in an application to a resting-state fMRI data set to model fluid intelligence and sex and to identify predictive links between brain regions.

Second, we consider the longitudinal data setting in which observation times are informative of the response. We show that under a general class of shared random effect models, a commonly used functional data method may lead to inconsistent model estimation, while another functional data method results in consistent and even rate-optimal estimation. Specifically, we show that the mean and covariance functions can be estimated appropriately via penalized splines and penalized tensor-product splines, respectively, both with specific choices of parameters. Theoretical results are provided for the proposed method, and simulation studies and a real data analysis are conducted to demonstrate its performance.

Finally, motivated by a longitudinal diffusion tensor imaging study of sports-related

concussion (SRC), we present the problem of biclustering multivariate longitudinal data in which subjects and features are grouped simultaneously based on longitudinal patterns. We propose a penalized regression-based method for solving this problem by exploiting the heterogeneity in the longitudinal patterns within subjects and features. We evaluate the performance of the proposed methods via a simulation study and perform an analysis of the motivating data set. In this analysis, we reveal subgroups of SRC cases that exhibit heterogeneous patterns of white-matter abnormalities.

Biostatistical Methods for Longitudinal and Neuroimaging Data

by
Caleb Joshua Weaver

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
2022

APPROVED BY:

Wenbin Lu	Eric Chi
Ana-Maria Staicu	Jaroslaw Harezlak External Member

Luo Xiao
Chair of Advisory Committee

DEDICATION

To Maureen.

BIOGRAPHY

Caleb Weaver was born in Columbia, South Carolina. He attended North Carolina State University (NCSU) and earned a Bachelor of Science in Mathematics in 2014. He began his professional career as a data scientist at SAS Institute in Cary, NC. Meanwhile, he continued his education at NCSU in the Department of Statistics and received a Master of Statistics degree in 2017. He proceeded to pursue his Ph.D. in the field of statistics at NCSU. He is currently a data science manager at MetLife Inc. and lives in Raleigh, North Carolina, with his wife, Maureen, and their dog, Goose.

ACKNOWLEDGEMENTS

I would first like to thank my advisor and friend, Dr. Luo Xiao, who has demonstrated such stoicism and fortitude in the face of my ignorance. Through his mentorship, I have become more thoughtful in my research, effective in my communication, and deliberate in my work. I will always be grateful for that.

I also wish to thank my committee members, Drs. Wenbin Lu, Jaroslaw (Jarek) Harezlak, Ana-Maria Staicu, and Eric Chi, as well as Dr. Martin Lindquist, each of whom has acted as a mentor to me and individually contributed to the quality of my research in unique ways. Their support and guidance have made the writing of this dissertation possible.

I am fortunate to have made lasting friendships during my time at NCSU. My friends Kade Young, Sarah Mislan, Peter Norwood, Caroline Norwood, Arielle Valint, and Tim Valint made what may have been a long, challenging, and unenjoyable undertaking into a long, challenging, but enjoyable one. I look forward to now having more time to spend with them.

I am also fortunate to have such a supportive family, viz., my parents, Michael and Rebecca, and my brother, Luke. I appreciate their love and encouragement. Year after year, they made me feel industrious rather than withdrawn for turning on my laptop after Thanksgiving dinner and on Christmas Eve. I promise last year was the last time.

Finally, I would like to express my appreciation and gratitude to my beloved wife, Maureen, for making me feel good about myself and my work when it was most difficult. I've already read some of this to you out loud, but I hope you appreciate the rest.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Outline and Contributions

The analysis of change is fundamental to the understanding of any biological process. Longitudinally collected data, which measure subjects repeatedly through time, allow for the study of temporal patterns within subjects and the factors that influence change. For example, white matter properties may be observed for subjects following brain injury to characterize recovery patterns. Because this kind of analysis plays such an essential role in the applied sciences, many statistical methods have been developed for analyzing longitudinal data; see, e.g., Diggle et al. (2002) and Fitzmaurice et al. (2008). However, when standard methods are inappropriately applied without consideration for the unique features of a data set, results can be biased and misleading. Data collected via techniques of neuroimaging, for example, are replete with such unique features; see Ombao et al. (2019) for a recent review of existing methodologies in neuroimaging data analysis. Such data's complex patterns and size can make appropriate analysis via traditional statistical methods infeasible. This dissertation addresses settings and questions of scientific interest in longitudinal and neuroimaging data for which existing methodologies are insufficient.

In Chapter 2, we propose a methodology for modeling nonlinear relationships between response variables and functional connectivity network-valued predictors using methods of semiparametric regression and structured sparsity. The proposed model exploits the network structure of functional connectivity by imposing meaningful sparsity constraints, which leads not only to the identification of association of interactions between brain regions with the response, but also the assessment of whether or not the functional

connectivity associated with a region is related to the response variable. We develop an iterative algorithm based on the alternating direction method of multipliers (ADMM; Boyd et al. (2011)) for estimating the components of the proposed model. We demonstrate the effectiveness of the proposed model in simulation studies and in an application to a resting-state functional magnetic resonance imaging rs-fMRI data set to model fluid intelligence and sex and to identify predictive links between brain regions. Software in the form of R code for implementing the proposed methodology is made available.

In Chapter 3, we consider the use of functional data analysis for longitudinal data in which observation times are informative of the response. We show that under a general class of shared random effect models, a commonly used functional data method may lead to inconsistent model estimation, while another functional data method results in consistent and even rate-optimal estimation. Specifically, we show that the mean and covariance functions can be estimated appropriately via penalized splines and penalized tensor-product splines, respectively, both with specific choices of parameters. Theoretical results are provided for the proposed method. We demonstrate the performance of the proposed method in a simulation study and in an analysis of a motivating dataset of longitudinal trajectories of patient-reported symptom severity in Parkinson's disease.

In Chapter 4, we present the problem of biclustering multivariate longitudinal data in which subjects and features are grouped simultaneously based on longitudinal patterns. We motivate this work by a longitudinal diffusion tensor imaging study of sports-related concussion (SRC). We propose a penalized regression-based method for solving this problem by exploiting the heterogeneity in the longitudinal patterns within subjects and features. We develop an iterative algorithm based on ADMM for estimating the components of the proposed model. We evaluate the performance of the proposed methods via a simulation study and perform an analysis of the motivating data set. In this analysis, we reveal subgroups of SRC cases that exhibit heterogeneous patterns of white-matter abnormalities. An R package for implementing the proposed methods is developed and made available.

The remainder of this chapter provides a brief overview of common methodologies in statistics employed in subsequent chapters. What follows is not intended to serve as a thorough review of the topics discussed but rather to provide the necessary background supporting the technical work in the following chapters.

1.2 Technical Background

1.2.1 Penalized Regression and Structured Sparsity

Consider the standard linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y} \in \mathbb{R}^n$ is a vector of response values, $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of coefficients to be estimated, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of error terms. The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ minimizes the sum of squared residuals; that is, $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, which has the well-known solution $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$. In many cases, particularly when the number of features approaches or exceeds the number of observations, the OLS estimate can be improved upon by imposing a penalty term $\mathcal{R}(\boldsymbol{\beta})$ on the values of the estimated coefficients. The penalized estimator solves a penalized sum of squares:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \mathcal{R}(\boldsymbol{\beta}),$$

where the tuning parameter $\lambda \geq 0$ balances the trade-off between model bias and the penalty term. While many others exist, the most common penalty in statistics is likely the ℓ_1 norm penalty, which leads to the lasso (Tibshirani, 1996),

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 ,$$

which, due to the geometry of the ℓ_1 norm, simultaneously performs coefficient estimation and variable selection by shrinking some coefficients to be exactly zero. A vector or matrix is said to be *sparse* if it has few nonzero entries, and a penalty is said to promote sparsity if it penalizes nonzero elements.

As sparsity patterns induce structure, useful models of real-world high-dimensional data tend to be sparse. This has led to significant interest in the topic of sparsity from the statistical and machine learning communities. In some applications, it can be beneficial to impose sparsity-inducing penalty functions that encourage coefficient estimation according to available structural information about the predictors. Examples of penalty functions that promote structured sparsity include the group lasso (Yuan & Lin, 2006), which allows groups of related covariates to be selected as a unit, and the fused lasso (Tibshirani et al., 2005), which encourages smoothness in coefficients that can be ordered in a meaningful way by imposing sparsity in their successive differences.

Of course, penalization is not restricted to the standard linear regression model. The general penalized regression framework (Hastie et al., 2009) is given by

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathcal{L}\{y_i, f(\mathbf{x}_i)\} + \lambda \mathcal{R}(f),$$

where y_i and $\mathbf{x}_i$ are the response and vector of covariates for the ith observation, respectively, $\mathcal{L}$ is a loss function, f belongs to a space of regression functions $\mathcal{F}$, and λ and $\mathcal{R}$ are as previously defined. Common loss functions include the squared error loss in linear regression and the negative log-likelihood in logistic regression. Many popular statistical learning methods fit into the penalized regression framework, including the lasso (Tibshirani, 1996), ridge regression (Hoerl & Kennard, 1970), elastic net (Zou & Hastie, 2005), graphical lasso (Friedman et al., 2008), adaptive lasso (Zou, 2006), convex clustering (Pelckmans et al., 2005), smoothing splines (e.g. Wahba, 1990; Gu, 2002), and penalized splines (e.g. Eilers & Marx, 1996; Ruppert, 2003). Penalized splines, which are briefly discussed in Section 1.2.2, are used in model estimation in Chapters 2 and 3. Additionally, the methods proposed in Chapters 2 and 4 fit into the penalized regression framework and involve penalty terms that induce structured sparsity patterns tailored to their data applications.

1.2.2 Nonparametric Regression

In the standard linear regression model discussed in the previous section, the structural relationship between the covariates and the response variable is assumed prior to model estimation. This can result in well-behaved estimates when the true relationship is consistent with model assumptions. However, when these assumptions are violated, the resulting estimates can be biased and misleading. In contrast, methods of nonparametric regression make no assumptions on the form of the response variable's dependence on the predictors. The nonparametric regression model of the relationship between the response y_i and covariate x_i of the ith observation of a data set is

$$y_i = f(x_i) + \epsilon_i,$$

for $i = 1, \ldots, n$, where f is an unknown and unspecified function.

Methods of estimating f include kernel regression (Nadaraya, 1964; Watson, 1964;

Wand & Jones, 1994), local polynomial methods (e.g. Fan, 1992), smoothing splines (e.g. Wahba, 1990; Gu, 2002), and penalized splines (e.g. Eilers & Marx, 1996; Ruppert, 2003). In kernel regression and local polynomial methods, f is estimated using weighted averages of neighboring observed data, such that closer points are given higher weights. However, in this dissertation, we will use spline-based methods. In regression splines, f is approximated by a collection of K basis functions; that is, $f(x) \approx \sum_{k=1}^{K} \theta_k B_k(x)$, where $\{B_1(x), \ldots, B_K(x)\}$ is a set of K basis functions and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$ is a vector of unknown coefficients. There are many possible choices of basis functions, including the Fourier basis and wavelet basis (e.g. Morris & Carroll, 2006), but the B-spline basis is the most common choice for regression splines. A spline of order $r+1$ is a piecewise polynomial function of degree r that is smoothly connected at its changepoints, called knots. B-splines, more specifically, are locally non-zero for at most $r+1$ adjacent intervals between knots, which makes them computationally faster and more stable (Boor, 1978). The estimate of f using spline-based methods is obtained as $\hat{f}(x) = \sum_{k=1}^{K} \widehat{\theta}_k B_k(x)$, where $\widehat{\boldsymbol{\theta}}$ is estimated from the observed data. In regression splines, $\widehat{\boldsymbol{\theta}}$ is estimated by solving a least squares problem.

Clearly, the selection of basis functions K, or equivalently, the number of knots, directly affects the quality of the resulting fit. If K is too small, the estimate will underfit the data, resulting in bias; if K is too large, the estimate will overfit the data. To alleviate the difficulty associated with choosing the number of basis functions for regression splines, smoothing splines place knots at all observed data points and impose a penalty term on the "roughness" of the estimated function as quantified by the second derivative; the smoothing spline estimator minimizes a penalized least squares:

$$\hat{f} = \arg\min_{f} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int f^{(2)}(x)^2 \, \mathrm{d}x.$$

Here, $\lambda > 0$ is called the smoothing parameter and controls the balance between goodness of fit and smoothness of the estimate.

Penalized splines represent a hybrid of regression splines and smoothing splines in that they use a reduced number of knots as well as impose a roughness penalty. In this way, penalized splines are computationally simpler than smoothing splines and ameliorate the overfitting of regression splines. Specifically, the penalized spline estimator

is $\hat{f}(x) = \sum_{k=1}^{K} \widehat{\theta}_k B_k(x)$, where

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left\{ y_i - \sum_{k=1}^{K} \theta_k B_k(x_i) \right\}^2 + \lambda \boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta},$$

where $\mathbf{P}$ is a penalty matrix that penalizes the roughness of the estimate. Penalized splines can be categorized based on their penalty term and basis function selection. One example is P-splines (Eilers & Marx, 1996), which use B-spline bases and a discrete difference penalty. We will use penalized splines in model estimation in Chapters 2 and 3.

While nonparametric regression methods alleviate the risk of misspecification error, they do have disadvantages that limit their usefulness in application. One important problem is that the precision of a nonparametric estimator decreases rapidly as the dimension of the predictor variable increases. This phenomenon is referred to as the *curse of dimensionality*. Semiparametric regression methods offer a compromise between parametric and nonparametric regression methods by incorporating both parametric and nonparametric components (e.g. Ruppert, 2003; Harezlak et al., 2018). The most commonly used semiparametric regression models are the partially linear model (Engle et al., 1986), varying coefficient models (Hastie & Tibshirani, 1993), and the single index model (Härdle & Stoker, 1989; Ichimura, 1993). In the single index model, the response variable is modeled as a nonparametric function of a linear combination of the predictors called an index; that is,

$$y_i = f(\mathbf{x}_i^\top \boldsymbol{\beta}) + \epsilon_i,$$

where y_i, $\mathbf{x}_i$, and ϵ_i are the response, vector of predictors, and error of the ith observation, respectively, $\boldsymbol{\beta}$ is a vector of coefficients to be estimated, and f is an unknown smooth function, also to be estimated. The advantage of the single index model over a fully nonparameteric regression model is that there is only one nonparametric dimension, so the curse of dimensionality is avoided. The methods proposed in Chapter 2 use a single index model to describe nonlinear relationships between response variables and high-dimensional predictors.

1.2.3 Functional Data Analysis

In functional data analysis (FDA), observations from a single subject over a continuum, such as time, are treated as realizations from a smooth random function; for a compre-

hensive review of FDA, see Ramsay & Silverman (2006). Consider a set of observations $\{(Y_{ij}, T_{ij}) : j = 1, \ldots, m_i; i = 1, \ldots, n;\}$, where Y_{ij} is the response of subject i observed at time point $T_{ij} \in \mathcal{T}$ and $\mathcal{T}$ is a closed interval. Methods of FDA assume that Y_{ij} is a realization of a smooth random function $X(\cdot)$, contaminated with some measurement error; that is, $Y_{ij} = X_i(T_{ij}) + \epsilon_{ij}$, where $X_i(\cdot)$ is a square-integrable random function in $\mathcal{L}^2(\mathcal{T})$ and ϵ_{ij} is a zero-mean measurement error with finite variance. The functions $X_1(t), \ldots, X_n(t)$ are typically assumed to be i.i.d. copies of a random function $X(t)$ with mean function $\mu(t) = \mathrm{E}\{X(t)\}$ and covariance function $\sigma(t, t') = \mathrm{Cov}\{X(t), X(t')\}$.

Functional data can be observed either (a) over a dense grid of regularly spaced time points or (b) at different and irregular time intervals for each individual. The observed data is called dense functional data in the former case and sparse functional data in the latter. The structure of observed dense functional data is comparable to high-frequency multivariate data, while the structure of sparse functional data is comparable to irregularly spaced longitudinal data. The assumption of smoothness is what distinguishes FDA from multivariate or longitudinal data analysis.

In the case of dense observations in which all subjects are sampled at the same time points, i.e., $T_{ij} = T_j$ and $m_i = m$ for all i, the mean function can be estimated empirically by $\widehat{\mu}(T_j) = n^{-1} \sum_{i=1}^{n} Y_{ij}$ for $j = 1, \ldots, m$. Similarly, the covariance function can be estimated empirically by $\widehat{\sigma}(T_j, T_{j'}) = n^{-1} \sum_{i=1}^{n} \{Y_{ij} - \widehat{\mu}(T_j)\}\{Y_{ij'} - \widehat{\mu}(T_{j'})\}$. Smooth estimates of the mean function can be obtained via smoothing over $\mathcal{T}$ using, for example, methods discussed in Section 1.2.2. Smooth estimates of the covariance function can be obtained via smoothing over $\mathcal{T} \times \mathcal{T}$ using methods of bivariate smoothing, such as bivariate P-splines (Eilers & 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 of dense functional data.

The previously discussed sample mean and covariance estimates cannot be obtained in the case of sparse observation points. For a study of FDA for sparse functional data, see Yao et al. (2003) and Yao et al. (2005). In this setting, nonparametric regression methods can be applied to smooth the observations of Y_{ij} over time, yielding an estimate of $\widehat{\mu}$. For example, the estimate of the mean based on the penalized splines method discussed in

Section 1.2.2 is $\widehat{\mu}(t) = \sum_{k=1}^{K} \widehat{\theta}_k B_k(t)$, where

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i \left\{ Y_{ij} - \sum_{k=1}^{K} \theta_k B_k(T_{ij}) \right\}^2 + \lambda \boldsymbol{\theta}^\top \mathbf{P} \boldsymbol{\theta},$$

where w_i are non-negative weights placed on each observation, and B_k and $\mathbf{P}$ are as previously described. Given a mean function estimate, the covariance function of sparse functional data can be estimated by smoothing the empirical covariance estimates. Let $\widetilde{e}_{ij} = Y_{ij} - \widehat{\mu}(T_{ij})$ and define the auxiliary variables $\widetilde{\sigma}_{ijj'} = \widetilde{e}_{ij}\widetilde{e}_{ij'}$; the covariance function can then be estimated by smoothing the values of these auxilary variables over $\mathcal{T} \times \mathcal{T}$ using methods of bivariate smoothing. Alternatively, Xiao et al. (2018) proposed the FACE method for directly estimating the covariance function of sparse functional data and longitudinal data.

By Mercer's theorem, the covariance function $\sigma(t, t')$ admits the spectral decomposition $\sigma(t, t') = \sum_{k=1}^{\infty} \lambda_k \psi_k(t) \psi_k(t')$ for non-negative eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and corresponding orthonormal eigenfunctions ψ_k (Bosq, 2000). Following the Karhunen-Loeve expansion, the random function X_i can be represented using the eigenfunctions as $X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \psi_k(t)$, where $\xi_{ik} = \int \{X_i(t) - \mu(t)\} \psi_k(t) \, \mathrm{d}t$ are uncorrelated random variables with zero mean and variance $\mathrm{E}(\xi_{ik}^2) = \lambda_k$. Given these eigencomponents, X_i can be approximated by truncating up to the first K terms: $X_i(t) \approx \mu(t) + \sum_{k=1}^{K} \xi_{ik} \psi_k(t)$. What distinguishes this decomposition of X_i from other basis decompositions is that the basis $\{\psi_k(t)\}_k$ depicts the dominant modes of variation in the data. The analysis of the eigencomponents of functional data is called functional principal component analysis (FPCA) and is likely the most commonly used technique in FDA; see, e.g., Rice & Silverman (1991) for details on FPCA for dense data, and see Yao et al. (2005) for details of FPCA for sparse functional data and longitudinal data, see . Of course, in order to estimate the eigencomponents of a functional data set, consistent estimates of the smooth mean and covariance functions are required. Indeed, the asymptotic properties of eigenfunction estimation depend on the theoretical properties of mean and covariance function estimation (Hall et al., 2006). For study of the theoretical properties of functional data analysis, see, e.g., Li & Hsing (2010), Cai & Yuan (2011), Zhang & Wang (2016), and Xiao (2020).

The methods of mean and covariance function estimation discussed in this section involve the formulation of a weighted least squares. A common choice of weights in the

formulation of least squares in FDA is such that each observation has equal weight in forming the least squares. Another choice of weights is such that each set of observations from each subject has equal weight. Zhang & Wang (2016) and Xiao (2020) derived the asymptotic rates of convergence for kernel smoothing estimators and penalized spline estimators, respectively, under a general weighting scheme for fixed and random observation times. In Chapter 3, we discuss a longitudinal data setting in which the equal observation weighting scheme leads to biased estimates, but we show that using the equal subject weighting scheme results in consistent and even rate-optimal estimation of the mean and covariance functions.

Chapter 2

Single Index Models with Functional Connectivity Network Predictors[1]

.

2.1 Introduction

A central goal in neuroimaging research is to relate inter-individual variability in phenotypes to properties of brain functionality. In functional magnetic resonance imaging (fMRI) studies, blood oxygen-level dependent (BOLD) signals are used as a proxy for neural activity. Functional connectivity is defined as the undirected association between the measured BOLD signal between two or more anatomically distinct brain regions. Analysis of functional connectivity has been shown to be useful in the assessment of mental illnesses and disorders including autism (Nair et al., 2013), Alzheimer's disease (Greicius et al., 2004), and Schizophrenia (Lynall et al., 2010).

Functional connectivity can be represented as a network, with nodes as brain regions and edges as relations between regions. While computationally easy to estimate, marginal correlation matrices quantify the indirect interactions between regions and hence are not quite interpretable. In contrast, partial correlation matrices quantify the direct interactions between regions and are therefore scientifically more meaningful (Marrelec et al., 2006; Varoquaux et al., 2010; Smith et al., 2011; Solo et al., 2018). However, they are

[1] A version of this chapter previously appeared as: Weaver, C. et al. (2021). "Single-index models with functional connectivity network predictors". *Biostatistics*.

computationally more challenging to estimate, especially in large dimensional settings, as their estimation often involves the inversion of the covariance matrix. Thus, regularization methods are required, with the most common approach being sparse network estimation (Friedman et al., 2008; Mazumder & Hastie, 2012b; Mazumder & Hastie, 2012a; Ryali et al., 2012; Liu & Luo, 2012). Sparsity is appealing from an interpretation standpoint as it implies the lack of a direct relationship between two brain regions. The joint estimation of multi-subject inverse covariance matrices for sharing information about the strength of connections and sparsity patterns has also been studied and applied to measuring brain region connectivity (Danaher et al., 2014; Liang et al., 2016; Qiu et al., 2016; Mejia et al., 2018; Colclough et al., 2018).

Recent literature has focused on identifying the association between brain networks and a response variable using functional connectome-based predictive methods (Woo et al., 2017; Dadi et al., 2019). Standard statistical models have previously been used to predict and classify clinical outcomes and patient attributes based on functional connectivity. For example, random forest classifiers have been used to classify schizophrenic patients (Shen et al., 2010) and support vector machines (SVM) have been used to discriminate between young and old subjects (Meier et al., 2012) and to identify depressed subjects (Zeng et al., 2012). Despite the promise of these high-performing predictive models, they often can be difficult to interpret, which impedes the analysis of brain pattern associations. Models that exploit the network structure of functional connectivity could lead to increased statistical efficiency and improved interpretation. Some recent literature has focused on developing classification models that incorporate network structure. Wang et al. (2020), Wang et al. (2019a), and Vogelstein et al. (2012) propose linear classification models that incorporate subgraph structure in predictors, Guha & Rodriguez (2020) propose a class of network shrinkage priors for Bayesian learning of undirected network predictor coefficients, and Relión et al. (2019) propose a linear classification model that encourages sparsity in the number of nodes and edges of the coefficient matrix, a pattern of sparsity we also seek to incorporate in our proposed model.

A key assumption common to the majority of the previously discussed models is the linear relationship between response variables and predictors, which can be seriously violated in practice. Nonparametric regression methods can be used to accommodate potential nonlinear relationships present in the data. In fully nonparametric regression, the response variable is generally assumed to satisfy some smoothness assumptions, but

no assumptions are made on the form of dependence on the predictors. While this offers flexibility in modeling, results are often undesirable due to the curse of dimensionality. A useful alternative to a fully nonparametric regression approach is the single index model, in which the response variable is modeled as a nonparametric function of a linear combination of the predictors called an index (Härdle & Stoker, 1989; Ichimura, 1993; Carroll et al., 1997; Wang & Yang, 2009).

In this chapter, we propose a single index model in which response variables and sparse functional connectivity network-valued predictors are linked by an unspecified smooth function in order to accommodate potential nonlinear relationships between the response and the predictors. The parameter space can be thought of as a network of parameters; we impose two types of sparsity in the effects of the elements of the networks, similar to Relión et al. (2019). The first type of sparsity, here referred to as *edge sparsity*, places constraints on individual edges in the parameter network. This allows for the pruning of the connection between two nodes while simultaneously allowing the two nodes to remain connected to other nodes in the network. The second type of sparsity, here referred to as *node sparsity*, places constraints on the nodes themselves. Hence, in this model, any node is either in or out of the network. In the latter case, all connections (edges) between the node and all other nodes are removed from the network. The proposed model is estimated via an iterative algorithm, which involves a sub-algorithm based on the alternating direction method of multipliers (ADMM; Boyd et al. (2011)). The proposed model and estimation procedure can be easily generalized for general matrix predictors with other types of sparsity-inducing penalties. Additionally, we consider various regularization approaches to estimating individual functional connectivity matrices as inputs to this model and evaluate their performances in predicting the responses. Software in the form of R code, along with a sample input data set and complete documentation, is available at `https://github.com/clbwvr/FC-SIM`.

The remainder of the chapter is organized as follows: Section 2.2 describes a motivating data set. Section 2.3 discusses the estimation of functional connectivity networks and introduces the proposed model and estimation methods. Section 2.4 assesses the numerical performance of our model with a simulation study. In Section 2.5, we apply the proposed methods to the motivating dataset. Finally, Section 2.6 concludes the chapter with a discussion.

2.2 Materials

Resting-state fMRI (rs-fMRI) has become increasingly popular for evaluating regional interactions between brain regions that occur when a subject is not performing an explicit task. It has been shown that fluctuations in BOLD signal show strong and reliable correlations at rest, even in distant brain regions. In addition, the spatial patterns of the correlations have been shown to be reliably observed across subjects and scanning sessions. Because of the lack of task demands, rs-fMRI removes the burden of subject compliance and training, making it an attractive option in many studies of development and clinical populations.

We analyzed data from 820 subjects with complete rs-fMRI scans from the 900-subject public data release from the Human Connectome Project (HCP).[2] The HCP provides the required ethics and consent needed for study and dissemination. The sample consists of healthy adults (22 to 35 years old, 453 females) scanned on a 3-T Siemens connectome-Skyra. For each subject, a total of four 15 min fMRI scans with a temporal resolution of 0.73 s, and a spatial resolution of 2-mm isotropic were available. The preprocessing pipeline followed the procedure outlined in Smith et al. (2013b) and Smith et al. (2013a). Spatial preprocessing was applied using the procedure described by Glasser et al. (2013). ICA, followed by FMRIBs ICA-based X-noisefier (FIX) from the FMRIB Software Library (FSL) (Griffanti et al., 2014), was used for structured artifact removal, removing more than 99% of the artifactual ICA components in the dataset. Group spatial ICA was used to obtain a parcellation of 100 components that cover both the cortical surfaces and the subcortical areas.

The parcellation was used to project the fMRI data into 100-dimensional data corresponding to 100 brain regions. Data from the 4 scans, each measuring the BOLD signal at 1200 time points, were concatenated into a single run. Hence, the final rs-fMRI data used in the analysis consisted of a 4800×100 matrix for each of 820 subjects. In addition, sex and fluid intelligence assessed using a form of Raven's Standard Progressive Matrices Test (Bilker et al., 2012) were extracted for each subject and were used as response variables in our model.

[2] Data were provided (in part) by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

2.3 Methods

2.3.1 Estimation of Functional Connectivity Networks

Functional connectivity networks describe how different regions of the brain interact. The strength of interaction is often characterized by partial correlations. Let $\mathbf{\Theta}_i = (\theta_{ijk})_{1\leq j\leq p, 1\leq k\leq p}$ be the ith subject's inverse covariance matrix. That subject's partial correlation matrix is then $\mathbf{P}_i = (P_{ijk})_{1\leq j\leq p, 1\leq k\leq p}$ with $P_{ijk} = -\theta_{ijk}/\sqrt{\theta_{ijj}\theta_{ikk}}$ for $j \neq k$ and $P_{ijj} = 1$. Thus, estimation of partial correlations can be derived from that of the inverse covariance matrix.

To estimate a subject's inverse covariance matrix, a simple method is to invert the sample covariance matrix, i.e. to let $\widehat{\mathbf{\Theta}}_i = \mathbf{S}_i^{-1}$, where $\mathbf{S}_i$ is the ith subject's sample covariance matrix. For example, for the rs-fMRI data, the 4800×100 data matrix for each subject leads to a 100×100 sample covariance matrix. Although not an issue for our data application, a drawback of this approach is that the sample covariance is not invertible when the number of regions exceeds the sample size (the number of time points for the rs-fMRI data), which could be the case if a high resolution of brain regions is desired. One solution is to instead invert the Ledoit-Wolf estimator (Ledoit & Wolf, 2004), which takes the form $\widehat{\mathbf{\Sigma}}_i = \delta_i \mathbf{I} + (1-\delta_i)\mathbf{S}_i$ where δ_i is a scalar between 0 and 1. In application, we will use the closed form optimal value of δ_i as derived by Ledoit & Wolf (2004). This optimal value minimizes the expected squared Frobenious norm of $\widehat{\mathbf{\Sigma}}_i - \mathbf{S}_i$ using theoretical results. Each subject will have a different value of δ_i. Then $\widehat{\mathbf{\Sigma}}_i^{-1}$ is the resulting inverse covariance for subject i.

Alternatively, one may consider the graphical lasso (Friedman et al., 2008), which gives $\widehat{\mathbf{\Theta}}_i = \arg\min_{\mathbf{\Theta}}\{\mathrm{Tr}(\mathbf{\Theta}\mathbf{S}_i) - \log\det(\mathbf{\Theta}) + \lambda_i \sum_{i\neq j} |\theta_{ij}|\}$, where $\mathrm{Tr}(\cdot)$ is the trace operator and λ_i is a regularization parameter controlling the level of regularization in the estimate. The estimate may be element-wise sparse, which, in the context of functional connectivity networks, implies conditional independence between the signals of two brain regions given the signals of all other regions. We will use 5-fold cross validation (Bien & Tibshirani, 2011) for selecting the value of λ_i, the value of which could vary across subjects. The glasso estimates for three subjects of the HCP study are shown in Figure 2.1.

The glasso estimate usually gives different sparsity patterns between subjects. It may sometimes be desirable to instead obtain the same sparsity pattern across subjects. This can be achieved by joint estimation with group penalties. The set of joint glasso estimates

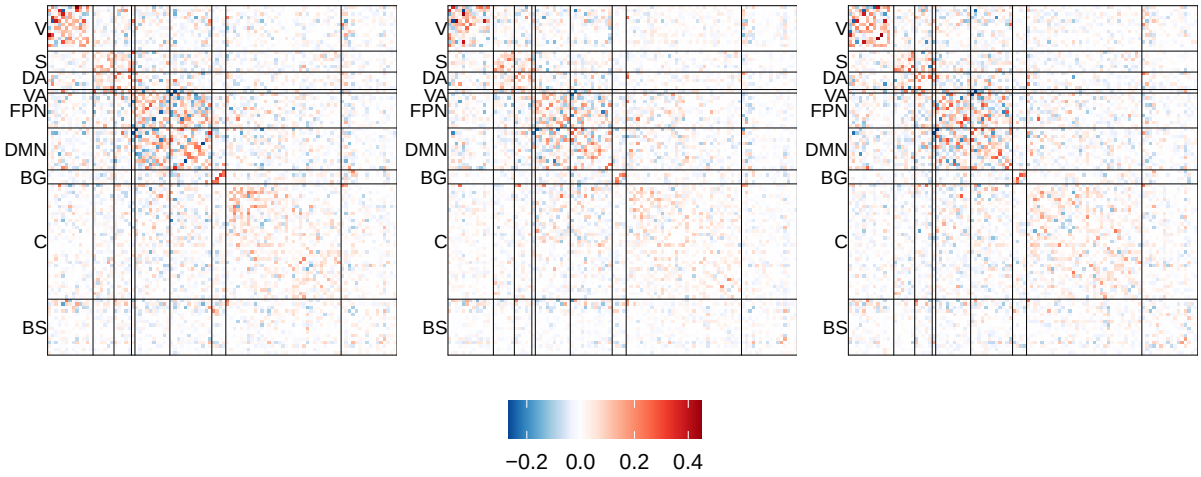

Figure 2.1 Partial correlation estimates via the glasso for three subjects of the HCP study. Nodes are grouped by functional network denoted in the rows of each matrix (V: visual network, S: somatomotor, DA: dorsal attention network, VA: ventral attention network, FPN: frontoparietal network, DMN: default mode network, BG: basal ganglia, C: cerebellum, BS: brain stem).

$\{\widehat{\boldsymbol{\Theta}}_i, i = 1, \ldots, n\}$ is the solution to

$$\underset{\{\boldsymbol{\Theta}_i=(\theta_{ijk})_{1\leq j\leq p, 1\leq k\leq p}\}}{\arg\min} \sum_{i=1}^{n} \mathrm{Tr}(\boldsymbol{\Theta}_i \mathbf{S}_i) - \log\det(\boldsymbol{\Theta}_i) + \lambda \sum_{j\neq k} \left(\sum_{i=1}^{n} \theta_{ijk}^2 \right)^{1/2},$$

where λ is a tuning parameter. We will use 5-fold cross-validation for selecting the value of λ. We will refer to this approach as *joint glasso* as it is a special case of the joint graphical lasso (Danaher et al., 2014).

The three methods provide three different sparsity patterns: (1) the Ledoit-Wolf estimator has no sparsity, (2) the glasso gives sparsity at the individual level, and (3) the joint glasso ensures sparsity at the group level. As a zero value of partial correlation implies conditional independence between the two associated brain regions, the Ledoit-Wolf estimator is the least interpretable, while the joint glasso gives the most parsimonious interpretation as the same sparsity pattern is enforced across subjects. However, the glasso is more flexible than the joint glasso as it accommodates potentially varying sparsity patterns across subjects. We will estimate individual functional connectivity

matrices using these methods and consider the estimates as inputs to our regression model. We shall investigate how these estimates compare when they are used as predictors to relate to response variables. Specifically, we are interested in their predictability of the response variable and the interpretability of the resulting model estimation for identifying meaningful associations of interactions between brain regions with the response variable.

2.3.2 Model Specification

Suppose that y_i is the response variable and $\mathbf{X}_i = (X_{ijk})_{1\leq j\leq p, 1\leq k\leq p}$ is the functional connectivity network for subject $i\,(1 \leq i \leq n)$. Assuming that the response variable is from a distribution in the exponential family, we propose a single index model in which the conditional mean of y_i given $\mathbf{X}_i$ is

$$g\{\mathrm{E}(y_i \mid \mathbf{X}_i)\} = f\left(\langle \mathbf{X}_i, \mathbf{B}\rangle\right), \tag{2.1}$$

where g is a known monotonic function, f is an unknown and unspecified index function, $\mathbf{B} = (B_{jk})_{1\leq j\leq p, 1\leq k\leq p}$ is an unknown coefficient matrix, and the inner product $\langle \mathbf{X}, \mathbf{B}\rangle$ is $\sum_{j,k} X_{jk}B_{jk}$. As $\mathbf{X}_i$ is a partial correlation matrix with 1s on the diagonal, it is reasonable to assume that the coefficient matrix $\mathbf{B}$ is also symmetric and has 0 values on the diagonal. Let vech$(\cdot)$ be the matrix operator that stacks the lower triangle (excluding the diagonal) of a square matrix into a vector. Then $\langle \mathbf{X}_i, \mathbf{B}\rangle = 2\boldsymbol{\chi}_i^\top \boldsymbol{\beta}$, where $\boldsymbol{\chi}_i = \mathrm{vech}(\mathbf{X}_i)$ and $\boldsymbol{\beta} = \mathrm{vech}(\mathbf{B})$.

As the number of coefficients in the coefficient matrix $\mathbf{B}$ is $\mathcal{O}(p^2)$, which could well exceed the sample size n, regularization is required for model estimation. More importantly, it is desirable to assume sparsity in the coefficient matrix for better interpretation. Indeed, if an edge is not associated with the response, then the corresponding coefficient should be zero; if a node is not associated with the response, then the coefficients corresponding to the edges that are associated with the node should all be zero. The former is called *edge sparsity* and can be achieved using the lasso, while the latter is called *node sparsity* and can be achieved by imposing a carefully constructed group lasso. Therefore, we consider the following optimization problem:

$$\underset{f,\mathbf{B}}{\text{minimize}}\ \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_f(\mathbf{B} \mid y_i, \mathbf{X}_i) + \mathcal{R}(\mathbf{B}) \tag{2.2}$$

subject to the identifiability constraints $\|\mathbf{B}\|_{\mathrm{F}} = 1$, $B_{12} \geq 0$, and $B_{ii} = 0$ for $i = 1, \ldots, p$, in addition to the symmetry constraint $\mathbf{B}^\top = \mathbf{B}$. Here, $\mathcal{L}_f$ is the negative log-likelihood function that depends on both f and $\mathbf{B}$, and $\mathcal{R}$ is a penalty function. To impose node and edge sparsity, we let

$$\mathcal{R}(\mathbf{B}) = \lambda_1 \sum_{j=1}^{p} \left(\sum_{k \neq j} B_{jk}^2 \right)^{1/2} + \lambda_2 \sum_{j,k} |B_{jk}|.$$

The term associated with the tuning parameter λ_1 is referred to as the *node penalty*, as it penalizes as a group all the coefficients associated with all edges connecting each node. The term associated with λ_2 is the referred to as the *edge penalty* as it penalizes separately each coefficient associated with individual edges. The node & edge penalty term has been used previously in Relión et al. (2019). In Section 2.4, we will compare the node & edge penalty to the lasso (which is equivalent to the edge penalty model) and the elastic net, which linearly combines the ℓ_1 and squared ℓ_2 norms of the elements of $\mathbf{B}$.

2.3.3 Model Estimation

Simultaneously estimating the unknown smooth function f and the coefficient matrix $\mathbf{B}$ is computationally challenging; sequentially estimating f given $\mathbf{B}$ and estimating $\mathbf{B}$ given f is computationally more efficient and a more commonly used strategy in fitting single index models. We will use this strategy for estimating the model parameters. Specifically, the first step will be to estimate f given $\mathbf{B}$ and the second step will be to estimate $\mathbf{B}$ given f. The solutions to both steps are as follows.

Given the coefficient matrix $\mathbf{B}$, we estimate f using penalized splines with a second-order difference penalty (Eilers & Marx, 1996). Let $t_i = \langle \mathbf{X}_i, \mathbf{B} \rangle$ be the index of the ith subject so that $g\{\mathrm{E}(y_i \mid \mathbf{X}_i)\} = f(t_i)$. The penalized splines approximate $f(t)$ by a linear combination of B-splines; that is, $f(t) \approx \sum_{k=1}^{K} \theta_k \phi_k(t)$, where $\{\phi_k\}_{k=1,\ldots,K}$ is a set of cubic B-spline basis functions defined in the range of the t_is, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$ is a vector of associated coefficients. The penalized spline estimator is $\hat{f}(t) = \sum_{k=1}^{K} \widehat{\theta}_k \phi_k(t)$, where

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\boldsymbol{\theta} \mid y_i, t_i) + \gamma \boldsymbol{\theta}^\top \boldsymbol{\Omega} \boldsymbol{\theta},$$

where $\boldsymbol{\Omega}$ is the second-order difference penalty matrix. The second term is a roughness

penalty controlled by the smoothing parameter γ. In our simulation studies and real data analysis, we use 10 basis functions and select γ via generalized cross-validation (GCV; Ruppert (2002)).

Given f, we estimate the coefficient matrix $\mathbf{B}$. To deal with the overlapping group lasso in the objective function (2.2), we adopt the so-called union of groups approach (Jacob et al., 2009), which allows variables to be selected as long as they belong to at least one group with positive coefficients. Then we solve the optimization using the alternating direction method of multipliers (ADMM; Boyd et al. (2011)). We reformulate (2.2) as

$$\begin{aligned} \underset{\boldsymbol{\beta},\, \mathbf{z}_1,\ldots,\mathbf{z}_p,\mathbf{c}}{\text{minimize}} \quad & \mathcal{L}_f(\boldsymbol{\beta}) + \lambda_2 \left\|\boldsymbol{\beta}\right\|_1 + \lambda_1 \sum_{j=1}^{p} \left\|\mathbf{z}_j\right\|_2 + \iota(\mathbf{c}), \\ \text{s.t.} \quad & \boldsymbol{\beta}_{g_j} - \mathbf{z}_j = \mathbf{0} \text{ for } j = 1, \ldots, p, \\ & \boldsymbol{\beta} - \mathbf{c} = \mathbf{0}, \end{aligned}$$

where g_j denotes the group of indices of coefficients in $\boldsymbol{\beta}$ that are associated with the jth node, $\mathbf{z}_j \in \mathbb{R}^{p-1}$ is a local variable, and ι is the indicator function of the constraint set given by

$$\iota(\mathbf{c}) = \begin{cases} 0, & \text{if } \left\|\mathbf{c}\right\|_2 = 1 \text{ and } c_1 \geq 0, \\ \infty, & \text{otherwise} \end{cases}.$$

The above reformulation leads to the following scaled-form ADMM updates:

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= \arg\min_{\boldsymbol{\beta}} \mathcal{L}_f(\boldsymbol{\beta}) + \lambda_2 \left\|\boldsymbol{\beta}\right\|_1 + \frac{\rho}{2} \left\| \boldsymbol{\beta} - \frac{\mathbf{D}\mathbf{z}^{(k)}}{2} + \frac{\mathbf{D}\mathbf{u}^{(k)}}{2} \right\|_2^2 + \frac{\rho}{2} \left\| \boldsymbol{\beta} - \mathbf{c}^{(k)} + \mathbf{w}^{(k)} \right\|_2^2, \\ \mathbf{c}^{(k+1)} &= \arg\min_{\mathbf{c}} \iota(\mathbf{c}) + \frac{\rho}{2} \left\| \boldsymbol{\beta}^{(k+1)} - \mathbf{c} + \mathbf{w}^{(k)} \right\|_2^2, \\ \mathbf{w}^{(k+1)} &= \mathbf{w}^{(k)} + \boldsymbol{\beta}^{(k+1)} - \mathbf{c}^{(k+1)}, \\ \mathbf{z}_j^{(k+1)} &= \arg\min_{\mathbf{z}_j} \lambda_1 \left\|\mathbf{z}_j\right\|_2 + \frac{\rho}{2} \left\| \boldsymbol{\beta}_{g_j}^{(k+1)} - \mathbf{z}_j + \mathbf{u}_j^{(k)} \right\|_2^2, \\ \mathbf{u}_j^{(k+1)} &= \mathbf{u}_j^{(k)} + \boldsymbol{\beta}_{g_j}^{(k+1)} - \mathbf{z}_j^{(k+1)}, \end{aligned}$$

where $\mathbf{z} \in \mathbb{R}^{p(p-1)}$ and $\mathbf{u} \in \mathbb{R}^{p(p-1)}$ are the stacked vectors of $\mathbf{z}_j$'s and $\mathbf{u}_j$'s respectively, and $\mathbf{D} \in \mathbb{R}^{p(p-1)/2 \times p(p-1)}$ is an indicator matrix with 1s in row r indicating the index

of two "copies" of the rth element of $\boldsymbol{\beta}$ in $\mathbf{z}$ and 0s otherwise. Essentially, the matrix multiplication $\mathbf{Dz}/2$ creates a vector of averages of the "copies" of the elements of $\boldsymbol{\beta}$, so that each element of $\boldsymbol{\beta}$ is shrunk towards the average of its two corresponding elements in $\mathbf{z}$.

The update of $\boldsymbol{\beta}$ is solved via the method of proximal gradient descent. Letting $\mathbf{b}$ denote the current iterate $\boldsymbol{\beta}^{(k)}$, we make the iterative update

$$\mathbf{b}^{+} = \text{prox}_{\lambda_2\|\cdot\|_1}\left\{\mathbf{b} - \alpha\nabla\left(\mathcal{L}_f(\mathbf{b}) + \frac{\rho}{2}\left\|\mathbf{b} - \frac{\mathbf{Dz}^{(k)}}{2} + \frac{\mathbf{Du}^{(k)}}{2}\right\|_2^2 + \frac{\rho}{2}\left\|\mathbf{b} - \mathbf{c}^{(k)} + \mathbf{w}^{(k)}\right\|_2^2\right)\right\}$$

until convergence of $\mathbf{b}$ with learning rate α. The derivation of the gradients in the case of normally distributed and binary responses is provided in Section A.1 of Appendix A. In practice, we use backtracking line search (Boyd et al., 2004) for determining the value of α for each step. The update of $\mathbf{c}$ is solved by

$$\mathbf{c}^{(k+1)} = \mathcal{P}\left(\boldsymbol{\beta}^{(k+1)} + \mathbf{w}^{(k)}\right)$$

where $\mathcal{P}(\mathbf{c})$ is the projection of $\mathbf{c}$ onto the surface of a unit ball with $c_1 \geq 0$. The update of $\mathbf{z}_j$ is solved by the proximal operator of the ℓ_2 norm. Letting $\mathbf{a}_j^{(k)} = \boldsymbol{\beta}_{g_j}^{(k+1)} + \mathbf{u}_j^{(k)}$,

$$\mathbf{z}_j^{(k+1)} = \max\left(0, 1 - \frac{\lambda_2}{\rho}\left\|\mathbf{a}_j^{(k)}\right\|_2^{-1}\right)\mathbf{a}_j^{(k)}.$$

The estimation of $\mathbf{B}$ is solved by computing these iterative updates until the convergence of the objective values. The convergence results of the algorithm for estimating $\mathbf{B}$ are similar to those of a similar algorithm used by Feng et al. (2020). For a given f, (2.2) may contain components of a non-convex function. Note, $\mathbf{X}_i$ is a partial correlation matrix with elements $-1 \leq X_{ijk} \leq 1$, so the index $\langle \mathbf{X}_i, \mathbf{B}\rangle$ is bounded over the identifiability constraint set. Then, given the cubic spline construction of f, we have that $\mathcal{L}_f$ is Lipschitz differential. Further, the node & edge penalty and the indicator function ι are both prox-regular functions (Poliquin & Rockafellar, 1996). Thus, by Wang et al. (2019b), we have that the proposed algorithm for estimating $\mathbf{B}$ will have at least one limit point for the sequence of $\mathbf{B}^{(k)}$s, with each limit point being a stationary point.

The overall minimization problem (2.2) is solved by iterating over the two-step algorithm until convergence. Each step of the two-step algorithm is guaranteed convergence.

If the tuning parameter γ in the estimation of f was fixed, the overall two-step algorithm is guaranteed convergence by properties of block coordinate descent. However, as we propose to select γ via GCV at each iteration, the overall algorithm does not have a guarantee of convergence. Nevertheless, our experiments found that the selected value of γ stabilizes quickly, which gives empirical evidence that the algorithm is well-behaved in practice.

For ease of selecting values for the tuning parameters λ_1 and λ_2 via cross-validation, it is convenient to reparameterize the tuning parameters λ_1 and λ_2 as $\delta\lambda$ and $(1-\delta)\lambda$ respectively, so that $\lambda \geq 0$ is the overall level of regularization and $\delta \in [0,1]$ controls the balance between the two sparsity inducing penalties. Optimal values of δ and λ are obtained through grid search in a two-dimensional sequence.

2.4 Simulation Study

2.4.1 Simulation Settings

We conduct a simulation study to evaluate the performance of the proposed regression model in terms of variable selection and out-of-sample prediction. We consider three link functions, $f_1(u) = 1 - u$, $f_2(u) = 1/(1+e^{-u})$, and $f_3(u) = \sin(u)$, and simulate responses from model (2.1). The predictors $\mathbf{X}_i$s are the 820 functional connectivity matrices estimated from the HCP data with the glasso estimator. We consider two settings for the number of nodes in the covariate matrices. In the first setting, the $\mathbf{X}_i$s will consist of only the estimated edges between the 25 nodes in either the visual and default mode network. In the second, the $\mathbf{X}_i$s will consist of the estimated edges between all 100 nodes. We will refer to these two settings as $p = 25$ and $p = 100$ respectively.

Continuous responses are simulated from a normal distribution with mean $f(\langle \mathbf{X}_i, \mathbf{B} \rangle)$ and variance $\text{Var}\{f(\langle \mathbf{X}_i, \mathbf{B} \rangle)\}/c_s$, with $c_s = 5$ in high signal-to-noise (SNR) settings and $c_s = 2$ in low SNR settings. Binary responses are simulated from a Bernoulli(p_i) distribution with $\log\{p_i/(1-p_i)\} = c_s \times f(\langle \mathbf{X}_i, \mathbf{B} \rangle)$, with $c_s = 5$ in high signal settings and $c_s = 2$ in low signal settings. We consider three cases for the coefficient matrix $\mathbf{B}^*$. The first case is the adjacency matrix of a circle network (Wang & Li, 2012), where $\mathbf{B}^* = (B^*_{jk})$ with $B^*_{j+1,j} = B^*_{p,1} = 1$ for $j = 1, \ldots, p-1$ and the upper triangle elements are matched with the lower triangle elements for symmetry. The second case is a random

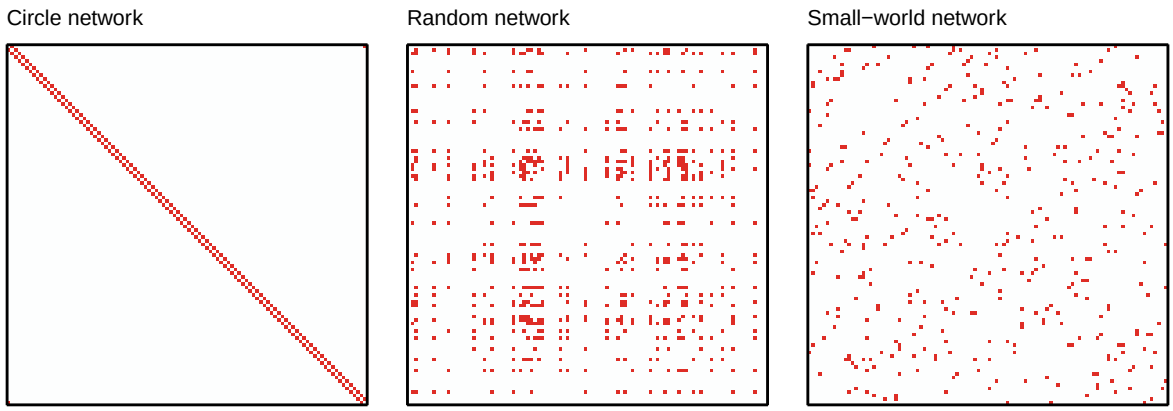

Figure 2.2 The three cases for the true coefficient matrix in the simulation settings with $p = 100$. Red elements indicate values of 1, and white elements indicate a value of 0.

symmetric matrix where each node (column/row) has a probability of 0.5 of being included in the network, and each edge has a probability of 0.5 being included, conditional on node inclusion. The value of each included element is 1. The third case is the adjacency matrix of a small-world network (Watts & Strogatz, 1998), a graph with a high clustering coefficient and a short average path length. The small-world network is generated according to the Watts-Strogatz model as implemented in the R package *igraph* (Csardi & Nepusz, 2006). The random symmetric matrix demonstrates node and edge sparsity, while the circle network adjacency matrix demonstrates only edge sparsity. The three cases for the coefficient matrix in the simulation settings with $p = 100$ are illustrated in Figure 2.2.

While we have proposed the node & edge sparse regularization term for the estimation of $\mathbf{B}$, other regularization terms may also be used to estimate the coefficient matrix in the second step of the two-step algorithm described in Section 2.3. Specifically, we may employ the lasso penalty $\mathcal{R}(\mathbf{B}) = \lambda \sum_{j,k} |B_{jk}|$ and the elastic net penalty $\mathcal{R}(\mathbf{B}) = \delta\lambda \sum_{j,k} |B_{jk}| + (1-\delta)\lambda \sum_{j,k} B_{jk}^2$ (Zou & Hastie, 2005), where $\delta \in (0, 1)$ is a weight parameter. A simple modification to the proposed estimation procedure of $\mathbf{B}$ can be used and hence the details are omitted. 5-fold cross-validation is used for tuning the parameter(s). A total of 50 simulations are conducted under each setting, and we compare the node & edge penalty with the lasso and elastic net penalties. In addition, we also fit linear models by setting the estimate of the smooth function f to the identity function and running only the second step of the two-step algorithm.

2.4.2 Simulation Results

The simulation results are summarized for the random network coefficient matrix in Section A.2 of Appendix A. Predictive performance is measured by the mean squared errors of the predictions in the case of continuous responses, and by area under the ROC curve in the case of the binary responses. True positive rate is defined as the proportion of elements of $\mathbf{B}$ that are correctly estimated as zero. Similarly, false positive rate is defined as the proportion of elements of $\mathbf{B}$ that are incorrectly estimated as zero. We do not include a measure of estimation error of the coefficient matrix as $\mathbf{B}$ is only identifiable up to a constant of proportionality.

As expected, there is no significant difference in the performances of the linear and nonlinear models when the index function is linear. However, when the index function is nonlinear, the nonlinear models clearly outperform the linear models in prediction error, true positive rate, and false positive rate. Further, the lasso and elastic net penalties both result in node sparsity less frequently than the node-edge penalty. While the results of the three penalties are often similar, the node & edge model results in the lowest prediction errors and best sparsity classifications more often than not. This, along with the desirable interpretation of the node & edge sparsity pattern, demonstrates the value of the proposed model.

2.5 Data Application

We apply the proposed methods to the HCP data discussed in Section 2.2. The predictors, i.e., the functional connectivity matrices, can be estimated using either the inverted Ledoit-Wolf estimator, the glasso, or the joint glasso as described in Section 2.3. To assess their resulting performances, we compute the three estimators for all subjects in the HCP study. The estimates for 80% of the subjects are used to fit the proposed single index models of sex and fluid intelligence and the estimated model components are then used to generate out-of-sample predictions for the remaining subjects.

The ROC curves for the fitted models of sex using the three estimators of functional connectivity are compared in Figure 2.3. We see no apparent difference between the estimator imposing individual level sparsity (glasso) and the estimator that does not impose sparsity (Ledoit-Wolf). The results suggest that while imposing sparsity in the estimation of functional connectivity leads to clearer interpretation of model results, the

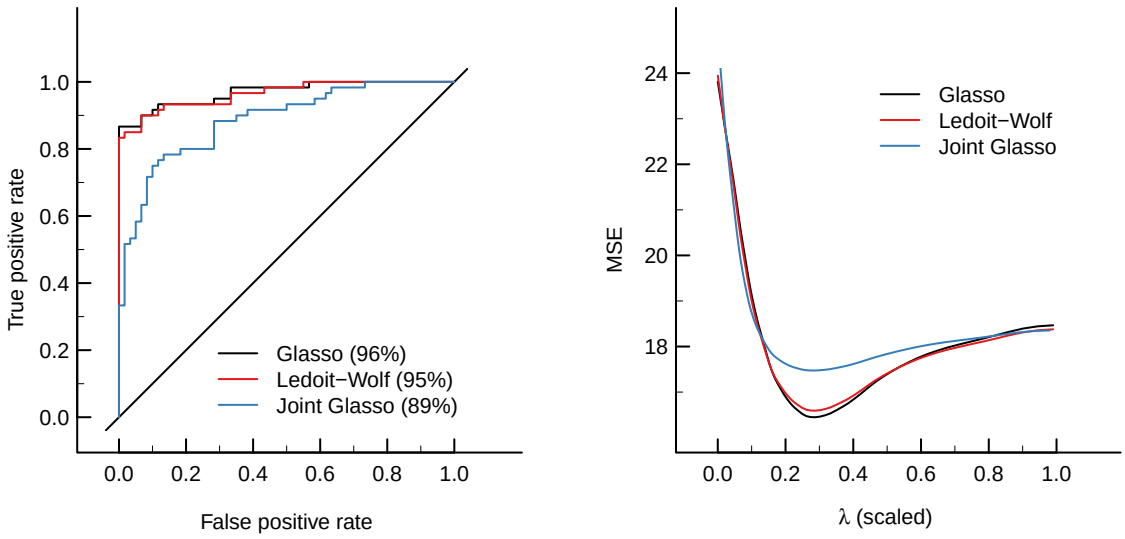

Figure 2.3 (Left) ROC curves of the single index model predictions of sex using the glasso, Ledoit-Wolf, and joint glasso. AUC values are reported along with legend labels. (Right) MSE of the single index model predictions of fluid intelligence by level of regularization.

predictive performance of the node & edge model is not affected by the inclusion of non-zero estimates of the node weights between conditionally independent nodes. We also see that the joint glasso leads to a lower area under the ROC curve (AUC), which might be due to individual differences in functional connectivity being dampened by the joint regularization. The mean squared errors (MSE) of the models of fluid intelligence are plotted against the level of regularization in Figure 2.3. The conclusions regarding the three estimators of functional connectivity are the same as in the model for the sex variable. Based on these results, we suggest the use of the glasso in application.

While the Ledoit-Wolf and glasso estimates result in similar out-of-sample predictive performance, the resulting coefficient matrix estimates do show a slight difference in levels of sparsity. In the model of sex, the estimate of the coefficient matrix using the Ledoit-Wolf estimates of functional connectivity has 77% edge-wise sparsity, while the estimate of the coefficient matrix using the glasso estimates has 79%. In the model of fluid intelligence, the Ledoit-Wolf estimates and glasso estimates have edge-wise sparsities of 89% and 92% respectively. Defining node-wise sparsity as the proportion of nodes with more than 95% of their associated edges estimated to be zero, the estimate of the

coefficient matrix using the Ledoit-Wolf estimates of functional connectivity has 34% node-wise sparsity, while the estimate of the coefficient matrix using the glasso estimates has 37%. In the model of fluid intelligence, the Ledoit-Wolf estimates and glasso estimates have edge-wise sparsities of 73% and 76% respectively.

Figure 2.4 illustrates the estimated coefficient matrix and mean function of the single index model of sex and of fluid intelligence. The selected nodes agree with with existing studies; for example, Hearne et al. (2016) identified a relationship between fluid intelligence and functional connectivity between the default mode and frontoparietal networks, and Finn et al. (2015) found that connectivity associated with the frontoparietal network was predictive of individuals' fluid intelligence. The effective degrees of freedom in the estimate of the f was 2.9 in the model of sex, and 4.8 in the model of fluid intelligence (compared to 2 degrees of freedom of linear and logistic fits), indicating nonlinear relationships between the response and covariates in both models.

We performed similar analyses on two additional response variables, namely, subject age and depression score assessed by the Beck Depression Inventory–II (Beck et al., 1996). The results are provided in Section A.3 of Appendix A.

2.6 Discussion

Here, we proposed a single index model that relates brain functional connectivity networks to response variables to deal with potential nonlinear relationships. We employed two types of sparsity in model estimation motivated by the network structure of the functional connectivity matrix. The induced sparsity patterns provided variable selection in both the edges and the nodes of the network associated with a sparse estimate of functional connectivity. Therefore, the proposed model could identify links between regions that relate to the response and directly assess whether the functional connectivity associated with a brain region is related to the response variable. We found in simulation studies that the proposed model has better prediction error and true positive and false positive rates on average in estimating network sparsity patterns than linear models and nonlinear models that do not exploit network structure. We also found that sparse estimation of functional connectivity matrices led to more interpretable model estimates via applications to the HCP data. Finally, we applied the proposed approach to perform sex discrimination and fluid intelligence prediction in the HCP dataset. The most discriminative connectivity

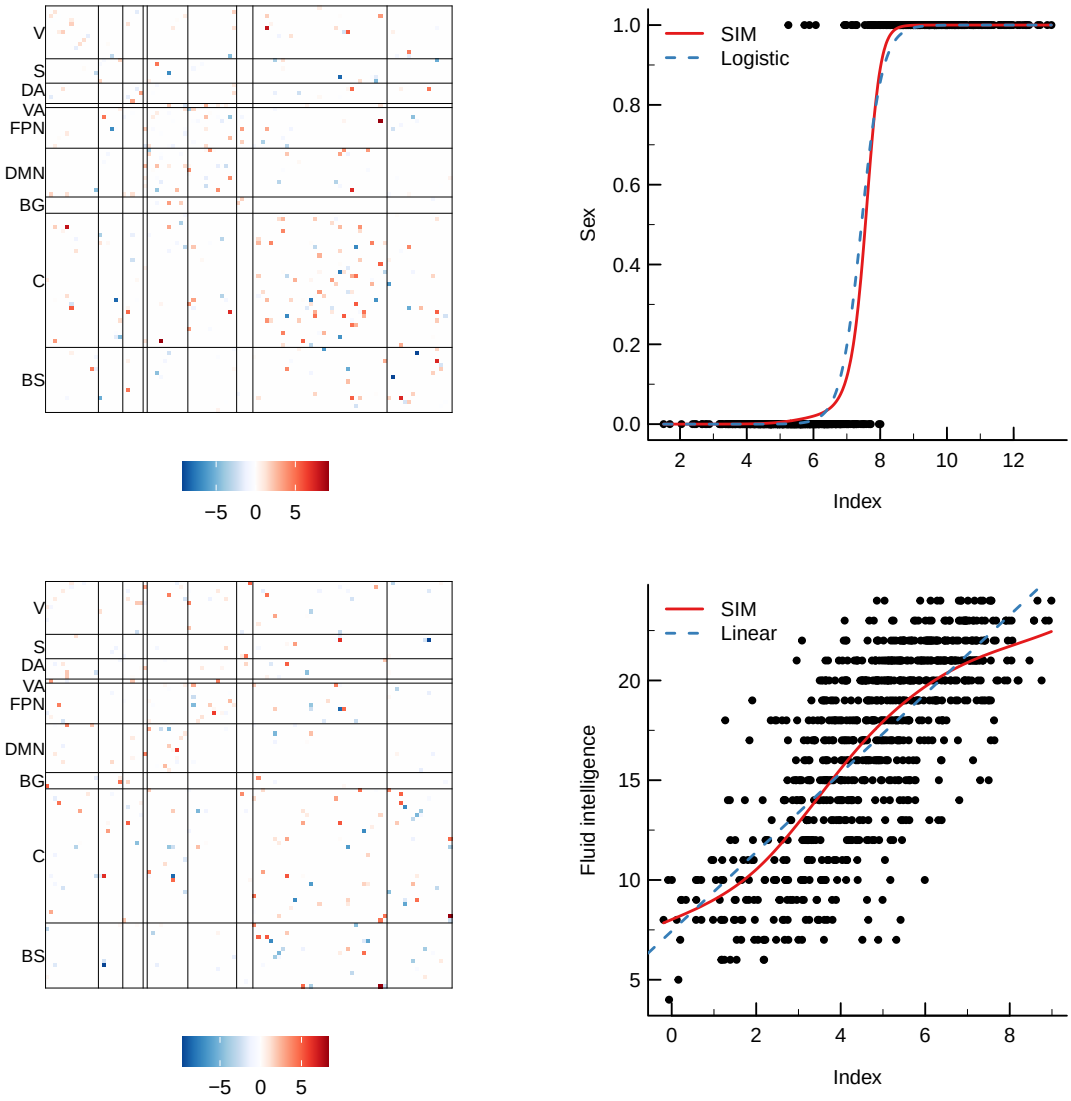

Figure 2.4 Estimated coefficient matrix and mean function for the single index models of sex (top; male is coded as 0 and female is coded as 1) and fluid intelligence (bottom).

between regions and their effects reported by our model were consistent with existing results.

While the proposed model was developed using functional connectivity matrix data, it can be applied to any predictor with a form of symmetric matrices. In this work, the proposed model was applied to rs-fMRI data in the HCP; however, we anticipate that it might be useful for task-based data as well. In fact, while the methods here were developed specifically for functional connectivity matrices, they are appropriate for modeling linear relationship between response variables and any other symmetric matrix or network-valued predictors.

Finally, the proposed model allows for more general matrix predictors and other regularization terms may be useful. Some popular matrix penalties include the row penalty, column penalty, and the rank penalty, which penalizes the nuclear norm of the coefficient matrix. Zhou et al. (2013) and Zhou & Li (2014) discuss regularization techniques in regression models with matrix-valued covariates and Feng et al. (2020) make use of general matrix regularization penalties in fitting single index models. If the observed data includes predictors that involve spatial relationships, it may be of interest to involve a penalty based on the fused lasso (Tibshirani et al., 2005), such as the GraphNet penalty (Grosenick et al., 2013).

Table 2.1 Mean squared error (standard error) of the normal response simulations and AUC (standard error) from simulations with a random network coefficient matrix.

			Normal response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.175 (0.119)	0.177 (0.121)	0.181 (0.124)	0.170 (0.111)	0.174 (0.113)	0.168 (0.109)
		High	0.079 (0.032)	0.079 (0.033)	0.080 (0.033)	0.080 (0.034)	0.081 (0.034)	0.079 (0.034)
	100	Low	0.896 (0.200)	0.930 (0.208)	0.965 (0.216)	0.997 (0.204)	1.033 (0.212)	0.963 (0.197)
		High	0.480 (0.163)	0.509 (0.173)	0.540 (0.182)	0.510 (0.165)	0.541 (0.174)	0.481 (0.155)
f_2	25	Low	0.021 (0.007)	0.021 (0.007)	0.021 (0.007)	0.150 (0.061)	0.153 (0.062)	0.149 (0.060)
		High	0.009 (0.005)	0.009 (0.005)	0.009 (0.005)	0.072 (0.038)	0.073 (0.038)	0.071 (0.037)
	100	Low	0.045 (0.029)	0.046 (0.031)	0.048 (0.032)	0.942 (0.226)	0.976 (0.233)	0.910 (0.217)
		High	0.030 (0.019)	0.031 (0.020)	0.033 (0.021)	0.564 (0.164)	0.594 (0.172)	0.536 (0.156)
f_3	25	Low	0.146 (0.087)	0.147 (0.087)	0.150 (0.088)	0.523 (0.204)	0.532 (0.206)	0.514 (0.201)
		High	0.088 (0.044)	0.089 (0.044)	0.091 (0.045)	0.494 (0.111)	0.503 (0.109)	0.484 (0.112)
	100	Low	0.756 (0.041)	0.761 (0.040)	0.764 (0.040)	0.962 (0.225)	0.998 (0.234)	0.928 (0.216)
		High	0.592 (0.024)	0.598 (0.025)	0.601 (0.025)	0.504 (0.124)	0.533 (0.130)	0.477 (0.119)

			Binary response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.700 (0.041)	0.697 (0.041)	0.697 (0.041)	0.770 (0.074)	0.746 (0.096)	0.786 (0.073)
		High	0.851 (0.036)	0.850 (0.036)	0.850 (0.036)	0.910 (0.042)	0.900 (0.045)	0.919 (0.040)
	100	Low	0.701 (0.036)	0.692 (0.037)	0.692 (0.037)	0.806 (0.085)	0.777 (0.098)	0.831 (0.082)
		High	0.839 (0.048)	0.826 (0.054)	0.826 (0.054)	0.892 (0.059)	0.873 (0.063)	0.910 (0.055)
f_2	25	Low	0.804 (0.061)	0.789 (0.063)	0.772 (0.065)	0.542 (0.033)	0.542 (0.033)	0.545 (0.033)
		High	0.923 (0.028)	0.914 (0.030)	0.905 (0.032)	0.655 (0.080)	0.655 (0.080)	0.664 (0.072)
	100	Low	0.843 (0.075)	0.819 (0.078)	0.794 (0.080)	0.540 (0.056)	0.540 (0.056)	0.542 (0.043)
		High	0.929 (0.050)	0.912 (0.055)	0.894 (0.059)	0.693 (0.165)	0.693 (0.165)	0.678 (0.161)
f_3	25	Low	0.801 (0.054)	0.784 (0.058)	0.767 (0.061)	0.568 (0.045)	0.568 (0.045)	0.576 (0.046)
		High	0.910 (0.036)	0.901 (0.037)	0.919 (0.034)	0.674 (0.133)	0.674 (0.133)	0.686 (0.127)
	100	Low	0.795 (0.097)	0.763 (0.116)	0.826 (0.083)	0.574 (0.099)	0.574 (0.099)	0.586 (0.099)
		High	0.910 (0.046)	0.889 (0.052)	0.929 (0.040)	0.602 (0.073)	0.602 (0.073)	0.612 (0.083)

Table 2.2 True positive rate (false positive rate) from simulations with a random network coefficient matrix.

			Normal response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.844 (0.028)	0.805 (0.019)	0.758 (0.011)	0.864 (0.028)	0.821 (0.020)	0.772 (0.012)
		High	0.827 (0.000)	0.780 (0.000)	0.727 (0.000)	0.828 (0.000)	0.785 (0.000)	0.731 (0.000)
	100	Low	0.985 (0.943)	0.982 (0.934)	0.979 (0.924)	0.987 (0.950)	0.984 (0.942)	0.980 (0.932)
		High	0.974 (0.895)	0.970 (0.881)	0.966 (0.869)	0.973 (0.893)	0.968 (0.880)	0.964 (0.867)
f_2	25	Low	0.887 (0.004)	0.848 (0.003)	0.799 (0.003)	0.879 (0.022)	0.842 (0.011)	0.798 (0.008)
		High	0.862 (0.001)	0.822 (0.001)	0.768 (0.001)	0.870 (0.012)	0.831 (0.009)	0.783 (0.008)
	100	Low	0.985 (0.943)	0.982 (0.932)	0.978 (0.922)	0.992 (0.973)	0.990 (0.968)	0.988 (0.961)
		High	0.973 (0.897)	0.969 (0.888)	0.965 (0.876)	0.988 (0.954)	0.986 (0.948)	0.984 (0.941)
f_3	25	Low	0.878 (0.013)	0.840 (0.006)	0.796 (0.005)	0.974 (0.662)	0.956 (0.612)	0.933 (0.575)
		High	0.833 (0.001)	0.791 (0.001)	0.749 (0.000)	0.974 (0.734)	0.959 (0.690)	0.940 (0.638)
	100	Low	0.981 (0.938)	0.978 (0.928)	0.974 (0.916)	1.000 (1.000)	0.999 (0.999)	0.998 (0.998)
		High	0.970 (0.889)	0.966 (0.877)	0.961 (0.866)	1.000 (1.000)	0.999 (0.999)	0.998 (0.998)

			Binary response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.940 (0.611)	0.915 (0.546)	0.880 (0.483)	0.932 (0.562)	0.901 (0.509)	0.864 (0.437)
		High	0.867 (0.090)	0.829 (0.058)	0.780 (0.036)	0.860 (0.090)	0.818 (0.059)	0.776 (0.038)
	100	Low	0.992 (0.980)	0.990 (0.975)	0.987 (0.969)	0.993 (0.984)	0.991 (0.978)	0.988 (0.973)
		High	0.986 (0.958)	0.983 (0.950)	0.980 (0.941)	0.985 (0.953)	0.982 (0.944)	0.978 (0.934)
f_2	25	Low	0.926 (0.590)	0.898 (0.521)	0.857 (0.457)	0.992 (0.980)	0.979 (0.944)	0.962 (0.912)
		High	0.862 (0.111)	0.821 (0.076)	0.772 (0.064)	0.949 (0.696)	0.922 (0.631)	0.895 (0.575)
	100	Low	0.991 (0.976)	0.989 (0.973)	0.987 (0.966)	0.999 (0.999)	0.999 (0.999)	0.998 (0.997)
		High	0.984 (0.948)	0.981 (0.939)	0.977 (0.930)	0.996 (0.991)	0.994 (0.989)	0.992 (0.985)
f_3	25	Low	0.932 (0.571)	0.906 (0.503)	0.868 (0.415)	0.988 (0.926)	0.976 (0.898)	0.953 (0.856)
		High	0.878 (0.100)	0.831 (0.067)	0.786 (0.043)	0.955 (0.707)	0.940 (0.660)	0.911 (0.624)
	100	Low	0.992 (0.981)	0.990 (0.977)	0.987 (0.970)	0.999 (0.999)	0.999 (0.999)	0.997 (0.998)
		High	0.984 (0.949)	0.980 (0.941)	0.977 (0.932)	0.999 (0.999)	0.999 (0.999)	0.997 (0.997)

Chapter 3

Functional Data Analysis for Longitudinal Data with Informative Observation Times

3.1 Introduction

In functional data analysis for longitudinal data, the observation process is typically assumed to be non-informative, i.e., it is assumed to be independent of longitudinal outcomes. However, this assumption is often violated in real applications. For example, in a longitudinal study, the frequency and timing of a subject's observations may be correlated with the values of longitudinal outcomes. Data collected in such a setting is referred to as longitudinal data with informative observation times.

A motivating example is a longitudinal dataset of trajectories of patient-reported symptom severity of Parkinson's disease (Mischley et al., 2017).[1] The data consist of self-reported severity of each of 33 symptoms on a scale of 0 to 100 for 371 patients, with the higher number reflecting greater symptom severity. The sum of these symptoms generates the total patient-reported outcome in Parkinson's disease (PRO-PD) score. The number of observations for each patient ranges from 2 to 38, and observations are made on a voluntary basis at varying times since a patient's initial diagnosis. Patients entered the study at arbitrary times since diagnosis. There are no instances of dropout and we only consider data observed less than 20 years following any patient's initial diagnosis.

[1] PRO-PD data were provided by Dr. Laurie Mischley from Bastyr University.

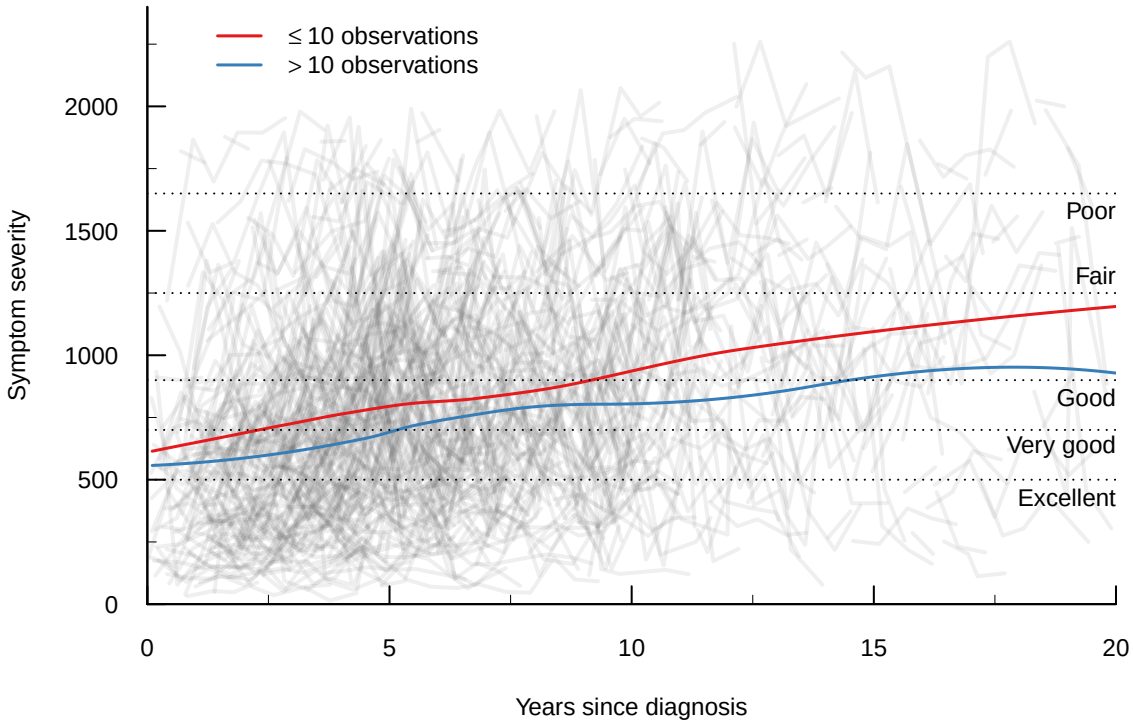

Figure 3.1 The 371 trajectories of symptom severity along with smoothed mean profiles of subjects with less than or equal to 10 observations and of subjects with more than 10 observations. Reference lines indicate the average PRO-PD of subjects answering the prompt "In general, would you say your quality of life is:" with either "poor", "fair", "good", "very good", or "excellent".

Figure 3.1 illustrates the 371 trajectories along with smoothed group-mean profiles of subjects with 10 or fewer observations and of subjects with more than 10 observations in relation to patient-reported quality of life (Cella et al., 2010). The average (standard error) PRO-PD score of subjects with 10 or fewer observations is 828.3 (16.5) and the average PRO-PD score of subjects with more than 10 observations is 752.4 (13.2). It is apparent that subjects reporting lower symptom severity tend to report more frequent observations, which provides evidence against the assumption of non-informative observation times.

It is well understood that standard methods for longitudinal data that ignore the informative observation processes may result in bias in parameter estimation. Hence, a number of methods have been proposed for the situation in which longitudinal outcomes are correlated with observation times (Lin et al., 2004; Sun et al., 2005; Sun et al., 2007;

Liang et al., 2009; Zhu et al., 2011; Fang et al., 2016; Lin et al., 2004; Song et al., 2012; Zhao et al., 2012; Sun et al., 2012; Cai et al., 2012). Among them, a commonly used modeling strategy is to consider shared random effects models (e.g. Sun et al., 2005; Liang et al., 2009). Such models can allow flexible dependence between longitudinal outcomes and observation times.

However, existing literature for longitudinal data with informative observation times mostly focuses on the estimation of an unspecified smooth mean function or integrated mean function. In the analysis of longitudinal data, besides the mean function, it might be also interesting to consider a nonparametric covariance function, which captures the patterns of subject-specific trajectories. A popular nonparametric modeling approach for estimating both mean and covariance functions of longitudinal data is via methods of functional data analysis (Ramsay & Silverman, 2006), in which one views the values of a longitudinal trajectory as observations from a random smooth function; see Guo (2002), Yao et al. (2003) and Yao et al. (2005). The advantage of functional data methods compared to traditional parametric longitudinal methods is that they allow for unspecified mean/covariance functions, which may better capture complex variation patterns of subject-specific trajectories.

Theoretical properties of functional data analysis have been well studied in several papers (Li & Hsing, 2010; Cai & Yuan, 2011; Zhang & Wang, 2016; Xiao, 2020). In particular, Zhang & Wang (2016) and Xiao (2020) considered general weighting schemes for aggregating multiple observations of each subject in the estimation of mean and covariance functions of functional data. The theories therein assume that the observation times are either fixed or randomly sampled and independent from the longitudinal outcomes. Hence, existing theories are not directly applicable to settings in which the observation process is informative.

In this chapter, we consider functional data methods for longitudinal data with informative observation times. We use a heuristic argument to show that commonly used functional data methods may be inconsistent for such data. Using the same argument, we identify one particular functional data method that could give consistent estimation under the considered shared random effect model. The difference in these methods lies in how the multiple observations of each subject are aggregated to formulate the weighted least square estimation. This contrasts dramatically with existing functional data theories, which found that both methods lead to consistent estimation. We derive the corresponding rates of

convergence for the suggested method using penalized splines. The derived convergence rates are either optimal or the best in the literature.

The remainder of this chapter is organized as follows. In Section 3.2, we begin by introducing notation and specifying the data model. Next, the estimation of the mean and covariance functions is discussed. In Section 3.3, we present the theoretical properties of the estimators. In Section 3.4, we present numerical results from a simulation study to assess the performance of the methods. In Section 3.5, we apply the methods to the dataset of trajectories of Parkinson's disease patient-reported symptom severity. In Section 3.6, we conclude with a discussion and comments on further research.

3.2 Methods

3.2.1 Model Specification

Let $Y_i(t)$ be subject i's underlying longitudinal outcome ($i = 1, \ldots, n$) at time $t \in \mathcal{T}$ where $\mathcal{T}$ is a compact time interval. Without loss of generality, let $\mathcal{T} = [0, 1]$. The observations $Y_{ij} = Y_i(T_{ij})$ of subject i are taken at time points $T_{i1} < T_{i2} < \ldots < T_{im_i}$. Let $C_i \in \mathcal{T}$ denote the censoring time of the ith subject. The cumulative number of observations for the ith subject by time t is $N_i(t) = N_i^*(t \wedge C_i)$, where $N_i^*(t) = \sum_{j \geq 1} I(T_{ij} \leq t)$. Here, I is the indicator function and $s \wedge t$ denotes the minimum of s and t. The observed values of $Y_i(t)$ are obtained at the jump points of $N_i(t)$. Consider the functional data model

$$Y_i(t) = \mu^*(t, Z_i) + X_i(t) + \epsilon_i(t), \tag{3.1}$$

where $\mu^*(\cdot, Z_i)$ is a conditional mean function parameterized by the latent frailty variable Z_i, $X_i(\cdot)$ is a zero-mean random function that models subject i's smooth deviation from the mean, and $\epsilon_i(\cdot)$ is zero-mean white noise with finite variance σ_ϵ^2. Define the marginal mean function $\mu(t) = \mathrm{E}\{\mu^*(t, Z_i)\}$, the covariance function of the conditional mean functions $\sigma_{\mu^*}(s, t) = \mathrm{Cov}\{\mu^*(s, Z_i), \mu^*(t, Z_i)\}$, and the covariance function of the random subject-specific functions $\sigma_X(s, t) = \mathrm{Cov}\{X_i(s), X_i(t)\}$. Then the responses can be considered as observations from a random function with mean function $\mu(t)$ and covariance function $\sigma(s, t) = \sigma_{\mu^*}(s, t) + \sigma_X(s, t)$, contaminated with measurement error.

As commonly considered in shared random effects models (e.g. Liang et al., 2009), we make the following assumptions: (1) censoring time C_i is independent of the observation

times $\{T_{ij}, j \geq 1\}$ and longitudinal outcomes $\{Y_i(T_{ij}), j \geq 1\}$; (2) given the frailty variable Z_i, the observation process $N_i^*(\cdot)$ is independent of the longitudinal outcomes $\{Y_i(T_{ij}), j \geq 1\}$; and (3) given the frailty variable Z_i, the observation process $N_i^*(t)$ follows a Poisson process with the intensity function $\lambda(t, Z_i) = Z_i\lambda_0(t)$, where $\lambda_0(t)$ is an unspecified intensity function. Define $\Lambda_0(t) = \int_0^t \lambda_0(s)\,\mathrm{d}s$, the baseline cumulative intensity function.

The dependence between longitudinal outcomes and observation times is induced by the latent frailty variable Z_i, but the form and magnitude of the dependence is left completely unspecified for flexibility. To simplify theoretical analysis, we allow m_i to be 0. In formulas and equations involving m_i^{-1}, the term m_i^{-1} should always be interpreted as $m_i^{-1}I(m_i \neq 0)$.

3.2.2 Estimation of Mean Function

The mean function $\mu(t)$ is approximated by a spline function $B^\top(t)\boldsymbol{\theta}$ where $B(t) = \{B_1(t), \ldots, B_K(t)\}^\top \in \mathbb{R}^K$ is a vector of rth order B-spline basis functions constructed from equally spaced knots in $\mathcal{T}$, and $\boldsymbol{\theta}$ is a vector of coefficients. The coefficient vector $\boldsymbol{\theta}$ is estimated via the minimization

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_i \left\{ Y_{ij} - B^\top(T_{ij})\boldsymbol{\theta} \right\}^2 + \lambda \boldsymbol{\theta}^\top P \boldsymbol{\theta} \right], \tag{3.2}$$

where w_is are nonnegative weights that may depend on m_i and C_i, λ is a smoothing parameter that balances fit and smoothness, and P is a penalty matrix such that $\boldsymbol{\theta}^\top P \boldsymbol{\theta}$ equals the squared sum of the qth order consecutive differences of $\boldsymbol{\theta}$ as in Eilers & Marx (1996). The mean function is then estimated by $\widehat{\mu}(t) = B^\top(t)\widehat{\boldsymbol{\theta}}$.

A common choice of the weights is $w_i = 1$, which means that each observation has equal weight in forming the least squares in (3.2). We will denote this choice of weights by "OBS", following Zhang & Wang (2016) and Xiao (2020). As far as we know, this is the most often used choice for formulating the least squares in functional data analysis. Another choice of weights is $w_i = m_i^{-1}$, which implies that the set of observations from each subject has equal weight (e.g. Li & Hsing, 2010). We will denote this choice of weights by "SUBJ". Both choices of weights could lead to consistent and even rate-optimal estimation in functional data analysis as shown in Zhang & Wang (2016) and Xiao (2020) for functional data with non-informative observation times. However, for functional

data with informative observation times, we shall derive heuristically different weighting schemes will affect the estimation in the following derivation.

Let $\mathcal{A}_i(t) = w_i\{Y_i(t) - \mu_n(t)\}\,\mathrm{d}N_i(t)$, where $\mu_n(t)$ is the solution to the estimating equation $\sum_{i=1}^n \mathcal{A}_i(t) = 0$. Then, we have

$$\mu_n(t) = \frac{\sum_{i=1}^n w_i Y_i(t)\,\mathrm{d}N_i(t)}{\sum_{i=1}^n w_i\,\mathrm{d}N_i(t)}. \tag{3.3}$$

As shown in Liang et al. (2009),

$$\mathrm{E}\left\{\mathrm{d}N_i(t) \mid m_i, C_i\right\} = I(C_i \geq t) m_i \frac{\mathrm{d}\Lambda_0(t)}{\Lambda_0(C_i)}.$$

Thus,

$$\begin{aligned}
&\mathrm{E}\left\{w_i Y_i(t)\,\mathrm{d}N_i(t) \mid C_i\right\} \\
=&\mathrm{E}\left[w_i \mathrm{E}\left\{Y_i(t)\,\mathrm{d}N_i(t) \mid m_i, C_i\right\} \mid C_i\right] \\
=&\mathrm{E}\left[w_i I(C_i \geq t) m_i \frac{\mathrm{d}\Lambda_0(t)}{\Lambda_0(C_i)} \mathrm{E}\left\{\mu^*(t, Z_i) \mid m_i, C_i\right\} \mid C_i\right] \\
=&I(C_i \geq t) \frac{\mathrm{d}\Lambda_0(t)}{\Lambda_0(C_i)} \mathrm{E}\left\{\mu^*(t, Z_i) w_i m_i \mid C_i\right\}.
\end{aligned}$$

Similarly, we derive that

$$\mathrm{E}\left\{w_i\,\mathrm{d}N_i(t) \mid C_i\right\} = I(C_i \geq t) \frac{\mathrm{d}\Lambda_0(t)}{\Lambda_0(C_i)} \mathrm{E}\left(w_i m_i \mid C_i\right).$$

It follows by equation (3.3) that as n increases to infinity, $\mu_n(t)$ converges to

$$\mu^o(t) \equiv \frac{\mathrm{E}\left[\frac{I(C_i \geq t)}{\Lambda_0(C_i)} \mathrm{E}\left\{\mu^*(t, Z_i) w_i m_i \mid C_i\right\}\right]}{\mathrm{E}\left\{\frac{I(C_i \geq t)}{\Lambda_0(C_i)} \mathrm{E}\left(w_i m_i \mid C_i\right)\right\}}. \tag{3.4}$$

We now evaluate expression (3.4) with either the OBS weights or the SUBJ weights. First, consider the SUBJ weights for which $w_i = m_i^{-1}$. Then $w_i m_i = 1$. Because Z_i is independent of C_i, $\mu^o(t) = \mathrm{E}\{\mu^*(t, Z_i) \mid C_i\} = \mathrm{E}\{\mu^*(t, Z_i)\} = \mu(t)$. Hence, for the SUBJ weights, $\mu_n(t)$ converges to $\mu(t)$ as desired. Second, consider the OBS weights for which $w_i = 1$. In general, in this case $\mu^o(t)$ in (3.4) is different from $\mu(t)$. To give an example,

consider the simulation setting in Section 3.4, where we let $\mu^*(t, Z) = Z(t^2 + 2/3)$, $Z = (1 - \rho) + \rho Z^*$ for some $0 < \rho < 1$, and Z^* is from a distribution with mean 1 and variance $\sigma^2_{Z^*}$. In addition, m follows a Poisson distribution with mean $10Z\Lambda_0(C)$. Here (Z, m, C) has the same distribution as (Z_i, m_i, C_i). Then we derive that $\mathrm{E}(m \mid C) = 10\mathrm{E}\{Z\Lambda_0(C) \mid C\} = 10\Lambda_0(C)$, $\mu^*(t, Z)m = Zm(t^2 + 2/3)$, and $\mathrm{E}\{\mu^*(t, Z)m \mid C\} = (t^2 + 2/3)\mathrm{E}(Zm \mid C) = 10(t^2 + 2/3)\mathrm{E}\{Z^2\Lambda_0(C) \mid C\} = 10\Lambda_0(C)(t^2 + 2/3)(1 + \rho^2\sigma^2_{Z^*})$. It follows that $\mu^o(t) = (t^2 + 2/3)(1 + \rho^2\sigma^2_{Z^*})$, which differs from $\mu(t) = t^2 + 2/3$. Indeed, our simulation study in Section 3.4 will provide empirical confirmation of the above derivations. Another general example is when $C_i = 1$ for all i, i.e., no censoring. Then it is easy to see that (3.4) reduces to $\mu^o(t) = \mathrm{E}\{\mu(t, Z)m\}/\mathrm{E}(m)$, which differs from $\mu(t) = \mathrm{E}\{\mu^*(t, Z)\}$ unless $\mu^*(t, Z)$ and m are uncorrelated.

The above derivations show that the OBS weights lead to inconsistent estimation of the mean function while in contrary the SUBJ weights might give consistent estimation. This differs from traditional functional data analysis, where both OBS and SUBJ weights will lead to consistent estimation (Zhang & Wang, 2016; Xiao, 2020). Thus, for functional data with informative observation times under the shared random effects model, we propose to use the SUBJ weights, and accordingly we shall rigorously derive the rate of convergence of the mean function estimation.

3.2.3 Estimation of Covariance Function

Let $\widehat{\mu}(t)$ be the estimate of the mean function using the SUBJ weights. Let $\widetilde{e}_{ij} = Y_{ij} - \widehat{\mu}(T_{ij})$ and define the auxiliary variables $\widetilde{\sigma}_{ij_1j_2} = \widetilde{e}_{ij_1}\widetilde{e}_{ij_2}$, which is an empirical estimate of $\sigma(T_{ij_1}, T_{ij_2})$. Then $\{(T_{ij_1}, T_{ij_2}, \widetilde{\sigma}_{ij_1j_2}) : 1 \le j_1 \ne j_2 \le m_i, i = 1, \ldots, n\}$ forms a data set for estimating the covariance function.

The covariance function $\sigma(s, t)$ is approximated by a tensor-product spline, $H(s, t) = \sum_{1 \le k,l \le K} \theta_{kl} B_k(s) B_l(t)$, where $\Theta = (\theta_{kl})_{1 \le k \le K, 1 \le l \le K}$ is a coefficient matrix. To simplify notation, we use the same univariate B-splines as in the mean function estimation. Since the covariance function is symmetric, it is natural to impose the constraint that $\Theta = \Theta^\top$, which ensures that $H(s, t) = H(t, s)$. Let $\mathrm{vec}(\cdot)$ be the vectorization operator, which is invertible, and define $\boldsymbol{\theta}_\sigma = \mathrm{vec}(\Theta)$. Let $B(s, t) = B(t) \otimes B(s)$, where $\otimes$ is the Kronecker product. Then $H(s, t) = B^\top(s, t)\boldsymbol{\theta}_\sigma$. The coefficient matrix Θ is estimated via

the minimization

$$\widehat{\Theta} = \arg\min_{\Theta:\,\Theta=\Theta^\top} \sum_{i=1}^{n} \left[v_i \sum_{1\le j_1\ne j_2\le m_i} \{\widetilde{\sigma}_{ij_1j_2} - H(T_{ij_1}, T_{ij_2})\}^2 \right] + \lambda K^{-1} \boldsymbol{\theta}_\sigma^\top (I_K \otimes P)\boldsymbol{\theta}_\sigma, \quad (3.5)$$

where v_is are nonnegative weights, λK^{-1} is a smoothing parameter, and P is the same penalty matrix for mean function estimation. We use the same notation λ as in the mean function estimation to simplify the notation in theoretical analysis. The above method was proposed in Xiao et al. (2018) and can be shown to be a special case of the bivariate tensor product P-splines in Eilers & Marx (2003).

Let $\widehat{\boldsymbol{\theta}}_\sigma = \mathrm{vec}(\widehat{\Theta})$. Then the covariance function $\sigma(s,t)$ is estimated by $\widehat{\sigma}(s,t) = B^\top(s,t)\widehat{\boldsymbol{\theta}}_\sigma$. As in the mean function estimation, we consider two choices of weights. We will denote the choice of $v_i = 1$ by "OBS" and the choice of $v_i = \{m_i(m_i-1)\}^{-1}$ by "SUBJ". Similar to the mean function estimation, the former weight implies each auxiliary variable carries the same weight in forming the weighted least squares while the latter implies the set of auxiliary variables from each subject carries the same weight. Then, by similar arguments (omitted) as in the mean function estimation, it can be shown that the SUBJ weights should be adopted. And we shall also derive rates of convergence of the above proposed estimator with the SUBJ weights.

3.3 Theoretical Properties

3.3.1 Notation

For a univariate function g over $\mathcal{T}$, $\|g\|$ denotes its supreme norm. The same notation is also used for a bivariate function. We also use $g^{(1)}$ to denote the derivative of a function g. For two scalars, let $a \wedge b = \min(a,b)$ and $a \vee b = \max(a,b)$. We let $\underline{m} = \{m_i, 1 \le i \le n\}$ and $\underline{C} = \{C_i, 1 \le i \le n\}$.

3.3.2 Asymptotic Properties of Mean Function Estimator

Assumption 3.1. *The nonnegative latent variables $Z_i\,(1 \le i \le n)$ are independent and identically distributed and satisfy, for any positive constant $c > 0$, $0 < \mathrm{E}(Z_1 e^{-cZ_1}) < \infty$.*

Assumption 3.2. *The censoring times $C_i\,(1 \le i \le n)$ are independent and identically distributed.*

Assumption 3.3. *The baseline hazard function $\Lambda_0(t)$ has a positive and continuous density function on $\mathcal{T}$.*

Assumption 3.4. *Let $\rho_1(t) = \mathrm{E}\{\Lambda_0^{(1)}(t\wedge C_1)/\Lambda_0(C_1)\}$ and $\rho_2(s,t) = \mathrm{E}\{\Lambda_0^{(1)}(t\wedge C_1)\Lambda_0^{(1)}(s\wedge C_1)/\Lambda_0^2(C_1)\}$. Then ρ_1 and ρ_2 are continuous and bounded away from 0 and infinity, over $\mathcal{T}$ and $\mathcal{T}^2$, respectively.*

Assumption 3.5. *(a) $\|\sigma\| < \infty$; (b) $\sigma_\epsilon^2 < \infty$.*

Assumption 3.6. *(a) The number of basis functions K satisfies $K \ge n^{\delta_1}$ for some constant $\delta_1 > 0$ and $K = o(n)$; (b) The smoothing parameter λ satisfies $\lambda = o(n^{-\delta_2})$ for some constant $\delta_2 > 0$; (c) $\log n/n = o(K^{-4})$.*

We now discuss Assumption 3.4. First, $\rho_1(t)$ is the density function of the theoretical counterpart of the empirical distribution of the observed time points. In nonparametric regression, it is usually required this function is bounded away from 0 and infinity. When there are no censoring, i.e., $C_i = 1$, then $\rho_1(t) = \Lambda_0^{(1)}(t)$ and $\rho_2(s,t) = \Lambda_0^{(1)}(s)\Lambda_0^{(1)}(t)$. Because of Assumption 3.3, Assumption 3.4 will always hold. When there are censoring, we could show that one generating distribution for C_i in the simulation study also ensures that Assumption 3.4 holds.

For theoretical results, we shall focus on the SUBJ weights and scale the weights w_i with $w_i = (nm_i)^{-1}$ so that $\sum_{i=1}^n w_i m_i = 1$. The notation r denotes the order of splines and q denotes the order of penalty. Let $h = K^{-1}$ and $h_e = h \vee \lambda^{1/(2q)}$. By Xiao (2019a), h_e plays the role of bandwidth parameter for penalized splines.

Theorem 3.1. *Suppose that Assumptions 3.1 - 3.6 hold. If $\mu \in \mathcal{C}^p(\mathcal{T})$ with $q \le p \wedge r$, then*

$$\mathrm{E}\left[\int_{\mathcal{T}} \{\widehat{\mu}(t) - \mu(t)\}^2 \,\mathrm{d}t | \underline{m}, \underline{C}\right] = O\left(h^{2r}\right) + o\left(h^{2p}\right) + O\left(\lambda^2 h_e^{-2q}\right) + O\left(n^{-1}h_e^{-1}\right), \ \textit{a.s.}$$

Proof. The proof of Theorem 3.3 in Xiao (2020) can be adapted if Lemmas A.7, A.8, and A.9 in Xiao (2020) hold almost surely, which are given in Lemma B.1 and proved in Section B.1. These lemmas deal with empirical distributions of the observed time points, which are informative, and hence different proofs are required. □

The derived L_2 convergence rate in Theorem 3.1 is the same as the those in Xiao (2020) for the scenario with a finite number observations per subject and randomly observed time points. Moreover, as shown in Claeskens et al. (2009), Xiao (2019a), and Huang & Su (2021), there are two types of asymptotics for penalized splines: (1) when $\lambda = o(h^{p+q})$, then $h_e = h$ and the rates become $O(h^{2r}) + o(h^{2p}) + O(n^{-1}h^{-1})$, which are the asymptotic rates for regression splines; (2) when $\lambda^{-1} = o(h^{-2q})$, then $h_e = \lambda^{-1/(2q)}$ and the rates become $O(\lambda) + O(n^{-1}\lambda^{-1/(2q)})$, the asymptotic rates for smoothing splines. Finally, both types of asymptotics could lead to the best rate, $n^{-2p/(2p+1)}$, which is the optimal rate for estimating univariate smooth functions (Stone, 1982).

3.3.3 Asymptotic Properties of Covariance Function Estimator

To simplify the theoretic analysis, we shall assume that the mean function $\mu(t)$ is known and the auxiliary variables in Section 3.2.3 are accordingly constructed.

Assumption 3.7. *(a)* $\sup_{t\in\mathcal{T}} \mathrm{E}\{\mu^*(t, Z_i) - \mu(t)\}^4 < \infty$*; (b)* $\sup_{t\in\mathcal{T}} \mathrm{E}\{X_i(t)\}^4 < \infty$*; (c)* $\sup_{t\in\mathcal{T}} \mathrm{E}\{\epsilon_i(t)\}^4 < \infty$.

Assumption 3.8. *Let*

$$\rho_3(s, t_1, t_2) = \mathrm{E}\{\Lambda_0^{(1)}(s \wedge C_1)\Lambda_0^{(1)}(t_1 \wedge C_1)\Lambda_0^{(1)}(t_2 \wedge C_1)/\Lambda_0^3(C_1)\}$$

and

$$\rho_4(s_1, s_2, t_1, t_2) = \mathrm{E}\{\Lambda_0^{(1)}(s_1 \wedge C_1)\Lambda_0^{(1)}(s_2 \wedge C_1)\Lambda_0^{(1)}(t_1 \wedge C_1)\Lambda_0^{(1)}(t_2 \wedge C_1)/\Lambda_0^4(C_1)\}.$$

Both ρ_3 and ρ_4 are continuous and bounded away from 0 and infinity, over $\mathcal{T}^3$ and $\mathcal{T}^4$, respectively.

Theorem 3.2. *Suppose that Assumptions 3.1 - 3.8 hold. If $\sigma \in \mathcal{C}^p(\mathcal{T}^2)$ with $q \leq p \wedge r$, then*

$$\mathrm{E}\left[\iint \{\widehat{\sigma}(s,t) - \sigma(s,t)\}^2 \,\mathrm{d}s\,\mathrm{d}t \mid \underline{m}, \underline{C}\right] = O_{\mathbb{P}}\left(h^{2r}\right) + o_{\mathbb{P}}\left(h^{2p}\right) + O_{\mathbb{P}}\left(\lambda^2 h_e^{-2q}\right) + O_{\mathbb{P}}\left(n^{-1}h_e^{-2}\right).$$

The proof of Theorem 3.2 is given in Section B.2 of Appendix B. The derived L_2 rate in Theorem 3.2 can be shown to achieve the best rate in the literature (Zhang & Wang,

2016; Xiao, 2020), $n^{-2p/(2p+2)}$, which is the optimal rate for estimating bivariate smooth functions (Stone, 1982).

3.4 Simulation Study

3.4.1 Simulation Settings

In this section, we conduct simulation studies to evaluate the performance of the proposed method, i.e., penalized splines with SUBJ weights, and compare it with that of penalized splines with OBS weights. In the following simulations, the maximum follow-up time is $\tau = 1$ and we consider three censoring time distributions: (1) $C_i = 1$, which corresponds to no censoring; (2) C_i is sampled from a mixture distribution of Uniform$(0.2, 1)$ and a point mass at 1, with mixing factors of 0.8 and 0.2, respectively; and (3) $C_i \sim$ Uniform$(0, 1)$. For the second generating distribution for the censoring times, which we call mixed censoring, Assumptions 3.4 and 3.8 still hold and hence our theoretical results will still hold; see Section B.3 for a proof. The third generating distribution, which we call uniform censoring, serves for the purpose of a sensitivity analysis. The frailty variable is chosen as $Z_i = (1 - \rho) + \rho Z_i^*$, where Z_i^* is sampled from a gamma distribution or lognormal distribution, both with mean 1 and variance $\sigma^2_{Z^*} = 0.5$ or $\sigma^2_{Z^*} = 1$, and $\rho = 0.5$ or 0.85, corresponding to the informative observation process settings, and $\rho = 0$, corresponding to the non-informative observation process settings. The observation process follows a Poisson process with the intensity function $10Z_i$. Given Z_i and C_i, the number of observations m_i of subject i then follows a Poisson distribution with the mean $10Z_iC_i$. The conditional mean function is $\mu^*(t, Z_i) = Z_i(t^2 + 2/3)$, implying a marginal mean function of $\mu(t) = t^2 + 2/3$. The covariance function of the random functions X_i is chosen as $\sigma_X(s, t) = .05 \exp(s^2) \times .05 \exp(t^2) \times (1 - \mid s - t \mid)$. The observed values $Y_i(T_{ij})$ are then generated from (3.1), where $\epsilon_i(T_{ij}) \sim \mathcal{N}(0, \sigma^2_\epsilon)$ and σ^2_ϵ is set to be 0.5 in the informative observation process settings, and 1 in the non-informative observation process settings. The number of subjects is set to $n = 150$ or $n = 300$.

We replicate each simulation setting 200 times. For univariate penalized splines, we use 10 cubic B-spline bases with equally-spaced knots in the unit interval and for the tensor-product penalized splines, we use 10 marginal cubic B-splines with equally-spaced knots in both dimensions. The smoothing parameters are selected using generalized

cross-validation via a modified version of the function *fpca.sparse* in the R package *face*, which performs the fast covariance estimation (FACE) method for sparse functional data (Xiao et al., 2018).

3.4.2 Simulation Results

Define the integrated squared error as either $\int\{\widehat{\mu}(t)-\mu(t)\}^2\,\mathrm{d}t$ or $\iint\{\widehat{\sigma}(s,t)-\sigma(s,t)\}^2\,\mathrm{d}s\,\mathrm{d}t$ where $\widehat{\mu}(t)$ and $\widehat{\sigma}(s,t)$ are estimates of $\mu(t)$ and $\sigma(s,t)$. For each simulation setting, we calculate the median and interquartile range of the integrated squared error of the estimates using the OBS and SUBJ choices of weights.

Figure 3.2 plots the averages of the estimated mean functions from penalized splines with either OBS or SUBJ weights for a few simulation settings with gamma frailty variables, for no censoring and mixed censoring times. In these plots, $\sigma^2_{Z^*} = 0.5$. A non-zero value of ρ in these plots indicate informative observation times. We first observe from the plots that penalized splines with OBS weights converge to $\mu^o(t) = (t^2 + 2/3)(1 + \rho^2\sigma^2_{Z^*})$, the derived expression in (3.4) for OBS weights, which differ from the true $\mu(t) = t^2 + 2/3$ and the difference becomes more pronounced as ρ increases. Next we see that penalized splines with SUBJ weights converge to the true mean function as desired, irrespective of the values of ρ. Thus, the plots demonstrate the inconsistent estimation of the mean function by penalized splines with OBS weights and the desired model estimation by penalized splines with SUBJ weights. The plots for uniform censoring and lognormal frailty variables are similar and are presented in Section B.4 of Appendix B.

Table 3.1 gives the median and interquantile range of integrated squared error for estimating the mean function using both methods in all simulation settings with informative observation times. The numerical results show that penalized splines with SUBJ weights have much smaller median integrated squared error for estimating the mean function, in agreement with the above plots. In addition, the estimators based on the SUBJ weights have much smaller variation in terms of interquartile range in the presence of information observation times. While there is some additional variability in the estimates at the larger time points in the mixed censoring settings compared to the no censoring settings, censoring seems to have a less significant effect on the estimates. Similar simulation results for the uniform censoring setting are found and given in Section B.4 of Appendix B.

The numerical results for the covariance function estimation under the informative observation times settings are summarized in Table 3.2, and the corresponding numerical

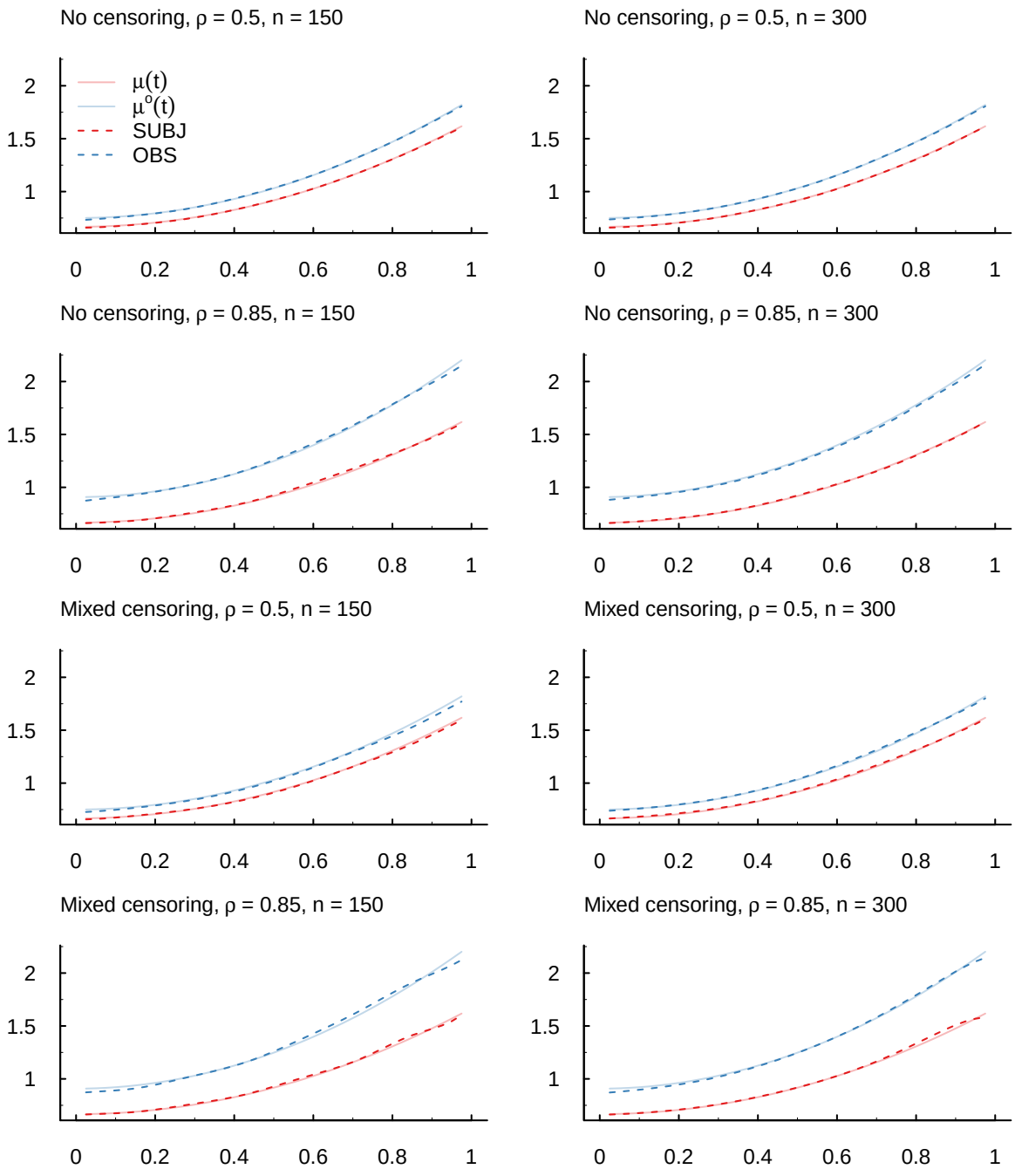

Figure 3.2 Average of 200 estimates of the mean function using penalized splines with both OBS and SUBJ weights for simulation settings in which the frailty variable Z_i^* follows a gamma distribution with variance $\sigma^2_{Z^*} = 0.5$. The true mean function is $\mu(t) = t^2 + 2/3$ and $\mu^o(t)$ is the formula in (3.4) associated with OBS weights, specifically $\mu^o(t) = (t^2 + 2/3)(1 + \rho^2\sigma^2_{Z^*})$.

Table 3.1 Median (interquartile range) of integrated squared error for estimating the mean function using penalized splines with both OBS and SUBJ weights.

			Gamma			
			No censoring		Mixed censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ
0.50	150	0.5	0.011 (0.017)	0.001 (0.002)	0.022 (0.021)	0.002 (0.002)
		1.0	0.059 (0.060)	0.002 (0.002)	0.068 (0.039)	0.004 (0.004)
	300	0.5	0.016 (0.008)	0.000 (0.000)	0.011 (0.007)	0.001 (0.002)
		1.0	0.069 (0.037)	0.001 (0.001)	0.073 (0.042)	0.003 (0.003)
0.85	150	0.5	0.171 (0.109)	0.003 (0.003)	0.102 (0.116)	0.007 (0.010)
		1.0	0.557 (0.467)	0.007 (0.009)	0.409 (0.343)	0.016 (0.038)
	300	0.5	0.141 (0.056)	0.001 (0.002)	0.102 (0.057)	0.005 (0.008)
		1.0	0.596 (0.184)	0.003 (0.003)	0.593 (0.297)	0.011 (0.014)

			Lognormal			
			No censoring		Mixed censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ
0.50	150	0.5	0.016 (0.024)	0.001 (0.002)	0.012 (0.027)	0.002 (0.003)
		1.0	0.073 (0.120)	0.001 (0.002)	0.075 (0.081)	0.005 (0.005)
	300	0.5	0.017 (0.014)	0.000 (0.000)	0.017 (0.017)	0.001 (0.001)
		1.0	0.048 (0.061)	0.001 (0.001)	0.054 (0.078)	0.002 (0.003)
0.85	150	0.5	0.141 (0.048)	0.003 (0.004)	0.098 (0.137)	0.009 (0.008)
		1.0	0.477 (0.344)	0.006 (0.009)	0.297 (0.684)	0.014 (0.012)
	300	0.5	0.102 (0.052)	0.001 (0.002)	0.107 (0.091)	0.005 (0.005)
		1.0	0.579 (0.348)	0.003 (0.003)	0.375 (0.283)	0.005 (0.009)

Table 3.2 Median (interquartile range) of integrated squared error for estimating the covariance function using penalized splines with both OBS and SUBJ weights.

			Gamma			
			No censoring		Mixed censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ
0.50	150	0.5	0.018 (0.015)	0.000 (0.001)	0.011 (0.066)	0.001 (0.002)
		1.0	0.224 (0.653)	0.004 (0.009)	0.242 (1.141)	0.008 (0.014)
	300	0.5	0.009 (0.010)	0.000 (0.000)	0.006 (0.041)	0.001 (0.001)
		1.0	0.220 (0.472)	0.003 (0.004)	0.211 (0.789)	0.005 (0.005)
0.85	150	0.5	0.324 (0.581)	0.008 (0.014)	0.033 (0.226)	0.011 (0.018)
		1.0	5.359 (21.07)	0.043 (0.157)	5.568 (22.08)	0.077 (0.049)
	300	0.5	0.235 (0.399)	0.003 (0.003)	0.068 (0.577)	0.006 (0.013)
		1.0	3.022 (15.30)	0.045 (0.060)	2.556 (5.849)	0.054 (0.059)

			Lognormal			
			No censoring		Mixed censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ
0.50	150	0.5	0.032 (0.109)	0.001 (0.002)	0.011 (0.160)	0.002 (0.004)
		1.0	2.885 (7.116)	0.011 (0.043)	0.341 (1.786)	0.012 (0.028)
	300	0.5	0.093 (0.113)	0.001 (0.002)	0.043 (0.088)	0.001 (0.002)
		1.0	0.614 (2.761)	0.005 (0.006)	0.747 (12.21)	0.010 (0.021)
0.85	150	0.5	0.691 (1.745)	0.007 (0.011)	0.652 (5.084)	0.031 (0.042)
		1.0	12.37 (74.19)	0.034 (0.106)	9.179 (96.96)	0.176 (0.238)
	300	0.5	0.550 (1.728)	0.004 (0.005)	0.562 (2.247)	0.009 (0.021)
		1.0	32.76 (93.44)	0.076 (0.133)	10.48 (129.7)	0.060 (0.078)

results under the non-informative observation times settings are included in Section B.4 of Appendix B. The estimate using the SUBJ weights clearly outperforms the estimate using the OBS weights in both cases. Similar to the results for the mean function estimation, the estimators based on the SUBJ weights have much smaller integrated squared error and variation than those based on the OBS weights under all the settings with information observation times. For illustration, Figure 3.3 presents the estimated covariance function from both methods along with the true covariance function from two simulated data sets.

In simulation settings with non-informative observations, i.e., $\rho = 0$, we find no substantial difference in the integrated squared errors for estimating the mean and

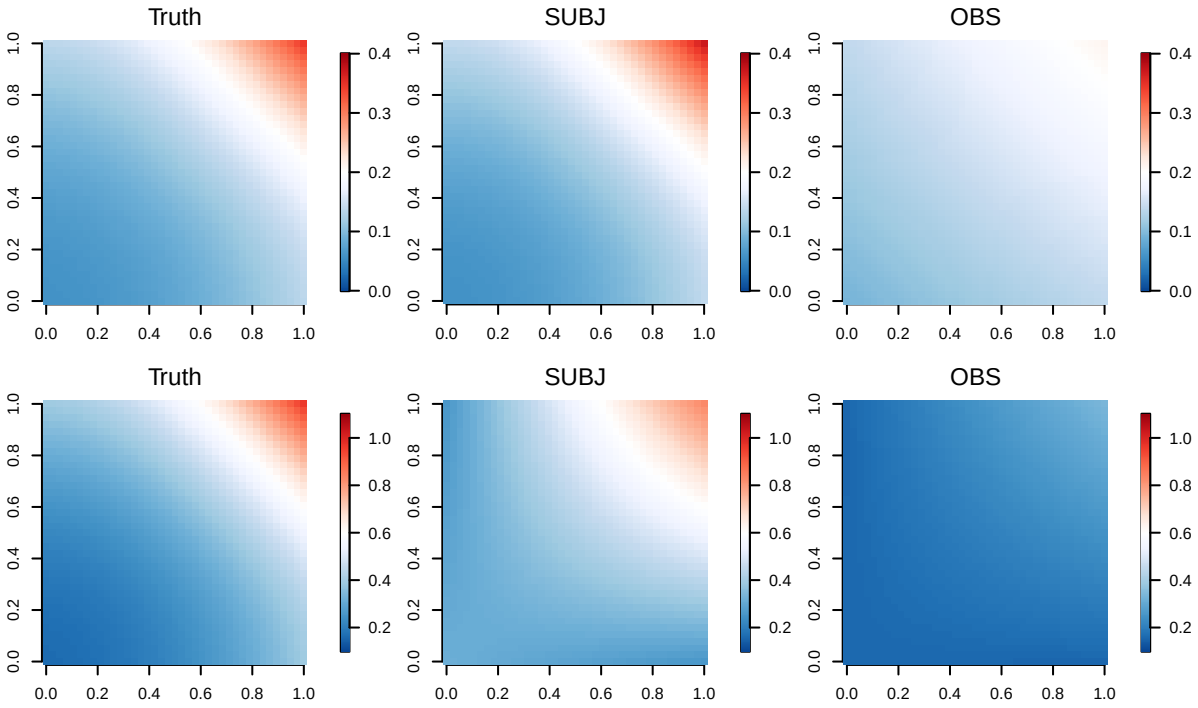

Figure 3.3 True covariance function and corresponding estimates using penalized splines with both OBS and SUBJ weights for two simulated data sets with $\rho = 0.5$ (top row) and $\rho = 0.85$ (bottom row).

covariance functions using either the SUBJ or OBS weights. Indeed, penalized splines with both types of weights perform similarly and both estimate accurately the mean function. The numerical results are given in Section B.4 of Appendix B.

Based on the above empirical results, we can recommend the use of the SUBJ weights for any longitudinal data with or without informative observations, without concern for sub-optimal estimation in the case that the observation process is non-informative.

3.5 Data Application

In this section, we apply the discussed methods to the Parkinson's disease symptom severity dataset described in Section 1. Recall that in this dataset, there was evidence of an informative observation process in which the average severity score of subjects with more than 10 observations is less than that of subjects with less than or equal to 10 observations. The intuition behind the selection of weights is clearly illustrated by this example. When weighting each observation equally, the estimates are more heavily influenced towards subjects with more observations. Thus, the mean function is likely to be underestimated. Similarly, the estimate of the covariance function is likely to be biased. Also recall that the time at which a patient entered the study is unrelated to the time since their initial diagnosis, there is no instance of dropout, and we only consider data observed less than 20 years following a patient's diagnosis. There is therefore no informative truncation or censoring present in the data. Therefore, it seems appropriate to use the proposed functional method with SUBJ weights.

The trajectories are plotted in Figure 3.4 along with the mean function estimates of the penalized spline estimator with the OBS and SUBJ weights. As expected, the mean estimate when using SUBJ weights is greater than the estimate when using OBS weights over the whole time domain. By the estimate from penalized splines with SUBJ weights, we see a monotonic increase in severity over the first 20 years after diagnosis, which agrees with existing research (Group, 2004; Venuto et al., 2016; Holden et al., 2018).

The estimated correlation and variance functions of the penalized spline estimator with SUBJ weights are displayed in Figure 3.4. It is interesting to see that the variance function takes its smallest value around 8 years after diagnosis and is greatest at the maximal followup time. Finally, we consider the spectral decomposition of the estimated covariance function. The first three eigenfunctions are shown in Figure 3.4. These first

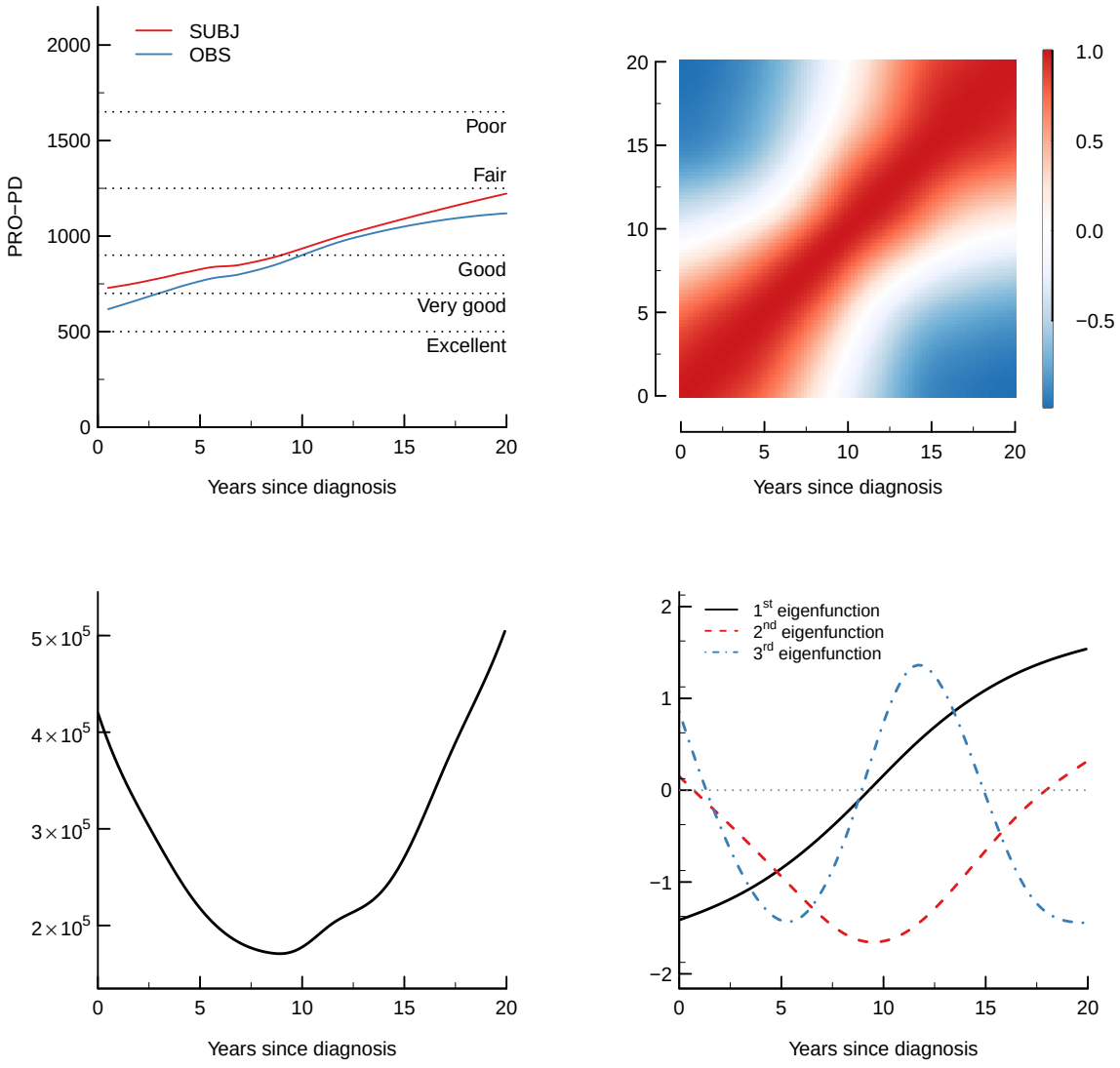

Figure 3.4 (Top left) Mean function estimates using both the SUBJ and OBS weights compared to average PRO-PD by reported quality of life. (Top right) Estimated correlation function using the SUBJ weights. (Bottom left) Estimated variance function using the SUBJ weights. (Bottom right) First three estimated eigenfunctions using the SUBJ weights.

three eigenfunctions account for 60%, 30%, and 6% of the total variation, respectively. The first eigenfunction corresponds to a contrast between early and late time after diagnosis. The second accounts for deviation from the mean trajectory around 8 years after diagnosis.

We conduct a multivariate longitudinal analysis of the PRO-PD dataset in Section C.2 of Appendix C.

3.6 Discussion

In this article, we studied some commonly used functional data methods for analysis of longitudinal data with informative observation times and identified a method with a proper subject-specific weighting scheme that can achieve the consistent estimation of the mean and covariance functions under a class of shared random effect models. The suggested method can be easily implemented using existing software. The theoretical properties of the estimator were studied under the setting with the informative observation process, and special care must be taken to handle the distribution of observation times for establishing the convergence rate results.

In our current work, no covariates are included in the model. But the proposed functional data method can be easily extended to accommodate covariates. In addition, censoring times are assumed to be independent of longitudinal outcomes and observation times, which we think is reasonable for our data application. It is certainly of interest to extend the proposed method to the setting with the informative observation process and informative dropout. This is a more challenging case that merits further investigation.

Chapter 4

Biclustering Multivariate Longitudinal Data with Application to Recovery Trajectories of White Matter After Sports-Related Concussion[1]

4.1 Introduction

In longitudinal studies, the goal of cluster analysis is to identify homogeneous subgroups of trajectories with similar characteristics. While all individuals may have unique patterns over time, the identification of groups subjects that have similar trajectories can reveal trends that may be generalized to larger populations. This kind of analysis has been applied in a wide range of application domains including psychology, medicine, and economics. Notable examples include the identification of patterns of childhood shyness

[1] The work detailed in this chapter was partially supported by the Office of the Assistant Secretary of Defense for Health Affairs through the Psychological Health and Traumatic Brain Injury Program under award W81XWH-14-2-0151 and by the NINDS award R01NS112303. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the DoD (DHP) funds or by the NIH.

that predict mental health problems in adulthood (Tang et al., 2017), the description of changes in adherence to treatment for obstructive sleep apnea (Babbin et al., 2015), and the classification of spatial–temporal patterns of neighborhood socioeconomic change (Delmelle, 2016).

Numerous methods of longitudinal cluster analysis have been developed. Some methods treat the trajectories as multivariate observations and disregard any temporal associations (Genolini & Falissard, 2010). Others describe groups of trajectories using mixed model parameterizations (Tan et al., 2012; Dziak et al., 2015). Often used are ad hoc methods such as the two-step procedure in which the components of a parametric model are first estimated, and subjects are then clustered based on their associated parameter estimates (Twisk & Hoekstra, 2012). A comparison of methods for longitudinal data clustering can be found in Den Teuling et al. (2020).

In contrast with traditional clustering methods, biclustering methods simultaneously group the samples and features of a data set. In doing so, subsets of samples that exhibit similar behaviors across subsets of features can be identified. In gene expression analysis, for example, bicluster analysis has been used to study breast cancer data in order to identify subgroups of subjects with distinct survival outcomes that could be distinguished by the expression levels of subgroups of genes (Sørlie et al., 2001).

Biclustering of multivariate data has received much attention over the past two decades, particularly in applied settings (Madeira & Oliveira, 2004; Tanay et al., 2005; Busygin et al., 2008). Some methods, like the plaid model proposed by Lazzeroni & Owen (2002), estimate bicluster membership as components of a parametric model of the observed data. Others, like those proposed by Kluger et al. (2003) and Lee et al. (2010), organize the observed data into a matrix indexed by samples in the rows and features in the columns, and apply matrix decomposition methods such as the singular value decomposition to uncover checkerboard-like structures in the matrix. Finally, some methods perform biclustering by solving restricted or penalized least squares problems. Two particular examples are the sparse biclustering method proposed by Tan & Witten (2014) and the convex biclustering algorithm (COBRA; Chi et al. (2017)). The latter, which solves a convex formulation of the biclustering problem for which a unique global solution exists. The authors use a generalization of the fused lasso (Tibshirani et al., 2005) to create a new sparsity penalty tailored to the biclustering problem.

In this chapter, we consider the extension of the biclustering problem to multivariate

longitudinal data. In the data set that motivated this study, four diffusion tensor imaging (DTI) measures are collected along each of 27 white matter tracts, including the genu and body of the corpus callosum, for college athletes at three time points following sports-related concussion (SRC). Among others, an important analysis of this data involves the identification of dynamics of progression in SRC cases. Identifying subgroups of subjects with similar progressions could lead to a better understanding of SRC recovery. We hypothesize that these subgroups may be distinguished by the measurements along certain subsets of white matter tracts. To address the novel problem of biclustering multivariate trajectories, we formulate a convex optimization problem, following Chi et al. (2017), in which the trends of trajectories are grouped into biclusters via the penalization of pairwise distances between subject and feature-specific model parameters. The model formulation lends guarantees of convergence and of uniqueness of the solution. The model is estimated via an iterative algorithm based on the alternating direction method of multipliers (ADMM; Boyd et al. (2004)). An R package is developed and made available at `https://github.com/clbwvr/longbc`.

The remainder of the chapter is organized as follows: Section 4.2 introduces the proposed model and our estimation methods. Section 4.3 assesses the numerical performance of our model with a simulation study. In Section 4.4, we perform an analysis of the motivating dataset. Finally, Section 4.5 concludes the chapter with a discussion.

4.2 Methods

4.2.1 Model Specification

Suppose response values are observed from p features for each of n subjects. Let y_{ijk} denote the observation of subject i $(i = 1, \ldots, n)$, feature j $(j = 1, \ldots, p)$ at the kth time point t_{ijk} $(k = 1, \ldots, m_{ij})$ such that there are $N = \sum_{i,j} m_{ij}$ total observations. We assume that the response values are translated such that they are zero-centered at time $t = 0$ within each feature and then scaled such that the magnitudes of the response values can be meaningfully compared. We consider a multivariate linear model

$$y_{ijk} = \alpha_{ij} + \beta_{ij} t_{ijk} + \epsilon_{ijk}, \tag{4.1}$$

where α_{ij} is a subject- and feature-specific intercept, β_{ij} is a subject- and feature-specific slope, and ϵ_{ijk} is an error term. To simplify notation, we will suppose the same number of observations m are made for each subject. Let $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ijm})^\top \in \mathbb{R}^m$, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ip})^\top \in \mathbb{R}^p$, $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^\top \in \mathbb{R}^p$, and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}, \ldots, \boldsymbol{\epsilon}_{ip})^\top \in \mathbb{R}^{pm}$. Then, let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_n^\top)^\top \in \mathbb{R}^{np}$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_n^\top)^\top \in \mathbb{R}^{np}$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_n^\top)^\top \in \mathbb{R}^N$.

To account for subject-specific heterogeneity, we consider $\boldsymbol{\alpha}_i$ to be random with $\boldsymbol{\alpha}_i \sim \mathrm{N}_p(\mathbf{0}, \mathbf{P})$ for $i = 1, \ldots, n$ such that $\boldsymbol{\alpha} \sim \mathrm{N}_{np}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{P})$. Here, the covariance matrix $\mathbf{P}$ is assumed to follow some specified correlation structure specific to the application. For example, we impose a compound symmetric structure on $\mathbf{P}$ in the simulation study. We assume that $\boldsymbol{\epsilon} \sim \mathrm{N}_N(\mathbf{0}, \sigma^2\mathbf{I})$ and that $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n$ are mutually independent. We consider $\boldsymbol{\beta}$ to be a vector of fixed effects. In this paper, we posit that the trend of the trajectory is of interest. Moreover, we assume that trajectories in the same bicluster share the same value of their slope coefficient and we will perform biclustering of the trajectories in order to explore the subgroup-specific heterogeneity in the trends.

Let $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijm})^\top \in \mathbb{R}^m$ be the vector of responses for subject i's jth feature and $\mathbf{t}_{ij} = (t_{ij1}, \ldots, t_{ijm})^\top \in \mathbb{R}^m$ be the vector of observation time points. Then, let $\mathbf{y}_i = (\mathbf{y}_{i1}^\top, \ldots, \mathbf{y}_{ip}^\top)^\top \in \mathbb{R}^{pm}$, $\mathbf{X}_i = \mathrm{blockdiag}(\mathbf{t}_{i1}, \ldots, \mathbf{t}_{ip}) \in \mathbb{R}^{pm \times p}$, and $\mathbf{Z}_i = \mathrm{blockdiag}(\mathbf{1}_m, \ldots, \mathbf{1}_m) \in \mathbb{R}^{pm \times p}$. Finally, let $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_n^\top)^\top \in \mathbb{R}^N$ be the vector of all observed responses, $\mathbf{X} = \mathrm{blockdiag}(\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{N \times np}$, and $\mathbf{Z} = \mathrm{blockdiag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_n) \in \mathbb{R}^{N \times np}$. Then, (4.1) can be written in matrix notation as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Equivalently, (4.1) can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$, where $\boldsymbol{\epsilon}^* \sim \mathrm{N}_N(0, \mathbf{V})$ with $\mathbf{V} = \mathbf{Z}(\mathbf{I}_n \otimes \mathbf{P})\mathbf{Z}^\top + \sigma^2\mathbf{I}_N$. Here, the parameters in specifying $\mathbf{P}$ and the error variance σ^2 will be denoted together by $\boldsymbol{\theta}$. Then the corresponding negative two log-likelihood is

$$-2\ell(\boldsymbol{\beta}, \mathbf{V}) = \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Based on the residuals of the ordinary least squares fit of model (4.1), we will compute the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ to use as a plug-in estimator for $\boldsymbol{\theta}$.

We propose to perform biclustering of the slope coefficients by minimizing the objective function

$$-2\ell(\boldsymbol{\beta}) + \lambda \left\{ \sum_{i<i'} w_{ii'} \left\| \begin{pmatrix} \beta_{i1} - \beta_{i'1} \\ \ldots \\ \beta_{ip} - \beta_{i'p} \end{pmatrix} \right\|_2 + \sum_{j<j'} \tilde{w}_{jj'} \left\| \begin{pmatrix} \beta_{1j} - \beta_{1j'} \\ \ldots \\ \beta_{nj} - \beta_{nj'} \end{pmatrix} \right\|_2 \right\}, \tag{4.2}$$

where $\lambda \geq 0$ is a tuning parameter and $w_{ii'}$ and $\tilde{w}_{jj'}$ are non-negative weights. The regularization term in (4.2) is the same as that proposed by Chi et al. (2017) for biclustering transposable data and is related to the ℓ_2-norm fusion lasso penalty used in convex clustering methods (Pelckmans et al., 2005; Hocking et al., 2011; Lindsten et al., 2011; Chi & Lange, 2015). The regularization term has two components. The first component penalizes the pairwise distances of the slope coefficient estimates between subjects. The second penalizes the pairwise distances between features. As the value of λ increases, the groups of slope coefficient estimates between pairs of subjects and pairs of features are shrunk towards each other.

By minimizing (4.2), we obtain coefficient estimates as well as bicluster detection. Specifically, to assign bicluster membership to the trajectories, we obtain two sets of vectors $\{\mathbf{u}_{ii'}\}_{i<i'}$ and $\{\mathbf{v}_{jj'}\}_{j<j'}$, where $\mathbf{u}_{ii'} = (\beta_{i1} - \beta_{i'1}, \ldots, \beta_{ip} - \beta_{i'p})^\top$ and $\mathbf{v}_{jj'} = (\beta_{1j} - \beta_{1j'}, \ldots, \beta_{nj} - \beta_{nj'})^\top$ measure coefficient group differences between subjects and features, respectively. When $\mathbf{u}_{ii'} = \mathbf{0}$ and $\mathbf{v}_{jj'} = \mathbf{0}$, we say that the trajectories of subject i's feature j and subject i''s feature j' belong to the same bicluster.

The minimization of (4.2) is a strictly convex problem. Hence, any algorithm that solves (4.2) will converge to a unique global minimizer (Boyd et al., 2004). This property ensures that results are reproducible, regardless of the initialization of the iterative algorithm, which is critical for cluster analysis in scientific research.

4.2.2 Model Estimation

The objective function (4.2) can be decomposed into two convex functions, namely, a negative loglikelihood function and a regularization term. Chi et al. (2017) proposed an algorithm for minimizing a convex function involving the same regularization term using the Dykstra-like proximal algorithm (DLPA; Bauschke & Combettes (2008)), which is used to solve best approximation problems. As (4.2) can not be written in the form of a best approximation problem as required by a DLPA, we will use the alternating direction method of multipliers (ADMM; Boyd et al. (2011)), which has become popular for its ability to solve optimization problems that can be decomposed into several convex problems with overlapping terms and constraint sets. We reformulate the minimization of

(4.2) as

$$\begin{aligned}\underset{\mathbf{a},\mathbf{b},\mathbf{c}}{\text{minimize}} \quad & -2\ell(\mathbf{a}) + \lambda \left\{ \sum_{i<i'} w_{ii'} \left\| \mathbf{b}_{ii'} \right\|_2 + \sum_{j<j'} \tilde{w}_{jj'} \left\| \mathbf{c}_{jj'} \right\|_2 \right\} \\ \text{s.t.} \quad & \mathbf{b}_{ii'} = \mathbf{D}_{ii'}\mathbf{a} \text{ for } i < i' \\ & \mathbf{c}_{jj'} = \widetilde{\mathbf{D}}_{jj'}\mathbf{a} \text{ for } j < j',\end{aligned}$$

where $\mathbf{D}_{ii'} \in \mathbb{R}^{p \times np}$ and $\widetilde{\mathbf{D}}_{jj'} \in \mathbb{R}^{n \times np}$ are differencing matrices such that the kth element of $\mathbf{D}_{ii'}\boldsymbol{\beta} = \beta_{ik} - \beta_{i'k}$ and the kth element of $\widetilde{\mathbf{D}}_{jj'}\boldsymbol{\beta} = \beta_{kj} - \beta_{kj'}$. This reformulation lends the following scaled-form ADMM updates:

$$\begin{aligned}\mathbf{a}^{(k+1)} &= \arg\min_{\mathbf{a}} -2\ell(\mathbf{a}) + \frac{\rho}{2} \sum_{i<i'} \left\| \mathbf{D}_{ii'}\mathbf{a} - \mathbf{b}_{ii'}^{(k)} + \boldsymbol{\delta}_{ii'}^{(k)} \right\|_2^2 + \frac{\rho}{2} \sum_{j<j'} \left\| \widetilde{\mathbf{D}}_{jj'}\mathbf{a} - \mathbf{c}_{jj'}^{(k)} + \boldsymbol{\eta}_{jj'}^{(k)} \right\|_2^2, \\ \mathbf{b}_{ii'}^{(k+1)} &= \arg\min_{\mathbf{b}} \lambda w_{ii'} \left\| \mathbf{b} \right\|_2 + \frac{\rho}{2} \left\| \mathbf{D}_{ii'}\mathbf{a}^{(k+1)} - \mathbf{b} + \boldsymbol{\delta}_{ii'}^{(k)} \right\|_2^2, \\ \mathbf{c}_{jj'}^{(k+1)} &= \arg\min_{\mathbf{c}} \lambda \tilde{w}_{jj'} \left\| \mathbf{c} \right\|_2 + \frac{\rho}{2} \left\| \widetilde{\mathbf{D}}_{jj'}\mathbf{a}^{(k+1)} - \mathbf{c} + \boldsymbol{\eta}_{jj'}^{(k)} \right\|_2^2, \\ \boldsymbol{\delta}_{ii'}^{(k+1)} &= \boldsymbol{\delta}_{ii'}^{(k)} + \mathbf{D}_{ii'}\mathbf{a}^{(k+1)} - \mathbf{b}_{ii'}^{(k+1)}, \\ \boldsymbol{\eta}_{jj'}^{(k+1)} &= \boldsymbol{\eta}_{jj'}^{(k)} + \widetilde{\mathbf{D}}_{jj'}\mathbf{a}^{(k+1)} - \mathbf{c}_{jj'}^{(k+1)}.\end{aligned}$$

Let $\boldsymbol{\Delta} = \sum_{i<i'} \mathbf{D}_{ii'}^\top \mathbf{D}_{ii'} + \sum_{j<j'} \widetilde{\mathbf{D}}_{jj'}^\top \widetilde{\mathbf{D}}_{jj'}$. Similarly, let $\mathbf{d}^{(k)} = \sum_{i<i'} \mathbf{D}_{ii'}^\top (\mathbf{b}_{ii'}^{(k)} - \boldsymbol{\delta}_{ii'}^{(k)}) + \sum_{j<j'} \widetilde{\mathbf{D}}_{jj'}^\top (\mathbf{c}_{jj'}^{(k)} - \boldsymbol{\eta}_{jj'}^{(k)})$. The update of $\mathbf{a}$ is solved by

$$\mathbf{a}^{(k+1)} = \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} + \rho \boldsymbol{\Delta} \right)^{-1} \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} + \rho \mathbf{d}^{(k)} \right).$$

The derivation is provided in Section C.1 of Appendix C. The update of $\mathbf{b}_{ii'}$ is solved by the proximal operator of the ℓ_2 norm. Letting $\tilde{\mathbf{b}}_{ii'}^{(k)} = \mathbf{D}_{ii'}\mathbf{a}^{(k+1)} + \boldsymbol{\delta}_{ii'}^{(k)}$,

$$\mathbf{b}_{ii'}^{(k+1)} = \max\left(0, 1 - \frac{\lambda w_{ii'}}{\rho} \left\| \tilde{\mathbf{b}}_{ii'}^{(k)} \right\|_2^{-1}\right) \tilde{\mathbf{b}}_{ii'}^{(k)}.$$

Similarly, letting $\tilde{\mathbf{c}}_{jj'}^{(k)} = \widetilde{\mathbf{D}}_{jj'}\mathbf{a}^{(k+1)} + \boldsymbol{\eta}_{jj'}^{(k)}$,

$$\mathbf{c}_{jj'}^{(k+1)} = \max\left(0, 1 - \frac{\lambda \tilde{w}_{jj'}}{\rho} \left\| \tilde{\mathbf{c}}_{jj'}^{(k)} \right\|_2^{-1}\right) \tilde{\mathbf{c}}_{jj'}^{(k)}.$$

The objective function (4.2) is minimized by iteratively performing these updates until convergence of the objective values. We set $\rho = 1$ as recommended by Boyd et al. (2011). The proposed algorithm is guaranteed convergence to the unique global minimizer (Boyd et al., 2011).

4.2.3 Implementation Details

In practice, we set the values of the weights $w_{ii'}$ and $w_{jj'}$ to be inversely proportional to $\left\|(\bar{\beta}_{i1} - \bar{\beta}_{i'1}, \ldots, \bar{\beta}_{ip} - \bar{\beta}_{i'p})^\top\right\|_2$ and $\left\|(\bar{\beta}_{1j} - \bar{\beta}_{1j'}, \ldots, \bar{\beta}_{nj} - \bar{\beta}_{nj'})^\top\right\|_2$, respectively, and scale such that $\sum_{i<i'} w_{ii'} = p^{-1/2}$ and $\sum_{j<j'} w_{jj'} = n^{-1/2}$. Here, $\bar{\beta}_{ii}$ indicates an initial OLS estimate of the coefficient β_{ii} in model (4.1). These weights more aggressively shrink coefficient groups between subjects and features that are more similar to each other.

We select the value of the tuning parameter λ from a sequence of grid points. Specifically, we select λ by minimizing $\mathrm{BIC}_\lambda = k \log N - 2\ell(\widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}}$ is the solution to the optimization problem (4.2) and k is the number of unique biclusters selected plus the number of estimated components in $\widehat{\boldsymbol{\theta}}$.

Following Meinshausen & Yu (2009) and Chi et al. (2017), we use a refinement of the thresholded lasso to group together subject and feature coefficient vectors that are almost but not exactly identical. Specifically, we hard-threshold the difference vectors $\mathbf{u}_{ii'}$ and $\mathbf{v}_{jj'}$ with parameters τ_u and τ_v. In the simulation and data analysis, we set τ_u and τ_v to be a 1/4 of the standard deviation of the values of $\{\|\mathbf{u}_{ii'}\|_2\}_{i<i'}$ and $\{\|\mathbf{v}_{jj'}\|_2\}_{j<j'}$, respectively.

4.3 Simulation Study

4.3.1 Simulation Settings

In this section, we conduct a simulation study to evaluate the performance of the proposed method. Responses are generated from model (4.1) with $n = 50$ or $n = 100$ and $p = 10$ or $p = 100$. We consider two settings for the observation times. In the fixed time point setting, we set the observation times for each subject at 0, 1, 2, and 3. In the random time point setting, we uniformly sample the number of observations m_i for subject i from the interval of 3 to 6. The values of the m_i observation times for subject i are then sampled without replacement from the integers from 0 to 5 with proportional probabilities

1, 0.8, 0.6, 0.4, 0.2, and 0.1, respectively. For example, if $m_i = 3$, two possible values of observation times are 0, 1, and 2, or 0, 1, and 5, with the former being more likely. This random time point setting mimics a longitudinal observation process that involves truncation and missingness.

We consider either $n_b = 9$ or $n_b = 16$ biclusters of slope coefficients. For each bicluster, the value of the coefficient is sampled uniformly from the integers between -10 and 10 without replacement. Each trajectory is then assigned randomly to one of the n_b biclusters. Each subject's vector of intercept coefficients is sampled from a zero-mean multivariate normal distribution with a compound symmetric correlation structure in which $\text{Cor}(\alpha_{ij}, \alpha_{ij'}) = 0.5$ when $j \neq j'$. We adjust the value of σ^2 to maintain one of two signal-to-noise ratios (SNR), SNR = 1 or SNR = 2, where SNR is defined as the variance of the expected response values divided by σ^2. Response values are centered such that the mean of each feature is zero at time $t = 0$. An example of one simulated dataset is illustrated in Figure 4.1.

We compare the proposed method to a two-step process in which the coefficients in model (4.1) are first estimated via linear mixed effect models using the R package *lme4* (Bates et al., 2014). The estimated slope coefficients are then arranged in a matrix indexed by subjects in the rows and features in the columns. Traditional biclustering is then applied to the matrix of coefficients, and the bicluster assignment of a trajectory is set to be the bicluster assignment of the coefficient. This procedure is analogous to the two-step procedures used in univariate longitudinal clustering as described in Section 1. We will use three methods of biclustering in the two-step procedure: (1) the clustered dendogram, which performs separate hierarchical clustering of the subjects and of the features; (2) the COBRA (Chi et al., 2017) implemented in the R package *cvxbiclustr*; and (3) the sparse biclustering algorithm (Tan & Witten, 2014) implemented in the R package *sparseBC*. Hard biclustering assignments are made for the clustered dendrogram using the dynamic tree cutting (DTC) algorithm (Langfelder et al., 2008) implemented in the R package *dynamicTreeCut*. We choose to use these algorithms in the two-step approach because their authors provide data-adaptive ways to control the number of detected biclusters. Many other biclustering algorithms require the number of biclusters as an input parameter, which can be prohibitive in practice. Our proposed methodology does not require the number of biclusters to be determined a priori and its simulation results are therefore directly comparable to those of the other methods. In carrying out

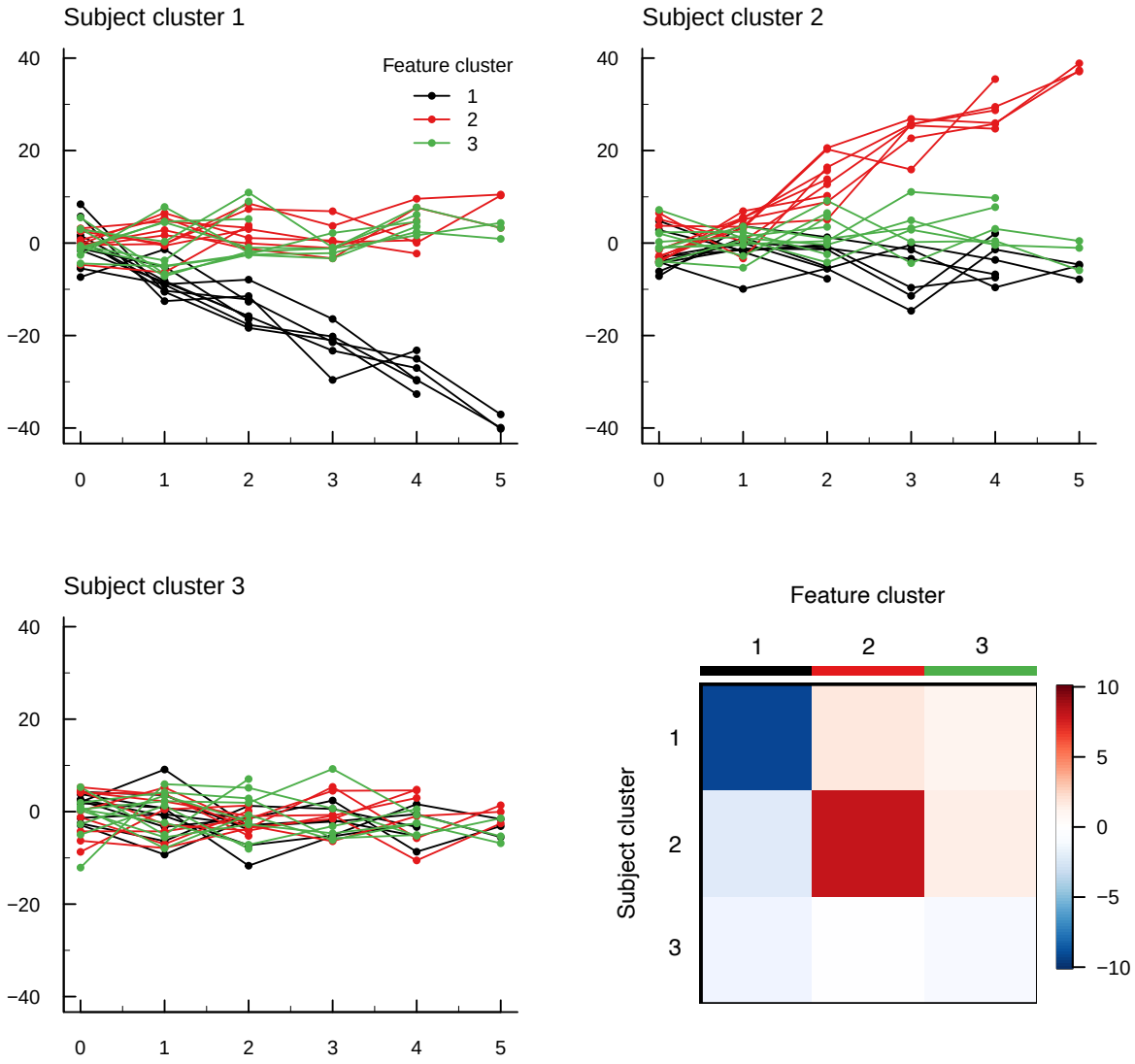

Figure 4.1 Trajectories of 15 features across 15 subjects from one simulated dataset with 9 biclusters and a SNR of 2. Sample clusters are indicated by panel and each include 5 subjects. Feature clusters are indicated by color and each include 5 features. (Bottom right) True slope coefficients arranged according to the true underlying biclustering structure with subjects represented by rows and features represented by columns.

the simulation study, we found that modifications of the model selection strategies for the sparse biclustering method and the COBRA led to better performance in some settings and worse performance in others. Thus, the tuning parameters were selected according to the default model selection strategies in the corresponding software packages, but some fine-tuning of the algorithms' parameter selection may lead to better performances.

4.3.2 Simulation Results

The adjusted Rand index (ARI; Hubert & Arabie (1985)) is calculated between the true biclusters and those obtained by each biclustering approach. The Rand index measures the similarity between two clusterings of a data set by considering all pairs of data points and calculating the proportion of agreements between the true and predicted clusterings. The ARI adjusts the Rand index by assuming a hypergeometric distribution for the null case in which two random clusterings are being compared and scales the Rand index such that expected value of the adjusted Rand index for a random clustering is 0. When there is perfect agreement between the predicted and true clusterings, the value of ARI is 1.

Table 4.1 gives the average and standard deviation of ARI over 50 replicates of the simulation settings. In the random observation time settings, the proposed method resulted in higher mean ARI values in most cases. In the fixed observation time settings, the sparse biclustering approach and the proposed method have comparable mean ARI, but higher mean ARI than the COBRA and DTC. While these results may seem to imply that the proposed method is only useful for settings in which there are missing or sparse observation times, recall that one of the advantages of the proposed method that, unlike the sparse biclustering algorithm, its convex formulation possesses a unique solution and thus its results are reproducible. Hence, the proposed methodology may be preferable to alternative approaches that give similar bicluster detection performance.

4.4 Data Application

In this section, we apply the proposed methods to data collected in a sub-study of the NCAA-DoD Concussion Assessment, Research and Education (CARE) Consortium. This data was analyzed by Wu et al. (2020) to study the recovery trajectories of white matter after sports-related concussion in collegiate athletes. We concentrate on the diffusion-weighted magnetic resonance imaging data analyzed by DTI methods obtained from

Random time points							
n	p	n_b	SNR	DTC	Sparse BC	COBRA	Long. BC
50	10	9	1	0.35 (0.08)	0.76 (0.22)	0.70 (0.21)	**0.82** (0.14)
			2	0.33 (0.10)	0.84 (0.14)	0.76 (0.19)	**0.88** (0.17)
		16	1	0.45 (0.08)	0.79 (0.18)	0.74 (0.14)	**0.84** (0.20)
			2	0.43 (0.09)	0.84 (0.16)	0.80 (0.23)	**0.89** (0.28)
	100	9	1	0.36 (0.06)	0.86 (0.19)	0.90 (0.20)	**0.95** (0.14)
			2	0.37 (0.10)	0.86 (0.17)	0.98 (0.03)	**1.00** (0.00)
		16	1	0.45 (0.05)	0.92 (0.14)	0.97 (0.09)	**0.99** (0.05)
			2	0.46 (0.08)	0.90 (0.14)	0.99 (0.06)	**1.00** (0.00)
100	10	9	1	0.46 (0.16)	0.89 (0.12)	0.77 (0.18)	**0.91** (0.14)
			2	0.46 (0.15)	**1.00** (0.00)	0.96 (0.07)	**1.00** (0.00)
		16	1	0.61 (0.17)	0.91 (0.16)	0.90 (0.17)	**0.94** (0.09)
			2	0.58 (0.13)	0.98 (0.07)	0.95 (0.11)	**1.00** (0.00)
	100	9	1	0.42 (0.12)	0.98 (0.02)	0.97 (0.01)	**0.99** (0.01)
			2	0.40 (0.08)	0.97 (0.08)	**1.00** (0.00)	**1.00** (0.00)
		16	1	0.58 (0.13)	**1.00** (0.00)	0.98 (0.03)	**1.00** (0.00)
			2	0.57 (0.16)	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)
Fixed time points							
n	p	n_b	SNR	DTC	Sparse BC	COBRA	Long. BC
50	10	9	1	0.12 (0.04)	0.66 (0.18)	0.55 (0.27)	**0.71** (0.13)
			2	0.13 (0.04)	**0.79** (0.21)	0.64 (0.23)	0.77 (0.18)
		16	1	0.21 (0.06)	**0.81** (0.21)	0.68 (0.24)	0.79 (0.17)
			2	0.24 (0.08)	0.87 (0.21)	0.79 (0.22)	**0.89** (0.32)
	100	9	1	0.13 (0.04)	**0.86** (0.16)	0.84 (0.10)	0.85 (0.11)
			2	0.14 (0.04)	0.84 (0.17)	0.84 (0.11)	**0.90** (0.10)
		16	1	0.22 (0.05)	**0.86** (0.18)	0.88 (0.20)	0.83 (0.12)
			2	0.25 (0.09)	0.90 (0.21)	0.93 (0.12)	**0.95** (0.14)
100	10	9	1	0.19 (0.08)	0.90 (0.04)	0.84 (0.15)	**0.94** (0.11)
			2	0.19 (0.09)	**0.97** (0.08)	0.92 (0.11)	0.94 (0.20)
		16	1	0.27 (0.14)	0.94 (0.12)	0.86 (0.14)	**0.96** (0.11)
			2	0.33 (0.11)	0.96 (0.11)	0.87 (0.20)	**1.00** (0.00)
	100	9	1	0.19 (0.09)	**1.00** (0.00)	0.98 (0.06)	**1.00** (0.00)
			2	0.18 (0.08)	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)
		16	1	0.33 (0.13)	**0.95** (0.06)	0.94 (0.06)	**0.95** (0.05)
			2	0.32 (0.13)	**1.00** (0.00)	**1.00** (0.00)	**1.00** (0.00)

Table 4.1 Mean adjusted Rand index (standard deviation) between the true biclusters and those obtained by each approach in the simulation settings. Numbers in boldface indicate the greatest mean value of each row.

football players from three US colleges.

In methods of DTI, the diffusion of water molecules in the brain is described at each voxel by a 3×3 symmetric, positive-definite matrix called the diffusion tensor, whose eigenvectors and eigenvalues λ_1, λ_2, and λ_3 describe the directions and diffusivities along the principal axes of diffusion. Mean diffusivity (MD), which describes the amount of diffusion irrespective of direction, is the equal to the average of the three eigenvalues, that is, $\text{MD} = (\lambda_1 + \lambda_2 + \lambda_3)/3$. Axial diffusivity (AD), which describes the amount of diffusion parallel to fiber tracts, is equal to the largest eigenvalue, that is, $\text{AD} = \lambda_1$. Radial diffusivity (RD), which describes the amount of diffusion perpendicular to fiber tracts, is equal to the average of the smaller two eigenvalues, that is, $\text{RD} = (\lambda_1 + \lambda_2)/2$. Lastly, fractional anisotropy (FA), which describes the directional restriction of the diffusion, is equal to the normalized variance of the three eigenvalues, that is,

$$\text{FA} = \sqrt{\frac{3}{2}} \frac{\sqrt{(\lambda_1 - \bar{\lambda})^2 + (\lambda_1 - \bar{\lambda})^2 + (\lambda_3 - \bar{\lambda})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}},$$

with $\bar{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$ being the average values of the eigenvalues. FA can be interpretted as the normalized variance of the eigenvalues and ranges from 0 to 1. In the motivating dataset, these four measurements are collected along 27 white matter tracts. As FA has a negative correlation with the other three measures, we use the values of 1-FA instead of FA for the sake of cluster analysis.

Measurements are summarized by quantile per tract for each subject. While tract-specific analyses typically involve averaging measurements over voxels, we found upper quantile values to better distinguish between concussed athletes and controls in the motivating dataset. Previous DTI studies have used similar approaches to summarizing measurements over voxels to assess relationships between clinical responses and DTI measurements (Bazarian et al., 2007; Bazarian et al., 2013; Pietrosanu et al., 2021). We use the 90th quantile, but our results are similar when using the 80th, 90th, and 95th quantiles.

Observations are made at four time points: (1) 24 to 48 hours post-injury, (2) the point at which the concussed athlete became asymptomatic, (3) 7 days after unrestricted return-to-play, and (4) 6 months post-injury. We examine the acute effects of SRC on white matter properties, which occur within 90 days after injury (McCrea et al., 2003), by only considering the first three time points. Further, we only consider athletes with

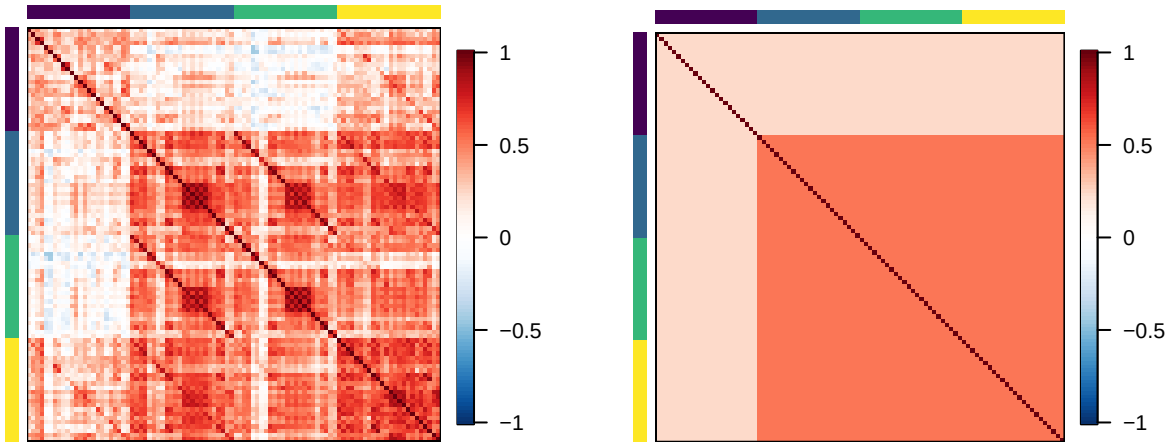

Figure 4.2 (Left) Empirical correlation matrix of the 108 features ordered by measure. Measurement is indicated in the left and upper color bars (AD: purple, RD: blue, 1-FA: green, MD: yellow). (Right) Estimated correlation matrix of the random intercepts in the fitted model based on imposed structure.

at least one observation at the second and third time points. This results in $p = 108$ longitudinal trajectories associated with each of $n = 60$ players (33 concussed athletes and 27 controls) observed at three time points. The median (interquartile range) number of days from injury to the second and third time points is 10 (6) and 26 (12.5), respectively. As we anticipate heterogeneity between the concussed athletes and the controls to be greatest immediately post-injury, we represent the third time point as $t = 0$ and represent the first and second time points as $t = -2$ and $t = -1$, respectively. The trajectories are scaled within each feature such that magnitudes can be compared.

Based on an inspection of the empirical correlation matrix of the features illustrated in Figure 4.2, we assume the covariance structure of the random intercept coefficients to be such that $\text{Cov}(\alpha_{ij}, \alpha_{ik}) = \theta_0[I(j = k) + I(j \neq k)\{\theta_1 + \iota_{jk}\theta_2\}]$, where $\iota_{jkl} = 1$ when either of features j and l are of AD features and 0 otherwise. Under this parameterization, θ_1 may be interpreted as the correlation between pairs of features not involving measures of AD, and θ_2 may be interpreted as the difference in correlation between features involving measures of AD. The estimated correlation matrix based on the imposed structure is shown in Figure 4.2. The values of the estimated correlation components are $\widehat{\theta}_1 = 0.52$ and $\widehat{\theta}_2 = -0.28$, indicating a lower correlation between features involving measures of AD.

The coefficients and detected biclustering structure estimated by the proposed method are illustrated in Figure 4.3. Three subgroups of subjects are apparent. The first subgroup (n=11, 8 SRC) consists primarily of concussed subjects, and exhibits decreasing MD, AD, and RD, and increasing FA. Elevated MD, AD, and RD and decreased FA have been shown to be associated with SRC (Chamard et al., 2013; Koerte et al., 2012; McAllister et al., 2014; Mustafi et al., 2018), and thus the average trend of the subgroup of concussed subjects in this cluster, denoted "SRC-1", may be considered as demonstrating a trajectory of recovery. The second subgroup (n=12, 9 SRC) also consists primarily of concussed subjects, and exhibits increasing MD, AD, and RD, and decreasing FA. We denote the subgroup of concussed subjects in this cluster by "SRC-2". The third subgroup (n=37, 16 SRC) exhibits no changes in white matter from time of injury to 7 days after unrestricted return-to-play. Wu et al. (2020) demonstrated that effects of SRC on white-matter differences persist beyond the point when the concussed athletes become asymptomatic. Our analysis supports this finding and identifies homogeneous subgroups defined by the persistency of these effects. The average trajectories by feature cluster are illustrated in Figure 4.4. Wu et al. (2020) also demonstrated significantly higher MD specifically in the genu and body of the corpus callosum of concussed athletes. It is interesting, therefore, that the feature cluster that demonstrates the greatest rate of change primarily consists of MD features and includes measurements made along the corpus collusum.

Next, we evaluate the differences between the two subgroups of concussed subjects using the clinical assessments followed by the CARE Consortium study. This collection of clinical outcome measures included the Standardized Assessment of Concussion (SAC; (McCrea et al., 1997)), the Sports Concussion Assessment Tool (SCAT; McCrory et al. (2017)), the Balance Glossary (BESS; Guskiewicz et al. (2001)), and the Brief Symptom Inventory (BSI; Derogatis & Spencer (1993)). Figure 4.5 illustrates the longitudinal changes in the SCAT total severity score, number of SCAT symptoms, BSI, each of the three subcategories of BSI (BSI-Soma, BSI-Anxiety, and BSI-Depression), BESS, and SAC.

We used generalized linear mixed models to compare scores between the subgroups. Specifically, we modeled the trajectories of scores as a function of subgroup and time point, and contrasts between subgroups at each time point were estimated. We considered a p-value of $p < 0.05$ to be significant after applying Tukey's method for multiple comparisons.

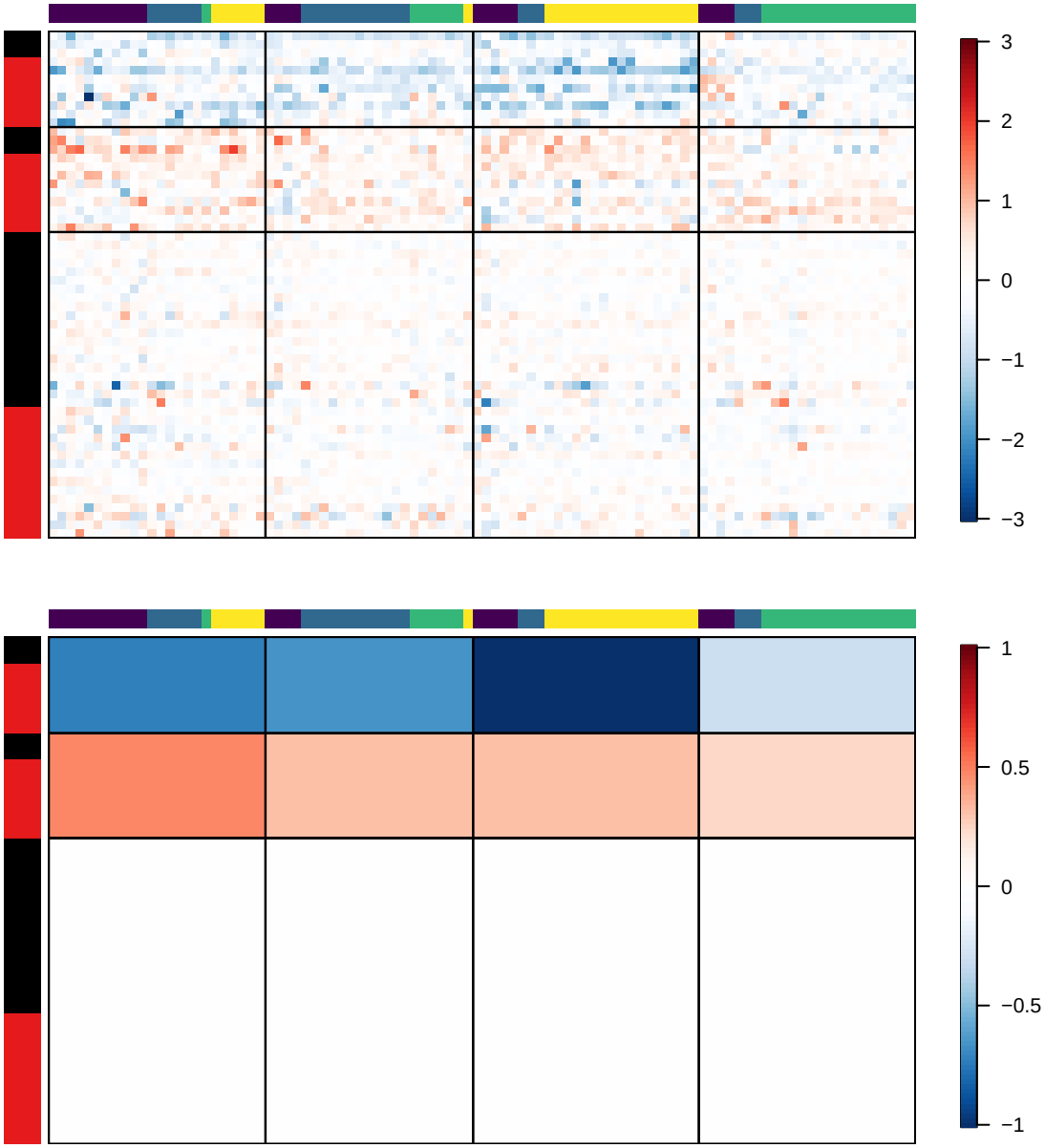

Figure 4.3 (Top) Estimated slope coefficients of the acute effect trajectories arranged according to the estimated biclustering structure. Subjects are represented by rows, and features are represented by columns. SRC status is indicated in the left color bar (SRC: red, Non-SRC: black). Measurement is indicated in the left color bar (AD: purple, RD: blue, 1-FA: green, MD: yellow). The estimated biclustering structure is indicated with grid lines. (Bottom) Estimated mean matrix of the slope coefficients of the acute effect trajectories.

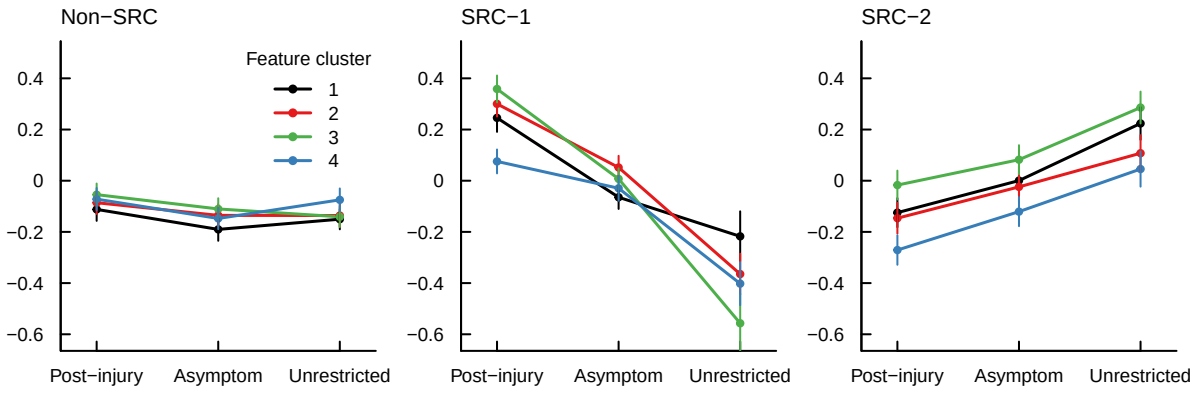

Figure 4.4 Mean quantile trajectories and standard error bars of the four feature subgroups for the control subjects (Non-SRC) and two subgroups of SRC subjects (SRC-1 and SRC-2).

As expected, both subgroups of SRC cases report significantly worse clincal assessments than the Non-SRC group. There are no significant differences in SCAT total severity and number of SCAT symptoms between the SRC-1 and SRC-2 subgroups. The SRC-1 subgroup had a lower average SAC score and greater average BSI and BESS scores post-injury. At the third time point, however, there are no significant differences between the SRC subgroups and the Non-SRC subgroup.

For comparison, we applied the two-step approach discussed in Section 4.2 for analyzing the DTI data using the sparse biclustering algorithm. The results included fewer detected biclusters. Specifically, there were two subject groups detected, one of which contained primarily SRC subjects with decreasing AD, RD, and MD, and increasing FA; in point, this approach failed to detect any subgrouping behavior of the SRC subjects.

To further demonstrate the utility of the proposed model, we perform a multivariate longitudinal biclustering analysis of the symptoms of Parkinson's disease data set discussed in Chapter 3. The results are provided in Section C.2 of Appendix C.

4.5 Discussion

In this chapter, we have introduced the task of bicluster analysis in multivariate longitudinal data and proposed a method for performing this analysis. The proposed method performs coefficient estimation and bicluster assignment simultaneously. We found in simulation studies that the proposed model has better performance in terms of cluster

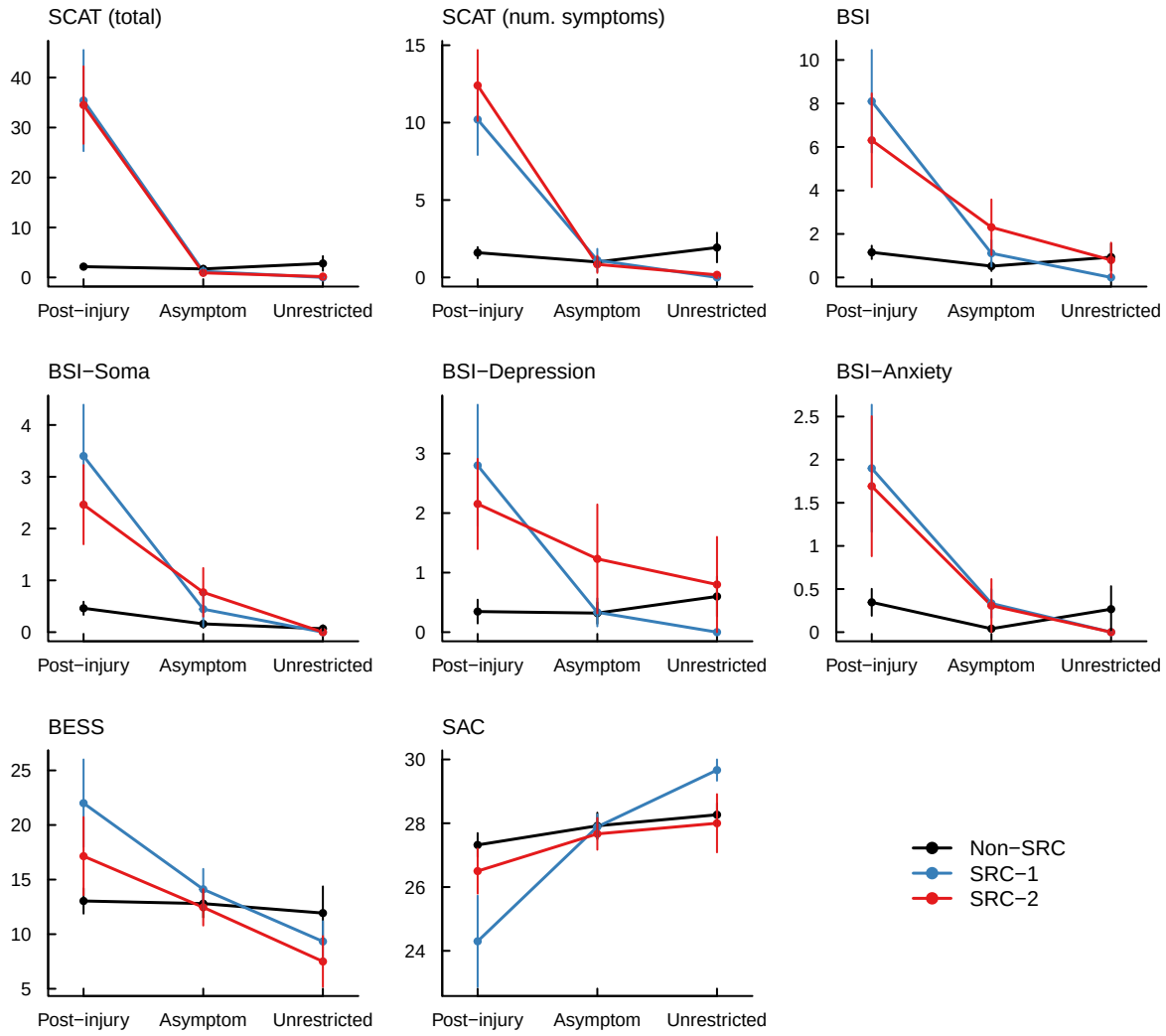

Figure 4.5 Mean and standard error bars for the SCAT total severity score, number of SCAT symptoms, BSI, each of the three subcategories of BSI (BSI-Soma, BSI-Anxiety, and BSI-Depression), BESS, and SAC for the control subjects (Non-SRC) and the two subgroups of SRC subjects (SRC-1 and SRC-2).

assignment and coefficient estimation when compared to more naive approaches. Through application of the proposed methodology to the motivating dataset, we identified subgroups of SRC cases with similar trajectories distinguished by the longitudinal patterns of measures along subsets of white matter tracts.

Our proposed algorithm minimizes a convex function and is guaranteed a unique solution, a property not often available to clustering algorithms. Further, the proposed methodology uses a single tuning parameter that is adaptive to the observed data and it does not require the number of biclusters to detect as an input parameter. These attributes, along with its performance in simulation studies, give the proposed methodology a substantial advantage to alternative approaches for use in scientific research. To our knowledge, bicluster analysis has not previously been applied to multivariate longitudinal data. Here we have performed one such data analysis. We anticipate that many other data sets may be analyzed similarly to produce scientifically interesting results.

Our model makes a few simplifying assumptions on the biclustering structure present in the data. First, we assume that each trajectory belongs to a single bicluster. However, some biclustering methods allow for overlapping groups, that is, observations can belong to more than one bicluster (Cheng & Church, 2000; Lazzeroni & Owen, 2002). Second, we assume that the parameter of interest for identifying subgroups of longitudinal trajectories is their linear trend. Our method can be extended to more complicated parametric models by incorporating the coefficients of higher order terms in the regularization term. However, while nonparametric regression is often useful for modeling more complex patterns in longitudinal data, there is no direct extension of our model to incorporate a nonparametric form of the relationship between response variables and time. The relaxing of these model assumptions merits further investigation.

REFERENCES

Babbin, S. F. et al. (2015). "Identifying longitudinal patterns for individuals and subgroups: an example with adherence to treatment for obstructive sleep apnea". *Multivariate behavioral research* **50**.1, pp. 91–108.

Bates, D. et al. (2014). "Fitting linear mixed-effects models using lme4". *arXiv preprint arXiv:1406.5823*.

Bauschke, H. H. & Combettes, P. L. (2008). "A Dykstra-like algorithm for two monotone operators". *Pacific Journal of Optimization* **4**.3, pp. 383–391.

Bazarian, J. J. et al. (2007). "Diffusion tensor imaging detects clinically important axonal damage after mild traumatic brain injury: a pilot study". *Journal of neurotrauma* **24**.9, pp. 1447–1459.

Bazarian, J. J. et al. (2013). "The relation between posttraumatic stress disorder and mild traumatic brain injury acquired during Operations Enduring Freedom and Iraqi Freedom". *The journal of head trauma rehabilitation* **28**.1, pp. 1–12.

Beck, A. et al. (1996). *BDI-II, Beck Depression Inventory: Manual*. Gutman Reserve Test Collection. Psychological Corporation.

Bien, J. & Tibshirani, R. J. (2011). "Sparse estimation of a covariance matrix". *Biometrika* **98**.4, pp. 807–820.

Bilker, W. B. et al. (2012). "Development of abbreviated nine-item forms of the Raven's standard progressive matrices test". *Assessment* **19**.3, pp. 354–369.

Boor, C. de (1978). *A Practical Guide to Splines*. Vol. Volume 27.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*. Lecture Notes in Statistics. Springer New York.

Boyd, S. et al. (2004). *Convex optimization*. Cambridge university press.

Boyd, S. et al. (2011). "Distributed optimization and statistical learning via the alternating direction method of multipliers". *Foundations and Trends® in Machine learning* **3**.1, pp. 1–122.

Bronner, G. et al. (2004). "Sexual dysfunction in Parkinson's disease". *Journal of sex & marital therapy* **30**.2, pp. 95–105.

Busygin, S. et al. (2008). "Biclustering in data mining". *Computers & Operations Research* **35**.9, pp. 2964–2987.

Cai, N. et al. (2012). "Time-varying latent effect model for longitudinal data with informative observation times". *Biometrics* **68**.4, pp. 1093–1102.

Cai, T. T. & Yuan, M. (2011). "Optimal estimation of the mean function based on discretely sampled functional data: Phase transition". *The annals of statistics* **39**.5, pp. 2330–2355.

Carroll, R. J. et al. (1997). "Generalized partially linear single-index models". *Journal of the American Statistical Association* **92**.438, pp. 477–489.

Cella, D. et al. (2010). "The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008". *Journal of clinical epidemiology* **63**.11, pp. 1179–1194.

Chamard, E. et al. (2013). "Neurometabolic and microstructural alterations following a sports-related concussion in female athletes". *Brain injury* **27**.9, pp. 1038–1046.

Cheng, Y. & Church, G. M. (2000). "Biclustering of expression data." *Ismb*. Vol. 8. 2000, pp. 93–103.

Chi, E. C. & Lange, K. (2015). "Splitting methods for convex clustering". *Journal of Computational and Graphical Statistics* **24**.4, pp. 994–1013.

Chi, E. C. et al. (2017). "Convex biclustering". *Biometrics* **73**.1, pp. 10–19.

Claeskens, G. et al. (2009). "Asymptotic properties of penalized spline estimators". *Biometrika* **96**.3, pp. 529–544.

Colclough, G. L. et al. (2018). "Multi-subject hierarchical inverse covariance modelling improves estimation of functional brain networks". *NeuroImage* **178**, pp. 370–384.

Csardi, G., Nepusz, T., et al. (2006). "The igraph software package for complex network research". *InterJournal, complex systems* **1695**.5, pp. 1–9.

Dadi, K. et al. (2019). "Benchmarking functional connectome-based predictive models for resting-state fMRI". *Neuroimage* **192**, pp. 115–134.

Danaher, P. et al. (2014). "The joint graphical lasso for inverse covariance estimation across multiple classes". *Journal of the Royal Statistical Society. Series B, Statistical methodology* **76**.2, p. 373.

Delmelle, E. C. (2016). "Mapping the DNA of urban neighborhoods: Clustering longitudinal sequences of neighborhood socioeconomic change". *Annals of the American Association of Geographers* **106**.1, pp. 36–56.

Den Teuling, N. et al. (2020). "A comparison of methods for clustering longitudinal data with slowly changing trends". *Communications in Statistics-Simulation and Computation*, pp. 1–28.

Derogatis, L. R. & Spencer, P. (1993). *Brief symptom inventory: BSI*. Vol. 18. Pearson Upper Saddle River, NJ.

Diggle, P. et al. (2002). *Analysis of Longitudinal Data*. Oxford Statistical Science Series. OUP Oxford.

Dziak, J. J. et al. (2015). "Modeling intensive longitudinal data with mixtures of nonparametric trajectories and time-varying effects." *Psychological methods* **20**.4, p. 444.

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

Eilers, P. H. & Marx, B. D. (1996). "Flexible smoothing with B-splines and penalties". *Statistical science*, pp. 89–102.

Engle, R. F. et al. (1986). "Semiparametric estimates of the relation between weather and electricity sales". *Journal of the American statistical Association* **81**.394, pp. 310–320.

Fan, J. (1992). "Design-adaptive Nonparametric Regression". *Journal of the American Statistical Association* **87**.420, pp. 998–1004.

Fang, S. et al. (2016). "Joint analysis of longitudinal data with additive mixed effect model for informative observation times". *Journal of Statistical Planning and Inference* **169**, pp. 43–55.

Feng, Y. et al. (2020). "Sparse Single Index Models for Multivariate Responses". *Journal of Computational and Graphical Statistics*, pp. 1–10.

Finn, E. S. et al. (2015). "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity". *Nature neuroscience* **18**.11, pp. 1664–1671.

Fitzmaurice, G. et al. (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Friedman, J. et al. (2008). "Sparse inverse covariance estimation with the graphical lasso". *Biostatistics* **9**.3, pp. 432–441.

Genolini, C. & Falissard, B. (2010). "KmL: k-means for longitudinal data". *Computational Statistics* **25**.2, pp. 317–328.

Glasser, M. F. et al. (2013). "The minimal preprocessing pipelines for the Human Connectome Project". *Neuroimage* **80**, pp. 105–124.

Greicius, M. D. et al. (2004). "Default-mode network activity distinguishes Alzheimer's disease from healthy aging: evidence from functional MRI". *Proceedings of the National Academy of Sciences* **101**.13, pp. 4637–4642.

Griffanti, L. et al. (2014). "ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging". *Neuroimage* **95**, pp. 232–247.

Grosenick, L. et al. (2013). "Interpretable whole-brain prediction analysis with GraphNet". *NeuroImage* **72**, pp. 304–321.

Group, P. S. (2004). "Levodopa and the progression of Parkinson's disease". *New England Journal of Medicine* **351**.24, pp. 2498–2508.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer Series in Statistics. Springer.

Guha, S. & Rodriguez, A. (2020). "Bayesian regression with undirected network predictors with an application to brain connectome data". *Journal of the American Statistical Association* just-accepted, pp. 1–34.

Guo, W. (2002). "Functional mixed effects models". *Biometrics* **58**.1, pp. 121–128.

Guskiewicz, K. M. et al. (2001). "Postural stability and neuropsychological deficits after concussion in collegiate athletes". *Journal of athletic training* **36**.3, p. 263.

Hall, P. et al. (2006). "Properties of principal component methods for functional and longitudinal data analysis". *The annals of statistics*, pp. 1493–1517.

Härdle, W. & Stoker, T. M. (1989). "Investigating smooth multiple regression by the method of average derivatives". *Journal of the American statistical Association* **84**.408, pp. 986–995.

Harezlak, J. et al. (2018). *Semiparametric Regression with R*. Use R! Springer New York.

Hastie, T et al. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. Springer.

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

Hearne, L. J. et al. (2016). "Functional brain networks related to individual differences in human intelligence at rest". *Scientific reports* **6**, p. 32328.

Hocking, T. D. et al. (2011). "Clusterpath an algorithm for clustering using convex fusion penalties". *28th international conference on machine learning*, p. 1.

Hoerl, A. E. & Kennard, R. W. (1970). "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* **12**.1, pp. 55–67.

Holden, S. K. et al. (2018). "Progression of MDS-UPDRS scores over five years in de novo Parkinson disease from the Parkinson's progression markers initiative cohort". *Movement disorders clinical practice* **5**.1, pp. 47–53.

Huang, J. Z. & Su, Y. (2021). "Asymptotic Properties of Penalized Spline Estimators in Concave Extended Linear Models: Rates of Convergence". *arXiv preprint arXiv:2105.06367*.

Hubert, L. & Arabie, P. (1985). "Comparing partitions". *Journal of classification* **2**.1, pp. 193–218.

Ichimura, H. (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models". *Journal of Econometrics* **58**.1-2, pp. 71–120.

Jacob, L. et al. (2009). "Group lasso with overlap and graph lasso". *Proceedings of the 26th annual international conference on machine learning*, pp. 433–440.

Kluger, Y. et al. (2003). "Spectral biclustering of microarray data: coclustering genes and conditions". *Genome research* **13**.4, pp. 703–716.

Koerte, I. K. et al. (2012). "A prospective study of physician-observed concussion during a varsity university hockey season: white matter integrity in ice hockey players. Part 3 of 4". *Neurosurgical focus* **33**.6, E3.

Langfelder, P. et al. (2008). "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". *Bioinformatics* **24**.5, pp. 719–720.

Lazzeroni, L. & Owen, A. (2002). "Plaid models for gene expression data". *Statistica sinica*, pp. 61–86.

Ledoit, O. & Wolf, M. (2004). "Honey, I shrunk the sample covariance matrix". *The Journal of Portfolio Management* **30**.4, pp. 110–119.

Lee, M. et al. (2010). "Biclustering via sparse singular value decomposition". *Biometrics* **66**.4, pp. 1087–1095.

Li, Y. & Hsing, T. (2010). "Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data". *The Annals of Statistics* **38**.6, pp. 3321–3351.

Liang, X. et al. (2016). "A novel joint sparse partial correlation method for estimating group functional networks". *Human brain mapping* **37**.3, pp. 1162–1177.

Liang, Y. et al. (2009). "Joint modeling and analysis of longitudinal data with informative observation times". *Biometrics* **65**.2, pp. 377–384.

Lin, H. et al. (2004). "Analysis of longitudinal data with irregular, outcome-dependent follow-up". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**.3, pp. 791–813.

Lindsten, F. et al. (2011). *Just relax and come clustering!: A convexification of k-means clustering*. Linköping University Electronic Press.

Liu, W. & Luo, X. (2012). "High-dimensional sparse precision matrix estimation via sparse column inverse operator". *arXiv preprint arXiv:1203.3896* **38**.

Lynall, M.-E. et al. (2010). "Functional connectivity and brain networks in schizophrenia". *Journal of Neuroscience* **30**.28, pp. 9477–9487.

Madeira, S. C. & Oliveira, A. L. (2004). "Biclustering algorithms for biological data analysis: a survey". *IEEE/ACM transactions on computational biology and bioinformatics* **1**.1, pp. 24–45.

Marrelec, G. et al. (2006). "Partial correlation for functional brain interactivity investigation in functional MRI". *Neuroimage* **32**.1, pp. 228–237.

Mazumder, R. & Hastie, T. (2012a). "The graphical lasso: New insights and alternatives". *Electronic journal of statistics* **6**, p. 2125.

Mazumder, R. & Hastie, T. (2012b). "Exact covariance thresholding into connected components for large-scale graphical lasso". *Journal of Machine Learning Research* **13**.Mar, pp. 781–794.

McAllister, T. W. et al. (2014). "Effect of head impacts on diffusivity measures in a cohort of collegiate contact sport athletes". *Neurology* **82**.1, pp. 63–69.

McCrea, M et al. (1997). "Standardized assessment of concussion in football players". *Neurology* **48**.3, pp. 586–588.

McCrea, M. et al. (2003). "Acute effects and recovery time following concussion in collegiate football players: the NCAA Concussion Study". *Jama* **290**.19, pp. 2556–2563.

McCrory, P. et al. (2017). "Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016". *British journal of sports medicine* **51**.11, pp. 838–847.

Meier, T. B. et al. (2012). "Support vector machine classification and characterization of age-related reorganization of functional brain networks". *Neuroimage* **60**.1, pp. 601–613.

Meinshausen, N. & Yu, B. (2009). "Lasso-type recovery of sparse representations for high-dimensional data". *The annals of statistics* **37**.1, pp. 246–270.

Mejia, A. F. et al. (2018). "Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage". *NeuroImage* **172**, pp. 478–491.

Mischley, L. K. et al. (2017). "Use of a self-rating scale of the nature and severity of symptoms in Parkinson's disease (PRO-PD): correlation with quality of life and existing scales of disease severity". *npj Parkinson's Disease* **3**.1, pp. 1–7.

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

Mustafi, S. M. et al. (2018). "Acute white-matter abnormalities in sports-related concussion: a diffusion tensor imaging study from the NCAA-DoD CARE Consortium". *Journal of neurotrauma* **35**.22, pp. 2653–2664.

Nadaraya, E. A. (1964). "On estimating regression". *Theory of Probability & Its Applications* **9**.1, pp. 141–142.

Nair, A. et al. (2013). "Impaired thalamocortical connectivity in autism spectrum disorder: a study of functional and anatomical connectivity". *Brain* **136**.6, pp. 1942–1955.

Ombao, H. et al. (2019). *Handbook of Neuroimaging Data Analysis*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press LLC.

Pelckmans, K. et al. (2005). "Convex clustering shrinkage". *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

Pietrosanu, M. et al. (2021). "Estimation for the bivariate quantile varying coefficient model with application to diffusion tensor imaging data analysis". *Biostatistics*.

Poliquin, R. & Rockafellar, R (1996). "Prox-regular functions in variational analysis". *Transactions of the American Mathematical Society* **348**.5, pp. 1805–1838.

Qiu, H. et al. (2016). "Joint estimation of multiple graphical models from high dimensional time series". *Journal of the Royal Statistical Society. Series B, Statistical Methodology* **78**.2, p. 487.

Ramsay, J. O. & Silverman, B. W. (2006). *Functional data analysis*. Springer.

Relión, J. D. A. et al. (2019). "Network classification with applications to brain connectomics". *The Annals of Applied Statistics* **13**.3, pp. 1648–1677.

Rice, J. A. & Silverman, B. W. (1991). "Estimating the mean and covariance structure nonparametrically when the data are curves". *Journal of the Royal Statistical Society: Series B (Methodological)* **53**.1, pp. 233–243.

Ruppert, D (2003). *Semiparametric regression*. Cambridge New York: Cambridge University Press.

Ruppert, D. (2002). "Selecting the number of knots for penalized splines". *Journal of computational and graphical statistics* **11**.4, pp. 735–757.

Ryali, S. et al. (2012). "Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty". *NeuroImage* **59**.4, pp. 3852–3861.

Shen, H. et al. (2010). "Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fMRI". *Neuroimage* **49**.4, pp. 3110–3121.

Smith, S. M. et al. (2011). "Network modelling methods for FMRI". *Neuroimage* **54**.2, pp. 875–891.

Smith, S. M. et al. (2013a). "Functional connectomics from resting-state fMRI". *Trends in cognitive sciences* **17**.12, pp. 666–682.

Smith, S. M. et al. (2013b). "Resting-state fMRI in the human connectome project". *Neuroimage* **80**, pp. 144–168.

Solo, V. et al. (2018). "Connectivity in fMRI: blind spots and breakthroughs". *IEEE transactions on medical imaging* **37**.7, pp. 1537–1550.

Song, X. et al. (2012). "Regression analysis of longitudinal data with time-dependent covariates and informative observation times". *Scandinavian Journal of Statistics* **39**.2, pp. 248–258.

Sørlie, T. et al. (2001). "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". *Proceedings of the National Academy of Sciences* **98**.19, pp. 10869–10874.

Stone, C. J. (1982). "Optimal global rates of convergence for nonparametric regression". *The annals of statistics*, pp. 1040–1053.

Sun, J. et al. (2005). "Semiparametric regression analysis of longitudinal data with informative observation times". *Journal of the American Statistical Association* **100**.471, pp. 882–889.

Sun, J. et al. (2007). "Regression analysis of longitudinal data in the presence of informative observation and censoring times". *Journal of the American Statistical Association* **102**.480, pp. 1397–1406.

Sun, L. et al. (2012). "Joint analysis of longitudinal data with informative observation times and a dependent terminal event". *Journal of the American Statistical Association* **107**.498, pp. 688–700.

Tan, K. M. & Witten, D. M. (2014). "Sparse biclustering of transposable data". *Journal of Computational and Graphical Statistics* **23**.4, pp. 985–1008.

Tan, X. et al. (2012). "A time-varying effect model for intensive longitudinal data." *Psychological methods* **17**.1, p. 61.

Tanay, A. et al. (2005). "Biclustering algorithms: A survey". *Handbook of computational molecular biology* **9**.1-20, pp. 122–124.

Tang, A. et al. (2017). "Shyness trajectories across the first four decades predict mental health outcomes". *Journal of abnormal child psychology* **45**.8, pp. 1621–1633.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society: Series B (Methodological)* **58**.1, pp. 267–288.

Tibshirani, R. et al. (2005). "Sparsity and smoothness via the fused lasso". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**.1, pp. 91–108.

Twisk, J. & Hoekstra, T. (2012). "Classifying developmental trajectories over time should be done with great caution: a comparison between methods". *Journal of clinical epidemiology* **65**.10, pp. 1078–1087.

Varoquaux, G. et al. (2010). "Brain covariance selection: better individual functional connectivity models using population prior". *Advances in neural information processing systems*, pp. 2334–2342.

Venuto, C. S. et al. (2016). "A review of disease progression models of Parkinson's disease and applications in clinical trials". *Movement Disorders* **31**.7, pp. 947–956.

Vogelstein, J. T. et al. (2012). "Graph classification using signal-subgraphs: Applications in statistical connectomics". *IEEE transactions on pattern analysis and machine intelligence* **35**.7, pp. 1539–1551.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wand, M. P. & Jones, M. C. (1994). *Kernel smoothing*. CRC press.

Wang, H., Li, S. Z., et al. (2012). "Efficient Gaussian graphical model determination under G-Wishart prior distributions". *Electronic Journal of Statistics* **6**, pp. 168–198.

Wang, L. & Yang, L. (2009). "Spline estimation of single-index models". *Statistica Sinica*, pp. 765–783.

Wang, L. et al. (2019a). "Symmetric bilinear regression for signal subgraph estimation". *IEEE Transactions on Signal Processing* **67**.7, pp. 1929–1940.

Wang, L. et al. (2020). "Learning Clique Subgraphs in Structural Brain Network Classification with Application to Crystallized Cognition". *BioRxiv*.

Wang, Y. et al. (2019b). "Global convergence of ADMM in nonconvex nonsmooth optimization". *Journal of Scientific Computing* **78**.1, pp. 29–63.

Watson, G. S. (1964). "Smooth regression analysis". *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 359–372.

Watts, D. J. & Strogatz, S. H. (1998). "Collective dynamics of 'small-world'networks". *nature* **393**.6684, pp. 440–442.

Woo, C.-W. et al. (2017). "Building better biomarkers: brain models in translational neuroimaging". *Nature neuroscience* **20**.3, p. 365.

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

Wu, Y.-C. et al. (2020). "Longitudinal white-matter abnormalities in sports-related concussion: A diffusion MRI study". *Neurology* **95**.7, e781–e792.

Xia, R. & Mao, Z.-H. (2012). "Progression of motor symptoms in Parkinson's disease". *Neuroscience bulletin* **28**.1, pp. 39–48.

Xiao, L. (2019b). "Asymptotics of bivariate penalised splines". *Journal of Nonparametric Statistics* **31**.2, pp. 289–314.

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

Xiao, L. (2020). "Asymptotic properties of penalized splines for functional data". *Bernoulli* **26**.4, pp. 2847–2875.

Xiao, L. et al. (2016). "Fast covariance estimation for high-dimensional functional data". *Statistics and computing* **26**.1-2, pp. 409–421.

Xiao, L. et al. (2018). "Fast covariance estimation for sparse functional data". *Statistics and computing* **28**.3, pp. 511–522.

Yao, F. et al. (2003). "Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate". *Biometrics* **59**.3, pp. 676–685.

Yao, F. et al. (2005). "Functional data analysis for sparse longitudinal data". *Journal of the American statistical association* **100**.470, pp. 577–590.

Yuan, M. & Lin, Y. (2006). "Model selection and estimation in regression with grouped variables". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**.1, pp. 49–67.

Zeng, L.-L. et al. (2012). "Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis". *Brain* **135**.5, pp. 1498–1507.

Zhang, X. & Wang, J.-L. (2016). "From sparse to dense functional data and beyond". *The Annals of Statistics* **44**.5, pp. 2281–2321.

Zhao, X. et al. (2012). "Joint analysis of longitudinal data with dependent observation times". *Statistica Sinica*, pp. 317–336.

Zhou, H. & Li, L. (2014). "Regularized matrix regression". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**.2, pp. 463–483.

Zhou, H. et al. (2013). "Tensor regression with applications in neuroimaging data analysis". *Journal of the American Statistical Association* **108**.502, pp. 540–552.

Zhu, L. et al. (2011). "Regression analysis of longitudinal data with informative observation times and application to medical cost data". *Statistics in medicine* **30**.12, pp. 1429–1440.

Zou, H. (2006). "The adaptive lasso and its oracle properties". *Journal of the American statistical association* **101**.476, pp. 1418–1429.

Zou, H. & Hastie, T. (2005). "Regularization and variable selection via the elastic net". *Journal of the royal statistical society: series B (statistical methodology)* **67**.2, pp. 301–320.

APPENDICES

Appendix A

Supplementary Material of Chapter 2

A.1 Derivation of Gradients

In this section, we derive the gradients used in Section 2.3 for normally distributed and binary responses. Given $f(t) = \sum_{l=1}^{K} \theta_l \phi_l(t)$ and its derivative $f^{(1)}(t) = \sum_{l=1}^{K} \theta_l \phi_l^{(1)}(t)$, we have that

$$\nabla \left\{ \mathcal{L}_f(\mathbf{b}) + \frac{\rho}{2} \left\| \mathbf{b} - \frac{\mathbf{D}\mathbf{z}^{(k)}}{2} + \frac{\mathbf{D}\mathbf{u}^{(k)}}{2} \right\|_2^2 + \frac{\rho}{2} \left\| \mathbf{b} - \mathbf{c}^{(k)} + \mathbf{w}^{(k)} \right\|_2^2 \right\} = \nabla\{\mathcal{L}_f(\mathbf{b})\} + \mathbf{e},$$

where $\mathbf{e} = \rho(2\mathbf{b} - \frac{\mathbf{D}\mathbf{z}^{(k)}}{2} + \frac{\mathbf{D}\mathbf{u}^{(k)}}{2} - \mathbf{c}^{(k)} + \mathbf{w}^{(k)})$. When $y_i \sim \text{Normal}(f(\boldsymbol{\chi}_i^\top \mathbf{b}), \sigma^2))$,

$$\begin{aligned} \nabla \mathcal{L}_f(\mathbf{b}) &= \frac{\partial}{\partial \mathbf{b}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{\chi}_i^\top \mathbf{b}))^2 \\ &= \frac{2}{n} \sum_{i=1}^{n} (f(\boldsymbol{\chi}_i^\top \mathbf{b}) - y_i) f^{(1)}(\boldsymbol{\chi}_i^\top \mathbf{b}) \boldsymbol{\chi}_i. \end{aligned}$$

When $y_i \sim \text{Bernoulli}(\pi_i)$ with $\pi_i = s(f(\boldsymbol{\chi}_i^\top \mathbf{b})) = 1/\{1 + e^{-f(\boldsymbol{\chi}_i^\top \mathbf{b})}\}$,

$$\begin{aligned}\nabla \mathcal{L}_f(\mathbf{b}) &= \frac{\partial}{\partial \mathbf{b}} \frac{1}{n} \sum_{i=1}^{n} y_i \log s(f(\boldsymbol{\chi}_i^\top \mathbf{b})) + (1 - y_i) \log(1 - s(f(\boldsymbol{\chi}_i^\top \mathbf{b}))) \\ &= \frac{1}{n} \sum_{i=1}^{n} y_i \frac{\partial}{\partial \mathbf{b}} \log s(f(\boldsymbol{\chi}_i^\top \mathbf{b})) + (1 - y_i) \frac{\partial}{\partial \mathbf{b}} \log(1 - s(f(\boldsymbol{\chi}_i^\top \mathbf{b}))).\end{aligned}$$

Note,

$$\frac{\partial}{\partial \mathbf{b}} \log s(f(\boldsymbol{\chi}_i^\top \mathbf{b})) = \frac{f^{(1)}(\boldsymbol{\chi}^\top \mathbf{b})\boldsymbol{\chi}}{e^{f(\boldsymbol{\chi}^\top \mathbf{b})} + 1}$$

and

$$\frac{\partial}{\partial \mathbf{b}} \log(1 - s(f(\boldsymbol{\chi}_i^\top \mathbf{b}))) = -\frac{e^{f(\boldsymbol{\chi}^\top \mathbf{b})} f^{(1)}(\boldsymbol{\chi}^\top \mathbf{b})\boldsymbol{\chi}}{e^{f(\boldsymbol{\chi}^\top \mathbf{b})} + 1}.$$

Thus,

$$\nabla \mathcal{L}_f(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \frac{f^{(1)}(\boldsymbol{\chi}^\top \mathbf{b})\boldsymbol{\chi}}{e^{f(\boldsymbol{\chi}^\top \mathbf{b})} + 1} - (1 - y_i) \frac{e^{f(\boldsymbol{\chi}^\top \mathbf{b})} f^{(1)}(\boldsymbol{\chi}^\top \mathbf{b})\boldsymbol{\chi}}{e^{f(\boldsymbol{\chi}^\top \mathbf{b})} + 1} \right\}.$$

A.2 Additional Simulation Results

The simulation results are summarized for the circle network coefficient matrix in Table A.1 and Table A.2, and for the small-world network coefficient matrix in Table A.3 and Table A.4.

A.3 Additional Data Application: Age and Depression

Figure A.1 illustrates the estimated coefficient matrix and mean function for the single index models of depression score and age in the HCP data set. The effective degrees of freedom in the estimate of the f was 4.4 in the model of depression, and 2.8 in the model of age.

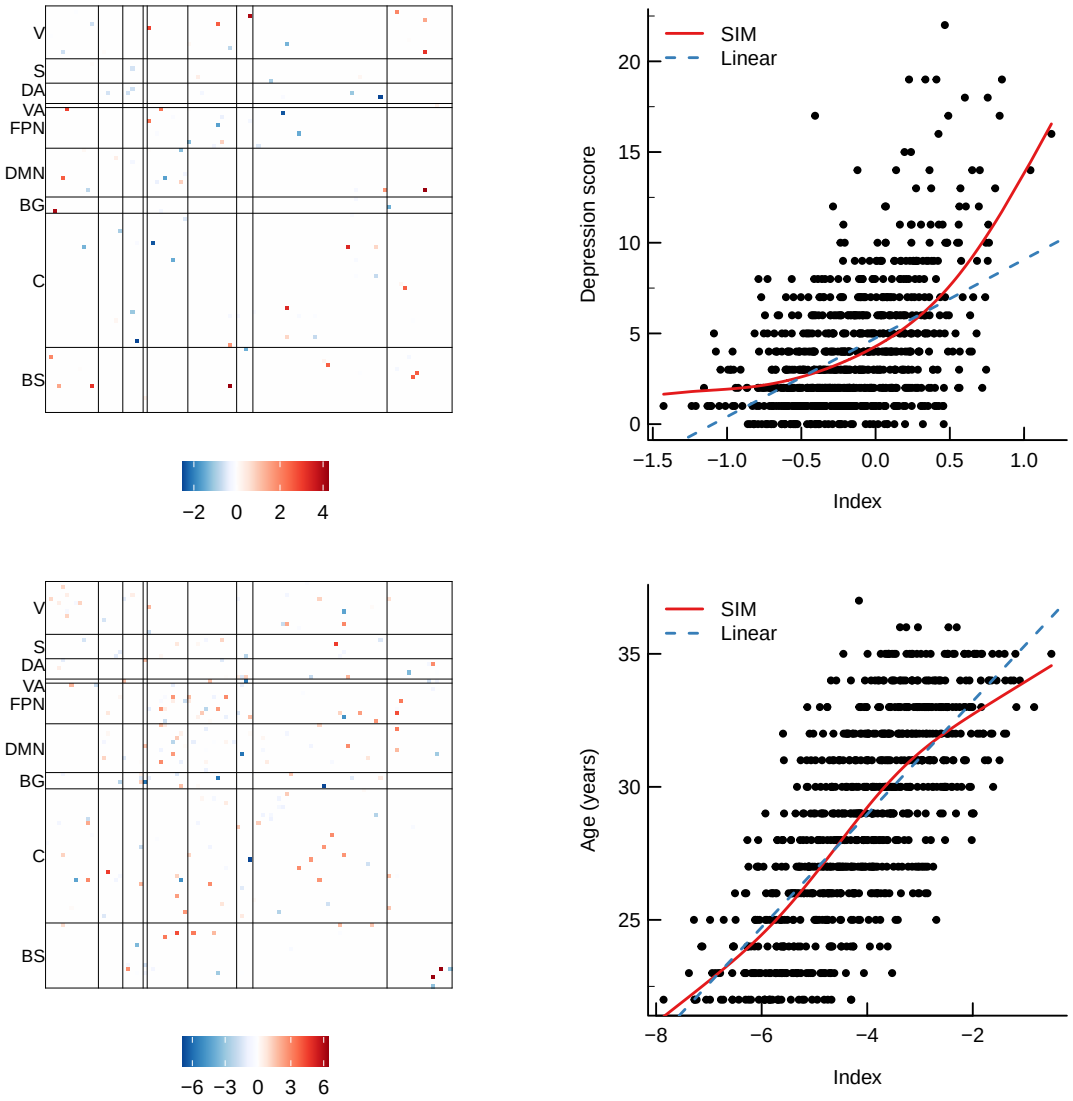

Figure A.1 Estimated coefficient matrix and mean function for the single index models of depression score (top) and age (bottom).

Table A.1 Mean squared prediction error (standard error) of the normal response simulations and AUC (standard error) from simulations with a small-world network coefficient matrix.

			Normal response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.121 (0.006)	0.122 (0.007)	0.124 (0.007)	0.121 (0.006)	0.124 (0.006)	0.121 (0.006)
		High	0.049 (0.001)	0.050 (0.002)	0.048 (0.001)	0.048 (0.002)	0.049 (0.002)	0.048 (0.001)
	100	Low	0.144 (0.009)	0.149 (0.010)	0.140 (0.008)	0.144 (0.009)	0.150 (0.010)	0.139 (0.007)
		High	0.060 (0.003)	0.063 (0.003)	0.059 (0.002)	0.060 (0.004)	0.062 (0.004)	0.058 (0.003)
f_2	25	Low	0.013 (0.000)	0.013 (0.000)	0.014 (0.000)	0.121 (0.006)	0.124 (0.006)	0.120 (0.006)
		High	0.006 (0.000)	0.006 (0.000)	0.006 (0.000)	0.048 (0.002)	0.049 (0.002)	0.048 (0.002)
	100	Low	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.140 (0.009)	0.145 (0.011)	0.137 (0.008)
		High	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.058 (0.003)	0.060 (0.004)	0.057 (0.002)
f_3	25	Low	0.124 (0.006)	0.125 (0.006)	0.128 (0.007)	0.588 (0.030)	0.598 (0.030)	0.577 (0.030)
		High	0.048 (0.002)	0.048 (0.002)	0.049 (0.002)	0.422 (0.019)	0.435 (0.018)	0.410 (0.019)
	100	Low	0.138 (0.008)	0.142 (0.009)	0.147 (0.010)	0.564 (0.040)	0.585 (0.040)	0.544 (0.039)
		High	0.057 (0.003)	0.058 (0.004)	0.061 (0.004)	0.371 (0.034)	0.393 (0.034)	0.350 (0.033)

			Binary response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.761 (0.030)	0.743 (0.029)	0.726 (0.028)	0.668 (0.019)	0.668 (0.019)	0.671 (0.021)
		High	0.915 (0.012)	0.905 (0.013)	0.895 (0.014)	0.846 (0.013)	0.846 (0.013)	0.846 (0.012)
	100	Low	0.707 (0.075)	0.664 (0.080)	0.585 (0.113)	0.538 (0.022)	0.538 (0.022)	0.544 (0.025)
		High	0.901 (0.043)	0.876 (0.048)	0.847 (0.052)	0.736 (0.020)	0.736 (0.020)	0.745 (0.015)
f_2	25	Low	0.772 (0.031)	0.755 (0.033)	0.739 (0.033)	0.523 (0.025)	0.523 (0.025)	0.528 (0.024)
		High	0.920 (0.010)	0.911 (0.011)	0.900 (0.012)	0.624 (0.073)	0.624 (0.073)	0.625 (0.077)
	100	Low	0.688 (0.081)	0.649 (0.081)	0.587 (0.102)	0.512 (0.017)	0.512 (0.017)	0.520 (0.016)
		High	0.918 (0.041)	0.894 (0.046)	0.868 (0.052)	0.569 (0.088)	0.569 (0.088)	0.591 (0.083)
f_3	25	Low	0.756 (0.036)	0.740 (0.035)	0.723 (0.034)	0.591 (0.022)	0.591 (0.022)	0.595 (0.024)
		High	0.916 (0.020)	0.907 (0.021)	0.897 (0.022)	0.643 (0.018)	0.643 (0.018)	0.654 (0.019)
	100	Low	0.691 (0.062)	0.649 (0.064)	0.566 (0.092)	0.582 (0.058)	0.582 (0.058)	0.598 (0.059)
		High	0.897 (0.046)	0.871 (0.051)	0.844 (0.056)	0.812 (0.115)	0.812 (0.115)	0.821 (0.099)

Table A.2 True positive rate (false positive rate) from simulations with a small-world network coefficient matrix.

			Normal response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.911 (0.000)	0.876 (0.000)	0.834 (0.000)	0.903 (0.000)	0.867 (0.000)	0.816 (0.000)
		High	0.916 (0.000)	0.880 (0.000)	0.830 (0.000)	0.903 (0.000)	0.864 (0.000)	0.817 (0.000)
	100	Low	0.978 (0.505)	0.973 (0.475)	0.968 (0.452)	0.978 (0.502)	0.973 (0.474)	0.968 (0.444)
		High	0.960 (0.227)	0.955 (0.210)	0.949 (0.198)	0.957 (0.224)	0.952 (0.207)	0.947 (0.192)
f_2	25	Low	0.921 (0.000)	0.888 (0.000)	0.850 (0.000)	0.906 (0.004)	0.867 (0.000)	0.820 (0.000)
		High	0.912 (0.000)	0.874 (0.000)	0.828 (0.000)	0.894 (0.000)	0.850 (0.000)	0.803 (0.000)
	100	Low	0.974 (0.474)	0.969 (0.452)	0.964 (0.428)	0.991 (0.729)	0.988 (0.695)	0.984 (0.655)
		High	0.957 (0.215)	0.952 (0.199)	0.947 (0.179)	0.981 (0.545)	0.977 (0.515)	0.972 (0.474)
f_3	25	Low	0.914 (0.004)	0.881 (0.002)	0.839 (0.000)	0.983 (0.870)	0.973 (0.820)	0.958 (0.748)
		High	0.909 (0.000)	0.868 (0.000)	0.824 (0.000)	0.939 (0.616)	0.917 (0.514)	0.892 (0.422)
	100	Low	0.976 (0.501)	0.970 (0.473)	0.965 (0.446)	0.992 (0.786)	0.990 (0.757)	0.986 (0.737)
		High	0.958 (0.222)	0.953 (0.205)	0.947 (0.189)	0.989 (0.697)	0.986 (0.661)	0.982 (0.632)

			Binary response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.963 (0.598)	0.940 (0.520)	0.911 (0.438)	0.958 (0.586)	0.934 (0.498)	0.905 (0.424)
		High	0.900 (0.056)	0.858 (0.032)	0.809 (0.022)	0.897 (0.046)	0.856 (0.032)	0.809 (0.020)
	100	Low	0.998 (0.988)	0.997 (0.978)	0.996 (0.967)	0.998 (0.988)	0.997 (0.981)	0.995 (0.971)
		High	0.991 (0.805)	0.988 (0.771)	0.983 (0.740)	0.990 (0.800)	0.987 (0.777)	0.983 (0.749)
f_2	25	Low	0.953 (0.550)	0.932 (0.466)	0.897 (0.382)	0.989 (0.968)	0.977 (0.928)	0.961 (0.898)
		High	0.887 (0.042)	0.849 (0.020)	0.797 (0.016)	0.975 (0.748)	0.958 (0.696)	0.933 (0.630)
	100	Low	0.998 (0.987)	0.998 (0.982)	0.996 (0.973)	0.999 (1.000)	0.998 (0.999)	0.997 (0.997)
		High	0.989 (0.776)	0.985 (0.748)	0.981 (0.717)	0.999 (1.000)	0.999 (0.999)	0.998 (0.999)
f_3	25	Low	0.962 (0.608)	0.944 (0.546)	0.916 (0.486)	0.979 (0.916)	0.966 (0.862)	0.947 (0.818)
		High	0.893 (0.052)	0.855 (0.038)	0.808 (0.018)	0.964 (0.842)	0.942 (0.780)	0.919 (0.692)
	100	Low	0.999 (0.990)	0.998 (0.978)	0.997 (0.969)	0.997 (0.964)	0.996 (0.950)	0.994 (0.933)
		High	0.991 (0.798)	0.988 (0.778)	0.984 (0.746)	0.990 (0.783)	0.986 (0.757)	0.982 (0.735)

Table A.3 Mean squared prediction error (standard error) of the normal response simulations and AUC (standard error) from simulations with a circle network coefficient matrix.

			Normal response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.108 (0.017)	0.109 (0.018)	0.111 (0.019)	0.110 (0.017)	0.112 (0.018)	0.109 (0.017)
		High	0.041 (0.009)	0.042 (0.009)	0.043 (0.010)	0.041 (0.010)	0.042 (0.010)	0.040 (0.010)
	100	Low	0.118 (0.015)	0.122 (0.016)	0.126 (0.017)	0.127 (0.014)	0.132 (0.014)	0.123 (0.014)
		High	0.047 (0.005)	0.048 (0.006)	0.050 (0.006)	0.048 (0.007)	0.050 (0.008)	0.047 (0.006)
f_2	25	Low	0.104 (0.019)	0.105 (0.020)	0.107 (0.020)	0.017 (0.002)	0.017 (0.002)	0.017 (0.002)
		High	0.041 (0.010)	0.041 (0.010)	0.042 (0.010)	0.006 (0.002)	0.006 (0.002)	0.006 (0.002)
	100	Low	0.120 (0.016)	0.124 (0.016)	0.128 (0.017)	0.013 (0.008)	0.014 (0.008)	0.013 (0.007)
		High	0.049 (0.008)	0.050 (0.008)	0.052 (0.008)	0.006 (0.003)	0.006 (0.003)	0.006 (0.003)
f_3	25	Low	0.111 (0.024)	0.113 (0.025)	0.110 (0.023)	0.511 (0.076)	0.524 (0.076)	0.499 (0.075)
		High	0.045 (0.010)	0.046 (0.010)	0.044 (0.009)	0.386 (0.066)	0.396 (0.063)	0.376 (0.067)
	100	Low	0.123 (0.016)	0.127 (0.017)	0.119 (0.015)	0.500 (0.058)	0.510 (0.063)	0.489 (0.055)
		High	0.052 (0.007)	0.053 (0.008)	0.050 (0.006)	0.367 (0.050)	0.377 (0.050)	0.356 (0.050)

			Binary response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.653 (0.036)	0.651 (0.037)	0.651 (0.037)	0.742 (0.043)	0.722 (0.045)	0.761 (0.040)
		High	0.843 (0.014)	0.897 (0.012)	0.843 (0.014)	0.843 (0.014)	0.887 (0.013)	0.906 (0.012)
	100	Low.	0.554 (0.024)	0.554 (0.024)	0.563 (0.024)	0.681 (0.076)	0.641 (0.087)	0.714 (0.080)
		High	0.706 (0.037)	0.706 (0.037)	0.716 (0.033)	0.846 (0.063)	0.821 (0.065)	0.871 (0.060)
f_2	25	Low	0.771 (0.029)	0.753 (0.030)	0.734 (0.032)	0.536 (0.024)	0.536 (0.024)	0.539 (0.025)
		High	0.916 (0.017)	0.908 (0.018)	0.899 (0.018)	0.655 (0.101)	0.655 (0.101)	0.651 (0.082)
	100	Low	0.675 (0.060)	0.641 (0.058)	0.566 (0.088)	0.523 (0.024)	0.523 (0.024)	0.519 (0.028)
		High	0.878 (0.064)	0.854 (0.067)	0.829 (0.070)	0.536 (0.036)	0.536 (0.036)	0.547 (0.036)
f_3	25	Low	0.742 (0.040)	0.718 (0.048)	0.688 (0.073)	0.617 (0.090)	0.617 (0.090)	0.626 (0.089)
		High	0.914 (0.013)	0.905 (0.014)	0.896 (0.015)	0.769 (0.114)	0.769 (0.114)	0.779 (0.106)
	100	Low	0.708 (0.081)	0.675 (0.080)	0.611 (0.111)	0.579 (0.058)	0.579 (0.058)	0.592 (0.054)
		High	0.859 (0.064)	0.829 (0.070)	0.795 (0.086)	0.634 (0.109)	0.634 (0.109)	0.648 (0.106)

Table A.4 True positive rate (false positive rate) from simulations with a circle network coefficient matrix.

			Normal response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.900 (0.004)	0.859 (0.002)	0.816 (0.000)	0.907 (0.002)	0.871 (0.000)	0.828 (0.000)
		High	0.904 (0.000)	0.866 (0.000)	0.819 (0.000)	0.900 (0.000)	0.863 (0.000)	0.815 (0.000)
	100	Low	0.974 (0.499)	0.969 (0.477)	0.964 (0.454)	0.979 (0.553)	0.975 (0.512)	0.971 (0.479)
		High	0.953 (0.198)	0.949 (0.180)	0.944 (0.170)	0.954 (0.204)	0.949 (0.184)	0.944 (0.169)
f_2	25	Low	0.902 (0.000)	0.866 (0.000)	0.819 (0.000)	0.908 (0.000)	0.870 (0.000)	0.828 (0.000)
		High	0.912 (0.000)	0.878 (0.000)	0.833 (0.000)	0.888 (0.000)	0.850 (0.000)	0.798 (0.000)
	100	Low	0.977 (0.526)	0.973 (0.492)	0.968 (0.461)	0.983 (0.618)	0.979 (0.588)	0.974 (0.559)
		High	0.959 (0.234)	0.953 (0.210)	0.948 (0.192)	0.964 (0.340)	0.960 (0.311)	0.955 (0.289)
f_3	25	Low	0.897 (0.000)	0.864 (0.000)	0.818 (0.000)	0.946 (0.396)	0.925 (0.348)	0.898 (0.292)
		High	0.908 (0.000)	0.874 (0.000)	0.829 (0.000)	0.955 (0.462)	0.932 (0.428)	0.903 (0.396)
	100	Low	0.976 (0.518)	0.971 (0.483)	0.966 (0.455)	0.996 (0.906)	0.994 (0.888)	0.991 (0.871)
		High	0.956 (0.217)	0.951 (0.198)	0.945 (0.184)	0.994 (0.885)	0.992 (0.869)	0.990 (0.855)

			Binary response					
			Nonlinear			Linear		
f	p	SNR	Node & edge	Elastic net	Lasso	Node & edge	Elastic net	Lasso
f_1	25	Low	0.948 (0.564)	0.924 (0.504)	0.889 (0.438)	0.949 (0.582)	0.927 (0.508)	0.893 (0.426)
		High	0.877 (0.062)	0.842 (0.042)	0.796 (0.030)	0.891 (0.054)	0.853 (0.034)	0.805 (0.026)
	100	Low	0.998 (0.984)	0.997 (0.978)	0.995 (0.971)	0.999 (0.991)	0.998 (0.987)	0.997 (0.977)
		High	0.991 (0.856)	0.988 (0.830)	0.985 (0.803)	0.992 (0.872)	0.989 (0.838)	0.986 (0.805)
f_2	25	Low	0.948 (0.542)	0.923 (0.466)	0.888 (0.388)	0.996 (0.978)	0.984 (0.936)	0.966 (0.900)
		High	0.874 (0.048)	0.841 (0.034)	0.792 (0.024)	0.962 (0.712)	0.944 (0.638)	0.918 (0.566)
	100	Low	0.999 (0.993)	0.998 (0.985)	0.997 (0.977)	0.999 (0.999)	0.998 (0.999)	0.997 (0.998)
		High	0.991 (0.855)	0.988 (0.835)	0.984 (0.819)	0.999 (0.993)	0.998 (0.987)	0.997 (0.983)
f_3	25	Low	0.969 (0.650)	0.948 (0.534)	0.913 (0.454)	0.975 (0.762)	0.955 (0.708)	0.928 (0.632)
		High	0.889 (0.076)	0.848 (0.050)	0.806 (0.032)	0.926 (0.440)	0.900 (0.388)	0.865 (0.356)
	100	Low	0.998 (0.989)	0.997 (0.982)	0.995 (0.974)	0.997 (0.983)	0.996 (0.979)	0.995 (0.972)
		High	0.992 (0.879)	0.990 (0.857)	0.986 (0.828)	0.995 (0.936)	0.994 (0.927)	0.992 (0.918)

Appendix B

Supplementary Material of Chapter 3

B.1 Technical Lemmas for the Proof of Theorem 3.1

First, we define some notation. Let $\tau_1 = \sum_{i=1}^n w_i^2 m_i$ and $\tau_2 = w_i^2 m_i(m_i - 1)$. Define $Q_{ni}(t) = m_i^{-1} N_i(t)$, $Q_n(t) = \sum_{i=1}^n w_i m_i Q_{ni}(t)$ and $\widetilde{Q}_n(t) = \tau_1^{-1} \sum_{i=1}^n w_i^2 m_i Q_{ni}(t)$. Both $Q_n(t)$ and $\widetilde{Q}_n(t)$ are empirical cumulative distribution functions (CDF) and they shall be proved to converge to proper CDFs. Moreover, let $R_n(s,t) = \tau_2^{-1} \sum_{i=1}^n \sum_{j_1 \neq j_2} w_i^2 I(T_{ij_1} \leq s) I(T_{ij_2} \leq t)$, which is a bivariate empirical CDF.

In Lemmas B.1 - B.3 below, suppose that Assumptions 3.1 - 3.6 hold.

Lemma B.1. *Define*

$$Q(t) = \mathrm{E}\{\Lambda_0(t \wedge C_1)/\Lambda_0(C_1)\}$$

and

$$R(s,t) = \mathrm{E}\{\Lambda_0(t \wedge C_1)\Lambda_0(s \wedge C_1)/\Lambda_0^2(C_1)\}.$$

If $\log n/n = o(h^4)$*, the following results hold almost surely:*

(i) $\|Q_n - Q\| = o(h)$.

(ii) $\|\widetilde{Q}_n - Q\| = o(h)$.

(iii) Let $U(s,t) = R_n(s,t) - R(s,t)$*. Then,* $\|U\| = o(h^2)$.

Proof. We first consider (i) and have

$$\mathrm{E}\{\mathrm{d}N_i(t) \mid m_i, C_i\} = I(C_i \geq t) m_i \frac{\mathrm{d}\Lambda_0(t)}{\Lambda_0(C_i)}.$$

Hence,

$$\begin{aligned}\mathrm{E}\{Q_{ni}(t) \mid m_i, C_i\} =& \mathrm{E}\left\{\frac{1}{m_i}\int_0^t \mathrm{d}N_i(s) \mid m_i, C_i\right\} \\ =& \int_0^t I(C_i \geq s)\frac{\mathrm{d}\Lambda_0(s)}{\Lambda_0(C_i)} \\ =& \frac{\Lambda_0(t \wedge C_i)}{\Lambda_0(C_i)}.\end{aligned}$$

Let

$$F_n(t) = n^{-1}\sum_{i=1}^n \frac{\Lambda_0(t \wedge C_i)}{\Lambda_0(C_i)}.$$

Define $S_{ni}(t) = Q_{ni}(t) - \Lambda_0(t \wedge C_i)/\Lambda_0(C_i)$ and $S_n(t) = Q_n(t) - F_n(t)$. As $w_i = (nm_i)^{-1}$, $S_n(t) = n^{-1}\sum_{i=1}^n S_{ni}(t)$. We have $\mathrm{E}[\{S_{ni}(t)\}^2 \mid m_i, C_i] \leq m_i^{-1}$ and thus $\mathrm{E}[\{S_n(t)\}^2 \mid m_i, C_i, 1 \leq i \leq n] \leq n^{-1}$ for any t. Then by using the same technique as in the proof of Lemma A.7 in Xiao (2020), $\|Q_n - F_n\| = o(h)$ almost surely as we assume $\log n/n = o(h^4)$. Note that $\Lambda_0(t \wedge C_i)/\Lambda_0(C_i)$ is a non-decreasing function of t and is bounded by 1. Then similarly, it can be shown that $\|F_n - Q\| = o(h)$ almost surely.

Next for (ii), let

$$\widetilde{F}_n(t) = \tau_1^{-1}\sum_{i=1}^n w_i^2 m_i \frac{\Lambda_0(t \wedge C_i)}{\Lambda_0(C_i)}.$$

Then, we derive that $\|\widetilde{Q}_n - \widetilde{F}_n\| = o(h)$ as long as Lemma B.2 (i) and (ii) hold almost surely. Similarly, we could show that $\|\widetilde{F}_n - Q\| = o(h)$ almost surely.

Finally, let

$$F_n(s,t) = \tau_2^{-1}\sum_{i=1}^n w_i^2 m_i(m_i - 1)\frac{\Lambda_0(s \wedge C_i)\Lambda_0(t \wedge C_i)}{\Lambda_0^2(C_i)}.$$

Then, we derive that $\|R_n - F_n\| = o(h^2)$ almost surely if Lemma B.2 (iii) and (iv) hold almost surely. Similarly, we could show that $\|F_n - R\| = o(h^2)$ almost surely. □

Lemma B.2. *If* $\log n/n = o(h^4)$*, the following results hold almost surely:*

(i) $\tau_1^{-2}\left(\sum_{i=1}^n m_i w_i^4\right) = o(h^2/\log n)$.

(ii) $\tau_1^{-1}\max_i m_i w_i^2 = o(h/\log n)$.

(iii) $\tau_2^{-2}\left\{\sum_{i=1}^n m_i(m_i-1)(4m_i-6)w_i^4\right\} = o(h^4/\log n)$.

(iv) $\tau_2^{-1}\max_i m_i(m_i-1)w_i^2 = o(h^2/\log n)$.

Proof. First consider (i) and (ii). We have $\tau_1 = n^{-2}\sum_i m_i^{-1}$. For (iii), we have

$$\tau_1^{-2}\left(\sum_i m_i w_i^4\right) \leq \left(\sum_i m_i^{-1}\right)^{-1} = O(n^{-1})$$

almost surely by Lemma B.3 and then (i) holds.

For (ii),

$$\tau_1^{-1}\max_i m_i w_i^2 \leq \left(\sum_i m_i^{-1}\right)^{-1} = O(n^{-1}).$$

Hence, (ii) holds.

Next consider (iii) and (iv). We have

$$\tau_2 = n^{-2}\sum_i \left(1 - m_i^{-1}\right).$$

For (iii),

$$\tau_2^{-2}\left\{\sum_{i=1}^n m_i(m_i-1)(4m_i-6)w_i^4\right\} \leq \frac{4\sum_i m_i^{-1}}{\left\{\sum_i \left(1-m_i^{-1}\right)\right\}^2}.$$

For (iv),

$$\tau_2^{-1}\max_i m_i(m_i-1)w_i^2 \leq \left\{\sum_i \left(1-m_i^{-1}\right)\right\}^{-1}.$$

Thus, by Lemma B.3 and the assumption that $\log n/n = o(h^4)$, both (iii) and (iv) hold. □

Lemma B.3. *There exist constants* $0 < c_1 < c_2 < 1$ *such that*

$$c_1 < n^{-1}\sum_i I(m_i \neq 0)m_i^{-1} < c_2, \quad \textit{almost surely.}$$

Proof. Let $a_i = I(m_i \geq 0)m_i^{-1}$, then marginally $a_i\,(1 \leq i \leq n)$ are independent and identically distributed. By the strong law of large numbers, $n^{-1}\sum_i I(m_i \neq 0)m_i^{-1}$ converges almost surely to $\mathrm{E}(a_1)$, which is bounded by $c_2 = (1 + e^{-2})/2$. Indeed, by Lemma B.4, $\mathrm{E}(a_1) = \mathrm{E}\{\mathrm{E}(a_1 \mid Z_1, C_1)\} \leq \mathrm{E}(c_2) = c_2$. The proof is complete if $\mathrm{E}(a_1) > c_1$ for some constant $c_1 > 0$. First, we have $\mathrm{E}(a_1 \mid Z_1, C_1) > \lambda_1 e^{-\lambda_1}$, where $\lambda_1 = c_0 Z_1 \Lambda_0(C_1)$. Thus,

$$\mathrm{E}(a_1 \mid Z_1) \geq \mathrm{E}\left\{c_0 Z_1 \Lambda_0(C_1) e^{-c_0 Z_1 \Lambda_0(C_1)} \mid Z_1\right\} \geq c_0 \mathrm{E}\{\Lambda_0(C_1)\} Z_1 e^{-c_0 Z_1 \Lambda_0(1)}$$

and

$$\mathrm{E}(a_1) \geq c_0 \mathrm{E}\{\Lambda_0(C_1)\} \mathrm{E}\left\{Z_1 e^{-c_0 Z_1 \Lambda_0(1)}\right\}.$$

Let c_1 be the right hand side of the above inequality. By Assumptions 3.1 and 3.2, $c_1 > 0$. The proof is now complete. □

Lemma B.4. *If m has a Poisson distribution with mean $\lambda \geq 0$. Then*

$$\mathrm{E}\{I(m \neq 0)m^{-1}\} \leq \left(1 + e^{-2}\right)/2.$$

Proof. We derive that

$$\begin{aligned}
\mathrm{E}\{I(m \neq 0)m^{-1}\} &= \sum_{k=1}^{\infty} \frac{1}{k} \frac{e^{-\lambda}\lambda^k}{k!} \\
&\leq \frac{1}{2} e^{-\lambda}\lambda + \frac{1}{2} \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \\
&\leq \frac{1}{2} e^{-\lambda}\lambda + \frac{1}{2}(1 - e^{-\lambda}) \\
&\leq \frac{1}{2}\{1 + (\lambda - 1)e^{-\lambda}\}.
\end{aligned}$$

Let $f(\lambda) = (\lambda - 1)e^{-\lambda}$. It is easy to check that f has maximum at $\lambda = 2$. Thus, for any $\lambda \geq 0$,

$$\mathrm{E}\{I(m \neq 0)m^{-1}\} \leq 2^{-1}\{1 + f(2)\} = 2^{-1}(1 + e^{-2}).$$

□

B.2 Proof of Theorem 3.2

First, we define some additional notation. The Euclidean norm is $\|\cdot\|_2$ and the operator norm is $\|\cdot\|_{op}$. For a matrix $A = (a_{ij})$, $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ and $\|A\|_\infty = \max_i \sum_j |a_{ij}|$.

Let a and b be a sequence of scalars that depend on n. The notation $a \preceq b$ denotes that there exists an absolute constant $c > 0$ such that $a \leq cb$ for sufficiently large n. And $a \simeq b$ means that $a \preceq b$ and $b \preceq a$, i.e., a and b are rate-wise equivalent. For two square matrices A and B, $A \leq B$ denotes that $B - A$ is elementwise nonnegative, $A \preceq B$ means that there exists an absolute constant $c > 0$ such that $A \leq cB$ for sufficiently large n, and $A \simeq B$ means that $A \preceq B$ and also $B \preceq A$.

Also to simplify notation, in the rest of the proof, unless otherwise noted, the expectation and variance are conditional on $\underline{m}$ and $\underline{C}$.

First, by symmetry of the auxiliary variables, i.e., $\widetilde{\sigma}_{ij_1j_2} = \widetilde{\sigma}_{ij_2j_1}$, it can be shown that $\widehat{\Theta}$, the solution to (5), is also the minimizer of the following optimization problem,

$$\widehat{\Theta} = \arg\min_{\Theta} \sum_{i=1}^{n} \left[v_i \sum_{1 \leq j_1 \neq j_2 \leq m_i} \{\widetilde{\sigma}_{ij_1j_2} - H(T_{ij_1}, T_{ij_2})\}^2 \right] + \boldsymbol{\theta}_\sigma' P_\sigma \boldsymbol{\theta}_\sigma, \tag{B.1}$$

where $P_\sigma = 2^{-1}\lambda K^{-1}(I_K \otimes P + P \otimes I_K)$. We shall use the above formulation to derive an explicit expression of $\widehat{\sigma}(s,t)$. To simplify notation, we shall drop the subscript in λ_σ in the following derivation.

Let $e_{ij} = Y_{ij} - \mu(T_{ij})$ and $\widetilde{\sigma}_{ij_1j_2} = e_{ij_1}e_{ij_2}$. Then let $\Sigma_i = \{\sigma(T_{ij_1}, T_{ij_2})\}_{1 \leq j_1, j_2 \leq m_i} \in \mathbb{R}^{m_i \times m_i}$ and $\boldsymbol{\sigma}_i = \text{vec}^*(\Sigma_i)$, where vec* is a matrix operator that is the same as vec$(\cdot)$ except it excludes the diagonal elements of a square matrix. Similarly, let $\widehat{\Sigma}_i = (\widetilde{\sigma}_{ij_1j_2})_{1 \leq j_1, j_2 \leq m_i}$ and $\widehat{\boldsymbol{\sigma}}_i = \text{vec}^*(\widehat{\Sigma}_i)$. Also let $B_i = [B(T_{i1}), \ldots, B(T_{im_i})]' \in \mathbb{R}^{m_i \times K}$ and A_i be the submatrix of $B_i \otimes B_i$ that excludes the rows corresponding to the same T_{ij}. Finally, let $V_i = v_i I_{m_i(m_i-1)}$. Then, the objective function in equation (B.1) becomes

$$\sum_{i=1}^{n} (\widetilde{\boldsymbol{\sigma}}_i - A_i\boldsymbol{\theta}_\sigma)^\top V_i (\widetilde{\boldsymbol{\sigma}}_i - A_i\boldsymbol{\theta}_\sigma) + \boldsymbol{\theta}_\sigma' P_\sigma \boldsymbol{\theta}_\sigma.$$

Now let $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1^\top, \ldots, \boldsymbol{\sigma}_n^\top)^\top$, $\widetilde{\boldsymbol{\sigma}} = (\widetilde{\boldsymbol{\sigma}}_1^\top, \ldots, \widetilde{\boldsymbol{\sigma}}_n^\top)^\top$, $V = \text{blockdiag}(V_1, \ldots, V_n)$, and $A =$

$[A_1^\top, \ldots, A_n^\top]^\top$. Finally, let $G_{\sigma,n} = A^\top VA$ and $H_{\sigma,n} = G_{\sigma,n} + P_\sigma$. Then we derive

$$\widehat{\sigma}(s,t) = B^\top(s,t)\widehat{\boldsymbol{\theta}}_\sigma = B^\top(s,t)H_{\sigma,n}^{-1}(A^\top V\widetilde{\boldsymbol{\sigma}}).$$

We now consider $\|\mathrm{E}\widehat{\sigma} - \sigma\|$ and derive that

$$\begin{aligned}\mathrm{E}\widehat{\sigma}(s,t) =& B^\top(s,t)H_{\sigma,n}^{-1}(A^\top V\boldsymbol{\sigma}) \\ =& B^\top(s,t)G_{\sigma,n}^{-1}(A^\top V\boldsymbol{\sigma}) - B^\top(s,t)H_{\sigma,n}^{-1}P_\sigma G_{\sigma,n}^{-1}(A^\top V\boldsymbol{\sigma}).\end{aligned}$$

Let η_σ be as in Lemma S.3.1 in the supplement of Xiao (2020) and such that

$$\|\sigma - \eta_\sigma\| = O(h_\sigma^r) + o\left(h_\sigma^p\right). \tag{B.2}$$

Note, $B^\top(s,t)G_{\sigma,n}^{-1}(A^\top V\boldsymbol{\eta}_\sigma) = \eta_\sigma(s,t)$, where $\boldsymbol{\eta}_\sigma$ is similarly defined as $\boldsymbol{\sigma}$. Thus,

$$\mathrm{E}\widehat{\sigma}(s,t) = \eta_\sigma(s,t) + B^\top(s,t)G_{\sigma,n}^{-1}\left\{A^\top V(\boldsymbol{\sigma} - \boldsymbol{\eta}_\sigma)\right\} - B^\top(s,t)H_{\sigma,n}^{-1}P_\sigma G_{\sigma,n}^{-1}(A^\top V\boldsymbol{\sigma}),$$

Let $\boldsymbol{\alpha}_\sigma = A^\top V(\boldsymbol{\sigma} - \boldsymbol{\eta})$ and $\boldsymbol{\gamma}_\sigma = G_{\sigma,n}^{-1}(A^\top V\boldsymbol{\sigma})$. Then,

$$(\mathrm{E}\widehat{\sigma} - \sigma)(s,t) = (\eta_\sigma - \sigma)(s,t) + B^\top(s,t)G_{\sigma,n}^{-1}\boldsymbol{\alpha}_\sigma - B^\top(s,t)H_{\sigma,n}^{-1}P_\sigma\boldsymbol{\gamma}_\sigma.$$

It follows that

$$\begin{aligned}&\frac{1}{3}\iint \left\{\mathrm{E}\widehat{\sigma}(s,t) - \sigma(s,t)\right\}^2 d_s d_t R(s,t) \\ \leq& \iint \left\{\eta_\sigma(s,t) - \sigma(s,t)\right\}^2 d_s d_t R(s,t) + \boldsymbol{\alpha}_\sigma^\top G_{\sigma,n}^{-1}G_\sigma G_{\sigma,n}^{-1}\boldsymbol{\alpha}_\sigma + \boldsymbol{\gamma}_\sigma' P_\sigma H_{\sigma,n}^{-1}G_\sigma H_{\sigma,n}^{-1}P_\sigma\boldsymbol{\gamma}_\sigma,\end{aligned} \tag{B.3}$$

where $G_\sigma = \iint \{B(s)B^\top(s)\} \otimes \{B(t)B^\top(t)\} d_s d_t R(s,t)$.

Let $\alpha_{\sigma,k\ell}$ be the (k,ℓ)th element of $\boldsymbol{\alpha}_\sigma$. Then

$$\begin{aligned}\alpha_{\sigma,k\ell} =& \sum_{i=1}^n \sum_{1\leq j_1\neq j_2\leq m_i} v_i B_k(T_{ij_1})B_\ell(T_{ij_2})\left\{\sigma(T_{ij_1},T_{ij_2}) - \eta_\sigma(T_{ij_1},T_{ij_2})\right\} \\ =& \iint B_k(s)B_\ell(t)\left\{\sigma(s,t) - \eta_\sigma(s,t)\right\} d_s d_t R_n(s,t).\end{aligned}$$

By Lemma S.3.2 in the supplement of Xiao (2020) and Lemma 1 (*iii*), it can be shown that $\|\boldsymbol{\alpha}_\sigma\|_{\max} = o(h^{p+2})$. Then, by Lemma B.6,

$$\boldsymbol{\alpha}_\sigma^\top G_{\sigma,n}^{-1} G_\sigma G_{\sigma,n}^{-1} \boldsymbol{\alpha}_\sigma = o(h^p), \text{ a.s.} \tag{B.4}$$

Next by a proof similar to that of (8.15) in Xiao (2019a), we derive that

$$\boldsymbol{\gamma}_\sigma' P_\sigma H_{\sigma,n}^{-1} G_\sigma H_{\sigma,n}^{-1} P_\sigma \boldsymbol{\gamma}_\sigma = O\left\{ \left(1 \vee \|G_{\sigma,n}^{-1/2} P_\sigma G_{\sigma,n}^{-1/2}\|_{op}\right) \boldsymbol{\gamma}_\sigma' P_\sigma \boldsymbol{\gamma}_\sigma \right\}.$$

First we have

$$\|G_{\sigma,n}^{-1/2} P_\sigma G_{\sigma,n}^{-1/2}\|_{op} \le \|G_{\sigma,n}^{-1}\|_{op} \|P_\sigma\|_{op} \simeq \lambda K^{2q-2} h^{-2} \simeq \lambda h^{-2q}, \text{ a.s.}$$

Then by a proof similar to that of Lemma A.1 in Xiao (2019b), we derive that $\boldsymbol{\gamma}_\sigma' P_\sigma \boldsymbol{\gamma}_\sigma = O(\lambda)$ almost surely. It follows that

$$\boldsymbol{\gamma}_\sigma' P_\sigma H_{\sigma,n}^{-1} G_\sigma H_{\sigma,n}^{-1} P_\sigma \boldsymbol{\gamma}_\sigma = O\left(\lambda \wedge \lambda^2 h^{-2q}\right) = O(\lambda^2 h_e^{-2q}), \text{ a.s.} \tag{B.5}$$

Combining equations (B.2) - (B.5),

$$\iint \left\{\mathrm{E}\widehat{\sigma}(s,t) - \sigma(s,t)\right\}^2 d_s d_t R(s,t) = O(h^{2r}) + o(h^{2p}) + O(\lambda^2 h_e^{-2q}), \text{ a.s.} \tag{B.6}$$

Now we consider the variance of $\widehat{\sigma}(s,t)$ and derive that

$$\mathrm{Var}\left\{\widehat{\sigma}(s,t)\right\} = B^\top(s,t) H_{\sigma,n}^{-1} A^\top V \mathrm{E}\left(\widehat{\boldsymbol{\sigma}}\widehat{\boldsymbol{\sigma}}^\top\right) V A H_{\sigma,n}^{-1} B(s,t).$$

Let

$$\widetilde{\Pi} = (\widetilde{\pi}_{k_1\ell_1 k_2\ell_2}) = A^\top V \mathrm{E}\left(\widehat{\boldsymbol{\sigma}}\widehat{\boldsymbol{\sigma}}^\top\right) V A = \sum_{i=1}^n A_i^\top V_i \mathrm{E}\left(\widehat{\boldsymbol{\sigma}}_i \widehat{\boldsymbol{\sigma}}_i^\top\right) V_i A_i.$$

Then, $\mathrm{Var}\left\{\widehat{\sigma}(s,t)\right\} = B^\top(s,t) H_{\sigma,n}^{-1} \widetilde{\Pi} H_{\sigma,n}^{-1} B(s,t)$ and

$$\iint \mathrm{Var}\left\{\widehat{\sigma}(s,t)\right\} d_s d_t R(s,t) = \mathrm{Tr}\left(H_{\sigma,n}^{-1} \widetilde{\Pi} H_{\sigma,n}^{-1} G_\sigma\right). \tag{B.7}$$

Since $G_\sigma \simeq h^2 I$ almost surely by Lemma B.6,

$$\mathrm{Tr}\left(H_{\sigma,n}^{-1}\widetilde{\Pi}H_{\sigma,n}^{-1}G_\sigma\right) \simeq h^2\mathrm{Tr}\left(H_{\sigma,n}^{-2}\widetilde{\Pi}\right), \text{ a.s.} \tag{B.8}$$

By Lemma B.8,

$$H_{\sigma,n}^{-2} \preceq h^{-2}H^{-1}\otimes H^{-1}, \text{ a.s.}$$

where H is defined in Lemma B.7. As a result,

$$\mathrm{Tr}\left(H_{\sigma,n}^{-2}\widetilde{\Pi}\right) = O\left[h^{-2}\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)\widetilde{\Pi}\right\}\right], \text{ a.s.} \tag{B.9}$$

Let $\Pi = (\pi_{k_1\ell_1k_2\ell_2})$ with

$$\pi_{k_1\ell_1k_2\ell_2} = \sum_{i=1}^n v_i^2 \sum_{j_1\neq j_2, j_3\neq j_4} B_{k_1}(T_{ij_1})B_{\ell_1}(T_{ij_2})B_{k_2}(T_{ij_3})B_{\ell_2}(T_{ij_4}). \tag{B.10}$$

Then, by the proof between equation (S.14) and (S.18) in the supplement of Xiao (2020), we get

$$\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)\widetilde{\Pi}\right\} = O\left[\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)_+\Pi\right\}\right], \tag{B.11}$$

where for a matrix $A = (a_{ij})$, $A_+ = (|a_{ij}|)$. Note that

$$\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)_+\Pi\right\} = O_{\mathbb{P}}\left[\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)_+\mathrm{E}(\Pi)\right\}\right],$$

and combining equations (B.7) - (B.11), we have

$$\iint \mathrm{Var}\left\{\widehat{\sigma}(s,t)\right\} d_s d_t R(s,t) = O_{\mathbb{P}}\left[\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)_+\mathrm{E}(\Pi)\right\}\right]. \tag{B.12}$$

By the definition of Π in (B.10), we have the matrix format

$$\begin{aligned}\Pi &= \sum_{i=1}^n v_i^2 \sum_{j_1\neq j_2, j_3\neq j_4} \{B(T_{ij_2})\otimes B(T_{ij_1})\}\{B(T_{ij_4})\otimes B(T_{ij_3})\}^\top \\ &= \sum_{i=1}^n v_i^2 \sum_{j_1\neq j_2, j_3\neq j_4} \{B(T_{ij_2})B^\top(T_{ij_4})\}\otimes\{B(T_{ij_1})B^\top(T_{ij_3})\}.\end{aligned}$$

To simplify notation, let $F_i(s) = \Lambda_0(s \wedge C_i)/\Lambda_0(C_i)$. It follows that

$$\mathrm{E}(\Pi) = \Pi_{11} + \Pi_{12} + \Pi_{21} + \Pi_{22} + \Pi_{23} + \Pi_{24} + \Pi_{31},$$

where

$$\begin{aligned}
\Pi_{11} &= \sum_{i=1}^{n} I(m_i \geq 2) v_i^2 m_i(m_i-1) \iint \{B(t)B^\top(t)\} \otimes \{B(s)B^\top(s)\} d_s F_i(s) d_t F_i(t), \\
\Pi_{12} &= \sum_{i=1}^{n} \frac{v_i^2 I(m_i \geq 2)}{m_i(m_i-1)} \iint \{B(t)B^\top(s)\} \otimes \{B(s)B^\top(t)\} d_s F_i(s) d_t F_i(t), \\
\Pi_{21} &= \sum_{i=1}^{n} I(m_i \geq 3) v_i^2 m_i(m_i-1)(m_i-2) \\
&\quad \times \iiint \{B(t_1)B^\top(t_2)\} \otimes \{B(s)B^\top(s)\} d_s F_i(s) d_{t_1} F_i(t_1) d_{t_2} F_i(t_2), \\
\Pi_{22} &= \sum_{i=1}^{n} I(m_i \geq 3) v_i^2 m_i(m_i-1)(m_i-2) \\
&\quad \times \iiint \{B(t_1)B^\top(s)\} \otimes \{B(s)B^\top(t_2)\} d_s F_i(s) d_{t_1} F_i(t_1) d_{t_2} F_i(t_2), \\
\Pi_{23} &= \sum_{i=1}^{n} I(m_i \geq 3) v_i^2 m_i(m_i-1)(m_i-2) \\
&\quad \times \iiint \{B(s)B^\top(s)\} \otimes \{B(t_1)B^\top(t_2)\} d_s F_i(s) d_{t_1} F_i(t_1) d_{t_2} F_i(t_2), \\
\Pi_{24} &= \sum_{i=1}^{n} I(m_i \geq 3) v_i^2 m_i(m_i-1)(m_i-2) \\
&\quad \times \iiint \{B(s)B^\top(t_1)\} \otimes \{B(t_2)B^\top(s)\} d_s F_i(s) d_{t_1} F_i(t_1) d_{t_2} F_i(t_2), \\
\Pi_{31} &= \sum_{i=1}^{n} I(m_i \geq 4) v_i^2 m_i(m_i-1)(m_i-2)(m_i-3) \\
&\quad \times \iiiint \{B(s_1)B^\top(t_1)\} \otimes \{B(s_2)B^\top(t_2)\} d_{s_1} F_i(s_1) d_{s_2} F_i(s_2) d_{t_1} F_i(t_1) d_{t_2} F_i(t_2).
\end{aligned}$$

As $v_i = \{nm_i(m_i-1)\}^{-1}$, it is easy to verify the following:

$$\begin{aligned} I(m_i \geq 2) v_i^2 m_i(m_i-1) &\leq n^{-2}, \\ I(m_i \geq 3) v_i^2 m_i(m_i-1)(m_i-2) &\leq n^{-2}, \\ I(m_i \geq 4) v_i^2 m_i(m_i-1)(m_i-2)(m_i-3) &\leq n^{-2}. \end{aligned}$$

Thus, we have $\Pi_{ij} \leq \widetilde{\Gamma}_{ij}$, where

$$\begin{aligned}
\widetilde{\Gamma}_{11} &= n^{-1} \iint \{B(t)B^\top(t)\} \otimes \{B(s)B^\top(s)\} d_s d_t \left\{ n^{-1} \sum_{i=1}^n F_i(s)F_i(t) \right\}, \\
\widetilde{\Gamma}_{12} &= n^{-1} \iint \{B(t)B^\top(s)\} \otimes \{B(s)B^\top(t)\} d_s d_t \left\{ n^{-1} \sum_{i=1}^n F_i(s)F_i(t) \right\}, \\
\Gamma_{21} &= n^{-1} \iiint \{B(t_1)B^\top(t_2)\} \otimes \{B(s)B^\top(s)\} d_s d_{t_1} d_{t_2} \left\{ n^{-1} \sum_{i=1}^n F_i(s)F_i(t_1)F_i(t_2) \right\}, \\
\Gamma_{22} &= n^{-1} \iiint \{B(t_1)B^\top(s)\} \otimes \{B(s)B^\top(t_2)\} d_s d_{t_1} d_{t_2} \left\{ n^{-1} \sum_{i=1}^n F_i(s)F_i(t_1)F_i(t_2) \right\}, \\
\widetilde{\Gamma}_{23} &= n^{-1} \iiint \{B(s)B^\top(s)\} \otimes \{B(t_1)B^\top(t_2)\} d_s d_{t_1} d_{t_2} \left\{ n^{-1} \sum_{i=1}^n F_i(s)F_i(t_1)F_i(t_2) \right\}, \\
\widetilde{\Gamma}_{24} &= n^{-1} \iiint \{B(s)B^\top(t_1)\} \otimes \{B(t_2)B^\top(s)\} d_s d_{t_1} d_{t_2} \left\{ n^{-1} \sum_{i=1}^n F_i(s)F_i(t_1)F_i(t_2) \right\}, \\
\widetilde{\Gamma}_{31} &= n^{-1} \iiiint \{B(s_1)B^\top(t_1)\} \otimes \{B(s_2)B^\top(t_2)\} d_{s_1} d_{s_2} d_{t_1} d_{t_2} \\
&\qquad \times \left\{ n^{-1} \sum_{i=1}^n F_i(s_1)F_i(s_2)F_i(t_1)F_i(t_2) \right\}.
\end{aligned}$$

Let

$$\widetilde{\Gamma} = \widetilde{\Gamma}_{11} + \widetilde{\Gamma}_{12} + \widetilde{\Gamma}_{21} + \widetilde{\Gamma}_{22} + \widetilde{\Gamma}_{23} + \widetilde{\Gamma}_{24} + \widetilde{\Gamma}_{31},$$

It follows by equation (B.12) that

$$\iint \mathrm{Var}\left\{\widehat{\sigma}(s,t)\right\} d_s d_t R(s,t) = O_{\mathbb{P}}\left[\mathrm{Tr}\left\{ \left(H^{-1} \otimes H^{-1}\right)_+ \widetilde{\Gamma} \right\} \right]. \tag{B.13}$$

Now we can define Γ_{ij} where the empirical CDFs in defining $\widetilde{\Gamma}_{ij}$ are replaced by their expectation with the censoring times. For example, $n^{-1}\sum_{i=1}^{n} F_i(s)F_i(t)$ in $\widetilde{\Gamma}_{11}$ is replaced by $\mathrm{E}\{F_1(s)F_1(t)\}$. We also define Γ, similar to $\widetilde{\Gamma}$. Then, we derive from (B.13) that

$$\iint \mathrm{Var}\left\{\widehat{\sigma}(s,t)\right\} d_s d_t R(s,t) = O_{\mathbb{P}}\left[\mathrm{Tr}\left\{\left(H^{-1}\otimes H^{-1}\right)_{+}\Gamma\right\}\right]. \tag{B.14}$$

By Assumption 3.8, those CDFs in defining Γ_{ij}s have partial derivatives that are bounded away from 0 and infinity. Then, (B.14) still holds when we replace the CDFs in defining Γ_{ij}s by the proper CDFs corresponding to uniform distribution in the unit cube. Then the rest of proof can be carried out similar to the that for equation (S.28) in the supplement of Xiao (2020) and hence is omitted.

Finally, by combining (B.6) and (B.14), we derive that

$$\iint \mathrm{E}\left\{\widehat{\sigma}(s,t)-\sigma(s,t)\right\}^2 d_s d_t R(s,t) = O(h^{2r}) + o(h^{2p}) + O(\lambda^2 h_e^{-2q}) + O_{\mathbb{P}}(nh_e^{-2}).$$

By Assumption 3.4, $\partial^2 R(s,t)/\partial s\partial t$ is bounded away from 0 and infinity, hence

$$\iint \mathrm{E}\left\{\widehat{\sigma}(s,t)-\sigma(s,t)\right\}^2 \mathrm{d}s\,\mathrm{d}t \simeq \iint \mathrm{E}\left\{\widehat{\sigma}(s,t)-\sigma(s,t)\right\}^2 d_s d_t R(s,t),$$

and the proof is complete.

In Lemmas B.5 - B.8, we assume Assumptions 3.1 - 3.8 always hold.

Lemma B.5. *Let* $\widetilde{\tau} = \sum_{i=1}^{n} m_i(m_i-1)v_i$ *with* $v_i = \{nm_i(m_i-1)\}^{-1}$ *and*

$$\widetilde{R}_n(s,t) = \widetilde{\tau}^{-1}\sum_{i=1}^{n} v_i \sum_{j_1\neq j_2} I(T_{ij_1}\le s)I(T_{ij_2}\le t)$$

be an empirical cumulative distribution function. Let $\widetilde{Z}(s,t) = \widetilde{R}_n(s,t) - R(s,t)$*, where* $R(s,t)$ *is defined in Lemma B.1. Then,* $\|\widetilde{Z}\| = o(h^2)$ *almost surely.*

Lemma B.6. *Let* $G_\sigma = \iint\{B(s)B^\top(s)\}\otimes\{B(t)B^\top(t)\}d_s d_t R(s,t)$*, where* $R(s,t)$ *is defined in Lemma B.1. Then,*

$$G_{\sigma,n} \simeq G_\sigma \simeq h^2 I, \text{ a.s.}$$

Proof. It can be shown that

$$G_{\sigma,n} = \iint \{B(s)B^\top(s)\} \otimes \{B(t)B^\top(t)\} d_s d_t \widetilde{R}_n(s,t),$$

where $\widetilde{R}_n(s,t)$ is defined in Lemma B.5. □

Lemma B.7. *Let $G = \int B(s)B^\top(s)d_sQ(s)$, where $Q(s)$ is defined in Lemma B.1. Also let $H = G + \lambda P$. Then, $G \simeq hI$ and $\|H\|_\infty = h_e^{-1}$.*

Proof. The lemma is just parts of Lemmas A.3 and A.6 in Xiao (2020). □

Lemma B.8. *Let H be as in Lemma B.7. Then,*

$$H_{\sigma,n}^{-2} \preceq h^{-2}H^{-1} \otimes H^{-1}, \text{ a.s.}$$

Proof. We derive that

$$H_{\sigma,n} = G_{\sigma,n} + P_\sigma \simeq h^2 I + \lambda h(I \otimes P + P \otimes I).$$

Thus,

$$\begin{aligned} H_{\sigma,n}^2 &\simeq \left\{h^2 I + \lambda h(I \otimes P + P \otimes I)\right\}\left\{h^2 I + \lambda h(I \otimes P + P \otimes I)\right\} \\ &\geq h^2(G + \lambda P) \otimes (G + \lambda P). \end{aligned}$$

It follows that

$$H_{\sigma,n}^{-2} \preceq h^{-2}H^{-1} \otimes H^{-1},$$

which proves the lemma. □

B.3 Proofs of Assumptions 3.4 and 3.8 for Random Censoring

We shall generate the censoring times C_i randomly according to the following cumulative distribution function

$$F_C(t) = \delta_1 I(t = 1) + \frac{1 - \delta_1}{1 - \delta_2}(t - \delta_2) I(\delta_2 \leq t \leq 1),$$

where δ_1 and δ_2 are two fixed constants in $(0, 1)$. The mixed censoring simulation setting in Section 4 is a special case in which $\delta_1 = \delta_2 = 0.2$. We derive that

$$\rho_1(t) = \delta_1 \Lambda_0^{(1)}(t) + \frac{1-\delta_1}{1-\delta_2} \int_{\delta_2}^1 \frac{\Lambda_0^{(1)}(t \wedge c)}{\Lambda_0(c)} dc,$$

which is continuous. Then we have $\rho_1(t) \geq \delta_1 \Lambda_0^{(1)}(t)$, which is bounded away from 0 because by Assumption 3.3, $\Lambda_0^{(1)}(t)$ is bounded away from 0. We also have

$$\begin{aligned}
\rho_1(t) &\leq \delta_1 \Lambda_0^{(1)}(t) + \frac{1-\delta_1}{1-\delta_2} \int_{\delta_2}^1 \frac{\Lambda_0^{(1)}(t \wedge c)}{\Lambda_0(c)} dc \\
&\leq \delta_1 \Lambda_0^{(1)}(t) + \frac{1-\delta_1}{1-\delta_2} \int_{\delta_2}^1 \frac{\|\Lambda_0^{(1)}\|}{\Lambda_0(\delta_2)} dc \\
&\leq \delta_1 \|\Lambda_0^{(1)}\| + (1-\delta_1) \frac{\|\Lambda_0^{(1)}\|}{\Lambda_0(\delta_2)}.
\end{aligned}$$

Hence, ρ_1 is also bounded away from infinity.

Next we consider ρ_2. We have

$$\rho_2(s,t) = \delta_1 \Lambda_0^{(1)}(s) \Lambda_0^{(1)}(t) + \frac{1-\delta_1}{1-\delta_2} \int_{\delta_2}^1 \frac{\Lambda_0^{(1)}(s \wedge x) \Lambda_0^{(1)}(t \wedge x)}{\Lambda_0^2(x)} dx$$

and the rest of the derivations are similar to the above.

Similarly we could show that ρ_3 and ρ_4 are also continuous and bounded away from 0 and infinity.

B.4 Additional Simulation Results

The mean function simulation results for all simulation settings are summarized in in Table B.1, and the covariance function simulation results are summarized in Table B.2. Figure B.1 and Figure B.2 plot the averages of the estimated mean functions from penalized splines with either OBS or SUBJ weights for all simulation settings with informative observations.

Table B.1 Median (interquartile range) of integrated squared error for estimating the mean function using penalized splines with both OBS and SUBJ weights.

			Gamma					
			No censoring		Mixed censoring		Uniform censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ	OBS	SUBJ
0.00	150	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.001 (0.001)	0.001 (0.001)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.001)	0.000 (0.001)
	300	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0.50	150	0.5	0.011 (0.017)	0.001 (0.002)	0.022 (0.021)	0.002 (0.002)	0.013 (0.067)	0.007 (0.016)
		1.0	0.059 (0.060)	0.002 (0.002)	0.068 (0.039)	0.004 (0.004)	0.082 (0.077)	0.006 (0.013)
	300	0.5	0.016 (0.008)	0.000 (0.000)	0.011 (0.007)	0.001 (0.002)	0.020 (0.019)	0.003 (0.009)
		1.0	0.069 (0.037)	0.001 (0.001)	0.073 (0.042)	0.003 (0.003)	0.059 (0.101)	0.006 (0.010)
0.85	150	0.5	0.171 (0.109)	0.003 (0.003)	0.102 (0.116)	0.007 (0.010)	0.123 (0.102)	0.021 (0.029)
		1.0	0.557 (0.467)	0.007 (0.009)	0.409 (0.343)	0.016 (0.038)	0.322 (0.706)	0.046 (0.081)
	300	0.5	0.141 (0.056)	0.001 (0.002)	0.102 (0.057)	0.005 (0.008)	0.120 (0.096)	0.014 (0.011)
		1.0	0.596 (0.184)	0.003 (0.003)	0.593 (0.297)	0.011 (0.014)	0.448 (0.528)	0.019 (0.025)

			Lognormal					
			No censoring		Mixed censoring		Uniform censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ	OBS	SUBJ
0.00	150	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.001)	0.000 (0.001)
	300	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.001)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0.50	150	0.5	0.016 (0.024)	0.001 (0.002)	0.012 (0.027)	0.002 (0.003)	0.023 (0.030)	0.008 (0.012)
		1.0	0.073 (0.120)	0.001 (0.002)	0.075 (0.081)	0.005 (0.005)	0.065 (0.192)	0.013 (0.019)
	300	0.5	0.017 (0.014)	0.000 (0.000)	0.017 (0.017)	0.001 (0.001)	0.019 (0.016)	0.002 (0.005)
		1.0	0.048 (0.061)	0.001 (0.001)	0.054 (0.078)	0.002 (0.003)	0.064 (0.081)	0.003 (0.009)
0.85	150	0.5	0.141 (0.048)	0.003 (0.004)	0.098 (0.137)	0.009 (0.008)	0.079 (0.164)	0.027 (0.023)
		1.0	0.477 (0.344)	0.006 (0.009)	0.297 (0.684)	0.014 (0.012)	0.422 (0.450)	0.041 (0.046)
	300	0.5	0.102 (0.052)	0.001 (0.002)	0.107 (0.091)	0.005 (0.005)	0.143 (0.150)	0.013 (0.014)
		1.0	0.579 (0.348)	0.003 (0.003)	0.375 (0.283)	0.005 (0.009)	0.420 (0.611)	0.010 (0.015)

Table B.2 Median (interquartile range) of integrated squared error for estimating the covariance function using penalized splines with both OBS and SUBJ weights.

			Gamma					
			No censoring		Mixed censoring		Uniform censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ	OBS	SUBJ
0.00	150	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	300	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0.50	150	0.5	0.018 (0.015)	0.000 (0.001)	0.011 (0.066)	0.001 (0.002)	0.005 (0.013)	0.003 (0.006)
		1.0	0.224 (0.653)	0.004 (0.009)	0.242 (1.141)	0.008 (0.014)	0.161 (0.451)	0.008 (0.020)
	300	0.5	0.009 (0.010)	0.000 (0.000)	0.006 (0.041)	0.001 (0.001)	0.011 (0.033)	0.001 (0.002)
		1.0	0.220 (0.472)	0.003 (0.004)	0.211 (0.789)	0.005 (0.005)	0.148 (0.567)	0.011 (0.017)
0.85	150	0.5	0.324 (0.581)	0.008 (0.014)	0.033 (0.226)	0.011 (0.018)	0.044 (0.400)	0.017 (0.014)
		1.0	5.359 (21.07)	0.043 (0.157)	5.568 (22.08)	0.077 (0.049)	0.407 (4.671)	0.129 (0.092)
	300	0.5	0.235 (0.399)	0.003 (0.003)	0.068 (0.577)	0.006 (0.013)	0.032 (0.129)	0.007 (0.019)
		1.0	3.022 (15.30)	0.045 (0.060)	2.556 (5.849)	0.054 (0.059)	2.950 (10.06)	0.070 (0.076)

			Lognormal					
			No censoring		Mixed censoring		Uniform censoring	
ρ	n	$\sigma^2_{Z^*}$	OBS	SUBJ	OBS	SUBJ	OBS	SUBJ
0.00	150	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
	300	0.5	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
		1.0	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
0.50	150	0.5	0.032 (0.109)	0.001 (0.002)	0.011 (0.160)	0.002 (0.004)	0.022 (0.251)	0.004 (0.006)
		1.0	2.885 (7.116)	0.011 (0.043)	0.341 (1.786)	0.012 (0.028)	0.801 (3.151)	0.027 (0.071)
	300	0.5	0.093 (0.113)	0.001 (0.002)	0.043 (0.088)	0.001 (0.002)	0.025 (0.124)	0.003 (0.005)
		1.0	0.614 (2.761)	0.005 (0.006)	0.747 (12.21)	0.010 (0.021)	0.778 (9.974)	0.013 (0.019)
0.85	150	0.5	0.691 (1.745)	0.007 (0.011)	0.652 (5.084)	0.031 (0.042)	0.156 (0.257)	0.032 (0.026)
		1.0	12.37 (74.19)	0.034 (0.106)	9.179 (96.96)	0.176 (0.238)	9.697 (29.30)	0.137 (0.254)
	300	0.5	0.550 (1.728)	0.004 (0.005)	0.562 (2.247)	0.009 (0.021)	0.941 (3.214)	0.024 (0.017)
		1.0	32.76 (93.44)	0.076 (0.133)	10.48 (129.7)	0.060 (0.078)	15.32 (71.81)	0.124 (0.244)

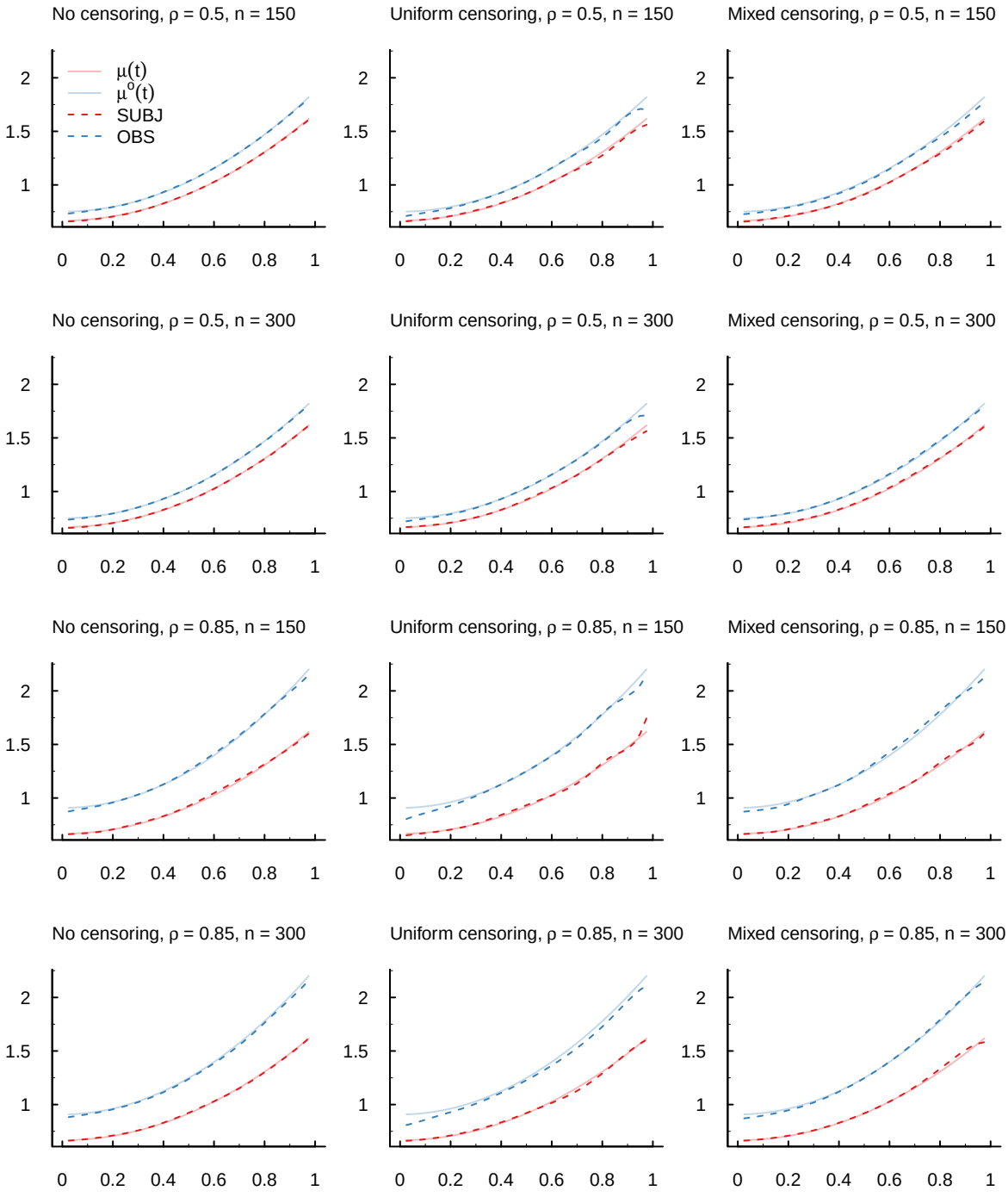

Figure B.1 Average of 200 estimates of the mean function using penalized splines with both OBS and SUBJ weights for simulation settings in which the frailty variable Z_i^* follows a gamma distribution. The true mean function is $\mu(t) = t^2 + 2/3$ and $\mu^o(t)$ is the formula in (4) associated with equal weights, specifically $\mu^o(t) = (t^2 + 2/3)(1 + \rho^2\sigma^2_{Z^*})$.

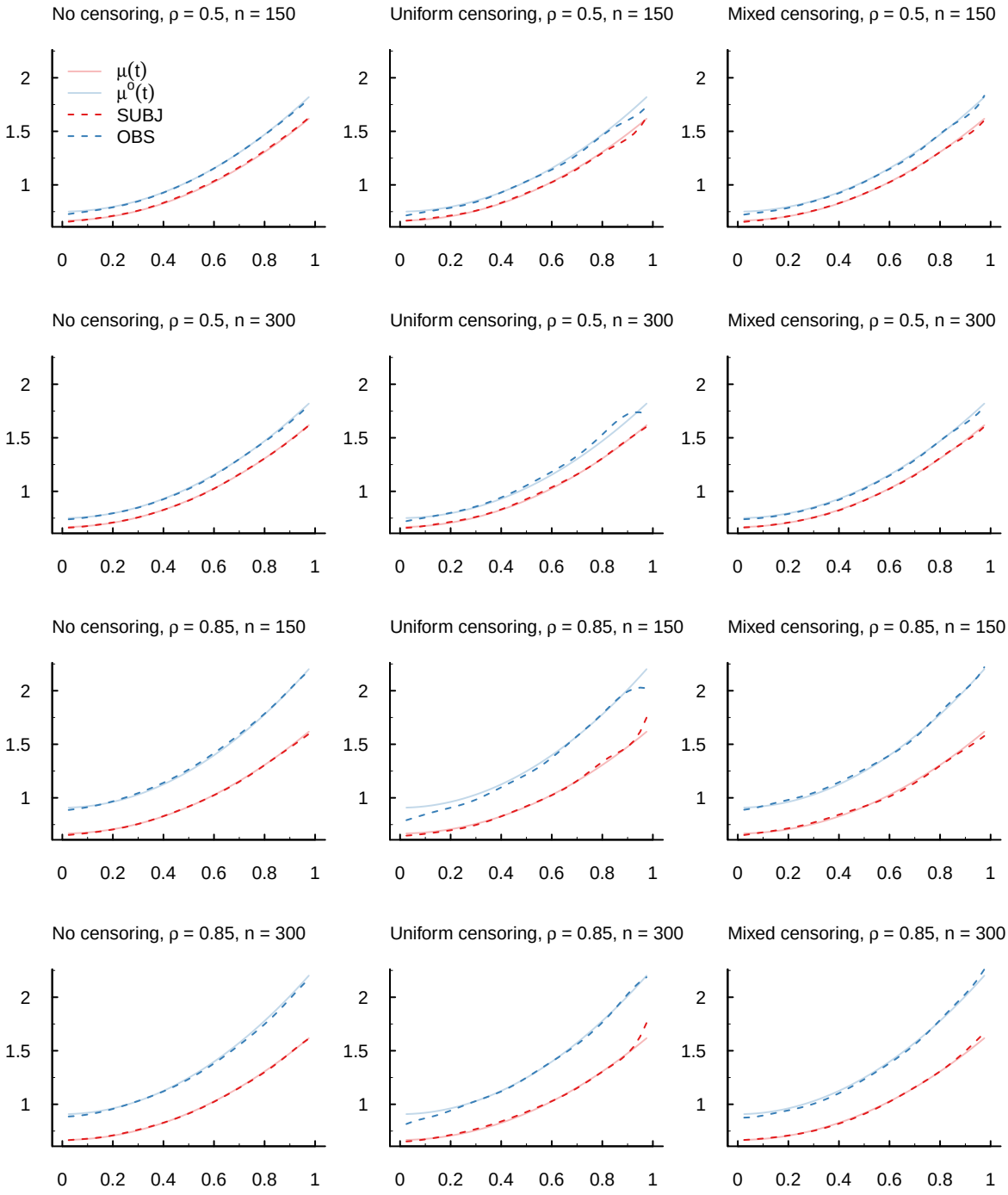

Figure B.2 Average of 200 estimates of the mean function using penalized splines with both OBS and SUBJ weights for simulation settings in which the frailty variable Z_i^* follows a lognormal distribution. The true mean function is $\mu(t) = t^2 + 2/3$ and $\mu^o(t)$ is the formula in (4) associated with equal weights, specifically $\mu^o(t) = (t^2 + 2/3)(1 + \rho^2\sigma^2_{Z^*})$.

Appendix C

Supplementary Material of Chapter 4

C.1 Derivation of Updates

In this section, we derive the update for the primal variable $\mathbf{a}$ in the iterative algorithm described in Section 4.2.

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{a}} \Bigg\{ \log|\mathbf{V}| + (\mathbf{y}-\mathbf{X}\mathbf{a})^\top \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{a}) + \frac{\rho}{2}\sum_{i<i'} \|\mathbf{D}_{ii'}\mathbf{a} - \mathbf{b}_{ii'} + \boldsymbol{\delta}_{ii'}\|_2^2 \\
&\qquad + \frac{\rho}{2}\sum_{j<j'} \left\|\widetilde{\mathbf{D}}_{jj'}\mathbf{a} - \mathbf{c}_{jj'} + \boldsymbol{\eta}_{jj'}\right\|_2^2 \Bigg\} \\
&= -2\mathbf{X}^\top \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\mathbf{a}) + \rho \sum_{i<i} \left\{ \mathbf{D}_{ii'}^\top \left(\mathbf{D}_{ii'}\mathbf{a} - \mathbf{b}_{ii'} + \boldsymbol{\delta}_{ii'}\right)\right\} \\
&\qquad + \rho \sum_{j<j'} \left\{ \widetilde{\mathbf{D}}_{jj'}^\top \left(\widetilde{\mathbf{D}}_{jj'}\mathbf{a} - \mathbf{c}_{jj'} + \boldsymbol{\eta}_{jj'}\right)\right\}.
\end{aligned}$$

Setting equal to zero and rearranging terms, we have that

$$\begin{aligned}
&\left\{ \rho \left(\sum_{i<i'} \mathbf{D}_{ii'}^\top \mathbf{D}_{ii'} + \sum_{j<j'} \widetilde{\mathbf{D}}_{jj'}^\top \widetilde{\mathbf{D}}_{jj'} \right) + \mathbf{X}^\top \mathbf{V}^{-1}\mathbf{X} \right\} \mathbf{a} \\
&= \mathbf{X}^\top \mathbf{V}^{-1}\mathbf{y} + \rho \Bigg[\sum_{i<i'} \left\{ \mathbf{D}_{ii'}^\top \left(\mathbf{b}_{ii'} - \boldsymbol{\delta}_{ii'}\right)\right\} + \sum_{j<j'} \left\{ \widetilde{\mathbf{D}}_{jj'}^\top \left(\mathbf{c}_{jj'} - \boldsymbol{\eta}_{jj'}\right)\right\} \Bigg].
\end{aligned}$$

Let $\boldsymbol{\Delta} = \sum_{i<i'} \left\{\mathbf{D}_{ii'}^\top \mathbf{D}_{ii'}\right\} + \sum_{j<j'} \left\{\widetilde{\mathbf{D}}_{jj'}^\top \widetilde{\mathbf{D}}_{jj'}\right\}$. Similarly, let $\mathbf{d} = \sum_{i<i'} \left\{\mathbf{D}_{ii'}^\top \left(\mathbf{b}_{ii'} - \boldsymbol{\delta}_{ii'}\right)\right\} + \sum_{j<j'} \left\{\widetilde{\mathbf{D}}_{jj'}^\top \left(\mathbf{c}_{jj'} - \boldsymbol{\eta}_{jj'}\right)\right\}$. Then

$$\begin{aligned}(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} + \rho \boldsymbol{\Delta})\mathbf{a} &= \mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} + \rho \mathbf{d} \\ \mathbf{a} &= \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X} + \rho \boldsymbol{\Delta}\right)^{-1} \left(\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{y} + \rho \mathbf{d}\right),\end{aligned}$$

which gives the solution to the update of $\mathbf{a}$.

C.2 Additional Data Application: Patient-reported Symptom Severity of Parkinson's Disease

In this section, we conduct a longitudinal biclustering analysis of the Parkinson's symptom severity dataset analyzed in Section 3. In the previous analysis, we considered the univariate PRO-PD score, which is the sum of severity scores of 33 symptoms of Parkinson's disease. 18 of these symptoms are *non-motor* symptoms, including fatigue and insomnia; the other symptoms are *motor* symptoms, including tremor and loss of balance. A multivariate longitudinal analysis of this data set may lead to a deeper understanding of disease trajectories.

Based on an inspection of the empirical correlation matrix of the 33 features illustrated in Figure C.1, we assume a compound symmetric correlation structure of the random intercept coefficients. The estimated correlation matrix based on the imposed structure is also shown in Figure C.1. The common correlation coefficient estimate is 0.41.

The coefficients and biclustering structure estimated by the proposed method are illustrated in Figure C.2. Four subgroups of subjects are apparent. The first subgroup ($n = 66$) is characterized by subjects with rapidly increasing symptom severities. The subjects in the second subgroup ($n = 106$) demonstrate a milder increase in most severities when compared to the first subgroup, but is unique in its rapid progression of one symptom: sexual dysfunction, a symptom well-studied in existing Parkinson's literature (Bronner et al., 2004). Five subgroups of symptom severities are also apparent; it is notable that motor and non-motor symptoms tend to cluster together.

It is interesting that the first subgroup of subjects is later on in their disease when compared to the second subgroup; specifically, the mean (standard error) of years since

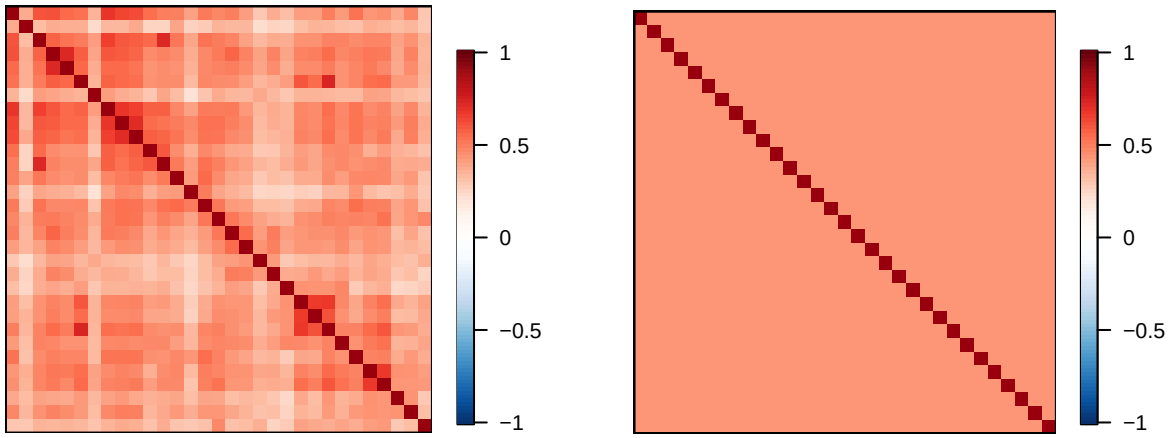

Figure C.1 (Left) Empirical correlation matrix of the 33 symptom severity scores. (Right) Estimated correlation matrix of the random intercepts in the fitted model based on imposed structure.

diagnosis is 6.9 (0.29) and 6.3 (0.23) for the first and second subgroups, respectively. These results agree with existing literature that found motor symptoms of Parkinson's Disease become progressively worse as the disease advances (Xia & Mao, 2012). Our results may suggest that rapid progression of sexual dysfunction in early stages of Parkinson's disease is unique when compared to other symptoms.

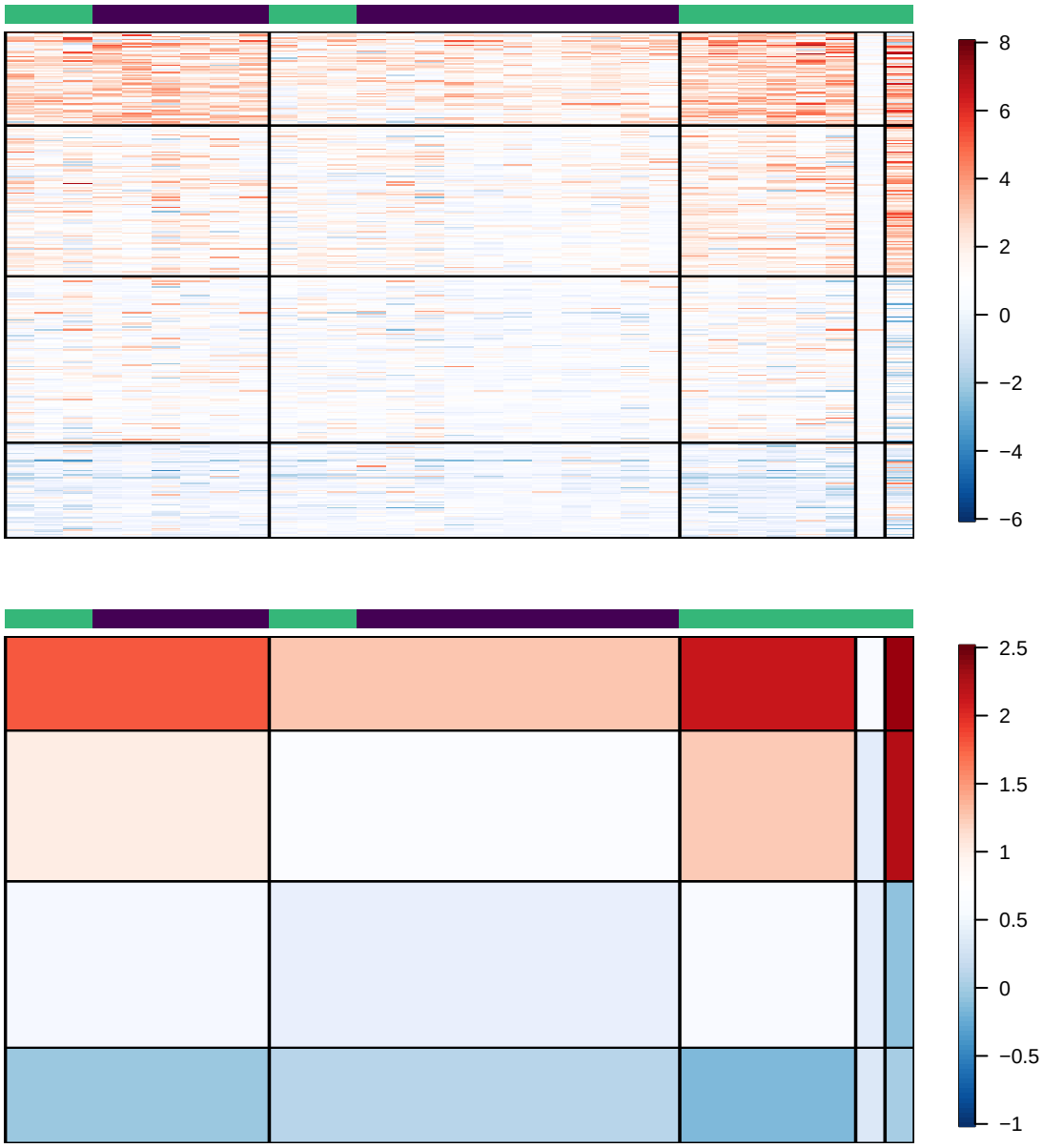

Figure C.2 (Top) Estimated slope coefficients of the symptom severity trajectories arranged according to the estimated biclustering structure. Symptoms are indicated as motor or non-motor in the upper color bar (motor: green, non-motor: blue). The estimated biclustering structure is indicated with grid lines. (Bottom) Mean matrix of the estimated slope coefficients.