Data Fusion and Systems Engineering Approaches for Quality and Performance Improvement of Health Care Systems: From Diagnosis to Care to System-level Decision-making

by

Bing Si

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved March 2018 by the
Graduate Supervisory Committee:

Jing Li, Chair
Douglas Montgomery
Todd Schwedt
Teresa Wu

ARIZONA STATE UNIVERSITY

May 2018

ABSTRACT

Technology advancements in diagnostic imaging, smart sensing, and health information systems have resulted in a data-rich environment in health care, which offers a great opportunity for Precision Medicine. The objective of my research is to develop data fusion and system informatics approaches for quality and performance improvement of health care. In my dissertation, I focus on three emerging problems in health care and develop novel statistical models and machine learning algorithms to tackle these problems from diagnosis to care to system-level decision-making.

The first topic is diagnosis/subtyping of migraine to customize effective treatment to different subtypes of patients. Existing clinical definitions of subtypes use somewhat arbitrary boundaries primarily based on patient self-reported symptoms, which are subjective and error-prone. My research develops a novel Multimodality Factor Mixture Model that discovers subtypes of migraine from multimodality imaging MRI data, which provides complementary accurate measurements of the disease. Patients in the different subtypes show significantly different clinical characteristics of the disease. Treatment tailored and optimized for patients of the same subtype paves the road toward Precision Medicine.

The second topic focuses on coordinated patient care. Care coordination between nurses and with other health care team members is important for providing high-quality and efficient care to patients. The recently developed Nurse Care Coordination Instrument (NCCI) is the first of its kind that enables large-scale quantitative data to be collected. My research develops a novel Multi-response Multi-level Model (M3) that enables transfer learning in NCCI data fusion. M3 identifies key factors that contribute to improving care coordination, and facilitates the design and optimization of nurses' training, workload assignment, and practice environment, which leads to improved patient outcomes.

The last topic is about system-level decision-making for Alzheimer's disease early detection at the early stage of Mild Cognitive Impairment (MCI), by predicting each MCI patient's risk of converting to AD using imaging and proteomic biomarkers. My research proposes a systems engineering approach that integrates the multi-perspectives, including prediction accuracy, biomarker cost/availability, patient heterogeneity and diagnostic efficiency, and allows for system-wide optimized decision regarding the biomarker testing process for prediction of MCI conversion.

DEDICATION

To my family, who have always been supportive.

ACKNOWLEDGMENTS

I am deeply grateful to my advisor, Dr. Jing Li, for her mentorship over the years. I sincerely appreciate her guidance of both academic and life advice, which has made this endeavor a reality, and will keep inspiring me in my future career.

Second, I want to acknowledge and thank my committee member, Dr. Douglas Montgomery, who offered me great guidance and suggestions for my Ph.D. research.

Third, I would like to thank Dr. Todd Schwedt for being a very encouraging and supportive mentor throughout the years. His professional and hard-working spirit will keep encouraging me in my future career development.

In addition, I am deeply grateful to Dr. Teresa Wu for her great help and valuable suggestions in my PhD career at ASU-Mayo Imaging Informatics Laboratory (AMIIL).

Finally, I would like to acknowledge and thank all of the faculty, staff, and students at AMIIL who have assisted, advised, and supported my research.

TABLE OF CONTENTS

APPENDIX Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1
INTRODUCTION

1.1 Background

Technology advancements in diagnostic imaging, smart sensing, and health information systems have resulted in a data-rich environment in health care. It is now possible to track every piece of information related to a patient's diagnosis, treatment, and care. This offers a great opportunity for Personalized Medicine (PM), i.e., to offer the right medical decision-making to the right person at the right time. On the other hand, the size and complexity of the data overwhelms the modeling capability of existing statistical methods.

The objective of my research is to develop data fusion and system informatics approaches for quality and performance improvement of healthcare system from diagnosis to care to system-level decision-making. In my dissertation, I focus on three emerging problems in health care and develop novel statistical models driven by the unique data structure and objectives of the specific problem domains. The three topics are (I) multimodality imaging data fusion and novel latent variable models for subtype discovery of migraine, (II) multi-source multi-level system-wide data fusion and novel transfer learning models for improving nurse care coordination and patient outcomes, and (III) Systems engineering approach for biomarker testing process optimization and Alzheimer's disease early intervention.

The three topics cover a full spectrum of decision makings in health care ranging from diagnosis (I), to care (II), and to process optimization (III). The objective of my dissertation research is not only to provide solutions to each of the three individual problems, but also to demonstrate that advanced statistical and machine learning development integrated with domain knowledge through collaboration with medical professionals shows great promise for tackling challenging issues at different levels of the complex health care system.

1.2 Summary of Research Topics and State of the Art

Topic (I): Multimodality imaging data fusion and novel latent variable models for subtype discovery of migraine. Migraine is a neurological disease that ranks in the top 20 of the world's most disabling medical illnesses. Over 10% of the population suffers from migraine and nearly 1 in 4 U.S. households includes someone with migraine. Treatment of migraine has not achieved much success because of not being tailored to different subtypes of the disease. Current clinical definitions of subtypes use somewhat arbitrary boundaries primarily based on patient self-reported symptoms, which are subjective and error-prone. Diagnostic structural MRI provides complementary, accurate multimodality measurements of the disease (Nan et al., 2013; Ung et al., 2012; Sundermann et al., 2014; Schwedt et al., 2015; Chong et al., 2016). However, the existing imaging-based migraine research is supervised, i.e., it aims to find structural imaging biomarkers to differentiate migraine from healthy controls. However, it is known that there is substantial heterogeneity among patients with migraine. Also, even for patients that are diagnosed as having migraine, the clinical diagnostic criteria are symptom-based and use somewhat arbitrary boundaries developed by expert consensus. As a result, it is possible that patients with different outcomes or prognostications are lumped together. This inability to delineate patient heterogeneity leaves clinicians with inadequate information for early determination of the most appropriate, personalized treatment strategy (i.e. more aggressive therapy vs. conservative therapy), and prevents them from accurately predicting functional outcomes for individual patients. To address this limitation of the existing research, I propose a multimodality factor mixture model (MFMM) for migraine subtype discovery.

Topic (II): Multi-source multi-level data fusion and novel statistical models for improving nurse care coordination. Care coordination has been found to be instrumental for decreasing adverse events, improving quality and efficiency of care, and enhancing patient satisfaction (McDonald et al., 2007; Sticker et al., 2009). In coordinating patient care within the hospital, staff nurses, as the patient's "ever-present" health care team members, play a vital role. In "Keeping Patients Safe", a recent report by the Institute of Medicine, the role of staff nurses in care coordination that promotes patient safety and quality outcomes was highlighted. Recent qualitative studies illuminated the considerable amount of time staff nurses spend coordinating patient care via a broad range of activities from admission to discharge (Hendrich et al., 2008; Lamb et al., 2008; Aiken et al., 2008; Friese, 2008; Kazanjian et al. 2005; Laschinger et al. 2006). However, little research is available to reveal the relationship between the care coordination activities conducted by nurses and their demographics and workload as well as the characteristics of their practice environment. Such research is important for nursing process improvement and designing of the best practices. The recently developed Nurse Care Coordination Instrument (NCCI) is the first of its kind that enables quantitative data to be collected to measure various aspects of nurse care coordination (Lamb et al., 2008). Driven by this new development, we propose a multi-response multilevel model with joint fixed effect selection and joint random effect selection across multiple responses. The proposed model is a combination of conventional multilevel models and modern variable selection techniques. Various variable selection techniques have been proposed in recent years, including lasso (Tibshirani, 1996), group lasso (Yuan et al., 2006), CAP (Zhao et al., 2009), just to name a few. However, these methods are for single-level predictors; there is much less research in the multilevel setting. There are a few existing efforts in introducing variable selection in the multilevel setting (Schelldorfer et al., 2011; Ibrahim et al., 2011; Bondell et al., 2010; Ahn et al., 2012). However, these methods are for a single response only. In all, there is a lack of research in multi-response multilevel models with variable selection in the existing literature, which motives my research development.

Topic (III): Imaging biomarker testing process optimization for prognostics of Mild Cognitive Impairment (MCI) conversion to AD. Important to early detection and prevention of AD is the use of biomarkers to precisely predict the conversion of MCI to AD within a clinical time of interest. According to the new diagnostic guidelines recommended by the National Institute on Aging and the Alzheimer's Association (1), the important biomarkers include those measuring Aβ deposition in plagues and those linked to downstream neuronal degeneration or injury processes, such as the phosphorylated tau (p-tau) level in cerebrospinal fluid (CSF), mean cerebral metabolism on ^{18}F fludeoxyglucose positron emission tomography (FDG-PET), and hippocampal volume on structural magnetic resonance imaging (MRI). There has been a vast amount of studies aiming at using biomarker data to predict the conversion of MCI patients to AD (Borroni et al., 2006; Davatzikos et al., 2011; Hinrichs et al., 2011; Jack et al., 2005; Llano et al., 2011; Misra et al., 2009; Stoub et al., 2004; Tondelli et al., 2012; Westman et al., 2011). A particular area of study with clear clinical relevance is to achieve this prediction using *baseline* biomarker measurements (Davatzikos et al., 2011; Hinrichs et al., 2011; Llano et al., 2011; Misra et al., 2009). Although using longitudinal repeated measurements of the same biomarkers has a potential to improve the prediction accuracy, this prolongs the diagnostic time span and makes clinical trials more time consuming and costly. In using baseline biomarkers to predict MCI conversion, most of the existing studies built statistical classification models that assign each MCI patient to be a converter or non-converter using a pre-trained model. The accuracy on large public datasets like the Alzheimer's Disease Neuroimaging Initiative (ADNI) has been reported to be between 60-72%. The existing studies have several limitations, including unsatisfactory accuracy due to MCI heterogeneity, use of conventional classification models that require biomarkers to be measured all at once instead of sequentially and as-needed, and use of raw numerical measurement of the biomarkers instead of discretized levels that are more robust to measurement errors and provide convenience for clinical utilization. To tackle these limitations, we propose a novel sequence tree-based classifier (STC) for predicting the conversion of MCI to AD.

1.3 Expected Original Contribution

The expected original contributions include:

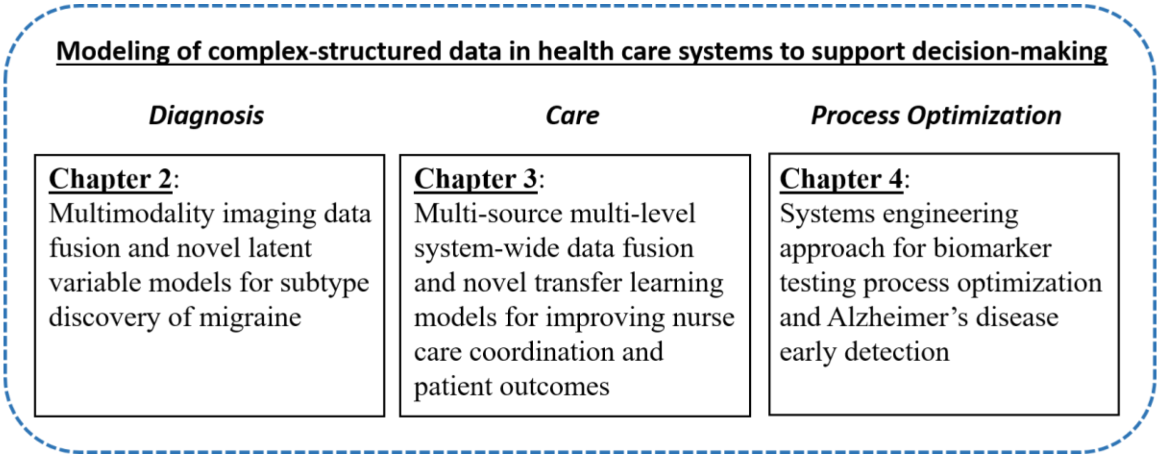
- We propose a multimodality factor mixture model (MFMM) for migraine subtype discovery. MFMM adopts a latent variable formulation that assumes there are latent variables of a much lower dimension underlying the observed features of each modality and joins the modality-wise latent variables in a unified framework for identifying the cluster structure of a patient cohort. MFMM employs a novel double-L_{21}-penalized likelihood formulation to achieve hierarchical selection of informative imaging modes and features. This formulation is proven to satisfy a Quadratic Majorization (QM) condition that allows for an efficient Group-wise Majorization Descent (GMD) algorithm to be developed for model estimation. Simulation studies are performed and show significantly better performance of MFMM than competing methods. MFMM is applied to migraine subtype discovery based on brain cortical area, cortical thickness, and volume measurements from structural magnetic resonance imaging (MRI). Two migraine subtypes are found, whose subjects significantly differ in clinical characteristics. This finding shows promise of using imaging data to help with patient stratification and with development of biomarkers for personalized management of migraine.

- We propose a novel multi-response multilevel model with joint fixed effect selection and joint random effect selection across multiple responses to reveal the relationship between the care coordination activities conducted by nurses and their demographics and workload as well as the characteristics of their practice environment. This model is particularly suitable for modeling the unique data structure of the NCCI due to its ability of jointly modeling of multilevel predictors, including demographic and workload variables at the individual/nurse level and characteristics of the practice environment at the unit level, and multiple response variables that measure the key components of nurse care coordination. We develop a Block Coordinate Descent (BCD) algorithm integrated with an Expectation-Maximization (EM) framework for model estimation, and perform theoretical analysis to reveal the reason why the proposed model is able to outperform the existing multilevel method that models each response variable in separation. Asymptotic properties of the proposed model are derived. Simulation studies are performed, showing that the proposed model outperforms competing methods. We apply the proposed model to a dataset collected across four U.S. hospitals using the NCCI. Our model achieves a significantly higher prediction accuracy compared with competing methods and also facilitates knowledge discovery.

- We propose a novel sequence tree-based classifier (STC) for predicting the conversion of MCI to AD. Different from conventional classification models, STC achieves a sequential, as-needed use of biomarkers and a three-category classification (high-risk converter, low-risk converter, and inconclusive diagnosis) by finding an optimal sequence of biomarkers and two-sided cutoffs of each biomarker that satisfy pre-specified accuracy requirements while minimizing the proportion of inconclusive diagnosis. STC is also a personalized approach as it allows patient characteristic variables to be included to help identify patient-specific cutoffs for each biomarker. We apply STC to two important clinical applications using the data from the worldwide Alzheimer's Disease Neuroimaging Initiative (ADNI) project: prediction of MCI conversion and patient selection for AD-related clinical trials. In the first application, STC achieves high prediction accuracy. It also allows multiple criteria, e.g., accuracy and efficiency, to be optimized using a Pareto optimal frontier. Compared with the conventional decision tree classifier, STC achieved higher PPV and NPV, saved biomarker testing costs and patient waiting time, facilitated timely medical decision making, and produced a model that is consistent with medical knowledge and biological principles and thus being clinically more trustworthy. In the other application, STC is able to identify a sub-cohort of MCI subjects with a high risk to convert to AD. The sub-cohort has a reasonable size appropriate for clinical trials.

1.4 Dissertation Organization

The proposed dissertation research will be presented in three chapters, followed by the conclusion in Chapter 5, as shown in Figure 1. Chapter 2 presents the development of topic (I): multimodality imaging data fusion and novel latent variable models for migraine subtype discovery. Chapter 3 presents the development of topic (II): multi-source multi-level data fusion and novel statistical models for improving nurse care coordination. Chapter 4 presents the development of topic (III): imaging biomarker testing process optimization for prognostics of MCI conversion to AD. Chapter 5 summarizes the dissertation with conclusion remarks and discussions on future work.



Modeling of complex-structured data in health care systems to support decision-making
Diagnosis
Care
Process Optimization
Chapter 2: Multimodality imaging data fusion and novel latent variable models for subtype discovery of migraine
Chapter 3: Multi-source multi-level system-wide data fusion and novel transfer learning models for improving nurse care coordination and patient outcomes
Chapter 4: Systems engineering approach for biomarker testing process optimization and Alzheimer's disease early detection

Figure 1 Dissertation framework

CHAPTER 2

A MULTI-MODE FACTOR MIXTURE MODEL WITH HIERARCHICALLY-STRUCTURED SPARSITY FOR IMAGING-BASED MIGRAINE SUBTYPE DISCOVERY

2.1 Introduction

Medical imaging technology has revolutionized health care over the past 30 years by greatly facilitating screening, early diagnosis, treatment planning, evaluation of response to therapy, and prognosis. That is why the New England Journal of Medicine ranked imaging as one of the top medical developments over the past 1,000 years. With the rapid advances in imaging technology, it is now possible to acquire imaging of different modes for the same patient. These modes consist of complementary information about the organ of interest, thus enabling better medical decision making.

Medicine is experiencing a paradigm shift toward precision medicine (PM), a shift that facilitates individualized patient evaluation and administration of precise treatment to the right patient at the right time. Imaging data from multiple modes plays a pivotal role in PM. For PM to succeed, it is critically important that subtle differences amongst patients with a disease be identified, especially if those differences are associated with prognoses and treatment responses. Multi-mode imaging data are likely to contribute to this subgroup/ subtype discovery (Giardino et al. 2017).

Subtypes exist for almost every complicated disease. Clinical definitions of subtypes are typically determined using somewhat arbitrary boundaries developed by expert consensus. These definitions, however, are generally inadequate for explaining the considerable heterogeneity among patients in terms of prognosis and response to treatment. For example, migraine is a neurological disease that ranks in the top 20 of the world's most disabling medical illnesses. Approximately 12% of the population suffers from migraine and nearly 1 in 4 U.S. households includes someone with migraine. The current clinical subtype classification of migraine is based on the International Classification of Headache Disorders 3 beta (ICHD-3 beta) criterion, according to which migraine is sub-classified into episodic vs. chronic migraine based on headache frequency and into migraine with aura vs. migraine without aura. Although this subtype classification can explain the patient heterogeneity to some extent, a large amount of the heterogeneity is left unexplainable, i.e., the patients within the same subtype can still have significantly different disease course, prognosis, and response to treatment. This is a strong indication that there may be undiscovered subtypes. Similar frustration exists for other diseases especially those that are either extremely fatal or currently lack effective disease management strategies for individual patients, such as Alzheimer's disease (Komarova et al. 2011), Parkinson's disease, (van Rooden et al. 2010), and Type-II diabetes (Li et al. 2015). If subtypes of these diseases could be more accurately identified, a focus on homogeneous groups would enhance the likelihood of success for understanding the underlying disease mechanisms and lead to tailored treatment strategies. This would pave the road toward PM in which medical care is designed to optimize diagnosis, prognosis and therapeutic benefit for each particular group of patients or even individual patients.

The focus of this research is to develop a data-driven method for subtype identification using imaging data of multiple modes. A method like this belongs to the general category of clustering or unsupervised learning methods in statistics. However, there are multifold challenges in developing a clustering method appropriate for our specific focus on multi-mode data: 1) There can be quite a few modes of data used in a study and some of them may be uninformative to the differentiation of subtypes. These modes should be automatically selected out such that they will not mask the underlying clustering structure. We would like to point out that the proposed method can be easily extendable to including non-imaging data of different modes such as patient demographics, disease history, clinical symptoms, and genetic signatures. This would lead to a more comprehensive discovery of subtypes, but with a greater likelihood for including uninformative/noise modes. This makes the ability of mode selection critically important. 2) Within each mode, it is commonplace that there are many features and some of them may be uninformative to the differentiation of subtypes. These features should be selected out.

To address these challenges, we propose a novel clustering method called Multi-mode Factor Mixture Model (MFMM) that enables an automatic, hierarchical selection of informative imaging modes and features. Here, "hierarchical selection" means that if a mode is uninformative, all the features it includes should be excluded; feature selection happens in the imaging modes that remain. This research contributes to both statistics and medicine:

- Contribution to statistics: MFMM intersects with several existing research areas in statistics, such as sparse learning, model-based clustering, and factor models, but none of these areas has a method that is capable of clustering data from multiple modes with hierarchical mode and feature selection. In this sense, MFMM is the first of its kind.

Specifically, we propose a novel double-L_{21}-penalized likelihood formulation for MFMM to achieve mode and feature selection. We prove that this formulation satisfies a Quadratic Majorization (QM) condition such that an efficient Group-wise Majorization Descent (GMD) algorithm can be developed to estimate the MFMM, which greatly speeds up the Expectation-Maximization (EM) iterations.

- Contribution to medicine: We applied MFMM to identification of potential subtypes of migraine by clustering subjects using their brain cortical area, cortical thickness, and volume measurements (treated as three modes) from structural Magnetic Resonance Imaging (MRI). Data were obtained from two medical institutions, Mayo Clinic at Arizona (MCA) and Washington University School of Medicine in St. Louis (WashU). MFMM found two clusters that are very well separated, indicating that subjects in the clusters have distinct imaging phenotypes. The imaging features selected by MFMM to produce the clustering result are well-documented in the literature to relate to migraine. Interestingly, we found that the two clusters also significantly differ in terms of clinical characteristics, with one cluster having more allodynia symptoms during migraine attacks, more migraine-related disability, and a greater number of years with migraine. In essence, this study contributes to understanding of migraine heterogeneity from an imaging perspective. The finding that the identified imaging subtypes were associated with distinct clinical characteristics shows promise of using imaging subtypes to help stratify patients and to serve as biomarkers for personalized management of migraine.

The rest of this paper is organized as follows: Section 2.2 provides a literature review. Section 2.3 presents the development of MFMM. Section 2.4 shows simulation experiments. Section 2.5 presents the migraine application. Section 2.6 is conclusion.

2.2 Literature Review

The proposed method intersects with a number of existing research areas. Next, we will review them one by one and point out their limitations, which highlights the need for new methodological development.

Sparse learning (SL): SL models (a.k.a. variable selection techniques) started to emerge a few decades ago, driven by the technological improvement on human genomic sequencing that produced high-dimensional genomic data, with the classic Least Absolute Shrinkage and Selection Operator (LASSO) model developed by Tibshirani in 1996 (Tibshirani 1996). The basic idea of LASSO is a L_1-penalized least squares method that results in the estimates of many irrelevant regression coefficients to be exactly zero. The following years have witnessed a booming development of SL models with different structural considerations or/and statistical properties, such as adaptive LASSO (Zou 2006), SCAD (Fan et al. 2001), elastic net (Zou et al 2005), group LASSO (Yuan et al. 2006), fused LASSO (Tibshirani et al. 2005), tree-guided LASSO (Kim et al. 2012), just to name a few. However, all these existing models are supervised learning methods, i.e., they aim to predict a response variable while our focus here on subtype discovery requires an unsupervised/clustering method.

Model-based clustering (MBC): Clustering analysis is a classic research area in statistical modeling and machine learning. Clustering methods generally fall into two categories: algorithm-based and model-based methods. The first category includes many methods such as hierarchical clustering, k-means, and more recently developed DBSCAN (Ram et al. 2010) and OPTICS (Ankerst et al. 1999) algorithms for big data. These algorithms are largely heuristic and not based on formal models. This is not necessarily a disadvantage since clustering, by nature, is exploratory. On the other hand, MBC methods are based on formal models, making it possible to study statistical properties and perform inferences (Melnykov et al. 2010). MBC methods assume that sample observations arise from a distribution that is a mixture of several components (i.e., the clusters) and each component can be described by a probability density function and has an associated probability or weight in the mixture. In principle, any probability model for the components can be adopted, while a multivariate Gaussian distribution is the most common.

In recent years, sparse learning has been introduced into MBC. The basic idea of sparse MBC is to maximize the log-likelihood function of the mixture model subject to a penalty that is chosen to yield sparsity in the features. For example, Pan and Shen (Pan et al. 2007) assumed that the cluster-wise covariance matrix was diagonal and the same, and imposed a L_1-penalty on the cluster-wise mean vectors. Xie et al. (Xie et al. 2008) proposed a more general approach that allowed for cluster-specific diagonal covariance matrices and penalized the variances together with the means. Wang and Zhu (Wang et al. 2008) proposed two models that allowed for the cluster-specific mean parameters associated with the same feature to be penalized as a group. Raftery and Dean (Raftery et al. 2006) recast the feature selection problem as a model selection problem by comparing models containing nested subsets of features and making sure the nested models are sparse in features. Witten and Tibshirani (Witten et al. 2010) developed a general framework for feature selection in MBC, and showed that k-means and hierarchical clustering can be represented as special cases of this framework with sparsity constraints. This work bridged algorithm-based and model-based clustering methods.

For clustering of high-dimensional datasets, one approach is the above-reviewed sparse MBC. An alternative approach assumes that the observed high-dimensional features lie on a low-dimensional latent space, which is the idea of factor models. For clustering of imaging data, factor models are more appropriate than sparse MBC, because imaging features typically embrace a complex correlation structure, suggesting the existence of latent factors. For example, the imaging features used in our migraine application correspond to anatomically defined brain regions that are structurally and functionally related.

Factor models: Earlier research adopts a two-step strategy in which a dimension reduction method such as Principal Component Analysis (PCA) and Correspondence Analysis (CA) is first used and clustering is then performed on the reduced space. However, treating dimension reduction and clustering as two separate steps may destroy the cluster structure in the data, as pointed out by Raftery (Raftery et al. 2006). More recent research developed the so-called factor mixture models (FMM). FMM is an extension of the classic factor analysis (FA). FA assumes that the sample observations are from a single distribution and aims at discovering the latent factors underlying the observed features. In contrast, FMM assumes that the factors are distributed as a mixture model and therefore represents a clustering approach. Different FMM models were developed based on different assumptions on the mixture distribution (Lubke et al. 2005, Muthen et al. 2006, Montanari et al. 2010, Baek et al. 2010): Some assume only varying component means; some additionally assume the component covariance matrices to be different. Muthen and Lubke (Lubke et al. 2005) presented different strategies for integrating covariates in FMM with a notation that the heterogeneity in the observed features is caused by not only the factor mixture structure but also by covariates.

FMM has been extensively used for subtype discovery of various diseases. For example, Lubke et al. (Lubke et al. 2007) used FMM to discover subtypes for Attention-Deficit/Hyperactivity Disorder using behavioral data collected by a parent questionnaire. Rainbow et al. (Ho et al. 2014) applied FMM to find subtypes of breast cancer-related fatigue using fatigue symptom data. Pattyn et al. (Pattyn et al. 2015) used FMM to identify panic disorder subtypes on a broad range of anxiety symptoms. Litpon et al. (Lipton et al. 2014) used FMM to identify migraine subtypes based on a broad collection of symptom measurements. However, the existing FMM models are essentially a single-mode approach, i.e., they forces all features to share the same latent factors even when the features are indeed from distinct modes (e.g., cortical area, thickness, and volume in the migraine application). This is not biologically valid and may lead to poor clustering performance because it destroys the inherent data structure. Also, the existing FMM does not have the ability for mode and feature selection.

2.3 Development of MFMM

2.3.1 MFMM Formulation

Consider M modes of imaging data and let $\boldsymbol{x}_m$ contain mean-centered features belonging to the m-th mode, $m = 1, \dots, M$. Consider $\boldsymbol{x}_m$ generated from low-dimensional factors $\boldsymbol{f}_m$, i.e.,

$$\boldsymbol{x}_m = \mathbf{H}_m \boldsymbol{f}_m + \mathbf{B}_m \boldsymbol{z} + \boldsymbol{\varepsilon}_m \tag{2.1}$$

$\boldsymbol{z}$ contains patient-specific covariates such as sex and age. $\boldsymbol{\varepsilon}_m$ contains random errors that follow a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Psi}_m$. $\mathbf{H}_m$ and $\mathbf{B}_m$ are coefficient matrices. In factor models, $\mathbf{H}_m$ is also known as the loading matrix.

Furthermore, let $\boldsymbol{s} = (s_1, \dots, s_K)^T$ contain indicator variables for K subtypes of a disease. $s_k = 1$ if the patient has the k-th subtype and 0 otherwise. $\boldsymbol{s}$ follows a multinomial distribution, i.e.,

$$f(\boldsymbol{s}) \quad = f(s_1, \dots, s_K) \quad = \prod_{k=1}^{K} (w_k)^{s_k}, \tag{2.2}$$

where w_k is the probability of the k-th subtype. $\boldsymbol{s}$ is linked with the latent factors $\boldsymbol{f}_m$ by

$$\boldsymbol{f}_m \ = \mathbf{A}_m \boldsymbol{s} + \boldsymbol{\xi}_m. \tag{2.3}$$

$\mathbf{A}_m$is a coefficient matrix and $\boldsymbol{\xi}_m$ contains random errors that follow a zero-mean Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}_m$. $\mathbf{a}_{m,k}$ is the k-th column of $\mathbf{A}_m$, representing the mean value of $\boldsymbol{f}_m | s_k = 1$. It is clear from (2.3) that the multiple imaging modes are coupled together through the shared latent subtype variables $\boldsymbol{s}$.

Put all the parameters into a set $\boldsymbol{\Theta}$, i.e., $\boldsymbol{\Theta} = \{\{\boldsymbol{\Theta}_m\}_{m=1}^{M}, \boldsymbol{w}\}$, where $\boldsymbol{\Theta}_m = \{\mathbf{H}_m, \mathbf{B}_m, \mathbf{A}_m, \boldsymbol{\Psi}_m, \boldsymbol{\Sigma}_m\}$ and $\boldsymbol{w} = (w_1, \dots, w_K)^T$. We can write the complete log-likelihood function as:

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \left\{ \sum_{m=1}^{M} log\left(f\left(\boldsymbol{x}_{m,i} | \boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \boldsymbol{\Theta}\right)\right) + \sum_{m=1}^{M} log\left(f\left(\boldsymbol{f}_{m,i} | \boldsymbol{s}_i; \boldsymbol{\Theta}\right)\right) + log\left(f(\boldsymbol{s}_i; \boldsymbol{\Theta})\right) \right\}, \tag{2.4}$$

where N is the sample size; $\boldsymbol{s}$ follows a multinomial distribution as shown in (2.2); $\boldsymbol{f}_{m,i}|\boldsymbol{s}_{k,i} = 1 \sim N(\mathbf{a}_{m,k}, \mathbf{\Sigma}_m)$ based on (2.3); $\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i \sim N(\mathbf{H}_m\boldsymbol{f}_{m,i} + \mathbf{B}_m\boldsymbol{z}_i, \mathbf{\Psi}_m)$ according to (2.1).

Because (2.4) involves latent variables, we could use an EM algorithm to estimate the parameters. However, this approach does not consider that some modes or some features within a mode may be uninformative to the differentiation of subtypes. To help eliminate uninformative modes and features, we propose to add two L_{21}-penalties to (2.4), which results in the following optimization problem:

$$\min_{\mathbf{\Theta}} -l(\mathbf{\Theta}) + \lambda_1 \sum_{m=1}^{M} \sum_{j=1}^{P_m} \left\|\mathbf{h}_m^j\right\|_2 + \lambda_2 \sum_{m=1}^{M} \|\mathbf{A}_m\|_2. \tag{2.5}$$

$\mathbf{h}_m^j$ is the j-the row of $\mathbf{H}_m$. $\|\cdot\|_2$ is the L_2-norm of a vector or matrix. λ_1 and λ_2 are penalty parameters. It is well-known that an L_{21}-penalty is able to zero out all the coefficients within the $\|\cdot\|_2$ as a group (Yuan, M. et al. 2006). Because of this property, the proposed MFMM in (2.5) can eliminate uninformative features *hierarchically*. That is, MFMM uses $\sum_{m=1}^{M}\|\mathbf{A}_m\|_2$ to eliminate uninformative modes (i.e., features of an uninformative mode will be eliminated altogether), and uses $\sum_{m=1}^{M}\sum_{j=1}^{P_m}\left\|\mathbf{h}_m^j\right\|_2$ to eliminate uninformative features within a mode. In this way, MFMM achieves both efficiency and flexibility. To see this more clearly, we can insert (2.3) into (2.1) and obtain the distribution of features $\boldsymbol{x}_m$ for the k-th subtype, i.e.,

$$\boldsymbol{x}_m|s_k = 1 \sim N(\mathbf{H}_m\mathbf{a}_{m,k} + \mathbf{B}_m\boldsymbol{z},\ \ \mathbf{H}_m\mathbf{\Sigma}_m\mathbf{H}_m^T + \mathbf{\Psi}_m). \tag{2.6}$$

$\mathbf{a}_{m,k}$ is the k-th column of $\mathbf{A}_m$. (2.6) indicates that the distributions of $\boldsymbol{x}_m$ for different subtypes differ in their means, because $\mathbf{a}_{m,k}$ is subtype-specific. If $\mathbf{A}_m = \mathbf{0}$, i.e., $\mathbf{a}_{m,k} = \mathbf{0}$

for all subtypes, then the distribution of $\boldsymbol{x}_m$ is the same regardless of the subtypes, i.e., all the features in $\boldsymbol{x}_m$ are uninformative. Furthermore, given that $\mathbf{A}_m \neq \mathbf{0}$, if $\mathbf{h}_m^j = \mathbf{0}$, then the distribution of the j-th feature in $\boldsymbol{x}_m$ is the same regardless of the subtypes, i.e., this feature is uninformative.

Finally, to ensure model identifiability, we impose the following constraints to the MFMM:

$$E(\boldsymbol{f}_m) = \mathbf{0} \text{ and } (\boldsymbol{f}_m) = \mathbf{I}\ .$$

I is an identify matrix of an appropriate size.

2.3.2 MFMM Estimation by EM Integrated with an Efficient GMD Alogorithm

2.3.2.1 The EM Framework

Because MFMM involves latent variables, we can adopt the EM framework for model estimation. Let $\{\mathbf{X}_m\}_{m=1}^M$ be a dataset for the features of M modes. Let $\{\mathbf{F}_m\}_{m=1}^M$ and $\mathbf{S}$ be the missing data for the latent factors and subtype indicators. Also let $g(\boldsymbol{\Theta})$ denote the objective function in (2.5). In the E-step, we will need to derive the expectation of $g(\boldsymbol{\Theta})$ with respect to the conditional distribution of $\{\mathbf{F}_m\}_{m=1}^M, \mathbf{S}$ given $\{\mathbf{X}_m\}_{m=1}^M$ and the current estimate $\widetilde{\boldsymbol{\Theta}}$, i.e.,

$$_{\{\mathbf{F}_m\}_{m=1}^M, \mathbf{S}\,|\{\mathbf{X}_m\}_{m=1}^M; \widetilde{\boldsymbol{\Theta}}}\big(g(\boldsymbol{\Theta})\big)$$

$$= \left\{\sum_{i=1}^N \sum_{m=1}^M E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}; \widetilde{\boldsymbol{\Theta}}}\big(-\log f\big(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \boldsymbol{\Theta}\big)\big) + \lambda_1 \sum_{m=1}^M \sum_{j=1}^{P_m} \left\|\mathbf{h}_m^j\right\|_2\right\} +$$

$$\left\{\sum_{i=1}^N \sum_{m=1}^M E_{\boldsymbol{f}_{m,i}, \boldsymbol{s}_i|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i}; \widetilde{\boldsymbol{\Theta}}}\big(-\log f\big(\boldsymbol{f}_{m,i}|\boldsymbol{s}_i; \boldsymbol{\Theta}\big)\big) + \lambda_2 \sum_{m=1}^M \|\mathbf{A}_m\|_2\right\} +$$

$$\sum_{i=1}^N E_{\boldsymbol{s}_i|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i}; \widetilde{\boldsymbol{\Theta}}}\big(-\log f(\boldsymbol{s}_i; \boldsymbol{\Theta})\big). \tag{2.7}$$

Please see Appendix I for explicit forms of the expectations in (2.7) and detailed derivations to get them. In the M-step, we minimize (2.7) and obtain an updated estimate for $\boldsymbol{\Theta}$, i.e.,

$$\boldsymbol{\Theta}^* = arg \min_{\boldsymbol{\Theta}} \ E_{\{\mathbf{F}_m\}_{m=1}^M, \mathbf{S}\,|\{\mathbf{X}_m\}_{m=1}^M; \widetilde{\boldsymbol{\Theta}}}\big(g(\boldsymbol{\Theta})\big). \tag{2.8}$$

The two steps will iterate until convergence. A nice property of MFMM is that it allows the optimization in (2.8) to be decomposed into separate sub-optimization problems each with a smaller set of parameters to estimate, i.e.,

$$\mathbf{H}_m^*, \mathbf{B}_m^* = \underset{\mathbf{H}_m, \mathbf{B}_m}{\operatorname{argmin}} \sum_{i=1}^N E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}}\Big(-\log f\big(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \boldsymbol{\Theta}\big)\Big) + \lambda_1 \sum_{j=1}^{P_m}\big\|\mathbf{h}_m^j\big\|_2, \tag{2.9}$$

$$\{\mathbf{A}_m^*\}_{m=1}^M = \underset{\{\mathbf{A}_m\}_{m=1}^M}{\operatorname{argmin}} \quad \sum_{i=1}^N \sum_{m=1}^M E_{\boldsymbol{f}_{m,i}, \boldsymbol{s_i}|\boldsymbol{x}_{1,i},\dots,\boldsymbol{x}_{M,i}\,;\widetilde{\boldsymbol{\Theta}}}\Big(-\log f\big(\boldsymbol{f}_{m,i}|\boldsymbol{s}_i; \boldsymbol{\Theta}\big)\Big) + \lambda_2 \sum_{m=1}^M \|\mathbf{A}_m\|_2, \tag{2.10}$$

$$w_k^* = \frac{\sum_{i=1}^N f\big(s_{k,i}=1|\boldsymbol{x}_{1,i},\dots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}\big)}{\sum_{i=1}^N \sum_{k=1}^K f\big(s_{k,i}=1|\boldsymbol{x}_{1,i},\dots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}\big)}, k = 1, \dots, K,$$

$$\boldsymbol{\Psi}_m^* = diag\left(\frac{1}{N}\left(\sum_{i=1}^N\big(\boldsymbol{x}_{m,i} - \mathbf{B}_m^*\boldsymbol{z}_i\big)\big(\boldsymbol{x}_{m,i} - \mathbf{B}_m^*\boldsymbol{z}_i\big)^T - \left(\sum_{i=1}^N\big(\boldsymbol{x}_{m,i} - \mathbf{B}_m^*\boldsymbol{z}_i\big)E\left(\big(\boldsymbol{f}_{m,i}\big)^T|\boldsymbol{x}_{m,i}; \widetilde{\boldsymbol{\Theta}}\right)\right)(\mathbf{H}_m^*)^T\right)\right),$$

$$\boldsymbol{\Sigma}_m^* = \frac{\sum_{k=1}^K \sum_{i=1}^N f\big(s_{k,i}=1|\boldsymbol{x}_{1,i},\dots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}\big)\left(E\left(\boldsymbol{f}_{m,i}\left(\boldsymbol{f}_{m,i}\right)^T\middle|\boldsymbol{x}_{1,i}, \dots, \boldsymbol{x}_{M,i}, s_{k,i} = 1; \widetilde{\boldsymbol{\Theta}}\right) - \mathbf{a}_{m,k}^*\big(\mathbf{a}_{m,k}^*\big)^T\right)}{\sum_{i=1}^N f\big(s_{k,i}=1|\boldsymbol{x}_{1,i},\dots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}\big)},$$

$= 1, \dots, M$. Except (2.9) and (2.10), all other parameters can be estimated analytically. Therefore, the key to speeding up the EM iterations is to develop an efficient algorithm to solve the optimization problems in (2.9) and (2.10).

2.3.2.2 The GMD Algorithm

The optimization problem in (2.9) or (2.10) involves an L_{21}-penalty (a.k.a. group-lasso penalty). Classic approaches for solving L_{21}-penalized optimization include the block-wise descent (BD) algorithm (Yuan et al. 2006), block coordinate gradient descent algorithm (BCGD) (Meier et al. 2008), and Nesterov's method (Liu et al. 2009). However, these approaches are computationally slow. Recently, Yang and Zou (Yang et al. 2015) developed an efficient GMD algorithm that is 5~10 times faster than the classic algorithms. To apply GMD, the optimization problem must satisfy a QM condition. In what follows, we will first present the definition of the QM condition, then prove that the optimization problems in (2.9) and (2.10) satisfy the QM condition, and finally derive the GMD algorithm used to solve (2.9) and (2.10).

Definition 1 (QM condition): Let $\mathbf{D}$ denote a dataset and $\boldsymbol{\beta}$ denote p-dimensional parameters to be estimated in a minimization problem. $\boldsymbol{\beta}$ is partitioned into J groups, $\boldsymbol{\beta}^{(1)}, \dots, \boldsymbol{\beta}^{(J)}$. The minimization takes the form of

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L(\boldsymbol{\beta}|\mathbf{D}) + \lambda \textstyle\sum_{j=1}^{J} \left\| \boldsymbol{\beta}^{(j)} \right\|_2. \tag{2.11}$$

(2.11) satisfies the QM condition if and only if the following two assumptions hold:

(i) $L(\boldsymbol{\beta}|\mathbf{D})$ is differentiable as a function of $\boldsymbol{\beta}$, i.e., $\nabla L(\boldsymbol{\beta}|\mathbf{D})$ exists everywhere.

(ii) There exists a $p \times p$ matrix $\boldsymbol{\Lambda}$, which may only depend on the data $\mathbf{D}$, such that for all $\boldsymbol{\beta}, \boldsymbol{\beta}^*$,

$$L(\boldsymbol{\beta}|\mathbf{D}) \leq L(\boldsymbol{\beta}^*|\mathbf{D}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla L(\boldsymbol{\beta}^*|\mathbf{D}) + \tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \boldsymbol{\Lambda}(\boldsymbol{\beta} - \boldsymbol{\beta}). \tag{2.12}$$

Proposition 1: The minimization problem in (2.9) satisfies the QM condition.

Proposition 2: The minimization problem in (2.10) satisfies the QM condition.

Please see the proof of Proposition 1 in Appendix II. The proof of Proposition 2 follows a similar idea so it is skipped due to space limit. Because the QM condition is satisfied, GMD can be used to solve (2.9) and (2.10). Next, we briefly describe the GMD algorithm: In step $(\omega+1)$ of the algorithm, we want to update the j-th group in $\boldsymbol{\beta}^{(\omega)}$ while keeping the other groups unchanged, i.e.

$$\boldsymbol{\beta}^{(\omega+1)}-\boldsymbol{\beta}^{(\omega)}=\left(0,\dots 0,\underbrace{\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}}_{j\text{−th group}},0,\dots 0\right)^{T}.$$

According to the QM condition in (ii), we can get following inequality:

$$L\left(\boldsymbol{\beta}^{(\omega+1)}|\mathbf{D}\right)\leq L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)+\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}\nabla L^{(j)}+\frac{1}{2}\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}\boldsymbol{\Lambda}^{(j)}\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right), \quad (2.13)$$

where $\nabla L^{(j)}$ and $\boldsymbol{\Lambda}^{(j)}$ are sub-matrices of $\nabla L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)$ and $\boldsymbol{\Lambda}$ only including the rows and columns corresponding to the j-th group. Furthermore, let τ_j be the largest eigenvalue of $\boldsymbol{\Lambda}^{(j)}$ and set $\rho_j=(1+10^{-6})\tau_j$. Then, (2.13) can be further relaxed as

$$L\left(\boldsymbol{\beta}^{(\omega+1)}|\mathbf{D}\right)\leq L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)+\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}\nabla L^{(j)}+\frac{1}{2}\rho_j\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right), \quad (2.14)$$

where the inequality strictly holds unless ${\boldsymbol{\beta}^{(j)}}^{(\omega+1)}={\boldsymbol{\beta}^{(j)}}^{(\omega)}$. Using (2.14), we can solve the optimization in (2.11) by solving

$$\underset{{\boldsymbol{\beta}^{(j)}}^{(\omega+1)}}{\operatorname{argmin}}\ L\left(\boldsymbol{\beta}^{(\omega)}|\mathbf{D}\right)-\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}\nabla L^{(j)}-\frac{1}{2}\rho_j\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)^{T}\left({\boldsymbol{\beta}^{(j)}}^{(\omega+1)}-{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)+\lambda\left\|{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right\|_2, \quad (2.15)$$

which has an analytical solution, i.e.,

$${\boldsymbol{\beta}^{(j)*}}^{(\omega+1)}=\frac{1}{\rho_j}\left(-\nabla L^{(j)}+\rho_j{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right)\left(1-\frac{\lambda}{\left\|-\nabla L^{(j)}+\rho_j{\boldsymbol{\beta}^{(j)}}^{(\omega)}\right\|_2}\right)_{+}. \quad (2.16)$$

This greatly reduces the computational time for solving the optimization problem. Also, this algorithm is guaranteed to converge (proof is skipped).

2.3.2.3 Model Selection

The numbers of subtypes and factors as well as the penalty parameters can be selected according to a model selection criterion that balances the model fit and complexity. The former is measured by the log-likelihood of the observed data, $\tilde{l}$, while the latter is given by the degree of freedom, df, that counts the number of non-zeros in the estimated parameters. There are various criteria to combine the fit and complexity in the literature, among which we found BIC works well in our simulation and real data experiments. BIC takes the form of $-2\tilde{l} + log(N) \times df$. Finally, once a model that minimizes the BIC has been identified, each subject/sample i will be classified to a cluster for which the posterior probability of belonging to that cluster, i.e., $P\left(s_{k,i} = 1|\boldsymbol{x}_{1,i}, \dots, \boldsymbol{x}_{M,i}; \boldsymbol{\Theta}^{*}\right)$, is maximized.

2.4 Simulation Studies

We generate simulation data to assess the performance of MFMM and compare it with competing methods. Consider 120 subjects who have one of two subtypes for a disease. The probability of each subtype is $w_1 = 0.4$ and $w_2 = 0.6$. This means that among the 120 patients, 48 have subtype 1 and 72 have subtype 2. Furthermore, suppose there are four imaging modes with two factors in each mode. In order to demonstrate the capability of MFMM in identifying informative or eliminating uninformative modes, we assume that the first two modes have a two-cluster structure corresponding to the two subtypes while the other two modes do not so they are uninformative modes. To accomplish this, we set the cluster-wise factor means of each mode according to Table 1. The factor means differ between two clusters in mode 1 and 2, but not in mode 3 and 4. We set the covariance matrices of the four modes according to (2.17), which do not differ between two clusters:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix},\ \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.1 \\ 0.1 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_3 = \mathbf{I}, \text{ and } \boldsymbol{\Sigma}_4 = \mathbf{I}. \qquad (2.17)$$

Also, we consider three patient-specific covariates that are sampled from $N(0,1)$. Once the data for the factors and covariates are generated from the aforementioned distributions, we proceed to generate the features using (2.1). The number of features is set to be 40 in each mode. In order to demonstrate the capability of MFMM in identifying informative or eliminating uninformative features within each mode, we assume that 10 features have a two-cluster structure corresponding to the two subtypes while the other 30 do not so they are uninformative features. To accomplish this, we set the coefficient matrices $\mathbf{H}_m$ in (2.1) to be

$$\mathbf{H}_m = \begin{pmatrix} \widetilde{\mathbf{H}}_{10\times 2} \\ \mathbf{0}_{30\times 2} \end{pmatrix},$$

with each element of $\widetilde{\mathbf{H}}$ generated as follows: Generate a number from $Uniform\ [1, 1.5]$, which is used as the magnitude of the element. To decide the sign of the element, generate another number from $N(0,1)$. If this number is greater than -0.5, create a positive sign; otherwise, create a negative sign. Finally, we sample the random errors $\boldsymbol{\varepsilon}_m$ in (2.1) from $N(\mathbf{0}, 0.1 \times \mathbf{I})$.

Table 1: Factor means of each cluster/subtype within each mode

Modes	Clusters/subtypes	
	$k = 1$	$k = 2$
$m = 1$	$\mathbf{a}_{1,1} = (1,1)^T$	$\mathbf{a}_{1,2} = (-1,-1)^T$
$m = 2$	$\mathbf{a}_{2,1} = (0.75,0.75)^T$	$\mathbf{a}_{2,2} = (-0.75,-0.75)^T$
$m = 3$	$\mathbf{a}_{3,1} = (0,0)^T$	$\mathbf{a}_{3,2} = (0,0)^T$
$m = 4$	$\mathbf{a}_{4,1} = (0,0)^T$	$\mathbf{a}_{4,2} = (0,0)^T$

MFMM is applied to the simulation data. The experiment is repeated for 20 times. The two informative modes are correctly selected in 18 out of the 20 experiments, resulting in 90% accuracy for mode selection by MFMM. Furthermore, to evaluate the feature selection accuracy of MFMM within each mode, we compute the sensitivity (i.e., the percentage of features with truly non-zero coefficients that are selected) and specificity (i.e., the percentage of features with truly zero coefficients that are not selected) of each experiment. Table 2 shows the average and standard deviation of sensitivity and specificity over all the experiments for each mode. These results show that MFMM achieves high accuracies in mode and feature selection.

Table 2: Sensitivity and specificity of feature selection by MFMM (average $\pm$ standard deviation over all experiments)

	Mode 1	Mode 2	Mode 3	Mode 4
Feature selection sensitivity (%)	100 ± 0	100 ± 0	100 ± 0	100 ± 0
Feature selection specificity (%)	92.8 ± 5.1	95.2 ± 4.1	93.3 ± 5.4	93.8 ± 4.5

Furthermore, we compare MFMM with several completing methods, which are methods people would typically adopt if MFMM were not available. One competing approach is to apply conventional FMM on one mode at a time; the other is to apply FMM on pooled features from all the modes. For a fair comparison with MFMM, we add feature selection to FMM using an L_{21}-penality, calle gFMM hereafter, and use BIC to select the penalty parameter. Because the ultimate performance measure for a clustering algorithm is its accuracy in discovering the true clustering structure, we compare the clustering accuracy of the competing methods with MFMM. Figure 2 shows the overall clustering accuracy of each method, defined as the percentage of subjects correctly classified to their ground-truth subtypes. For the first completing approach of applying gFMM on each mode alone, we only show the results of mode 1 and 2, because the accuracies of mode 3 and 4 (i.e., the uninformative modes) are poor. It can be seen that MFMM achieves an overall accuracy of $90.8\% \pm 10.7\%$, which is significantly higher than the competing methods in Figure 2 whose accuracies are $66.4\% \pm 13.2\%, 60\% \pm 0\%$, and $61.6\% \pm 7.3\%$, respectively (p values < 0.001).

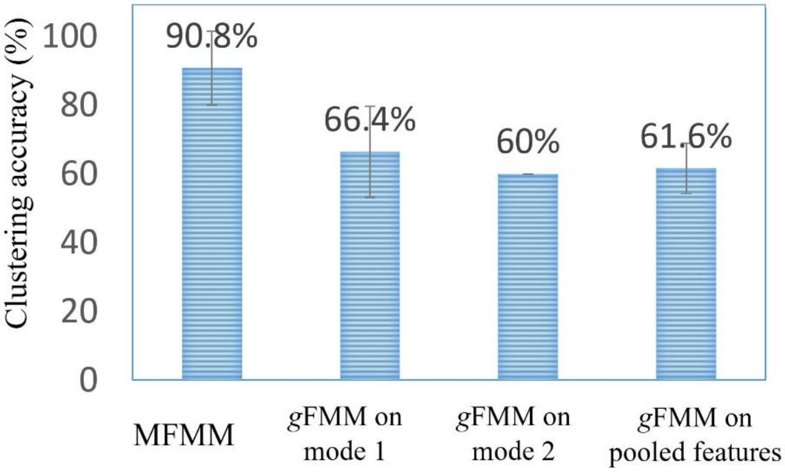


Figure 2: Clustering accuracy of MFMM in comparison with three competing methods

2.5 Application in Migraine Subtype Discovery from Multi-mode Imaging Data

2.5.1 Data Collection and Image Processing

The data used for this application were obtained from MCA and WashU through our clinical collaborators. A total of 120 subjects were included in this analysis. Migraine was diagnosed in accordance with the diagnostic criteria defined by the International Classification of Headache Disorders (Tibshirani 1996). Data collected from all subjects included demographics such as age and sex, and clinical characteristics and symptoms measured through a number of instruments such as Beck Depression Inventory (BDI), State-Trait Anxiety Inventory (STAI), Allodynia Symptom Checklist 12 (ASC-12), Migraine Disability Assessment (MIDAS), Hyperacusis Questionnaire, Photosensitivity Assessment Questionnaire, together with a few individually measured key symptom variables such as headache frequency, number of years with migraine, and aura status.

Structural MRI data were obtained from two Siemens 3T MRI machines. Details of the MRI acquisition were described in prior publications (Schwedt et al. 2017, Schwedt et al. 2015). Using a cortical reconstruction and segmentation program in the FreeSurfer image analysis suite (version 5.3, http://www.surfer.nmr.mgh.harvard.edu/), cortical area, thickness, and volume measurements of 68 Region of Interests (ROIs) were extracted. In our study, cortical area, cortical thickness, and volume are treated as three modes. Within each mode, 34 features correspond to ROIs at the right brain hemisphere, while the other 34 correspond to same-name ROIs at the left hemisphere. Our analysis found no difference in the clustering structure between using 68 features in each mode and using 34 by averaging the features at the left and right hemispheres corresponding to the same-name ROI. Therefore, we will only present the result for the latter situation in this paper.

2.5.2 Data Augmentation with Nuisance Modes and Features

A challenge in applying any clustering method to real data is that the ground-truth clustering structure is unknown. This prohibits rigorous performance assessment for the clustering result using accuracy metrics like what can be done in a simulation study. On the other hand, this is what makes a clustering method appealing because it can lead to new discovery to extend the boundary of the existing knowledge in a domain. In this paper, we design our study in a way that allows for performance assessment. Specifically, we add artificially generated nuisance modes and features to the real modes and features, apply MFMM to the combined data, and examine if MFMM is able to identify the real modes and features. In a prior study (Schwedt et al. 2017), we applied a non-penalized version of MFMM to the same dataset and found all three modes and 34 features within each mode to be relevant to a two-cluster (subtype) structure. This result was validated with the medical knowledge of our clinical collaborators and the existing literature of migraine studies. Therefore, the three modes are treated as real modes and 34 features as real features in the present study.

To add nuisance modes and features, we employ the following steps: First, we add 68 nuisance features $\widetilde{\boldsymbol{x}}_m$ to each real mode, which are generated by

$$\widetilde{\boldsymbol{x}}_m = \mathbf{0}_{68\times 1}\boldsymbol{f}_m + \widetilde{\mathbf{B}}_m\boldsymbol{z} + \widetilde{\boldsymbol{\varepsilon}}_m, m = 1,2,3. \tag{2.18}$$

(2.18) takes the same form as (2.1) with $\mathbf{H}_m = \mathbf{0}_{68\times 1}$ because nuisance features are not supposed to have any clustering structure. To make the distribution of nuisance data as close as possible to the real data, we do not give the coefficient matrix $\widetilde{\mathbf{B}}_m$ arbitrarily but sample each row of $\widetilde{\mathbf{B}}_m$ with replacement from the rows of $\mathbf{B}_m$. Although the true $\mathbf{B}_m$ is unknown, we have a reliable estimate $\widehat{\mathbf{B}}_m$ from a previous study that applied a non-penalized version of MFMM to the real dataset (Schwedt et al. 2017). Similarly, we sample $\widetilde{\boldsymbol{\varepsilon}}_m$ from $N(\mathbf{0},\ \widehat{\boldsymbol{\Psi}}_m)$, where $\widehat{\boldsymbol{\Psi}}_m$ is from the previous study. Furthermore, we add three nuisance modes each having 34+68=102 features that match the size of augmented real modes. The features in a nuisance mode are generated by

$$\tilde{\tilde{\boldsymbol{x}}}_m = \tilde{\tilde{\mathbf{H}}}_m\tilde{\tilde{\boldsymbol{f}}}_m + \tilde{\tilde{\mathbf{B}}}_m\boldsymbol{z} + \tilde{\tilde{\boldsymbol{\varepsilon}}}_m, m = 1,2,3, \tag{2.19}$$

where each row of $\tilde{\tilde{\tilde{\mathbf{B}}}}_m$ is sampled with replacement from the rows of $\widehat{\mathbf{B}}_m$. Each of the first 34 rows of $\tilde{\tilde{\tilde{\mathbf{H}}}}_m$ are sampled from the rows of $\widehat{\mathbf{H}}_m$ while the remaining 68 rows are all zeros. $\tilde{\tilde{\boldsymbol{\varepsilon}}}_m$ is sampled from $N(\mathbf{0},\ \widehat{\boldsymbol{\Psi}}_m)$. We sample $\tilde{\tilde{\boldsymbol{f}}}_m$ from $N(0, 1)$ because nuisance modes are not supposed to have any clustering structure. Through this procedure, we create an augmented dataset of six modes (three real and three nuisance modes) and 102 features within each mode (34 real features in each real mode).

2.5.3 Results from Application of MFMM

We apply MFMM to the augmented data together with two patient-specific covariates, sex and age. All three real modes are correctly selected, resulting in 100% accuracy. Within the two real modes, the sensitivity and specificity of selecting out the real features is 100% and 96.6%, respectively. MFMM found two clusters/subtypes among the 120 subjects, each consisting of 53 and 67 subjects, respectively. Call these subtypes A and B hereafter. A total of seven factors (2, 3, 2 from cortical area, cortical thickness, and volume modes, respectively) are found to differentiate subtypes A and B. The correspondence between these factors and the imaging features is encoded in the estimated loading matrix $\widehat{\mathbf{H}}_m$ and is shown in Figure 3. A clear pattern is that the within each mode, loadings that reflect the contributions to the imaging features from one factor (i.e., bars of one color) are different from another factor (i.e., bars of another color). This indicates that there may be more than one biological underpinning underlying the observed imaging features, and thus supporting the validity of multiple factors found in each mode. Furthermore, we highlight the ROIs in each mode whose measurements most contribute to differentiation of the two subtypes in

Figure 4. These ROIs are those whose loading magnitudes are greater than the 80-th percentile of all the loadings estimated by the MFMM.

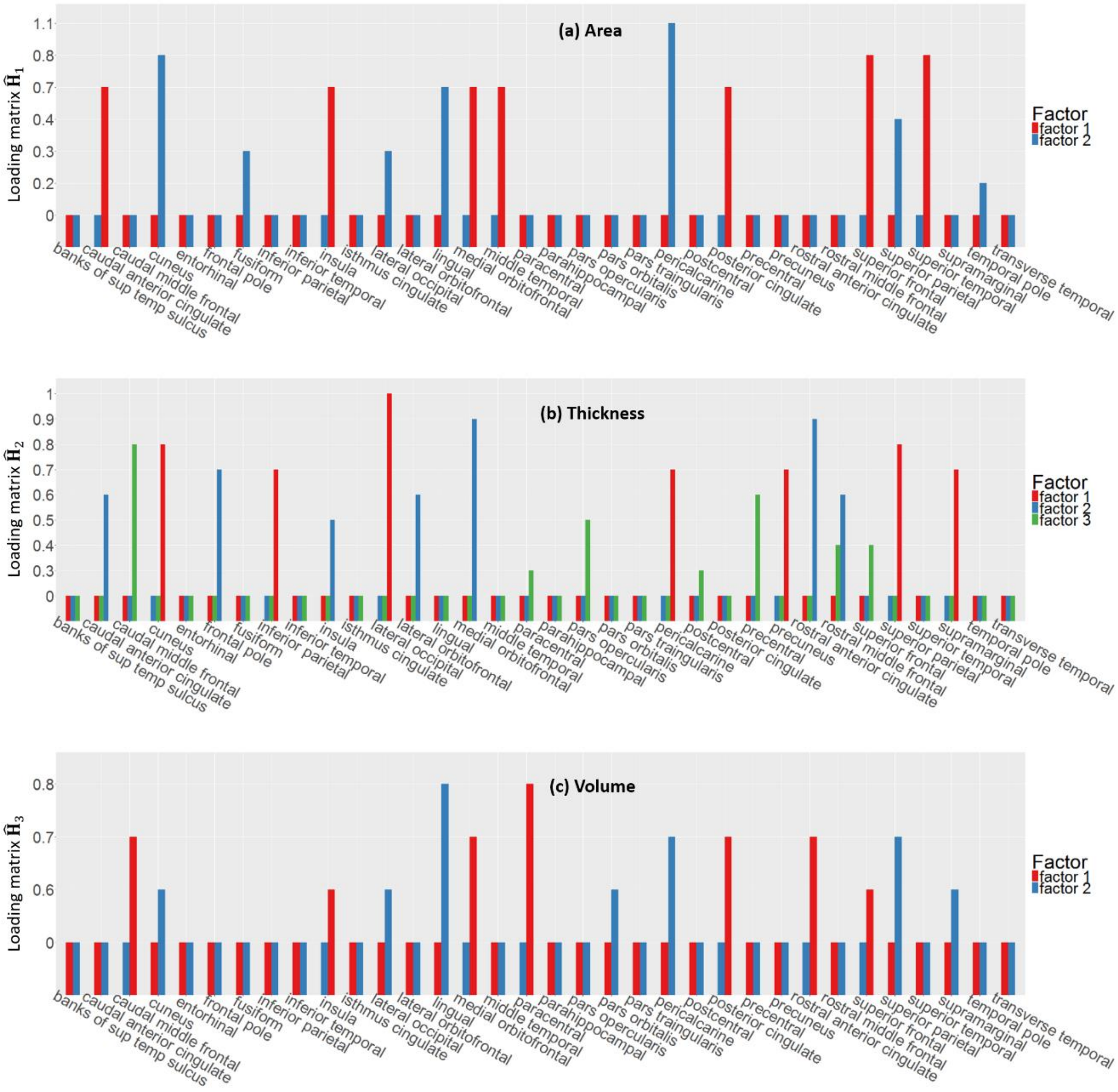


Figure 3: Estimated loadings (y axis) that show the contribution of factors to original features (x axis) for (a) area, (b) thickness, and (c) volume. Loadings whose magnitudes are less than the 80-th percentile of all loadings are suppressed and represented by short bars for better visualization.

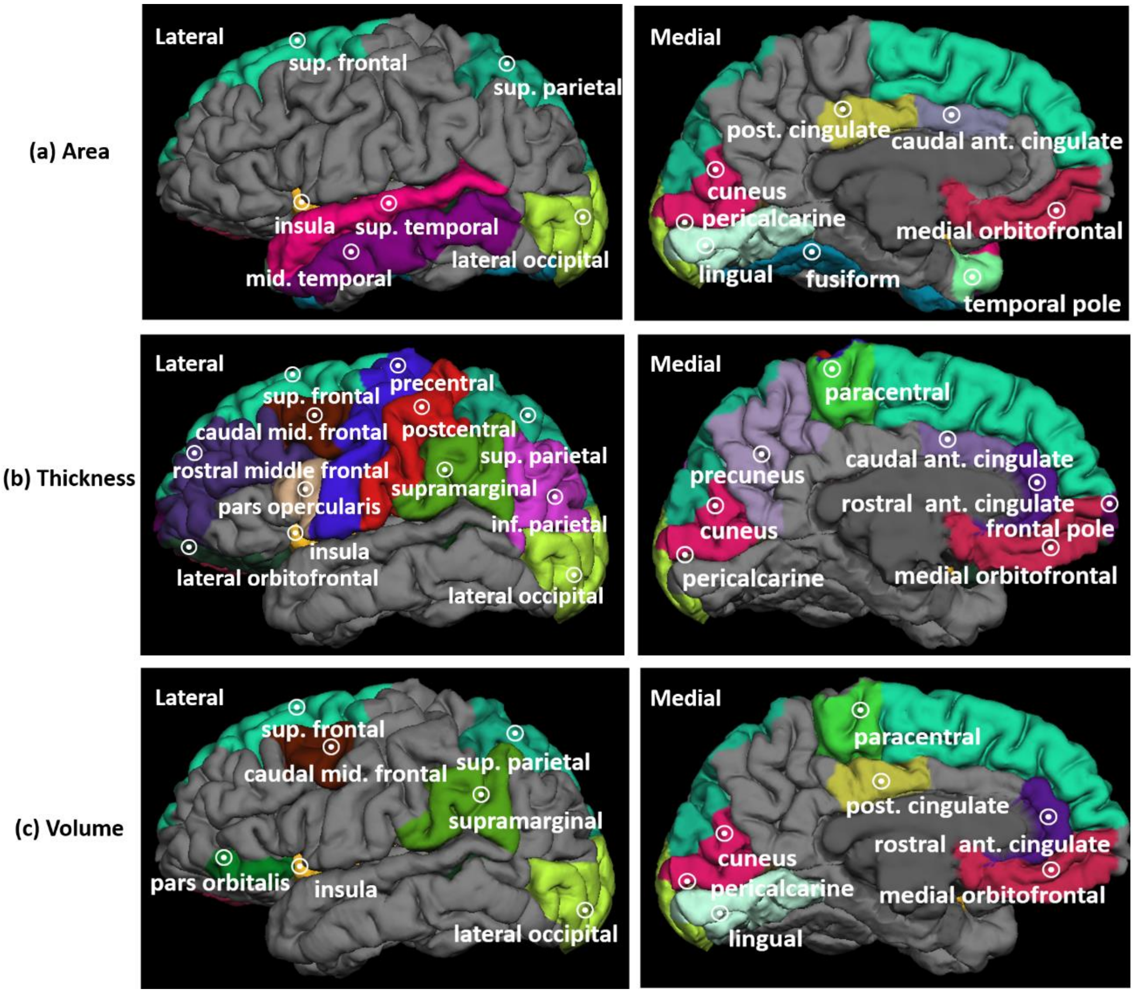


Figure 4: The ROIs whose (a) area, (b) thickness, or (c) volume measurements most contribute to differentiation of the two subtypes are color-highlighted on a 3-D reading of the brain.

Next, we would like to see how well subjects with subtype A and B are separated in terms of the seven factors. Since it is impossible to visualize the separation on a seven-dimensional space, we choose to visualize it mode by mode. Figure 5(a)-(c) plot the subjects in terms of the two, three, and two factors within area, thickness, and volume modes, respectively. Figure 5(d) plots the posterior probability of each subject being subtype B, which reflects the joint effect of the seven factors in separating subjects with the two subtypes. These results demonstrate that all the factors in each mode and all three modes contribute to the subtype separation. Also, the two clusters are separated very well, as the vast majority of the subjects in each cluster have a high posterior probability of being in the cluster they are assigned to, as shown in Figure 5(d).

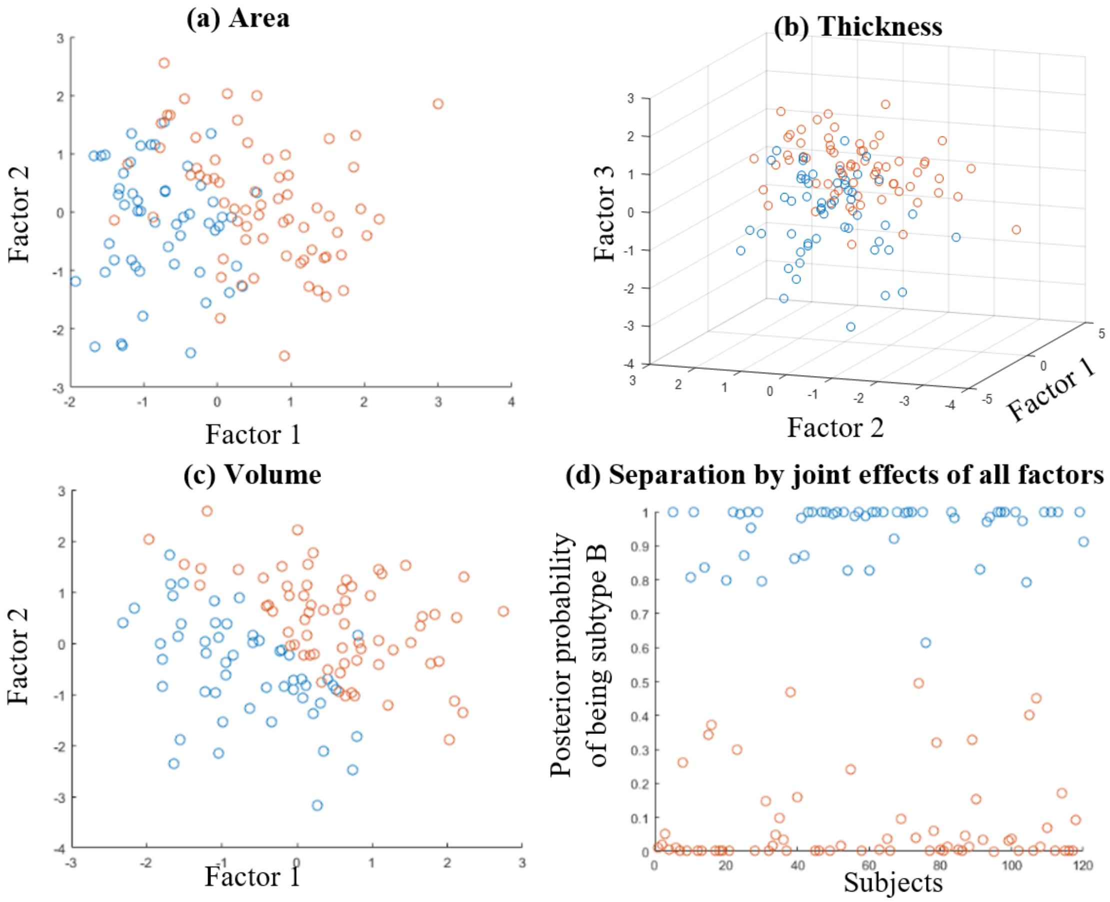


Figure 5: Separation of subjects with subtype A (red) and B (blue) in terms of the factors in each mode and the posterior probability of cluster membership.

Finally, we would like to see how the two imaging-defined subtypes differ in clinical characteristics and symptoms. We focus on a panel of variables including the number of headache days per month, number of years with migraine, aura status, MIDAS score, STAI score, BDI score, allodynia during and between migraine attacks, hyperacusis, and photophobia. We perform hypothesis testing to compare subtypes A and B in terms of each variable. Three variables are found to have statistically significant subtype difference: migraine subjects with subtype A have a greater number of years with migraine (p value = 0.01), more migraine-related disability as measured by the MIDAS score (p value = 0.04), and greater symptoms of allodynia during migraine attacks (p value = 0.03).

2.5.4 Discussion on Medical Implications

The main finding of this application was identification of two clusters (i.e., subtype A and B) of the study cohort based on structural MRI measurements of brain cortical area, cortical thickness, and volume. The two clusters significantly differ in a number of clinical characteristics including the number of years with migraine, allodynia during migraine attacks, and migraine-related disability. These clinical variables have been previously reported to relate to brain imaging findings in migraine. For example, a number of studies have shown that the number of years with migraine is associated with brain structure and function (Chong et al. 2016, Chong et al. 2015, Jin et al. 2013). In general, the longer a person has had migraine and the more attacks they have had, the greater the brain differences. Allodynia symptom severity was measured using the ASC-12, a questionnaire that collects information about cutaneous allodynia – the sensation of pain to normally non-noxious stimulation of the skin (Lipton et al. 2008). Several imaging studies have demonstrated associations between brain structure and function with symptoms of allodynia (Moulton et al. 2008, Schwedt et al. 2014, Chong et al. 2016, Russo et al. 2016). Disability could be a marker for the severity of migraine symptoms as well as the person’s ability to cope with their migraine symptoms [Ford et al. 2008]. Migraine severity and coping mechanisms could both associate with measures of brain structure and function. The structural measurements that differentiated the two clusters in this study were of brain regions that have previously been shown to be aberrant in migraine and/or in individuals who have allodynia. (Schwedt et al. 2014, Russo et al. 2016, Schwedt et al. 2015, Hadjikhani et al. 2013, Russo et al. 2012, Schwedt et al. 2014, Schwedt et al. 2015, Liu et al. 2012, Zhao et al. 2013, Mickleborough et al. 2016, Schmitz et al. 2008). Of note, the clusters did not differ for headache frequency or aura status – characteristics that are currently used to subtype migraine in the ICHD 3 beta. Further investigations are needed to determine if brain imaging based subtyping of migraine is superior to current classification in regards to prognosticating outcomes, predicting development of co-morbidities, and predicting treatment responses, which important components of precision medicine.

2.6 Conclusion

In this paper, we proposed a new method, MFMM, for clustering multi-mode image data to enable subtype identification. MFMM employed a double-L_{21}-penalized likelihood formulation to enable imaging mode and feature selection. We developed an efficient GMD algorithm embedded in the EM framework to estimate the model parameters. We performed simulation experiments to compare MFMM with competing methods and found significantly better performance of MFMM in terms of mode selection accuracy, feature selection accuracy, and clustering accuracy. We applied MFMM to migraine subtype discovery based on brain cortical area, cortical thickness, and volume measurements from MRI. Two clusters/subtypes were found and well separated using a total of seven factors.

Subjects in the two clusters had significant different clinical characteristics. Findings from this study showed promise for imaging-based subtyping of migraine and patient stratification toward PM.

There are a number of extensions for the current study. In terms of statistical modeling, MFMM could be extended to include mixed-type features. In terms of migraine, functional imaging data such as fMRI could be combined with the currently used structural MRI for subtype identification on a broader range of structural and functional measurements. Also, MFMM and its extensions can be applied to subtype discovery of other diseases.

CHAPTER 3

A MULTI-RESPONSE MULTILEVEL MODEL WITH APPLICATION IN NURSE CARE COORDINATION

3.1 Introduction

Due to the increasing prevalence of chronic illnesses and aging of our society, patients hospitalized for acute care episodes nowadays are likely to have at least one chronic illness (Anderson, 2007; Boltz et al., 2008). This has created a tremendous new challenge for the already-heavily-burdened health care system: treating and caring for the acute episode of a patient who has chronic comorbidities is complex, requiring well-planned interventions and involving numerous providers. To tackle this challenge, care coordination has been recommended as one fundamental approach (Institute for Healthcare Improvement, 2004; MedPac, 2007) and effective care coordination has been found to decrease adverse events, improve quality and efficiency of care, and enhance patient satisfaction (McDonald et al., 2007; Sticker et al., 2009). In coordinating patient care within the hospital, staff nurses, as the patient's "ever-present" health care team members, play a vital role. In "Keeping Patients Safe", a recent report by the Institute of Medicine, the role of staff nurses in care coordination that promotes patient safety and quality outcomes was highlighted. Recent qualitative studies illuminated the considerable amount of time staff nurses spend coordinating patient care via a broad range of activities from admission to discharge (Hendrich et al., 2008; Lamb et al., 2008; Aiken et al., 2008; Friese, 2008; Kazanjian et al. 2005; Laschinger et al. 2006).

Until recently, study of staff nurse care coordination was hampered by the lack of an operational definition of staff nurse care coordination and the absence of tools to measure the process. In a recent project sponsored by the Robert Wood Johnson Foundation, Lamb (one of the co-authors of this paper) and her team developed, for the first time, an operational definition for staff nurse care coordination through systematic analysis of extensive observations and interviews of staff nurses and members of their nursing and interdisciplinary teams. The definition of nurse care coordination, according to Lamb et al. (2008), is "the actions initiated by nurses with patients, families, and/or members of their health care team to manage and correct the sequence, timing, and/or effectiveness of patient care from hospital admission to discharge". Based on this definition, Lamb further identified six categories of staff nurse care coordination activities: "organizing", "checking", "mobilizing", "exchanging", "assisting", and "backfilling", referred to as "o", "c", "m", "e", "a", and "b" in this paper. Detailed definitions of the six categories can be found in Appendix A.1. Furthermore, Lamb led the design and validation of an instrument called the Nurse Care Coordination Instrument (NCCI) that allows for quantitative data to be collected to measure the coordination activities. This effort provided groundwork for advancing the understanding and improvement of nurse care coordination in the hospital.

Capitalizing on the newly developed NCCI, we present a study in this paper that aims to examine and reveal how nurses' care coordination is related to their practice environment, demographics, and workload. To achieve this goal, a multilevel model (Demidenko 2013) is a natural choice for analyzing the NCCI data, because the predictors in the data come from two levels: demographic and workload variables at the individual/nurse level and characteristics of the practice environment at the organizational/unit level. However, simply adopting the existing multilevel model would not suffice. There are a number of challenges inspiring new model development in this paper. First, nurse care coordinate is not a univariate concept but includes multiple categories describing the multi-faceted coordination activities, such as “m”, “e”, and “a”. This results in multiple response variables to be modeled simultaneously. Simply applying the existing multilevel model to each response separately overlooks the correlation between the multiple responses. This correlation inherently exists and could be strong in our problem domain because prior research has found that nurses who engage more in one type of care coordination activities tend to engage more in another type of activities (Duva, 2010). From the point of view of statistical modeling, joint modeling of multiple responses allows for the multiple models to borrow strength from each other and mitigates sample size limitation. Second, considering that the statistical model should be ultimately helpful for guiding the improvement and best practices of nurse care coordination, the model should be able to identify a small number of significant predictors out of the originally included predictors, many of which could be noise or have only a small effect on the responses. This provides convenience for practical implementation of the modeling results. Furthermore, with similar prediction accuracy, a model that uses the same subset of predictors to predict the multiple responses is more desirable than a model that uses different subsets of predictors to predict different responses. The former model typically requires a smaller number of total predictors to be measured in order to predict the multiple responses, thus saving cost and effort for data acquisition. Also, if the predictors can be confirmed to causally affect the response variables of care coordination, the former model means a potential saving in the cost of intervention by adjusting fewer predictors to improve multiple aspects of the care coordination.

To address the aforementioned challenges in modeling the NCCI data, we propose a multi-response multilevel model that uses two adaptive l_{21}-penalties to enable *joint* fixed effect selection and *joint* random effect selection across multiple responses. To our best knowledge, such a model is not available in the existing literature. The contribution of this research is two-fold: To the field of statistical modeling, we propose a new formulation for a multi-response multilevel model driven by a newly emerged problem in the health care system, develop an efficient Block Coordinate Descent (BCD) algorithm integrated with an Expectation-Maximization (EM) framework for model estimation, perform theoretical analysis to reveal the insight as to how the proposed method "joins" the estimation of the model for each response variable together and the benefit of such a joint estimation, and demonstrate asymptotic properties. To the field of nursing research, our study is the first-of-its-kind that elucidates the quantitative relationship between nurses' practice environment, demographics, and workload and their multi-faceted care coordination activities. We anticipate that the knowledge and insights generated from our study could facilitate the design and optimization of nurses' workload and practice environment, which leads to better care coordination and eventually better patient outcomes.

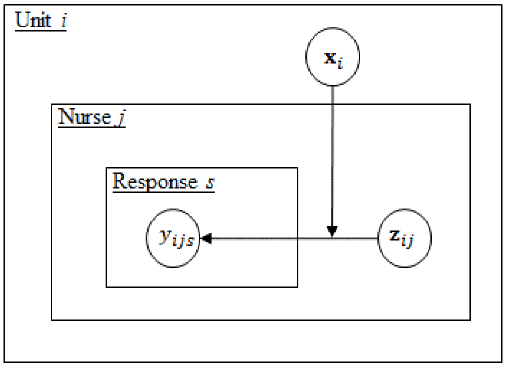
The rest of the paper is organized as follows: Section 3.2 reviews the existing research related to the proposed statistical model; Section 3.3 presents the model formulation; Section 3.4 presents the estimation algorithm; Section 3.5 investigates asymptotic properties; Section 3.6 presents simulation studies; Section 3.7 presents the application; Section 3.8 concludes the chapter.

3.2 Literature Review

The proposed model is a combination of conventional multilevel models and modern variable selection techniques. Conventional multilevel models have been extensively discussed in numerous papers and books, which do not include “variable selection” in their formulations. Variable selection techniques are modern statistical modeling and machine learning developments that were driven by the emergence of high-dimensional datasets in various domains. The basic idea of variable selection is to add penalties to the regression coefficients in order to shrink the estimates for the regression coefficients of insignificant predictors to be exactly zero. Various forms of penalties have been proposed, which can model different structures of the predictors and/or have different statistical properties. The classic lasso model was proposed as a penalized least squares method with an l_1-penalty that resulted in the estimates of some regression coefficients to be exactly zero (Tibshirani, 1996). Fan et al. (2001) conjectured the asymptotic inconsistency of lasso and proposed an SCAD penalty that enjoyed the oracle properties. Zhao et al. (2006) further discussed the consistency of lasso and proved an almost sufficient and necessary condition for lasso to select the true model. Zou (2006) proposed an adaptive lasso model that applied adaptive weights to the l_1-penalty and proved the oracle properties of this model. To handle data with grouped predictors, Yuan et al. (2006) proposed a group lasso model capable of selecting a sparse set of groups by imposing an l_1-penalty on the regression coefficients of predictors from each group. Zou et al. (2005) proposed an elastic net model that encouraged a grouping effect among strongly correlated predictors. As an integration of some existing methods and bootstrap, random lasso was proposed to alleviate some of the limitations of lasso, elastic net, and related methods (Wang et al., 2011). A hierarchical lasso was proposed to not only remove unimportant groups of predictors but also select important predictors within a group (Zhou and Zhu, 2010). To achieve grouped and hierarchical variable selection, a Composite Absolute Penalties (CAP) family was proposed to add side information to boost the estimation of a regression or classification model (Zhao et al. 2009). In addition to frequentist methods, Bayesian methods for variable selection were also developed. George et al. (1997) discussed and compared a variety of approaches for Bayesian variable selection. Fahrmeir et al. (2010) provided a unified view between Bayesian methods and penalized frequentist methods for variable selection. Note that the review in this section by no means provides a complete list of existing variable selection approaches. However, a vast majority of the existing approaches including all the aforementioned ones are for single-level predictors; there is much less research in the multilevel setting.

Among the few existing efforts in introducing variable selection in the multilevel setting, Schelldorfer et al. (2011) proposed a method that adds an l_1-penalty to the fixed effects. This achieves variable selection on fixed effects alone, but not on random effects. A significant difficulty in variable selection on random effects is that, not like fixed effects that are characterized by regression coefficients, random effects are characterized by a covariance matrix, $\boldsymbol{\Psi}$. Therefore, variable selection on random effects will have to be done through penalizing the covariance matrix, which is not straightforward. To achieve this, Ibrahim et al. (2011) proposed a Cholesky decomposition on $\mathbf{\Psi}$, i.e., $\mathbf{\Psi} = \mathbf{\Lambda}\mathbf{\Lambda}^T$, where $\mathbf{\Lambda}$ is a lower triangular matrix. Then, the elements in each row of $\mathbf{\Lambda}$ are penalized as a group. If some rows of $\mathbf{\Lambda}$ are estimated to be zero, this will result in some rows and columns of $\mathbf{\Psi}$ to be zero, which has an effect of removing the random effects corresponding to these rows/columns. Bondell et al. (2010) proposed a modified Cholesky decomposition to decompose $\mathbf{\Psi}$ into a lower triangular matrix, $\mathbf{\Gamma}$, whose diagonal elements are all ones and a diagonal matrix $\mathbf{D}$, i.e., $\mathbf{\Psi} = \mathbf{D}\mathbf{\Gamma}(\mathbf{D}\mathbf{\Gamma})^T$. Then, an adaptive l_1-penalty is put on the diagonal elements of $\mathbf{D}$ to shrink some elements to be zero, which has an effect of removing the rows/columns of $\mathbf{\Psi}$ corresponding to these elements and thereby excluding the random effects corresponding to the removed rows/columns. Ahn et al. (2012) proposed a moment-based loss function for estimating the covariance matrix of random effects. Then, two types of penalties including a hard threshsholding operator and a sandwich-type soft thresholding penalty are imposed to achieve variable selection of random effects. However, all these existing methods are for a single response only.

3.3 A Multi-response Multilevel Model with Joint Fixed Effect Selection and Joint Random Effect Selection



y_{ijs}: the s–th response variable of nurse j in unit i (e.g., the amount of time spent on the s–th care coordination activity by nurse j in unit i)

$\mathbf{z}_{ij}$: individual-level predictors (e.g., nurse demographic and workload variables)

$\mathbf{x}_i$: unit-level predictors (e.g., organizational characteristics such as availability of certain technology and policies in the unit that can potentially facilitate care coordination)

$s = 1, \dots, \mathbb{S}$ (responses), $i = 1, \dots, N$ (units), $j = 1, \dots, n_i$ (nurses)

Figure 6: Data structure targeted by the proposed model

It is evident from the literature review that little work has been done on multi-response multilevel models with variable selection in both fixed and random effects. However, such a model is needed for properly modeling the NCCI data. This motivates our new model development. Our proposed model is aimed for a nested multilevel multi-response data structure as depicted in Figure 6. First, y_{ijs} is related to $\boldsymbol{z}_{ij}$ by a linear model, i.e., $y_{ijs} = \boldsymbol{\alpha}_{is}^T \boldsymbol{z}_{ij} + \varepsilon_{ijs}$. This is called a level-one model, which characterizes how nurses' demographics and workload impact their care coordination. Then, this impact, reflected by $\boldsymbol{\alpha}_{is}$, is related to $\boldsymbol{x}_i$ by a level-two model, i.e., $\boldsymbol{\alpha}_{is} = \mathbf{B}_s \boldsymbol{x}_i + \boldsymbol{e}_{is}$. This model characterizes how units' organizational characteristics affect the relationship between nurses' demographics/workload and their coordination. Combining level-one and level-two models, we can get

$$y_{ijs} = \boldsymbol{x}_i^T \mathbf{B}_s^T \boldsymbol{z}_{ij} + \boldsymbol{e}_{is}^T \boldsymbol{z}_{ij} + \varepsilon_{ijs}, \tag{3.1}$$

where $\boldsymbol{e}_{is} \sim N(\mathbf{0}, \sigma_s^2 \boldsymbol{\Psi}_s)$ and $\varepsilon_{ijs} \sim N(0, \sigma_s^2)$ are between-unit and within-unit random errors. $\mathbf{B}_s$ and $\boldsymbol{e}_{is}$ are known as fixed and random effects, respectively. Apply a modified Cholesky decomposition (Chen and Dunson, 2003) to the covariance matrix of the random effects, i.e., $\boldsymbol{\Psi}_s = \mathbf{D}_s \boldsymbol{\Gamma}_s (\mathbf{D}_s \boldsymbol{\Gamma}_s)^T$, where $\mathbf{D}_s$ is a diagonal matrix and $\boldsymbol{\Gamma}_s$ is a lower triangular matrix. Then, the random effects can be re-parameterized as $\boldsymbol{e}_{is} = \mathbf{D}_s \boldsymbol{\Gamma}_s \tilde{\boldsymbol{e}}_{is}$, where $\tilde{\boldsymbol{e}}_{is} \sim \mathrm{N}(\mathbf{0}, \sigma_s^2 \mathbf{I})$. For the ease of subsequent discussion, we also re-parameterize the fixed effects as $\boldsymbol{x}_i^T \mathbf{B}_s^T \boldsymbol{z}_{ij} = \boldsymbol{\beta}_s^T \boldsymbol{w}_{ij}$, where $\boldsymbol{w}_{ij}$ is a vector that concatenates $\boldsymbol{x}_i$, $\boldsymbol{z}_{ij}$, and the interactions between them, and $\boldsymbol{\beta}_s$ is a vector consisting of the elements of $\mathbf{B}_s$. Considering these re-parameterizations, (3.1) becomes

$$y_{ijs} = \boldsymbol{\beta}_s^T \boldsymbol{w}_{ij} + (\mathbf{D}_s \boldsymbol{\Gamma}_s \tilde{\boldsymbol{e}}_{is})^{\mathrm{T}} \boldsymbol{z}_{ij} + \varepsilon_{ijs}, \tag{3.2}$$

Stacking up the data of all the nurse within the i-th unit, we can get $\boldsymbol{y}_{is} = \mathbf{W}_i \boldsymbol{\beta}_s + \boldsymbol{Z}_i (\mathbf{I} \otimes \mathbf{D}_s)(\mathbf{I} \otimes \boldsymbol{\Gamma}_s) \tilde{\boldsymbol{e}}_s + \boldsymbol{\varepsilon}_{is}$, which corresponds to the grey blocks in Figure 7. Further stacking up the data of all the units as illustrated in Figure 7, we can get

$$\boldsymbol{y}_s = \mathbf{W} \boldsymbol{\beta}_s + \mathbf{Z}(\mathbf{I} \otimes \mathbf{D}_s)(\mathbf{I} \otimes \boldsymbol{\Gamma}_s) \tilde{\boldsymbol{e}}_s + \boldsymbol{\varepsilon}_s. \tag{3.3}$$

The parameters to be estimated for the model in (3.3) can be put into a vector $\boldsymbol{\phi}_s = (\boldsymbol{\beta}_s^T, \boldsymbol{d}_s^T, \boldsymbol{\gamma}_s^T)^T$, where $\boldsymbol{d}_s$ is a vector consisting of the diagonal elements of $\mathbf{D}_s$ and $\boldsymbol{\gamma}_s$ is a vector consisting of the elements in $\boldsymbol{\Gamma}_s$. Considering all the responses, the total parameters to be estimated are $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_{\mathbb{S}}^T)^T$.

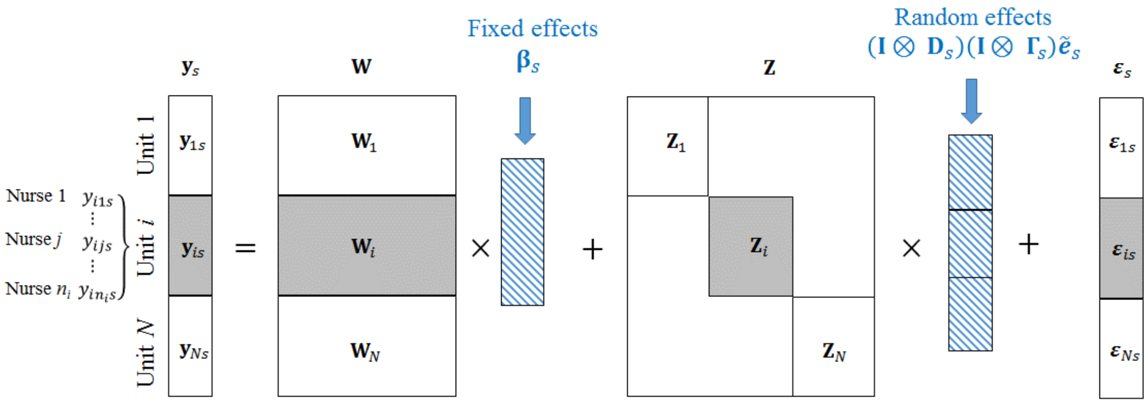


Figure 7: A graph illustration for the model in (3.3)

According to (3.3), $\boldsymbol{y}_s$ follows a normal distribution with a mean $\mathbf{W}\boldsymbol{\beta}_s$ and a covariance matrix $\widetilde{\mathbf{V}}_s = Diag(\mathbf{V}_{1s}, \dots, \mathbf{V}_{is}, \dots, \mathbf{V}_{Ns})$, where $\mathbf{V}_{is} = \sigma_s^2\left(\mathbf{Z}_i\mathbf{D}_s\boldsymbol{\Gamma}_s\boldsymbol{\Gamma}_s{}^T\mathbf{D}_s{}^T\mathbf{Z}_i{}^T + \mathbf{I}\right)$. Then, dropping constants, the log-likelihood function of the parameter set $\boldsymbol{\phi}$ can be written as:

$$l\left(\boldsymbol{\phi}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}\right) = -\frac{\mathbf{1}}{\mathbf{2}}\sum_{s=1}^{\mathbb{S}}\left\{log\left|\widetilde{\mathbf{V}}_s\right| + (\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s)^T\widetilde{\mathbf{V}}_s{}^{-1}(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s)\right\}. \qquad (3.4)$$

Furthermore, by treating the random effects in $\tilde{\boldsymbol{e}}_s$ as observed data and dropping constants, we can write the complete-data log-likelihood function as

$$l\left(\boldsymbol{\phi}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}, \{\tilde{\boldsymbol{e}}_s\}_{s=1}^{\mathbb{S}}\right) = -\frac{\sum_{i=1}^{N} n_i + N\mathbb{Q}}{2}\sum_{s=1}^{\mathbb{S}} log\,\sigma_s{}^2 - \frac{1}{2}\sum_{s=1}^{\mathbb{S}}\frac{1}{\sigma_s{}^2}\left(\begin{matrix}\|\boldsymbol{y}_s - \mathbf{Z}(\mathbf{I} \otimes \mathbf{D}_s)(\mathbf{I} \otimes \boldsymbol{\Gamma}_s)\tilde{\boldsymbol{e}}_s - \mathbf{W}\boldsymbol{\beta}_s\|^2 \\ +\tilde{\boldsymbol{e}}_s^T\,\tilde{\boldsymbol{e}}_s\end{matrix}\right). \qquad (3.5)$$

Let $\mathbb{P}$ and $\mathbb{Q}$ denote the dimensions of the fixed and random effects, respectively. To enable joint variable selection in fixed effects across all the responses, we impose an l_{21}-penalty on $\boldsymbol{\beta}_s$, i.e., $\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\beta_{ps}^2}$. This l_{21}-penalty allows for the fixed effects corresponding to the same predictor across all the responses to be selected as a group, i.e., these fixed effects are either all kept in or dropped out of the model. To achieve the same purpose for random effects, we impose another l_{21}-penalty on $\mathbf{d}_s$, i.e., $\sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} d_{qs}^2}$. The consequence of this l_{21}-penalty is that if the d_{qs}'s corresponding to a predictor across all the responses (e.g., the $d_{q'1}, \dots, d_{q'\mathbb{S}}$ corresponding to the q'-th predictor across all the responses) are zero, then the q'-th row and q'-th column of $\boldsymbol{\Psi}_1, \dots, \boldsymbol{\Psi}_{\mathbb{S}}$ are automatically zero by definition of the modified Cholesky decomposition. As a result, the random effects corresponding to the q'-th predictor across all the responses are dropped out of the model as a group. Furthermore, we want to have adaptive weights to penalize different coefficients and thus using an adaptive l_{21}-penalty in the form of $\sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2}$ and $\sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2}$, where $\tilde{\beta}_{ps}$ is an adaptive weight for β_{ps} and $\tilde{d}_{qs}$ is an adaptive weight for d_{qs}. The purpose is to have a large amount of shrinkage for zero coefficients and a smaller amount of shrinkage for nonzero coefficients, thus achieving improved estimator efficiency and variable selection properties. With all these considerations, we define an adaptive l_{21}-penalized complete-data log-likelihood criterion as follows:

$$f(\boldsymbol{\phi}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}, \{\tilde{\boldsymbol{e}}_s\}_{s=1}^{\mathbb{S}}) = -l(\boldsymbol{\phi}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}, \{\tilde{\boldsymbol{e}}_s\}_{s=1}^{\mathbb{S}}) + \lambda_1 \sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2} + \lambda_2 \sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2} . \quad (3.6)$$

λ_1 and λ_2 are regularization parameters for the fixed and random effects, respectively. In the next section, we present model estimation based on (3.6).

3.4 Model Estimation by EM Integrated with a BCD Optimization Algorithm

The proposed adaptive l_{21}-penalized complete-data log-likelihood function in (3.6) involves unobserved variables, $\tilde{\boldsymbol{e}}_s$. This makes the EM algorithm a proper choice for model estimation. EM is a general method for finding the maximum likelihood estimate of model parameters from data with missing values (Dempster et al., 1977). It has also been used when optimizing the likelihood function is intractable analytically but is possible if some quantities in the likelihood function can be assumed known. These quantities are treated as missing/unobserved data in EM. EM works by iteratively conducting two steps. The E-step is to find the expectation of the complete-data log-likelihood with respect to the unobserved data given the observed data and the current parameter estimates. The M-step is to find parameter estimates that maximize the expectation in the E-step. The two steps are repeated until convergence. The EM framework has a nice property that it is guaranteed to converge to a local maximum of the likelihood function (Wu, 1983).

The challenges in using the general EM framework in specific model estimation are to derive the expectation specific to that model formulation in the E-step and to develop an efficient optimization algorithm in the M-step. In what follows, we will discuss the two steps specific to our problem setting in (3.6):

In the E-step at the ω-th iteration, our goal is to compute the expectation of the criterion in (3.6) with respect to the conditional distribution of $\{\tilde{\boldsymbol{e}}_s\}_{s=1}^{\mathbb{S}}$ given $\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}$ and the current estimate for $\boldsymbol{\phi}$, $\boldsymbol{\phi}^{(\omega)}$. It can be derived that this conditional distribution is normal with a mean and a covariance matrix given by (please see the derivation in Appendix A.2):

$$\hat{\boldsymbol{e}}_s^{(\omega)} = (\tilde{\boldsymbol{\Gamma}}_s^{T(\omega)}\,\widetilde{\mathbf{D}}_s^{(\omega)}\mathbf{Z}^T\mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\tilde{\boldsymbol{\Gamma}}_s^{(\omega)} + \mathbf{I})^{-1}(\mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\tilde{\boldsymbol{\Gamma}}_s^{(\omega)})^T\left(\boldsymbol{y}_s - \mathbf{W}\,\boldsymbol{\beta}_s^{(\omega)}\right), \quad (3.7)$$

$$\mathbf{U}_s^{(\omega)} = \sigma_s^{2(\omega)}\,(\tilde{\boldsymbol{\Gamma}}_s^{T(\omega)}\,\widetilde{\mathbf{D}}_s^{(\omega)}\mathbf{Z}^T\mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\tilde{\boldsymbol{\Gamma}}_s^{(\omega)} + \mathbf{I})^{-1}, \quad (3.8)$$

respectively. Here, $\widetilde{\mathbf{D}}_s^{(\omega)} = \mathbf{I} \otimes \mathbf{D}_s^{(\omega)}$, $\tilde{\boldsymbol{\Gamma}}_s^{(\omega)} = \mathbf{I} \otimes \boldsymbol{\Gamma}_s^{(\omega)}$, and the $\sigma_s^{2(\omega)}$ in (3.8) is given by

$$\sigma_s^{2(\omega)} = \frac{\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right)^T (\mathbf{Z}\tilde{\boldsymbol{\Gamma}}_s^{T(\omega)}\,\widetilde{\mathbf{D}}_s^{(\omega)}\mathbf{Z}^T\mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\tilde{\boldsymbol{\Gamma}}_s^{(\omega)}\mathbf{Z}^T + \boldsymbol{I})^{-1} \times \left(\boldsymbol{y}_s - \mathbf{W}\,\boldsymbol{\beta}_s^{(\omega)}\right)}{\sum_{i=1}^{N} \mathrm{n}_i}. \tag{3.9}$$

Then, the expectation of the criterion in (3.6) can be obtained as:

$$g\left(\boldsymbol{\phi}\middle|\boldsymbol{\phi}^{(\omega)}\right) =$$

$$\sum_{s=1}^{\mathbb{S}} \frac{1}{2\sigma_s^{2(\omega)}} \left(\begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{d}_s \end{bmatrix}^T \begin{bmatrix} \mathbf{W}^T\mathbf{W} & \mathbf{W}^T\mathbf{Z}\,Diag\left(\tilde{\boldsymbol{\Gamma}}_s\hat{\boldsymbol{e}}_s^{(\omega)}\right)(\mathbf{1}_N \otimes \mathbf{I}) \\ (\mathbf{1}_N \otimes \mathbf{I})^T Diag(\tilde{\boldsymbol{\Gamma}}_s\hat{\boldsymbol{e}}_s^{(\omega)})\mathbf{Z}^T\mathbf{W} & (\mathbf{1}_N \otimes \mathbf{I})^T(\mathbf{R} \circ \tilde{\boldsymbol{\Gamma}}_s\widehat{\mathbf{G}}_s^{(\omega)}\tilde{\boldsymbol{\Gamma}}_s^T)\,(\mathbf{1}_N \otimes \mathbf{I}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{d}_s \end{bmatrix} - \right.$$

$$\left. 2\,\boldsymbol{y}_s^T \begin{bmatrix} \mathbf{W} & \mathbf{Z}\,Diag\left(\tilde{\boldsymbol{\Gamma}}_s\hat{\boldsymbol{e}}_s^{(\omega)}\right)(\mathbf{1}_N \otimes \mathbf{I}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{d}_s \end{bmatrix} \right) + \lambda_1 \sum_{p=1}^{P} \sqrt{\sum_{s=1}^{S} \left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2} + \lambda_2 \sum_{q=1}^{Q} \sqrt{\sum_{s=1}^{S} \left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2}, \tag{3.10}$$

where $\tilde{\boldsymbol{\Gamma}}_s = \mathbf{I} \otimes \boldsymbol{\Gamma}_s$, $\mathbf{1}_\mathrm{N}$ is a $N \times 1$ vector of ones, $Diag(\tilde{\boldsymbol{\Gamma}}_s\hat{\boldsymbol{e}}_s^{(\omega)})$ is a diagonal matrix whose diagonal elements being $\tilde{\boldsymbol{\Gamma}}_s\hat{\boldsymbol{e}}_s^{(\omega)}$, $\mathbf{R} = \mathbf{Z}^T\mathbf{Z}$, $\widehat{\mathbf{G}}_s^{(\omega)} = \mathrm{E}(\tilde{\boldsymbol{e}}_s\tilde{\boldsymbol{e}}_s^T) = \mathbf{U}_s^{(\omega)} + \hat{\boldsymbol{e}}_s^{(\omega)}\hat{\boldsymbol{e}}_s^{T(\omega)}$. “ ° ” represents the Hadamard product operator.

In the M-step, our goal is to minimize $g\left(\boldsymbol{\phi}\middle|\boldsymbol{\phi}^{(\omega)}\right)$ with respect to $\boldsymbol{\phi}$. Recall that $\boldsymbol{\phi}$ includes $\boldsymbol{\beta}_s$, $\boldsymbol{d}_s$, and $\boldsymbol{\Gamma}_s$, $s = 1, .., \mathbb{S}$. Therefore, the optimization of $g\left(\boldsymbol{\phi}\middle|\boldsymbol{\phi}^{(\omega)}\right)$ with respect to $\boldsymbol{\phi}$ can be done by iterating between two sub-optimizations: One sub-optimization is to minimize $g\left(\boldsymbol{\phi}\middle|\boldsymbol{\phi}^{(\omega)}\right)$ with respect to $\boldsymbol{\Gamma}_s$, treating $(\boldsymbol{\beta}_s^T, \boldsymbol{d}_s^T)^T$ as given. This sub-optimization has a closed-form solution. The other sub-optimization is to minimize $g\left(\boldsymbol{\phi}\middle|\boldsymbol{\phi}^{(\omega)}\right)$ with respect to $(\boldsymbol{\beta}_s^T, \boldsymbol{d}_s^T)^T$, treating $\boldsymbol{\Gamma}_s$ as given. This sub-optimization takes the following form:

$$min_{\{\boldsymbol{\beta}_s\}_{s=1}^{\mathbb{S}}, \{\boldsymbol{d}_s\}_{s=1}^{\mathbb{S}}} \left\{ \begin{array}{c} \sum_{s=1}^{\mathbb{S}} \frac{1}{2\sigma_s^{2(\omega)}} \left(\begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{d}_s \end{bmatrix}^T \mathbf{A}_s \begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{d}_s \end{bmatrix} - 2\,\boldsymbol{b}_s^T \begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{d}_s \end{bmatrix} \right) + \\ \lambda_1 \sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2} + \lambda_2 \sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2} \end{array} \right\}, \tag{3.11}$$

where $\mathbf{A}_s$ and $\boldsymbol{b}_s^T$ are known and their forms can be obtained by comparing (3.11) and (3.10). To solve the optimization problem in (3.11), we note that (3.11) is a convex optimization whose non-smooth parts, i.e., the adaptive l_{21}-norms, are separable. This property motivates us to develop a BCD algorithm that is guaranteed to converge to a global minimum. Next, we describe the proposed BCD algorithm:

Each “coordinate” in the BCD algorithm corresponds to a fixed or random effect, so there are $\mathbb{P}+\mathbb{Q}$ coordinates. BCD cycles through the coordinates until convergence. In what follows, we will discuss one cycle of BCD that estimates the p-th fixed effect. Other cycles that estimates the other fixed effects and the random effects share a similar procedure. Specifically, in the cycle that estimates the p-th fixed effect, $\boldsymbol{\beta}^p = (\beta_{p1}, \dots, \beta_{p\mathbb{S}})$ is the parameter to be estimated, whereas all other fixed effects and all random effects are treated as known. Considering this, the optimization in (3.11) can be written as:

$$min_{\boldsymbol{\beta}^p} \ \sum_{s=1}^{\mathbb{S}} \frac{1}{2\sigma_s^{2(\omega)}} \begin{pmatrix} [\hat{\beta}_{1s} \cdots \hat{\beta}_{p-1,s}\, \beta_{ps}\, \hat{\beta}_{p+1,s} \cdots \hat{\beta}_{Ps}\, \widehat{\boldsymbol{d}}_s^T]\mathbf{A}_s[\hat{\beta}_{1s} \cdots \hat{\beta}_{p-1,s}\, \beta_{ps}\, \hat{\beta}_{p+1,s} \cdots \hat{\beta}_{Ps}\, \widehat{\boldsymbol{d}}_s^T]^T \\ -2\, \boldsymbol{b}_s^T[\hat{\beta}_{1s} \cdots \hat{\beta}_{p-1,s}\, \beta_{ps}\, \hat{\beta}_{p+1,s} \cdots \hat{\beta}_{Ps}\, \widehat{\boldsymbol{d}}_s^T]^T \end{pmatrix} +$$

$$\lambda_1 \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2}. \tag{3.12}$$

Denote the objective function in (3.12) by $g(\boldsymbol{\beta}^p)$ and optimal solution by $\boldsymbol{\beta}^{\boldsymbol{p}^*}$, i.e., $\boldsymbol{\beta}^{\boldsymbol{p}^*} = argmin_{\boldsymbol{\beta}^p}\, g(\boldsymbol{\beta}^p)$. The subgradients of $g(\boldsymbol{\beta}^p)$ are $\partial g(\boldsymbol{\beta}^p) = [l_{p1}, \dots, l_{p\mathbb{S}}]^T + \lambda_1 \boldsymbol{k}_p$, where $l_{ps} = \frac{1}{\sigma_s^{2(\omega)}}\left(\mathbf{A}_s^p[\hat{\beta}_{1s} \cdots \hat{\beta}_{p-1,s}\, \beta_{ps}\, \hat{\beta}_{p+1,s} \cdots \hat{\beta}_{\mathbb{P}s}\, \widehat{\boldsymbol{d}}_s^T]^T - b_{ps}\right)$, $s = 1, \dots, \mathbb{S}$. $\mathbf{A}_s^p$ is the p-th row of $\mathbf{A}_s$ and b_{ps} is the p-th element of $\boldsymbol{b}_s^T$. Furthermore, the necessary and sufficient condition for $\boldsymbol{\beta}^{\boldsymbol{p}^*}$ to be zero is that the equations $[\tilde{l}_{p1}, \dots, \tilde{l}_{p\mathbb{S}}]^T + \lambda_1 \boldsymbol{k}_p = \mathbf{0}$ have a solution with $\boldsymbol{k}_p \in \left\{\left(\frac{t_{p1}}{\tilde{\beta}_{p1}}, \dots, \frac{t_{p\mathbb{S}}}{\tilde{\beta}_{p\mathbb{S}}}\right)^T \mid (t_{p1}, \dots, t_{p\mathbb{S}}) \triangleq \boldsymbol{t}_p, ||\boldsymbol{t}_p||_2 \leq 1\right\}$, where $\tilde{l}_{ps} =$

$\frac{1}{\sigma_s^{2(\omega)}}\left(\mathbf{A}_s^p\left[\hat{\beta}_{1s} \cdots \hat{\beta}_{p-1,s}\; 0\; \hat{\beta}_{p+1,s} \cdots \hat{\beta}_{\mathbb{P}s}\; \widehat{\boldsymbol{d}}_s^T\right]^T - b_{ps}\right)$. One equivalent criterion for $\boldsymbol{\beta^{p*}}$ to be zero is $\left\|\left(\tilde{l}_{p1} \times \tilde{\beta}_{p1}, \dots, \tilde{l}_{p\mathbb{S}} \times \tilde{\beta}_{p\mathbb{S}}\right)\right\|_2 \leq \lambda_1$. If this criterion is not satisfied, we minimize (3.12) by a one-dimensional search over $\boldsymbol{\beta}^p = \left(\beta_{p1}, \dots, \beta_{p\mathbb{S}}\right)$ as follows: Focus on the step in the search that estimates β_{ps} while treating other elements in $\boldsymbol{\beta}^p$ as known. Then, (3.12) becomes an optimization problem with respect to β_{ps}, i.e.,

$$min_{\beta_{ps}} \left\{h\left(\beta_{ps}\right) + \lambda_1 \sqrt{\left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2 + c_{-ps}^2}\right\}. \tag{3.13}$$

$h\left(\beta_{ps}\right)$ is a quadratic convex function of β_{ps}, i.e.,

$$h\left(\beta_{ps}\right) = \frac{\{\mathbf{A}_s\}_{pp}{}^2}{\sigma_s^{2(\omega)}}\, \beta_{ps}^2 + 2\tilde{l}_{ps}\, \beta_{ps}, \tag{3.14}$$

where $\{\mathbf{A}_s\}_{pp}$ is the p-th diagonal element of $\mathbf{A}_s$. c_{-ps}^2 is the sum of squared adaptive estimates for the elements in $\boldsymbol{\beta}^p$ except β_{ps}, i.e.,

$$c_{-ps}^2 = \left(\frac{\widehat{\beta}_{p1}}{\widetilde{\beta}_{p1}}\right)^2 + \cdots + \left(\frac{\widehat{\beta}_{p,s-1}}{\widetilde{\beta}_{p,s-1}}\right)^2 + \left(\frac{\widehat{\beta}_{p,s+1}}{\widetilde{\beta}_{p,s+1}}\right)^2 + \cdots + \left(\frac{\widehat{\beta}_{p,\mathbb{S}}}{\widetilde{\beta}_{p,\mathbb{S}}}\right)^2. \tag{3.15}$$

So c_{-ps}^2 is a non-negative known constant at this step. The solution to the optimization in (3.13) is motivated by the following proposition (proof of the proposition is not shown due to space limit):

Proposition 2.2.1: Let β_{ps}^0 be the minimizer of (3.14), i.e., $\beta_{ps}^0 = -\frac{\tilde{l}_{ps}}{\{\mathbf{A}_s\}_{pp}{}^2/\sigma_s^{2(\omega)}}$. A sufficient and necessary condition for β_{ps}^* to be the solution to (3.13) is:

$$\beta_{ps}^* = sign\left(\beta_{ps}^0\right)\left(\left|\beta_{ps}^0\right| - \frac{\lambda_1}{2\left|\widetilde{\beta}_{ps}\right|\{\mathbf{A}_s\}_{pp}{}^2/\sigma_s^{2(\omega)}}\right)^+, \text{ if } c_{-ps}^2 = 0 \tag{3.16}$$

$$\beta^*_{ps} + \lambda_1 \frac{\beta^*_{ps}}{\sqrt{{\beta^*_{ps}}^2 + c^2_{-ps}}} = \beta^0_{ps}, \qquad \text{if } c^2_{-ps} \neq 0 \tag{3.17}$$

Based on Proposition 2.2.1, to solve (3.13), we can apply a simple soft-thresholding rule to β^0_{ps} if $c^2_{-ps} = 0$; otherwise, the solution that satisfies the equation in (3.17) will have to be obtained numerically (no close-form exists). This completes our discussion on the parameter estimation. Interested readers can find the pseudo code of our EM and BCD algorithms in Supplementary Material.

Furthermore, we would like to examine the parameter estimation process described previously and reveal the insight as to how the proposed method "joins" the estimation of the model for each response variable together and the benefit of such a joint estimation. As a matter of fact, Proposition 2.2.1 reveals how the estimation for one response is joined with other responses. Specifically, (3.17) shows that the estimate for an effect in the s-th response, i.e., β^*_{ps}, is related to c^2_{-ps}, which is a sum of squares of the same effects in other responses. Corollary 2.2.1 further shows that the relationship is monotonic, i.e., the larger the c^2_{-ps}, the smaller the shrinkage on β^*_{ps}. This is indeed an advantage of the proposed method: Consider the situation when a predictor has a non-trivial fixed or random effect on all the responses, but the effects on *some* responses are small. If each response was modelled separately, these small-effects would be easily missed. In contrast, the proposed method is able to borrow strength from other responses with larger effects to help identify the small effects.

Corollary 2.2.1: Let $c_{-ps,1}^{2}$ and $c_{-ps,2}^{2}$ be two values for c_{-ps}^{2}. Let $\beta_{ps,1}^{*}$ and $\beta_{ps,2}^{*}$ be the β_{ps}^{*} that satisfies (3.17) corresponding to $c_{-ps,1}^{2}$ and $c_{-ps,2}^{2}$, respectively. If $c_{-ps,1}^{2} < c_{-ps,2}^{2}$, then $\frac{|\beta_{ps,1}^{*}|}{|\beta_{ps}^{0}|} < \frac{|\beta_{ps,2}^{*}|}{|\beta_{ps}^{0}|}$.

Finally in this section, we discuss the choice of tuning parameters. There are two tuning parameters in the proposed method, i.e., λ_1 and λ_2 corresponding to the fixed and random effects, respectively. The previously presented EM and BCD algorithms apply to a given pair of (λ_1, λ_2). To find the best pair, a common practice is to choose one that minimizes a certain model selection criterion such as BIC, AIC, and cross-validated prediction errors. We propose a BIC-type criterion and found it to work well in simulation studies and the application. For a given pair of (λ_1, λ_2), the criterion takes the following format:

$$BIC(\lambda_1, \lambda_2) = \mathbf{-2}\left(l\left(\widehat{\boldsymbol{\phi}}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}\right)\right) + log(\textstyle\sum_{i=1}^{N} n_i) \times \, df(\lambda_1, \lambda_2), \quad (3.18)$$

where $l\left(\widehat{\boldsymbol{\phi}}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}\right)$ is the log-likelihood function defined in (3.4) when the parameters in $\boldsymbol{\phi}$ take their estimated values and $df(\lambda_1, \lambda_2)$ is the number of non-zero fixed and random effects in $\widehat{\boldsymbol{\phi}}$. The pair of (λ_1, λ_2) that minimizes this criterion is used to produce the final parameter estimation.

3.5 Asymptotic Properties

Having presented the formulation and estimation for the proposed model in Sections 3 and 4, respectively, we further study the asymptotic properties of the model in this section. First, we define some new notations. Let $Q\left\{(\boldsymbol{\phi}_s)_{s=1}^{\mathbb{S}}\right\}$ be the adaptive l_{21}-penalized log-likelihood function, i.e.,

$$Q\left(\{\boldsymbol{\phi}_s\}_{s=1}^{\mathbb{S}}\right) = l\left(\{\boldsymbol{\phi}_s\}_{s=1}^{\mathbb{S}}\right) - \lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2} - \lambda_{N2}\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2}\ . \quad (3.19)$$

$l\left\{(\boldsymbol{\phi}_s)_{s=1}^{\mathbb{S}}\right\}$ is the log-likelihood function defined in (3.4).

Let $\tilde{\boldsymbol{\phi}}_s$ be the true value for $\boldsymbol{\phi}_s$. $\tilde{\boldsymbol{\phi}}_s = (\tilde{\boldsymbol{\phi}}_{s1}^T, \tilde{\boldsymbol{\phi}}_{s2}^T)^T$. $\tilde{\boldsymbol{\phi}}_{s1}^T = (\tilde{\boldsymbol{\beta}}_{s1}^T, \tilde{\boldsymbol{d}}_{s1}^T, \tilde{\boldsymbol{\gamma}}_{s1}^T)^T$ is a vector whose elements are non-zero. Without loss of generality, assume that the first $\mathbb{P}'$ elements of $\boldsymbol{\beta}_s^T$ are non-zero, which are stored in $\tilde{\boldsymbol{\beta}}_{s1}^T$, and the first $\mathbb{Q}'$ elements of $\boldsymbol{d}_s^T$ are non-zero, which are stored in $\tilde{\boldsymbol{d}}_{s1}^T$. Let $\left|\tilde{\boldsymbol{\gamma}}_{s1}\right|$ denote the dimension of $\tilde{\boldsymbol{\gamma}}_{s1}^T$. So the dimension of $\tilde{\boldsymbol{\phi}}_{s1}^T$ is $\mathbb{P}' + \mathbb{Q}' + \left|\tilde{\boldsymbol{\gamma}}_{s1}\right|$. $\tilde{\boldsymbol{\phi}}_{s2}^T = (\tilde{\boldsymbol{\beta}}_{s2}^T, \tilde{\boldsymbol{d}}_{s2}^T, \tilde{\boldsymbol{\gamma}}_{s2}^T)^T$ consists the remaining elements of $\tilde{\boldsymbol{\phi}}_s$, i.e., $\tilde{\boldsymbol{\phi}}_{s2}^T = \mathbf{0}$. Put the $\tilde{\boldsymbol{\phi}}_{s1}^T$, $s = 1, \dots, \mathbb{S}$, into one vector, i.e., $\tilde{\boldsymbol{\phi}}^{\mathbf{1}} = \left(\tilde{\boldsymbol{\phi}}_{11}^T, \dots, \tilde{\boldsymbol{\phi}}_{\mathbb{S}1}^T\right)^T$, and the $\tilde{\boldsymbol{\phi}}_{s2}^T$, $s = 1, \dots, \mathbb{S}$, into another vector, i.e., $\tilde{\boldsymbol{\phi}}^{\mathbf{2}} = \left(\tilde{\boldsymbol{\phi}}_{12}^T, \dots, \tilde{\boldsymbol{\phi}}_{\mathbb{S}2}^T\right)^T$. Let $\tilde{\boldsymbol{\phi}} = \begin{pmatrix}\tilde{\boldsymbol{\phi}}^1 \\ \tilde{\boldsymbol{\phi}}^2\end{pmatrix}$.

Following the same decomposition as $\tilde{\boldsymbol{\phi}}_s$, let $\boldsymbol{\phi}_s = (\boldsymbol{\phi}_{s1}^T, \boldsymbol{\phi}_{s2}^T)^T$. Similarly, define $\boldsymbol{\phi}^1 = (\boldsymbol{\phi}_{11}^T, \dots, \boldsymbol{\phi}_{\mathbb{S}1}^T)^T$, $\boldsymbol{\phi}^2 = (\boldsymbol{\phi}_{12}^T, \dots, \boldsymbol{\phi}_{\mathbb{S}2}^T)^T$, and $\boldsymbol{\phi} = \begin{pmatrix}\boldsymbol{\phi}^{\mathbf{1}} \\ \boldsymbol{\phi}^{\mathbf{2}}\end{pmatrix}$. Let $Q\left\{\begin{pmatrix}\boldsymbol{\phi}^1 \\ \mathbf{0}\end{pmatrix}\right\}$ denote the $Q\left(\{\boldsymbol{\phi}_s\}_{s=1}^{\mathbb{S}}\right)$ in (3.19) with $\boldsymbol{\phi}^{\mathbf{2}} = \mathbf{0}$. According to the sequence of $\boldsymbol{\phi} = \begin{pmatrix}\boldsymbol{\phi}^1 \\ \boldsymbol{\phi}^2\end{pmatrix}$, we rearrange $\mathbf{W}_s$, $\mathbf{Z}_s$, and $\tilde{\mathbf{V}}_s$ to be $\mathbf{W}_{s(1)}$, $\mathbf{Z}_{s(1)}$ and $\tilde{\mathbf{V}}_{s(1)} = \mathbf{Z}_{s(1)}\left(\mathbf{I} \otimes \mathbf{D}_{s(1)}\right)\left(\mathbf{I} \otimes \boldsymbol{\Gamma}_{s(1)}\right)\left(\mathbf{I} \otimes \boldsymbol{\Gamma}_{s(1)}\right)^T\left(\mathbf{I} \otimes \mathbf{D}_{s(1)}\right)^T\mathbf{Z}_{s(1)}{}^T + \mathbf{I}$. (3.19) can be written as the following equation.

$$Q\left(\{\boldsymbol{\phi}_s\}_{s=1}^{\mathbb{S}}\right) = \sum_{s=1}^{\mathbb{S}} l(\boldsymbol{\phi}_s) - \lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps(1)}}{\tilde{\beta}_{ps(1)}}\right)^2} - \lambda_{N2}\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs(1)}}{\tilde{d}_{qs(1)}}\right)^2}\ , \quad (3.20)$$

where $l(\boldsymbol{\phi}_s) = -\frac{\mathbf{1}}{\mathbf{2}}\left\{log\left|\tilde{\mathbf{V}}_{s(1)}\right| + \left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)^T\tilde{\mathbf{V}}_{s(1)}{}^{-1}\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right\}$.

The following Theorems hold under common regularity conditions. Theorems and 1 and 2 together show that the proposed method can identify the true model with probably tending to one. Theorem 2.3 indicates that the estimator in the proposed method enjoys the oracle property. Proofs of the Theorems can be found in Supplementary Material.

Theorem 2.2.1: If $\frac{\lambda_{N1}}{\sqrt{N}} \longrightarrow 0$ and $\frac{\lambda_{N2}}{\sqrt{N}} \longrightarrow 0$, then there exists a local maximizer $\widehat{\boldsymbol{\phi}} = \begin{pmatrix} \widehat{\boldsymbol{\phi}}^1 \\ \mathbf{0} \end{pmatrix}$ of $Q\left\{\begin{pmatrix} \boldsymbol{\phi}^1 \\ \mathbf{0} \end{pmatrix}\right\}$ such that $\widehat{\boldsymbol{\phi}}^1$ is $\sqrt{N}$ consistent for $\widetilde{\boldsymbol{\phi}}^1$.

Theorem 2.2: If $\lambda_{N1} \longrightarrow \infty$ and $\lambda_{N2} \longrightarrow \infty$, then with probability tending to one for any given $\boldsymbol{\phi}^1$satisfying $\left\|\boldsymbol{\phi}^1 - \widetilde{\boldsymbol{\phi}}^1\right\| \le \frac{C}{\sqrt{N}}$ and some constant $C > 0$,

$$Q\left\{\begin{pmatrix} \boldsymbol{\phi}^1 \\ \mathbf{0} \end{pmatrix}\right\} = \underset{\|\boldsymbol{\phi}^2\| \le \frac{C}{\sqrt{N}}}{max} \; Q\left\{\begin{pmatrix} \boldsymbol{\phi}^1 \\ \boldsymbol{\phi}^2 \end{pmatrix}\right\}.$$

Theorem 2.3: If $\lambda_{N1} \longrightarrow \infty$, $\lambda_{N2} \longrightarrow \infty$, $\frac{\lambda_{N1}}{\sqrt{N}} \longrightarrow 0$ and $\frac{\lambda_{N2}}{\sqrt{N}} \longrightarrow 0$, then

$$\sqrt{N}\, I(\widetilde{\boldsymbol{\phi}}^1)\left(\left(\widehat{\boldsymbol{\phi}}^1 - \widetilde{\boldsymbol{\phi}}^1\right) + I(\widetilde{\boldsymbol{\phi}}^1)^{-1}(\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_{\mathbb{S}}^T)^T\right) \longrightarrow_d N\left(0, I(\widetilde{\boldsymbol{\phi}}^1)\right),$$

where

$$\boldsymbol{v}_s = \left(\frac{\lambda_{N1}}{N} \times \frac{\frac{\beta_{1s}}{\left|\widetilde{\beta}_{1s}\right|^2}}{\sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{1s}}{\widetilde{\beta}_{1s}}\right)^2}}, \ldots, \frac{\lambda_{N1}}{N} \times \frac{\frac{\beta_{\mathbb{P}'s}}{\left|\widetilde{\beta}_{\mathbb{P}'s}\right|^2}}{\sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{\mathbb{P}'s}}{\widetilde{\beta}_{\mathbb{P}'s}}\right)^2}}, \frac{\lambda_{N2}}{N} \times \frac{\frac{d_{1s}}{\left|\widetilde{d}_s\right|^2}}{\sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{1s}}{\widetilde{d}_{1s}}\right)^2}}, \ldots, \frac{\lambda_{N2}}{N} \times \frac{\frac{d_{\mathbb{Q}'s}}{\left|\widetilde{d}_{\mathbb{Q}'s}\right|^2}}{\sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{\mathbb{Q}'s}}{\widetilde{d}_{\mathbb{Q}'s}}\right)^2}}, 0, \ldots, 0 \right)^T,$$

$s = 1, \ldots\ , \mathbb{S}$. The number of zeros at the end of $\boldsymbol{v}_s$ is $\left|\widetilde{\boldsymbol{\gamma}}_{s1}\right|$. $I(\widetilde{\boldsymbol{\phi}}^1)$ is the Fisher Information with $\boldsymbol{\phi}^2 = \mathbf{0}$.

3.6 Simulation Studies

In this section, we present the performance of our proposed method. For each simulation setting, we report the True Positive Rate (TPR), True Negative Rate (TNR), and accuracy for the fixed effect identification, and those for the random effect identification, under the optimal λ_1 and λ_2 chosen by the BIC criterion in (3.18). The TPR measures the proportion of identified non-zero fixed (random) effects that are truly non-zero. The TNR measures the proportion of identified zero fixed (random) effects that are truly zero. The accuracy measures the proportion of fixed (random) effects that are correctly identified. For comparison, we also fit a multilevel model with an adaptive l_1-penalty for each response variable separately and use a BIC criterion to choose the tuning parameter for the adaptive l_1-penalty. This is the method proposed by Bondell *et al.* (2010), and is referred to as the "competing method" in the rest of this paper.

3.6.1 Study for the Impact of Effect Size, Sample Size, and Sample Distribution on Performance

We conduct four experiments, in all of which there are three response variables. For each response, we consider the multilevel model to consist of ten and four fixed and random effects, respectively. Because both the proposed and competing methods enable variable selection, we set three out of the ten fixed effects and three out of the four random effects to be non-zero. Furthermore, to induce correlation between the models of the three responses, we set the same fixed and random effects to be non-zero across all the responses.

In the first experiment, we set the fixed effects for the three responses to be $\boldsymbol{\beta}_1 = (0.1,1,1,0,\dots,0)^T$, $\boldsymbol{\beta}_2 = (1,0.1,1,0,\dots,0)^T$, and $\boldsymbol{\beta}_3 = (1,1,0.1,0,\dots,0)^T$, and the random effects to be $\mathbf{d}_1 = (1,1,0.3,0)^T$, $\mathbf{d}_2 = (1,0.3,1,0)^T$, $\mathbf{d}_3 = (1,1,1,0)^T$, $\boldsymbol{\Gamma}_{\mathrm{s}} = \mathbf{I}$, and $\sigma_{\mathrm{s}}^2 = 1$ for $\mathrm{s} = 1,2,3$. The sample sizes are set to be $N = 30$ units and $n_i = 5$ individuals per unit. Furthermore, for each unit i, the data of predictors corresponding to the fixed effects, i.e.,

$\mathbf{W}_i$, are generated as follows: To account for the possible correlation structure among the predictors, we first generate a random matrix of size $n_i \times \mathbb{P} = 5 \times 10$, whose elements are independently sampled from a $N(0,1)$ distribution. Denote this random matrix by $\mathbf{V}_i$. Then, let $\mathbf{W}_{i,p} = \frac{\mathbf{V}_{i,p}+\mathbf{V}_{i,\mathbb{P}+1}}{\sqrt{2}}$, $p = 1, \dots, \mathbb{P}$. $\mathbf{W}_{i,p}$ and $\mathbf{V}_{i,p}$ are the p–th column of matrix $\mathbf{W}_i$ and $\mathbf{V}_i$. , respectively. This idea of generating corrected predictors has also been adopted by a few other papers (Yuan et al, 2006). Furthermore, we generate the data of predictors corresponding to the random effects, i.e., $\mathbf{Z}_i$, by setting $\mathbf{Z}_i = \left(\mathbf{1}_{n_i}, \mathbf{W}_{i,1}, \mathbf{W}_{i,2}, \mathbf{W}_{i,3}\right)$. Finally, the data for each response variable are generated according to (3.3). We apply the proposed and competing methods to the data. Table 3 summarizes the result based on 200 simulation runs. Both methods have high TNRs. The proposed method also has high TPRs, while the TPRs for the competing method are significantly lower. This is because the competing method fails to identify the small fixed and random effects of 0.1 and 0.3^2, while the proposed method is able to do so due to its ability of performing a joint estimation across all the responses.

Table 3: Comparison between the proposed and competing methods ($N = 30$, $n_i = 5$, $\mathbf{d}_3 = (1,1,1,0)^T$): average (standard deviation) of TPR/TNR/accuracy

	Fixed effect identification			Random effect identification		
	TPR	TNR	Accuracy	TPR	TNR	Accuracy
Proposed method	99.4% (2.4%)	99.0% (4.1%)	99.1% (3.0%)	96.0% (8.5%)	98.9% (7.4%)	96.7% (6.5%)
Competing method	69.3% (5.8%)	96.2% (3.8%)	88.2% (2.9%)	75.0% (8.8%)	99.4% (4.3%)	81.1% (6.7%)

The second experiment aims to evaluate the performance with a greater number of small random effects. To this end, we modify the setting of the first experiment by changing

$\mathbf{d}_3 = (1,1,1,0)^T$ to $\mathbf{d}_3 = (1,0.3,0.3,0)^T$. Table 4 reports the performance. Compared with Table 3, we can see that the TPR of random effect identification for the competing method is deteriorated significantly, while this performance of the proposed method remains high.

Table 4: Comparison between the proposed and competing methods ($N = 30$, $n_i = 5$, $\mathbf{d}_3 = (1,0.3,0.3,0)^T$): average (standard deviation) of TPR/TNR/accuracy

	Fixed effect identification			Random effect identification		
	TPR	TNR	Accuracy	TPR	TNR	Accuracy
Proposed method	99.6% (2.0%)	99.1% (4.3%)	99.3% (3.1%)	90.8% (14.2%)	99.7% (3.0%)	93.1% (10.8%)
Competing method	69.8% (5.9%)	96.2% (4.1%)	88.3% (3.5%)	59.6% (7.2%)	99.5% (4.2%)	69.6% (5.5%)

The third experiment aims to show the sample size impact. We keep the setting of the second experiment but increase the unit sample size from $n_i = 5$ to $n_i = 10$. Table 5 reports the performance. Compared with Table 4, we observe that doubling the sample size increases the average and decreases the standard deviation of the TPR for random effect identification by the proposed method. This performance improvement is less obvious by the competing method.

Table 5: Comparison between the proposed and competing methods ($N = 30$, $n_i = 10$, $\mathbf{d}_3 = (1,0.3,0.3,0)^T$): average (standard deviation) of TPR/TNR/accuracy

	Fixed effect identification			Random effect identification		
	TPR	TNR	Accuracy	TPR	TNR	Accuracy
Proposed method	99.4% (2.4%)	100% (0.0%)	99.8% (0.7%)	97.4% (5.0%)	98.7% (8.1%)	97.7% (4.7%)
Competing method	70.3% (5.7%)	97.3% (3.3%)	89.2% (2.9%)	62.5% (7.8%)	99.3% (4.7%)	71.7% (5.9%)

In the fourth experiment, noting that the total sample size is a product of the number of units and the unit sample size, i.e., $N \times n_i$, we would like to study the impact of the sample distribution between N and n_i on the performance. To this end, we keep the setting of the second experiment, which has a total sample size of $N \times n_i = 30 \times 5 = 150$, but re-distribute the samples to have $N = 15$ and $n_i = 10$. Table 6 reports the performance. Compared with Table 4, we observe that the sample re-distribution does not change the performance of both methods. Therefore, it is more likely that the performance of the methods is affected by the total sample size.

Table 6: Comparison between the proposed and competing methods ($N = 15, n_i = 10$, $\mathbf{d}_3 = (1{,}0.3{,}0.3{,}0)^T$): average (standard deviation) of TPR/TNR/accuracy

	Fixed effect identification			Random effect identification		
	TPR	TNR	Accuracy	TPR	TNR	Accuracy
Proposed method	99.6% (2.0%)	99.8% (1.9%)	99.7% (1.4%)	90.9% (16.5%)	99.4% (4.3%)	93.0% (12.3%)
Competing method	70.1% (5.2%)	97.2% (3.6%)	89.1% (3.0%)	60.3% (7.8%)	98.3% (7.4%)	69.8% (6.1%)

3.6.2 Study for the Impact of the Number of Response Variables 2.1 Introduction

All the experiments in the previous section have three response variables. The focus of this section is to study performance with respect to different numbers of response variables. We consider scenarios with two, five, and eight responses, in all of which there are six fixed effects (two being non-zero) and three random effects (two being non-zero). Specifically, the first scenario has the fixed effects for the two responses as $\boldsymbol{\beta}_1 = (0.1{,}1{,}0, \dots ,0)^T$ and $\boldsymbol{\beta}_2 = (-1{,}0.1{,}0, \dots ,0)^T$, and random effects as $\mathbf{d}_1 = (1{,}0.3{,}0)^T$ and

$\mathbf{d}_2 = (1,0.3,0)^T$. The second scenario has the same fixed and random effects for the first two responses as the first scenario, and $\boldsymbol{\beta}_3 = (0.5,1,0,\dots,0)^T$, $\boldsymbol{\beta}_4 = (-1,1,0,\dots,0)^T$, $\boldsymbol{\beta}_5 = (1,0.5,0,\dots,0)^T$, $\mathbf{d}_3 = (1,1,0)^T$, $\mathbf{d}_4 = (1,0.8,0)^T$, and $\mathbf{d}_5 = (1,1,0)^T$ for the remaining three responses. The third scenario has the same fixed and random effects for the first five responses as the second scenario, and $\boldsymbol{\beta}_6 = \boldsymbol{\beta}_3$, $\boldsymbol{\beta}_7 = \boldsymbol{\beta}_4$, $\boldsymbol{\beta}_8 = \boldsymbol{\beta}_5$, $\mathbf{d}_6 = \mathbf{d}_3$, $\mathbf{d}_7 = \mathbf{d}_4$, and $_8 = \mathbf{d}_5$ for the remaining three responses. The sample sizes are set to be $N = 20$ units and $n_i = 5$ individuals per unit. Note that we purposely choose a small sample size so that the performance in the two-response scenario is not good. This would allow us to see if adding more responses could remedy the sample size shortage. Under these settings, the data is generated in the same way as Section 6.1. Table 7 summarizes the results. With two responses, the TPR for random effect identification by the proposed method is low ($53.8\% \pm 0.12\%$). This is significantly improved by having five responses ($90\% \pm 0.17\%, p_{value} < 0.0001$), which is further improved by having eight responses (($94.7\% \pm 0.1\%, p_{value} = 0.015$). There is no significant difference in the TPR for fixed effect identification across the three scenarios. However, this does not mean that having more responses would not improve the TPR for fixed effects. It is simply because the TPR with two responses is already very high, leaving little room to demonstrating improvement. In contrast, the competing method has low TPR for both random and fixed effects, and adding more responses does not help. Furthermore, in terms of TNR, both methods perform well across all three scenarios with the proposed method having slightly higher TNR for fixed effects.

Table 7: Comparison between the proposed and competing method with varying numbers of responses

Number of responses		Fixed effect identification			Random effect identification		
		TPR	TNR	Accuracy	TPR	TNR	Accuracy
2	Proposed method	100% (0.0%)	98.8% (0.06%)	99.2% (0.04%)	53.8% (0.12%)	100% (0.0%)	69.2% (0.08%)
	Competing method	60% (0.15%)	94.4% (0.09%)	82.9% (0.08%)	53.8% (0.09%)	100% (0.0%)	69.2% (0.06%)
5	Proposed method	98.8% (0.06%)	100% (0.0%)	99.6% (0.02%)	90% (0.17%)	100% (0.0%)	93.3% (0.11%)
	Competing method	61.2% (0.13%)	93.8% (0.09%)	82.9% (0.07%)	51.3% (0.06%)	100% (0.0%)	67.5% (0.04%)
8	Proposed method	98.7% (0.06%)	98.7% (0.06%)	98.7% (0.04%)	94.7% (0.1%)	97.4% (0.1%)	95.6% (0.09%)
	Competing method	57.9% (0.15%)	93.4% (0.08%)	81.6% (0.08%)	56.6% (0.1%)	100% (0.0%)	71.1% (0.08%)

3.6.3 BIC vs. AIC

While BIC has been used for choosing the tuning parameters in previous simulation studies, it is of interest to study if other criteria such as AIC may offer some advantage. To this end, we repeat the experiments in Section 6.2 but selecting the tuning parameters of the proposed and competing methods using AIC. The results are shown in Table 8. Compared with Table 7, we can see that AIC has significantly lower TNR than BIC in both the proposed and competing methods, while the TPR performances of the two criteria are similar. Also, the standard deviations of TPR and TNR under AIC are much higher than BIC, indicating a less stable performance of AIC. While several previous studies have suggested advantages of AIC over BIC (Burnham & Anderson, 2002; 2004; Yang, 2005), these studies did not specifically compare the two criteria for multilevel models. Our experiments, on the other hand, empirically demonstrate better performance of BIC for multilevel models. Theoretical explanation behind this empirical observation is left for future research.

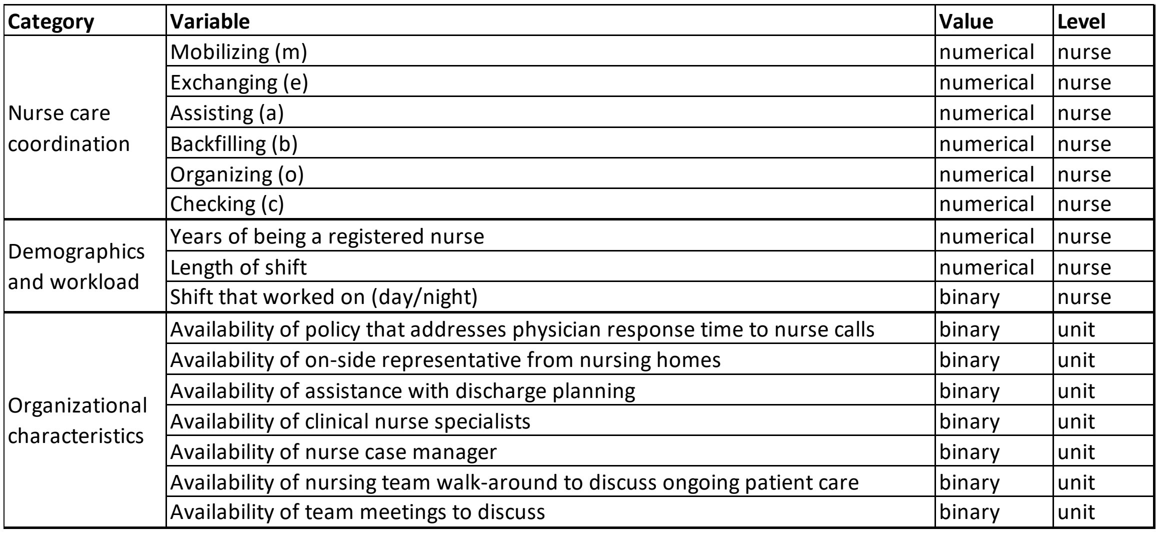
Table 8: Results of the experiments in Section 6.2 using AIC for tuning parameter selection

Number of responses		Fixed effect identification			Random effect identification		
		TPR	TNR	Accuracy	TPR	TNR	Accuracy
2	Proposed method	100% (0.0%)	71.3% (0.45%)	80.8% (0.3%)	58.8% (0.19%)	97.5% (0.11%)	71.7% (0.11%)
	Competing method	66.3% (0.15%)	80.6% (0.15%)	75.8% (0.13%)	62.5% (0.15%)	100% (0.0%)	75% (0.1%)
5	Proposed method	98.8% (0.06%)	76.3% (0.35%)	83.8% (0.23%)	95% (0.13%)	77.5% (0.41%)	89.2% (0.17%)
	Competing method	66.3% (0.12%)	83.8% (0.12%)	77.9% (0.09%)	57.5% (0.12%)	97.5% (0.11%)	70.8% (0.07%)
8	Proposed method	98.7% (0.06%)	90.8% (0.24%)	93.4% (0.16%)	94.7% (0.1%)	79% (0.3%)	89.5% (0.14%)
	Competing method	61.8% (0.19%)	82.9% (0.15%)	75.9% (0.13%)	67.1% (0.19%)	97.4% (0.11%)	77.2% (0.13%)

3.7 Application in Nurse Care Coordination

We apply the proposed method to a dataset created using the NCCI (Duva, 2010; Shuai et al, 2014). The dataset includes 614 nurses from 32 medical-surgical units of four hospitals in the metro Atlanta area. These data are used in this paper with permission. Three categories of variables are measured in the dataset, as shown in Table 9. The first category consists of nurse care coordination activities belonging to six constructs, “m”, “e”, “a”, “b”, “o”, and “c”. For example, “I initiate actions to get my nursing team members to do what is needed to keep my patients on their plan” is an activity belonging to “m”. “I communicate information to my interdisciplinary team members they need to know to carry out their patient care activities or to make changes in their plan of care” is an activity belonging to “e”. “I ask my nursing team members to assist me with my patient activities when I am tied up with one or more of my patients” is an activity belonging to “a”. “I prompt my interdisciplinary team to do the work they are responsible for so I can get my own work done and keep patients on their plan of care” is an activity belonging to “b”. “I organize the supplies that I need to be able to keep the care of my patients on track” is an activity belonging to “o”. “I perform my patient assessments so that they will be useful to everyone on the team” is an activity belonging to “c”. A total of 45 activities were measured in the form of 45 questions asked to the nurses in a questionnaire. The answer to each question is a five-point likert-type scale with higher scores representing greater amounts of the corresponding activity. The correspondence between each question and the latent construct is known by the design of the NCCI. To get a measurement for each construct, we average the scores of the corresponding questions. The second category of variables in the dataset are nurse demographic and workload variables, and three of them are included in this study as shown in Table 9. The third category consists of organizational characteristic variables of nurses’ practice environment, i.e., their units. Seven variables are included, which measure the availability of certain infrastructure, technology, and policies that may potentially facilitate nurse care coordination.

Table 9: Description of the NCCI data

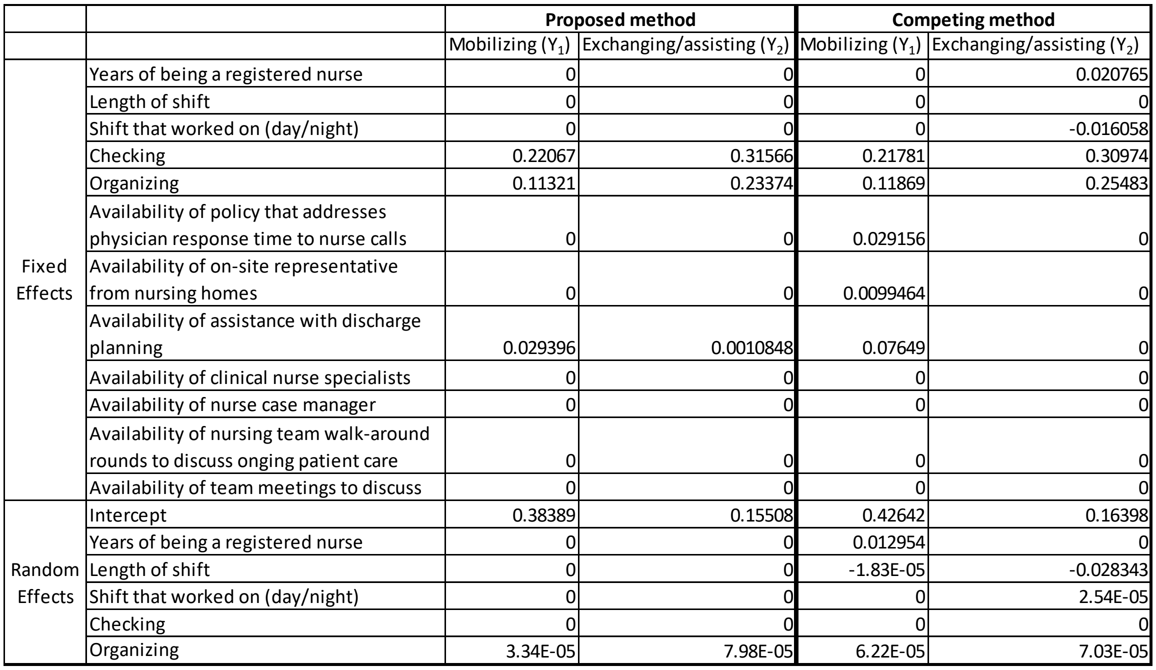


Category	Variable	Value	Level
Nurse care coordination	Mobilizing (m)	numerical	nurse
	Exchanging (e)	numerical	nurse
	Assisting (a)	numerical	nurse
	Backfilling (b)	numerical	nurse
	Organizing (o)	numerical	nurse
	Checking (c)	numerical	nurse
Demographics and workload	Years of being a registered nurse	numerical	nurse
	Length of shift	numerical	nurse
	Shift that worked on (day/night)	binary	nurse
Organizational characteristics	Availability of policy that addresses physician response time to nurse calls	binary	unit
	Availability of on-side representative from nursing homes	binary	unit
	Availability of assistance with discharge planning	binary	unit
	Availability of clinical nurse specialists	binary	unit
	Availability of nurse case manager	binary	unit
	Availability of nursing team walk-around to discuss ongoing patient care	binary	unit
	Availability of team meetings to discuss	binary	unit

Next, we discuss roles of the variables in the proposed model. Among the six constructs, “m”, “e”, “a”, and “b” measure the inter-dependent activities among nurses, i.e., their coordination, while “o” and “c” measure their independent activities that are instrumental to their coordination. We focus on three out of the four inter-dependent constructs, “m”, “e”, and “a”, because there has been some controversy over whether “b” plays a positive or negative role in care coordination (Duva, 2010). Among “m”, “e”, and “a”, “m” is treated as one response variable. Originally, we tried making “e” and “a” two other response variables, but the result was not as good as combining them into one response variable. This is an interesting finding: On the one hand, “a” and “e” are different in the sense that the former concerns coordination between nurses while the latter concerns coordination initiated by nurses but with other professionals in the patient care team. On the other hand, the fact that combining “e” and “a” gives better model performance is an indication that there might be a higher-level abstraction making “e” and “a” more similar to each other in terms of characterizing care coordination than to other constructs like “m”. Pending further investigation, we focus on presenting the results of modeling two response variables in this paper, i.e., “m” and combined “e/a”. Furthermore, we include demographic and workload variables as well as “o” and “c” as individual-level predictors, and organizational characteristics as unit-level predictors.

We apply the proposed method to link the predictors with the two responses, “m” and “e/a”. Two tuning parameters λ_1 and λ_2 are chosen by BIC. The responses and predictors are standardized, so that the fixed effects do not include an intercept but the random effects still do. In Table 10 under “Proposed method”, we show the estimated fixed effects and the variances of the estimated covariance matrix of random effects. Several observations can be drawn: 1) Same effects are found to be zero (non-zero) in both responses, although the magnitudes of non-zero effects are somewhat different. This is expected because of the use of an adaptive l_{21}-penalty in our method. 2) “O” and “c” are found to have non-zero fixed effects among the five individual-level predictors. This makes sense because independent nurse care activities like “c” and “o” form the basis for nurses to perform interdependent coordination activities. The positive signs of the effects of “c” and “o” suggest that such independent activities have a positive impact on interdependent coordination activities. 3) Among the seven unit-level predictors, “the availability of assistance with discharge planning in the unit” is found to have a non-zero fixed effect. The positive sign of this effect suggests that providing assistance with discharge planning in a unit helps create a positive practice environment for the nurses in the unit to conduct care coordination. 4) Among the random effects, “o” is found to be non-zero in addition to the intercept. This reinforces the important role of independent nurse care activities on care coordination especially “o”.

Table 10: Estimated fixed effect regression coefficients and random effect variances by the proposed and competing methods



		Proposed method		Competing method	
		Mobilizing (Y_1)	Exchanging/assisting (Y_2)	Mobilizing (Y_1)	Exchanging/assisting (Y_2)
Fixed Effects	Years of being a registered nurse	0	0	0	0.020765
	Length of shift	0	0	0	0
	Shift that worked on (day/night)	0	0	0	-0.016058
	Checking	0.22067	0.31566	0.21781	0.30974
	Organizing	0.11321	0.23374	0.11869	0.25483
	Availability of policy that addresses physician response time to nurse calls	0	0	0.029156	0
	Availability of on-site representative from nursing homes	0	0	0.0099464	0
	Availability of assistance with discharge planning	0.029396	0.0010848	0.07649	0
	Availability of clinical nurse specialists	0	0	0	0
	Availability of nurse case manager	0	0	0	0
	Availability of nursing team walk-around rounds to discuss onging patient care	0	0	0	0
	Availability of team meetings to discuss	0	0	0	0
Random Effects	Intercept	0.38389	0.15508	0.42642	0.16398
	Years of being a registered nurse	0	0	0.012954	0
	Length of shift	0	0	-1.83E-05	-0.028343
	Shift that worked on (day/night)	0	0	0	2.54E-05
	Checking	0	0	0	0
	Organizing	3.34E-05	7.98E-05	6.22E-05	7.03E-05

For comparison, we also apply the competing method to model each response variable separately. The result is presented in Table 10 under “Competing method”. The competing method also finds “o” and “c” to have non-zero fixed effects for the two responses, but two additional non-zero fixed effects for the response “e/a”. Furthermore, the competing method finds three unit-level predictors to have non-zero fixed effects, including the predictor of “the availability of assistance with discharge planning in the unit” that is also found by the proposed method, for the response “m”. However, all unit-level predictors for the response “e/a” are found to have zero fixed effects. Lastly, although the competing method also finds “o” to have a non-zero random effect, it finds more non-zero random effects for the two responses than the proposed method. In summary, the proposed method finds 10 non-zero fixed and random effects, whereas the competing method finds 17.

Finally, we would like to compare the prediction performances of the two methods. A common metric is the average Mean Squared Prediction Error (MSPE), $\overline{\text{MSPE}}$, through a K-fold Cross Validation (CV). Because both our proposed method and the competing method are multilevel models, the generic CV procedure needs some modification. Specifically, the K-fold division is done within each unit, i.e., the samples within each of the 32 units are divided into K folds. Then, the samples of $K-1$ folds within each unit are pooled together and used for training. Furthermore, the trained model is applied to the pooled remaining one fold from each unit to compute a MSPE. This modified CV procedure is to make sure that the training model includes at least some samples from each unit. We apply this modified CV procedure to the proposed and competing methods. Noting that the results could vary depending on the number of folds in the CV, we vary K from 2 to 8. Also noting that the results could vary even with a fixed K because of the randomness of the CV partition, we run the CV partition three times for each fixed K. This procedure results in $8 \times 3 = 24$ pairs of $\overline{\text{MSPE}}$ to compare the proposed and competing methods. We count the proportion of times the proposed method has a lower $\overline{\text{MSPE}}$ than the competing method and conduct hypothesis testing for this proportion. The hypothesis testing yields a p value of 0.037, indicating that the proposed method has a significantly smaller $\overline{\text{MSPE}}$. A smaller prediction error provides evidence to support the appropriateness of the joint modelling of two responses by the proposed method.

3.8 Conclusion

In this paper, we developed a multi-response multilevel model to characterize the relationship between nurse care coordination and nurses' practice environment, demographics, and workload. Our model development was driven by the recently available NCCI that was the first of its kind allowing quantitative data to be collected to measure nurse care coordination, and was further driven by the unique data structure that required a joint modeling of multiple response variables in relation to predictors at two levels (unit level and individual/nurse level). Our model development included a unique formulation that used two adaptive l_{21}-penalties to enable joint fixed effect selection and joint random effect selection across the multiple responses, and an efficient BCD algorithm integrated with an EM framework for parameter estimation. We performed theoretical analysis to reveal that the reason for the proposed method to outperform a separate modeling of each response was the consideration of a sum of squares of the effects in all other responses when estimating the effect in one response. In this way, the estimate was less shrunk, leading to better identification of small non-zero effects. We conducted simulation studies, which demonstrated that the proposed method achieved high TFRs and TNRs for both fixed and random effect identification, and the TFRs of our method were consistently higher than the competing method that modelled each response variable separately, especially when the non-zero effects were small. Our simulation studies also revealed the sample size and sample distribution influences on the proposed method in comparison with the competing method. We applied our method to the dataset created using the NCCI to model the relationship between two response variables measuring nurse care coordination and five level-one predictors characterizing nurses' demographics and workload and seven level-two predictors characterizing the practice environment of the nurses' residing units. Our method achieved a significantly higher prediction accuracy on the two response variables compared with the competing method. Our method also identified a significantly smaller number of predictors to have non-zero effects than the competing method and these predictors were shared by the two responses. A model that requires fewer predictors without sacrificing the prediction accuracy and that is able to use the same subset of predictors to predict multiple response variables is desirable in practice. This means a greater saving in the effort and cost for data acquisition, and a potential saving in the cost of intervention if the significant predictors can be confirmed to "causally" affect the response variables of care coordination. In particular, our result showed that two independent nurse care activities, "c" and "o", and one unit characteristic, "the availability of assistance with discharge planning in the unit", had a significant positive relation to care coordination.

We would like to point out several limitations of the study in this paper, which also open opportunities for future research. First, although our study identified three significant predictors for nurse care coordination, the causality between them and care coordination is yet to be established. Finding not only predictors but also causal factors that influence care coordination is important for designing improvement strategies and the best practices. Second, we cannot rule out the possibility that the predictors not selected by our method may have a non-zero effect on the responses. They are just not as significant as the predictors that were selected under the *limited* and *specific* samples in our dataset. More data, especially data across hospitals at different geographical locations beyond the Atlanta area, are needed to validate and generalize the current findings. Third, our model formulation allows a selection of the same subset of predictors across all responses, which is a strength and a restriction. An immediate future extension of our model is to adopt a different choice for the penalties to allow for both joint variable selection across the responses and unique variable section within each response, such as the sparse group lasso penalty (Simon et al. 2013).

The long-term goal of this research is to inform interventions to improve staff nurse care coordination within hospital units that would in turn lead to improved patient outcomes, e.g., shorter length of stay, few medication errors, less likelihood for re-admission, and greater satisfaction. Achieving this goal is important to the current health care system because many hospitalized patients nowadays have multiple co-existing chronic illnesses demanding a great amount of coordinated care within the health care team especially the nurses who are the patients' "ever-present" care professionals. Without effective nurse care coordination, these patients would be at an elevated risk for poor outcomes, which not only decrease their quality of life but also result in unnecessary costs to the health care system.

CHAPTER 4

A SEQUENTIAL TREE-BASED CLASSIFICATION FOR PERSONALIZED BIOMARKER TESTING OF ALZHEIMER’S DISEASE RISK

4.1 Introduction

Alzheimer’s disease (AD) is an irreversible neurodegenerative disease of the brain characterized by debilitating impairment in daily activities and cognitive decline. More than five million people in the U.S. currently have AD, and the number is expected to increase to 16 million by 2050. The direct health care cost is over $200 billion per year and projected to reach $1.2 trillion by 2050. Recent clinical trials designed to treat AD at the mild-to-moderate dementia phase have been largely unsuccessful. There is a growing consensus that treatment should target the disease in its early phases before irreversible brain damage occurs. Mild cognitive impairment (MCI) is a prodromal phase of AD at which patients experience cognitive decline but have not developed dementia. Treatment at the MCI phase could potentially delay the progression to AD or even prevent the patient from developing AD, and therefore has considerable interest.

Important to early detection and prevention of AD is the use of biomarkers to precisely predict the conversion of MCI to AD within a clinical time of interest. According to the new diagnostic guidelines recommended by the National Institute on Aging and the Alzheimer’s Association [Albert et al. 2011], the important biomarkers include those measuring Aβ deposition in plagues and those linked to downstream neuronal degeneration or injury processes, such as the phosphorylated tau (p-tau) level in cerebrospinal fluid

(CSF), mean cerebral metabolism on ^{18}F fludeoxyglucose positron emission tomography (FDG-PET), and hippocampal volume on structural magnetic resonance imaging (MRI).

There has been a vast amount of studies aiming at using biomarker data to predict the conversion of MCI patients to AD (Barnes et al. 2014, Cui et al. 2011, Heister et al. 2011, Hinrichs et al. 2011, Jack et al. 2010, Risacher et al. 2009, Wee et al. 2013, Ye et al. 2012, Yu et al. 2012, Zhang et al. 2012, Zhang et al. 2012). A particular area of study with clear clinical relevance is to achieve this prediction using *baseline* biomarker measurements (Barnes et al. 2014, Cui et al. 2011, Heister et al. 2011, Hinrichs et al. 2011, Jack et al. 2010, Risacher et al. 2009, Wee et al. 2013, Ye et al. 2012, Yu et al. 2012, Zhang et al. 2012, Zhang et al. 2012)Although using longitudinal repeated measurements of the same biomarkers has a potential to improve the prediction accuracy, this prolongs the diagnostic time span and makes clinical trials more time consuming and costly. In using baseline biomarkers to predict MCI conversion, most of the existing studies built statistical classification models that assign each MCI patient to be a converter or non-converter using a pre-trained model. The accuracy on large public datasets like the Alzheimer's Disease Neuroimaging Initiative (ADNI) has been reported to be between 60-72%. The existing research has a few limitations:

First, the prediction accuracy is unsatisfactory. This can be attributed to the heterogeneity of MCI patients. That is, there may be subgroups across which different biomarkers or different combinations of biomarkers are useful for predicting conversion to AD. MCI heterogeneity is a known challenge in AD studies and has been reported in many papers (Cerami et al. 2015, Yu et al. 2012). A recent study using the comprehensive dataset collected through the worldwide ADNI project revealed that there is little agreement in using different biomarkers for predicting the conversion of MCI to AD, such as the p-tau level in CSF, mean cerebral metabolism on FDG-PET, and hippocampal volume on MRI. Conflicting predictions by the different biomarkers happen in roughly every third MCI patient (Alexopoulos et al. 2014). This provides strong evidence that MCI is a heterogeneous group and that the existing research of "one-model-fits-all (OMFA)" is unlikely to work well. Here, OMFA means building one classification model, which assumes the same multivariate association of biomarkers with conversion/non-conversion, across all the MCI patients.

Second, the existing research is bounded by an inherent limitation of conventional classification models that require the biomarkers to be measured *all at once*. This is because a conventional classification model takes the form of $Y = f\left(X_1, \dots, X_p\right)$, where $X_1, \dots, X_p$ are biomarkers and Y is a binary variable of conversion or. non-conversion. When using this model to predict an MCI patient, data on all the biomarkers included in the model, i.e. $X_1, \dots, X_p$, must be available. Otherwise, the model cannot be applied. Almost all the commonly used classification models have this limitation, such as logistic regression, discriminate analysis, support vector machine, and artificial neural network. However, requiring biomarkers to be available all at once at the time of making a prediction/diagnosis does not reflect the reality of clinical practices in which biomarkers are typically measured *sequentially*. That is, the most predictive biomarker is first tested for a patient. If the result is conclusive, e.g., the patient is predicted to be a converter or non-converter with a high confidence, no other biomarkers need to be tested. Otherwise, if the result from the first biomarker is inconclusive, an additional biomarker may be tested. More biomarkers may be added until a conclusive diagnosis is reached. It is also possible that no conclusive diagnosis can be reached even with all the biomarkers having been tested, which is common for early stages of a disease. If this happens, the patient will be asked to come back to re-test during a follow up visit.

Lastly, in most existing research that uses biomarkers to predict MCI conversion, biomarkers are treated as numerical variables. Although the raw biomarker measurement is on a numerical scale, clinical interpretation is typically based on a cutoff that dichotomizes the biomarker into “positive” and “negative”. For example, $1.21, 3260\ mm^3$, and 23 $\mathrm{pg/mL}$ are the currently used clinical cutoffs for the mean cerebral metabolism on FDG-PET, hippocampal volume on MRI, and p-tau level in CSF, respectively (Jack et al. 2008, Kim et al. 2011). Both approaches have limitations: Using the raw, numerical measurement of biomarkers is clinically inconvenient. Also, there may be measurement errors associated with the testing instrument and bias due to patient’s health condition and exposure to environmental factors that potentially confound with the target disease. This makes the use of raw biomarker measurement a less robust approach. On the other hand, using a single cutoff as in the current clinical practice is an over-simplification by ignoring the quantitative relationship between biomarker values and disease risks. Between using the numerical measurement and a single cutoff, a “middle” approach that uses more discretized levels of a biomarker may be more appropriate.

To overcome the aforementioned limitations of the existing research, we propose a sequential tree-based classifier (STC) for predicting MCI patients’ risks of converting to

AD in this paper. Compared with conventional classification models, STC does not require all the biomarkers be available for every patient at the time of the prediction, but sequentially adds biomarkers only when necessary. Another difference is that not like conventional classification models that enforce a binary decision (conversion vs. non-conversion) for each patient, STC classifies patients into three categories: a clinically-defined high-risk (HR) category, a clinically-defined low-risk (LR) category, and an inconclusive category. The HR and LR categories includes MCI patients that will convert to AD within a clinical time of interest with a high and a low probability, e.g., 80% and 20%, respectively. HR patients need immediate medical attention. LR patients can be cleared of the disease or put on long-term observation. Patients falling into the inconclusive category at the baseline may be asked for a re-test in a follow up visit. In essence, STC achieves the sequential, as-needed use of biomarkers and the three-category classification by finding an optimal sequence of biomarkers and two-sided cutoffs of each biomarker that satisfy the HR and LR requirement while minimizing the proportion of MCI patients classified as inconclusive. Also, STC is personalized because it allows patient-specific information such as age, gender, education level, and genotyping to be included to help identify patient-specific cutoffs for each biomarker. Additionally, STC is flexible in the sense that it can be developed depending on the available biomarkers in a clinic. Each clinic has a different level of resources, which may limit its biomarker testing capability. A model has limited use if it has to assume the same biomarkers to be tested across different clinics. Finally, we would like to stress that STC approaches the challenge of low accuracy in predicting MCI conversion, which is faced by the existing research, from a different angle.

That is, a target prediction accuracy is first defined, which is reflected by HR and LR, and it is then used by STC for identifying groups of patients for which this accuracy can be reached. This capability has tremendous value for disease management and patient selection in clinical trials.

We apply STC to two important clinical applications using the ADNI data. One is to predict/classify MCI patients into HR, LR, or inconclusive categories so that appropriate medical decisions can be made for each patient. The other application is to help patient selection in clinical trials, i.e., identify a sub-cohort of MCI patients with a HR of converting to AD, as these patients are more likely to benefit from the intervention being tested. The remaining of this paper is organized as follows: Section 4.2 provides a literature review of the statistical methods used for prediction of MCI conversion to AD. Section 4.3 presents the formulation, estimation, and algorithm of the proposed STC model. Section 4.4 presents the application. Section 5 concludes the chapter.

4.2 Literature Review

One of the most prominent findings on AD is that AD patients have significant hippocampal atrophy that can be seen on an MRI scan. Because of this, abundant research has been devoted to using MRI imaging data for prediction of MCI conversion to AD. Risacher et al. (Risacher et al. 2009) analyzed MRI data using voxel-based morphometry and automated parcellation methods, and identified the degree of neurodegeneration in medial temporal structures as the best antecedent MRI marker of imminent conversion, with decreased hippocampal volume being the most robust. Zhang et al. (Zhang et al., 2012) applied a logistic regression model on MRI imaging, and found that combining a medical temporal lobe atrophy scale (MTAS) and a brain atrophy and lesion index (BALI) results in an improved predictive accuracy for MCI conversion. Wee et al. (Wee et al. 2013) proposed a novel approach to extract correlative morphological information from MRI, and demonstrated that combining this information with the conventional ROI-based information via multi-kernel support vector machines improves the prediction of MCI conversion.

Due to the complicated nature of MCI, it has been acknowledged that using MRI data alone may not suffice. As a result, abundant research has been done to integrate MRI with other data sources such as CSF measurement, cognitive test scores, and functional imaging like FDG-PET. Barnes et al., (Barnes et al. 2014) proposed a point-based risk score for prediction of MCI conversion, which combines MRI hippocampal subcortical volume and middle temporal cortical thinning together with the scores from several cognitive test instruments. Heister et al. (Heister et al. 2011) used a cox proportional hazard model to predict MCI conversion, which integrated medial temporal atrophy measured by MRI, CSF biomarker levels, and the degree of learning impairment measured by the Rey Auditory Verbal Learning Test. Jack et al. (Jack et al. 2010) proposed to integrate hippocampal volumes on MRI with CSF Aβ42 levels and Pittsburgh compound B PET measures in prediction of time-to-conversion from MCI to AD. Ye et al., (Ye et al. 2012) proposed a sparse learning model that integrated 15 features from MRI scans, cognitive measures, and APOE genotype.

Because multi-source data have been used in prediction of MCI conversion, there is a growing interest to evaluate which source carries the most weight. Toward this end, a number of comparative studies have been performed. Landau et al. (Landau et al, 2010) compared APOE ϵ4 allele frequency, CSF measurement, FDG-PET, hippocampal volume on MRI, and episodic memory performance at baseline. Their result showed that FDG-PET and episodic memory best predicted MCI conversion to AD. Cui et al. (Cui et al. 2011) compared MRI morphometry features, CSF measurement, and neuropsychological and functional measures (NMs). Their result showed that NMs outperformed CSF and MRI features. Yu et al., (Yu et al. 2012) compared MRI, FDG-PET, and CSF measurement, and found that MRI measures had the best predictive power. Overall, the existing comparative studies reached inconsistent conclusions regarding the relative importance of different data sources. The inconsistency might be caused by the difference in the subject pools included in each study and in the statistical methods used for the data analysis. Another possible reason may be the inherent heterogeneity of the MCI population. However, almost all the studies reached the same conclusion that integrating multi-source information yields a significantly better accuracy than using a single data source alone.

In additional to the aforementioned studies using baseline data, longitudinal data has also been used for MCI prediction. Zhang et al. (Zhang et al. 2012) developed a longitudinal feature selection method to jointly select brain regions across multiple time points and proposed a multi-kernel support vector machine for MCI prediction based on MRI, FDG-PET and cognitive scores. Misra et al. (Misra et al. 2009) investigated baseline and longitudinal patterns of brain atrophy in MCI patients, and found MCI converters displayed significantly lower volume in a number of white matter and grey matter regions.

Hinrichs et al. (Hinrichs et al. 2011) developed predictive markers for MCI conversion using a multi-kernel learning (MKL) framework.

4.3 Proposed method – a sequential tree-based classifier (STC)

4.3.1 Formulation of STC

Suppose there are p biomarkers $\mathbf{X}$ = $\{X_1, \ldots, X_p\}$, q patient characteristic variables/risk factors $\mathbf{Z}$ =$\{Z_1, \ldots, Z_q\}$, and a binary diagnostic outcome Y. For example, in diagnosing/predicting the conversion of MCI to AD, commonly used biomarkers include the p-tau level in CSF, mean cerebral metabolism on FDG-PET, and hippocampal volume on MRI, referred to as P-tau, FDG-PET, and MRI hereafter. Risk factors may include age, education level, and status of APOE e4 gene (Landau et al. 2012). $Y = 1$ if an MCI patient converts to AD within a clinical time of interest and $Y = 0$ otherwise. Our objective is to find a testing sequence for the biomarkers as well as a lower and an upper cutoff value for each biomarker adjusted for patient difference in terms of the risk factors, in order to classify patients into a HR, a LR, or an inconclusive category.

First, we focus on a less complicated problem in which the sequence of biomarkers is given. Without loss of generality, assume the sequence to be $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_p$. Also assume a positive correlation between each biomarker and the disease risk, i.e., a higher value of a biomarker means a higher risk of the disease. Although negative correlations exist for some biomarkers, we can always turn the correlations into positive by transforming the biomarkers. This assumption was made for simplicity of the subsequent discussion. We would like to sequentially find two cutoffs for each biomarker. That is, we would like to first find a lower and an upper cutoff for X_1, $l_1(\mathbf{Z})$ and $u_1(\mathbf{Z})$, that are functions of the risk factors $\mathbf{Z}$, such that a patient will have a HR of having the disease if $X_1 \geq u_1(\mathbf{Z})$, a LR if $X_1 \leq l_1(\mathbf{Z})$, and be inconclusive otherwise. HR and LR patients will need no more biomarker testing. Inconclusive patients will be further tested for the second biomarker X_2. Therefore, we will need to find a lower and an upper cutoffs for X_2, $l_2(\mathbf{Z})$ and $u_2(\mathbf{Z})$, such that an inconclusive patient from the previous biomarker testing will have a HR of having the disease if $X_2 \geq u_2(\mathbf{Z})$, a LR if $X_2 \leq l_2(\mathbf{Z})$, and continuously be inconclusive otherwise. The inconclusive patients at the current step will be further tested for X_3. This process will continue until all the biomarkers have been tested.

In a mathematically rigorous way, we can formulate the i-th step of the above process as follows: Let D_{i-1} be the cohort of inconclusive patients from the previous step. The goal of the i-th step is to find $l_i(\mathbf{Z})$ and $u_i(\mathbf{Z})$ for X_i that:

$$\min_{l_i(\mathbf{Z}),u_i(\mathbf{Z})} p(\, l_i(\mathbf{Z}) \leq X_i \ \leq u_i(\mathbf{Z})|D_{i-1}) \tag{4.1}$$

$$s.t. \quad p(Y = 1|X_i \ \geq u_i(\mathbf{Z}), \mathbf{Z}, D_{i-1}) \geq r_h,$$

$$p(Y = 1|X_i \ \leq l_i(\mathbf{Z}), \mathbf{Z}, D_{i-1}) \leq r_l\ . \tag{4.2}$$

The objective function is to minimize the proportion of inconclusive patients. This is important to patients by reducing the need and the associated cost and waiting time for another biomarker testing before a conclusive diagnosis can be made. It is also important to the clinic by reducing the overall cost including the labor and resource spent on the diagnosis. r_h and r_l are clinically-defined HR and LR thresholds, respectively, and are typically known to specific applications. For example, in diagnosis, r_h is typically 80-85% and r_l 10-20%. r_h is not necessarily equal to $1 - r_l$. Proposition 1 shows that the optimization problem in (4.1) is equivalent to two simpler sub-optimization problems.

Proposition 1: Let $\tilde{l}_i(\mathbf{Z})$ and $\tilde{u}_i(\mathbf{Z})$ denote the optimal solutions to (4.1). Let $l_i^*(\mathbf{Z})$ and $u_i^*(\mathbf{Z})$ be the optimal solutions to the optimization problems in (4.2) and (4.3), respectively.

$$l_i^*(\mathbf{Z}) = \begin{cases} \max\limits_{l_i(\mathbf{Z})} l_i(\mathbf{Z}) \\ s.t.\ \ p(Y = 1|X_i\ \le l_i(\mathbf{Z}), \mathbf{Z}, D_{i-1}) \le r_l \end{cases}. \tag{4.3}$$

$$u_i^*(\mathbf{Z}) = \begin{cases} \min\limits_{u_i(\mathbf{Z})} u_i(\mathbf{Z}) \\ s.t.\ \ p(Y = 1|X_i\ \ge u_i(\mathbf{Z}), \mathbf{Z}, D_{i-1}) \ge r_h \end{cases}. \tag{4.4}$$

Then, $l_i^*(\mathbf{Z}) = \tilde{l}_i(\mathbf{Z})$ and $u_i^*(\mathbf{Z}) = \tilde{u}_i(\mathbf{Z})$. (Proof skipped.)

Proposition 1 implies that $l_i^*(\mathbf{Z})$ can be obtained by first finding the feasible region of $l_i(\mathbf{Z})$, following which the $l_i^*(\mathbf{Z})$ can be naturally obtained by using the maximum value in that region. The same implication applies to $u_i^*(\mathbf{Z})$.

Furthermore, to facilitate identification of the feasible region for $u_i(\mathbf{Z})$, we apply Bayes' rule to the constraints in (4.4) and (4.3) and get

$$\frac{1-\varphi_{X_i|Y=1,\mathbf{Z}=\mathbf{z}}\left(u_i(\mathbf{Z})\right)}{1-\varphi_{X_i|Y=0,\mathbf{Z}=\mathbf{z}}\left(u_i(\mathbf{Z})\right)} \ge \frac{r_h}{1-r_h} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}, \text{ and} \tag{4.5}$$

$$\frac{\varphi_{X_i|Y=1,\mathbf{Z}=\mathbf{z}}\left(l_i(\mathbf{Z})\right)}{\varphi_{X_i|Y=0,\mathbf{Z}=\mathbf{z}}\left(l_i(\mathbf{Z})\right)} \le \frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}, \tag{4.6}$$

respectively. D_{i-1} was dropped for notation simplicity. $\varphi_{X_i|Y,\mathbf{Z}}(x)$ denotes the cumulative distribution function (CDF) of X_i given Y and $\mathbf{Z}$. $\pi(\mathbf{Z}) = p(Y = 1|\mathbf{Z})$ is the prior of Y before the biomarker X_i is tested. In (4.5) and (4.6), r_h and r_l are given constants. $\pi(\mathbf{Z})$ can be known from population statistics, i.e., the probability for people with a certain demographic profile (e.g., female, older than 65, and APOE e4 carrier) to have the disease. Therefore, the key to identifying the feasible regions of $l_i(\mathbf{Z})$ and $u_i(\mathbf{Z})$ is to know the distribution of $X_i|Y, \mathbf{Z}$. Because biomarkers are typically measured on a continuous scale, we assume a Gaussian distribution for $X_i|Y,\mathbf{Z}$. Note that even though the distribution of a biomarker may not be strictly Gaussian, we can apply Box-Cox transformation (Box et al. 1964) to make it approximately Gaussian. Under the Gaussian distribution, we can further link X_i and $Y,\mathbf{Z}$ by a linear model, i.e.,

$$X_i = \beta_{0,i} + \beta_{y,i}Y + \boldsymbol{\beta}_{z,i}^T\mathbf{Z} + \varepsilon_i, \tag{4.7}$$

where $\varepsilon_i \sim N(0,\sigma_i^2)$. Then, $\varphi_{X_i|Y,\mathbf{Z}}(x)$ becomes $\Phi\left(\frac{x-(\beta_{0,i}+\beta_{y,i}Y+\boldsymbol{\beta}_{z,i}^T\mathbf{Z})}{\sigma_i}\right)$, where $\Phi(\cdot)$ is the CDF of $N(0,1)$. Inserting this into (4.5) and (4.6) and further into the optimization problems in (4.2) and (4.3), we get:

$$l_i^*(\mathbf{Z}) = \begin{cases} \max\limits_{l_i(\mathbf{Z})} l_i(\mathbf{Z}) \\ s.t. \quad \dfrac{\Phi\left(\frac{l_i(\mathbf{Z})-\left(\beta_{0,i}+\beta_{y,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)}{\Phi\left(\frac{l_i(\mathbf{Z})-\left(\beta_{0,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)} \leq \frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \cdot \end{cases} \tag{4.8}$$

$$u_i^*(\mathbf{Z}) = \begin{cases} \min\limits_{u_i(\mathbf{Z})} u_i(\mathbf{Z}) \\ s.t. \quad \dfrac{1-\Phi\left(\frac{u_i(\mathbf{Z})-\left(\beta_{0,i}+\beta_{y,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)}{1-\Phi\left(\frac{u_i(\mathbf{Z})-\left(\beta_{0,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)} \geq \frac{r_h}{1-r_h} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \cdot \end{cases} \tag{4.9}$$

Next, we present an important property of the solutions to the optimization problems in (4.8) and (4.9) in Propositions 2 and 3, respectively. The proof for Proposition 2 is given in the Appendix. The proof for Proposition 3 is similar and therefore not provided.

Proposition 2: The solution to (4.8) exists and is unique. When $\frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \in (0,1)$, $l_i^*(\mathbf{Z})$ is the feasible solution at which the equality of the constraint is achieved. When $\frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \geq 1$, $l_i^*(\mathbf{Z}) = \infty$.

Proposition 3: The solution to (4.9) exists and is unique. When $\frac{r_h}{1-r_h} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} > 1$, $u_i^*(\mathbf{Z})$ is the feasible solution at which the equality of the constraint is achieved. When $\frac{r_h}{1-r_h} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \in (0,1]$, $u_i^*(\mathbf{Z}) = -\infty$.

4.3.2 Model Estimation for STC

Proposition 2 sheds some light on how to find the lower cutoff of the biomarker, $l_i^*(\mathbf{Z})$. Before the patient takes the biomarker testing, his/her $\frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}$ will be computed. If it is greater than or equal to one, the lower cutoff of the biomarker for this patient is infinity. This means that the patient can be considered LR regardless of the biomarker value. In other words, this patient does not need to be tested for the biomarker. Such situations rarely happen in practice, except for people with extremely high resistance to a certain disease, e.g., people carrying some genes that are disease-protective. In most cases, people coming to a clinic for diagnosis of a disease usually bear a fairly extensive amount of suspicion or risk for the disease. Therefore, we focus on the condition when $\frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \in (0,1)$. Then, the problem becomes finding the $l_i(\mathbf{Z})$ satisfying the equality of

$$\frac{\Phi\left(\frac{l_i(\mathbf{Z})-\left(\beta_{0,i}+\beta_{y,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)}{\Phi\left(\frac{l_i(\mathbf{Z})-\left(\beta_{0,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)} = \frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}. \tag{4.10}$$

Unfortunately, this problem does not have an analytical solution. To solve it, we may adopt one of two approaches: a numerical approach that finds the $l_i(\mathbf{Z})$ satisfying (4.10) for any given $\mathbf{Z}$. This approach can achieve any required precision for the solution, but is computationally intensive. An alternative approach is to use an approximation for $\Phi(x)$ proposed by (Bowling et al., 2009), i.e.,

$$\Phi(x) \approx \frac{1}{1+exp(-1.702x)}. \tag{4.11}$$

By substituting (4.11) into (4.10), $l_i^*(\mathbf{Z})$ can be solved analytically as

$$l_i^*(\mathbf{Z}) = -\frac{\sigma_i}{1.702} ln\left(\frac{1-\frac{r_h}{1-r_h}\times\frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}}{\frac{r_h}{1-r_h}\times\frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}exp\left(1.702\frac{\left(\beta_{0,i}+\beta_{y,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)-exp\left(1.702\frac{\left(\beta_{0,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)}\right). \tag{4.12}$$

Likewise, proposition 3 sheds some light on how to find the higher cutoff of the biomarker, i.e., $u_i^*(\mathbf{Z})$. Following similar reasoning and using the approximation in (4.11), $u_i^*(\mathbf{Z})$ can be solved analytically as

$$u_i^*(\mathbf{Z}) = \frac{\sigma_i}{1.702} ln\left(\frac{1-\frac{r_h}{1-r_h}\times\frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}}{\frac{r_h}{1-r_h}\times\frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}exp\left(1.702\frac{\left(\beta_{0,i}+\beta_{y,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)-exp\left(1.702\frac{\left(\beta_{0,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right)}{\sigma_i}\right)}\right). \tag{4.13}$$

Finally, we would like to point out that the $\beta_{0,i}$, $\beta_{y,i}$, $\boldsymbol{\beta}_{z,i}$, and σ_i in (4.12) and (4.13) are unknown but can be estimated from a training dataset. For example, under the linear model in (4.7), $\beta_{0,i}$, $\beta_{y,i}$, $\boldsymbol{\beta}_{z,i}$, and σ_i can be estimated by an maximum likelihood estimation (MLE). If **Z** is high-dimensional, variable selection techniques may be adopted to select a small subset of **Z** that have non-zero coefficients, such as the well-known lasso model (Wee et al. 2013), followed by an MLE on the non-zero coefficients. However, regardless of the estimation method, there is sampling uncertainty in the estimated $\beta_{0,i}$, $\beta_{y,i}$, $\boldsymbol{\beta}_{z,i}$, and σ_i due to the finite sample size of the training dataset, which will further introduce uncertainty into $u_i^*(\mathbf{Z})$ and $l_i^*(\mathbf{Z})$. To better account for the sampling uncertainty, we use Monte Carlo simulation to generate an empirical sampling distribution for $\hat{u}_i^*(\mathbf{Z})$

and $\hat{l}_i^*(\mathbf{Z})$, respectively, and then use the empirical means as the solutions to (4.13) and (4.12). This approach is found to be more robust to sampling uncertainty and have better accuracy in our case studies. Specifically, the empirical sampling distribution for $\hat{u}_i^*(\mathbf{Z})$ is generated as follows (a similar procedure can be used for $\hat{l}_i^*(\mathbf{Z})$): Let $\tilde{\beta}_{0,i}$, $\tilde{\beta}_{y,i}$, $\tilde{\boldsymbol{\beta}}_{z,i}$, and $\tilde{\sigma}_i$ be the estimated model parameters from the training dataset through MLE. We use Monte Carlo simulation to generate N samples from the following empirical distributions:

$$\begin{pmatrix} \hat{\beta}_{0,i}^{(t)} \\ \hat{\beta}_{y,i}^{(t)} \\ \hat{\boldsymbol{\beta}}_{z,i}^{(t)} \end{pmatrix} \sim N\left(\begin{pmatrix} \tilde{\beta}_{0,i} \\ \tilde{\beta}_{y,i} \\ \tilde{\boldsymbol{\beta}}_{z,i} \end{pmatrix}, \tilde{\sigma}_i^2 \left((\mathbf{1} \quad \mathbf{y} \quad \mathbf{z})^T (\mathbf{1} \quad \mathbf{y} \quad \mathbf{z}) \right)^{-1} (\mathbf{1} \quad \mathbf{y} \quad \mathbf{z})^T \mathbf{x}_i \right), \tag{4.14}$$

$$\hat{\sigma}_i^{2(t)} \sim \frac{\left(\mathbf{x}_i - (\tilde{\beta}_{0,i} + \tilde{\beta}_{y,i}\mathbf{y} + \tilde{\boldsymbol{\beta}}_{z,i}^T \mathbf{z})\right)^T \left(\mathbf{x}_i - (\tilde{\beta}_{0,i} + \tilde{\beta}_{y,i}\mathbf{y} + \tilde{\boldsymbol{\beta}}_{z,i}^T \mathbf{z})\right)}{\chi_{n-p}^2}, \tag{4.15}$$

$t = 1, \ldots, N$. $\mathbf{x}_i$, $\mathbf{y}$, and $\mathbf{z}$ are training data for n patients. $\mathbf{1}$ is a $n \times 1$ vector of ones. p is the column dimension of the predictor matrix $(\mathbf{1} \quad \mathbf{y} \quad \mathbf{z})$. Then, each sample generated from (4.14) and (4.15), i.e., $\hat{\beta}_{0,i}^{(t)}$, $\hat{\beta}_{y,i}^{(t)}$, $\hat{\boldsymbol{\beta}}_{z,i}^{(t)}$, and $\hat{\sigma}_i^{2(t)}$, is inserted into (4.13) to obtain $\hat{u}_i^*(\mathbf{Z})^{(t)}$. The average, $\bar{\hat{u}}_i^*(\mathbf{Z}) = \frac{\sum_{t=1}^N \hat{u}_i^*(\mathbf{Z})^{(t)}}{N}$, is used as the final solution to (4.13).

4.3.3 Algorithm for STC

Section 3.1 and 3.2 assumed that the biomarker sequence is known and the discussion was focused on the i-th step (i.e., the i-th biomarker) of the modeling building process of the STC. In this section, we present the full algorithm. The input to the algorithm includes a specification on the biomarkers that are allowed to be used in a clinic. This may be clinic-specific depending on availability and resource constraints. The input also includes a training and a validation set on the biomarkers $\mathbf{X}$, patient characteristic variables/risk factors $\mathbf{Z}$, and the diagnostic outcome Y, the HR and LR thresholds, r_h and r_l, and the prior, $\pi(\mathbf{Z})$. Suppose p biomarkers are available. Then, the objective or output of the algorithm is to find an optimal sequence of the biomarkers with cutoffs for each biomarker, $u_i^*(\mathbf{Z})$ and $l_i^*(\mathbf{Z})$, $i = 1, \dots, p$. Since the number of biomarkers for a particular disease is usually small, we will perform an exhaustive search over all possible sequences. We will report three metrics computed on the validation set for comparing the sequences: positive prediction value (PPV), negative prediction value (NPV), and the percentage of patients classified as inconclusive. The first two metrics reflect the accuracy, where PPV measures the proportion of patients classified as HR that are true converters and NPV measures the proportion of patients classified as LR that are true non-converters. The last metric reflects the efficiency: the lower the inconclusive percentage, the more efficient the biomarker sequence.

Specifically, given that p biomarkers are available in a clinic, our algorithm performs three major steps for each of $p!$ possible biomarker sequences. Without loss of generality, denote each sequence by $X_1 \to X_2 \to \cdots \to X_p$.

Step 1 (initialization): Initialize the algorithm by having $i \leftarrow 1$ and putting the entire training set into D_{i-1}.

Step 2 (sequential estimation)

Sub-step 2.1 (identification of the cutoffs for $\boldsymbol{X_i}$): Fit a linear model as (4.7) for X_i using the training data in D_{i-1}, and obtain estimates for the model coefficients, $\tilde{\beta}_{0,i}$, $\tilde{\beta}_{y,i}$, $\tilde{\boldsymbol{\beta}}_{z,i}$, and $\tilde{\sigma}_i$. Check the normality assumption of the model and apply box-cox transformation to X_i if needed. Use the estimated model coefficients to obtain

N Monte Carlo samples $\hat{\beta}_{0,i}^{(t)}, \hat{\beta}_{y,i}^{(t)}, \widehat{\boldsymbol{\beta}}_{z,i}^{(t)}$, and $\hat{\sigma}_i^{2(t)}$, $t = 1, \dots, N$. Insert each sample into (4.13) and (4.12) and obtain sample realizations for the cutoffs, i.e., $\hat{u}_i^*(\mathbf{Z})^{(t)}$ and $\hat{l}_i^*(\mathbf{Z})^{(t)}$, $t = 1, \dots, N$. Use the sample averages, $\bar{\hat{u}}_i^*(\mathbf{Z})$ and $\bar{\hat{l}}_i^*(\mathbf{Z})$, as the estimated cutoffs for X_i.

Sub-step 2.2 (subsetting of the training set): Apply the estimated cutoffs in sub-step 2.1 to the patients in D_{i-1} and only keep patients with $\bar{\hat{l}}_i^*(\mathbf{Z}) < X_i < \bar{\hat{u}}_i^*(\mathbf{Z})$ in the training set. Denote the current training set by D_i.

Sub-step 2.3 (continuation or stopping): Move onto the next biomarker by having $i \leftarrow i + 1$ and going to sub-step 2.1, until $i + 1 = p$.

Step 3 (evaluation): Apply the estimated cutoffs for each biomarker, i.e., $\bar{\hat{u}}_i^*(\mathbf{Z})$ and $\bar{\hat{l}}_i^*(\mathbf{Z})$, $i = 1, \dots, p$, to the validation set and compute PPV, NPV, and the percentage of patients classified as inconclusive.

This three-step algorithm will be applied to each of the $p!$ possible biomarker sequences. These sequences will then be compared in terms of the diagnostic accuracy (PPV and NPV) and efficiency (percentage of inconclusive patients) evaluated on the validation set. Because multiple metrics are used in the comparison, an integrated metric may be used to help select the optimal sequence. Alternatively, a Pareto optimal frontier may be provided to practitioners to show the tradeoffs between multiple Pareto optimal solutions/sequences.

4.3.4 Extension to non-Gaussian biomarkers

When the biomarkers do not follow Gaussian distributions, one approach is to apply Box-Cox transformation to make them approximately Gaussian, which was mentioned in

Section 3.1. An alternative approach is to deal with the non-Gaussian distributions directly. Specifically, instead of linking the biomarker X_i with $Y, \mathbf{Z}$ by a linear model as in (4.7), we can use a Generalized Linear Model (GLM), i.e.,

$$E(X_i) = g^{-1}\left(\beta_{0,i} + \beta_{y,i}Y + \boldsymbol{\beta}_{z,i}^T\mathbf{Z}\right), \tag{4.16}$$

where $g(\cdot)$ is an appropriate link function depending on the distribution of the biomarker. Consequently, (4.8) and (4.9) change to

$$l_i^*(\mathbf{Z}) = \begin{cases} \max_{l_i(\mathbf{Z})} l_i(\mathbf{Z}) \\ s.t. \quad \frac{\varphi_{X_i|Y=1,\mathbf{Z}=\mathbf{z}}(l_i(\mathbf{Z}))}{\varphi_{X_i|Y=0,\mathbf{Z}=\mathbf{z}}(l_i(\mathbf{Z}))} \leq \frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \end{cases}, \tag{4.17}$$

$$u_i^*(\mathbf{Z}) = \begin{cases} \min_{u_i(\mathbf{Z})} u_i(\mathbf{Z}) \\ s.t. \quad \frac{1-\varphi_{X_i|Y=1,\mathbf{Z}=\mathbf{z}}(u_i(\mathbf{Z}))}{1-\varphi_{X_i|Y=0,\mathbf{Z}=\mathbf{z}}(u_i(\mathbf{Z}))} \geq \frac{r_h}{1-r_h} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})} \end{cases}. \tag{4.18}$$

$\varphi_{X_i|Y,\mathbf{Z}}(x)$ is the CDF of X_i given Y and $\mathbf{Z}$, which can be specificed according to the GLM in (4.16). $\varphi_{X_i|Y,\mathbf{Z}}(x)$ is not Gaussian, so the approximation in (4.11) cannot be used. Consequently, (4.17) and (4.18) cannot be solved analytically but by a numerical search, which is computationally more intensive. The modified STC algorithm is the following:

Step 1 (initialization): Initialize the algorithm by having $i \leftarrow 1$ and putting the entire training set into D_{i-1}.

Step 2 (sequential estimation)

Sub-step 2.1 (identification of the cutoffs for $g\left(E(\boldsymbol{X}_i)\right)$): Fit a GLM as (4.16) using the training data in D_{i-1}, and obtain estimates for the model coefficients, $\tilde{\beta}_{0,i}$, $\tilde{\beta}_{y,i}$, and $\tilde{\boldsymbol{\beta}}_{z,i}$. In order to solve the optimization problems in (4.17), we can start from a small $l_i(\mathbf{Z})$ for which the constraint holds, and increase $l_i(\mathbf{Z})$ in small steps until the constraint is violated. The last value of $l_i(\mathbf{Z})$ before the constrain is violated is the optimal solution $l_i^*(\mathbf{Z})$. Likewise, we can obtain the optimal solution $u_i^*(\mathbf{Z})$ in (4.18).

Sub-step 2.2 (subsetting of the training set): Apply the estimated cutoffs in sub-step 2.1 to the patients in D_{i-1} and only keep patients with $l_i^*(\mathbf{Z}) < g\big(E(\boldsymbol{X}_i)\big) < u_i^*(\mathbf{Z})$ in the training set. Denote the current training set by D_i.

Sub-step 2.3 (continuation or stopping): Move onto the next biomarker by having $i \leftarrow i + 1$ and going to sub-step 2.1, until $i + 1 = p$.

Step 3 (evaluation): Apply the estimated cutoffs for each biomarker, i.e., $u_i^*(\mathbf{Z})$ and $l_i^*(\mathbf{Z})$, $i = 1, \ldots, p$, to the validation set and compute PPV, NPV, and the percentage of patients classified as inconclusive.

4.4 Case studies in prediction of MCI conversion to AD

In this section, we present two clinical applications using the proposed STC: Sub-section 4.1 presents an application in clinical diagnosis, i.e., prediction/classification of MCI patients into HR, LR, or inconclusive categories so that appropriate medical decisions can be made for each patient. Sub-section 4.1 presents another application of using STC to help patient selection in clinical trials. As mentioned in Introduction, there has been a growing consensus in the medical society that treatment of AD should target on its early phases before irreversible brain damage occurs. MCI is such an early phase and therefore has been targeted by drug companies to develop treatment for slowing down or even stopping the progression to AD. However, it is well-known that MCI patients are heterogeneous and not all of them will eventually develop AD. To be able to appropriately assess the efficacy of an AD-defeating drug, it is important to identify a sub-cohort of MCI patients with a HR of converting to AD and enter these patients into the drug trial. This important task is known as patient selection in clinical trials and can be accomplished with the help of STC.

The data used in this section was obtained from the ADNI database (http:// adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W.Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for at least 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information, see http://www.adni-info.org/.

Specifically, our study includes 187 MCI patients included in the ADNI database who have complete data on three biomarkers, P-tau, FDG-PET, and Hippo, at their baseline visits, patient-specific variables/risk factors such as age, gender, education level, APOE e4 status, and cognitive test scores, as well as conversion vs. non-conversion to AD at the end of their clinical follow-up time periods. A detailed description of the data is shown in Table 11.

Table 11: Description of the data

Variable	**Non-Converters**	**Converters**
Sample size	87	100
Gender: female %	59.8	58
Age: ave. (std.)	73.1 (7.3)	73.6 (7.6)
Education years: ave. (std.)	16.6 (2.6)	16.3 (2.7)
APOE e4 status: carriers %	41.4	68
Mini-mental State Examination score: ave. (std.)	27.9 (1.6)	26.7 (1.7)
P-tau, pg/mL: ave. (std.)	35.8 (22.8)	50.6 (25.1)
FDG-PET, relative counts: ave. (std.)	1.25 (0.13)	1.16 (0.11)
Hippo, mm^3: ave. (std.)	3449 (551)	3067 (497)

Standardized biomarker acquisition and performance methods of ADNI are described at www.loni.ucla.edu/ADNI. Protocols of image and CSF analyses are reported in detail elsewhere (Jack et al. 2010, Jagust et al. 2009, Kim et al. 2011, Landau et al. 2010). In brief, the mean FDG count per subject (i.e., biomarker “FDG-PET”) was extracted from a composite region of interest on the basis of the AD-typical hypometabolic pattern (Jack et al. 2008, Kim et al. 2011). Hippocampal volumes (i.e., biomarker “Hippo”) were extracted from structural MRI scans (1.5 T) using the FreeSurfer software http://surfer.nmr.mgh.harvard.edu (Kim et al. 2011). Peptide concentrations (i.e., biomarker “P-tau”) were measured in CSF using aliquots obtained from the same vial at the same thaw (Jagust et al. 2010).

4.4.1 Clinical diagnosis of MCI conversion to AD

4.4.1.1 Diagnosis of MCI conversion to AD with three biomarkers

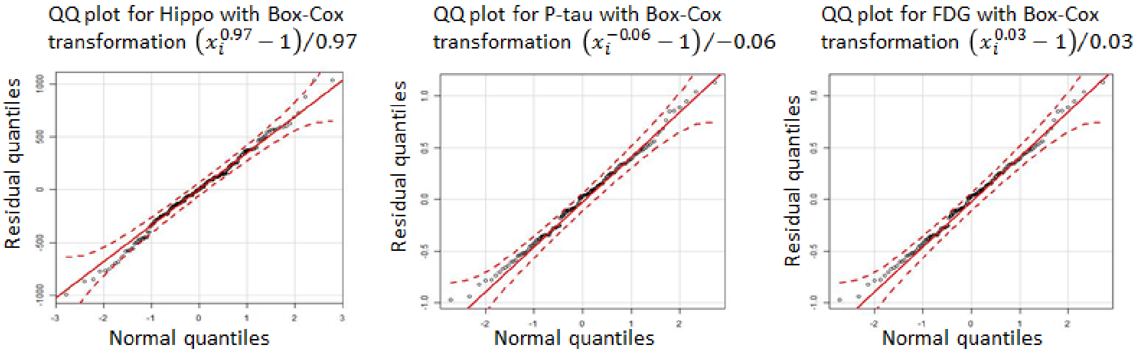


Figure 8: QQ plots for biomarkers after Box-Cox transformation in the sequence “P-tau->FDG-PET->Hippo”

We first focus on a scenario that all three biomarkers, P-tau, FDG-PET, and Hippo, are available in the clinic. Then, the goal is to find an optimal sequence of the biomarkers with cutoffs for each biomarker, i.e., $u_i^*(\mathbf{Z})$ and $l_i^*(\mathbf{Z})$, $i = 1,2,3$. Among known risk factors such as age, gender, education level, and APOE e4 status, only age is found to be significant in this dataset. Therefore, $\mathbf{Z}$ includes age. The HR and LR thresholds are set to be $r_h = 0.8$ and $r_l = 0.2$, which are common choices in clinical diagnosis. Also, a uniform prior is adopted, i.e., $\pi(\mathbf{Z}) = 0.5$. Three biomarkers compose $3! = 6$ possible sequences. For each sequence, we apply the algorithm in Section 3.3 with a minor modification of using cross validation (CV) instead of arbitrarily splitting the entire dataset into a training and a validation set. The CV-based PPV, NPV, and percentage of inconclusive patients for each sequence are summarized in Table 12. Box-Cox transformation on the biomarkers is used and the transformed biomarkers in each sequence follow Gaussian distributions. For example, Figure 8 shows the QQ plot of each transformed biomarker in the sequence “P-tau->FDG-PET->Hippo”, which demonstrates clear normality.

Table 12: CV-based PPV, NPV, and percentage of inconclusive patients for all possible sequences of three biomarkers using STC

Sequence of biomarkers	**PPV**	**NPV**	**% inconclusive patients**
P-tau->FDG-PET->Hippo	74%	81%	59%
FDG-PET->P-tau->Hippo	70%	74%	52%
P-tau->Hippo->FDG-PET	71%	78%	56%
Hippo->P-tau->FDG-PET	73%	78%	52%
FDG-PET->Hippo->P-tau	72%	77%	58%
Hippo->FDG-PET->P-tau	72%	77%	63%

A clear trend of the results in Table 12 is that the NPVs are higher than PPVs regardless of the sequence of biomarkers. This suggests that the three biomarkers have a better capability for identifying non-converters than converters. Another observation is that the PPVs are lower than the HR threshold $r_h = 0.8$. This is reasonable because $r_h = 0.8$ is set for model training and the PPVs are computed based on CV which reflect the accuracy of the trained model applied to unseen data. The fact that the PPVs are only slightly lower than 0.8 implies that STC has good generalization capability. Likewise, the NPVs are only slightly lower than or almost equal to $1 - r_l = 0.8$, which also indicates good generalization capability of STC. Last but not least, we observe that over half of the MCI patients in the dataset are found to be inconclusive no matter which sequence of the biomarkers is used. This is expected because this study only uses *baseline* biomarker measurements to predict the conversion of MCI to AD. Use of baseline biomarkers for the prediction has clear clinical benefits as it enables early decision making for patients classified as HR and LR converters. On the other hand, it is highly likely that a conclusive classification is not possible for some MCI patients using baseline biomarker measurements alone. These patients need to be followed up and kept tracked of for the changes in their biomarker measurements over time before a conclusive prediction can be reached.

To select an optimal biomarker sequence among the six possible ones in Table 12, we need to make a tradeoff between accuracy (measured by PPV and NPV) and efficiency (measured by the percentage of inconclusive patients) because no sequence optimizes the two criteria simultaneously. If accuracy is the primary consideration, the sequence “P-tau->FDG-PET->Hippo” should be selected because it has the highest PPV (74%) and NPV (81%). This sequence, on the other hand, classifies 59% of MCI patients as inconclusive, which makes it the second least efficient sequence among the six. If efficiency is the primary consideration, the sequence “Hippo->P-tau->FDG-PET” should be selected as it has the lowest percentage of inconclusive patients (52%), although its accuracy is sub-optimal.

A commonly used approach in optimization when multiple criteria need to be optimized is to examine the Pareto optimal frontier. Figure 9 shows the Pareto optimal frontier for the six biomarker sequences. The vertical axis “efficiency” is defined as $1 -$ percentage of inconclusive patients or the percentage of patients classified as HR or LR

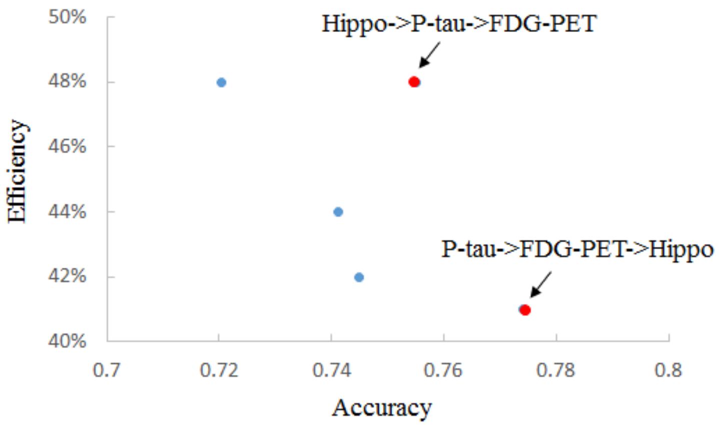
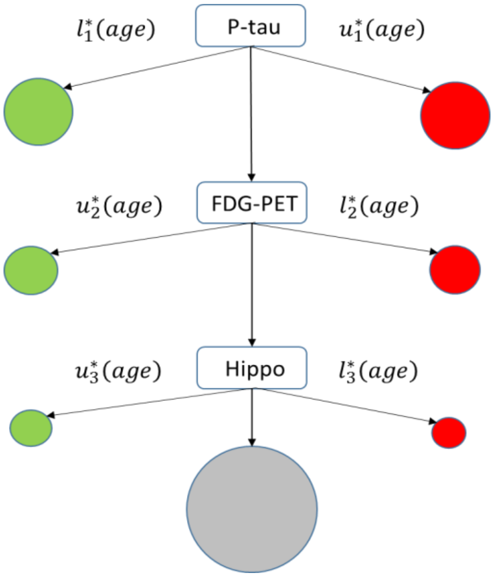


Figure 9: Efficiency vs. accuracy of six possible sequences given all three biomarkers. Sequences in red are on the Pareto optimal frontier.

by STC. The horizontal axis “accuracy” is defined as a weighted average of PPV and NPV, where the weights are proportions of samples classified as HR and LR, respectively. Each sequence is represented by a dot. Two dots in red are sequences on the Pareto optimal frontier. In particular, the sequence “P-tau->FDG-PET->Hippo” optimizes the accuracy criterion while “Hippo->P-tau->FDG-PET” optimizes the efficiency.



(a)

Cutoffs as functions of age	Cutoffs at median age
$l_1^*(age) = exp\left(\frac{log\left(-0.01 \times ln\left(\frac{1}{3}\times e^{(16.83-0.04\times age)} - \frac{4}{3}\times e^{(15.43-0.04\times age)}\right)+1\right)}{-0.04}\right)$	19.4
$u_1^*(age) = exp\left(\frac{log\left(0.01 \times ln\left(-\frac{4}{3}\times e^{(16.83-0.04\times age)} + \frac{1}{3}\times e^{(15.43-0.04\times age)}\right)+1\right)}{-0.04}\right)$	68.4
$l_2^*(age) = exp\left(\frac{log\left(0.01 \times ln\left(-\frac{4}{3}\times e^{(-3.87+0.02\times age)} + \frac{1}{3}\times e^{(-5.24+0.02\times age)}\right)+1\right)}{0.15}\right)$	1.054
${}_2^*(age) = exp\left(\frac{log\left(-0.01 \times ln\left(\frac{1}{3}\times e^{(-3.87+0.02\times age)} - \frac{4}{3}\times e^{(-5.24+0.02\times age)}\right)+1\right)}{0.15}\right)$	1.360
$l_3^*(age) = exp\left(\frac{log\left(80.17 \times ln\left(-\frac{4}{3}\times e^{(-20.42+0.11\times age)} + \frac{1}{3}\times e^{(-21.70+0.11\times age)}\right)+1\right)}{0.86}\right.$	2471.8
$u_3^*(age) = exp\left(\frac{log\left(-80.17 \times ln\left(\frac{1}{3}\times e^{(-20.42+0.11\times age)} - \frac{4}{3}\times e^{(-21.70+0.11\times age)}\right)+1\right)}{0.86}\right.$	3976.2

(b)

Figure 10: Cutoffs found by STC for biomarker sequence “P-tau->FDG-PET->Hippo” represented by (a) a tree-like plot in which green/red/grey circles represent LR/HR/inconclusive categories and sizes of the circles are in proportion to the sample size of each branch. (b) Cutoffs of each biomarker as functions of “age”

Next, we would like to show the cutoffs of each biomarker found by STC. We choose to show these for the sequence “P-tau->FDG-PET->Hippo” as an example using a tree-like plot in Figure 10. Specifically, In Figure 10(a), branches in green/red represent HR/LR converters classified by STC. The branch in grey represents the inconclusive category. Sizes of the branches/circles are in proportion to the sample sizes of the branches.

A clear observation is that less samples are classified as HR and LR as the tree goes deeper. This is a result from the sequential nature of STC, i.e., a later biomarker needs to classify samples that are failed to be classified (i.e., the inconclusive samples) by a previous biomarker so it has a "tougher" mission to accomplish. Figure 10(b) shows the cutoffs as functions of "age" using the approximations in (4.13) and (4.15). Values of the cutoffs at the median age of the dataset are also provided for illustration purposes.

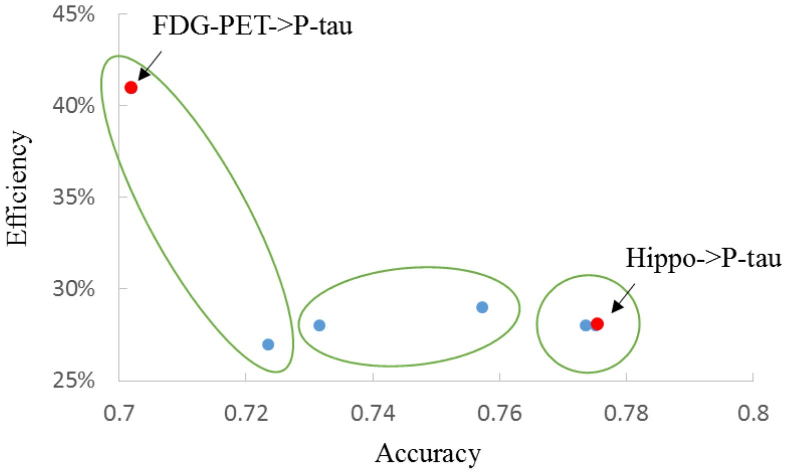


Figure 11: Efficiency vs. accuracy of six possible sequences given two out of three biomarkers. Sequences in red are on the Pareto optimal frontier. Each ellipse highlights two sequences with the same pair of biomarkers but in different orders.

Moreover, we would like to point out that the findings from STC can help not only clinical diagnosis but also knowledge discovery such as discovering disease sub-types. Using the tree in Figure 10 as an example, there seem to exist three distinct sub-types of HR converters, i.e., the sub-types of P-tau-abnormality (P-tau $\geq u_1^*(age)$), FDG-PET-abnormality (P-tau $< u_1^*(age)$ & FDG-PET $\leq l_2^*(age)$), and Hippo-abnormality (P-tau $< u_1^*(age)$ & FDG-PET $> l_2^*(age)$ & Hippo $\leq l_3^*(age)$). Indeed, there has been medical evidence that the three biomarkers track distinct aspects of the AD pathophysiological process (Jack et al. 2010). That is, FDG-PET, as a measure for AD-related glucose hypometabolism, reflects reduction in synaptic density/activity and phenomena of diaschisis, Hippos, as a measure for hippocampal atrophy, reflects neural loss, while P-tau reflects intracellular hyperphosphorylation of tau. STC allows for a finer distinction of HR converters into different sub-types according to specific biomarker abnormality, which may lead to more targeted and effective treatment. Likewise, STC can help discover sub-types of LR converters. This would facilitate the study of different pathophysiological mechanisms that lead to disease protection or resistance.

4.4.1.2 Diagnosis of MCI conversion to AD with two biomarkers – a limited-resource scenario

Next, we present the results of STC in a "limited-resource" scenario, e.g., when only two out of the three biomarkers are available. This situation is common in many clinics. We use the same setting as the previously-presented three-biomarker scenario, i.e., $r_h = 0.8$, $r_l = 0.2$, $\pi(\mathbf{Z}) = 0.5$, and $\mathbf{Z} = \{age\}$. Two biomarkers compose six possible sequences. For each sequence, we apply the algorithm in Section 3.3 and compute the CV-based PPV, NPV, and percentage of inconclusive patients. Figure 10 shows the Pareto optimal frontier for the six sequences, in which efficiency and accuracy are defined in the same way as Figure 9. The sequence "Hippo->P-tau" optimizes the accuracy criterion while "FDG-PET->P-tau" optimizes the efficiency. Furthermore, each ellipse includes two sequences with the same pair of biomarkers but in different orders. If a clinic only has the resource for testing two specific biomarkers, we can compare the two dots/sequences within the same ellipse and select an order of the two biomarkers that is more appropriate in terms of efficiency or/and accuracy. For example, if a clinic only has FDG-PET and Hippos, we can compare the two dots within the middle ellipse. The dot at the upper-right corner corresponds to the sequence “Hippo->FDG-PET” and is clearly better because it has better efficiency and accuracy.

4.4.1.3 Comparison between STC and decision tree

Finally, we compare STC with the conventional decision tree. Specifically, we apply the C4.5 algorithm in the Weka software (Hall et al. 2009) to the same dataset as that used by STC. Because STC uses age in addition to three biomarkers, we include the same variables in C4.5 for a fair comparison. Parameters of C4.5 are tuned to optimize the CV accuracy. Figure 11 shows the decision tree generated by C4.5. Compared with the tree generated by STC in Figure 10, we can obtain the following observations: Both methods find P-tau as the first biomarker to be used for the classification. This suggests that P-tau may be more informative than the other two biomarkers. The differences between the two methods are summarized as follows: 1) The CV-based PPV and NPV of the decision tree are 68% and 69%, respectively, which are significantly lower than the PPV (74%) and NPV (81%) of the optimal sequence found by STC. This is because the decision tree, by design, must assign a class membership to every sample, even when a sample does not have a significantly higher probability of belonging to one class than the other. This leads to potentially large classification errors. In contrast, STC only classifies samples with a HR or LR of conversion while putting samples with only a mild risk in either direction (i.e., a risk between LR and HR) in an inconclusive category. From a disease management point of view, STC is more appropriate by allowing HR patients to receive immediate medical attention, LR patients to be put on long-term observation, and patients in between to be followed up to track the changes in their disease risks. 2) According to the decision tree in Figure 11, no patients can be classified using a single biomarker. In contrast, according to the tree in Figure 10 produced by STC, 52.3% of the patients classified as HR and LR only need to be tested by P-tau. In this sense, STC means less diagnostic costs, less patient waiting time, and more timely medical decision making. 3) The decision tree in Figure 11 is somewhat counter-intuitive. Biomarkers are expected to have a monotonic relationship with the risk of disease. For example, a higher P-tau, lower FDG-PET, or lower Hippo indicates a higher risk of AD pathology. However, there are several branches in Figure 11 whose biomarker ranges are contrary to this expectation. For instance, the top-right green circle represents non-converters whose classification rule is P-tau> 28.5 and FDG-PET> 1.19. This higher value range for P-tau is expected to indicate a higher risk of AD pathology. From a clinical utilization's point of view, clinicians would be very reluctant to adopt such a model as the decision tree in Figure 11 regardless of the accuracy, because the model is against their medical knowledge and thus being difficult to understand and trust. In essence, decision tree is a pure data-driven model that does not integrate medical knowledge and biological principles into its model building process. In contrast, STC, by its unique design, honors the monotonic relationship between a biomarker and disease risk, and therefore is able to provide a model with good interpretability and clinical utility.

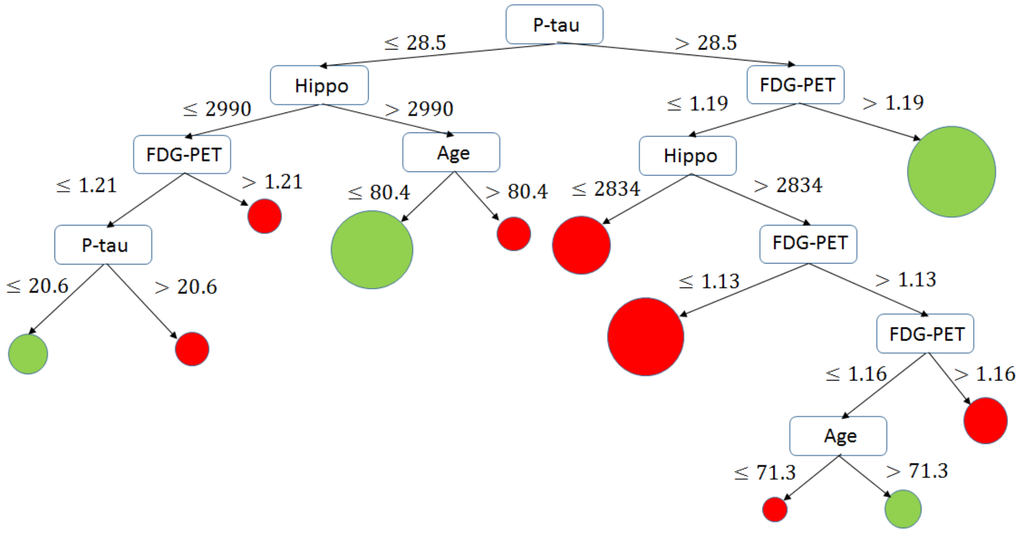


Figure 12: Decision tree generated by C4.5. Green/red circles represent non-convert/converter categories and sizes of the circles are in proportion to the sample size of each branch.

4.4.2 Selection of HR converters for clinical trials

Here, our objective is to identify a sub-cohort of MCI patients with a HR of converting to AD and enter these patients into a drug trial. This objective is different from clinical diagnosis as presented in sub-section 4.1 in the sense that we only care about maximizing PPV, as opposed to accuracy that includes both PPV and NPV, and maximizing the number/proportion of patients classified as HR, as opposed to efficiency that includes patients classified as HR or LR. To serve this objective, we modify the STC algorithm by treating r_l as a tuning parameter ranging from 0.05 to 0.5 in increment of 0.05. We adopt the same setting as that in sub-section 4.1, i.e., $r_h = 0.8$, $\pi(\mathbf{Z}) = 0.5$, and $\mathbf{Z} = \{ge\}$. Figure 12 shows the Pareto optimal frontier for the biomarker sequences, in which each dot represents a sequence at a specific r_h (a total of 6 sequences × 10 r_h values = 60 dots). On the frontier, the sequence "Hippo->P-tau->FDG-PET" at $r_l = 0.35$ is probably the one achieving the best tradeoff between the CV-based PPV (87%) and number of HR patients (30), and therefore recommended as the biomarker testing sequence used for HR converter patient selection in AD-related clinical trials. Finally, Figure 13 shows the cutoffs of each biomarker for the sequence “Hippo->P-tau->FDG-PET” at $r_l = 0.35$ found by STC.

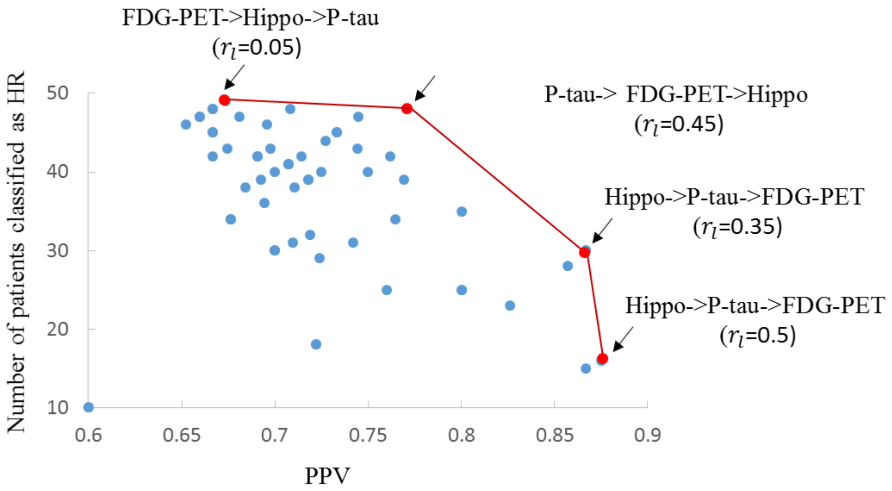
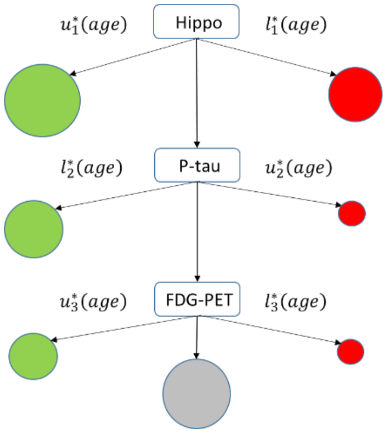


Figure 13: Number of HR patients vs. PPV of six possible sequences given all three biomarkers at r_l ranging from 0.05 to 0.5 in increment of 0.05. Sequences in red are on the Pareto optimal frontier.



(a)

Cutoffs as functions of age	**Cutoffs at median age**
$l_1^*(age)$ $= exp\left(\dfrac{log\left(209.97 \times ln\left(-\frac{4}{3} \times e^{(-19.80+0.11\times age)} + \frac{1}{3} \times e^{(-21.12+0.11\times age)}\right) + 1\right.}{0.97}\right.$	2613.8
$u_1^*(age)$ $= exp\left(\dfrac{log\left(-209.97 \times ln\left(\frac{7}{5} \times e^{(-19.80+0.11\times age)} - \frac{13}{5} \times e^{(-21.12+0.11\times age)}\right) +\right.}{0.97}\right.$	3470.1
$l_2^*(age)$ $= exp\left(\dfrac{log\left(-0.01 \times ln\left(\frac{7}{5} \times e^{(16.28-0.03\times age)} - \frac{13}{5} \times e^{(15.13-0.03\times age)}\right) + 1\right)}{-0.04}\right)$	29.7
$u_2^*(age)$ $= exp\left(\dfrac{log\left(0.01 \times ln\left(-\frac{4}{3} \times e^{(16.28-0.03\times age)} + \frac{1}{3} \times e^{(15.13-0.03\times age)}\right) + 1\right)}{-0.04}\right)$	91.5
$l_3^*(age)$ $= exp\left(\dfrac{log\left(0.01 \times ln\left(-\frac{4}{3} \times e^{(-0.85-0.03\times age)} + \frac{1}{3} \times e^{(-1.86-0.03\times age)}\right) + 1\right)}{0.19}\right)$	0.992
${}_3^*(age)$ $= exp\left(\dfrac{log\left(-0.01 \times ln\left(\frac{7}{5} \times e^{(-0.85-0.03\times age)} - \frac{13}{5} \times e^{(-1.86-0.03\times age)}\right) + 1\right)}{0.19}\right)$	1.261

(b)

Figure 14: Cutoffs found by STC for biomarker sequence “Hippo->P-tau->FDG-PET” represented by (a) a tree-like plot in which green/red/grey circles represent LR/HR/inconclusive categories and sizes of the circles are in proportion to the sample size of each branch. (b) Cutoffs of each biomarker as functions of “age”

4.5 Conclusion

In this paper, we developed a STC for predicting the conversion of MCI to AD. The uniqueness of the STC is to find an optimal testing sequence of the biomarkers and two-sided cutoffs of each biomarker that satisfy pre-specified accuracy requirements while minimizing the proportion of inconclusive diagnosis. The cutoffs can be customized for each individual patient by taking into account patient demographic and genetic variables that are potential risk factors for AD. We formulated STC into an optimization problem and performed theoretical analysis to prove the existence and uniqueness of the solution to STC. Then, we proposed two approaches for estimating the cutoffs of the biomarkers, including a numerical approach and an approximate-analytical approach, with consideration of sampling uncertainty. Next, we presented the full algorithm integrating the estimation approaches for the cutoffs with a search of the optimal sequence. Finally, we presented two applications of STC using the ADNI data. In the first application, we used STC to identify an optimal sequence of three and two biomarkers (as an example of a resource-limited situation) and the associated cutoffs for classifying MCI patients into HR converters, LR converters, or the inconclusive category. The CV-based PPV and NPV of the optimal sequence are close to the pre-specified HR and LR thresholds that reflected the expected accuracy. STC also allowed multiple criteria, e.g., accuracy and efficiency, to be optimized using a Pareto optimal frontier. The results also helped identify subtypes within HR converters. Compared with the conventional decision tree classifier, STC achieved higher PPV and NPV, saved biomarker testing costs and patient waiting time, facilitated timely medical decision making, and produced a model that is consistent with medical knowledge and biological principles and thus being clinically more trust-worthy. In the other application, we used STC to identify a sub-cohort of MCI patients with a HR of converting to AD. With a slight modification of the STC algorithm, we were able to identify such a sub-cohort with a high CV-based PPV (87%) and a reasonable size appropriate for clinical trials.

Finally, we would like to point out that STC is applicable to other disease diagnosis for which multiple biomarkers need to be tested, such as the Parkinson’s disease and cancer. The key benefit of STC is to allow physicians to test the biomarkers sequentially with a known sequence optimized for each patient’s demographic profile, and on an as-needed basis. This would save the diagnostic time – a benefit to the patient, and the resources – a benefit to the health care provider. We plan to explore the application of STC to other diseases in future work.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

There are rich data available in today's healthcare systems due to technology advancements, such as diagnostic imaging, smart sensing, and health information systems, which offers a great opportunity of Precision Medicine. My research focus on developing data fusion and system informatics approaches for quality and performance improvement of healthcare systems from diagnosis to care to system-level decision-making. In my dissertation, I focus on three emerging problems in healthcare and develop novel statistical models and machine learning algorithms. In collaboration with healthcare domain experts, my research has explored a few healthcare domains, including imaging-based disease diagnosis, coordinated patient care, and system-level medical decision-making.

For disease diagnosis/subtyping, I proposed a new method, MFMM, for clustering multi-mode image data to discover migraine subtypes. MFMM employed a double-L_{21}-penalized likelihood formulation to enable hierarchical selection of imaging modes and features. We applied MFMM to migraine subtype discovery based on brain cortical area, cortical thickness, and volume measurements from MRI. Two clusters/subtypes were found and well separated using a total of seven factors. Subjects in the two clusters had significantly different clinical characteristics. Findings from this study showed promise for imaging-based subtyping of migraine and patient stratification toward Precision Medicine. In my future research, MFMM could be extended to include mixed-type features, and even applied to subtype discovery of other diseases.

For coordinated patient care, I developed a Multi-response Multi-level Model to fuse NCCI data and reveal how care coordination activities conducted by nurses are related to their demographics, workload, and characteristics of their practice environment. The long-term goal of this research is to inform interventions to improve staff nurse care coordination within hospital units that would in turn lead to improved patient outcomes, e.g., shorter length of stay, few medication errors, less likelihood for re-admission, and greater satisfaction.

There are many opportunities in healthcare systems for data science research, and as industrial engineer, I would like to consider medical problems from a system level and involve multi-perspectives, such as accuracy, efficiency, safety and quality, to develop novel systems engineering approaches and support system-level decision-making in healthcare.

REFERENCES

Ahn, M., Zhang, H. H., & Lu, W. 2012. "Moment-based method for random effects selection in linear mixed models." Statistica Sinica. 22(4), 1539.

Aiken, L. H., Clarke, S. P., Sloane, D. M., Lake, E. T., & Cheney, T. 2008. “Effects of hospital care environment on patient mortality and nurse outcomes.” The Journal of nursing administration, 38(5), 223.

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Snyder, P. J. 2011. "The diagnosis of mild cognitive impairment due to Alzheimer’s disease: Recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer's disease." Alzheimer's & dementia, 7(3), 270-279.

Alexopoulos, P., Kriett, L., Haller, B., Klupp, E., Gray, K., Grimmer, T., Drzezga, A. 2014. "Limited agreement between biomarkers of neuronal injury at different stages of Alzheimer's disease." Alzheimer's & Dementia, 10(6), 684-689.

Anderson, G. 2007. “Chronic Conditions: Making the case for ongoing care Chronic care chartbook". Baltimore: Johns Hopkins Bloomberg School of Public Health. pp. 1-74

Ankerst M, Breunig MM, Kriegel HP, Sander J. 1999. "OPTICS: ordering points to identify the clustering structure." ACM. 2: 49-60.

Baek J, McLachlan GJ, Flack LK. 2010. "Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data." IEEE transactions on pattern analysis and machine intelligence. 32: 1298-1309.

Barnes, D. E., Cenzer, I. S., Yaffe, K., Ritchie, C. S., Lee, S. J., Alzheimer's Disease Neuroimaging Initiative. 2014. "A point-based tool to predict conversion from mild cognitive impairment to probable Alzheimer's disease." Alzheimer's & Dementia, 10(6), 646-655.

Bondell, H. D., Krishna, A., & Ghosh, S. K. 2010. "Joint Variable Selection for Fixed and Random Effects in Linear Mixed‐Effects Models." Biometrics, 66(4), 1069-1077.

Bowling, S. R., Khasawneh, M. T., Kaewkuekool, S., Cho, B. R. 2009. "A logistic approximation to the cumulative normal distribution." Journal of Industrial Engineering and Management, 2(1), 114-127.

Box, G. E., & Cox, D. R. 1964. "An analysis of transformations." Journal of the Royal Statistical Society. Series B (Methodological), 211-252.

Burnham, K. P., & Anderson, D. R. 2004. "Multimodel inference understanding AIC and BIC in model selection." Sociological methods & research, 33(2), 261-304.

Burnham, K. P., Anderson, D. R. 2003. "Model selection and multimodel inference: a practical information-theoretic approach." Springer Science & Business Media.

Cerami, C., Della Rosa, P. A., Magnani, G., Santangelo, R., Marcone, A., Cappa, S. F., & Perani, D. 2015. "Brain metabolic maps in mild cognitive impairment predict heterogeneity of progression to dementia." NeuroImage: Clinical, 7, 187-194.

Chen, Z., Dunson, D. B. 2003. "Random effects selection in linear mixed models." Biometrics, 59(4), 762-769.

Chong CD and Schwedt TJ. 2015. "Migraine affects white-matter tract integrity: A diffusion-tensor imaging study." Cephalalgia. 35: 1162-1171.

Chong CD, Gaw N, Fu Y, et al. 2016. "Migraine classification using magnetic resonance imaging resting-state functional connectivity data." Cephalalgia.

Chong CD, Plasencia, J.D., Frakes, D.H., Schwedt, T.J. 2016. "Structural alterations of the brainstem in migraine." NeuroImage: Clinical. epub ahead of print.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Alzheimer's Disease Neuroimaging Initiative. 2011. "Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors." PloS one, 6(7), e21896.

Demidenko, E. 2013. "Mixed models: theory and applications with R." John Wiley & Sons.

Dempster, A. P., Laird, N. M., Rubin, D. B. 1977. "Maximum likelihood from incomplete data via the EM algorithm." Journal of the royal statistical society. Series B (methodological), 1-38.

Duva, I. H. 2010. "Factors impacting staff nurse care coordination (Doctoral dissertation, Emory University)."

Fahrmeir, L., Kneib, T., Konrath, S. 2010. "Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection." Statistics and Computing, 20(2), 203-219.

Fan J, Li R. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association. 96: 1348-1360.

Fan, J., & Li, R. 2001. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association, 96(456), 1348-1360.

Ford S, Calhoun A, Kahn K, et al. 2008. "Predictors of disability in migraineurs referred to a tertiary clinic: neck pain, headache characteristics, and coping behaviors." Headache. 48: 523-528.

George, E. I., & McCulloch, R. E. 1997. "Approaches for Bayesian variable selection." Statistica sinica, 7(2), 339-373.

Giardino A, Gupta S, Olson E, Sepulveda K, Lenchik L, Ivanidze J, Rakow-Penner R, Patel MJ, Subramaniam RM, Ganeshan D. 2017. "Role of Imaging in the Era of Precision Medicine." Academic Radiology. 24: 639-649.

Hadjikhani N, Ward N, Boshyan J, et al. 2013. "The missing link: Enhanced functional connectivity between amygdala and visceroceptive cortex in migraine." Cephalalgia. 33: 1264-1268.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. 2009. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Heister, D., Brewer, J. B., Magda, S., Blennow, K., McEvoy, L. K., Alzheimer's Disease Neuroimaging Initiative. 2011. "Predicting MCI outcome with clinically available MRI and CSF biomarkers." Neurology, 77(17), 1619-1628.

Hinrichs, C., Singh, V., Xu, G., Johnson, S. C., Alzheimers Disease Neuroimaging Initiative. 2011. "Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population." Neuroimage, 55(2), 574-589.

Ho RT, Fong TC, Cheung IK. 2014. "Cancer-related fatigue in breast cancer patients: factor mixture models with continuous non-normal distributions." Quality of Life Research. 23: 2909-2916.

Institute for Healthcare Improvement. 2004. "Transitions home how-to guide Innovation Series 2004." Cambridge Institute for Healthcare Improvement.

Jack, C. R., Bernstein, M. A., Fox, N. C., Thompson, P., Alexander, G., Harvey, D., Dale, A. M. 2008. "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods." Journal of Magnetic Resonance Imaging, 27(4), 685-691.

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., ... & Trojanowski, J. Q. 2010. "Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade." The Lancet Neurology, 9(1), 119-128.

Jack, C. R., Wiste, H. J., Vemuri, P., Weigand, S. D., Senjem, M. L., Zeng, G., Weiner, M. W. 2010. "Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer’s disease." Brain, 133(11), 3336-3348.

Jagust, W. J., Bandy, D., Chen, K., Foster, N. L., Landau, S. M., Mathis, C. A., Alzheimer's Disease Neuroimaging Initiative. 2010. "The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core." Alzheimer's & Dementia, 6(3), 221-229.

Jagust, W. J., Landau, S. M., Shaw, L. M., Trojanowski, J. Q., Koeppe, R. A., Reiman, E. M., Mathis, C. A. 2009. "Relationships between biomarkers in aging and dementia." Neurology, 73(15), 1193-1199.

Jin C, Yuan K, Zhao L, et al. 2013. "Structural and functional abnormalities in migraine patients without aura." NMR Biomed. 26: 58-64.

Kim S, Xing EP. 2012. "Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping." The Annals of Applied Statistics. 6: 1095-1117.

Kim, S., Swaminathan, S., Shen, L., Risacher, S. L., Nho, K., Foroud, T., Craig, D. W. 2011. "Genome-wide association study of CSF biomarkers Aβ1-42, t-tau, and p-tau181p in the ADNI cohort." Neurology, 76(1), 69-79.

Komarova NL, Thalhauser CJ. High Degree of Heterogeneity in Alzheimers Disease Progression Patterns. 2011. "PLoS Computational Biology." 7: e1002251.

Lake, E. T. 2002. "Development of the practice environment scale of the nursing work index." Research in nursing & health, 25(3), 176-188.

Lamb, G. S., Schmitt, M. H., Edwards, P., Sainfort, F., Duva, I., Higgins, M. 2008. "Measuring staff nurse care coordination in the hospital." In Invited presentation, National State of the Science Congress on Nursing Research. Washington, DC.

Landau, S. M., Harvey, D., Madison, C. M., Reiman, E. M., Foster, N. L., Aisen, P. S., Weiner, M. W. 2010. "Comparing predictors of conversion and decline in mild cognitive impairment." Neurology, 75(3), 230-238.

Landau, S. M., Mintun, M. A., Joshi, A. D., Koeppe, R. A., Petersen, R. C., Aisen, P. S., Jagust, W. J. 2012. "Amyloid deposition, hypometabolism, and longitudinal cognitive decline." Annals of neurology, 72(4), 578-586.

Laschinger, H. K. S., & Leiter, M. P. 2006. "The impact of nursing work environments on patient safety outcomes: The mediating role of burnout engagement." Journal of Nursing Administration, 36(5), 259-267.

Li L, Cheng WY, Glicksberg BS, Gottesman O, Tamler R, Chen R, Bottinger EP, Dudley JT. 2015. "Identification of type 2 diabetes subgroups through topological analysis of patient similarity." Science translational medicine. 7: 311ra174-311ra174.

Lipton RB, Bigal ME, Ashina S, et al. 2008. "Cutaneous allodynia in the migraine population. Ann Neurol." 63: 148-158.

Lipton RB, Munjal S, Buse DC, et al. 2016. "Predicting Inadequate Response to Acute Migraine Medication: Results From the American Migraine Prevalence and Prevention (AMPP) Study." Headache. 56: 1635-1648.

Lipton RB, Serrano D, Pavlovic JM, Manack AN, Reed ML, Turkel CC, Buse DC. 2014. "Improving the classification of migraine subtypes: an empirical approach based on factor mixture models in the American Migraine Prevalence and Prevention (AMPP) Study." Headache: The Journal of Head and Face Pain. 54: 830-849.

Liu J, Ji S, Ye J. 2009. "SLEP: Sparse learning with efficient projections. Arizona State University." 6:491.

Liu J, Zhao L, Li G, et al. 2012. "Hierarchical alteration of brain structural and functional networks in female migraine sufferers." PLoS One. 7: e51250.

Liu, C. C., Kanekiyo, T., Xu, H., and Bu, G. 2013. "Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy." Nature Reviews Neurology, 9(2), 106-118.

Louter MA, Bosker JE, van Oosterhout WP, et al. 2013. "Cutaneous allodynia as a predictor of migraine chronification." Brain. 136: 3489-3496.

Lubke GH, Muthen B, Moilanen IK, McGOUGH JJ, Loo SK, Swanson JM, Yang MH, Taanila A, Hurtig T, Järvelin MR, Smalley SL. 2007. "Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort." Journal of the American Academy of Child & Adolescent Psychiatry. 46: 1584-1593.

Lubke GH, Muthén B. 2005. "Investigating population heterogeneity with factor mixture models." Psychological methods. 10: 21-39.

Medicare Payment Advisory Commission. 2005. "Report to Congress: Promoting Greater Efficiency in Medicare (Washington, DC: MedPAC, June 2007) (Doctoral dissertation, These data refer to)."

Meier L, Van De Geer S, Bühlmann P. 2008. "The group lasso for logistic regression." Journal of the Royal Statistical Society: Series B (Statistical Methodology). 70: 53-71.

Melnykov V, Maitra R. 2010. "Finite mixture models and model-based clustering." Statistics Surveys. 4: 80-116.

Mickleborough MJ, Ekstrand C, Gould L, et al. 2016. "Attentional Network Differences Between Migraineurs and Non-migraine Controls: fMRI Evidence." Brain Topogr. 29: 419-428.

Misra, C., Fan, Y., and Davatzikos, C. 2009. "Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI." Neuroimage, 44(4), 1415-1422.

Montanari A, Viroli C. 2010. "Heteroscedastic factor mixture analysis." Statistical Modelling. 10: 441-460.

Moulton EA, Becerra L, Maleki N, et al. 2011. "Painful heat reveals hyperexcitability of the temporal pole in interictal and ictal migraine States." Cereb Cortex. 21: 435-448.

Moulton EA, Burstein R, Tully S, et al. 2008. "Interictal dysfunction of a brainstem descending modulatory center in migraine patients." PLoS One. 3: e3799.

Muthen B, Asparouhov T. 2006. "Item response mixture modeling: Application to tobacco dependence criteria." Addictive behaviors. 31: 1050-1066.

Page, A. (Ed.). 2004. "Keeping Patients Safe: Transforming the Work Environment of Nurses." National Academies Press.

Pan W, Shen X. 2007. "Penalized model-based clustering with application to variable selection." Journal of Machine Learning Research. 8: 1145-1164.

Pattyn T, Eede F, Lamers F, Veltman D, Sabbe BG, Penninx BW. 2015. "Identifying panic disorder subtypes using factor mixture modeling." Depression and anxiety. 32: 509-517.

Raftery AE, Dean N. 2006. "Variable selection for model-based clustering." Journal of the American Statistical Association. 101: 168-178.

Ram A, Jalal S, Jalal AS, Kumar M. 2010. "A density based algorithm for discovering density varied clusters in large spatial databases." International Journal of Computer Applications. 3: 1-4.

Risacher, S. L., Saykin, A. J., Wes, J. D., Shen, L., Firpi, H. A., and McDonald, B. C. 2009. "Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort." Current Alzheimer Research, 6(4), 347-361.

Russo A, Esposito F, Conte F, et al. 2016. "Functional interictal changes of pain processing in migraine with ictal cutaneous allodynia." Cephalalgia.

Russo A, Tessitore A, Esposito F, et al. 2012. "Pain processing in patients with migraine: an event-related fMRI study during trigeminal nociceptive stimulation." J Neurol. 259: 1903-1912.

Schelldorfer, J., Bühlmann, P., DE, G., VAN, S. 2011. "Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ1-Penalization." Scandinavian Journal of Statistics, 38(2), 197-214.

Schmitz N, Admiraal-Behloul F, Arkink EB, et al. 2008. "Attack frequency and disease duration as indicators for brain damage in migraine." Headache. 48: 1044-1055.

Schwedt TJ, Berisha V., Chong C.D. 2015. "Temporal lobe cortical thickness correlations differentiate the migraine brain from the healthy brain." PLoS One. 10: e0116687.

Schwedt TJ, Chiang CC, Chong CD, et48 al. 2015. "Functional MRI of migraine." The Lancet Neurology. 14: 81-91.

Schwedt TJ, Chong CD, Chiang CC, et al. 2014. "Enhanced pain-induced activity of pain-processing regions in a case-control study of episodic migraine." Cephalalgia. 34: 947-958.

Schwedt TJ, Chong CD, Wu T, Gaw N, Fu Y, Li J. 2015. "Accurate classification of chronic migraine via brain magnetic resonance imaging." Headache: The Journal of Head and Face Pain. 55: 762-777.

Schwedt TJ, Larson-Prior L, Coalson RS, et al. 2014. "Allodynia and Descending Pain Modulation in Migraine: A Resting State Functional Connectivity Analysis." Pain Med. 15: 154-165.

Schwedt TJ, Si B, Li J, et al. 2017. "Migraine sub-classification via a data-driven automated approach using multi-modality factor mixture modeling of brain structure measurements." Headache: The Journal of Head and Face Pain.

Schwedt TJ. 2013. "Multisensory integration in migraine." Curr Opin Neurol. 26: 248-253.

Simon, N., Friedman, J., Hastie, T., Tibshirani, R. 2013. "A sparse-group lasso." Journal of Computational and Graphical Statistics, 22(2), 231-245.

Sundaram, V., Bravata, D. M., Lewis, R., Lin, N., Kraft, S. A., Owens, D. K. 2007. "Closing the quality gap: a critical analysis of quality improvement strategies (Vol. 7: Care Coordination)."

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. 2005. "Sparsity and smoothness via the fused lasso." Journal of the Royal Statistical Society: Series B (Statistical Methodology). 67: 91-108.

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

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Tibshirani, R. 1996. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Trittschuh, E., Larson, E., Crane, P., Cholerton, B., McCurry, S., McCormick, W., and Craft, S. 2010. "Heterogeneity of MCI characterization across two time points in a community-based population." Alzheimer's & Dementia, 6(4), S78-S79.

van Rooden SM, Heiser WJ, Kok JN, Verbaan D, van Hilten JJ, Marinus J. 2010. "The identification of Parkinson's disease subtypes using cluster analysis: a systematic review. Movement Disorders. 25: 969-978.

Wang S, Zhu J. 2008. "Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data." Biometrics. 64: 440-448.

Wang, S., Nan, B., Rosset, S., Zhu, J. 2011. "Random lasso." The annals of applied statistics, 5(1), 468.

Wee, C. Y., Yap, P. T., and Shen, D. 2013. "Prediction of Alzheimer's disease and mild cognitive impairment using cortical morphological patterns." Human brain mapping, 34(12), 3411-3425.

Witten DM, Tibshirani R. 2010. "A framework for feature selection in clustering." Journal of the American Statistical Association. 105: 713-726.

Wu, C. J. 1983. "On the convergence properties of the EM algorithm." The Annals of statistics, 95-103.

Xie B, Pan W, Shen X. 2008. "Variable Selection in Penalized Model-Based Clustering Via Regularization on Grouped Parameters." Biometrics. 64: 921-930.

Yang Y, Zou H. 2015. "A fast unified algorithm for solving group-lasso penalize learning problems." Statistics and Computing. 25: 1129-1141.

Yang, Y. 2005. "Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation." Biometrika, 92(4), 937-950.

Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., and Narayan, V. A. 2012. "Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data." BMC neurology, 12(1), 1.

Yu, P., Dean, R. A., Hall, S. D., Qi, Y., Sethuraman, G., Willis, B. A., and Schwarz, A. J. 2012. "Enriching amnestic mild cognitive impairment populations for clinical trials: optimal combination of biomarkers to predict conversion to dementia. Journal of Alzheimer's Disease, 32(2), 373-385.

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: 49-67.

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), 49-67.

Zhang, D., Shen, D., and Alzheimer's Disease Neuroimaging Initiative. 2012. "Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers." PloS one, 7(3), e33182.

Zhang, N., Song, X., and Zhang, Y. 2012. "Combining structural brain changes improves the prediction of Alzheimer’s disease and mild cognitive impairment." Dementia and geriatric cognitive disorders, 33(5), 318-326.

Zhao L, Liu J, Dong X, et al. 2013. "Alterations in regional homogeneity assessed by fMRI in patients with migraine without aura stratified by disease duration." J Headache Pain. 14: 85.

Zhao, P., & Yu, B. 2006. "On model selection consistency of Lasso." The Journal of Machine Learning Research, 7, 2541-2563.

Zhao, P., Rocha, G., Yu, B. 2009. "The composite absolute penalties family for grouped and hierarchical variable selection." The Annals of Statistics, 3468-3497.

Zhou, N., Zhu, J. 2010. "Group variable selection via a hierarchical lasso and its oracle property." arXiv preprint arXiv:1006.2871.

Zou H, Hastie T. 2005. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology). 67: 301-320.

Zou, H. 2006. "The adaptive lasso and its oracle properties." Journal of the American statistical association. 101: 1418-1429.

APPENDIX A

SUPPLEMETNAL MATERIALS FOR CHAPTER 2

Appendix A-I: Deriving the expectations in (2.7)

(A-I.1) Deriving $E_{\boldsymbol{f}_{m,i},\boldsymbol{s}_i|\boldsymbol{x}_{1,i},\dots,\boldsymbol{x}_{M,i};\widetilde{\mathbf{\Theta}}}\left(-\log f\left(\boldsymbol{f}_{m,i}|\boldsymbol{s}_i;\mathbf{\Theta}\right)\right)$.

According to (2.3), the distribution of $\boldsymbol{f}_{m,i}|\ s_{k,i} = 1; \mathbf{\Theta}$ is $N\left(\mathbf{a}_{m,k}, \mathbf{\Sigma}_m\right)$. Inserting the probability density function of this distribution into the above expectation and ignoring constants, we can get

$$\sum_{k=1}^{K}\left\{\begin{array}{c}\frac{1}{2}log\left|\mathbf{\Sigma}_m\right| + \\ \frac{1}{2}\left(E\left(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}, s_{k,i} = 1; \widetilde{\mathbf{\Theta}}\right) - \mathbf{a}_{m,k}\right)^T \mathbf{\Sigma}_m^{-1}\left(E\left(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}, s_{k,i} = 1; \widetilde{\mathbf{\Theta}}\right) - \mathbf{a}_{m,k}\right) + \\ \frac{1}{2}tr\left(\mathbf{\Sigma}_m^{-1}E\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m.i}^T|\boldsymbol{x}_{m,i}, s_{k,i} = 1; \widetilde{\mathbf{\Theta}}\right)\right)\end{array}\right\}$$

$$f\left(s_{k,i} = 1|\boldsymbol{x}_{1,i}, \dots, \boldsymbol{x}_{M,i,}; \widetilde{\mathbf{\Theta}}\right). \qquad \text{(A-1)}$$

The distribution of $\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}, s_{k,i} = 1; \widetilde{\mathbf{\Theta}}$ is $N\left(\widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right), \widetilde{\mathbf{\Upsilon}}_{mk}\right)$, where $\widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right) = \widetilde{\mathbf{\Upsilon}}_{mk}\left(\widetilde{\mathbf{H}}_m^T\widetilde{\mathbf{\Psi}}_m^{-1}\boldsymbol{x}_{m,i} + \widetilde{\mathbf{\Sigma}}_m^{-1}\tilde{\mathbf{a}}_{m,k}\right)$ and $\widetilde{\mathbf{\Upsilon}}_{mk} = \left(\widetilde{\mathbf{H}}_m^T\widetilde{\mathbf{\Psi}}_m^{-1}\widetilde{\mathbf{H}}_m + \widetilde{\mathbf{\Sigma}}_m^{-1}\right)^{-1}$. Therefore, $E\left(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}, s_{k,i} = 1; \widetilde{\mathbf{\Theta}}\right) = \widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right)$ and $E\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m.i}^T|\boldsymbol{x}_{m,i}, s_{k,i} = 1; \widetilde{\mathbf{\Theta}}\right) = \widetilde{\mathbf{\Upsilon}}_{mk} - \widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right)^T\widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right)$ in (A-1). Finally, to derive the $f\left(s_{k,i} = 1|\boldsymbol{x}_{1,i}, \dots, \boldsymbol{x}_{M,i,}; \widetilde{\mathbf{\Theta}}\right)$ in (A-1), we can use the Bayes' theorem and get

$$f\left(s_{k,i} = 1|\boldsymbol{x}_{1,i}, \dots, \boldsymbol{x}_{M,i,}; \widetilde{\mathbf{\Theta}}\right) = \frac{\widetilde{w}_k\ \prod_{m=1}^{M} f\left(\boldsymbol{x}_{m,i}|s_{k,i}=1;\widetilde{\mathbf{\Theta}}\right)}{\sum_{k=1}^{K}\widetilde{w}_k\ \prod_{m=1}^{M} f\left(\boldsymbol{x}_{m,i}|s_{k,i}=1;\widetilde{\mathbf{\Theta}}\right)},$$

where $\boldsymbol{x}_{m,i}|s_{k,i} = 1; \widetilde{\mathbf{\Theta}} \sim N\left(\widetilde{\mathbf{H}}_m\tilde{\mathbf{a}}_{m,k} + \widetilde{\mathbf{B}}_m\boldsymbol{z}_i\,, \widetilde{\mathbf{H}}_m\widetilde{\mathbf{\Sigma}}_m\widetilde{\mathbf{H}}_m^T + \widetilde{\mathbf{\Psi}}_m\right)$.

(A-I.2) Deriving $E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\widetilde{\mathbf{\Theta}}}\left(-\log f\left(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \mathbf{\Theta}\right)\right)$.

According to (2.1), the distribution of $\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i}, \boldsymbol{z}_i; \mathbf{\Theta}$ is $N\left(\mathbf{H}_m\boldsymbol{f}_{m,i} + \mathbf{B}_m\boldsymbol{z}_i\,, \mathbf{\Psi}_m\right)$. Inserting the probability density function of this distribution into the above expectation and ignoring constants, we can get

$$\frac{1}{2}log|\boldsymbol{\Psi}_m| +$$
$$\frac{1}{2}\left(\boldsymbol{x}_{m,i} - \mathbf{H}_m E\left(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{B}_m\boldsymbol{z}_i\ \right)^T \boldsymbol{\Psi}_m^{-1}\left(\boldsymbol{x}_{m,i} - \mathbf{H}_m E\left(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right) - \mathbf{B}_m\boldsymbol{z}_i\ \right) +$$
$$\frac{1}{2}tr\left(\mathbf{H}_m^T\boldsymbol{\Psi}_m^{-1}\mathbf{H}_m\left(E\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m.i}^T\middle|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right) - E\left(\boldsymbol{f}_{m,i}\middle|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right)E\left(\boldsymbol{f}_{m,i}\middle|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right)^T\right)\right)$$

(A-2)

Using the result in (A-I.1), we can get $E\left(\boldsymbol{f}_{m,i}\middle|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right) = \sum_{k=1}^{K}\widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right)f\left(s_{k,i} = 1|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right)$, and $E\left(\boldsymbol{f}_{m,i}\boldsymbol{f}_{m.i}^T\middle|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}\right) = \sum_{k=1}^{K}\left(\widetilde{\boldsymbol{\Upsilon}}_{mk} - \widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right)^T\widetilde{\boldsymbol{\rho}}_{mk}\left(\boldsymbol{x}_{m,i}\right)\right)f\left(s_{k,i} = 1|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}\right)$ in (A-2).

(A-I.3) Deriving $E_{\boldsymbol{s_i}|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}}(-\log f(\boldsymbol{s}_i;\boldsymbol{\Theta}))$.

According to (2.2), $\log f(\boldsymbol{s}_i;\boldsymbol{\Theta}) = \sum_{k=1}^{K}s_{k,i}log(w_k)$. Inserting it into the above expectation, we can get $E_{\boldsymbol{s_i}|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}}(-\log f(\boldsymbol{s}_i;\boldsymbol{\Theta})) = -\sum_{k=1}^{K}log(w_k)\,f\left(s_{k,i} = 1|\boldsymbol{x}_{1,i},\ldots,\boldsymbol{x}_{M,i};\widetilde{\boldsymbol{\Theta}}\right)$. Δ

Appendix A-II: Proof of Proposition 1

We first need to write the minimization problem in (2.9) into the form of (2.11). This can be achieved by making $J = P_m$, $\boldsymbol{\beta}^{(j)} = \mathbf{h}_m^j$, and

$$L(\boldsymbol{\beta}|\mathbf{D}) = \sum_{i=1}^{N}E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\widetilde{\boldsymbol{\Theta}}}\left(-\log f\left(\boldsymbol{x}_{m,i}|\boldsymbol{f}_{m,i},\boldsymbol{z}_i;\boldsymbol{\Theta}\right)\right). \quad \text{(A-3)}$$

Since the QM condition requires $L(\boldsymbol{\beta}|\mathbf{D})$ satisfying two assumption, we will need to write $L(\boldsymbol{\beta}|\mathbf{D})$ into a format that facilitates checking of the assumptions. Through some derivation and dropping the terms not involving $\boldsymbol{\beta}$, we can write $L(\boldsymbol{\beta}|\mathbf{D})$ as

$$L(\boldsymbol{\beta}|\mathbf{D}) = \boldsymbol{\beta}^T\left(\sum_{i=1}^{N}\mathbf{C}_{mi}\right)\boldsymbol{\beta} - 2\left(\sum_{i=1}^{N}\boldsymbol{b}_{m,i}^T\right)\boldsymbol{\beta} \quad \text{(A-4)}$$

where $\mathbf{C}_{mi} = \frac{1}{2} tr(\boldsymbol{\Psi}_m^{-1}) \Bigg[\Bigg(\mathbf{1}_{P_M \times P_M} \otimes \Big(E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}}(\boldsymbol{f}_{m,i}) \Big)^T \Bigg)^T \Bigg(\mathbf{1}_{P_M \times P_M} \otimes$ $\Big(E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}}(\boldsymbol{f}_{m,i}) \Big)^T \Bigg) + (\mathbf{1}_{P_m}^T \otimes \mathbf{1}_{r_m}^T)(\mathbf{1}_{P_m}^T \otimes \mathbf{1}_{r_m}^T)^T tr \Big(E \Big(\boldsymbol{f}_{m,i}(\boldsymbol{f}_{m,i})^T \Big| \boldsymbol{x}_{m,i}; \boldsymbol{\Theta}^{(\omega)} \Big) -$ $E(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}; \boldsymbol{\Theta}^{(\omega)}) E(\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i}; \boldsymbol{\Theta}^{(\omega)})^T \Big) \Bigg]$ and $\boldsymbol{b}_{m,i}^T = \frac{1}{2} tr(\boldsymbol{\Psi}_m^{-1}) \boldsymbol{x}_{m,i}^T \Bigg(\mathbf{1}_{P_M \times P_M} \otimes$ $\Big(E_{\boldsymbol{f}_{m,i}|\boldsymbol{x}_{m,i};\boldsymbol{\Theta}^{(\omega)}}(\boldsymbol{f}_{m,i}) \Big)^T \Bigg).$

Next, we will prove (A-4) satisfy the two assumptions required by the QM condition:

(i) It is straightforward to get $\nabla L(\boldsymbol{\beta}|\mathbf{D}) = 2\big(\sum_{i=1}^{N} \mathbf{C}_{mi}\, \boldsymbol{\beta} - \sum_{i=1}^{N} \boldsymbol{b}_{m,i}^T\big)$, which exist everywhere.

(ii) To prove this assumption, we define a function $l(t) = L(\boldsymbol{\beta}^* + t(\boldsymbol{\beta} - \boldsymbol{\beta}^*)|\mathbf{D})$. By the mean value theorem, there exists $a \in (0, 1)$ such that

$$l(1) = l(0) + l'(a) = l(0) + l'(0) + \big(l'(a) - l'(0)\big). \qquad \text{(A-5)}$$

Using the $L(\boldsymbol{\beta}|\mathbf{D})$ in (A-4), we can get

$$l'(0) = 2(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \big(\textstyle\sum_{i=1}^{N} \mathbf{C}_{mi}\, \boldsymbol{\beta}^* - \sum_{i=1}^{N} \boldsymbol{b}_{m,i}^T\big) = (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla L(\boldsymbol{\beta}^*|\mathbf{D}), \quad \text{(A-6)}$$

and

$$l'(a) - l'(0) = 2a(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \textstyle\sum_{i=1}^{N} \mathbf{C}_{mi}\, (\boldsymbol{\beta} - \boldsymbol{\beta}^*) = \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T (4a \cdot \sum_{i=1}^{N} \mathbf{C}_{mi})(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \le (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T (4 \cdot \sum_{i=1}^{N} \mathbf{C}_{mi})(\boldsymbol{\beta} - \boldsymbol{\beta}^*). \qquad \text{(A-7)}$$

Substituting (A-6) and (A-7) into (A-5), we have

$$l(1) \le l(0) + (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T \nabla L(\boldsymbol{\beta}^*|\mathbf{D}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^T (4 \cdot \textstyle\sum_{i=1}^{N} \mathbf{C}_{mi})(\boldsymbol{\beta} - \boldsymbol{\beta}^*).$$

Noting that $l(1) = L(\boldsymbol{\beta}|\mathbf{D})$, $l(0) = L(\boldsymbol{\beta}^*|\mathbf{D})$, and let $\boldsymbol{\Lambda} = 4 \cdot \sum_{i=1}^{N} \mathbf{C}_{mi}$, we proved the second assumption of the QM condition, i.e., (3.12), holds for our problem. Δ

APPENDIX B

SUPPLEMETNAL MATERIALS FOR CHAPTER 3

Appendix B-I: Pseudo code for the EM and BCD algorithms for parameter estimation.

```
Input: data of predictors, $\mathbf{W}$, $\mathbf{Z}$; data of $\mathbb{S}$ response variables, $\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}$; regularization
    parameters, $\lambda_1$, $\lambda_2$; adaptive weights $\tilde{\beta}_{ps}$, $\tilde{d}_{qs}$, $p = 1, \dots, \mathbb{P}, q = 1, \dots, \mathbb{Q}, s =$
    $1, \dots, \mathbb{S}$.
Initialize:
        $\boldsymbol{\beta}_s^{(0)}$, $\boldsymbol{d}_s^{(0)}$, $\boldsymbol{\Gamma}_s^{(0)}$ by fitting a multilevel model for each response separately,
        $s = 1, \dots, \mathbb{S}$.
        $\omega \leftarrow 0$.
Iterate until convergence:
        E-step: compute $\hat{\boldsymbol{e}}_s^{(\omega)}$, $\mathbf{U}_s^{(\omega)}$, $\sigma_s^{2(\omega)}$ using (10), (11), and (12), respectively.
        M-step: alternate between following two sub-steps to get $\boldsymbol{\beta}_s^{(\omega+1)}$, $\boldsymbol{d}_s^{(\omega+1)}$,
        $\boldsymbol{\Gamma}_s^{(\omega+1)}$.
                (i) Solve $\boldsymbol{\Gamma}_s^{(\omega+1)}$ analytically, given $\boldsymbol{\beta}_s^{(\omega)}$, $\boldsymbol{d}_s^{(\omega)}$
                (ii) Solve $\boldsymbol{\beta}_s^{(\omega+1)}$, $\boldsymbol{d}_s^{(\omega+1)}$ using the BCD algorithm in Figure 3, given
                $\boldsymbol{\Gamma}_s^{(\omega+1)}$
        $\omega \leftarrow \omega + 1$.
Output: estimates for $\boldsymbol{\beta}_s$, $\boldsymbol{d}_s$, $\boldsymbol{\Gamma}_s$, $\sigma_s^2$, $s = 1, \dots, \mathbb{S}$.
```

Figure B1: An EM framework for estimating the proposed model in (3.6)

```
Input: regularization parameters, $\lambda_1, \lambda_2$; current estimates for $\sigma_s^2$, $\mathbf{A}_s$, $\boldsymbol{b}_s^T$, $s = 1, \dots, \mathbb{S}$;
    adaptive weights $\tilde{\beta}_{ps}, \tilde{d}_{qs}, p = 1, \dots, \mathbb{P}, q = 1, \dots, \mathbb{Q}, s = 1, \dots, \mathbb{S}$.
Iterate until convergence:
  At each iteration, update $\mathbb{P} + \mathbb{Q}$ coordinates one-by-one (let $\hat{\boldsymbol{\beta}}_s$, $\hat{\boldsymbol{d}}_s$ denote the estimates obtained in the previous iteration):
    Update the p-th fixed effect:
    $\tilde{l}_{ps} \leftarrow \frac{1}{\sigma_s^2}\left(\mathbf{A}_s^p[\hat{\beta}_{1s} \quad \cdots \quad \hat{\beta}_{p-1,s} \quad 0 \quad \hat{\beta}_{p+1,s} \quad \cdots \quad \hat{\beta}_{\mathbb{P}s} \quad \hat{\boldsymbol{d}}_s^T]^T - b_{ps}\right), s = 1, \dots, \mathbb{S}$
    If $\left\|\left(\tilde{l}_{p1} \times \tilde{\beta}_{p1}, \dots, \tilde{l}_{p\mathbb{S}} \times \tilde{\beta}_{p\mathbb{S}}\right)\right\|_2 \le \lambda_1$
      $\boldsymbol{\beta^p} \leftarrow \mathbf{0}$
    Else
      Do a one-dimensional search over $\boldsymbol{\beta}^p = \left(\beta_{p1}, \dots, \beta_{p\mathbb{S}}\right)$ as follows:
      If $\hat{\beta}_{p1}, \dots, \hat{\beta}_{p,s-1}, \hat{\beta}_{p,s+1}, \dots, \hat{\beta}_{p,\mathbb{S}}$ are all zeros and $\left|\tilde{l}_{ps} \times \tilde{\beta}_{ps}\right| \le \lambda_1$
        $\beta_{ps} \leftarrow 0$
      Else if $\hat{\beta}_{p1}, \dots, \hat{\beta}_{p,s-1}, \hat{\beta}_{p,s+1}, \dots, \hat{\beta}_{p,\mathbb{S}}$ are not all zeros and $\tilde{l}_{ps} = 0$
        $\beta_{ps} \leftarrow 0$
      Else
        Use standard software to solve the minimization problem with respect to $\beta_{ps}$
      End if
    End if
    Update the q-th random effect in a similar way.
```

Figure B2: A BCD algorithm for solving the optimization in (3.12)

Appendix B-II: Proof of Theorem 1

We will show the existence of a local maximizer of $Q\left(\{\boldsymbol{\phi}_s\}_{s=1}^{\mathbb{S}}\right)$ in the neighborhood of the true value $\tilde{\boldsymbol{\phi}}^1$. To achieve this purpose, we show that for an arbitrary positive ε, there exits a sufficiently large non-zero constant C such that for a sufficiently large N,

$$P\left\{\begin{matrix} sup \\ \|u\| = C \end{matrix} \; Q\begin{pmatrix} \tilde{\boldsymbol{\phi}}^1 + \frac{u}{\sqrt{N}} \\ \mathbf{0} \end{pmatrix} < Q\begin{pmatrix} \tilde{\boldsymbol{\phi}}^1 \\ \mathbf{0} \end{pmatrix}\right\} \ge 1 - \varepsilon. \tag{B-1}$$

Note that

$$D_m(u) = \; Q\begin{pmatrix} \widetilde{\boldsymbol{\phi}}^1 + \frac{u}{\sqrt{N}} \\ \mathbf{0} \end{pmatrix} - Q\begin{pmatrix} \widetilde{\boldsymbol{\phi}}^1 \\ \mathbf{0} \end{pmatrix} = \sum_{s=1}^{S} \left\{ l\left(\widetilde{\boldsymbol{\phi}}_{s1}^T + \frac{u_s}{\sqrt{N}}\right) - \right.$$

$$\left. l\left(\widetilde{\boldsymbol{\phi}}_{s1}^T\right) \right\} - \lambda_{N1}\left(\sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps(1)} + \frac{u\,ps}{\sqrt{N}}}{\widetilde{\beta}_{ps(1)}}\right)^2} - \sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps(1)}}{\widetilde{\beta}_{ps(1)}}\right)^2} \right) -$$

$$\lambda_{N2}\left(\sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs(1)} + \frac{u\,qs}{\sqrt{N}}}{\tilde{d}_{qs(1)}}\right)^2} - \sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs(1)}}{\tilde{d}_{qs(1)}}\right)^2} \right).$$

Using a Taylor series expansion, we have

$$D_m(u) = \sum_{s=1}^{S} \left\{ \frac{1}{\sqrt{N}} \left(\nabla l\left(\widetilde{\boldsymbol{\phi}}_{s1}\right)\right)^T u_s + \frac{1}{2N} {u_s}^T \left[\nabla^2 l\left(\widetilde{\boldsymbol{\phi}}_{s1}\right)\right] u_s \right\}$$

$$-\lambda_{N1}\left(\sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps(1)} + \frac{u\,ps}{\sqrt{N}}}{\widetilde{\beta}_{ps(1)}}\right)^2} - \sum_{p=1}^{\mathbb{P}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{\beta_{ps(1)}}{\widetilde{\beta}_{ps(1)}}\right)^2} \right) -$$

$$\lambda_{N2}\left(\sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs(1)} + \frac{u\,qs}{\sqrt{N}}}{\tilde{d}_{qs(1)}}\right)^2} - \sum_{q=1}^{\mathbb{Q}} \sqrt{\sum_{s=1}^{\mathbb{S}} \left(\frac{d_{qs(1)}}{\tilde{d}_{qs(1)}}\right)^2} \right). \qquad \text{(B-2)}$$

Under common regularity conditions, the remainder term vanishes. For the first partial derivatives of $l(\boldsymbol{\phi}_{s1}), \nabla l(\boldsymbol{\phi}_{s1})$, the e-th partial derivative for each corresponding $\boldsymbol{\beta}_{s(1)}$, $\boldsymbol{d}_{s(1)}$, and $\boldsymbol{\gamma}_{s(1)}$ satisfies

$$E\left\{\frac{\partial}{\partial \beta_{e_{s(1)}}} l(\boldsymbol{\phi}_{s1})\right\} = E\left[\mathbf{W}_{(1)e}^T \widetilde{\mathbf{V}}_{s(1)}^{-1} \left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]|_{\boldsymbol{\phi}_{s1}=\widetilde{\boldsymbol{\phi}}_{s1}} = 0,$$

$$E\left\{\frac{\partial}{\partial d_{e_{s(1)}}} l(\boldsymbol{\phi}_{s1})\right\}$$

$$= E\left[\frac{1}{2}\left[-Tr\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{S}_{s(1)}^{e}\right)\right.\right.$$

$$\left.\left. + \left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)^T \left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{S}_{s(1)}^{e}\widetilde{\mathbf{V}}_{s(1)}^{-1}\right)\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]\right] |_{\boldsymbol{\phi}_{s1}=\widetilde{\boldsymbol{\phi}}_{s1}} = 0,$$

$E\left\{\frac{\partial}{\partial \gamma_{e_{s(1)}}} l(\boldsymbol{\phi}_{s1})\right\} = E\left[\frac{1}{2}\left[-Tr\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{T}_{s(1)}^{e}\right) + \left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)^T\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{T}_{s(1)}^{e}\widetilde{\mathbf{V}}_{s(1)}^{-1}\right)\left(\boldsymbol{y}_s -\right.\right.\right.$

$\left.\left.\left.\mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]\right] |_{\boldsymbol{\phi}_{s1}=\widetilde{\boldsymbol{\phi}}_{s1}} = 0,$

where $\mathbf{W}_{(1)e}$ corresponds to the e-th column of stacked matrix $\mathbf{W}_{(1)}$, and $\mathbf{S}_{s(1)}^{e}$ and $\mathbf{T}_{s(1)}^{e}$ are block diagonal matrices of the partial derivatives of $\tilde{V}_{s(1)}$ and are given by

$\mathbf{S}_{s(1)}^{e} = \mathbf{Z}_{s(1)}\left\{\frac{\partial}{\partial d_{es(1)}}\left(\left(\mathbf{I}\otimes\ \mathbf{D}_{s(1)}\right)\left(\mathbf{I}\otimes\ \boldsymbol{\Gamma}_{s(1)}\right)\left(\mathbf{I}\otimes\ \boldsymbol{\Gamma}_{s(1)}\right)^T\left(\mathbf{I}\otimes\ \mathbf{D}_{s(1)}\right)^T\right)\right\}\mathbf{Z}_{s(1)}^{T}$ and

$\mathbf{T}_{s(1)}^{e} = \mathbf{Z}_{s(1)}\left(\mathbf{I}\otimes\ \mathbf{D}_{s(1)}\right)\left\{\frac{\partial}{\partial \gamma_{es(1)}}\left(\left(\mathbf{I}\otimes\ \boldsymbol{\Gamma}_{s(1)}\right)\left(\mathbf{I}\otimes\ \boldsymbol{\Gamma}_{s(1)}\right)^T\right)\right\}\left(\mathbf{I}\otimes\ \mathbf{D}_{s(1)}\right)^T\mathbf{Z}_{s(1)}^{T}.$

For total number of response $0 \leq S < \infty$, we have

$$\frac{1}{\sqrt{N}}\sum_{s=1}^{S}\left[\mathbf{W}_{(1)e}^{T}\widetilde{\mathbf{V}}_{s(1)}^{-1}\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right] |_{\boldsymbol{\phi}_{s1}=\widetilde{\boldsymbol{\phi}}_{s1}} = \boldsymbol{O}_p(1)$$

$$\frac{1}{\sqrt{N}}\sum_{s=1}^{S}\left[\frac{1}{2}\left[-Tr\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{S}_{s(1)}^{e}\right) + \left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)^T\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{S}_{s(1)}^{e}\widetilde{\mathbf{V}}_{s(1)}^{-1}\right)\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]\right] |_{\boldsymbol{\phi}_{s1}=\widetilde{\boldsymbol{\phi}}_{s1}} = \boldsymbol{O}_p(1)$$

$$\frac{1}{\sqrt{N}}\sum_{s=1}^{S}\left[\frac{1}{2}\left[-Tr\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{T}_{s(1)}^{e}\right) + \left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)^T\left(\widetilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{T}_{s(1)}^{e}\widetilde{\mathbf{V}}_{s(1)}^{-1}\right)\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]\right] |_{\boldsymbol{\phi}_{s1}=\widetilde{\boldsymbol{\phi}}_{s1}} = \boldsymbol{O}_p(1). \tag{B-3}$$

Also,

$$\frac{1}{N}\sum_{s=1}^{S}\nabla^2 l\left(\widetilde{\boldsymbol{\phi}}_{s1}\right) \longrightarrow_p -\sum_{s=1}^{S} I\left(\widetilde{\boldsymbol{\phi}}_{s1}\right), \tag{B-4}$$

where $I\left(\widetilde{\boldsymbol{\phi}}_{s1}\right)$ is the Fisher information evaluated at $\widetilde{\boldsymbol{\phi}}_{s1}$.

Substituting (S-3) and (S-4) into (S-2), we have

$$D_m(u) = \sum_{s=1}^{S}\left\{\boldsymbol{O}_p(1)u_s - \frac{1}{2}u_s{}^T\left[I\left(\widetilde{\boldsymbol{\phi}}_{s1}\right) + o_p(1)\right]u_s\right\}$$

$$-\lambda_{N1}\left(\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps(1)}+\frac{u\,ps}{\sqrt{N}}}{\tilde{\beta}_{ps(1)}}\right)^2}-\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps(1)}}{\tilde{\beta}_{ps(1)}}\right)^2}\right)-$$

$$\lambda_{N2}\left(\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs(1)}+\frac{u\,qs}{\sqrt{N}}}{\tilde{d}_{qs(1)}}\right)^2}-\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs(1)}}{\tilde{d}_{qs(1)}}\right)^2}\right).$$

For the penalty term, if $\frac{\lambda_{N1}}{\sqrt{N}} \longrightarrow 0$ and $\frac{\lambda_{N2}}{\sqrt{N}} \longrightarrow 0$ as $N \longrightarrow \infty$,

$$\lambda_{N1}\left(\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps(1)}+\frac{u\,ps}{\sqrt{N}}}{\tilde{\beta}_{ps(1)}}\right)^2}-\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps(1)}}{\tilde{\beta}_{ps(1)}}\right)^2}\right)\longrightarrow_p 0, \text{ and}$$

$$\lambda_{N2}\left(\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs(1)}+\frac{u\,qs}{\sqrt{N}}}{\tilde{d}_{qs(1)}}\right)^2}-\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs(1)}}{\tilde{d}_{qs(1)}}\right)^2}\right)\longrightarrow_p 0.$$

For $D_m(u)$, under regularity conditions, $I(\widetilde{\boldsymbol{\phi}}_{s1})$ is finite and positive definite, hence the second term dominates the first term and the penalty term uniformly in $\|u\| = C$ for a sufficiently large number C. Hence, there exists a local maximum in the ball $\left\{\begin{pmatrix}\widetilde{\boldsymbol{\phi}}^1+\frac{u}{\sqrt{N}}\\ \mathbf{0}\end{pmatrix} \mid \|u\| \leq C\right\}$ with probability $1-\ \varepsilon$, and hence there exists a local maximizer $\widehat{\boldsymbol{\phi}} = \begin{pmatrix}\widehat{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}$ of $\widetilde{\boldsymbol{\phi}} = \begin{pmatrix}\widetilde{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}$ such that $\left\|\widehat{\boldsymbol{\phi}}^1 - \widetilde{\boldsymbol{\phi}}^1\right\| \ \ = \boldsymbol{O}_p(\frac{1}{\sqrt{N}})$. Δ

Appendix B-III: Proof of Theorem 2

To be clear, let’s fix the notation first. For the parameter of response s, $\boldsymbol{\phi}_s = (\boldsymbol{\phi}_{s1}^T, \boldsymbol{\phi}_{s2}^T)^T$, the sum of lengths corresponding to each parameter is $k_s = k_{s1} + k_{s2} = k_\beta + k_d + k_\gamma = k_{\beta 1} + k_{d1} + k_{\gamma 1} + k_{\beta 2} + k_{d2} + k_{\gamma 2}$.

To prove theorem 2, it’s sufficient to show that with probability tending to 1 as $N \longrightarrow \infty$, for any $\boldsymbol{\phi}^1$ satisfying $\left\|\boldsymbol{\phi}^1 - \widetilde{\boldsymbol{\phi}}^1\right\| \leq \frac{C}{\sqrt{N}}$ and for some small $\varepsilon_N = \frac{C}{\sqrt{N}}$ and for each $e_s = (k_{s1}+1), \cdots, (k_{s1}+k_{s2})$, we have

$$\frac{\partial}{\partial \phi_{e_s}} Q\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix} < 0, \text{ for } 0 < \phi_{e_s} < \varepsilon_N \text{ ,}$$

$$\frac{\partial}{\partial \phi_{e_s}} Q\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix} > 0, \text{ for } -\varepsilon_N < \phi_{e_s} < 0. \qquad \text{(B-5)}$$

Note that

$$\frac{\partial}{\partial \phi_{e_s}} Q\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix} = \frac{\partial}{\partial \phi_{e_s}}\, l(\boldsymbol{\phi}_s) - \frac{\partial}{\partial \phi_{e_s}}\left(\lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\widetilde{\beta}_{ps}}\right)^2} + \lambda_{N2}\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2}\right).$$

Using Taylor expansion about $\frac{\partial}{\partial \phi_{e_s}}\, l(\boldsymbol{\phi}_s)$, we have

$$\frac{\partial}{\partial \phi_{e_s}} Q\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix} = \frac{\partial}{\partial \phi_{e_s}}\, l(\widetilde{\boldsymbol{\phi}}_{s1}) - \sum_{f_s=1}^{k_s}\frac{\partial^2}{\partial \phi_{e_s}\partial \phi_{f_s}} l(\widetilde{\boldsymbol{\phi}}_{s1})(\phi_{f_s} - \tilde{\phi}_{f_s}) +$$

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{f_s=1}^{k_s}\sum_{g_s=1}^{k_s}\frac{\partial^3}{\partial \phi_{e_s}\partial \phi_{f_s}\partial \phi_{g_s}} l_i(\boldsymbol{\phi}_{s1}^*)(\phi_{f_s} - \tilde{\phi}_{f_s})(\phi_{g_s} - \tilde{\phi}_{g_s}) -$$

$$\frac{\partial}{\partial \phi_{e_s}}\left(\lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\widetilde{\beta}_{ps}}\right)^2} + \lambda_{N2}\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs}}{\tilde{d}_{qs}}\right)^2}\right), \qquad \text{(B-6)}$$

where $\boldsymbol{\phi}_{s1}^*$ lies between $\widetilde{\boldsymbol{\phi}}_{s1}$ and $\boldsymbol{\phi}_{s1}$.

In the proof of Theorem 1, the first order partial derivative for the ${e_s}^{th}$term of $\beta_{s(1)}$ and $d_{s(1)}$ is

$$\frac{\partial}{\partial \beta_{e_{s(1)}}} l\left(\widetilde{\boldsymbol{\phi}}_{s1}\right) = \mathbf{W}_{(1)e_s}^T \widetilde{\widetilde{\mathbf{V}}}_{s(1)}^{-1}\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\widetilde{\boldsymbol{\beta}}_{s(1)}\right).$$

$$\frac{\partial}{\partial d_{e_{s(1)}}} l\left(\widetilde{\boldsymbol{\phi}}_{s1}\right) = 0,$$

respectively. The second order derivatives are

$$\frac{1}{N}\nabla^2 l(\boldsymbol{\phi}_s)|_{\boldsymbol{\phi}_s=\widetilde{\boldsymbol{\phi}}_s} \longrightarrow_p - I(\boldsymbol{\phi}_s)|_{\boldsymbol{\phi}_s=\widetilde{\boldsymbol{\phi}}_s} = \frac{1}{N}E\left(\nabla^2 l(\boldsymbol{\phi}_s)\right)|_{\boldsymbol{\phi}_s=\widetilde{\boldsymbol{\phi}}_s},$$

where $E\left(\nabla^2 l(\boldsymbol{\phi}_s)\right)$ is given as

$$E\left(\nabla^2 l(\boldsymbol{\phi}_s)\right) = E\begin{bmatrix} L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{\beta}_{s(1)}} & L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{d}_{s(1)}} & L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{\gamma}_{s(1)}} \\ L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{d}_{s(1)}} & L_{\boldsymbol{d}_{s(1)}\boldsymbol{d}_{s(1)}} & L_{\boldsymbol{d}_{s(1)}\boldsymbol{\gamma}_{s(1)}} \\ L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{\gamma}_{s(1)}} & L_{\boldsymbol{d}_{s(1)}\boldsymbol{\gamma}_{s(1)}} & L_{\boldsymbol{\gamma}_{s(1)}\boldsymbol{\gamma}_{s(1)}} \end{bmatrix},$$

where $\left\{L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{\beta}_{s(1)}}\right\} = -\mathbf{W}_{(1)}^T\tilde{\mathbf{V}}_{s(1)}^{-1}\ \mathbf{W}_{(1)}$, and $E\left\{L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{d}_{s(1)}}\right\}$ and $E\left\{L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{\gamma}_{s(1)}}\right\}$ have the ${e_s}^{th}$

column being

$$E\left\{L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{d}_{s(1)}}\right\}_{e_s} = -E\left[\mathbf{W}_{(1)e_s}^T\left(\tilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{S}_{s(1)}^{e_s}\tilde{\mathbf{V}}_{s(1)}^{-1}\right)\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]|_{\boldsymbol{\phi}_s=\widetilde{\boldsymbol{\phi}}_s} = 0,$$

$$E\left\{L_{\boldsymbol{\beta}_{s(1)}\boldsymbol{\gamma}_{s(1)}}\right\}_{e_s} = -E\left[\mathbf{W}_{(1)e_s}^T\left(\tilde{\mathbf{V}}_{s(1)}^{-1}\mathbf{T}_{s(1)}^{e_s}\tilde{\mathbf{V}}_{s(1)}^{-1}\right)\left(\boldsymbol{y}_s - \mathbf{W}_{(1)}\boldsymbol{\beta}_{s(1)}\right)\right]|_{\boldsymbol{\phi}_s=\widetilde{\boldsymbol{\phi}}_s} = 0\ ,$$

respectively.

Considering $\phi_{e_s} = \beta_{e_s(1)}$, the expansion given in (B-6) yields

$$\frac{1}{\sqrt{N}}\frac{\partial}{\partial \beta_{e_s(1)}} Q\left\{\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix}\right\}$$

$$= \frac{1}{\sqrt{N}}\Bigg(\boldsymbol{o}_p(\sqrt{N}) - \sum_{f_s=1}^{k_\beta}\left\{-\mathbf{w}_{(1)e_s}^T \tilde{\mathbf{V}}_{s(1)}^{-1} \mathbf{w}_{(1)f_s} + o_p(1)\right\}\left(\beta_{f_s} - \tilde{\beta}_{f_s}\right)$$

$$- \sum_{f_s=k_\beta+1}^{k_\beta+k_d} o_p(1)\left(d_{f_s} - \tilde{d}_{f_s}\right) - \sum_{f_s=k_\beta+k_d+1}^{k_\beta+k_d+k_\gamma} o_p(1)\left(\gamma_{f_s} - \tilde{\gamma}_{f_s}\right)$$

$$+ \sum_{i=1}^{N}\sum_{f_s=1}^{k_\beta}\sum_{g_s=k_\beta+1}^{k_\beta+k_d} \mathbf{w}_{(1)i,f_s}^T\left(\tilde{\mathbf{V}}_{s(1)i*}^{-1}\mathbf{S}_{s(1)i}^{g_s}\, \tilde{\mathbf{V}}_{s(1)i*}^{-1}\right)\mathbf{w}_{(1)i,g_s}\big(\beta_{f_s}$$

$$- \tilde{\beta}_{f_s}\big)\left(d_{g_s} - \tilde{d}_{g_s}\right)$$

$$+ \sum_{i=1}^{N}\sum_{f_s=1}^{k_\beta}\sum_{g_s=k_\beta+k_d+1}^{k_\beta+k_d+k_\gamma} \mathbf{w}_{(1)i,f_s}^T\left(\tilde{\mathbf{V}}_{s(1)i*}^{-1}\mathbf{T}_{s(1)i}^{g_s}\, \tilde{\mathbf{V}}_{s(1)i*}^{-1}\right)\mathbf{w}_{(1)i,g_s}\big(\beta_{f_s}$$

$$- \tilde{\beta}_{f_s}\big)\left(\gamma_{g_s} - \tilde{\gamma}_{g_s}\right)$$

$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{f_s=k_\beta+1}^{k_\beta+k_d}\sum_{g_s=k_\beta+k_d+1}^{k_\beta+k_d+k_\gamma} \mathbf{w}_{(1)i,e_s}^T \frac{\partial^2 \tilde{\mathbf{V}}_{s(1)i*}^{-1}}{\partial d_{f_s}\,\partial d_{g_s}}\left(\boldsymbol{y}_{is} - \mathbf{w}_{(1)}\boldsymbol{\beta}_{s(1)*}\right)\big(d_{f_s}$$

$$- \tilde{d}_{f_s}\big)\left(\gamma_{g_s} - \tilde{\gamma}_{g_s}\right)$$

$$+ \frac{1}{2}\sum_{i=1}^{N}\sum_{f_s=k_\beta+k_d+1}^{k_\beta+k_d+k_\gamma}\sum_{g_s=k_\beta+k_d+1}^{k_\beta+k_d+k_\gamma} \mathbf{w}_{(1)i,e_s}^T \frac{\partial^2 \tilde{\mathbf{V}}_{s(1)i*}^{-1}}{\partial \gamma_{f_s}\,\partial \gamma_{g_s}}\big(\boldsymbol{y}_{is}$$

$$- \mathbf{w}_{i(1)}\boldsymbol{\beta}_{s(1)*}\big)\left(\gamma_{f_s} - \tilde{\gamma}_{f_s}\right)\left(\gamma_{g_s} - \tilde{\gamma}_{g_s}\right)$$

$$+ \sum_{i=1}^{N}\sum_{f_s=k_\beta+1}^{k_\beta+k_d}\sum_{g_s=k_\beta+k_d+1}^{k_\beta+k_d+k_\gamma} \mathbf{w}_{(1)i,e_s}^T \frac{\partial^2 \tilde{\mathbf{V}}_{s(1)i*}^{-1}}{\partial d_{f_s}\,\partial \gamma_{g_s}}\left(\boldsymbol{y}_{is} - \mathbf{w}_{i(1)}\boldsymbol{\beta}_{s(1)*}\right)\big(d_{f_s}$$

$$- \tilde{d}_{f_s}\big)\left(\gamma_{g_s} - \tilde{\gamma}_{g_s}\right)\Bigg) - \frac{\partial}{\partial \beta_{e_s(1)}}\left(\lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2}\right),$$

where $\|\boldsymbol{\phi}^* - \tilde{\boldsymbol{\phi}}\| \le \|\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}\|$. Since $\|\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}\| \le \frac{C}{\sqrt{N}}$,

$$\frac{1}{\sqrt{N}}\frac{\partial}{\partial \beta_{e_s(1)}} Q\left\{\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix}\right\} = \boldsymbol{o}_p(1) - \frac{1}{\sqrt{N}}\frac{\partial}{\partial \beta_{e_s(1)}}\left(\lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2}\right).$$

Considering the last term,

$$\frac{1}{\sqrt{N}}\frac{\partial}{\partial \beta_{e_s(1)}}\left(\lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2}\right) = \begin{cases} \lambda_{N1}\dfrac{sign(\beta_{pe_s})}{\sqrt{N}|\tilde{\beta}_{pe_s}|} & if \sum_{s'\neq e_s}\left(\dfrac{\beta_{ps'}}{\tilde{\beta}_{ps'}}\right)^2 = 0 \\ \lambda_{N1}\dfrac{\frac{\beta_{pe_s}}{\tilde{\beta}_{pe_s}^{2}}}{\sqrt{N}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps}}{\tilde{\beta}_{ps}}\right)^2}} & if \sum_{s'\neq e_s}\left(\dfrac{\beta_{ps'}}{\tilde{\beta}_{ps'}}\right)^2 \neq 0 \end{cases}.$$

In either case, the sign of $\frac{1}{\sqrt{N}}\frac{\partial}{\partial \beta_{e_s(1)}} Q\left\{\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix}\right\}$ is determined by that of $\beta_{e_s(1)}$. Similarly, the sign of $\frac{1}{\sqrt{N}}\frac{\partial}{\partial d_{e_s(1)}} Q\left\{\begin{pmatrix}\boldsymbol{\phi}^1\\ \boldsymbol{\phi}^2\end{pmatrix}\right\}$ is also determined by that of $d_{e_s(1)}$. (B-5) is proved.

Δ

Appendix B-IV: Proof of Theorem 3

From Theorem 1, we proved that there exists a local maximizer $\widehat{\boldsymbol{\phi}} = \begin{pmatrix}\widehat{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}$ of $\widetilde{\boldsymbol{\phi}} = \begin{pmatrix}\widetilde{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}$ such that $\left\|\widehat{\boldsymbol{\phi}}^1 - \widetilde{\boldsymbol{\phi}}^1\right\|_F = \boldsymbol{o}_p(\frac{1}{\sqrt{N}})$ and the local maximizer satisfies the set of penalized likelihood equation

$$\frac{}{\partial \boldsymbol{\phi}^1} Q(\boldsymbol{\phi})|_{\boldsymbol{\phi}=\begin{pmatrix}\widehat{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}} = \frac{\partial}{\partial \boldsymbol{\phi}^1}\left\{\sum_{s=1}^{\mathbb{S}} l(\boldsymbol{\phi}_s)\right\}|_{\boldsymbol{\phi}=\begin{pmatrix}\widehat{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}} - \frac{\partial}{\partial \boldsymbol{\phi}^1}\left\{\lambda_{N1}\sum_{p=1}^{\mathbb{P}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{\beta_{ps(1)}}{\tilde{\beta}_{ps(1)}}\right)^2} + \right.$$

$$\left.\lambda_{N2}\sum_{q=1}^{\mathbb{Q}}\sqrt{\sum_{s=1}^{\mathbb{S}}\left(\frac{d_{qs(1)}}{\tilde{d}_{qs(1)}}\right)^2}\right\}|_{\boldsymbol{\phi}=\begin{pmatrix}\widehat{\boldsymbol{\phi}}^1\\ \mathbf{0}\end{pmatrix}} = 0\ .$$

Using the Taylor series expansion and multiplying through by $\frac{1}{N}$, we have

$$\frac{1}{N}\left(\nabla l\left(\tilde{\boldsymbol{\phi}}_{11}\right)^T, \cdots, \nabla l\left(\tilde{\boldsymbol{\phi}}_{S1}\right)^T\right)^T - \left(\left(I(\tilde{\boldsymbol{\phi}}_{11})(\hat{\boldsymbol{\phi}}_{11} - \tilde{\boldsymbol{\phi}}_{11})\right)^T, \cdots, \left(I(\tilde{\boldsymbol{\phi}}_{S1})(\hat{\boldsymbol{\phi}}_{S1} - \tilde{\boldsymbol{\phi}}_{S1})\right)^T\right)^T - (\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_{\mathbb{S}}^T)^T = 0$$

$$\sqrt{N}\left(\left(I(\tilde{\boldsymbol{\phi}}_{11})(\hat{\boldsymbol{\phi}}_{11} - \tilde{\boldsymbol{\phi}}_{11})\right)^T, \cdots, \left(I\left(\tilde{\boldsymbol{\phi}}_{S1}\right)\left(\hat{\boldsymbol{\phi}}_{S1} - \tilde{\boldsymbol{\phi}}_{S1}\right)\right)^T + (\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_{\mathbb{S}}^T)\right)^T =$$

$$\sqrt{N}\{I\left(\tilde{\boldsymbol{\phi}}^1\right)\left(\hat{\boldsymbol{\phi}}^1 - \tilde{\boldsymbol{\phi}}^1\right) + (\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_{\mathbb{S}}^T)^T\} = \frac{1}{\sqrt{N}}\left(\nabla l\left(\tilde{\boldsymbol{\phi}}_{11}\right)^T, \cdots, \nabla l\left(\tilde{\boldsymbol{\phi}}_{S1}\right)^T\right)^T$$

Because of the proof of Theorem 1, it follows from the multivariate central limit theorem that

$\frac{1}{\sqrt{N}}\left(\nabla l\left(\tilde{\boldsymbol{\phi}}_{11}\right)^T, \cdots, \nabla l\left(\tilde{\boldsymbol{\phi}}_{S1}\right)^T\right)^T \longrightarrow_d N\left(0, I(\tilde{\boldsymbol{\phi}}^1)\right)$, where $I\left(\tilde{\boldsymbol{\phi}}^1\right) = diag(I\left(\tilde{\boldsymbol{\phi}}_{11}\right), \cdots, I\left(\tilde{\boldsymbol{\phi}}_{S1}\right))$.

Therefore,

$$\sqrt{N}\{I\left(\tilde{\boldsymbol{\phi}}^1\right)\left(\hat{\boldsymbol{\phi}}^1 - \tilde{\boldsymbol{\phi}}^1\right) + (\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_{\mathbb{S}}^T)^T\} \longrightarrow_d N\left(0, I(\tilde{\boldsymbol{\phi}}^1)\right)$$

which can be written as

$$\sqrt{N}\, I(\tilde{\boldsymbol{\phi}}^1)\left(\left(\hat{\boldsymbol{\phi}}^1 - \tilde{\boldsymbol{\phi}}^1\right) + I(\tilde{\boldsymbol{\phi}}^1)^{-1}(\boldsymbol{v}_1^T, \cdots, \boldsymbol{v}_{\mathbb{S}}^T)^T\right) \longrightarrow_d N\left(0, I(\tilde{\boldsymbol{\phi}}^1)\right). \qquad \Delta$$

Appendix B-V: Definitions of "organizing", "checking", "mobilizing", "exchanging", "assisting", and "backfilling" in the NCCI

"Organizing" is creating a structure that allows care coordination to be carried out in a safe and timely way. "Checking" is evaluating accuracy, timeliness, and completion of steps required in the sequence to carry out care coordination processes. "Mobilizing" is directly and indirectly getting others take actions for which they are accountable and are required to carry out care coordination processes. "Exchanging" is giving and receiving information needed to carry out care coordination processes. “Assisting” is getting or giving help to carry out one or more steps in care coordination process that a nurse would ordinarily do themselves. “Backfilling” is doing the work of other members of the care team for which they were responsible but did not do to carry out care coordination processes.

Appendix B-VI: The derivation for obtaining (3.7), (3.8), and (3.9)

$\mathrm{p}\left(\tilde{\boldsymbol{e}}_s \middle| \boldsymbol{y}_s, \boldsymbol{\phi}_s^{(\omega)}\right) = \frac{\mathrm{p}\left(\boldsymbol{y}_s, \tilde{\boldsymbol{e}}_s \middle| \boldsymbol{\phi}_s^{(\omega)}\right)}{\mathrm{p}\left(\boldsymbol{y}_s \middle| \boldsymbol{\phi}_s^{(\omega)}\right)}$. Taking logarithm for both sides,

$$log\,\mathrm{p}\left(\tilde{\boldsymbol{e}}_s \middle| \boldsymbol{y}_s; \boldsymbol{\phi}_s^{(\omega)}\right) = log\,\mathrm{p}\left(\boldsymbol{y}_s, \tilde{\boldsymbol{e}}_s \middle| \boldsymbol{\phi}_s^{(\omega)}\right) - log\,\mathrm{p}\left(\boldsymbol{y}_s \middle| \boldsymbol{\phi}_s^{(\omega)}\right). \quad \text{(B-7)}$$

According to (3.5),

$$log\,\mathrm{p}\left(\boldsymbol{y}_s, \tilde{\boldsymbol{e}}_s \middle| \boldsymbol{\phi}_s^{(\omega)}\right) = -\frac{\sum_{i=1}^{N} \mathrm{n}_i + NQ}{2} log(2\pi) - \frac{\sum_{i=1}^{N} \mathrm{n}_i + NQ}{2} \log\left(\sigma_s^{2^{(\omega)}}\right) - \frac{1}{2\sigma_s^{2^{(\omega)}}}\left(\left\|\boldsymbol{y}_s - \mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\widetilde{\boldsymbol{\Gamma}}_s^{(\omega)}\tilde{\boldsymbol{e}}_s - \mathbf{W}\boldsymbol{\beta}_{\boldsymbol{s}}^{(\omega)}\right\|^2 + \tilde{\boldsymbol{e}}_s^T\tilde{\boldsymbol{e}}_s\right). \quad \text{(B-8)}$$

Furthermore, we can get:

$$og\,\mathrm{p}\left(\boldsymbol{y}_s \middle| \boldsymbol{\phi}_s^{(\omega)}\right) = -\frac{\sum_{i=1}^{N} \mathrm{n}_i}{2} \log(2\pi) - \frac{1}{2}\log\left(\left|\widetilde{\mathbf{V}}_s\right|\right) - \frac{1}{2}\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right)^T \widetilde{\mathbf{V}}_s^{-1}\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right), \quad \text{(B-9)}$$

where $\widetilde{\mathbf{V}}_s = Diag(\mathbf{V}_{\mathrm{s}1}, \dots, \mathbf{V}_{\mathrm{sN}})$ and $\mathbf{V}_i = \sigma_s^{2^{(\omega)}}\left(\mathbf{Z}_i\ \mathbf{D}_s^{(\omega)}\boldsymbol{\Gamma}_s^{(\omega)}\boldsymbol{\Gamma}_s^{T(\omega)}\mathbf{D}_s^{(\omega)}\mathbf{Z}_i^T + \mathbf{I}\right)$, $i = 1, \dots, N$.

Inserting (B-8) and (B-9) into (B-7), we can get

$log\,\mathrm{p}\left(\tilde{\boldsymbol{e}}_s \middle| \boldsymbol{y}_s; \boldsymbol{\phi}_s^{(\omega)}\right)$

$$= -\frac{\sum_{i=1}^{N} \mathrm{n}_i + NQ}{2}\log(2\pi) - \frac{\sum_{i=1}^{N} \mathrm{n}_i + NQ}{2}\log\left(\sigma_s^{2^{(\omega)}}\right) - \frac{1}{2\sigma_s^{2^{(\omega)}}}\left(\left\|\boldsymbol{y}_s - \mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\tilde{\mathbf{\Gamma}}_s^{(\omega)}\tilde{\boldsymbol{e}}_s - \right.\right.$$

$$\left.\left.\mathbf{W}\boldsymbol{\beta}_{\boldsymbol{s}}^{(\omega)}\right\|^2 + \tilde{\boldsymbol{e}}_s^T\tilde{\boldsymbol{e}}_s\right) + \frac{\sum_{i=1}^{N} \mathrm{n}_i}{2}\log(2\pi) + \frac{1}{2}\log\left(\left|\widetilde{\mathbf{V}}_s\right|\right) + \frac{1}{2}\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right)^T\widetilde{\mathbf{V}}_s^{-1}\left(\boldsymbol{y}_s - \right.$$

$$\left.\mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right)$$

$$= -\frac{NQ}{2}\log(2\pi) - \frac{1}{2}\log\left(\left|\mathbf{U}_s^{(\omega)}\right|\right) - \frac{1}{2}(\tilde{\boldsymbol{e}}_s - \hat{\boldsymbol{e}}_s^{(\omega)})^T\mathbf{U}_s^{(\omega)^{-1}}(\tilde{\boldsymbol{e}}_s - \hat{\boldsymbol{e}}_s^{(\omega)}), \qquad \text{(B-10)}$$

where $\hat{\boldsymbol{e}}_s^{(\omega)}$ and $\mathbf{U}_s^{(\omega)}$ follow the definition in (7) and (8). The form of (S-10) means that $\tilde{\boldsymbol{e}}_s|\boldsymbol{y}_s, \boldsymbol{\phi}^{(\omega)} \sim N(\hat{\boldsymbol{e}}_s^{(\omega)}\ ,\ \mathbf{U}_s^{(\omega)})$.

Furthermore, by (3.4),

$$l\left(\boldsymbol{\phi}^{(\omega)}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}\right) = -\frac{1}{2}\sum_{s=1}^{\mathbb{S}}\left\{log\left\lfloor\widetilde{\mathbf{V}}_s^{\ (\omega)}\right\rfloor + \left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_{\boldsymbol{s}}^{(\omega)}\right)^T\widetilde{\mathbf{V}}_s^{\ (\omega)^{-1}}\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_{\boldsymbol{s}}^{(\omega)}\right)\right\}, \qquad \text{(B-11)}$$

where $\widetilde{\mathbf{V}}_s^{\ (\omega)} = Diag\left(\mathbf{V}_{s1}^{\ (\omega)}, \dots, \mathbf{V}_{sN}^{\ (\omega)}\right)$ and $\mathbf{V}_{si}^{\ (\omega)} = \sigma_s^{2^{(\omega)}}\left(\mathbf{Z}_i\ \mathbf{D}_s^{(\omega)}\mathbf{\Gamma}_s^{(\omega)}\mathbf{\Gamma}_s^{T(\omega)}\mathbf{D}_s^{(\omega)}\mathbf{Z}_i^T + \mathbf{I}\right)$, $i = 1, \dots, N$. Taking derivative of (B-11) with respect to $\sigma_s^{2^{(\omega)}}$,

$$\frac{\partial l\left(\boldsymbol{\phi}^{(\omega)}|\{\boldsymbol{y}_s\}_{s=1}^{\mathbb{S}}\right)}{\partial\sigma_s^{2^{(\omega)}}} = -\frac{1}{2}\left\{\frac{\sum_{i=1}^{N} \mathrm{n}_i}{\sigma_s^{2^{(\omega)}}} - \frac{\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right)^T\left(\mathbf{Z}\widetilde{\mathbf{D}}_s^{(\omega)}\tilde{\mathbf{\Gamma}}_s^{(\omega)}\tilde{\mathbf{\Gamma}}_s^{(\omega)^T}\widetilde{\mathbf{D}}_s^{(\omega)^T}\mathbf{Z}^T\right)^{-1}\left(\boldsymbol{y}_s - \mathbf{W}\boldsymbol{\beta}_s^{(\omega)}\right)}{\sigma_s^{4^{(\omega)}}}\right\}. \qquad \text{(B-12)}$$

Making (B-12) equal to zero and solving for $\sigma_s^{2^{(\omega)}}$, we can get (3.9). Δ

APPENDIX C

SUPPLEMETNAL MATERIALS FOR CHAPTER 4

Appendix C-I: Proof of Proposition 2

Proof: Let $= \frac{l_i(\mathbf{Z})-(\beta_{0,i}+\boldsymbol{\beta}_{z,i}^T\mathbf{Z})}{\sigma_i}$, $\delta = \frac{\beta_{y,i}}{\sigma_i}$, $r(x) = \frac{\Phi(x-\delta)}{\Phi(x)}$, and $r_0 = \frac{r_l}{1-r_l} \times \frac{1-\pi(\mathbf{Z})}{\pi(\mathbf{Z})}$. Then, the constraint in (4.8) becomes $r(x) \le r_0$. Here, $\delta > 0$ because $\beta_{y,i}$ represents the increase in the biomarker value as Y changes from 0 (non-diseased) to 1 (diseased). Recall that we made an assumption earlier on that there is a positive correlation between each biomarker and the disease risk, which suggests that $\beta_{y,i} > 0$. Also, $r_0 > 0$ by definition.

Next, we will show that $r(x)$ is strictly monotonically increasing from 0 to 1 as x increases from $-\infty$ to $+\infty$. When $x \to +\infty$, we have

$$\lim_{x\to+\infty} r(x) = \frac{\lim_{x\to+\infty} \Phi(x-\delta)}{\lim_{x\to+\infty} \Phi(x)} = \frac{1}{1} = 1.$$

When $x \to -\infty$, using L'Hospital's Rule, we have

$$\lim_{x\to-\infty} r(x) = \lim_{x\to-\infty} \frac{\Phi(x-\delta)}{\Phi(x)} = \lim_{x\to-\infty} \frac{\int_{-\infty}^{x-\delta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}{\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt} = \lim_{x\to-\infty} \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\delta)^2}{2}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} =$$

$$\lim_{x\to-\infty} e^{\delta x - \frac{\delta^2}{2}} = 0.$$

For finite x, $r(x)$ is strictly monotonically increasing because

$$\frac{d\,r(x)}{dx} = d\left(\frac{\int_{-\infty}^{x-\delta} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}{\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt}\right) \Big/ dx = \frac{e^{-\frac{(x-\delta)^2}{2}} \times \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt - e^{-\frac{x^2}{2}} \times \int_{-\infty}^{x-\delta} e^{-\frac{t^2}{2}} dt}{\left(\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt\right)^2}$$

$$= \frac{e^{-\frac{x^2}{2}} \times \int_{x-\delta}^{x} e^{-\frac{t^2}{2}} dt}{\left(\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt\right)^2} + \frac{e^{\left(-\frac{\delta^2}{2}+x\delta\right)} \times \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt}{\left(\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt\right)^2} > 0.$$

Hence, when $0 < r_0 < 1$, the feasible region of x is $(-\infty, x_0]$, where x_0 satisfies $r(x_0) = r_0$. Because $r(x)$ strictly monotonically increases with respect to x, the maximum

$r(x)$ is achieved at x_0 and this solution is unique. When $r_0 \geq 1$, the feasible region of x is $[-\infty, +\infty]$. The maximum $r(x)$ is achieved at $+\infty$ and this solution is unique. Δ