INDIVIDUALIZED LEARNING AND INTEGRATION FOR MULTI-MODALITY DATA

BY

XIWEI TANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Annie Qu, Chair
Professor Xiaofeng Shao
Professor Douglas Simpson
Assistant Professor Ruoqing Zhu

Abstract

Individualized modeling and multi-modality data integration have experienced an explosive growth in recent years, which have many important applications in biomedical research, personalized education and marketing. Conventional statistical models usually fail to capture significant variation due to subject-specific effects and heterogeneity of data from multiple sources. Consequently, it has become very critical to incorporate individuals' and modalities' heterogeneous characteristics in order to efficiently explore the data structure and enhance the prediction power. In this thesis, we address three challenging issues: mixture modeling for longitudinal data, individualized variable selection and multi-modality tensor learning with an application in medical imaging analysis.

In the first part of the thesis, we develop a model-based subgrouping method for longitudinal data. Specifically, we propose an unbiased estimating equation approach for a two-component mixture model with correlated response data. In contrast to most existing longitudinal data clustering methods, the proposed model allows subgroup membership change for each individual over time. Furthermore, we incorporate correlation structure on unobservable latent indicator variables. Another advantage our approach is that we do not require any information about joint likelihood function for each subject. The proposed model is shown to have more efficient parameter estimators in both mixing proportions and component densities. In addition, by utilizing within-subject serial correlations, the proposed approach enhances classification power compared to existing methods, especially for those boundary observations.

In the second part of the thesis, we propose an individualized variable selection approach to select different relevant variables for different individuals. The conventional homogeneous model, which assumes all subjects share the same effects of certain predictors, may wash out important information due to heterogeneous variation. For example, in personalized medicine, some individuals could have positive responses to the treatment while some individuals could have negative ones. Hence the population average effect could be close to zero. In this thesis, we construct a

separation penalty with multi-directional shrinkages including zero, which facilitates individualized modeling to distinguish strong signals from noisy ones. As a byproduct, the proposed model identifies subgroups among which individuals share similar effects, and thus improves estimation efficiency and personalized prediction accuracy. Finite sample simulation studies and an application to HIV longitudinal data demonstrate the model efficiency and the prediction power of the new approach compared to a variety of existing penalization models.

In the third part of the thesis, we are interested in employing medical imaging data for diagnosis. This work is motivated by breast cancer imaging data produced by a multimodality multiphoton optical imaging technique. We develop an innovative multilayer tensor learning method to predict disease status effectively through utilizing subject-wise imaging information. In particular, we propose an individualized multilayer model which leverages an additional layer of individual structure of imaging shared by multiple modalities in addition to employing a high-order tensor decomposition shared by populations. One major advantage of our approach is that we are able to capture the spatial information of microvesicles observed in certain modalities of optical imaging through integrating multimodality imaging data. Our simulation studies and real data analysis both indicate that the proposed multilayer learning method improves prediction accuracy significantly compared to existing competitive statistical and machine learning methods.

To my family.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my advisor Professor Annie Qu for her guidance on all the aspects of my Ph.D life. Her patience, enthusiasm, passion and immense knowledge builds the foundation of my research career and broadens my horizon in statistics and data science. In addition to detailed technical skills, I learned much more from her philosophy in always seizing the most important thing, which also applies to life. Facing all kinds of difficulties over the past five years, I would never be so confident and smooth without her countless inspirations and mentorship.

Special thanks to my thesis committee members: Prof. Douglas Simpson, Prof. Xiaofeng Shao and Prof. Ruoqing Zhu for their generous help and constant support in my career development. Their insightful comments and valuable suggestions have greatly facilitated my research. Furthermore, I owe my appreciation to Xuan Bi and Christopher Kinson for being wonderful collaborators. Besides, my thanks also goes to Professor Peng Wang for his help discussions. In addition, I would sincerely express my appreciation to Christopher Vecoli, who has generously give his time and expertise in scientific writing to better my work.

My time at Champaign and Urbana has become a most important part of my life for having so many memorable friends here. My thanks go to all the faculty and staff members in the Statistics Department for making the Illini Hall as a big family. I also want to thank the fellow students from the department, especially Xiaolu Zhu, Peibei Shi, Xuan Bi, Fei Xue, Xichen Huang, Weihong Huang, Yunbo Ouyang and Shun Yao. You are the brightest stars in the my sky.

Finally, I want to thank my family: my grandparents, my parents, my aunt, uncle and cousin, little panda and my fiancée for their constant accompany, encouragement, and support. My grandparents and parents are always behind me with their unconditional love raising me up. And last of all for my loving fiancée Xiaoxiao, I am so grateful to have you in my life, which brings me the faith that I would never walk alone.

Contents

Chapter 1

Introduction

There has been a growing demand to develop effective and efficient methods to capture and utilize data heterogeneity from specific individuals, subgroups of subjects, or multiple data sources. For example, personalized medicine requires to identify different treatment effect groups, which enables us to assign a more efficient treatment to each specific patient. In addition, in biomedical imaging analysis, multiple imaging techniques, such as CT scan, MRI, fMRI and optical imaging, are usually applied together for diagnosing disease status. To achieve a better diagnosis power, it is very crucial to effectively integrate information from different modalities of imaging data. In this thesis, we propose methods and theory for individualized modeling and integration of multi-modality data. Our contributions are mainly from three perspectives: longitudinal mixture modeling, individualized feature selection, and tensor learning for multi-modality imaging data.

1.1 Longitudinal Mixture Modeling

Mixture modeling is a major technique to model the subgroup structure, which draws more and more attention recently. A direct application of mixture modeling is on clustering. Compared to other clustering methods, e.g., K-means, mixture modeling provides a soft prediction on subgroup membership, which is more informative. In the past two decades, many mixture modeling tools have been developed to incorporate covariates information in addition to outcomes. A mixture model could be viewed as a hieratical structure consisting of a subgroup membership indicator and component outcomes given the indicator. However, the indicator variable is unable to be observed directly and could be only inferred from the outcomes and the covariates. Therefore the mixture modeling is also treated as an incomplete-data modeling. This unique challenge of the mixture

modeling prevents its extension to more complicated data structure, e.g., longitudinal data.

Longitudinal data has well known within subject correlation information, which is very important. It is very challenging to incorporate the correlation information in the mixture modeling while allowing time-varying subgroup membership, especially in latent subgroup membership's level. The conventional parametric mixture modeling would encounter difficulties since the full joint distribution of categorical latent variables is far more than complicated.

In Chapter 2, we propose an unbiased estimating equation approach for a two-component mixture model with correlated response data. We adapt the mixture-of-experts model and a generalized linear model for component distribution and mixing proportion, respectively. The new approach only requires marginal distributions of both component densities and latent variables. We utilize serial correlations from subjects' subgroup memberships, which improves estimation efficiency and classification accuracy, and show that estimation consistency does not depend on the choice of the working correlation matrix. The proposed estimating equation is solved by an Expectation-Solving estimating equation (ES) algorithm. In the E-step of the ES algorithm, we propose a joint imputation based on the conditional linear property for the multivariate Bernoulli distribution. In addition, we establish asymptotic properties for the proposed estimators and the convergence property using the ES algorithm. Our method is compared to an existing competitive clustering approach in both simulation studies and 2008 election data application.

1.2 Individualized Feature Selection

In recent years, the arise of precision medicine and wide-spread electronic health record data motivate us to develop a more effective personalized treatment. This has widely applications in personalized medicine, personalized education program and personalized marketing. Consequently, the increasing demand of personalized prediction requires personalized modeling. The traditional one-model-fits-the-whole-population may not have power to detect some important predictors for subgroups of interest. For example, different individuals may have different prognostic factors

associated with the same disease.

In Chapter 3, We propose a novel individualized variable selection method which performs coefficient estimation, subgroup identification and variable selection simultaneously. In contrast to traditional feature selection approaches, an individualized regression model allows different individuals to have different relevant variables. The key component of the new approach is to construct a multi-directional separation penalty which shrinks weak signals to zero and aggregates strong signals to a subgroup-shared homogeneous effect. This allows us to borrow information from subjects within the same subgroup, and therefore improve the estimation efficiency and variable selection accuracy for each individual. Another advantage of the proposed model is that it can incorporate within-subject correlation for longitudinal data.

We provide a general theoretical foundation under a double-divergence modeling framework where the number of subjects and the number of repeated measurements both go to infinity, and therefore involves high-dimensional individual parameters. In addition, we present the oracle property for the proposed estimator to ensure its optimal large sample property. Simulation studies and an application to HIV longitudinal data are illustrated to compare the new approach to existing penalization methods.

1.3 Tensor Learning for Imaging Data Analysis

Imaging analysis has drawn great attention and encounters an explosive growth, due to wide applications in medical images for diagnosing, especially in neuroimaging and cancer radiotherapy. It is in great demand to develop efficient statistical tools to utilize the image information to predict interested outcomes, for example, disease status and treatment responses. However, the imaging data has many challenges to be fitted into a traditional statistical model, including high-dimensional data structure, heterogeneous background noise and multiple data modalities.

Chapter 4 is motivated by multiphoton optical imaging data for breast cancer diagnosis produced by Boppart Lab at University of Illinois at Urbana-Champaign, where there are four imag-

ing modalities at each subject capturing different microenvironments. In Chapter 4, we treat the image data along with additional information (e.g., time) as a higher-order multi-dimensional array, which is called a tensor as well. We propose a multi-layer tensor learning model to efficiently extract image's information and then use the extracted information to fit a regression model associated to the interested outcomes. Specifically, we construct a low-rank decomposition for the image tensor, which consists of the individualized layers, which capture the subject-specific information over each individual's multiple modalities, and the population-shared layers, which model the modality-specific background and achieve an effective dimension reduction. Then the both layers' information are incorporated to predict the outcome responses.

Due to individualized signals, traditional homogeneous dimension reduction methods could loss their power in capturing those images' structures under a conventional low-rank model framework. By decomposing a tensor into different specific layers, the proposed method is capable of capturing the unique spatial information from the tensor structure and reduces the complex data's dimensionality efficiently. Numerical studies illustrate the power of the proposed method, especially when signal regions vary a lot among different images, which often occurs in breast cancer diagnosis. The proposed approach is also applied to four-modality optical imaging data and achieves a significantly better prediction power on breast cancer diagnosis.

Chapter 2

Mixture Modeling for Longitudinal Data

2.1 Introduction

The mixture model has been extensively applied in many fields due to its flexibility to capture the heterogeneity arising from subgroups (components) in the whole population. The mixture model can also be viewed as a two-level hierarchical structure with incomplete data, where the first level consists of latent variables indicating subjects' subgroup memberships, and the second level consists of outcome variables. However, the individual's subgroup membership has to be inferred from the location of the outcome response due to unobservable latent variables.

Existing methods of mixture model with covariates for independent data include, but are not limited to, [30]'s mixture-of-experts for a mixture of component regressions, [32]' extension on the generalized linear model for mixing proportion. For correlated or longitudinal outcome data, [70] introduce a linear random-effects model for components' densities under the mixture framework. Alternatively, [67] propose a generalized estimating equation for component distributions to incorporate correlations. These approaches all assume that the latent variables are independent. However, modeling correlation from outcome variables only is not sufficient to address correlations from subgroup membership over time.

One notable approach to model the correlated latent structure is the hidden Markov chain model ([62]). However, this is not applicable for longitudinal data since the Markov chain assumption dose not hold or approximate some common correlation structure in longitudinal data such as exchangeable structure. Another well-known approach of mixture modeling for longitudinal data is related to growth-curve mixture modeling [53], where the subject's group membership is fixed

over time, while different trajectory classes represent different mixture components. [29] propose a kernel smoothing method for a mixture of Gaussian processes incorporating both functional and heterogeneous types of dense longitudinal data. In addition, to incorporate the individual effects, [80] propose a multivariate Bernoulli mixture model by utilizing random effects in the generalized linear model for mixing proportion.

Although the mixed-effects model is widely used to handle dependent data, incorporating serial correlation for latent variables in mixture modeling for longitudinal data is still limited. In this chapter, we are interested in developing an efficient method in mixture modeling of longitudinal data where within-subject subgroup memberships could be correlated.

One challenge in formulating a mixture model for dependent data is that the joint likelihood function usually does not have an explicit form, because the latent variables are correlated categorical variables. In addition, the latent indicator variables are unobservable and require imputation. Although one can assume independent structure for the latent variables, the estimation efficiency will be compromised.

In this chapter, we allow group memberships to change over time in addition to taking serial correlation into account. We adopt the generalized estimating equation (GEE) approach ([44]) for both component distributions and mixing proportion in a two-component mixture model. This can be accomplished by treating the estimating equations for incomplete data as the conditional expectations of those for complete data. Specifically, we apply the Expectation-Estimating-Equation (EEE) algorithm to solve the equations. To impute the latent variable in the E-step, we provide an approximation method to calculate the joint conditional expectation utilizing the conditional linear property for the correlated binary variables ([61]; [63]).

In contrast to the joint likelihood approach, the proposed method only requires the marginal distributions for both components' densities and latent variables. In the estimating step, we fully utilize the serial correlation while treating it as a working structure. Allowing different working correlation structures enables one to incorporate various correlated latent structures, although the proposed method does not require to know the true correlation structure in order to produce con-

sistent estimators of mean parameters. However, if the correlation structure is correctly or closely specified, we gain efficiency of the parameter's estimation and improve the classification accuracy from a model-based clustering.

The rest of this chapter is organized as follows: Section 2.2 introduces some notation and background knowledge; Section 2.3 presents our method and provides some theoretical results, as well as the EEE algorithm and imputation methods; Section 2.4 provides simulation studies; Section 2.5 considers an application to the Election data; and Section 2.6 offers a brief summary and some further discussion.

2.2 Background and Notation

For longitudinal data, let y_{it} be the response of subject i at time t, and $\boldsymbol{x}_{it}$ be a p-dimensional covariate vector, where $i = 1, ..., n$ and $t = 1, ..., T_i$. For ease of notation, we first assume $T_i = T$ for all i representing a balanced data case, where an unbalanced data case will be discussed later. Denote $\boldsymbol{y}_i = (y_{i1}, ..., y_{iT})$ as a $T \times 1$ response vector, $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT})$ as a $p \times T$ covariate matrix. In a two-component mixture model, let z_{it} denote the binary latent variable associated with y_{it}. Let $\mu_r(\cdot)$ be an inverse link function satisfying $\mathbf{E}[y_{it}|\boldsymbol{x}_{it}, z_{it} = 2 - r] = \mu_r(\boldsymbol{\beta}_r'\boldsymbol{x}_{it})$ $(r = 1, 2)$, where $\boldsymbol{\beta}_r$ is a p-dimensional parameter, and $\boldsymbol{\mu}_{ri} = \mathbf{E}[\boldsymbol{y}_i|\boldsymbol{x_i}, \boldsymbol{z}_i = \mathbf{2} - \boldsymbol{r}]$.

In this chapter, we consider a two-component mixture model. The outcome variable y_{it} is from either one of the two subgroup populations and thus is assumed to follow a mixture distribution $\pi_{it} f_1(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_1}, \phi_1) + (1 - \pi_{it}) f_2(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_2}, \phi_2)$, where π_{it} is a mixing proportion defined as the probability of a response y_{it} from a specified subgroup, $f_1(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_1}, \phi_1)$ and $f_2(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_2}, \phi_2)$ are the components' distributions, and $\boldsymbol{\phi} = (\phi_1, \phi_2)$ is the dispersion parameter. This two-component mixture model can be regarded as a hierarchical structure. Specifically,

$$
\begin{aligned}
&z_{it} \sim Bernoulli(\pi_{it}),\\
&y_{it}|(z_{it} = 1) \sim f_1(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_1}, \phi_1) \quad and \quad y_{it}|(z_{it} = 0) \sim f_2(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_2}, \phi_2),
\end{aligned}
$$

where a latent variable z_{it} represents a subgroup membership indicator.

The most common assumption for this hierarchical model is that the outcome variable y_{it}'s are conditionally independent given the subgroup membership label z_{it}, then the joint likelihood density function of complete data $(\boldsymbol{y_i}, \boldsymbol{z_i})$ can be written as

$$\begin{aligned} f_c(\boldsymbol{y_i}, \boldsymbol{z_i}) &= f(\boldsymbol{y_i}|\boldsymbol{z_i}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\phi}, \boldsymbol{x_i}) p(\boldsymbol{z_i}|\boldsymbol{\pi_i}, \boldsymbol{x_i}) \\ &= \Big(\prod_{t=1}^{T} f_1(y_{it}|\boldsymbol{\beta_1}, \phi_1, \boldsymbol{x_{it}})^{z_{it}} f_2(y_{it}|\boldsymbol{\beta_2}, \phi_2, \boldsymbol{x_{it}})^{1-z_{it}} \Big) p(\boldsymbol{z_i}|\boldsymbol{\pi_i}, \boldsymbol{x_i}). \end{aligned} \tag{2.1}$$

Under the mixture framework, the latent variable z_{it} is missing. The simplest way to get around the latent structure is to assume independence of subgroup memberships at different times within each subject and thus $p(\boldsymbol{z_i}|\boldsymbol{\pi_i}) = \prod_{t=1}^{T} \pi_{it}^{z_{it}} (1-\pi_{it})^{1-z_{it}}$. In the independent mixture model, the correlations among longitudinal observations and their subgroup memberships are not fully utilized.

[32] propose a hierarchical mixture-of-experts model which takes covariates into consideration for both latent variable and component distributions, where the latent variable z_{it} is assumed to follow a logistic model: $\pi_{it} = exp(\boldsymbol{\eta' x_{it}})/(1 + exp(\boldsymbol{\eta' x_{it}}))$, and components' densities have a mean regression form. Here we denote $\boldsymbol{\theta} = (\boldsymbol{\eta}', \boldsymbol{\beta_1}', \boldsymbol{\beta_2}')'$ as a grand mean parameter vector, then the log of joint likelihood for complete data is:

$$\begin{aligned} L_c(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \sum_{t=1}^{T} \Big(& z_{it}\{log\pi_{it} + log f_1(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_1}, \phi_1)\} + (1-z_{it})\{log(1-\pi_{it}) \\ & + log f_2(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_2}, \phi_2)\} \Big), \end{aligned} \tag{2.2}$$

under the independence assumption within subject.

For classical mixture models where the latent variable z_{it} is missing, the maximum likelihood estimator (MLE) is typically computed through the Expectation-Maximization (EM) algorithm ([12]). The EM algorithm proceeds iteratively in two steps. In the E-step, we take the conditional expectation of complete-data log likelihood given observed outcome response $\boldsymbol{y}$ and current parameters' estimates $(\boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)})$. Since the complete-data log likelihood $L_c(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})$ in (2.2) is a linear function of latent variable z_{it}, the E-step only requires us to impute z_{it} by its conditional ex-

pectation $w_{it}^{(k)} = \boldsymbol{E}[z_{it}|y_{it}, \boldsymbol{x_{it}}, \boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)}]$ denoted as the mixing weight. In the M-step, we obtain the $k+1$th updates of parameters' estimations by maximizing

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)}) = \sum_{i=1}^{n}\sum_{t=1}^{T}\Bigg(w_{it}^{(k)}\{log\pi_{it} + logf_1(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_1}, \phi_1)\} + (1-w_{it}^{(k)})\{log(1-\pi_{it}) + logf_2(y_{it}|\boldsymbol{x_{it}}, \boldsymbol{\beta_2}, \phi_2)\}\Bigg). \tag{2.3}$$

It has been shown that the obtained series $(\boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)})$ in the EM algorithm converges to the MLE estimator of the mixture distribution $f(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \int_{\boldsymbol{z}} f_c(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\phi})d\boldsymbol{z}$ ([12]).

2.3 Unbiased Estimating Equations for Mixture Modeling

2.3.1 Unbiased estimating equations

Due to the correlated nature of longitudinal data, it is important to incorporate correlation for mixture modeling as it provides more efficient estimation and benefits the longitudinal subgrouping. In this section, we first propose an efficient unbiased estimating equation for complete data, and then extend it to the incomplete-data mixture model.

To account for within-subject correlation among group memberships in (2.1), we have to deal with the joint likelihood function $p(\boldsymbol{z_i}|\boldsymbol{\pi_i})$ of the latent multivariate Bernoulli variable $\boldsymbol{z_i}$. Unlike the multivariate Gaussian distribution, there is no explicit form of likelihood function for the multivariate Bernoulli distribution. The unknown membership in the mixture models makes the incomplete-data likelihood function even more complicated since the latent $\boldsymbol{z_i}$ needs to be integrated out from the complete-data likelihood function. One possible approximation is to apply the Bahardur representation ([1]) for the multivariate Bernoulli distribution by ignoring high-order moments. However, there are several drawbacks such that the correlations are constrained through marginal means and covariates in a complicated way ([15]). In addition, the dimension of the correlation parameters could be very high when the repeated measurement size T is large, which leads to increasing computational demands.

To address this problem, we first introduce an unbiased estimating equation approach and establish it for the complete-data model. The proposed unbiased estimating equation regrading the mean parameter $\boldsymbol{\theta}$ is formulated through the log-likelihood function L_c in (2.2). Maximizing L_c with respect to $\boldsymbol{\theta}$ by solving the corresponding score equation is equivalent to solving the quasi-likelihood equation ([89]; [51]) as follows:

$$S^c(\boldsymbol{\theta}) = \sum_{i=1}^{n} S_i^c(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi_i}}{\partial \boldsymbol{\eta}}\right)^T \boldsymbol{V}_i^{-1}(\boldsymbol{z}_i - \boldsymbol{\pi}_i) \\ \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu_{1i}}}{\partial \boldsymbol{\beta_1}}\right)^T \boldsymbol{U}_{1i}{}^{-1} \boldsymbol{Z}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_{1i}) \\ \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu_{2i}}}{\partial \boldsymbol{\beta_2}}\right)^T \boldsymbol{U}_{2i}{}^{-1}(1 - \boldsymbol{Z}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i}) \end{pmatrix} = 0, \tag{2.4}$$

where $\boldsymbol{\mu_{ri}} = (\mu_{ri1}, ..., \mu_{riT})'$, $\mu_{rit} = \mathbf{E}[y_{it}|z_{it} = 2 - r] = \mu_r(\boldsymbol{\beta_r'} \boldsymbol{x_{it}})$ $(r = 1, 2)$ are the corresponding mean functions for components' densities, $\boldsymbol{Z_i} = \text{diag}(z_{i1}, ..., z_{iT})$ is a diagonal matrix of corresponding latent labels, and $\boldsymbol{U}_{1i}$ and $\boldsymbol{U}_{2i}$ are the diagonal covariance matrices of component densities respectively: $\boldsymbol{U}_{ri} = \text{Var}(\boldsymbol{y_i}|\boldsymbol{z_i} = \boldsymbol{2 - r})$. The covariance matrices $\boldsymbol{U}_{1i}$ and $\boldsymbol{U}_{2i}$ could be functions of both mean parameters $\boldsymbol{\mu_{1i}}$, $\boldsymbol{\mu_{2i}}$ and dispersion parameter $\boldsymbol{\phi}$. As a result, given $\boldsymbol{\theta}$, the dispersion parameter $\boldsymbol{\phi}$ can be consistently estimated via the second moment conditions of component distributions. For the independent model, $\boldsymbol{V}_i = \text{diag}\Big(\text{Var}(z_{i1}), ..., \text{Var}(z_{iT})\Big)$.

To account for serial correlation induced by latent variable $\boldsymbol{z_i}$, we assume $\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}(\boldsymbol{\rho}) \boldsymbol{A}_i^{\frac{1}{2}}$ analogue to the generalized estimating equation (GEE) ([44]), where $\boldsymbol{A}_i = \text{diag}\{\pi_{i1}(1-\pi_{i1}), ..., \pi_{iT}(1-\pi_{iT})\}$ is a diagonal matrix of marginal variance of $\boldsymbol{z}_i$, and $\boldsymbol{R}(\boldsymbol{\rho})$ is a working correlation matrix. If $\boldsymbol{R}$ is the true correlation matrix for $\boldsymbol{z}_i$, then $\boldsymbol{V}_i = \text{Cov}(\boldsymbol{z_i})$.

In the following, we extend the proposed estimating equation to the incomplete-data model. To handle the missing latent indicator $\boldsymbol{z_i}$, we take the conditional expectations for the equations in (2.4) given the outcome observation $\boldsymbol{y}$, and therefore construct the unbiased estimating equations

for the incomplete data as:

$$S(\boldsymbol{\theta}) = \mathbf{E}_{\boldsymbol{z}}[\sum_{i=1}^{n} S_i^c(\boldsymbol{\theta})|\boldsymbol{y}, \boldsymbol{\theta}] = \sum_{i=1}^{n} S_i(\boldsymbol{\theta}|\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\pi_i}}{\partial \boldsymbol{\eta}}\right)^T \boldsymbol{V}_i^{-1}(\boldsymbol{w}_i - \boldsymbol{\pi}_i) \\ \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu_{1i}}}{\partial \boldsymbol{\beta_1}}\right)^T {\boldsymbol{U}_{1i}}^{-1} \boldsymbol{W}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_{1i}) \\ \sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu_{2i}}}{\partial \boldsymbol{\beta_2}}\right)^T {\boldsymbol{U}_{2i}}^{-1}(1 - \boldsymbol{W}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i}) \end{pmatrix} = 0, \tag{2.5}$$

where $\boldsymbol{w}_i = \mathbf{E}_{\boldsymbol{z}}[\boldsymbol{z_i}|\boldsymbol{y}, \boldsymbol{\theta}]$ is the mixing weight representing the inferred probability of the subgroup memberships, and $\boldsymbol{W}_i$ is a diagonal matrix of corresponding mixing weights associated with $\boldsymbol{y_i}$. Similar to the GEE method, our approach can handle unbalanced data if the missing mechanism is missing completely at random (MCAR) ([69]). If the missing mechanism is not MCAR, then the estimating equation estimators could be biased and inefficient, and weighted generalized estimating equations (WGEE) can be employed to deal with the missing not completely at random cases ([66]; [68]).

To solve the estimating equation (2.5), we present a two-step Expectation-Estimating-Equation (EEE) algorithm. Analogue to the EM algorithm, the EEE algorithm follows an iterative estimating process. Specifically, at the $(k+1)$th step, based on the current estimation of parameter $\boldsymbol{\theta^{(k)}}$, we impute $\boldsymbol{w_i^{(k)}} = \mathbf{E}_{\boldsymbol{z}}[\boldsymbol{z_i}|\boldsymbol{y}, \boldsymbol{\theta}^{(k)}]$, and then update $\boldsymbol{\theta}^{(k+1)}$ by solving the estimating equation:

$$S(\boldsymbol{\theta}|\boldsymbol{\theta^{(k)}}) = \sum_{i=1}^{n} S_i(\boldsymbol{\theta}|\boldsymbol{\theta^{(k)}}) = \begin{pmatrix} \sum_{i=1}^{n} (\frac{\partial \boldsymbol{\pi_i}}{\partial \boldsymbol{\eta}})^T \boldsymbol{V}_i^{-1}(\boldsymbol{w}_i^{(k)} - \boldsymbol{\pi}_i) \\ \sum_{i=1}^{n} (\frac{\partial \boldsymbol{\mu_{1i}}}{\partial \boldsymbol{\beta_1}})^T {\boldsymbol{U}_{1i}}^{-1} \boldsymbol{W}_i^{(k)}(\boldsymbol{y}_i - \boldsymbol{\mu}_{1i}) \\ \sum_{i=1}^{n} (\frac{\partial \boldsymbol{\mu_{2i}}}{\partial \boldsymbol{\beta_2}})^T {\boldsymbol{U}_{2i}}^{-1}(1 - \boldsymbol{W}_i^{(k)})(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i}) \end{pmatrix} = 0. \tag{2.6}$$

A more detailed EEE algorithm is provided in Section 2.3.3.

The main difference between the above method and [67]'s approach is that we incorporate an additional estimating equation associated with the mixing proportion. This enables us to utilize the longitudinal correlations among the unobservable subgroup memberships from the same subject. Furthermore, in our estimating equation (2.5), we can also utilize the correlation information from the outcome variable $\boldsymbol{y_i}$ simultaneously through the working correlation structures of $\boldsymbol{U}_{1i}$ and $\boldsymbol{U}_{2i}$. In this chapter, we focus on utilizing the subject's group membership correlation over time. We

assume that y_{it}'s given z_{it}'s are conditionally independent within the same subject, and thus $\boldsymbol{U}_{1i}$ and $\boldsymbol{U}_{2i}$ are set to be diagonal.

Indeed, the proposed model assuming working correlation structure has the same identifiability problem as the independent model, which is equivalent to the mixtures-of-experts model. We can refer to [31]'s discussion for the identifiability problem of the mixtures-of-experts model.

2.3.2 Asymptotic Properties

In this chapter, we utilize the latent variable's serial correlation information to improve the parameter estimation and the subject's group membership prediction. We first examine the optimization properties for the complete-data equations.

Once we establish the optimal estimating equation $S^c(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta})$ for complete data, then the conditional expectation $S(\boldsymbol{y}, \boldsymbol{\theta}) = \mathbf{E}[S^c|\boldsymbol{y}]$ is the best estimating equation for the incomplete data such that the L_2 norm $\|S^c(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}) - S(\boldsymbol{y}, \boldsymbol{\theta})\|$ is minimized. In other words, if the proposed $S^c(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta})$ is efficient enough, then $S(\boldsymbol{y}, \boldsymbol{\theta})$ will inherit some efficiency from the complete-data model. In the following proposition, we establish the optimality of the unbiased estimating equation (2.4) for complete data.

Proposition 1. *If the working correlation structure is correctly specified, that is, $\mathbf{V}_i = Var(\boldsymbol{z}_i)$, the estimating equations in (2.4) are the optimal linear estimating equations with respect to $(\boldsymbol{z_i}, \boldsymbol{y_i})$, in the sense that the asymptotic variance of the estimator solved by (2.4) reaches the minimum.*

The proof is provided in the Section 2.7. Indeed, incorporating the serial correlation information is more important for the mixture model since $\boldsymbol{z_i}$ can not be observed. For the complete-data estimating equations in (2.4), the first set of equations with respect to the group membership and the other two equations associated with component densities are uncorrelated with each other. Therefore the induced serial correlations only affect the estimation of mixing proportion parameters in $\boldsymbol{\pi(\eta)}$, but have no influence on estimating the component parameters in $\boldsymbol{\mu_1(\beta_1)}$ and $\boldsymbol{\mu_2(\beta_2)}$. However, for the incomplete data, the three equations in (2.5) are correlated because the

imputed term $\boldsymbol{w_i}$ is a function of $\boldsymbol{y_i}$. For the longitudinal data of the subjects, taking advantage of the information at the other time points through accounting for the serial correlation improves the estimation efficiency of both the component parameters and the mixing proportion parameters.

It is known that the consistency of the above estimator only depends on correct specification of the mean functions, but does not rely on correct specification of the working correlation structure nor on the estimation of the correlation parameters ρ. This is one desirable property for the proposed approach since the serial correlation induced from unobservable latent variables could be difficult to model and estimate precisely. In contrast to the full joint likelihood approach, we only require the second moment approximation. Furthermore, the following theorem establishes the asymptotic properties of the proposed estimators.

Theorem 1. *Let $\boldsymbol{\gamma} = (\boldsymbol{\theta}', \boldsymbol{\phi}')'$ denote a grand parameter vector, in the proposed unbiased estimating equation (2.5), assuming that the following regularity conditions are satisfied:*
(i) $\hat{\boldsymbol{\phi}}$ is consistently estimated given $\boldsymbol{\theta}$ via some unbiased estimating equation $\sum_{i=1}^{n} H_i(\boldsymbol{\phi}|\boldsymbol{\theta}) = 0$;
(ii) $\hat{\rho}$ is $n^{1/2}$-consistent given $\boldsymbol{\gamma}$, denoted as $\hat{\rho}(\boldsymbol{\gamma})$;
(iii) $\|\partial\hat{\rho}/\partial\boldsymbol{\gamma}\| \leq O_p(1)$;
Let $G_i(\boldsymbol{\gamma}) = (S_i(\boldsymbol{\theta})', H_i(\boldsymbol{\phi}|\boldsymbol{\theta})')'$ be the augmented unbiased estimating equation, then the asymptotic distribution of $\hat{\boldsymbol{\gamma}}$ obtained from $\sum_{i=1}^{n} G_i = 0$ is: $n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \to N(\mathbf{0}, V_g)$, where the asymptotic covariance matrix V_g is given by:

$$V_g = lim_{n\to\infty} n \Big(\sum_{i=1}^{n} \nabla G_i(\boldsymbol{\gamma}) \Big)^{-1} \Big(\sum_{i=1}^{n} cov(G_i(\boldsymbol{\gamma})) \Big) \Big(\sum_{i=1}^{n} \nabla G_i(\boldsymbol{\gamma}) \Big)^{-T},$$

∇G_i is the gradient of G_i with respect to $\boldsymbol{\gamma}$.

Theorem 1 is established based on the following unbiased estimating equations:

$$\begin{aligned}
\mathbf{E}[(\boldsymbol{w}_i - \boldsymbol{\pi}_i)] = \mathbf{E}[(\mathbf{E}[\boldsymbol{z}_i|\boldsymbol{y}_i] - \boldsymbol{\pi}_i)] = \mathbf{E}[\boldsymbol{z}_i] - \boldsymbol{\pi}_i = 0; \\
\mathbf{E}[\boldsymbol{W}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_{1i})] = \mathbf{E}[\mathbf{E}[\boldsymbol{Z}_i|\boldsymbol{y}_i](\boldsymbol{y}_i - \boldsymbol{\mu}_{1i})] = \mathbf{E}[\boldsymbol{Z}_i(\boldsymbol{y}_i - \boldsymbol{\mu}_{1i})] = 0; \\
\mathbf{E}[(1 - \boldsymbol{W}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i})] = \mathbf{E}[(1 - \mathbf{E}[\boldsymbol{Z}_i|\boldsymbol{y}_i])(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i})] = \mathbf{E}[(1 - \boldsymbol{Z}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i})] = 0.
\end{aligned}$$

A sketch of the proof of Theorem 1 is given in the Section 2.7. Noting that even if $\boldsymbol{w}_i$ is imputed through the marginal expectation $w_{it} = \mathbf{E}[z_{it}|y_{it}]$, the estimating equations in (2.5) are still unbiased. The estimating equation H_i for the dispersion parameter is established based on the second moment condition provided in the Section 2.7.

In practice, a consistent variance estimator of $\hat{\boldsymbol{\gamma}}$ can be obtained by replacing $\text{Cov}(G_i(\boldsymbol{\gamma}))$ by $\hat{G}_i'\hat{G}_i$, and ∇G_i by $\nabla \hat{G}_i$ at the estimates of $\hat{\boldsymbol{\theta}}, \hat{\rho}, \hat{\boldsymbol{\phi}}$. One practical difficulty arises from getting an explicit form of the gradient $\nabla \hat{G}_i$ in the calculation of mixing weights $\boldsymbol{w}_i$. The imputation term $\boldsymbol{w}_i$ is the posterior probability of $\boldsymbol{z}_i$ containing both the response value $\boldsymbol{y}_i$ and all the parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and ρ. The complicated form of $\boldsymbol{w}_i$ makes it difficult to calculate the exact derivatives, especially when the conditional expectation $\boldsymbol{w}_i = \mathbf{E}[\boldsymbol{z}_i|\boldsymbol{y}_i]$ requires specifying the joint likelihood function for $\boldsymbol{z}_i$. We present some numerical approximation methods to calculate $\boldsymbol{w}_i$ and ∇G_i. More details are provided in Section 2.3.3.

2.3.3 Algorithm and Implementation

In this section, we provide the two-step iterative EEE algorithm and its theoretical properties. In addition, the details of marginal imputation and the joint imputation methods are provided.

Algorithm 1

Step 1: Set the initial values of the mean regression parameter $\boldsymbol{\theta}$ and the dispersion parameter $\boldsymbol{\phi}$: $\boldsymbol{\gamma}^{(0)} = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\phi}^{(0)})$;
Step 2 (E-Step): Impute the mixing weights $w_{it}^{(k)} = \mathbf{E}[z_{it}|\boldsymbol{y_i}]$ given current estimates $\boldsymbol{\gamma}^{(k)}$;
Step 3 (M-Step): Given the current imputed mixing weights $w_{it}^{(k)}$,
(i) update $\boldsymbol{\theta}^{(k+1)}$ by solving the estimating equation in (2.6), and
(ii) update $\boldsymbol{\phi}^{(k+1)}$ given $\boldsymbol{\theta}^{(k+1)}$;
Step 4: Return to Step 2 if $\|\boldsymbol{\gamma}^{(k+1)} - \boldsymbol{\gamma}^{(k)}\| > \epsilon$, where ϵ is the chosen tolerance level.

In Step 3 of Algorithm 1, with the current mixing weights, we apply the Newton-Raphson method to solve the estimating equations in (2.6) to obtain the updates $\boldsymbol{\theta}^{(k+1)}$ for the mean parameter. The updated dispersion parameter $\boldsymbol{\phi}^{(k+1)}$ can be estimated using the second moment conditions given $\boldsymbol{\theta}^{(k+1)}$. See Section 2.7 for more details. In the Newton-Raphson algorithm, the first derivative of the estimating function can be calculated as a simple diagonal matrix:

$$\text{diag}\Big(\sum_{i=1}^{n} \big(\tfrac{\partial \boldsymbol{\pi_i}}{\partial \boldsymbol{\eta}}\big)^T \boldsymbol{V}_i^{-1} \big(\tfrac{\partial \boldsymbol{\pi_i}}{\partial \boldsymbol{\eta}}\big), \sum_{i=1}^{n} \big(\tfrac{\partial \boldsymbol{\mu_{1i}}}{\partial \boldsymbol{\beta_1}}\big)^T {\boldsymbol{U}_{1i}}^{-1} \boldsymbol{W}_i^{(k)} \big(\tfrac{\partial \boldsymbol{\mu_{1i}}}{\partial \boldsymbol{\beta_1}}\big), \sum_{i=1}^{n} \big(\tfrac{\partial \boldsymbol{\mu_{2i}}}{\partial \boldsymbol{\beta_2}}\big)^T {\boldsymbol{U}_{2i}}^{-1} (1 - \boldsymbol{W}_i^{(k)}) \big(\tfrac{\partial \boldsymbol{\mu_{2i}}}{\partial \boldsymbol{\beta_2}}\big) \Big).$$

For the EEE Algorithm 1, the estimating function in (2.5) can be regarded as a bivariate function $S(\boldsymbol{\theta}|\boldsymbol{\lambda})$ restricted on the subspace $\boldsymbol{\theta} = \boldsymbol{\lambda}$, which is denoted as $S(\boldsymbol{\theta}|\boldsymbol{\theta})$. In fact, the first argument $\boldsymbol{\theta}$ of $S(\boldsymbol{\theta}|\boldsymbol{\theta})$ comes from the mean regression part, while the second argument comes from the conditional expectations. The EEE Algorithm 1 is simply updating the estimation of $\boldsymbol{\theta}$ by solving $S(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = 0$ in (2.6). If the iterative sequence of the estimator obtained from the EEE algorithm converges, it will converge to the solution of the original equation $S(\boldsymbol{\theta}|\boldsymbol{\theta}) = 0$ which is guaranteed by the continuity. The following lemma provides a local convergence of the estimator based on the Algorithm 1 in a more general form.

Lemma 1. *Suppose that $S(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is a bivariate function on the space $\Theta \times \Theta$ satisfying the following regularity conditions:*
(a) $S(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is continuous and both $\frac{\partial S}{\partial \boldsymbol{\theta}}$ and $\frac{\partial S}{\partial \boldsymbol{\lambda}}$ exist;
(b) In a neighborhood of $(\boldsymbol{\theta_0}, \boldsymbol{\theta_0})$, where $S(\boldsymbol{\theta_0}|\boldsymbol{\theta_0}) = 0$, we have $\| \, (\frac{\partial S}{\partial \boldsymbol{\theta}})^{-1} \cdot \frac{\partial S}{\partial \boldsymbol{\lambda}} \, \|_2 < 1$ hold;
then the iterative sequence of estimator $\{\boldsymbol{\theta}^{(k)}\}$ solved by $S(\boldsymbol{\theta}^{(k+1)}|\boldsymbol{\theta}^{(k)}) = 0$ converges to $\boldsymbol{\theta_0}$.

The proof of Lemma 1 is provided in the Section 2.7. Lemma 1 provides a sufficient but not necessary condition for the algorithm's convergence. Indeed, the algorithm convergence problem is also associated with general convergence theory regarding the iterative solutions to nonlinear equations. More detailed discussion of this type of algorithm can be found in [55].

In practice, it is possible to have multiple roots for the proposed estimating equations. One suggested method is to choose the root closest to the independent estimator which is consistent ([78]). [73] proposed to exclude all singular solutions and therefore reduce the risk of selecting spurious roots of the likelihood equation. Their method is powerful in selecting reasonable roots for the independent finite normal mixture model. In addition, we may also need to provide a "good" initial value for the iterations to converge. One way is to select different initial values randomly until the algorithm converges; another choice is to set the initial value as the estimator obtained from the existing independent mixture model.

In Algorithm 1, the E-step (Step 2) requires one to impute the missing latent variable $\boldsymbol{z_i}$ through its conditional expectation $\boldsymbol{w_i} = \mathbf{E}[\boldsymbol{z_i}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi}]$. However, this joint conditional expectation involves the joint density of the multivariate binary variable $\boldsymbol{z_i}$, which is very difficult to calculate. One possible way is to use the marginal imputation $w_{it} = \mathbf{E}[z_{it}|y_{it}, \boldsymbol{\theta}, \boldsymbol{\phi}]$ since we only assume a marginal distribution for the individual observation (y_{it}, z_{it}) in our approach. The marginal imputation

$$w_{it} = \frac{\pi_{it} f_1(y_{it}, \boldsymbol{\theta_1}, \phi_1)}{\pi_{it} f_1(y_{it}, \boldsymbol{\theta_1}, \phi_1) + (1 - \pi_{it}) f_2(y_{it}, \boldsymbol{\theta_2}, \phi_2)}$$

guarantees the estimating equations in (2.5) to be unbiased, where f_1 and f_2 are component densities. In fact, the imputation w_{it} provides the inferred subgroup membership of the outcome observation y_{it}. The drawback of the marginal imputation is that it only utilizes local information to infer the group membership. If two subgroups are well-separated, i.e., most w_{it} are either close to 1 or 0, then the marginal imputation is sufficient since the local information y_{it} dominates the subgroup membership's prediction. However, if two subgroups are not well-separated, then the marginal imputation does not fully utilize the within-subject correlation to improve the subgroup classification.

Therefore, we provide an alternative approximation of the joint imputation $w^*_{it} = \mathbf{E}[z_{it}|\boldsymbol{y_i}]$ which relies only on the second moment condition of $\boldsymbol{z}_i$. To predict the subgroup membership of y_{it}, we consider the conditional expectation $\mathbf{E}[z_{it}|y_{it}, \boldsymbol{z_{i(-t)}}]$ based on local observation y_{it} and all latent group labels $\boldsymbol{z_{i(-t)}}$ at other time points for the ith subject, where $\boldsymbol{z_{i(-t)}} = (z_{i1}, ...z_{i(t-1)}, z_{i(t+1)}, ..., z_{iT})$. Since

$$P(z_{it}|y_{it}, \boldsymbol{z_{i(-t)}}) \propto f(y_{it}|z_{it}) P(z_{it}|\boldsymbol{z_{i(-t)}}),$$

then $w^*_{it} = \pi^*_{it} f_1(y_{it}, \boldsymbol{\beta_1}, \phi_1) / [\pi^*_{it} f_1(y_{it}, \boldsymbol{\beta_1}, \phi_1) + (1 - \pi^*_{it}) f_2(y_{it}, \boldsymbol{\beta_2}, \phi_2)]$, where $\pi^*_{it} = \mathbf{E}[z_{it}|\boldsymbol{z_{i(-t)}}]$. By taking account of $\boldsymbol{z_{i(-t)}}$, this imputation w^*_{it} allows us to utilize the group membership information from other time points within the same subject as well.

However, the exact value of $\mathbf{E}[z_{it}|\boldsymbol{z_{i(-t)}}]$ still requires the joint likelihood function for $\boldsymbol{z}_i$.

Motivated by [61]'s conditional linear family for the multivariate binary distribution in generating correlated binary data, we consider a linear approximate:

$$\pi_{it}^* = \mathbf{E}[z_{it}|\boldsymbol{z_{i(-t)}}] = \pi_{it} + \boldsymbol{b_{it}^T}(\boldsymbol{z_{i(-t)}} - \boldsymbol{\pi_{i(-t)}}),$$

where $\pi_{it} = \mathbf{E}[z_{it}]$, $\boldsymbol{\pi_{i(-t)}} = \mathbf{E}[\boldsymbol{z_{i(-t)}}]$, and $\boldsymbol{b_{it}}$ is a $(T-1)$-dimensional coefficient vector. Based on the fact that for any two random variable X and Y, $\text{Cov}(X, Y) = \text{Cov}(X, \mathbf{E}[Y|X])$, we have:

$$\text{Cov}(\boldsymbol{z_{i(-t)}}, z_{it}) = \text{Cov}(\boldsymbol{z_{i(-t)}}, \pi_{it}^*) = \text{Cov}(\boldsymbol{z_{i(-t)}})\boldsymbol{b_{it}}.$$

Denote $\text{Cov}(\boldsymbol{z_{i(-t)}}, z_{it})$ as $\boldsymbol{s_{it}}$ and $\text{Cov}(\boldsymbol{z_{i(-t)}})$ as $\boldsymbol{V_{i(-t)}}$, then $\boldsymbol{b_{it}} = \boldsymbol{V_{i(-t)}^{-1}}\boldsymbol{s_{it}}$. Here both $\boldsymbol{s_{it}}$ and $\boldsymbol{V_{i(-t)}}$ could be obtained from the second moment condition $\text{Cov}(\boldsymbol{z_i})$.

To implement the Algorithm 1, at the $(k+1)$th step, we use the current estimators $\boldsymbol{\theta^{(k)}}$ and $\boldsymbol{V_{i(-t)}^{(k)}}$ to obtain $\pi_{it}^{*(k+1)}$ and thus impute $w_{it}^{*(k+1)}$. Here the true latent label $\boldsymbol{z_{i(-t)}}$ is replaced by the last prediction $\boldsymbol{\hat{z}_{i(-t)}^{(k)}}$. A similar technique is also used by [62] to approximate $P(z_{it}|\boldsymbol{\hat{z}_{i(-t)}^{(k)}})$ in a hidden Markov field modeling.

In Section 2.3.2, we mention that in order to obtain the robust variance estimator of $\hat{\boldsymbol{\gamma}}$, it requires us to calculate the gradient of $G_i(\boldsymbol{\gamma})$. Here we take the numerical approximation of the gradient $\nabla G_i(\boldsymbol{\gamma})$ by $\frac{G_i(\boldsymbol{\gamma}^{(k+1)}) - G_i(\boldsymbol{\gamma}^{(k)})}{\boldsymbol{\gamma^{(k+1)}} - \boldsymbol{\gamma^{(k)}}}$, where $\boldsymbol{\gamma^{(k)}}$ and $\boldsymbol{\gamma^{(k+1)}}$ are obtained from the previous two adjacent iterations, and the ijth component of $\nabla G_i(\boldsymbol{\gamma})$ corresponds to the ratio between the ith and jth components of $S_i(\boldsymbol{\gamma}^{(k+1)}) - S_i(\boldsymbol{\gamma}^{(k)})$ and $\boldsymbol{\gamma^{(k+1)}} - \boldsymbol{\gamma^{(k)}}$ respectively.

2.4 Numerical Study

In this section, we conduct two simulation studies to illustrate the performance of the proposed method on mixture modeling for longitudinal data. We are particularly interested in comparing the performance of our method to other approaches when the serial correlation is induced by latent variables. Our simulation results show that we can gain efficiency on parameter estimation by

utilizing correlation information.

In the first simulation study, we consider a two-component mixture of univariate normal distribution, and mainly focus on the estimating performance of the mixing proportion parameters with different dependence structures. In the second simulation study, we consider a mixture of two regression models, where the separation levels of two subgroups will change over time. The proposed method has extra power in both estimation and prediction, especially at poorly-separated time points.

2.4.1 Study 1: Two-component mixture of univariate normal densities

In this simulation study, we first generate component indicator variables Z_i from Bernoulli distributions for subjects $i = 1, ..., n$. For each Z_i, we assume there are T repeated measurements over time, where $Z_i = (Z_{i1}, ..., Z_{iT})$, and Z_{it} is a binary variable following a logistic regression with time covariate. That is, $\mathbf{E}[Z_{it}] = \pi_{it}$ and $\log(\frac{\pi_{it}}{1-\pi_{it}}) = \eta_0 + \eta_1 \frac{t}{T}$. Conditional on Z_{it}, we generate outcome response Y_{it} from two univariate normal distributions: $Y_{it}|(Z_{it} = 1) \sim N(\mu_1, \sigma_1^2)$ and $Y_{it}|(Z_{it} = 0) \sim N(\mu_2, \sigma_2^2)$. For latent variable Z_i, we assume independence among different subjects, but a common correlation structure (either AR-1 or exchangeable) within a subject over time. For outcome response Y_i, we assume independence among different subjects and conditional independence within a subject given component indicator Z_i. In our simulation studies, we generate binary responses through the R package "mvtBinaryEP".

We let the sample size $n = 100$ and time points $T = 10$, and the mixing proportion parameters in the logistic model are set to be $(\eta_0, \eta_1) = (-0.3, 0.5)$, which allows certain correlations among multivariate binary variables. In addition, we set variance parameters as $(\sigma_1^2, \sigma_2^2) = (1, 1)$.

The separation between two normal homoscedastic components could be accessed by $\Delta = |\mu_1 - \mu_2|/\sigma$, defined as the Mahalanobis distance between two normal mixture distributions ([52]). We investigate two settings with component means $(\mu_1, \mu_2) = (-1.5, 1.5)$ and $(\mu_1, \mu_2) = (-1.2, 1.2)$ to represent well-separated and poorly-separated bimodal densities, respectively. In addition, we also simulate a heterogeneous case with $(\mu_1, \mu_2) = (-1.5, 1.5)$ and $(\sigma_1, \sigma_2) = (1, 1.5)$. The sim-

ulation results are similar to those for the homogeneous case, and are thus omitted due to space limitations.

In the following simulation studies, we choose the initial value randomly in the neighborhood of the independent estimator until the EEE algorithm converges. The estimation results are summarized based on 1000 replicates. In our simulations, the empirical standard errors are quite close to the standard errors calculated from the robust sandwich variance, and therefore we only provide the empirical standard errors in our tables. In practice, label-switching issues might arise, especially for Bayesian mixture models. [92] and [93] discussed many feasible labeling methods. In our simulation study, we solve the labeling problem by imposing an ordered constraint on components' mean parameters.

We compare the estimators based on the proposed unbiased estimating equation with either joint imputation (UEE_{Joint}) or marginal imputation (UEE_{Mar}), with working correlation structure of either exchangeable, AR-1 or unstructured, the mixtures-of-experts model (Jacobs et al., 1991) which is the same as the independent estimating equation (UEE_{Ind}), and the oracle estimators ($Oracle$) assuming the true values of latent variable Z_i are known. The oracle estimator serves as a benchmark estimator, where the generalized estimating equation (GEE) utilizes the correlation structure for the binary data Z_i, and the MLE estimators of component parameters are obtained given Z_i. In the following tables, we denote the correlation structure in the superscript such as UEE^{AR1}_{Joint}. In the text, to avoid redundant notation, we omit the superscript if the correlation structure is correctly specified.

In addition, we also compare the proposed method with the random-effects model ([80]. In the random-effects model, we assume $\mathbf{E}[Z_{it}] = \pi_{it}$ and $\log(\frac{\pi_{it}}{1-\pi_{it}}) = \eta_0 + \eta_1 \frac{t}{T} + \gamma_i$, where $\gamma_i \sim N(0, \sigma^2_\gamma)$ is the random intercept accounting for the subject effect. Given γ_i, the latent variables within the ith subject are independent. The random-effects estimator (RE) are obtained by the EM algorithm.

In Tables 2.1 and 2.2, we observe that all the estimating equation estimators are consistent as discussed in Section 2.3.2, including the estimators using misspecified correlation structures.

However, we can improve the estimating efficiency if the correct correlation information is incorporated especially for joint-imputed model. This is reflected in that the UEE_{Joint} estimators have much smaller standard errors compared with the other estimators in both well-separated and poorly-separated cases. Indeed, the UEE_{Joint} performs almost the same as the $Oracle$ estimator. For a well-separated case or if an exchangeable correlation structure is assumed, the performance of the UEE_{Mar} is quite similar to that of the UEE_{Ind} since the longitudinal data generated here is balanced ([44]). But when two subgroups are poorly-separated and the latent variable has an AR-1 correlation structure, then the UEE_{Mar} has smaller standard errors than UEE_{Ind}.

In addition, if we compare Table 2.1 and Table 2.2, we notice that the standard errors of the independent estimators increase significantly from a well-separated case to a poorly-separated case, while the proposed UEE_{Joint} is much more stable and performs similarly to the $Oracle$ estimator. Table 2.3 provides the classification error rates from the model-based clustering. The proposed UEE_{Joint} model has sufficient power in predicting the subgroup membership since we incorporate the information from other time points for the same subject.

In general, it is difficult to know the true correlation structure for latent variables. The unstructured working correlation is always a possible choice because of its flexibility, as it does not assume any pattern for the correlation structure. However, the unstructured correlation leads to additional computational cost, as it has more correlation parameters $\frac{T*(T-1)}{2}$ to estimate compared with the AR-1 and exchangeable working correlations. In addition, the variation introduced by unstructured correlation leads to less efficient estimations for regression parameters and increases the chance of the convergence problem in the EEE algorithm. The unstructured model is recommended when the sample size n is large and the repeated measurement size T is relatively small in the well-separated case. In Tables 2.1 and 2.2, UEE_{Mar}^{Uns} and UEE_{Joint}^{Uns} do not show significant improvement in estimations, but they have more power in prediction compared with the independent and misspecified models in Table 2.3.

In Tables 2.1-2.2, the random-effects estimators perform poorly with large bias and standard errors. This is because [80] approach can only incorporate the random intercept which might not

be sufficient to handle the within-subject dependence if group membership changes over time. In addition, we also conduct a simulation study where the latent variable Z_i is generated from the random-effects model only (2.4). The random-effects model generates correlations which do not have an obvious patten. Therefore the proposed estimating equation assuming a certain pattern of working structure for serial correlations might not be the most efficient in estimation. Nevertheless, Table 2.4 indicates that although the standard errors of the estimating equation estimators are slightly overestimated, they are still acceptable. In all, regardless of which source the dependence within subjects is induced from, the proposed UEE estimators are robust and efficient in general.

2.4.2 Study 2: Two-component mixture of linear regression models

In simulation study 2, we consider a mixture of two linear regression models. In this case, not only the mixing proportion, but also the component densities follow a mean regression model with time covariates.

Similar to the first simulation study, we generate component indicator variables Z_i's from a logistic model, and for each Z_i, we assume there are T repeated measurements over time with $Z_i = (Z_{i1}, ..., Z_{iT})$. The mixing proportion $\mathbf{E}[Z_{it}] = \pi_{it}$ is generated from the logistic model:

$$\log(\frac{\pi_{it}}{1-\pi_{it}}) = \eta_0 + \eta_1 \frac{t}{T},$$

with either an AR-1 or exchangeable correlation structure. Conditional on Z_{it}, we generate the outcome response Y_{it} from two normal distributions:

$$Y_{it}|(Z_{it}=1) \sim N\Big(\beta_0^{(1)} + \beta_1^{(1)}\frac{t}{T}, \quad \sigma_1^2\Big),$$
$$Y_{it}|(Z_{it}=0) \sim N\Big(\beta_0^{(2)} + \beta_1^{(2)}\frac{t}{T}, \quad \sigma_2^2\Big).$$

In this study, we let the sample size $n = 100$ and time points $T = 5$, the mixing proportion parameters in logistic model are set to be $(\eta_0, \eta_1) = (-0.3, 0.5)$, and the component's regression parameters are set as $(\beta_0^{(1)}, \beta_1^{(1)}) = (-3, 2)$ and $(\beta_0^{(2)}, \beta_1^{(2)}) = (3, -2)$, the variance parameters are set as $(\sigma_1^2, \sigma_2^2) = (1, 1)$. In contrast to the first simulation setting, the mixture components

are regression functions of time covariates, where the component means are changing over time, leading to different separation levels at different times.

Figure 2.1 illustrates four different separations of two components at different time points in this simulation study. This motivates us to take advantage of serial correlations within subjects to improve accuracy in predicting class memberships and thus to improve the estimations of the regression parameters. In addition, we allow subjects to change group memberships over time (see Figure 2.2 as an illustration), which makes our approach more flexible compared to growth-curve mixture modeling.

The results in Table 2.5 show that we gain extra efficiency with smaller standard errors in estimating both the mixing proportion parameters and the component regression parameters for the proposed UEE_{Joint} estimators especially on slope estimators. Table 2.6 indicates that our approach utilizing the within-subject correlation information provides more predictive power, especially at poorly-separated times. In general, when within-subject correlation is large, our approach is able to "borrow" information from well-separated observations to enhance membership prediction for poorly-separated observations by incorporating the correlations within each individual subject. Consequently, improvement of the prediction of the subgroup's membership leads to better estimation of the slope parameter associated with time effect.

2.5 Real Data Application: 2008 Election Data

In this section, we apply the proposed method to the 2007-2008 AP-YAHOO NEWS election panel study (*http://www.knowledgenetworks.com/GANP/election2008/index.html*). The study was conducted by Knowledge Networks on behalf of the Associated Press and Yahoo! News (APYN) which intends to measure opinion changes starting with the primary elections through the presidential election in November 2008. The data consists of 4965 participants over eleven waves from November 2007 to December 2008, with nine waves before the election, one wave on election day, and the last wave after the election. The primary goal of the study is to investigate the change of

participants' interest in the election over time and important factors associated with their interest. One important factor associated with the interest in the election is interest in election news.

Therefore we choose one of the survey questions: "Question LV3: How much interest do you have in the following news about the campaign for president, a great deal, quite a bit, only some, very little, or no interest at all." The five categories of opinions are recorded as an ordinal response variable: 1, 2, 3, 4 and 5, where a smaller value corresponds to a high level of interest in the election news. In order to measure the opinion change towards the election, we focus on the first nine waves occurring before the election date. There are 1200 participants who have completed the "Question LV3" over the first nine waves. In the following, we have $n = 1200$ and the repeated measurement size $T = 9$.

In this study, we intend to classify all the participants into two groups, whether they actively follow the election or not, based on their responses to the question "LV3" in the AP-Yahoo survey. Since the survey collects participants' responses longitudinally, it is not surprising that their interest towards the election could be different at different time points. Consequently, this results in changes of group membership over time. The covariates include time, gender and race, where gender is 1 for female and 0 for male, and "race" consists of "white," "black" and "the other" coded as dummy variables with "white" as the reference. In addition, we also include an interaction term between "time" and the "gender."

We formulate a two-component mixture model as follows. Let the latent variable Z_{it} indicate whether the participant i at the time point t is interested in election news ($Z_{it} = 1$) or not ($Z_{it} = 0$). We model the mixing proportion using a logistic regression to capture the change in group membership, and the univariate Gaussian for the component distribution. We compare estimators via the proposed unbiased estimating equations using different working correlation structures: independent (UEE^{Ind}), exchangeable (UEE^{Exch}), AR-1 (UEE^{AR1}) and unstructured (UEE^{Uns}), in addition to the random-effects model (RE). The joint imputation is applied in all cases. We focus on modeling the mixing proportion since we are interested in opinion changes over time. The estimates of mixing proportion parameters with corresponding p-values are summarized in Table

7. The p-values are calculated based on the asymptotic normal distribution since the sample, size $n = 1200$ is quite large.

Table 2.7 indicates that the participants become more and more interested in the election campaign as the time gets closer to election day. Male participants are more interested in election news than females on average; however, females become increasingly more interested in election news than their male counterparts as the election gets closer. In addition, the "black" group is more interested in election news compared to the "white" group, while the "other" group is slightly less interested in election news than the "white" group. In total, about 43.5% of the participants changed their group memberships of showing interest in election news over time.

Furthermore, Table 2.7 shows that the proposed estimating equation estimators using different working correlations are similar in identifying the effects for "gender" and the interaction of "gender" and "time," which are highly significant. However, the estimators for the gender and the interaction term are not significant using the random-effects model with much larger p-values. In addition, the [59] reports that the "race" factor played a significant role in the 2008 presidential election, where "black" voters showed more interest in the election than other races in general. For this survey study, the "black" group is only about 7.5% of the participant population (90 out of 1200) which makes it more difficult to detect the "race" factor. Neither the random-effects model nor the independent estimating equations are able to detect a significant difference between the "black" group and the "white" group. However, the proposed method using the unstructured correlation is capable of identifying that the "black" group is significantly different from the "white" group with a p-value of 0.04. This implies that the proposed method accounting for the serial correlation can improve the estimation efficiency and increase testing power to detect a factor which might not be picked by other approaches.

2.6 Discussion

In this chapter, we propose an unbiased estimating equation approach for mixture modeling in longitudinal data, and illustrate how to induce correlation at the level of latent subgroup indicator variables. The proposed method does not require that each subject belong to the same group at different time. To circumvent the complicated form of joint likelihood of the multivariate Bernoulli variables, we propose an unbiased estimating equation approach utilizing the first two moment approximations of the full likelihood, where the serial correlation is incorporated through a working correlation structure. The proposed estimating equations can be regarded as a projection of optimal estimating equations obtained from the complete data via taking the conditional expectation for the missing latent variables.

Our numerical studies show that incorporating correlation information allows one to gain estimation efficiency for the mean regression parameters and the mixing proportion parameters compared to the independent models. The efficiency improvement is significant if the correlation from the longitudinal data is strong and the working structure is correctly specified. In addition, we can improve the classification accuracy for the boundary observations through joint imputation for the missing latent variable.

We can further generalize the current method for more than two subgroups in the population through the extended generalized estimating equation, applying the cumulative logit model for a multinomial latent variable. In addition, we can extend the univariate outcome variable to the multivariate component distribution. If the dimension of the multivariate distribution is high, we can employ the variable selection method by incorporating some penalty terms ([90]. In this chapter, we mainly focus on modeling serial correlation arising from the latent variable, it would be worthwhile to consider a more complicated setting where within-subject dependence is induced by both latent variable and outcome variable.

2.7 Proofs of Theoretical Results

A.1 Estimate dispersion parameter

The dispersion parameter $\boldsymbol{\phi} = (\phi_1, \phi_2)$ is associated with the second moments of the components' distributions. For components' densities in the exponential family, we have $\text{Var}(y_{it}|z_{it} = 2-r) = \nu(\mu_{rit})\phi_r$, $r = 1, 2$. In the complete-data case, given the true subgroup label z_{it}, for instance, the second moment of the first component distribution could be estimated by $\widehat{\text{Var}}(y_{it}|z_{it} = 1) = \sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it}(y_{it} - \mu_{1it})^2 / \sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it}$. Then we could establish an unbiased estimating equation for ϕ_1 given mean parameter $\mu_{1it}(\boldsymbol{\beta_1})$:

$$\sum_{i=1}^{n} H_i^c(\phi_1|\boldsymbol{\beta}_1) = \sum_{i=1}^{n}\sum_{t=1}^{T_i} z_{it}[(y_{it} - \mu_{1it})^2 - \nu(\mu_{1it})\phi_1] = 0.$$

Similarly, we could establish the unbiased estimating equation for incomplete data by taking the conditional expectation:

$$\sum_{i=1}^{n} H_i(\phi_1|\boldsymbol{\beta}_1) = \sum_{i=1}^{n}\sum_{t=1}^{T_i} w_{it}[(y_{it} - \mu_{1it})^2 - \nu(\mu_{1it})\phi_1] = 0,$$

where w_{it} is the imputed mixing weight, and ϕ_2 is estimated in the same way.

For example, in the two-component Gaussian mixture model, the variance parameter (σ_1^2, σ_2^2) for normal components could be estimated by $\hat{\sigma_1}^2 = \sum_{i=1}^{n}\sum_{t=1}^{T_i} w_{it}(y_{it} - \mu_{1it})^2 / \sum_{i=1}^{n}\sum_{t=1}^{T_i} w_{it}$.

A.2 Proof of Theorem 1

Note that the estimating equation G_i in (2.5) contains both interest parameter $\boldsymbol{\gamma} = (\boldsymbol{\theta}', \boldsymbol{\phi}')'$ and nuisance correlation parameter ρ. By assumptions, the correlation parameter ρ could be estimated consistently given $\boldsymbol{\gamma}$, written as $\hat{\rho}(\boldsymbol{\gamma})$. Then the augmented unbiased estimating equation in Theorem 1 has the form

$$\sum_{i=1}^{n} \hat{G}_i(\boldsymbol{\gamma}, \hat{\rho}(\boldsymbol{\gamma})) = 0.$$

From the theory of unbiased estimating equations ([25]), under some regularity conditions, $n^{1/2}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})$ could be approximated by the one-step Taylor expansion:

$$[-n^{-1}\sum_{i=1}^{n}\frac{dG_i}{d\boldsymbol{\gamma}}]^{-1}\cdot[n^{1/2}\sum_{i=1}^{n}\hat{G}_i],$$

where

$$\frac{d\hat{G}_i(\boldsymbol{\gamma},\hat{\rho}(\boldsymbol{\gamma}))}{d\boldsymbol{\gamma}}=\frac{\partial\hat{G}_i(\boldsymbol{\gamma},\hat{\rho}(\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}+\frac{\partial\hat{G}_i(\boldsymbol{\gamma},\hat{\rho}(\boldsymbol{\gamma}))}{\partial\hat{\rho}}\cdot\frac{\partial\hat{\rho}}{\partial\boldsymbol{\gamma}}.$$

With marginal imputation $w_{it}=\mathbf{E}[z_{it}|y_{it}]$, the nuisance parameter ρ is only contained in the working correlation matrix $\boldsymbol{R}$, therefore $\frac{\partial\hat{G}_i(\boldsymbol{\gamma},\hat{\rho}(\boldsymbol{\gamma}))}{\partial\hat{\rho}}$ is linear of unbiased estimating equations $(\boldsymbol{w_i}-\boldsymbol{\pi_i})$, thus $\sum_{i=1}^{n}\frac{\partial\hat{G}_i(\boldsymbol{\theta},\hat{\rho}(\boldsymbol{\gamma}))}{\partial\hat{\rho}}=o_p(\boldsymbol{n})$, and by condition (iii) $\|\frac{\partial\hat{\rho}}{\partial\boldsymbol{\gamma}}\|=O_p(1)$, therefore $n^{-1}\sum_{i=1}^{n}\frac{d\hat{G}_i(\boldsymbol{\gamma},\hat{\rho}(\boldsymbol{\gamma}))}{d\boldsymbol{\gamma}}=n^{-1}\sum_{i=1}^{n}\frac{\partial\hat{G}_i(\boldsymbol{\gamma},\hat{\rho}(\boldsymbol{\gamma}))}{\partial\boldsymbol{\gamma}}+o_p(\mathbf{1})$.

Further, fix $\boldsymbol{\gamma}$ and from Taylor expansion again:

$$n^{-1/2}\sum_{i=1}^{n}\hat{G}_i(\boldsymbol{\gamma},\hat{\rho})=n^{-1/2}\sum_{i=1}^{n}G_i(\boldsymbol{\gamma},\rho)+[n^{-1}\sum_{i=1}^{n}\frac{\partial G_i(\boldsymbol{\gamma},\rho)}{\partial\rho}]\cdot[n^{1/2}(\hat{\rho}-\rho)]+o_p(\mathbf{1}).$$

By condition (ii) $n^{1/2}(\hat{\rho}-\rho)=O_p(1)$ and also $n^{-1}\sum_{i=1}^{n}\frac{\partial G_i(\boldsymbol{\gamma},\rho)}{\partial\rho}=o_p(\mathbf{1})$, suggests that $n^{-1/2}\sum_{i=1}^{n}\hat{G}_i(\boldsymbol{\gamma},\hat{\rho})$ is asymptomatically equivalent to $n^{-1/2}\sum_{i=1}^{n}G_i(\boldsymbol{\gamma},\rho)$. Hence $n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ could be approximated by

$$[-n^{-1}\sum_{i=1}^{n}\frac{\partial G_i}{\partial\boldsymbol{\gamma}}]^{-1}\cdot[n^{1/2}\sum_{i=1}^{n}G_i],$$

which would be asymptotically Gaussian with mean vector of $\mathbf{0}$ and asymptotic variance of V_g.

A.3 Proof of Proposition 1

It is well-known from the theory of optimal estimating equations ([25]), that for unbiased estimating equation $g_i(\boldsymbol{\theta})$, the optimal weights would be $\mathrm{Var}(g_i)^{-1}\dot{g}_i$, where $\dot{g}_i=\frac{\partial g_i}{\partial\boldsymbol{\theta}}$, $\boldsymbol{\theta'}=$

$(\boldsymbol{\eta}', \boldsymbol{\beta_1}, \boldsymbol{\beta_2}')'$. As for the complete data $(\boldsymbol{y}_i, \boldsymbol{z}_i)$, the unbiased linear equation is

$$g_i = \begin{pmatrix} \boldsymbol{z}_i - \boldsymbol{\pi}_i(\boldsymbol{\eta}) \\ \boldsymbol{Z}_i\Big(\boldsymbol{y}_i - \boldsymbol{\mu}_{1i}(\boldsymbol{\beta_1})\Big) \\ (1-\boldsymbol{Z}_i)\Big(\boldsymbol{y}_i - \boldsymbol{\mu}_{2i}(\boldsymbol{\beta_2})\Big) \end{pmatrix}.$$

Firstly, it is easy to show that $\dot{g}_i$ has the form

$$\dot{g}_i = \begin{pmatrix} \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\eta}} & 0 & 0 \\ 0 & \boldsymbol{Z}_i \frac{\partial \boldsymbol{\mu}_{1i}}{\partial \boldsymbol{\beta_1}} & 0 \\ 0 & 0 & (1-\boldsymbol{Z}_i)\frac{\partial \boldsymbol{\mu}_{2i}}{\partial \boldsymbol{\beta_2}} \end{pmatrix}.$$

Also we could show that $\text{Var}(g_i)$ is a diagonal matrix

$$\text{Var}(g_i) = \begin{pmatrix} \text{Var}(\boldsymbol{z}_i) = V_i & 0 & 0 \\ 0 & \text{Var}(\boldsymbol{Z_i}(\boldsymbol{y}_i - \boldsymbol{\mu_{1i}})) & 0 \\ 0 & 0 & \text{Var}((1-\boldsymbol{Z_i})(\boldsymbol{y}_i - \boldsymbol{\mu_{2i}})) \end{pmatrix}.$$

This is because

$$\mathbf{E}[(z_{it} - w_{it})z_{ij}(y_{ij} - \mu_{1ij})] = \mathbf{E}\Big[(z_{it} - w_{it})z_{ij}\mathbf{E}[(y_{ij} - \mu_{1ij})|(z_{it}, z_{ij})]\Big] = 0$$

always holds for any t and j since $\mathbf{E}[(y_{ij} - \mu_{1ij})|(z_{it}, z_{ij})] = (1 - z_{ij})(\mu_{2ij} - \mu_{1ij})$. In addition, from the joint log-likelihood function (2.2) we could see that the component densities are estimated given the true values of latent variable in the complete-data case, and thus $\text{Var}(\boldsymbol{Z_i}(\boldsymbol{y}_i - \boldsymbol{\mu_{1i}})) = \boldsymbol{Z_i}\boldsymbol{U_{1i}}\boldsymbol{Z_i^T}$. Noting that $\boldsymbol{Z}_i$ is a diagonal matrix and $\boldsymbol{U_{1i}}, \boldsymbol{U_{2i}}$ are diagonal covariance matrices, therefore $\boldsymbol{Z}_i^T\text{Var}(\boldsymbol{Z_i}(\boldsymbol{y}_i - \boldsymbol{\mu_{1i}}))^{-1}\boldsymbol{Z}_i = \boldsymbol{U_{1i}^{-1}}\boldsymbol{Z}_i$ and $(1-\boldsymbol{Z}_i)^T\text{Var}((1-\boldsymbol{Z_i})(\boldsymbol{y}_i - \boldsymbol{\mu_{2i}}))^{-1}(1-\boldsymbol{Z}_i) = \boldsymbol{U_{2i}^{-1}}(1-\boldsymbol{Z}_i)$. Then the optimal equation has the form (2.4).

A.4 Proof of Lemma 1

Consider the bivariate function $S(\boldsymbol{\theta}|\boldsymbol{\lambda})$ on $\boldsymbol{\Theta}\otimes\boldsymbol{\Theta}$, we have the first order partial Taylor's expansion with respect to $\boldsymbol{\theta}$ in a neighborhood of $(\boldsymbol{\theta_0}, \boldsymbol{\theta_0})$:

$$S(\boldsymbol{\theta}|\boldsymbol{\lambda}) \approx S(\boldsymbol{\theta_0}|\boldsymbol{\lambda}) + \frac{\partial S(\cdot|\boldsymbol{\lambda})}{\partial \boldsymbol{\theta}} \cdot (\boldsymbol{\theta} - \boldsymbol{\theta_0}).$$

If we obtain $\hat{\boldsymbol{\theta}}$ by solving the equation $S(\boldsymbol{\theta}|\boldsymbol{\lambda}) = 0$, then $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}) \approx (\frac{\partial S(\cdot|\boldsymbol{\lambda})}{\partial \boldsymbol{\theta}})^{-1} \cdot S(\boldsymbol{\theta_0}|\boldsymbol{\lambda})$.
Apply partial Taylor's expansion again with respect to $\boldsymbol{\lambda}$:

$$S(\boldsymbol{\theta}_0|\boldsymbol{\lambda}) \approx S(\boldsymbol{\theta}_0|\boldsymbol{\theta}_0) + \frac{\partial S(\boldsymbol{\theta}_0|\cdot)}{\partial \boldsymbol{\lambda}} \cdot (\boldsymbol{\lambda} - \boldsymbol{\theta}_0),$$

which indicates that $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx A(\boldsymbol{\lambda} - \boldsymbol{\theta}_0)$, where $A = (\frac{\partial S(\cdot|\boldsymbol{\lambda})}{\partial \boldsymbol{\theta}})^{-1} \cdot \frac{\partial S(\boldsymbol{\theta}_0|\cdot)}{\partial \boldsymbol{\lambda}}$. Therefore we have

$$\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \| \leq \| A \| \cdot \| \boldsymbol{\lambda} - \boldsymbol{\theta}_0 \| .$$

Hence, the iteratively constructed sequence $\{\boldsymbol{\theta}^{(k)}\}$ satisfies $\| \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}_0 \| \leq \| A \| \cdot \| \boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}_0 \|$ and thus converges to $\boldsymbol{\theta}_0$ based on the condition that $\| A \| < 1$.

2.8 Tables and Figures

Table 2.1: The parameter estimators and their empirical standard errors (provided in the subscripts) for a well-separated two-component univariate normal mixture model, based on 1000 replicates. The latent variable $\boldsymbol{z}_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$.

True		UEE_{Ind}	UEE^{Ex}_{Mar}	UEE^{Ex}_{Joint}	UEE^{AR1}_{Mar}	UEE^{AR1}_{Joint}	UEE^{Uns}_{Mar}	UEE^{Uns}_{Joint}	RE	$Oracle$
Exch	μ_1	$-1.50_{0.09}$	$-1.50_{0.09}$	$\mathbf{-1.50_{0.05}}$	$-1.48_{0.09}$	$-1.50_{0.06}$	$-1.50_{0.10}$	$-1.49_{0.07}$	$-1.52_{0.05}$	$-1.50_{0.05}$
	σ_1	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.01_{0.06}$	$0.98_{0.04}$	$1.00_{0.03}$
	μ_2	$1.50_{0.09}$	$1.50_{0.09}$	$\mathbf{1.50_{0.05}}$	$1.48_{0.10}$	$1.49_{0.06}$	$1.50_{0.09}$	$1.49_{0.06}$	$1.52_{0.05}$	$1.50_{0.05}$
	σ_2	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$	$1.01_{0.06}$	$1.01_{0.06}$	$1.01_{0.06}$	$0.99_{0.06}$	$0.98_{0.04}$	$1.00_{0.03}$
	η_0	$-0.31_{0.22}$	$-0.31_{0.22}$	$\mathbf{-0.30_{0.19}}$	$-0.32_{0.23}$	$-0.32_{0.20}$	$-0.31_{0.22}$	$-0.31_{0.21}$	$-0.88_{0.62}$	$-0.30_{0.19}$
	η_1	$0.52_{0.16}$	$0.52_{0.16}$	$\mathbf{0.50_{0.13}}$	$0.51_{0.20}$	$0.51_{0.17}$	$0.50_{0.20}$	$0.50_{0.18}$	$1.50_{0.48}$	$0.50_{0.12}$
AR1	μ_1	$-1.50_{0.10}$	$-1.50_{0.10}$	$-1.50_{0.10}$	$-1.50_{0.09}$	$\mathbf{-1.50_{0.05}}$	$-1.50_{0.09}$	$-1.50_{0.07}$	$-1.51_{0.05}$	$-1.50_{0.05}$
	σ_1	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$	$1.00_{0.06}$	$1.00_{0.05}$	$1.00_{0.06}$	$1.00_{0.03}$
	μ_2	$1.50_{0.11}$	$1.50_{0.09}$	$1.50_{0.09}$	$1.50_{0.09}$	$\mathbf{1.50_{0.05}}$	$1.50_{0.09}$	$1.50_{0.09}$	$1.52_{0.05}$	$1.50_{0.05}$
	σ_2	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$
	η_0	$-0.31_{0.26}$	$-0.33_{0.27}$	$-0.33_{0.27}$	$-0.31_{0.26}$	$\mathbf{-0.31_{0.22}}$	$-0.29_{0.25}$	$-0.30_{0.23}$	$-0.41_{0.32}$	$-0.31_{0.21}$
	η_1	$0.51_{0.36}$	$0.53_{0.38}$	$0.53_{0.38}$	$0.51_{0.36}$	$\mathbf{0.51_{0.34}}$	$0.52_{0.36}$	$0.51_{0.36}$	$0.83_{0.49}$	$0.51_{0.34}$

Table 2.2: The parameter estimators and their empirical standard errors (provided in the subscripts) for a poorly-separated two-component univariate normal mixture model, based on 1000 replicates. The latent variable $\boldsymbol{z}_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$.

True		UEE_{Ind}	UEE^{Ex}_{Mar}	UEE^{Ex}_{Joint}	UEE^{AR1}_{Mar}	UEE^{AR1}_{Joint}	UEE^{Uns}_{Mar}	UEE^{Uns}_{Joint}	RE	$Oracle$
Exch	μ_1	$-1.21_{0.14}$	$-1.21_{0.14}$	$\mathbf{-1.20_{0.05}}$	$-1.21_{0.16}$	$-1.21_{0.09}$	$-1.210.16$	$-1.21_{0.07}$	$-1.22_{0.05}$	$-1.20_{0.05}$
	σ_1	$1.00_{0.07}$	$1.00_{0.07}$	$0.99_{0.04}$	$1.00_{0.08}$	$1.00_{0.05}$	$0.99_{0.08}$	$1.00_{0.05}$	$0.98_{0.04}$	$1.00_{0.03}$
	μ_2	$1.21_{0.14}$	$1.21_{0.14}$	$\mathbf{1.20_{0.05}}$	$1.21_{0.16}$	$1.20_{0.09}$	$1.20_{0.16}$	$1.20_{0.07}$	$1.22_{0.06}$	$1.20_{0.05}$
	σ_2	$1.00_{0.08}$	$1.00_{0.08}$	$1.00_{0.04}$	$1.01_{0.08}$	$1.01_{0.05}$	$1.00_{0.08}$	$1.00_{0.05}$	$0.98_{0.05}$	$1.00_{0.03}$
	η_0	$-0.33_{0.32}$	$-0.33_{0.31}$	$\mathbf{-0.30_{0.19}}$	$-0.32_{0.34}$	$-0.32_{0.31}$	$-0.31_{0.35}$	$-0.31_{0.23}$	$-0.64_{0.46}$	$-0.31_{0.18}$
	η_1	$0.52_{0.20}$	$0.52_{0.20}$	$\mathbf{0.50_{0.14}}$	$0.52_{0.20}$	$0.51_{0.19}$	$0.49_{0.23}$	$0.50_{0.20}$	$1.11_{0.50}$	$0.50_{0.12}$
AR1	μ_1	$-1.22_{0.19}$	$-1.22_{0.15}$	$-1.21_{0.10}$	$-1.21_{0.14}$	$\mathbf{-1.22_{0.06}}$	$-1.22_{0.14}$	$-1.22_{0.09}$	$-1.23_{0.05}$	$-1.20_{0.04}$
	σ_1	$1.00_{0.09}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.04}$	$1.00_{0.07}$	$1.00_{0.05}$	$0.97_{0.06}$	$1.00_{0.03}$
	μ_2	$1.21_{0.19}$	$1.19_{0.15}$	$1.19_{0.11}$	$1.20_{0.14}$	$\mathbf{1.20_{0.06}}$	$1.21_{0.14}$	$1.20_{0.10}$	$1.22_{0.06}$	$1.20_{0.05}$
	σ_2	$1.00_{0.09}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.04}$	$1.00_{0.07}$	$1.00_{0.05}$	$0.98_{0.04}$	$1.00_{0.04}$
	η_0	$-0.31_{0.39}$	$-0.32_{0.32}$	$-0.31_{0.28}$	$-0.31_{0.33}$	$\mathbf{-0.31_{0.23}}$	$-0.31_{0.34}$	$-0.31_{0.25}$	$-0.51_{0.43}$	$-0.31_{0.21}$
	η_1	$0.51_{0.38}$	$0.53_{0.37}$	$0.48_{0.39}$	$0.51_{0.36}$	$\mathbf{0.51_{0.33}}$	$0.52_{0.36}$	$0.53_{0.36}$	$0.83_{0.62}$	$0.51_{0.34}$

Table 2.3: Classification error rate based on a two-component univariate normal mixture model. The latent variable $\boldsymbol{z}_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$.

Separation	Correlation	UEE_{Ind}	UEE^{Ex}_{Mar}	UEE^{Ex}_{Joint}	UEE^{AR1}_{Mar}	UEE^{AR1}_{Joint}	UEE^{Uns}_{Mar}	UEE^{Uns}_{Joint}	RE
Well	AR-1	0.070	0.069	0.055	0.069	**0.038**	0.069	0.051	0.055
	Exch	0.069	0.069	**0.030**	0.069	0.045	0.069	0.042	0.041
Poorly	AR-1	0.127	0.122	0.095	0.121	**0.068**	0.121	0.089	0.097
	Exch	0.121	0.121	**0.051**	0.123	0.092	0.122	0.088	0.093

Table 2.4: The parameter estimators and their empirical standard errors (provided in the subscripts) for a two-component univariate normal mixture model, based on 1000 replicates. The latent variable $\boldsymbol{z}_i$ is generated by random-effects model.

Separation		UEE_{Ind}	UEE^{AR1}_{Mar}	UEE^{AR1}_{Joint}	UEE^{Ex}_{Mar}	UEE^{Ex}_{Joint}	UEE^{Uns}_{Mar}	UEE^{Uns}_{Joint}	RE
Well	μ_1	$-1.51_{0.09}$	$-1.51_{0.09}$	$-1.51_{0.09}$	$-1.50_{0.09}$	$-1.50_{0.09}$	$-1.51_{0.09}$	$-1.51_{0.09}$	$-1.50_{0.06}$
	σ_1	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.01_{0.06}$	$1.00_{0.04}$
	μ_2	$1.51_{0.09}$	$1.50_{0.09}$	$1.49_{0.09}$	$1.51_{0.09}$	$1.51_{0.09}$	$1.49_{0.09}$	$1.50_{0.08}$	$1.50_{0.06}$
	σ_2	$1.00_{0.08}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$0.99_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$
	η_0	$-0.31_{0.18}$	$-0.31_{0.18}$	$-0.29_{0.18}$	$-0.31_{0.18}$	$-0.31_{0.18}$	$-0.28_{0.19}$	$-0.29_{0.19}$	$-0.32_{0.16}$
	η_1	$0.51_{0.25}$	$0.51_{0.25}$	$0.51_{0.25}$	$0.51_{0.25}$	$0.51_{0.25}$	$0.49_{0.27}$	$0.49_{0.26}$	$0.54_{0.27}$
Classification errors		0.069	0.068	0.067	0.069	0.069	0.069	0.067	0.067
Poorly	μ_1	$-1.21_{0.13}$	$-1.20_{0.14}$	$-1.20_{0.14}$	$-1.20_{0.14}$	$-1.20_{0.13}$	$-1.22_{0.15}$	$-1.21_{0.14}$	$-1.20_{0.05}$
	σ_1	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.07}$	$0.99_{0.07}$	$0.99_{0.07}$	$1.00_{0.04}$
	μ_2	$1.22_{0.14}$	$1.20_{0.14}$	$1.21_{0.14}$	$1.21_{0.14}$	$1.22_{0.14}$	$1.21_{0.15}$	$1.19_{0.16}$	$1.20_{0.05}$
	σ_2	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.06}$	$1.00_{0.07}$	$1.00_{0.07}$	$1.00_{0.04}$
	η_0	$-0.33_{0.26}$	$-0.31_{0.28}$	$-0.30_{0.26}$	$-0.33_{0.26}$	$-0.33_{0.26}$	$-0.33_{0.26}$	$-0.32_{0.26}$	$-0.30_{0.17}$
	η_1	$0.54_{0.28}$	$0.51_{0.27}$	$0.51_{0.27}$	$0.54_{0.28}$	$0.53_{0.28}$	$0.55_{0.28}$	$0.54_{0.28}$	$0.50_{0.26}$
Classification errors		0.116	0.118	0.116	0.120	0.116	0.130	0.128	0.114

Table 2.5: The parameter estimators and their empirical standard errors (provided in the subscripts) for a mixture of two regression models, based on 1000 replicates. Latent variable $\boldsymbol{z}_i$ is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$

True		UEE_{Ind}	UEE^{Ex}_{Mar}	UEE^{Ex}_{Joint}	UEE^{AR1}_{Mar}	UEE^{AR1}_{Joint}	UEE^{Uns}_{Mar}	UEE^{Uns}_{Joint}	$Oracle$
Exch	$\beta_0^{(1)}$	$-3.00_{0.17}$	$-3.00_{0.17}$	$-3.00_{0.16}$	$-2.99_{0.17}$	$-3.00_{0.16}$	$-3.00_{0.17}$	$-3.00_{0.16}$	$-3.00_{0.14}$
	$\beta_1^{(1)}$	$2.00_{0.30}$	$2.00_{0.30}$	$\mathbf{2.00_{0.24}}$	$1.95_{0.30}$	$1.98_{0.29}$	$1.98_{0.30}$	$1.97_{0.25}$	$2.00_{0.21}$
	$\sigma^{(1)}$	$0.99_{0.06}$	$0.99_{0.06}$	$0.99_{0.05}$	$0.99_{0.06}$	$0.99_{0.06}$	$0.99_{0.06}$	$0.99_{0.06}$	$1.00_{0.04}$
	$\beta_0^{(2)}$	$2.99_{0.16}$	$2.99_{0.16}$	$3.00_{0.15}$	$2.99_{0.17}$	$2.99_{0.16}$	$2.99_{0.16}$	$2.99_{0.16}$	$2.99_{0.13}$
	$\beta_1^{(1)}$	$-1.97_{0.29}$	$-1.97_{0.29}$	$\mathbf{-2.00_{0.23}}$	$-1.98_{0.32}$	$-1.99_{0.24}$	$-1.99_{0.29}$	$-1.99_{0.24}$	$-1.99_{0.22}$
	$\sigma^{(2)}$	$0.99_{0.06}$	$0.99_{0.06}$	$0.99_{0.04}$	$0.99_{0.06}$	$0.99_{0.06}$	$0.99_{0.06}$	$1.00_{0.06}$	$1.00_{0.04}$
	η_0	$-0.31_{0.22}$	$-0.31_{0.22}$	$\mathbf{-0.30_{0.19}}$	$-0.31_{0.26}$	$-0.32_{0.23}$	$-0.31_{0.24}$	$-0.30_{0.22}$	$-0.30_{0.19}$
	η_1	$0.52_{0.36}$	$0.52_{0.36}$	$\mathbf{0.49_{0.19}}$	$0.47_{0.40}$	$0.52_{0.28}$	$0.47_{0.36}$	$0.46_{0.25}$	$0.51_{0.18}$
AR-1	$\beta_0^{(1)}$	$-2.99_{0.17}$	$-2.99_{0.17}$	$-3.03_{0.17}$	$-2.99_{0.17}$	$-3.00_{0.16}$	$-3.02_{0.17}$	$-3.02_{0.16}$	$-3.00_{0.15}$
	$\beta_1^{(1)}$	$1.98_{0.29}$	$1.98_{0.32}$	$2.04_{0.29}$	$1.98_{0.30}$	$\mathbf{2.00_{0.24}}$	$1.99_{0.34}$	$2.00_{0.28}$	$2.00_{0.21}$
	$\sigma^{(1)}$	$0.99_{0.06}$	$0.99_{0.06}$	$1.01_{0.05}$	$0.99_{0.05}$	$0.99_{0.04}$	$0.98_{0.06}$	$0.98_{0.05}$	$1.00_{0.04}$
	$\beta_0^{(2)}$	$3.00_{0.17}$	$3.00_{0.17}$	$3.01_{0.17}$	$3.01_{0.17}$	$3.00_{0.16}$	$2.99_{0.17}$	$3.01_{0.16}$	$3.00_{0.15}$
	$\beta_1^{(1)}$	$-2.00_{0.32}$	$-1.97_{0.31}$	$-2.02_{0.29}$	$-2.00_{0.34}$	$\mathbf{-2.00_{0.26}}$	$-1.96_{0.32}$	$-1.98_{0.28}$	$-2.00_{0.22}$
	$\sigma^{(2)}$	$0.99_{0.06}$	$0.99_{0.06}$	$1.01_{0.05}$	$0.99_{0.05}$	$0.99_{0.04}$	$1.01_{0.06}$	$0.99_{0.05}$	$1.00_{0.04}$
	η_0	$-0.31_{0.26}$	$-0.31_{0.27}$	$-0.33_{0.26}$	$-0.30_{0.26}$	$\mathbf{-0.31_{0.23}}$	$-0.28_{0.28}$	$-0.29_{0.26}$	$-0.31_{0.23}$
	η_1	$0.51_{0.45}$	$0.50_{0.48}$	$0.54_{0.44}$	$0.51_{0.45}$	$\mathbf{0.50_{0.37}}$	$0.46_{0.47}$	$0.48_{0.40}$	$0.51_{0.33}$

Table 2.6: Classification error rate at each time point for a mixture of two regression models. Latent variable z_i is generated by exchangeable (Exch) and AR-1 structures with serial correlation parameter $\rho = 0.7$.

True Corr	Model	Time Points $t=1$	$t=2$	$t=3$	$t=4$	$t=5$	True Corr	Model	Time Points $t=1$	$t=2$	$t=3$	$t=4$	$t=5$
	UEE_{Ind}	0.005	0.014	0.037	0.086	0.165		UEE_{Ind}	0.004	0.014	0.036	0.084	0.164
	UEE^{Ex}_{Mar}	0.006	0.013	0.035	0.081	0.165		UEE^{AR1}_{Mar}	0.005	0.017	0.034	0.076	0.168
	UEE^{Ex}_{Joint}	0.005	0.009	0.022	0.057	0.140		UEE^{AR1}_{Joint}	0.003	0.011	0.028	0.060	0.098
AR1	UEE^{Uns}_{Mar}	0.005	0.009	0.037	0.083	0.162	Exch	UEE^{Uns}_{Mar}	0.005	0.013	0.036	0.084	0.164
	UEE^{Uns}_{Joint}	0.003	0.008	0.022	0.058	0.124		UEE^{Uns}_{Joint}	0.003	0.010	0.023	0.045	0.080
	UEE^{AR1}_{Mar}	0.005	0.014	0.037	0.085	0.161		UEE^{Ex}_{Mar}	0.004	0.014	0.036	0.084	0.164
	UEE^{AR1}_{Joint}	**0.003**	**0.007**	**0.019**	**0.055**	**0.115**		UEE^{Ex}_{Joint}	**0.002**	**0.008**	**0.019**	**0.040**	**0.069**

Table 2.7: Mixing proportion estimators and corresponding p-values (in subscripts) for the two-component mixture model of the election data.

	$\hat{\eta}_0$	$\hat{\eta}_{Time}$	$\hat{\eta}_{Gender}$	$\hat{\eta}_{Black}$	$\hat{\eta}_{Other}$	$\hat{\eta}_{G*T}$
UEE^{Ind}	$-0.71_{0.00}$	$1.24_{0.00}$	$-0.22_{0.00}$	$0.17_{0.18}$	$-0.02_{0.81}$	$0.29_{0.00}$
UEE^{Exch}	$-0.70_{0.00}$	$1.24_{0.00}$	$-0.20_{0.01}$	$0.17_{0.19}$	$-0.01_{0.91}$	$0.29_{0.00}$
UEE^{AR1}	$-0.83_{0.00}$	$1.28_{0.00}$	$-0.24_{0.02}$	$\mathbf{0.21_{0.09}}$	$-0.03_{0.76}$	$0.31_{0.00}$
UEE^{Uns}	$-0.77_{0.00}$	$1.23_{0.00}$	$-0.19_{0.01}$	$\mathbf{0.25_{0.04}}$	$0.02_{0.86}$	$0.24_{0.00}$
RE	$-0.47_{0.03}$	$3.01_{0.00}$	$-0.37_{0.19}$	$0.38_{0.42}$	$-0.08_{0.82}$	$0.44_{0.09}$

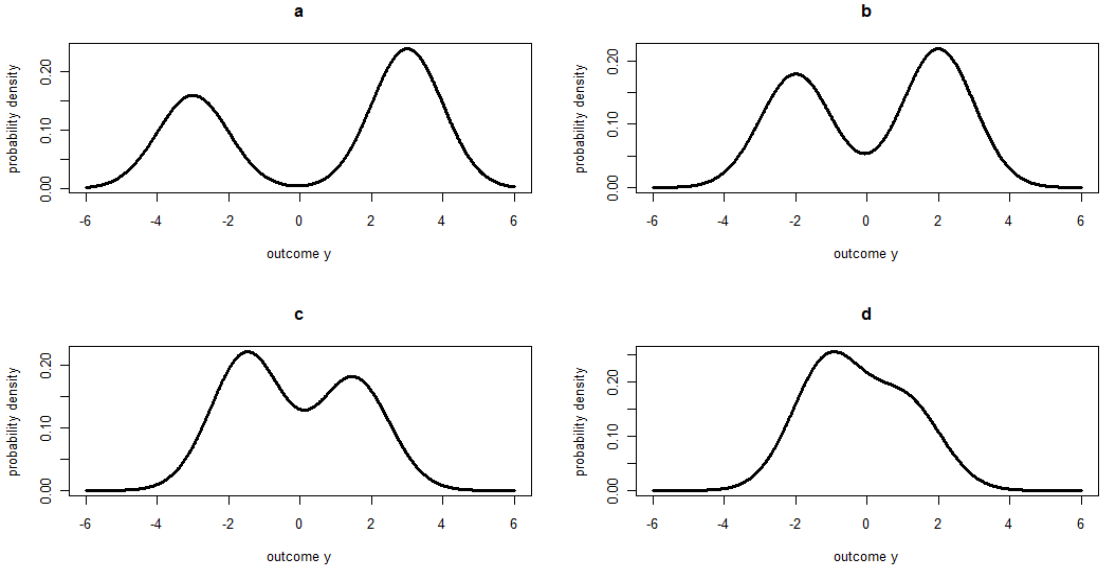

Figure 2.1: Plots of the mixture density of two univariate normal components with mixing proportion (corresponding to the negative-mean component): 0.4, 0.45, 0.55 and 0.6 for (a) to (d), a common variance $\sigma^2 = 1$, and the mean parameters are: (a) $(\mu_1, \mu_2) = (-3, 3)$, (b) $(\mu_1, \mu_2) = (-2, 2)$, (c) $(\mu_1, \mu_2) = (-1.5, 1.5)$, (d) $(\mu_1, \mu_2) = (-1.1, 1.1)$.

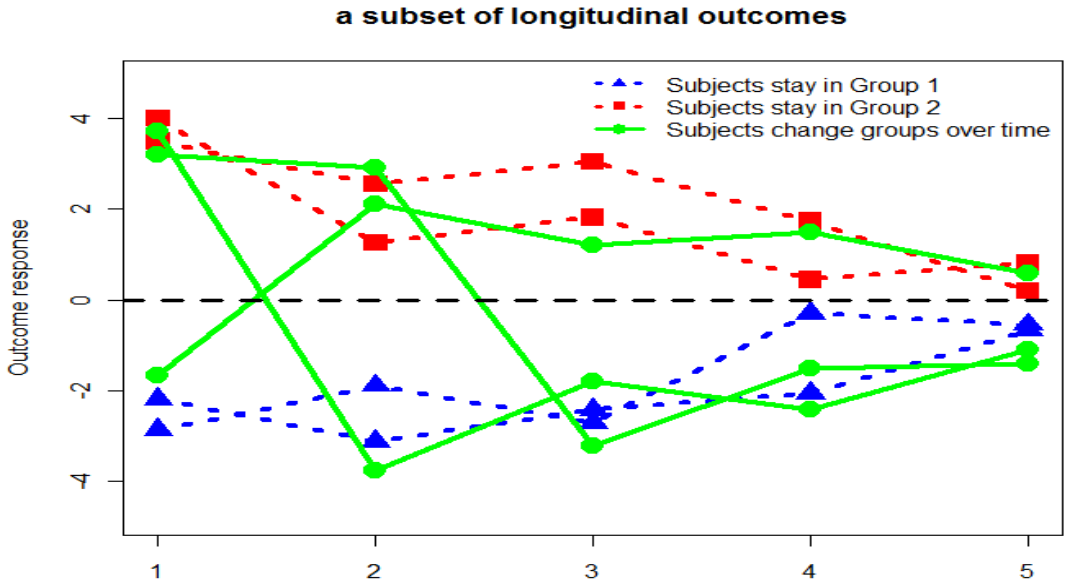

Figure 2.2: Plots of longitudinal responses from a subset of subjects in simulation study 2.4.2.

Chapter 3

Individualized Multi-directional Variable Selection

3.1 Introduction

In recent years there has been a growing demand for exploring individualized modeling, which has broad applications in personalized medicine, personalized education and personalized marketing. The traditional one-model-fits-the-whole-population approach is unable to detect important patterns and make personalized predictions for specific individuals. In addition, the rise of precision medicine and wide-spread electronic health record data also motivate us to develop more effective personalized treatment. The collection of rich data information makes it feasible and compelling to utilize individualized models as traditional population models cannot incorporate heterogeneous effects from different individuals.

In this chapter, we consider an individualized model based on a double-divergence framework, where the number of subjects and the amount of individual information increase together. Consequently, this introduces a diverging number of parameters as the sample size of subjects increases. In addition, one unique challenge of individualized model selection is that there could be different relevant or important predictors for different subjects. For instance, different individuals may have different prognostic factors associated with the same disease. Therefore it is important to develop new statistical methodology and theory for variable selection and estimation for individualized modeling.

In the past two decades several penalized model selection methods have been developed, e.g.,

the Lasso [81], the smoothly clipped absolute deviation (SCAD) [17], the elastic net [99], the adaptive Lasso [100], the group Lasso [94], the minimax concave penalty (MCP) [95] and the truncated L_1-penalty (TLP) [76]. However, the above methods are based on a homogeneous model setting which selects predictors for entire populations. For the individualized model, we can employ traditional variable selection methods separately for each subject, if there are multiple observations from each subject as in longitudinal data settings. However, in practice, the number of measurements for particular individuals could be limited. In addition, it is likely that some variables are invariant for the same subject, such as demographic information variables, e.g., race and gender, which impose restrictions and additional obstacles to performing individualized variable selection based on a standard subject-wise model framework.

Another limitation of applying standard subject-wise variable selection is that it ignores information from other subjects which might share similar effects on important predictors of interest. Moreover, assuming each individual to have unique effects for all covariates is practically unrealistic and computationally infeasible. In contrast, it is more sensible to assume that a subpopulation of individuals share common effects on selected predictors. In addition, borrowing information from homogeneous subgroups allows one to increase estimation efficiency and model selection accuracy.

In order to utilize cross-subject information, one may assume that an underlying subpopulation structure depends on unobserved covariates. Existing approaches dealing with clustering on regression coefficients include mixture modeling for regression, such as the mixture-of-experts model [30]. However, most model selection approaches under this framework including [64], [56] and [23] only focus on choosing informative variables to distinguish different subgroups, rather than on selecting relevant predictors for different individuals.

Alternative approaches to model-based clustering on regression coefficients employ grouping penalization. For example, [83] propose a fused Lasso by adding an L_1-penalty to the pair of adjacent coefficients; [5] propose a clustering algorithm for regression by imposing a special octagonal shrinkage penalty on each pair of coefficients; [75] develop a grouping pursuit algorithm utiliz-

ing the truncated L_1-penalty for fusions, and [35] propose a data-driven segmentation method to explore homogeneous groups with regression. Nevertheless, these are all still under the population-regression model, and do not allow different individuals to have different features. For the purpose of subgrouping different individuals, [26] and [45] formulate clustering as a penalized regression problem by adopting an L_p-fusion penalty. [57] and [47] apply non-convex fusion penalties to solve the bias problem. However, the fusion-type of penalty focuses on subgrouping rather than on model selection for individual coefficients.

In this chapter, we propose an effective individualized model selection approach utilizing multi-directional shrinkage to select unique relevant variables for different individuals. To the best of our knowledge, this is a new approach which has not been offered in the existing literature.

Specifically, the proposed penalty allows multiple possible shrinking directions including the one towards zero, which differs from conventional penalty functions with shrinking direction towards zero only. The consequence of conventional penalty functions is that non-zero signals could suffer from zero-directional shrinkage, although a variety of penalty methods have been proposed to solve the bias problem such as non-concave penalties (e.g., SCAD, MCP and TLP) or adaptive weights (e.g., adaptive Lasso). Instead we propose a rather different approach which shrinks penalized parameters to one of the multiple directions including zero, where the best shrinking direction is determined by the data itself. One advantage of the proposed method is that, as long as the candidate directions contain the one closest to the truth, the optimal large sample properties such as the oracle property hold by applying the L_1-type of penalty function in each direction.

Another advantage of the proposed method is that it separates different groups of individuals based on their effects on the same covariates. Indeed, the proposed penalty function is analogous to an objective function from center-based clustering, which can be viewed as a "separation penalty" among different individuals. As a byproduct, we identify subgroups with individuals sharing similar covariate effects, where the centers of subgroups provide a set of estimated shrinking directions. In addition, through utilizing cross-subject information, the proposed model improves estimation efficiency and thus enhances personalized prediction power.

Another contribution of this chapter is that we lay out a theoretical framework for the double-divergence individualized model with serial correlation. [91] and [2] established rigorous large sample theory for the generalized estimating equation [44] (GEE) estimator when the number of clusters and the cluster size are both large while the dimension of parameters is fixed; and [88] investigate the GEE model with high-dimensional covariates, but bounded cluster size. In contrast we establish theoretical properties in a framework when the number of clusters and the cluster size are both increasing, which involves high-dimensional parameters. We develop asymptotic theory for the oracle estimator and demonstrate the subpopulation effects on model estimation. In addition, we show the advantage of utilizing the multi-directional penalty for establishing the oracle property. Moreover, the proposed method is capable of incorporating within-subject correlation to achieve efficient estimation.

The rest of this chapter is organized as follows. Section 3.2 introduces the model framework and presents the proposed methodology. Section 3.3 establishes the theoretical results. Section 3.4 proposes an efficient algorithm with implementation. Section 3.5 provides simulation studies. Section 3.6 illustrates an application for HIV data. The last section provides concluding remarks and discussion.

3.2 Model Framework and Methodology

3.2.1 The individualized model and subject-wise variable selection

We formulate the problem under the clustered data setting, where each subject has multiple observations. Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{im_i})'$ be an m_i-dimensional response variable for the ith individual, $\boldsymbol{X_i} = (\boldsymbol{x_{i,1}}, \ldots, \boldsymbol{x_{i,p}})$ be an $m_i \times p$ covariates matrix corresponding to individual predictors, and $\boldsymbol{Z_i} = (\boldsymbol{z_{i,1}}, \ldots, \boldsymbol{z_{i,q}})$ be an $m_i \times q$ covariates matrix corresponding to population-shared predictors, where $i = 1, \ldots, N$. For ease of notation, we assume that the clustered data is balanced with cluster size $m_i = m$, although the development of the method does not require a balanced data structure.

We consider a regression model:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{Z_i}\boldsymbol{\alpha} + \boldsymbol{\varepsilon_i}, \quad i = 1, \ldots, N,$$

where each individual has its own regression parameter vector $\boldsymbol{\beta_i} = (\beta_{i1}, \ldots, \beta_{ip})'_{p\times 1}$, in addition to the population-shared parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)'_{q\times 1}$, and random errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})'_{m\times 1}$ independent over different subjects. Within a subject, ε_{ij}'s ($j = 1, \ldots, m$) have mean 0 and variance σ^2, and they could be correlated such as in the longitudinal date setting.

In general, to identify unique features for different individuals, we select and estimate the regression parameters $\boldsymbol{\beta_i}$'s and $\boldsymbol{\alpha}$ through minimizing the penalized objective function

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{N} L(\boldsymbol{y_i} - \boldsymbol{\mu_i}) + \sum_{i=1}^{N}\sum_{k=1}^{p} h^{(1)}_{\lambda_1}(\beta_{ik}) + \sum_{l=1}^{q} h^{(2)}_{\lambda_2}(\alpha_l), \tag{3.1}$$

where $\boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha}) = \boldsymbol{X_i}\boldsymbol{\beta_i} + \boldsymbol{Z_i}\boldsymbol{\alpha}$, $L(\cdot)$ is a loss function, $h^{(1)}_{\lambda_1}(\cdot)$ and $h^{(2)}_{\lambda_1}(\cdot)$ are feature-selection penalties for individualized parameters and population-shared parameters respectively, and λ_1, λ_2 are the corresponding tuning parameters. The selection of population parameter $\boldsymbol{\alpha}$ is regular and thus, in this chapter, we focus on individualized variable selection. To simplify the model, with a squared-error loss, the objective function in (3.1) becomes

$$\frac{1}{2}\sum_{i=1}^{N} \parallel \boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel_2^2 + \sum_{i=1}^{N}\sum_{k=1}^{p} h_{\lambda_{N,m}}(\beta_{ik}), \tag{3.2}$$

where $\parallel \cdot \parallel_2$ is the Euclidean norm. Then we could employ different penalties $h_{\lambda_{N,m}}(\cdot)$ to adopt traditional penalized selection approaches (e.g. Lasso, adaptive Lasso, MCP and SCAD).

Without the penalty term $h_{\lambda_{N,m}}(\cdot)$, minimizing (3.2) leads to the ordinary least squares (OLS) estimator. Let $\boldsymbol{\beta} = (\boldsymbol{\beta_1'}, \ldots, \boldsymbol{\beta_N'})'$ be the individualized coefficients vector and $\boldsymbol{Y} = (\boldsymbol{y_1'}, \ldots, \boldsymbol{y_N'})'$. We denote $\boldsymbol{X} = \text{diag}(\boldsymbol{X_1}, \ldots, \boldsymbol{X_N})$, a block-diagonal matrix, and $\boldsymbol{Z} = (\boldsymbol{Z_1'}, \ldots, \boldsymbol{Z_N'})'$. The OLS estimator is

$$(\hat{\boldsymbol{\beta}}^{Sub'}, \hat{\boldsymbol{\alpha}}^{Sub'})' = [(\boldsymbol{X}, \boldsymbol{Z})^T(\boldsymbol{X}, \boldsymbol{Z})]^{-1}(\boldsymbol{X}, \boldsymbol{Z})^T\boldsymbol{Y},$$

whose dimension $(Np + q)$ diverges as subject size N increases.

Note that if there are no population-shared predictors, minimizing (3.2) is the same as minimizing the objective function for each individual (subject) separately. We call this approach subject-wise modeling; however, it only utilizes within-subject information. As a result, this leads to inefficient estimation and over-fitting of a model, especially when the sample size N is large and m is relatively small.

3.2.2 The proposed model with multi-directional separation penalty

We propose a novel penalized variable selection approach by providing multiple shrinking directions for individualized parameters and utilizing homogeneity information within the subpopulation, which performs parameter estimation, variable selection and subgrouping simultaneously.

For the kth ($k = 1, \ldots, p$) individualized predictor corresponding to the ith subject, we assume that there are $G_k + 1$ subgroups in the population such that

$$\beta_{ik} = \begin{cases} \boldsymbol{\gamma}_k^{(g)}, & \text{if} \quad i \in \mathcal{G}_k^{(g)}, \quad g = 1, \ldots, G_k \\ 0, & \text{if} \quad i \in \mathcal{G}_k^{(0)} \end{cases}, \tag{3.3}$$

where $\boldsymbol{\gamma}_k^{(g)}$ $(g \neq 0)$ is an unknown non-zero parameter corresponding to the homogeneous coefficient for the gth subgroup, and $\mathcal{G}_k^{(g)}$'s are the index sets representing the subgroup memberships with respect to the kth predictor.

For ease of notation, in the following, we focus on the setting where there are two subgroups with respect to each individualized covariate: the non-zero-coefficient group ($\beta_{ik} = \gamma_k$) and the zero-coefficient group ($\beta_{ik} = 0$). We denote $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ as the sub-homogeneous effect vector. The extension to multiple subgroups is straightforward.

We first consider a model assuming within-subject independence. The extension to correlated data will be discussed later. The main idea is to encourage grouping of the subjects with similar effects on specific individualized predictors, by inducing the sub-homogeneous effect $\boldsymbol{\gamma}$ in the

proposed objective function

$$Q^{ind}_{N,m}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{N} \parallel \boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel_2^2 + \lambda_{N,m} \sum_{i=1}^{N}\sum_{k=1}^{p} s(\beta_{ik}, \gamma_k), \tag{3.4}$$

where $\lambda_{N,m}$ is the tuning parameter. Here the key part is the proposed multi-directional separation penalty (MDSP) function $s(\beta_{ik}, \gamma_k)$, defined as

$$s(\beta_{ik}, \gamma_k) = \min\Big(|\beta_{ik}|, |\beta_{ik} - \gamma_k|\Big), \tag{3.5}$$

which is a piece-wise L_1-penalization function (Figure 3.1).

The multi-directional penalty term in 3.4 essentially contains a double-summation providing two different perspectives of the proposed model. From a subject's point of view, the penalty term is $\sum_{k=1}^{p} s(\beta_{ik}, \gamma_k)$. In contrast to the traditional penalized variable selection approaches, the proposed MDSP function $s(\cdot)$ provides an alternative shrinking direction in addition to 0. Given γ_k, the $s(\cdot)$ penalty can be viewed as shrinking a weak signal of β_{ik} towards zero, while pulling the strong magnitude signals to γ_k. This reduces the bias for large coefficient estimators introduced by the L_p-penalty. Figure 3.1 illustrates the MDSP function $s(\beta_{ik}, \gamma_k)$ for a given γ_k, and Figure 3.2 provides plots of the thresholding functions of the Lasso and the proposed method. Without loss of generality, we assume $\gamma_k > 0$. Figure 3.2 indicates that when $\beta_{ik} > \gamma_k$ or $\beta_{ik} < 0$, $|\beta_{ik}|$ and $|\beta_{ik} - \gamma_k|$ have the same shrinking effect; and when $0 < \beta_{ik} < \gamma_k$, the two penalties produce different shrinking directions, which separates strong signals from weak signals.

From the other perspective, for one individualized predictor over different subjects, the MDSP term is $\sum_{i=1}^{N} s(\beta_{ik}, \gamma_k)$. Given $\beta'_{ik}s$, the proposed method leads to subgrouping the coefficients of individuals, where the separation-penalty term serves the role of centering, similar to K-means clustering. Compared to pairwise grouping penalization such as the fusion penalty, the center-based one has less computational cost, with $O(Np)$ penalty terms in contrast to the fusion-type of clustering containing $O(N^2p)$ penalty terms. This also implies that the computational cost of the proposed approach increases more slowly as the sample size N increases.

In addition, the unknown true effects γ_k's can be obtained simultaneously through minimizing the objective function in (3.4), where the estimation of γ_k utilizes information from individuals within the subgroup. By pulling the coefficients' estimators towards the center $\hat{\gamma}_k$, it allows us to borrow cross-subject information for individuals' coefficient estimation, and therefore reduces the estimation bias and variance for non-zero coefficients.

Furthermore, the above two-subgroup model can be extended to multiple subgroups and even with additional constraints in practice. We illustrate the extension of three subgroups which allows positive and negative effects of personalized coefficients. The separation penalty imposed for three groups is

$$s(\beta_{ik}, \gamma_k^+, \gamma_k^-) = \min\Big(|\beta_{ik}|, |\beta_{ik} - \gamma_k^+|, |\beta_{ik} - \gamma_k^-|\Big), \quad \text{s.t.} \quad \gamma_k^+ > 0, \quad \gamma_k^- < 0, \tag{3.6}$$

which shrinks the coefficient of the individualized predictor either to zero, a positive effect γ_k^+, or a negative effect γ_k^-.

For correlated data structure, we can incorporate correlations of errors to obtain more efficient estimation ([44]), and introduce within-subject correlations through a weighting matrix $\boldsymbol{V}_i$ to the weighted squared-loss in the objective function

$$Q_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta}))^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) + \lambda_{N,m} \sum_{i=1}^{N} \sum_{k=1}^{p} s(\beta_{ik}, \gamma_k) \tag{3.7}$$

$$= L_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + S_{\lambda_{N,m}}(\boldsymbol{\beta}, \boldsymbol{\gamma}), \tag{3.8}$$

where $\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i \boldsymbol{A}_i^{\frac{1}{2}}$, $\boldsymbol{A}_i$ is a diagonal matrix of marginal variance of $\boldsymbol{y}_i$ and $\boldsymbol{R}_i$ is a working correlation matrix.

3.3 Theoretical Results

In this section, we establish the theoretical properties of the proposed estimator, and the connection to the oracle estimator and the subject-wise least squares estimator. One unique aspect here is that our theory is established under a general double-divergence framework which assumes that both sample size N and cluster size m go to infinity, and therefore the number of individualized parameters also diverges.

We introduce some notation as follows. For any symmetric matrix $\boldsymbol{A}_{n\times n}$, let $\lambda_{min}(\boldsymbol{A})$ and $\lambda_{max}(\boldsymbol{A})$ be the smallest and the largest eigenvalues of $\boldsymbol{A}$, respectively. For an arbitrary matrix $\boldsymbol{A}_{m\times n}(b_{ij})$, denote $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{max}(\boldsymbol{A}^T\boldsymbol{A})}$ as its L_2-norm, $\|\boldsymbol{A}\|_1 = \max_{1\le j\le n}(\sum_{i=1}^m |b_{ij}|)$ as its L_1-norm and $\|\boldsymbol{A}\|_\infty = \max_{1\le i\le m}(\sum_{j=1}^n |b_{ij}|)$ as its L_∞-norm. For a vector $\boldsymbol{a} = (a_1, \ldots, a_n)'$, $\|\boldsymbol{a}\|_2$ reduces to its Euclidean norm and $\|\boldsymbol{a}\|_\infty = \max_{1\le i\le n}(|a_i|)$. Moreover, we denote $\|\boldsymbol{a}\|_0 = \sum_{i=1}^n I_{\{a_i\neq 0\}}$.

In addition, we define the order between two $n \times n$ square matrices as $\boldsymbol{A} > \boldsymbol{B}$ if $\forall \boldsymbol{x} \in \mathbf{R}^n$, $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}$ holds. Let $\boldsymbol{A} \asymp \boldsymbol{B}$ denote $c_1\boldsymbol{A} \le \boldsymbol{B} \le c_2\boldsymbol{A}$ for some constants $0 < c_1 \le c_2 < \infty$. Then we define a sequence of $m \times m$ matrices $\boldsymbol{A_n}$ as $\boldsymbol{A_n} = O(n)$ if $c_1 n\boldsymbol{I}_m \le \boldsymbol{A}_n \le c_2 n\boldsymbol{I}_m$ when n is large. Moreover, let $\boldsymbol{A} \circ \boldsymbol{B}$ denote the entrywise Hadamard product between two same-dimension matrices (see details in Appendix A.1), and "$\otimes$" denote the Kronecker product.

For unbalanced data, we define $\min(m_i) = m$ and assume $m_i = O(m)$ for $1 \le i \le N$. To simplify the notation, we let $m_i = m$ in the following discussion. In addition, without loss of generality, we consider the two-subpopulation structure with respect to each individualized predictor. The theory for a structure with more than two subpopulations can be shown similarly. Let $\mathcal{G}_k \subset \{i : 1 \le i \le N\}$ denote a signal-group index set for the kth individualized predictor such that $\beta_{ik} = \gamma_k \neq 0$ if $i \in \mathcal{G}_k$ and $\beta_{ik} = 0$ otherwise. For any set $\mathcal{G}$, let $|\mathcal{G}|$ be the cardinal of $\mathcal{G}$. Moreover, we denote $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ and let $\boldsymbol{\theta}^0 = ((\boldsymbol{\beta}^0)', (\boldsymbol{\alpha}^0)')'$ be its true value. Let the true value of $\boldsymbol{\beta}_i$ be $\boldsymbol{\beta_i^0} = (\boldsymbol{\beta^0_{i,\mathcal{A}_i}}', \boldsymbol{\beta^0_{i,\mathcal{A}_i^c}}')'$, where $\mathcal{A}_i$ and $\mathcal{A}_i^c$ denote the index sets such that $\boldsymbol{\beta}^0_{i,\mathcal{A}_i} = \boldsymbol{\gamma}^0_{\mathcal{A}_i} \neq \mathbf{0}$ and $\boldsymbol{\beta}^0_{i,\mathcal{A}_i^c} = \mathbf{0}$.

The proposed objective function (3.7) consists of a loss function $L_{N,m}(\cdot)$ and a penalty function

$S_{\lambda_{N,m}}(\cdot)$, where the squared loss function $L_{N,m}(\boldsymbol{\theta})$ in (3.8) can accommodate diverging N and m. Both the oracle estimator and the subject-wise least squares estimator are obtained by minimizing $L_{N,m}(\boldsymbol{\theta})$, but with different design matrices, where the corresponding quasi-likelihood estimating equation is

$$\boldsymbol{G}_{N,m}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{g}_i(\boldsymbol{\theta}) = \sum_{i=1}^{N} \boldsymbol{U}_i(\boldsymbol{\theta})^T \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\theta})) = 0, \tag{3.9}$$

with $\boldsymbol{U}_i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}$. Due to the linear mean function, $\boldsymbol{U}_i(\boldsymbol{\theta})$ does not depend on unknown parameters and thus is suppressed as $\boldsymbol{U}_i$ in the following, and we also denote $\boldsymbol{G}_{N,m} = \boldsymbol{G}_{N,m}(\boldsymbol{\theta}^0)$ for ease of notation. In addition, let

$$\begin{aligned}
\boldsymbol{D}_{N,m} &= -\frac{\partial \boldsymbol{G}_{N,m}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \sum_{i=1}^{N} \boldsymbol{U}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{U}_i, \\
\boldsymbol{H}_{N,m} &= \mathrm{Cov}(\boldsymbol{G}_{N,m}(\boldsymbol{\theta})) = \sum_{i=1}^{N} \boldsymbol{U}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} \boldsymbol{U}_i,
\end{aligned}$$

where $\boldsymbol{\Sigma}_i = \mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i^0 \boldsymbol{A}_i^{\frac{1}{2}}$ and $\boldsymbol{R}_i^0$ is the true correlation matrix. Note that $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ do not depend on unknown mean regression parameter $\boldsymbol{\theta}$. We require some common regularity conditions

(A1) The unknown parameter $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ belongs to a compact subset $\mathcal{B} \subseteq \mathbf{R}^{p_{\boldsymbol{\theta}}}$ and its true value $\boldsymbol{\theta}^0$ lies in the interior of $\mathcal{B}$;

(A2) $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ are positive definite when N or m is large.

Note that the standard assumptions of $\boldsymbol{R}_i$ such as converging to a constant positive definite matrix with eigenvalues bounded away from zero and infinity ([88]) might not be valid in the proposed framework, since the dimension of $\boldsymbol{R}_i$ increases as m increases. Here we only require the following general regularity condition for $\boldsymbol{R}_i$ and $\boldsymbol{R}_i^0$:

(A3) There exist $\nu_l > 0, \nu_l' > 0$, such that $\lambda_{min}(\boldsymbol{R}_i^0) > \nu_l$ and $\lambda_{min}(\boldsymbol{R}_i) > \nu_l'$ for all i and m.

The estimating equation $\boldsymbol{G}_{N,m}(\boldsymbol{\theta})$ contains double summations with the sample size N and the cluster size m, which both can diverge. Consequently, the standard asymptotic results for M-

estimators are not applicable here even with a fixed number of parameters ([91]). In general, for an estimator $\hat{\boldsymbol{\theta}}$ obtained by solving the estimating equation (3.9), under regularity conditions (A1)-(A2), by Taylor's expansion, $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = -\boldsymbol{D}_{N,m}^{-1}\boldsymbol{G}_{N,m}$. This implies that the consistency of $\hat{\boldsymbol{\theta}}$ relies on the following condition on the information matrix $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$,

$(\mathcal{C}_a) \quad \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}) \to \infty.$

In the independent model ($\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$), $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$ reduces to $\boldsymbol{D}_{N,m}$ as $\boldsymbol{H}_{N,m} = \boldsymbol{D}_{N,m}$.

The condition $\mathcal{C}_a$ is a standard condition analogous to [91] condition to establish the weak consistency of a fixed-dimensional GEE estimator. However, in contrast to [91]'s setting, the proposed method results in a diverging dimension of the information matrix which is more complicated. In addition, to utilize subpopulation information, the convergence rates for estimators of different parameters are of great importance and interest in this chapter. The following lemma provides a convergence property for the estimating equation estimator from (3.9).

Lemma 2. *Under regularity condition (A2), for any $\delta > 0$, there exists a solution $\hat{\boldsymbol{\theta}}$ of (3.9) such that*

$$P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2 > \delta\right) < \tfrac{1}{\delta^2},$$

where $p_{\boldsymbol{\theta}}$ is the dimension of $\boldsymbol{\theta}$. Moreover, if condition ($\mathcal{C}_a$) holds, we have

$$P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2 > \delta\right) \to 0.$$

Lemma 2 presents the consistency result under all settings. It indicates that the estimator's convergence rate depends on the divergence rate of $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}$'s eigenvalues.

Remark 1. Note that Lemma 2 provides consistency under the spectral norm (L_2-norm). For any fixed-dimensional estimator, for example, the oracle estimator and the subject-wise estimator when N is bounded, the consistency in Lemma 2 is equivalent to $P\left(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_\infty > \delta\right) \to 0$. However, if $p_{\boldsymbol{\theta}}$ is diverging, we need additional conditions to ensure the stronger consistency under the L_∞-norm. More discussion will be provided later regarding the proposed estimator when $N \to \infty$.

In addition, we assume that a few general regularity conditions hold for the design matrix,

(A4) $\tilde{\boldsymbol{X}}_{ij} = (\boldsymbol{X}'_{ij}, \boldsymbol{Z}'_{ij})'_{(p+q)\times 1}$ belongs to a compact set $\mathcal{X} \subset \mathbf{R}^{p+q}$ for $1 \leq i \leq N$ and $1 \leq j \leq m$;

(A5) Let $\tilde{\boldsymbol{X}}_{i,k}$ denote the kth column of $\tilde{\boldsymbol{X}}_i$, assume $\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(m)$ and $\sum_{i=1}^N m^{-1}\|\tilde{\boldsymbol{X}}_{i\cdot,k}\|_2^2 = O_p(N)$, for $1 \leq k \leq p+q$;

(A6) $m^{-1}\lambda_{min}(\boldsymbol{X}_i^T\boldsymbol{X}_i) > c_3$ for any i and $\frac{1}{Nm}\lambda_{min}\left(\sum_{i=1}^N \boldsymbol{Z}_i^T(\boldsymbol{I}_m - \boldsymbol{H}_{\boldsymbol{X}_i})\boldsymbol{Z}_i\right) > c_4$, where $\boldsymbol{H}_{\boldsymbol{X}_i} = \boldsymbol{X}_i(\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T$, for some constants $0 < c_3 < \infty$, $0 < c_4 < \infty$.

Conditions (A4)-(A6) are regularity conditions which are typically required for the bounded regressors. However, these are less restrictive than other assumptions, e.g., $\frac{1}{m}\boldsymbol{X}_i^T\boldsymbol{X}_i$ converges to a positive constant matrix. Note that condition (A6) allows within-subject invariant covariates, and is less restrictive since it does not require $\tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i$ to be positive definite.

The regularity conditions (A1)- (A6) are assumed to hold in this section. In Condition (A2) and Lemma 2, matrices $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ represent a general form according to the estimating equation (3.9). For different estimators using the same data, for example, the oracle estimator or the subject-wise estimator, $\boldsymbol{D}_{N,m}$ and $\boldsymbol{H}_{N,m}$ can be different due to their different formulating.

3.3.1 Asymptotic results for the oracle estimator with group effects

In the proposed framework, the oracle estimator assumes that all subpopulation information ($\mathcal{G}_k$, $1 \leq k \leq p$) with respect to the individualized predictors is known. This is equivalent to assuming that the true signal sets $\mathcal{A}_i$'s ($1 \leq i \leq N$) for all subjects are known.

The individualized parameter $\boldsymbol{\beta_i}$ for each subject is linked to the sub-homogeneous parameter $\boldsymbol{\gamma}$ as $\boldsymbol{\omega_i} \circ \boldsymbol{\gamma} = \boldsymbol{\beta_i}$ through an indicator vector $\boldsymbol{\omega_i} = (\omega_{i1}, \ldots, \omega_{ip})' \in \mathbf{R}^p$, where $\omega_{ik} = I_{\{i\in\mathcal{G}_k\}} = I_{\{k\in\mathcal{A}_i\}}$. Hence there exists a mapping linking two parameter spaces, which is $\mathbf{R}^p \mapsto \mathbf{R}^{Np} : \boldsymbol{\Omega\gamma} = \boldsymbol{\beta}$, where $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_N)'$ is a $Np \times p$ matrix and $\boldsymbol{\Omega}_i = \text{diag}(\boldsymbol{\omega}_i)$ is a diagonal matrix. We define $L^{or}_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = L_{Nm}(\boldsymbol{\alpha}, \boldsymbol{\beta}(\boldsymbol{\gamma}))$. By noting that $S_{\lambda_{N,m}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ with $\boldsymbol{\beta} = \boldsymbol{\Omega\gamma}$ and $\boldsymbol{\Omega}$ is

known, the oracle estimator can be obtained by minimizing $L^{or}_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ as

$$\left(((\hat{\boldsymbol{\gamma}}^{or})', \hat{\boldsymbol{\alpha}}^{or})'\right)' = \underset{\boldsymbol{\alpha},\boldsymbol{\gamma}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\boldsymbol{y}_i - \boldsymbol{X_i}(\boldsymbol{\omega}_i \circ \boldsymbol{\gamma}) - \boldsymbol{Z_i}\boldsymbol{\alpha}\right)^T \boldsymbol{V}_i^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X_i}(\boldsymbol{\omega}_i \circ \boldsymbol{\gamma}) - \boldsymbol{Z_i}\boldsymbol{\alpha}\right).$$

The oracle individualized estimator for each subject is obtained by $\hat{\boldsymbol{\beta}}_i^{or} = \boldsymbol{\omega}_i \circ \hat{\boldsymbol{\gamma}}^{or}$.

Let $\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\tilde{\boldsymbol{\omega}}_i = (\boldsymbol{\omega}_i', \mathbf{1}_q')'$, and $\tilde{\boldsymbol{X}}_i^{or} = \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{\Omega}}_i$ where $\tilde{\boldsymbol{\Omega}}_i = \operatorname{diag}(\tilde{\boldsymbol{\omega}}_i)$. We denote $\boldsymbol{H}^{or}_{N,m} = \sum_{i=1}^{N} (\tilde{\boldsymbol{X}}_i^{or})^T \boldsymbol{V}_i^{-1} \boldsymbol{\Sigma}_i \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i^{or}$, $\boldsymbol{D}^{or}_{N,m} = \sum_{i=1}^{N} (\tilde{\boldsymbol{X}}_i^{or})^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i^{or}$, and Lemma 2 directly applies for the oracle estimator by replacing $\boldsymbol{H}_{N,m}$ and $\boldsymbol{D}_{N,m}$ with $\boldsymbol{H}^{or}_{N,m}$ and $\boldsymbol{D}^{or}_{N,m}$, respectively. Let $\hat{\boldsymbol{\theta}}^{or} = \left((\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})'\right)'$ and $\tilde{\boldsymbol{\theta}}^0 = \left((\boldsymbol{\gamma}^0)', (\boldsymbol{\alpha}^0)'\right)'$, according to Lemma 2 we have

$$(\boldsymbol{H}^{or}_{N,m})^{-\frac{1}{2}} (\boldsymbol{D}^{or}_{N,m})(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0) = O_p(1). \tag{3.10}$$

Note that the divergence rates of $\boldsymbol{H}^{or}_{N,m}$ and $\boldsymbol{D}^{or}_{N,m}$ are associated with the subpopulation size $|\mathcal{G}_k|$'s as N goes to infinity. However, in contrast to other clustering approaches based on an entire set of coefficient vector $\boldsymbol{\beta}_i$ (e.g., [57]; [47]), the proposed model allows the subgroup partitions corresponding to different individualized predictors to be different. Therefore the design matrix for the oracle estimator here cannot be formulated as a block diagonal form, which leads to non-trivial subgroup effects on divergence rates.

Remark 2. A few comments about the eigenvalues of the matrices are worth mentioning. For two square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same dimension, $\boldsymbol{AB}$ and $\boldsymbol{BA}$ have the same non-zero eigenvalues. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular and $\boldsymbol{A} \leq \boldsymbol{B}$, for any matrix $\boldsymbol{C}$ we have $\boldsymbol{C}^T\boldsymbol{AC} \leq \boldsymbol{C}^T\boldsymbol{BC}$, and $\boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$. The proofs of these results are provided in Section 3.8.

To get a better understanding of the group effects on the oracle estimator, we reformulate $\boldsymbol{D}^{or}_{N,m} = \sum_{i=1}^{N} \tilde{\boldsymbol{\Omega}}_i^T \tilde{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i \tilde{\boldsymbol{\Omega}}_i = \sum_{i=1}^{N} (\tilde{\boldsymbol{\Omega}}_i \tilde{\boldsymbol{\Omega}}_i^T) \circ (\boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)$, where $\tilde{\boldsymbol{\Omega}}_i^T \tilde{\boldsymbol{\Omega}}_i^T$ is a symmetric square matrix with entries to be zero or one. Suppose

(R1). $$\kappa_m^l \leq \lambda_{min}(\tilde{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i) \leq \lambda_{max}(\tilde{\boldsymbol{X}}_i^T \boldsymbol{V}_i^{-1} \tilde{\boldsymbol{X}}_i) \leq \kappa_m^u$$

holds uniformly for any subject i with some positive constant sequences $\{\kappa_m^l\}_{m=1}^{\infty}$ and $\{\kappa_m^u\}_{m=1}^{\infty}$, then we have $\kappa_m^l \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i \leq \boldsymbol{D}_{N,m}^{or} \leq \kappa_m^u \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ by noting $\tilde{\boldsymbol{\Omega}}_i^2 = \tilde{\boldsymbol{\Omega}}_i$. Under a similar condition to (R1), we could show that $\phi_m^l \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i \leq \boldsymbol{H}_{N,m}^{or} \leq \phi_m^u \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ for some positive constant sequences $\{\kappa_m^l\}_{m=1}^{\infty}$ and $\{\kappa_m^u\}_{m=1}^{\infty}$. If $\sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ is non-singular, then

$$(\phi_m^u)^{-1}(\kappa_m^l)^2 \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i \leq \boldsymbol{D}_{N,m}^{or}(\boldsymbol{H}_{N,m}^{or})^{-1}\boldsymbol{D}_{N,m}^{or} \leq (\phi_m^l)^{-1}(\kappa_m^u)^2 \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i. \tag{3.11}$$

Let $\boldsymbol{\Lambda}_{N,m} = \sum_{i=1}^N \tilde{\boldsymbol{\Omega}}_i$ and note that $\boldsymbol{\Lambda}_{N,m} = \text{diag}(N\mathbf{1}_q', |\mathcal{G}_1|, \ldots, |\mathcal{G}_p|)$ is a diagonal matrix, where $|\mathcal{G}_k|$'s ($1 \leq k \leq p$) are signal-subgroup sizes corresponding to p individualized predictors. It is clear that $\boldsymbol{\Lambda}_{N,m}$ contains the group effects on estimation. In particular, the group size for the population-shared parameter is N.

Remark 3. The condition (R1) could be relaxed by replacing $\tilde{\boldsymbol{X}}i$ with $\boldsymbol{X}_i$ since we allow within-subject invariant covariates, especially for the population-shared predictors. Moreover, if m is bounded, it is straightforward to show that $c_l m \leq \kappa_m^l \leq \kappa_m^u \leq c_u m$ and $c_l' m \leq \phi_m^l \leq \phi_m^u \leq c_u' m$ hold for some constants $0 < c_l \leq c_m < \infty$, $0 < c_l' \leq c_m' < \infty$, which immediately implies that $\boldsymbol{D}_{N,m}^{or} \asymp m\boldsymbol{\Lambda}_{\boldsymbol{N},\boldsymbol{m}}$ and $\boldsymbol{H}_{N,m}^{or} \asymp m\boldsymbol{\Lambda}_{\boldsymbol{N},\boldsymbol{m}}$. This conclusion also holds for the independent model even when m goes to infinity.

Let $N_k = \sum_{i \in \mathcal{G}_k} m_i = m|\mathcal{G}_k|$ denote the number of observations in group $\mathcal{G}_k$ and $N_a = \sum_{i=1}^N m_i = mN$ denote the total number of observations. For the independent error model, we establish asymptotic normality for the oracle estimators with convergence rates associated to the sample size N and the cluster size m.

Theorem 2. *Under regularity conditions, suppose $\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$ holds for any i, as either $m \to \infty$ or $\min_{1 \leq k \leq p}(|\mathcal{G}_k|) \to \infty$, we have*

$$(\boldsymbol{H}_{N,m}^{or})^{\frac{1}{2}}\left(\{(\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})'\}' - \{(\boldsymbol{\gamma}^0)', (\boldsymbol{\alpha}^0)'\}'\right) \to_d N\left(\mathbf{0}, \boldsymbol{I}_{p+q}\right),$$

where $\boldsymbol{H}_{N,m}^{or} \asymp \boldsymbol{M}_{N,m}$, and $\boldsymbol{M}_{N,m} = diag(\underbrace{N_1, \ldots, N_p}_{p}, \underbrace{N_a, \ldots, N_a}_{q})$ is a $(p+q) \times (p+q)$-dim diagonal matrix.

Theorem 2 indicates that the convergence rates of the oracle estimator benefit from both increasing N and m, which implies that incorporating subgroup information is able to improve estimation efficiency as we utilize additional number of observations from each subgroup. In addition, Theorem 2 allows both m and N to go to infinity and has no restriction on their divergence rates.

However, in the correlated model with cluster size m diverging, the analysis of the estimator's asymptotic behavior becomes more complicated , since it involves the working correlation matrix $\boldsymbol{R}_i$ and the unknown true correlation matrix $\boldsymbol{R}_i^0$, which makes it difficult to verify the condition $(\mathcal{C}_a)$ and to figure out the estimators' convergence rates.

Similar to [91], we consider a sufficient condition which may simplify the verification and the discussion. Let $\eta_{N,m} = \max_{1\leq i\leq N}\{\lambda_{max}(\boldsymbol{R}_i^{-1}\boldsymbol{R}_i^0)\}$, an alternative condition for consistency is

$(\mathcal{C}_a^*)$ $$\eta_{N,m}^{-1}\lambda_{min}(\boldsymbol{D}_{N,m}) \to \infty.$$

The sufficiency of $(\mathcal{C}_a^*)$ that implies $(\mathcal{C}_a)$ is trivial by noting $\boldsymbol{H}_{N,m} \leq \eta_{N,m}\boldsymbol{D}_{N,m}$. Based on (3.10), we present the asymptotic theory for the oracle estimator with the the condition $\mathcal{C}_a^*$.

Theorem 3. *Under regularity conditions, for the oracle estimator $\hat{\boldsymbol{\theta}}^{or} = ((\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})')'$, we have*

$$\eta_{N,m}^{-\frac{1}{2}}\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^0)\|_2 \leq O_p(1),$$

and if $\eta_{N,m}^{-1}\lambda_{min}(\boldsymbol{D}_{N,m}^{or}) \to \infty$, $\hat{\boldsymbol{\theta}}^{or} \to_p \tilde{\boldsymbol{\theta}}^0$.

The proof of Theorem 3 is straightforward by following (3.11) and condition $\mathcal{C}_a^*$. Theorem 3 indicates that the convergence of the estimator depends on the divergence rate of $\eta_{N,m}$ and $\boldsymbol{D}_{N,m}^{or}$. Without considering the group effects, the oracle estimator reduces to a fixed-dimensional GEE estimator by [91] and [2]. Therefore, in the following, we only focus on a few common cases and some useful conditions.

Remark 4. For any N and m, according to regularity condition (A3), note that

$$\eta_{N,m} \leq (\min_{1\leq i\leq N}\{\lambda_{min}(\boldsymbol{R}_i)\})^{-1}\max_{1\leq i\leq N}\{\lambda_{max}(\boldsymbol{R}_i^0)\} \leq (\nu_l')^{-1}\text{tr}(\boldsymbol{R}_1^0) \leq (\nu_l')^{-1}m.$$

If m is bounded, then $\eta_{N,m}$ is bounded, which implies that the condition $\mathcal{C}_a^*$ does not depend on unknown true correlation structure $\boldsymbol{R}_i^0$. As $N \to \infty$, we have $\lambda_{min}(\boldsymbol{D}_{N,m}^{or}) \to \infty$ regardless of the choice of working correlation $\boldsymbol{R}_i$. Hence, similar to standard results for the GEE estimator, the oracle estimator $\hat{\boldsymbol{\theta}}^{or}$ has asymptotic normality, although it may not achieve optimal efficiency if $\boldsymbol{R}_i \neq \boldsymbol{R}_i^0$.

Remark 5. If $m \to \infty$, $\eta_{N,m}$ is not always bounded. For example, if $\boldsymbol{R}_i^0$ admits an exchangeable correlation structure and we choose working correlation $\boldsymbol{R}_i$ as an identity matrix, we have $\eta_{N,m} = O(m)$. For any bounded N, $\boldsymbol{D}_{N,m}^{or} = O(m)$, which implies that the condition ($\mathcal{C}_a^*$) fails. Although the condition ($\mathcal{C}_a$) may still hold with some constraints on the design matrix to ensure consistency (see following Example 1), the convergence rate could be slower than the optimal rate $\sqrt{m}$ and it may not converge to a normal distribution asymptotically ([91]).

We use the following example of a simple linear regression to illustrate some details about the conditions $\mathcal{C}_a$ and $\mathcal{C}_a^*$ with specific covariates design.

Example 1. Consider a subject-wise model with homogeneous effect,

$$y_{ij} = x_{ij}\beta + \varepsilon_{ij}, \quad i = 1, \ldots, N; j = 1, \ldots, m,$$

where $\boldsymbol{\varepsilon_i} = (\varepsilon_{i1}, \ldots, \varepsilon_{im})' \sim N(\mathbf{0}, \sigma^2 \boldsymbol{R^0})$ and $\boldsymbol{R^0}$ admits an exchangeable structure with parameter $\rho > 0$, x_{ij}'s are iid $N(\mu, 1)$. For the case of bounded N, without loss of generality, we assume $N = 1$. By using an independent working correlation $\boldsymbol{R}_i = \boldsymbol{I}_m$, we have $\boldsymbol{D}_m = \boldsymbol{x}_1^T \boldsymbol{x}_1 = O(m)$ and $\eta_m = \lambda_{max}(\boldsymbol{R}^0) = m\rho + 1 - \rho$, where $\boldsymbol{x}_1 = (x_{11}, \ldots, x_{1m})'$. Thus condition $\mathcal{C}_a^*$ fails. However, note that $\boldsymbol{R}^0(\rho) = (1-\rho)\boldsymbol{I}_m + \rho \mathbf{1}_m \mathbf{1}_m^T$. We have $\boldsymbol{H}_m = \sigma^2 \boldsymbol{x}_1^T \boldsymbol{R}^0 \boldsymbol{x}_1 = \sigma^2 \boldsymbol{x}_1^T((1-\rho)\boldsymbol{I}_m + \rho \mathbf{1}_m \mathbf{1}_m^T)\boldsymbol{x}_1^T = \sigma^2(1-\rho)\boldsymbol{x}_1^T \boldsymbol{x}_1 + m\rho(m^{-\frac{1}{2}} \sum_{i=1}^m x_{1j})^2 = O(m) + O(m)$ if $\mu = 0$, and thus $\lambda_{min}(\boldsymbol{D}_m \boldsymbol{H}_m^{-1} \boldsymbol{D}_m) = O(m) \to \infty$ as $m \to \infty$. But if $\mu > 0$, it is clear that $m\rho(m^{-\frac{1}{2}} \sum_{i=1}^m x_{1j})^2 = O(m^2)$ and thus $\lambda_{min}(\boldsymbol{D}_m \boldsymbol{H}_m^{-1} \boldsymbol{D}_m) = O(1)$.

Corollary 1. *Suppose $\eta_{N,m} \leq C_1$ holds uniformly for some constant $0 < C_1 < \infty$, under regularity conditions, we have*

$$\|\boldsymbol{M}_{N,m}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or} - \tilde{\boldsymbol{\theta}}^{0})\|_2 \leq O_p(1),$$

where $\boldsymbol{M}_{N,m}$ is defined in Theorem 2.

The condition of uniformly bounded $\eta_{N,m}$ in Corollary 1 naturally holds when m is bounded or for the independent model. However, as m goes to infinity, it implies that either we choose a working correlation matrix $\boldsymbol{R}_i$ close to the true one, or the correlation is not too strong. The first case involves a consistent and efficient estimator of the correlation structure, which has been discussed in [**?**], [33] and [24]. For the second case, a variety of conditions can be imposed on the correlation structures to ensure a "weak" dependency.

In the following, we provide a sufficient condition which can be verified easily in practice. For an arbitrary correlation matrix $\boldsymbol{R}_{m\times m}(\rho_{ij})$, assume

($\mathcal{R}_a$) $\quad |\rho_{ij}| \leq \rho_{|i-j|}$ for $i \neq j$ and $\sum_{k=1}^{\infty} \rho_k < \infty$.

We show in the Appendix that if condition ($\mathcal{R}_a$) holds for the true correlation matrix $\boldsymbol{R}_i^0$, then $\eta_{N,m}$ is bounded uniformly for any working correlation structures. This indicates that $\boldsymbol{R}_i^0$ is bounded as the within-subject correlation decays rapidly as m increases. In practice, a wide family of correlation structures satisfy the conditions ($\mathcal{R}_a$) including the AR-1 and the M-dependent correlation matrices.

3.3.2 Asymptotic results for the proposed estimator

In general, the least squares estimator plays an important intermediate role in investigating the large sample theory of the penalized estimator. Hence, prior to presenting the theoretical results for the proposed estimator, we provide the asymptotic theory for the subject-wise least squares estimator $\hat{\boldsymbol{\theta}}^{Sub} = ((\hat{\boldsymbol{\beta}}^{Sub})', (\hat{\boldsymbol{\alpha}}^{Sub})')'$ obtained by minimizing $L_{N,m}(\boldsymbol{\theta})$.

Note that, for the proposed estimator and the subject-wise least squares estimator, each term of $\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{U}_i$ in $\boldsymbol{D}_{N,m}$ does not equal to $\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, but is a block sparse, matrix as $\boldsymbol{\mu}_i$ does not contain any other individualized parameter $\boldsymbol{\beta}_j$ for $j \neq i$. We denote

$$\boldsymbol{D}^s_{N,m} = \begin{pmatrix} \boldsymbol{D}^s_{xx}(Np \times Np) & \boldsymbol{D}^s_{xz}(Np \times q) \\ \boldsymbol{D}^s_{zx}(q \times Np) & \boldsymbol{D}^s_{zz}(q \times q) \end{pmatrix},$$

for the subject-wise estimator, where $\boldsymbol{D}^s_{xx} = \text{bdiag}\Big(\{\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\}_{i=1}^N\Big)$ and bdiag$(\cdot)$ denotes a block-diagonal matrix. Similarly, we have $\boldsymbol{H}^s_{xx} = \text{bdiag}\Big(\{\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\}_{i=1}^N\Big)$ in $\boldsymbol{H}^s_{N,m}$ (see Appendix for details), and both $\boldsymbol{D}^s_{xx}$ and $\boldsymbol{H}^s_{xx}$ will expand as N increases. Following Lemma 2, we obtain the following result:

Lemma 3. *Under regularity conditions, for any $\delta > 0$ and $\boldsymbol{a} \in \mathbf{R}^{Np+q}$, we have*

$$P\Big(|\boldsymbol{a}^T(\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0)|^2 > \delta\Big) \leq \delta^{-2}\boldsymbol{a}^T(\boldsymbol{D}^s_{N,m}(\boldsymbol{H}^s_{N,m})^{-1}\boldsymbol{D}^s_{N,m})^{-1}\boldsymbol{a}.$$

If we choose $\boldsymbol{a}$ as a coordinate indicator for $\boldsymbol{\beta}_i$ in $\boldsymbol{\theta}$, that is, $\boldsymbol{a} = (\mathbf{0}_q', \boldsymbol{a}_1', \ldots, \boldsymbol{a}_N')'$, where $\boldsymbol{a}_j \in \mathbf{R}^p$, $1 \leq j \leq N$, $\boldsymbol{a}_j = \mathbf{1}_p$ if $j = i$ or $\boldsymbol{a}_j = \mathbf{0}_p$ if $j \neq i$, Lemma 3 implies the following corollary, which provides a detailed view of the convergence property for each subject-wise estimator $\hat{\boldsymbol{\beta}}_i^{Sub}$ and the population-shared estimator $\hat{\boldsymbol{\alpha}}^{Sub}$.

Corollary 2. *Under regularity conditions, for any $\delta > 0$ and individualized estimator $\hat{\boldsymbol{\beta}}_i^{Sub}$,*

$$P\Big(\|\hat{\boldsymbol{\beta}}_i^{Sub} - \boldsymbol{\beta}_i^0\|_2 > \delta\Big) \leq p\delta^{-2}\eta_{Nm}\lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{X}_i})^{-1},$$

where $\boldsymbol{D}^s_{\boldsymbol{X}_i} = \boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, $i = 1, \ldots, N$, and for the population-shared estimator $\hat{\boldsymbol{\alpha}}^{Sub}$,

$$P\Big(\|(\hat{\boldsymbol{\alpha}}^{Sub} - \boldsymbol{\alpha}^0)\|_2 > \delta\Big) \leq q\delta^{-2}\eta_{Nm}\lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{Z}})^{-1},$$

where $\boldsymbol{D}^s_{\boldsymbol{Z}} = \sum_{i=1}^N \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i$.

Note that the condition $((C)_a)$ requires that $m \to \infty$. In the case of bounded m and diverging N, it is straightforward that the consistency of any individualized parameter cannot be achieved since $\lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{X}_i})$ does not diverge. Intuitively, the increasing number of subjects does not accumulate additional information for the subject-wise parameters. However, the estimator of

population-shared parameter $\hat{\boldsymbol{\alpha}}$ could still be consistent as $N \to \infty$ by noting that η_{Nm} is bounded and $\lambda_{min}(\boldsymbol{D}^s_{\boldsymbol{Z}}) \to \infty$.

Lemma 3 and Corollary 2 provide consistent estimations under the L_2-norm, which depend on the dimension of parameters. Furthermore, we pursue a stronger uniform consistency with additional conditions on either the random errors' distributions or the divergence rates of N and m. In addition to the basic assumptions of zero mean and finite second moment σ^2 for random error ε_{ij}'s, let $\boldsymbol{\varepsilon}^*_{\boldsymbol{i}} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}_{\boldsymbol{i}}$ and denote $\tau^s_{N,m} = \lambda_{min}(\boldsymbol{D}^s_{N,m}(\boldsymbol{H}^s_{N,m})^{-1}\boldsymbol{D}^s_{N,m})$

($\mathcal{I}_a$) $\quad N = o(\tau^s_{N,m})$,

($\mathcal{I}_b$) (i) $\boldsymbol{\varepsilon}^*_{\boldsymbol{i}}$ is a sub-Gaussian vector, that is, $\mathrm{P}(|\boldsymbol{a}^T\boldsymbol{\varepsilon}^*_{\boldsymbol{i}}| > t) < 2\exp(-\frac{t^2}{c^2_\sigma\|\boldsymbol{a}\|^2_2})$ for any $\boldsymbol{a} \in \mathbf{R}^m$ and $t > 0$, where c_σ is a positive constant; and (ii) $\log(N) = o(\tau^s_{N,m})$.

In the independent model where $\boldsymbol{\Sigma} = \boldsymbol{I}_m$, the condition (i) in $\mathcal{I}_b$ is equivalent to assuming marginal sub-Gaussian tails for ε_{ij}'s, which is a standard assumption in high-dimensional model. Alternatively, if the random errors are assumed to be normally distributed, then the condition (i) in $\mathcal{I}_b$ holds naturally for both independent and correlated models.

Under condition ($\mathcal{I}_a$) or ($\mathcal{I}_b$), we achieve a stronger uniform consistency for the diverging number of parameters when $N \to \infty$ as $m \to \infty$.

Lemma 4. *Under regularity conditions, if either condition ($\mathcal{I}_a$) or ($\mathcal{I}_b$) is satisfied, for any $\delta > 0$, as $m \to \infty$, we have*

$$P\left(\|\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0\|_\infty > \delta\right) \to 0.$$

Theorem 4 indicates that if N diverges at a limited rate compared to m, or the tails of the random errors' distribution decay fast enough, we are able to achieve a stronger consistency under the L_∞ norm. Note that the $\tau^s_{N,m}$ in conditions ($\mathcal{I}_a$) and ($\mathcal{I}_b$) could also be replaced with $\eta^{-1}_{N,m}\lambda_{min}(\boldsymbol{D}^s_{N,m})$ analogous to the above discussion, which leads to a sufficient condition.

Based on the above conditions and results, we establish the large sample theory for the proposed estimator. We first provide insight into the proposed multi-directional separation penalty. Consider

a simple independent linear regression model for one subject with the objective function

$$Q_{i,m}(\boldsymbol{\beta_i}|\hat{\boldsymbol{\gamma}}) = \frac{1}{2} \parallel \boldsymbol{y}_i - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel_2^2 + \lambda_m \sum_{k=1}^{p} s(\beta_{ik}, \hat{\gamma}_k), \tag{3.12}$$

given an estimator of the sub-homogeneous effects $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \cdots, \hat{\gamma}_p)'$. Therefore, the proposed penalty function $s(\cdot, \hat{\gamma}^k)$ provides an alternative shrinking direction besides zero. The following theorem presents the asymptotic property for the individualized estimator obtained by minimizing (3.12).

Theorem 4. *Under regularity conditions, there exists a local minimizer* $\hat{\boldsymbol{\beta}}_i = (\hat{\boldsymbol{\beta}}'_{i,\mathcal{A}_i}, \hat{\boldsymbol{\beta}}'_{i,\mathcal{A}_i^c})'$ *of (3.12), if* $\lambda_m \to 0$, *as* $m \to 0$, *we have* $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \to_p \boldsymbol{\beta_i^0}$. *In addition, if* $\lambda_m/\sqrt{m} \to \infty$, *suppose* $\sqrt{m}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) = O_p(1)$, *then we have*

$$P(\hat{\boldsymbol{\beta}}_{i,\mathcal{A}_i^c} = 0) \to 1 \qquad \textit{and} \qquad P(\hat{\boldsymbol{\beta}}_{i,\mathcal{A}_i} = \hat{\boldsymbol{\gamma}}_{\mathcal{A}_i}) \to 1.$$

It is worth noting that the condition of consistency on $\hat{\boldsymbol{\gamma}}$ can be relaxed. The proof of Theorem 4 shows that both estimation consistency and selection consistency still hold even if $\hat{\boldsymbol{\gamma}}$ is not consistent. However, if $a_m(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) = O_p(1)$ and $a_m/\sqrt{m} \to \infty$ hold for some a_m, then the estimator $\hat{\boldsymbol{\beta}}_{i\mathcal{A}_i}$ can achieve a faster convergence rate than $\sqrt{m}$, which is optimal for any subject-wise model. In the proposed model, $\hat{\boldsymbol{\gamma}}$ is estimated over different subjects via the subgrouping and gains efficiency from increasing number of subjects N.

In another perspective, we investigate group separation as both N and m go to infinity. Denote $\boldsymbol{B}_{\boldsymbol{\beta}_i^0}(r)$ as a ball in $\mathbf{R}^p$ centered at $\boldsymbol{\beta}_i^0$ with a radius $r > 0$.

Lemma 5. *Suppose either condition* $(\mathcal{I}_a)$ *or* $(\mathcal{I}_b)$ *holds. Under regularity conditions, for any constant* $r > 0$, *as* $\tau_{N,m}^s \to \infty$, *there exists a local minimizer* $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ *of* $Q_{N,m}$ *in (3.7) such that*

$$P\bigg(\bigcap_{1 \le i \le N} \{\hat{\boldsymbol{\beta}}_i \in \boldsymbol{B}_{\boldsymbol{\beta}_i^0}(r)\} \bigcap \{\hat{\boldsymbol{\alpha}} \in \boldsymbol{B}_{\boldsymbol{\alpha}^0}(r)\} \bigcap \{\hat{\boldsymbol{\gamma}} \in \boldsymbol{B}_{\boldsymbol{\gamma}^0}(r)\} \bigg) \to 1.$$

As both sample size N and cluster size m increase, if N diverges at a limited rate, the speed of separation over subjects dominates the speed of increasing subjects. Lemma 5 essentially implies

group identification consistency and thus we obtain more information about the correct direction of the true individualized parameters.

In the spirit of Theorem 4 and Lemma 5, we present the oracle property for the proposed estimator under a general double-divergence setting.

Theorem 5. *Under regularity conditions, suppose either condition ($\mathcal{I}_a$) or ($\mathcal{I}_b$) holds, assuming* $\frac{\lambda_{N,m}}{\tau^s_{N,m}} \to 0$ *and* $\frac{\lambda_{N,m}}{\sqrt{\tau^s_{N,m}}} \to \infty$*, then there exists a local minimizer* $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ *of* $Q_{N,m}$ *in (3.7); as* $\tau^s_{N,m} \to \infty$*, we have*

$$P\left(\{\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T\}^T = \{(\hat{\boldsymbol{\alpha}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T, (\hat{\boldsymbol{\gamma}}^{or})^T\}^T\right) \to 1.$$

Corollary 3 (Uniform selection consistency)**.** *Under the same conditions as in Theorem 5, as* $\tau^s_{N,m} \to \infty$*, we have* $P\left(\bigcap_{i=1}^N \{\hat{\mathcal{A}}_i = \mathcal{A}_i\}\right) \to 1$.

Theorem 5 indicates that the proposed estimator is the same as the oracle estimator, which utilizes the most information. In fact, by providing additional shrinking directions, the proposed model enables us to separate the strong signals from the weak ones. Consequently, we achieve the oracle information about the underlying subpopulation structure, which ensures that the proposed estimator inherits the optimal efficiency from the oracle estimator. From the other perspective, Corollary 3 also implies subgroup identification consistency.

In addition, in the independent error model, by noting $\tau^s_{N,m} = m$, the conditions ($\mathcal{I}_a$) and ($\mathcal{I}_b$) can be simplified as follows:

($\mathcal{I}^*_a$) $\quad N = o(m)$,

($\mathcal{I}^*_b$) $\quad \varepsilon_{ij}$ has sub-Gaussian tails and $\log(N) = o(m)$.

Hence, we have a simplified result for the independent model.

Corollary 4. *Under regularity conditions, if* $\boldsymbol{R}_i = \boldsymbol{R}_i^0 = \boldsymbol{I}_m$*, suppose either condition ($\mathcal{I}^*_a$) or ($\mathcal{I}^*_b$) holds, assuming* $\frac{\lambda_{N,m}}{m} \to 0$ *and* $\frac{\lambda_{N,m}}{\sqrt{m}} \to \infty$*, then there exists a local minimizer* $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T)^T$ *of* $Q_{N,m}$ *in (3.7); as* $m \to \infty$*, we have*

$$P\left(\{\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T\}^T = \{(\hat{\boldsymbol{\alpha}}^{or})^T, (\hat{\boldsymbol{\beta}}^{or})^T, (\hat{\boldsymbol{\gamma}}^{or})^T\}^T\right) \to 1.$$

Combining Theorem 2 and Corollary 4, we have the asymptotic normality of the independent estimator with the optimal efficiency.

The proofs of the theorems and associated lemmas, corollaries and remarks are provided in the Section 3.8.

3.4 Computation

Compared to traditional penalized variable selection methods, the proposed method is more complex to implement since the proposed objective function $Q_{N,m}(\cdot)$ in (3.7) involves an unknown homogeneous effect $\boldsymbol{\gamma}$ in addition to a non-convex penalty function. We propose an iterative algorithm as follows to simplify the optimization process.

3.4.1 Algorithm and convergence property

Note that the first term of the quadratic loss function in (3.7) does not involve the subgroup homogeneous effect $\boldsymbol{\gamma}$. Therefore we first fix $\boldsymbol{\gamma}$ to minimize (3.7) with respect to $\boldsymbol{\beta}, \boldsymbol{\alpha}$. Next, given an estimator of $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}$, we update estimator of $\boldsymbol{\gamma}$ by minimizing the grouping loss through the separation penalty term in (3.7). We iterate these two steps until the algorithm converges. The specific algorithm is described as follows:

In Algorithm 1, under the homogeneous variance assumption, the $\boldsymbol{V}_i$ in the quadratic loss could be replaced by a working correlation matrix $\boldsymbol{R}_i$. Specifically, we recommend one-step moment estimation for the $\boldsymbol{R}_i$ using the subject-wise least squares estimator $\boldsymbol{\beta_i}$ from an independent model.

Note that at Step 3 in Algorithm 1, the objective function (3.13) is a Lasso-type penalized loss function, which is convex. We can solve the optimization problem by using existing algorithms developed for Lasso. In addition, Step 4 can be implemented mimicking K-means algorithm with one subgroup centered at zero.

The following theorem provides the convergence property of Algorithm 2.

Algorithm 2

Step 1. (Initialization) Start with initial estimators: $\hat{\boldsymbol{\beta}}^{(0)}, \hat{\boldsymbol{\alpha}}^{(0)}$, e.g. the OLS or Lasso estimators.
Step 2. Estimate an initial value of $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}^{(0)} = \text{argmin}_{\boldsymbol{\gamma}} \sum_{i=1}^{N} \sum_{k=1}^{p} \min(|\hat{\beta}_{ik}^{(0)}|, |\hat{\beta}_{ik}^{(0)} - \gamma_k|)$.
Step 3. (Penalized Regression) At the nth iteration, given $\hat{\boldsymbol{\gamma}}^{(n-1)}$, update $\hat{\boldsymbol{\beta}}^{(n)}$, $\hat{\boldsymbol{\alpha}}^{(n)}$ via minimizing the objective function:

$$\frac{1}{2}\sum_{i=1}^{N}\Big(\boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha})\Big)^T \boldsymbol{V}_i^{-1}\Big(\boldsymbol{y_i} - \boldsymbol{\mu_i}(\boldsymbol{\beta_i}, \boldsymbol{\alpha})\Big) + \lambda_{N,m}\sum_{i=1}^{N}\sum_{k=1}^{p} s^{(n-1)}(\beta_{ik}, \hat{\gamma_k}^{(n-1)}), \quad (3.13)$$

where $s^{(n-1)}(\beta_{ik}, \gamma_k) = (1 - \hat{\xi}_{ik}^{(n-1)})|\beta_{ik}| + \hat{\xi}_{ik}^{(n-1)}|\beta_{ik} - \gamma_k|$, $\hat{\xi}_{ik}^{(n-1)} = I(|\hat{\beta}_{ik}^{(n-1)}| > |\hat{\beta}_{ik}^{(n-1)} - \gamma_k|)$.
Step 4. (Grouping) Given $\hat{\boldsymbol{\alpha}}^{(n)}$, $\hat{\boldsymbol{\beta}}^{(n)}$, update $\hat{\boldsymbol{\gamma}}^{(n)} = \underset{\boldsymbol{\gamma}}{\text{argmin}} \sum_{i=1}^{N} \sum_{k=1}^{p} \min(|\hat{\beta}_{ik}^{(n)}|, |\hat{\beta}_{ik}^{(n)} - \gamma_k|)$.
Step 5. (Stopping Criterion) Iterate Step 3 and Step 4 until $\| \hat{\boldsymbol{\beta}}^{(n)} - \hat{\boldsymbol{\beta}}^{(n-1)} \|_2 + \| \hat{\boldsymbol{\alpha}}^{(n)} - \hat{\boldsymbol{\alpha}}^{(n-1)} \|_2$ is less than a small predetermined threshold value.

Theorem 6. *For a sequence of estimators $\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)}$ obtained in Algorithm 1, the objective function $Q_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)})$ in (3.7) is non-increasing as the number of iterations m increases, which leads to the convergence of $\hat{\boldsymbol{\beta}}^{(n)}$, $\hat{\boldsymbol{\alpha}}^{(n)}$ and $\hat{\boldsymbol{\gamma}}^{(n)}$.*

However, the iterative estimator may converge to a local minimizer since the objective function is non-convex. Multiple initial values are recommended so that the optimum value can be identified. In fact, the proposed piece-wise convex penalty function produces local minimums corresponding to different subgroups. However, not all individuals are sensitive to initial values except the corresponding coefficients close to boundary. Heuristically, if $\lambda_{N,m}/\gamma_k$ is small, implying that the true effects $\boldsymbol{\gamma}$ are strong, then the coefficient estimators for these individuals are stable. In addition, we recommend a step-wise tuning in practice, that is, we initialize the tuning parameter by a very small value and increase it to the specified value as the number of iterations increases.

3.4.2 Tuning parameter and select number of subgroups

In this chapter, we apply the generalized cross-validation (GCV) method to select an appropriate tuning parameter $\lambda_{N,m}$. The GCV can be regarded as an approximation of leave-one-out cross-validation (CV) and thus provides an approximately unbiased estimator of the prediction error

([57]). The GCV is defined as

$$GCV(\text{df}) = \frac{RSS}{(N_0 - \text{df})^2} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{m_i}(y_{ij} - \hat{y}_{ij})^2}{(N_0 - \text{df})^2},$$

where $N_0 = \sum_{i=1}^{N} m_i$ is the total sample size anddf is the degrees of freedom used in estimating the $\hat{y}_{ij}$'s. In our setting, the degrees of freedom cannot be considered as the number of non-zero parameters, since some of the $\hat{\beta}_{ik}$'s are shrunk to the exact sub-homogeneous effect $\hat{\gamma}_k$. [57] suggest the generalized degrees of freedom (GDF) which is computationally costly. Alternatively, we define the df as the number of homogeneous effects plus the number of remaining non-zero coefficient estimators which are not equal to $\hat{\gamma}_k$'s. To select a tuning parameter $\lambda_{N,m}$, we search from a sequence of grid points which minimizes the GCV.

The proposed method allows a multiple subgroups case as defined in (3.3), and the number of subgroups is usually unknown. In practice, we could specify the number of the subgroups according to known scientific information or a particular target such as exploring the positive effect, the negative effect and no effect.

In practice, we can select the number of subgroups based on a data-driven approach. One approach is to adopt the idea of the jump statistic ([79]) with a K-means clustering based on some pre-estimators, e.g., the subject-wise least squares estimator. This is easy to implement but might not be reliable, as in the two-step procedure, the pre-estimators are treated as observed responses which do not change as the number of subgroups changes.

Here we provide the modified Bayesian Information Criterion (BIC, [87]) for high-dimensional data settings to select the number of subgroups. We use one individualized covariate as an illustration. The number of subgroups G_k is selected by minimizing

$$\text{BIC}(G_k) = \log\left(\sum_{i=1}^{N}\sum_{j=1}^{m}(y_{ij} - \hat{\mu}_{ij}(G_k))^2/(mN)\right) + b_{N,m}\frac{\log(mN)}{mN}(G_k + q), \tag{3.14}$$

where $b_{N,m}$ is a positive number and depends on N and m. When $b_{N,m} = 1$, the modified BIC reduces to the traditional BIC ([72]). For the high-dimensional setting, we follow [86] with $b_{N,m} =$

$c\log(\log(p_{\boldsymbol{\theta}}))$, where $p_{\boldsymbol{\theta}} = N + q$ and $c = 2$. To extend to more than one individualized covariate, we adopt a strategy of selecting the number of subgroups for one predictor while fixing other individualized coefficients with the subject-wise least squares estimators.

3.5 Numerical Study

In this section, we provide simulation studies to investigate the numerical performance of the proposed method in finite samples. Specifically, we compare the proposed model with the subject-wise model, the homogeneous model and five other regularization models in Section 3.5.1. In addition, we demonstrate the benefit of incorporating within-subject correlations. In Section 3.5.2, we investigate the subgroup number selection of the proposed model and test the robustness against model misspecification.

3.5.1 Individualized regression with correct-specified subgroup numbers

In this simulation study, we simulate two cases to evaluate the proposed model when the number of subgroups is correctly specified. In the first case, we consider a heterogeneous regression model with one individualized variable and two population-shared variables:

$$y_{ij} = \alpha_0 + \alpha_1 z_{ij1} + \alpha_2 z_{ij2} + \beta_i x_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m. \tag{3.15}$$

We set the sample size $N = 40, 100$, and the cluster size $m = 10, 20$. The individualized coefficients are set as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)' = (\underbrace{\gamma, \ldots, \gamma}_{N/2}, \underbrace{0, \ldots, 0}_{N/2})'$, where γ is the true subgroup homogeneous effect chosen as 1 or 2, and the population parameters are $\boldsymbol{\alpha}' = (\alpha_0, \alpha_1, \alpha_2) = (1, 1, 1)$. The covariates z_{ij1}, z_{ij2} and x_{ij} are generated from $N(0, 1)$. The random error ε_{ij}'s are independently generated from $N(0, 1)$.

We compare the performance of the proposed model (MDSP) with five regularized variable selection approaches, namely, the Lasso ([81]) implemented by R package *glmnet* (version 2.0-2) ([19]), the adaptive Lasso (AdapL) ([100]) solved by R package *parcor* (version 0.2-6) ([39]), the

SCAD ([17]) and the MCP ([95]) implemented by R package *ncvreg* (version 3.5-1) ([8]), and the fused Lasso (FusedL) ([83]) solved by R package *penalized* (version 0.9-50) ([22]). Note that there are $N+3$ variables and Nm observations for the above five conventional regularization models. For the fused Lasso, we order estimators of the individualized coefficients based on the least squares estimation as the fused Lasso only imposes L_1-penalties on adjacent coefficients. In addition, we also compare two non-variable-selection models, namely, the heterogeneous model (Sub) assuming subject-wise coefficients β_i's, and the homogeneous model (Homo) assuming homogeneous effect $\beta_i = \beta_h$ $(i = 1, \ldots, N)$. Both of them are based on the least squares estimators.

To evaluate the performance of these approaches on individual variable selection and prediction, we calculate the correct variable selection rate (CVSR), sensitivity and specificity, and the root mean square error (RMSE) for coefficient estimators, where the correct variable selection rate (CVSR) of the individualized variable is defined as the rate of correctly classifying β_i's $(i = 1, \ldots, N)$ to be either zero or non-zero among all individuals, and sensitivity and specificity are the true positive rate $P(\hat{\beta}_i \neq 0 | \beta_i \neq 0)$ and the true negative rate $P(\hat{\beta}_i = 0 | \beta_i = 0)$, respectively. The root mean square error is defined as $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_2$, where $\boldsymbol{\beta}^0 = (\beta^0_{i1}, \ldots, \beta^0_{iN})'$ are the true values.

Table 3.1 provides the mean of root mean square errors (RMSE) based on 100 simulations. Figures 3.3 and 3.4 are the boxplots of the RMSE for all approaches. The proposed method has the smallest RMSE in all settings, which has an improvement of at least 20% $(m = 10)$ and 71% $(m = 20)$ compared to other methods for both sample sizes $N = 40, 100$ when $\gamma = 1$. The improvement is more significant and reaches 150% $(m = 10)$ and 250% $(m = 20)$ when subgroups are separated well $(\gamma = 2)$. This is because that the proposed method is able to borrow strength from different individuals within the same subgroup in estimating individualized coefficients.

The CVSR, sensitivity and specificity for the above simulations are summarized in Table 3.3. The proposed method (MDSP) clearly outperforms the other conventional penalization approaches in terms of the highest CVSR, especially when the subgroup homogeneous effect is large $(\gamma = 2)$. Although all models achieve similar rates on sensitivity, the proposed model leads to higher

specificity rates. Figures 3.5–3.8 provide the boxplots of CVSR, sensitivity and specificity for all of the variable selection approaches.

In addition, Table 3.2 summarizes the estimators and the empirical standard errors of the subgroup homogeneous effects γ from the proposed model. Specifically, the estimators $\hat{\gamma}$'s are consistent as the cluster size m increases. The estimators of the population-shared coefficients $\hat{\boldsymbol{\alpha}}$ are quite similar for all methods and thus are omitted here.

In the second simulation case, we consider a subject-wise model of two individualized predictors with serial correlations:

$$y_{ij} = \beta_{i1}x_{ij1} + \beta_{i2}x_{ij2} + \varepsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, \ldots, m.$$

The individualized coefficients $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{N1})'$ and $\boldsymbol{\beta}_2 = (\beta_{12}, \ldots, \beta_{N2})'$ are

$$\boldsymbol{\beta}_1 = (\underbrace{\gamma_1, \ldots, \gamma_1}_{N/2}, \underbrace{0, \ldots, 0}_{N/2}), \qquad \boldsymbol{\beta}_2 = (\underbrace{0, \ldots, 0}_{N/2}, \underbrace{\gamma_2, \ldots, \gamma_2}_{N/2}),$$

where $\gamma_1 = 1$ and $\gamma_2 = -2$. We choose the sample size $N = 20, 80$ and the cluster size $m = 10, 20$. The covariates x_{ij1} and x_{ij2} are generated from $N(0, 1)$. The random error $\boldsymbol{\varepsilon}_{\boldsymbol{i}}' = (\varepsilon_{i1}, \ldots, \varepsilon_m)$'s are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\sigma^2\boldsymbol{R}(\rho)$, where $\boldsymbol{R}(\rho)$ is the correlation matrix which has either an AR-1 or exchangeable structure. We set $\sigma = 1$ and $\rho = 0.5$.

We compare the performance of the proposed model using different working correlation structures to the independent model. Table 3.4 summarizes the average root mean square errors (RMSE) based on 100 simulations under various settings. Overall, the proposed model utilizing within-subject correlation information achieves smaller RMSE than the independent model. In particular, if the correct working structure is correctly specified, the RMSE can be reduced at least 40% compared to the one obtained using independent structure.

3.5.2 Subgroup number selection and robustness

In this simulation study, we first investigate the performance of the data-driven method discussed in Section 3.4 to select the number of shrinkage centers (subgroups). We compared the proposed method (MDSP) based on BIC-type criterion with a two-stage approach (OLSK) which employs the gap statistic ([82]) to choose the number of subgroups for the K-means algorithm based on the least squares estimators of individualized coefficients. The OLSK method is implemented by R package *cluster* (version 2.0.5) ([48]). The number of bootstrap samples in calculating the gap statistic is set as 100.

We generate the data following (3.15) under various scenarios. Scenario 1 has only a noise individualized variable ($\beta_i = 0, i = 1, \ldots, N$), while Scenarios 2 and 3 have two ($\beta_i = 0, 1$) or three subgroups ($\beta_i = 0, 2, 5$) for one individualized predictor, respectively, and Scenario 4 assumes a model of two individualized predictors with two ($\beta_{i1} = 0, 2$) or three ($\beta_{i2} = 0, -2, 1$) subgroups, respectively. The subgroup size in each scenario is balanced.

Table 3.5 provides the mean estimated number of subgroups and proportion of selecting the correct number of subgroups based on 100 replications. Overall, the proposed method is able to select the correct number of subgroup with more than 85% probability over all scenarios with different sample sizes ($N = 60, 120$) and cluster sizes ($m = 5, 10, 20$). The chance of selecting the correct number of subgroups increases as the cluster size increases. In addition, the proposed method consistently outperforms the two-stage OLSK method, especially when the cluster size is small ($m = 5$).

Next we test the robustness of the proposed model when the number of subgroups is mis-specified. We generate the data as in model (3.15) under two scenarios: one has a population homogeneous predictor ($\beta_i = \gamma = 2, i = 1, \ldots, N$) and the other generates an individualized variable with three subgroups ($\gamma_0 = 0, \gamma_1 = -3, , \gamma_2 = 1$) with balanced size. We set the sample size $N = 60$ and the cluster size $m = 10$. For both cases, we fit the proposed model assuming two subgroups ($\beta_i = 0, \gamma$).

Table 3.6 provides the mean of RMSE and CVSR for the proposed method, the subject-wise model and the five other regularized methods described in Section 3.5.1. In general, the proposed method is robust against the misspecification of subgroup numbers. In the case of homogeneous effect, all models perform similarly in selecting the true variable for all individuals. However, the proposed method has the smallest RMSE among all methods with a 170% reduction. In addition, in the case when there are fewer assumed subgroups than is true, the proposed method still has the best correct variable selection rate, and reduces the RMSE at least 14% compared to the other methods.

Figure 3.9 illustrates the estimation of individualized coefficients from the proposed model. In the setting where the true effect is homogeneous with individuals separate from zero, all subjects are identified correctly as one group, and are shrunk towards a non-zero group. In the scenario with three true subgroups, the subgroup with a relatively stronger signal ($\gamma_1 = -3$) is successfully identified, and therefore we gain more estimation efficiency for the individuals in this subgroup. Moreover, the subgroup with the weaker effect ($\gamma_2 = 1$) is shrunk towards zero since it is the only other shrinking direction we provide, where the proposed estimator is equivalent to the Lasso estimator.

3.6 Real Data Application

In this section, we illustrate the proposed individualized variable selection method using the Harvard longitudinal AIDS clinical trial group (ACTG) data. One of the goals from this study is to test the treatment effect of Zidovudine on CD4 cell counts (e.g., [16]). The 140 patients from this study are repeated measured over 14 time points with a missing rate of 8.5% and maintain CD4 counts above 50 at the baseline measures.

The demographic information includes age and gender for each patient. We denote ZDV=1 if the patient receives the treatment and ZDV=0 if the patient is in the control group. Let y_{it} be the CD4 counts for the ith patient at time t. Each individuals' CD4 measurements are standardized by

within-individual standard deviation to achieve a uniform scale. A marginal model to incorporate time, treatment, interaction of time and treatment, age and gender is provided as follows:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{zt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}. \quad (3.16)$$

We are particularly interested in the treatment effect of Zidovudine over time. The standard analysis concludes that the marginal treatment effect over time $\hat{\beta}_{zt}$ is not significant with p-value$= 0.113$.

However, if we examine the time trend of CD4 counts from individuals, there exist subgroups for the treatment group. Given the treatment ZDV, some individuals' CD4 counts are more stable over time while some patients' CD4 counts decrease more rapidly than the average of the control group over time. This could be interpreted that some patients respond more positively, while some respond more negatively, and the remaining patients have no effects from receiving ZDV treatment compared to the average effect of the control group.

Clearly, the subgroup differences are washed out if we apply the above marginal model in (3.16). Therefore, we employ an individualized regression model which accommodates the personalized treatment effects ZDV over time as the following:

$$y_{it} = \beta_0 + \beta_t * Time + \beta_z * ZDV + \beta_{izt} * ZDV * Time + \beta_a * Age + \beta_g * Gender + \varepsilon_{it}.$$

We assume for the β_{izt} coefficient, that it falls into three subgroups ($\beta_{izt} = \gamma^+ > 0$, $\beta_{izt} = \gamma^- < 0$ or $\beta_{izt} = 0$). Note that for patients in the control group, we set $\beta_{izt} = 0$ since their personalized effects corresponding to the treatment are unobserved. Since the treatment variable is constant over time, we compare our proposed method with the subject-wise Lasso model, the standard population homogeneous model, the random-effects model assuming a random slope of ZDV and time interaction and the fused Lasso model.

We choose observations at times $t = 1, \ldots, 12$ as the training set and the remaining observations at $t = 13, 14$ as the testing set. On the testing set, we calculate the root mean square prediction error for each individual at $t = 13, 14$, where the median of the individuals' prediction errors is reported. Table 3.7 shows that the proposed method has the smallest median prediction error among all methods. For example, the proposed method has 16.0%, 13.9% and 18.1% im-

provement in prediction accuracy compared to the marginal model, the random-effects model and the Lasso model, respectively.

Furthermore, Figure 3.10 shows the individuals corresponding to no effect, positive effect and negative effect in the treatment group identified by the Lasso method and the proposed method respectively. The proposed method is able to detect more individuals with significant responses to the treatment than the Lasso method does, as the proposed separation penalty enables us to shrink the estimated coefficients in multiple directions.

To examine whether subgrouping provides more informative treatment effect over time, we refit a marginal regression model in (3.16) for each subgroup, where each subgroup consists of the corresponding individuals identified in the treatment group and all individuals in the control group. Table 3.8 illustrates that the treatment effect over time from the positive-effect subgroup selected by the Lasso method is still not significant, while the negative-effect subgroup is significant with p-value of 0.02. In contrast, the proposed method identifies both positive and negative subgroups with significant p-values of 0.02 and 0.00 respectively.

3.7 Discussion

In this chapter, we consider an individualized regression model where both the number of subjects and the number of subject-wise repeated measurements increase. To select different important predictors for different individuals, we propose a novel multi-directional separation penalty to implement individualized variable selection. In addition, by utilizing subpopulation structure, we induce within-subgroup homogeneous effects and borrow cross-subject information to achieve a good balance of parsimonious modeling and heterogeneous interpretation.

In contrast to the conventional penalized variable selection approaches, the proposed method provides multiple shrinking directions to overcome estimation bias from L_1-regularization, where the alternative shrinking directions in addition to zero are automatically selected through grouping of subjects with similar effects from predictors. Consequently, for any subject, the proposed

model achieves estimation consistency and selection consistency, even with the L_1-penalty on each shrinking direction.

In addition, compared to subject-wise modeling, the proposed method is able to achieve the population-wise oracle property when the number of the individualized parameters increases along with the sample size. Consequently, the proposed estimator inherits the optimal convergence rate from the oracle estimator due to increasing sizes of within-subject measurements and subgroups. Moreover, by incorporating within-subject serial correlation, the proposed method is able to gain more efficiency than the model assuming independence.

In this chapter, the individualized and the population-shared predictors are pre-specified in the model. Therefore it is also essential to develop a method to identify individualized variables from population-shared variables prior to applying the proposed method. One possible solution is to impose an additional penalty on sub-homogeneous effects. In addition, it is worth investigating the possibility of linking subgroup membership to population-shared covariates, such as demographic information, which could be useful for making predictions for new subjects without much prior information.

3.8 Proofs of Theoretical Results

3.8.1 Some Notation and Matrix Algebra

(N1). Denote $a \wedge b = \min(a, b)$.

(N2). Define " $\circ$ " as the Hadamard product, that is, for two matrices, A and B, of the same dimension $m \times n$, then $A \circ B$ is a matrix, of the same dimension of A and B, with elements given by $(A \circ B)_{ij} = A_{ij} \cdot B_{ij}$.

Next, we provide the proofs for some matrix algebra. For two square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with the same dimension,

(M1). $\boldsymbol{AB}$ and $\boldsymbol{BA}$ have the same non-zero eigenvalues.

Proof: For any eigenvalue λ of $\boldsymbol{AB}$, there exists a non-zero vector $\boldsymbol{\mu}$ such that $\boldsymbol{AB\mu} = \lambda\boldsymbol{\mu}$. It implies that $\boldsymbol{BAB\mu} = \lambda\boldsymbol{B\mu}$. Let $\boldsymbol{B\mu} = \boldsymbol{\mu}^*$ and we have $\boldsymbol{BA\mu}^* = \lambda\boldsymbol{\mu}^*$ indicating that λ is also an eigenvalue of $\boldsymbol{BA}$.

(M2). If $\boldsymbol{A}$ and $\boldsymbol{B}$ are non-singular and $\boldsymbol{A} \leq \boldsymbol{B}$, for any matrix $\boldsymbol{C}$, we have $\boldsymbol{C}^T\boldsymbol{AC} \leq \boldsymbol{C}^T\boldsymbol{BC}$, and $\boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$.

Proof: 1) Note that $\boldsymbol{A} \leq \boldsymbol{B}$ is equivalent to $\boldsymbol{x}^T\boldsymbol{Ax} \leq \boldsymbol{x}^T\boldsymbol{Bx}$ for any vector $\boldsymbol{x}$. $\forall$ $\boldsymbol{x}$, denote $\boldsymbol{Cx} = \boldsymbol{x}^*$ such that $\boldsymbol{x}^T\boldsymbol{C}^T\boldsymbol{ACx} = (\boldsymbol{x}^*)^T\boldsymbol{Ax}^* \leq (\boldsymbol{x}^*)^T\boldsymbol{Bx}^* = \boldsymbol{x}^T\boldsymbol{C}^T\boldsymbol{BCx}$, implies $\boldsymbol{C}^T\boldsymbol{AC} \leq \boldsymbol{C}^T\boldsymbol{BC}$.

2) It is trivial that if $\boldsymbol{A} \geq \boldsymbol{I}$, then we have $\boldsymbol{A}^{-1} \leq \boldsymbol{I}$. Hence, $\boldsymbol{A} \leq \boldsymbol{B} \Rightarrow \boldsymbol{B}^{-\frac{1}{2}}\boldsymbol{AB}^{-\frac{1}{2}} \leq \boldsymbol{I} \Rightarrow \boldsymbol{B}^{\frac{1}{2}}\boldsymbol{A}^{-1}\boldsymbol{B}^{\frac{1}{2}} \geq \boldsymbol{I} \Rightarrow \boldsymbol{A}^{-1} \geq \boldsymbol{B}^{-1}$.

3.8.2 Proof of Lemma 2 and Theorem 2

Proof of Lemma 2

For an estimator $\hat{\boldsymbol{\theta}}$ obtained by solving the estimating equation $G_{N,m}(\boldsymbol{\theta}) = 0$ in (3.9), under regularity condition (A2), by Taylor's expansion, we have $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = -\boldsymbol{D}_{N,m}^{-1}\boldsymbol{G}_{N,m}$ and thus $\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) = -\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}$.

By the Chebyshev inequality,

$$
\begin{aligned}
P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)\|_2 > \delta\right) &= P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}\|_2 > \delta\right) \\
&\leq p_{\boldsymbol{\theta}}^{-1}\delta^{-2}E(\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}\|_2^2) \\
&= p_{\boldsymbol{\theta}}^{-1}\delta^{-2}E(tr(\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{G}_{N,m}\boldsymbol{G}_{N,m}^T\boldsymbol{H}_{N,m}^{-\frac{1}{2}})) \\
&= p_{\boldsymbol{\theta}}^{-1}\delta^{-2}tr(\boldsymbol{H}_{N,m}^{-\frac{1}{2}}E(\boldsymbol{G}_{N,m}\boldsymbol{G}_{N,m}^T)\boldsymbol{H}_{N,m}^{-\frac{1}{2}}) \\
&= p_{\boldsymbol{\theta}}^{-1}\delta^{-2}tr(\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{H}_{N,m}\boldsymbol{H}_{N,m}^{-\frac{1}{2}}) = \delta^{-2}.
\end{aligned}
$$

Furthermore, noting that $\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2 \geq \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2$ and thus

$$\begin{aligned}
P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2 > \delta\right) &= P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2 > \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\delta\right) \\
&\leq P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|\boldsymbol{H}_{N,m}^{-\frac{1}{2}}\boldsymbol{D}_{N,m}^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2 > \lambda_{min}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{\frac{1}{2}}\delta\right) \\
&\leq \lambda_{max}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m})^{-1}\sigma^{-2}.
\end{aligned}$$

As $\lambda_{max}(\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}) \to \infty$, we have $P\left(p_{\boldsymbol{\theta}}^{-\frac{1}{2}}\|(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^0)\|_2 > \delta\right) \to 0$.

Proof of Theorem 2

Following the matrix algebra in Remark 2, we have

$$(\phi_m^u)^{-1}(\sum_{i=1}^{N}\tilde{\boldsymbol{\Omega}}_i)^{-1} \leq (\boldsymbol{H}_{N,m}^{or})^{-1} \leq (\phi_m^u)^{-1}(\sum_{i=1}^{N}\tilde{\boldsymbol{\Omega}}_i)^{-1},$$

and therefore 3.11 holds.

Recall that $\hat{\boldsymbol{\theta}}^{or} = \left((\hat{\boldsymbol{\gamma}}^{or})', (\hat{\boldsymbol{\alpha}}^{or})'\right)'$, by Taylor's expansion, we note that $(\tilde{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^0) = -(\boldsymbol{D}_{N,m}^{or})^{-1}\boldsymbol{G}_{N,m}^{or} = -(\boldsymbol{H}_{N,m}^{or})^{-1}\boldsymbol{G}_{N,m}^{or}$, where

$$\boldsymbol{G}_{N,m}^{or} = \sum_{i=1}^{N}\tilde{\boldsymbol{X}}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\theta}}^0),$$

since $\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$ holds for any i. By the standard central limit theorem, we have $(\boldsymbol{H}_{N,m}^{or})^{-1/2}\boldsymbol{G}_{N,m}^{or} \to N(\boldsymbol{0}, \boldsymbol{I}_{p+q})$, implying that $(\boldsymbol{H}_{N,m}^{or})^{1/2}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^0) \to N(\boldsymbol{0}, \boldsymbol{I}_{p+q})$, as either $m \to \infty$ or $\min_{1\leq k\leq p}(|\mathcal{G}_k|) \to \infty$. In addition, if $\boldsymbol{R}_i^0 \neq \boldsymbol{I}_m$ but m is bounded, then the asymptotic normality still holds when N goes to infinity regardless of the choice of working correlation matrix $\boldsymbol{R}_i$.

Moreover, under regularity conditions (A5)-(A6), we have $\lambda_{min}(\sum_i^N \tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i) = O(mN)$ and $\lambda_{max}(\sum_i^N \tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i) = O(mN)$. When $\boldsymbol{R}_i^0 = \boldsymbol{R}_i = \boldsymbol{I}_m$, it is trivial that $\boldsymbol{H}_{N,m}^{or} = \boldsymbol{D}_{N,m}^{or} \asymp m\boldsymbol{\Lambda}_{\boldsymbol{N},\boldsymbol{m}} = \boldsymbol{M}_{N,m}$.

3.8.3 Proof of Theorem 3 and Corollary 1, and condition $\mathcal{R}_a$

Following Lemma 2, we have

$$P\bigg((p+q)^{-\frac{1}{2}}\|(\boldsymbol{H}_{N,m}^{or})^{-\frac{1}{2}}\boldsymbol{D}_{N,m}^{or}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^{0})\|_2>\delta\bigg)<\tfrac{1}{\delta^2}.$$

Note that

$$\begin{aligned}\boldsymbol{H}_{N,m}&=\sum_{i=1}^{N}\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{U}_i\\&=\sum_{i=1}^{N}\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1/2}\boldsymbol{V}_i^{-1/2}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1/2}\boldsymbol{V}_i^{-1/2}\boldsymbol{U}_i\\&\leq\lambda_{max}(\boldsymbol{R}_i^{-1/2}\boldsymbol{R}_i^{0}\boldsymbol{R}_i^{-1/2})\sum_{i=1}^{N}\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1/2}\boldsymbol{V}_i^{-1/2}\boldsymbol{U}_i\\&=\lambda_{max}(\boldsymbol{R}_i^{-1}\boldsymbol{R}_i^{0})\boldsymbol{D}_{N,m}=\eta_{N,m}\boldsymbol{D}_{N,m}.\end{aligned}$$

Therefore we have $\boldsymbol{D}_{N,m}\boldsymbol{H}_{N,m}^{-1}\boldsymbol{D}_{N,m}\geq\eta_{N,m}^{-1}\boldsymbol{D}_{N,m}$, which implies that

$$\|(\boldsymbol{H}_{N,m}^{or})^{-\frac{1}{2}}\boldsymbol{D}_{N,m}^{or}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^{0})\|_2\geq\eta_{N,m}^{-\frac{1}{2}}\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^{0})\|_2,$$

and thus

$$P\bigg(\eta_{N,m}^{-\frac{1}{2}}\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^{0})\|_2>\delta\bigg)<c_0\frac{1}{\delta^2}$$

for some $c_0>0$. The proof of Theorem 3 is completed.

As a result, if $\eta_{N,m}\leq C_1$ holds uniformly for some positive constant C_1, it is straightforward that $\|(\boldsymbol{D}_{N,m}^{or})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^{0})\|_2=O_p(1)$. Note that $\boldsymbol{D}_{N,m}^{or}\leq\lambda_{max}(\boldsymbol{R}_i^{-1})\sum_i^N\tilde{\boldsymbol{\Omega}}_i\tilde{\boldsymbol{X}}_i^T\tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\Omega}}_i\leq(\nu_l^{'})^{-1}\boldsymbol{M}_{N,m}$, and therefore $\|(\boldsymbol{M}_{N,m})^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}^{or}-\tilde{\boldsymbol{\theta}}^{0})\|_2\leq O_p(1)$.

The correlation matrix $\boldsymbol{R}(\rho_{ij})$ is symmetric, which implies that $\|\boldsymbol{R}_{m\times m}\|_1=\|\boldsymbol{R}_{m\times m}\|_\infty\leq\sum_{k=0}^{m-1}|\rho_k|<\sum_{k=0}^{\infty}|\rho_k|<\infty$. By noting that $\|\boldsymbol{R}_{m\times m}\|_2^2\leq\|\boldsymbol{R}_{m\times m}\|_1\|\boldsymbol{R}_{m\times m}\|_\infty$, we have $\lambda_{max}(\boldsymbol{R})=\|\boldsymbol{R}\|_2$ uniformly bounded, and thus $\eta_{N,m}\leq(\nu_l^{\prime})^{-1}\sum_{k=0}^{\infty}|\rho_k|<\infty$.

3.8.4 Proof of Lemma 3, Corollary 2

Note that, for the proposed estimator and the subject-wise least squares estimator, each term of $\boldsymbol{U}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{U}_i$ in $\boldsymbol{D}_{N,m}$ does not equal to $\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i$, but is a block sparse matrix as $\boldsymbol{\mu}_i$ does not contain any other individualized parameter $\boldsymbol{\beta}_j$ for $j \neq i$. We denote

$$\boldsymbol{D}_{N,m}^s = \begin{pmatrix} \boldsymbol{D}_{xx}^s(Np \times Np) & \boldsymbol{D}_{xz}^s(Np \times q) \\ \boldsymbol{D}_{zx}^s(q \times Np) & \boldsymbol{D}_{zz}^s(q \times q) \end{pmatrix},$$

for the subject-wise estimator. Specifically,

$$\boldsymbol{D}_{N,m}^s = \begin{pmatrix} \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & 0 & \dots & 0 & \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{Z}_1 \\ 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{Z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{Z}_N \\ \boldsymbol{Z}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{Z}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & \boldsymbol{Z}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \sum_{i=1}^N \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \end{pmatrix},$$

Similarly we have

$$\boldsymbol{H}_{N,m}^s = \begin{pmatrix} \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & 0 & \dots & 0 & \boldsymbol{X}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{-1}\boldsymbol{Z}_1 \\ 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & 0 & \boldsymbol{X}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{-1}\boldsymbol{Z}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\Sigma}_N\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \boldsymbol{X}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\Sigma}_N\boldsymbol{V}_N^{-1}\boldsymbol{Z}_N \\ \boldsymbol{Z}_1^T\boldsymbol{V}_1^{-1}\boldsymbol{\Sigma}_1\boldsymbol{V}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{Z}_2^T\boldsymbol{V}_2^{-1}\boldsymbol{\Sigma}_2\boldsymbol{V}_2^{-1}\boldsymbol{X}_2 & \dots & \boldsymbol{Z}_N^T\boldsymbol{V}_N^{-1}\boldsymbol{\Sigma}_N\boldsymbol{V}_N^{-1}\boldsymbol{X}_N & \sum_{i=1}^N \boldsymbol{Z}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{\Sigma}_i\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \end{pmatrix},$$

Since $\boldsymbol{H}_{N,m}^s \leq \eta_{N,m}\boldsymbol{D}_{N,m}^s$, we have $\boldsymbol{a}^T(\boldsymbol{D}_{N,m}^s(\boldsymbol{H}_{N,m}^s)^{-1}\boldsymbol{D}_{N,m}^s)^{-1} \leq \eta_{N,m}\boldsymbol{a}^T(\boldsymbol{D}_{N,m}^s)^{-1}\boldsymbol{a}$. Note that $\boldsymbol{D}_{N,m}^s$ can be decomposed as

$$\boldsymbol{D}_{N,m}^s = \begin{pmatrix} \boldsymbol{I}_{Np} & \boldsymbol{0} \\ \boldsymbol{D}_{zx}^s(\boldsymbol{D}_{xx}^s)^{-1} & \boldsymbol{I}_q \end{pmatrix} \begin{pmatrix} \boldsymbol{D}_{xx}^s & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}_{zz}^s - \boldsymbol{D}_{zx}^s(\boldsymbol{D}_{xx}^s)^{-1}\boldsymbol{D}_{xz}^s \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{Np} & (\boldsymbol{D}_{xx}^s)^{-1}\boldsymbol{D}_{xz}^s \\ \boldsymbol{0} & \boldsymbol{I}_q \end{pmatrix},$$

and hence

$$(\boldsymbol{D}^s_{N,m})^{-1} = \begin{pmatrix} \boldsymbol{I}_{Np} & \boldsymbol{0} \\ -\boldsymbol{D}^s_{zx}(\boldsymbol{D}^s_{xx})^{-1} & \boldsymbol{I}_q \end{pmatrix} \begin{pmatrix} (\boldsymbol{D}^s_{xx})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{D}^s_{zz} - \boldsymbol{D}^s_{zx}(\boldsymbol{D}^s_{xx})^{-1}\boldsymbol{D}^s_{xz})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{Np} & -(\boldsymbol{D}^s_{xx})^{-1}\boldsymbol{D}^s_{xz} \\ \boldsymbol{0} & \boldsymbol{I}_q \end{pmatrix}.$$

Therefore, for any coordinate indicator $\boldsymbol{a}$ of $\boldsymbol{\beta}_i$, we have $\boldsymbol{a}^T(\boldsymbol{D}^s_{N,m})^{-1}\boldsymbol{a} = \boldsymbol{1}_p^T(\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}\boldsymbol{1}_p \leq p\lambda_{min}(\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}$. The result for the population-shared parameter $\hat{\boldsymbol{\alpha}}$ could be obtained following the same argument.

3.8.5 Proof of Lemma 4

Denote

$$\tilde{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{Z}) = \begin{pmatrix} \boldsymbol{X}_1 & \dots & \boldsymbol{0} & \boldsymbol{Z}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \boldsymbol{0} & \dots & \boldsymbol{X}_N & \boldsymbol{Z}_N \end{pmatrix},$$

and $\tilde{\boldsymbol{V}} = \text{diag}(\boldsymbol{V}_1, \dots, \boldsymbol{V}_N)$, $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_N)$, $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\boldsymbol{\varepsilon}}_1', \dots, \tilde{\boldsymbol{\varepsilon}}_N')'$.

Denote $\hat{\boldsymbol{\theta}}^{Sub} = \left((\hat{\boldsymbol{\beta}}^{Sub})', (\hat{\boldsymbol{\alpha}}^{Sub})'\right)'$, we have the least squares estimator

$$\begin{aligned}
(\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0) &= \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\varepsilon}} \\
&= \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\boldsymbol{\Sigma}}^{-1/2}\tilde{\boldsymbol{\varepsilon}} \\
&= \left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\boldsymbol{\varepsilon}}^*.
\end{aligned}$$

Under condition ($\mathcal{I}_a$) that $N = o(\tau^s_{N,m})$, by Chebyshev's inequality,

$$\begin{aligned}
\mathrm{P}(\|\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0\|_\infty > \delta) &= \mathrm{P}(\|(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\varepsilon}}\|_\infty > \delta) \\
&\leq \delta^{-2} tr\left((\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\right) \\
&= \delta^{-2} tr\left((\boldsymbol{D}^s_{N,m})^{-1}\boldsymbol{H}^s_{N,m}(\boldsymbol{D}^s_{N,m})^{-1}\right) \\
&\leq \delta^{-2}(Np+q)\lambda_{max}\left((\boldsymbol{D}^s_{N,m})^{-1}\boldsymbol{H}^s_{N,m}(\boldsymbol{D}^s_{N,m})^{-1}\right) \\
&\leq \delta^{-2}(Np+q)(\tau^s_{N,m})^{-1} \to 0
\end{aligned}$$

as $\tau^s_{N,m} \to \infty$.

Moreover, let $\boldsymbol{a}_t = ((\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}^{1/2})_{t\cdot}$ denote the tth row of $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}^{1/2}$, $t = 1, \ldots, (Np+q)$. By condition (i) in ($\mathcal{I}_b$), we have

$$\mathrm{P}(|\boldsymbol{a}_t^T\boldsymbol{\varepsilon}^*_{\boldsymbol{i}}| > \delta) < 2\mathrm{exp}(-\frac{\delta^2}{c_\sigma^2\|\boldsymbol{a}_t\|_2^2}).$$

Hence

$$\begin{aligned}
\mathrm{P}(\|\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0\|_\infty > \delta) &= \mathrm{P}(\|(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}^{1/2}\tilde{\boldsymbol{\varepsilon}}^*\|_\infty > \delta) \\
&\leq \sum_{t=1}^{Np+q} \mathrm{P}(|\boldsymbol{a}_t^T\boldsymbol{\varepsilon}^*| > \delta) \\
&\leq \sum_{t=1}^{Np+q} 2\mathrm{exp}(-\frac{\delta^2}{c_\sigma^2\|\boldsymbol{a}_t\|_2^2}) \\
&\leq (Np+q)\max_{1\leq t\leq Np+q}(2\mathrm{exp}(-\frac{\delta^2}{c_\sigma^2\|\boldsymbol{a}_t\|_2^2})) \\
&= 2(Np+q)\mathrm{exp}(-\frac{\delta^2}{c_\sigma^2\max_{1\leq t\leq Np+q}(\|\boldsymbol{a}_t\|_2^2)}).
\end{aligned}$$

Note that

$$\max_{1\le t\le Np+q}(\|\boldsymbol{a}_t\|_2^2) \le \lambda_{max}\left((\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{X}})^{-1}\right)$$
$$= \lambda_{max}\left((\boldsymbol{D}^s_{N,m})^{-1}\boldsymbol{H}^s_{N,m}(\boldsymbol{D}^s_{N,m})^{-1}\right) = (\tau^s_{N,m})^{-1}.$$

By condition (ii) in ($\mathcal{I}_b$) that $\log(N) = o(\tau^s_{N,m})$,

$$\mathrm{P}(\|\hat{\boldsymbol{\theta}}^{Sub} - \boldsymbol{\theta}^0\|_\infty > \delta) \le 2(Np+q)\exp(-\frac{\delta^2\tau^s_{N,m}}{c_\sigma^2}) \to 0$$

as $\tau^s_{N,m} \to \infty$.

3.8.6 Proof of Theorem 4

First, we prove the estimation consistency as $\lambda_m = o(m)$. Recall that $\boldsymbol{\theta_i} = (\boldsymbol{\beta}_i', \boldsymbol{\alpha}')'$ and $\tilde{\boldsymbol{X}}_i = (\boldsymbol{X}_i, \boldsymbol{Z}_i)$. Given $\hat{\boldsymbol{\gamma}}$, let

$$Q_{i,m}(\boldsymbol{\theta_i}|\hat{\boldsymbol{\gamma}}) = \| \boldsymbol{y_i} - \tilde{\boldsymbol{X}}_i\boldsymbol{\theta_i} \|_2^2 + \lambda_m\sum_{k=1}^{p} s(\beta_{ik}, \hat{\gamma}_k)$$
$$= L_{i,m}(\boldsymbol{\theta_i}) + S_{\lambda_m}(\boldsymbol{\beta_i}|\hat{\boldsymbol{\gamma}}),$$

which is minimized at $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{(m)}$, where $L_{i,m}(\cdot)$ is the squared loss function and $S_{\lambda_m}(\cdot)$ is the MDSP function.

Suppose $\frac{1}{m}\tilde{\boldsymbol{X}}_i{}^T\tilde{\boldsymbol{X}}_i \to \boldsymbol{C}_i$ where C_i is a positive definite matrix. Following [36], we define another function not related to m

$$Q_i(\boldsymbol{\theta_i}|\boldsymbol{\gamma}^0) = (\boldsymbol{\theta_i} - \boldsymbol{\theta_i^0})^T C_i(\boldsymbol{\theta_i} - \boldsymbol{\theta_i^0}) + \lambda_0\sum_{k=1}^{p} s(\beta_{ik}, \gamma_k^0),$$

and $\lambda_m/m \to \lambda_0$. Since C_i is not singular, if $\lambda_0 = 0$, then Q_i has a unique minimizer $\boldsymbol{\theta_i^0}$. Following

[36], we need to show

$$\sup_{\boldsymbol{\theta_i}\in\boldsymbol{\Theta}} \mid \frac{1}{m}Q_{i,m}(\boldsymbol{\theta_i}|\hat{\boldsymbol{\gamma}}) - Q_i(\boldsymbol{\theta_i}|\boldsymbol{\gamma}^0) - \sigma^2 \mid \rightarrow_p 0, \tag{3.17}$$

for any compact set $\boldsymbol{\Theta}$ and also that

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{(m)} = O_p(1). \tag{3.18}$$

The result in (3.17) follows

$$\frac{1}{m} \parallel \boldsymbol{y_i} - \tilde{\boldsymbol{X}}_i\boldsymbol{\theta_i} \parallel_2^2 \rightarrow_p (\boldsymbol{\theta_i} - \boldsymbol{\theta_i^0})^T C_i(\boldsymbol{\theta_i} - \boldsymbol{\theta_i^0}) + \sigma^2$$

according to standard results ([60]) and also

$$\begin{aligned}\sup_{\boldsymbol{\theta_i}\in\boldsymbol{\Theta}} |\frac{1}{m}S_{\lambda_m}(\boldsymbol{\beta_i}|\hat{\boldsymbol{\gamma}}) - S_0(\boldsymbol{\beta_i}|\boldsymbol{\gamma}^0)| &\leq \sup_{\boldsymbol{\theta_i}\in\boldsymbol{\Theta}} \frac{1}{m}|S_{\lambda_m}(\boldsymbol{\beta_i}|\hat{\boldsymbol{\gamma}}) - S_{\lambda_m}(\boldsymbol{\beta_i}|\boldsymbol{\gamma}^0)| + \sup_{\boldsymbol{\theta_i}\in\boldsymbol{\Theta}} |\frac{1}{m}S_{\lambda_m}(\boldsymbol{\beta_i}|\boldsymbol{\gamma}^0) - S_0(\boldsymbol{\beta_i}|\boldsymbol{\gamma}^0)| \\ &\leq \frac{\lambda_m p}{m} \parallel \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0 \parallel_2 + c|\frac{\lambda_m}{m} - \lambda_0| \rightarrow 0,\end{aligned}$$

where $c > 0$ is a constant. Although $H_{i,m}$ is not convex, we note that $\operatorname{argmin}(L_{i,m}) = O_p(1)$ and $\operatorname{argmin}(S_{\lambda_m}) = O_p(1)$. It follows that $\hat{\boldsymbol{\theta}}_{\boldsymbol{i}}^{(n)} = \operatorname{argmin}(Q_{i,m}) = O_p(1)$. Under(3.17) and (3.18), we have

$$\operatorname{argmin}(Q_{i,m}) \rightarrow_p \operatorname{argmin}(Q_i).$$

Next, we prove the selection consistency as $\lambda_m/\sqrt{m} \rightarrow \infty$. Let $\boldsymbol{\beta}_i = \boldsymbol{\beta}_i^0 + \frac{\boldsymbol{u}}{\lambda_m}$ and $\boldsymbol{\alpha} =$

$\boldsymbol{\alpha}^0 + \frac{\boldsymbol{v}}{\lambda_m}$, where $\boldsymbol{u} = O_p(1)$ and $\boldsymbol{v} = O_p(1)$. Let

$$\begin{aligned} D_{i,m}(\boldsymbol{u}, \boldsymbol{v}) &= Q_{i,m}(\boldsymbol{\beta}_i, \boldsymbol{\alpha}|\hat{\boldsymbol{\gamma}}) - Q_{i,m}(\boldsymbol{\beta}_i^0, \boldsymbol{\alpha}^0|\hat{\boldsymbol{\gamma}}) \\ &= L_{i,m}(\boldsymbol{\beta}_i, \boldsymbol{\alpha}) - L_{i,m}(\boldsymbol{\beta}_i^0, \boldsymbol{\alpha}^0) + S_{\lambda_m}(\boldsymbol{\beta}_i|\hat{\boldsymbol{\gamma}}) - S_{\lambda_m}(\boldsymbol{\beta}_i^0|\hat{\boldsymbol{\gamma}}) \\ &= \|\boldsymbol{\varepsilon}_i - \boldsymbol{X}_i\frac{\boldsymbol{u}}{\lambda_m} - \boldsymbol{Z}_i\frac{\boldsymbol{v}}{\lambda_m}\|_2^2 - \|\boldsymbol{\varepsilon}_i\|_2^2 + \lambda_m\sum_{k=1}^{p}[s(\beta_{ik}^0 + \frac{u_k}{\lambda_m}, \hat{\gamma}_k) - s(\beta_{ik}^0, \hat{\gamma}_k)] \\ &= \frac{\sqrt{m}}{\lambda_m}\frac{1}{\sqrt{m}}\boldsymbol{\varepsilon}_i^T(\boldsymbol{X}_i\boldsymbol{u} + \boldsymbol{Z}_i\boldsymbol{v}) + \frac{m}{\lambda_m^2}(\boldsymbol{u}^T, \boldsymbol{v}^T)\left(\frac{1}{m}(\boldsymbol{X}_i, \boldsymbol{Z}_i)^T(\boldsymbol{X}_i, \boldsymbol{Z}_i)\right)(\boldsymbol{u}^T, \boldsymbol{v}^T)^T \\ &\quad + \lambda_m\sum_{k=1}^{p}[s(\beta_{ik}^0 + \frac{u_k}{\lambda_m}, \hat{\gamma}_k) - s(\beta_{ik}^0, \hat{\gamma}_k)]. \end{aligned}$$

The first two terms vanish as $\lambda_m/\sqrt{m} \to \infty$. Let $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}^{(0)}$, it follows that

$$D_{i,m}(\boldsymbol{u}, \boldsymbol{v}) \to \sum_{k\in\mathcal{A}_i^c}|u_k| + \sum_{k\in\mathcal{A}_i}u_k sgn(\gamma_k^{(0)} - \gamma_k^0).$$

Since $\sqrt{m}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) = O_p(1)$, that is, $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\gamma}^0$, then the second term above also vanishes, therefore $D_{i,m}(\boldsymbol{u}, \boldsymbol{v})$ is minimized at $u_k = 0, k \in \mathcal{A}_i^c$. Note that $\boldsymbol{u} = \lambda_m(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^0)$ and thus $\mathrm{argmin}(Q_{i,m}) = \mathrm{argmin}(D_{i,m})$, the proof is hence completed.

In general, the regularity condition (A6) only guarantees that $\frac{1}{m}\boldsymbol{X}_i^T\boldsymbol{X}_i$ is positive definite, but not for $\frac{1}{m}\tilde{\boldsymbol{X}}_i{}^T\tilde{\boldsymbol{X}}_i$ since there could be invariant population-shared covariates $\boldsymbol{Z}_i$ within the subject. However, the above argument still holds by taking a transformation $\boldsymbol{Z}_i^* = \boldsymbol{Z}_i\boldsymbol{T}_i$ such that $\frac{1}{m}\boldsymbol{Z}_i^{*T}\boldsymbol{Z}_i^*$ is positive definite.

3.8.7 Proof of Lemma 5, Theorem 5 and Corollary 4

We first establish the following result.

Lemma 6. *Suppose there is a sequence of numbers $\{a_i\}_{i=1,\ldots,N}$ associated with a partition of index sets $\mathcal{G}_l$ ($l = 1, \ldots, L$), such that $|a_i - b_l| \le \epsilon$ for any $i \in \mathcal{G}_l$, where ϵ is a small positive value. Then*

there is a local minimizer $\hat{\boldsymbol{b}}$ of following objective function

$$S(\boldsymbol{b}|\boldsymbol{a}) = \sum_{i=1}^{N} \left(\underset{1\le l\le L}{\wedge} |a_i - b_l| \right),$$

such that $\|\hat{\boldsymbol{b}} - \boldsymbol{b}\|_\infty \le 2\epsilon$, where $\underset{1\le l\le L}{\wedge} |a_i - b_l| = \min_{1\le l\le L}(|a_i - b_l|)$.

Proof: Without loss of generality, assume $b_1 = 0$, we have $|a_i| \le \epsilon$ for any $i \in \mathcal{G}_1$ and hence $\sum_{i\in\mathcal{G}_1} |a_i| \le |\mathcal{G}_1|\epsilon$. Moreover, note that $\sum_{i\in\mathcal{G}_1} |a_i - 2\epsilon| = \sum_{i\in\mathcal{G}_1}(2\epsilon - a_i) \ge \sum_{i\in\mathcal{G}_1} \epsilon = |\mathcal{G}_1|\epsilon$ and $\sum_{i\in\mathcal{G}_1} |a_i + 2\epsilon| = \sum_{i\in\mathcal{G}_1}(2\epsilon - a_i) \ge \sum_{i\in\mathcal{G}_1} \epsilon = |\mathcal{G}_1|\epsilon$. Therefore there is a minimizer $|\hat{b}_1| \le 2\epsilon$ and the proof of Lemma 6 is completed.

The proposed objective function is

$$\begin{aligned} Q_{N,m}(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) &= \frac{1}{2}\sum_{i=1}^{N} \parallel \boldsymbol{y}_i - \boldsymbol{X_i}\boldsymbol{\beta_i} - \boldsymbol{Z_i}\boldsymbol{\alpha} \parallel_2^2 + \lambda_{N,m} \sum_{i=1}^{N}\sum_{k=1}^{p} s(\beta_{ik}, \gamma_k) \\ &= L_{N,m}(\boldsymbol{\theta}) + S_{\lambda_{N,m}}(\boldsymbol{\beta}, \boldsymbol{\gamma}). \end{aligned}$$

Let $\boldsymbol{\theta}^* = \boldsymbol{\theta}^0 + (\tau^s_{N,m})^{-1/2}\boldsymbol{u}$, $\boldsymbol{\gamma}^* \in \mathbb{R}^p$, where $\parallel \boldsymbol{u} \parallel_2 = d$. Note that $S_{\lambda_{N,m}}(\boldsymbol{\beta}^0, \boldsymbol{\gamma}^0) = 0$, by Taylor's expansion, we have

$$\begin{aligned} D_{N,m}(\boldsymbol{u}) &= Q_{N,m}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) - Q_{N,m}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) \\ &= L_{N,m}(\boldsymbol{\theta}^*) - L_{N,m}(\boldsymbol{\theta}^0) + S_{\lambda_{N,m}}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) \\ &= (\tau^s_m)^{-1/2}\dot{L}^T_{N,m}(\boldsymbol{\theta}^0)\boldsymbol{u} + \frac{1}{2}(\tau^s_m)^{-1}\boldsymbol{u}^T\ddot{L}_{N,m}(\boldsymbol{\theta}^0)\boldsymbol{u} + S_{\lambda_{N,m}}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*), \\ &= (\tau^s_m)^{-1/2}(\boldsymbol{G}^s_{N,m})^T\boldsymbol{u} + \frac{1}{2}(\tau^s_m)^{-1}\boldsymbol{u}^T\boldsymbol{D}^s_{N,m}\boldsymbol{u} + S_{\lambda_{N,m}}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*), \end{aligned}$$

where $\dot{L}_{N,m}$ is the gradient vector of $L_{N,m}(\boldsymbol{\theta})$ and $\ddot{L}_{N,m}$ is the Jacobian matrix. Note that

$$\begin{aligned} P(\boldsymbol{u}^T(\boldsymbol{H}^s_{N,m})^{-1/2}\boldsymbol{G}^s_{N,m}| > \delta) &\le \delta^{-2}\boldsymbol{u}^T\mathbf{E}((\boldsymbol{H}^s_{N,m})^{-1/2}\boldsymbol{G}^s_{N,m}(\boldsymbol{G}^s_{N,m})^T(\boldsymbol{H}^s_{N,m})^{-1/2})\boldsymbol{u} \\ &\le \delta^{-2}d^2, \end{aligned}$$

implying that $\boldsymbol{u}^T(\boldsymbol{H}^s_{N,m})^{-1/2}\boldsymbol{G}^s_{N,m} = O_p(d)$. Moreover, we have

$$\begin{aligned}
(\boldsymbol{H}^s_{N,m})^{1/2} &= (\boldsymbol{D}^s_{N,m})^{1/2}(\boldsymbol{D}^s_{N,m})^{-1/2}(\boldsymbol{H}^s_{N,m})^{1/2}(\boldsymbol{D}^s_{N,m})^{-1/2}(\boldsymbol{D}^s_{N,m})^{1/2} \\
&\leq (\boldsymbol{D}^s_{N,m})^{1/2}\lambda_{max}\left((\boldsymbol{D}^s_{N,m})^{-1/2}(\boldsymbol{H}^s_{N,m})^{1/2}(\boldsymbol{D}^s_{N,m})^{-1/2}\right)(\boldsymbol{D}^s_{N,m})^{1/2} \\
&= \lambda_{min}\left((\boldsymbol{D}^s_{N,m})^{1/2}(\boldsymbol{H}^s_{N,m})^{-1/2}(\boldsymbol{D}^s_{N,m})^{1/2}\right)^{-1}\boldsymbol{D}^s_{N,m} \\
&= (\tau^s_{N,m})^{-1/2}\boldsymbol{D}^s_{N,m},
\end{aligned}$$

and thus $(\tau^s_{N,m})^{-1/2}(\boldsymbol{H}^s_{N,m})^{1/2} \leq (\tau^s_{N,m})^{-1}\boldsymbol{D}^s_{N,m}$. Consequently, if d is sufficiently large, then the second term in $D_{N,m}(\boldsymbol{u})$ dominates the first term, which implies that, with probability tending to 1, $D_{N,m}(\boldsymbol{u}) > 0$ at $\| \boldsymbol{u} \|_2 = d$. Hence we have

$$P\left\{ \inf_{\|\boldsymbol{u}\|_2=d} D_{N,m}(\boldsymbol{u}) > 0 \right\} \to 1.$$

This implies that, with probability tending to 1, there exists a local minimizer $\hat{\boldsymbol{\theta}}$ in the ball $B(\boldsymbol{\theta}^0, (\tau^s_{N,m})^{-1/2}d)$. In particular, this indicates that the convergence rate for estimator of any individualized parameter $\hat{\boldsymbol{\beta}}_i$ is $(\tau^s_{N,m})^{1/2}$. Following the proof of Lemma 4, under condition $\mathcal{I}_a$ or $\mathcal{I}_b$, we have $P(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_\infty > p^{-1}r) \to 0$ for any positive constant r. By Lemma 6, given $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|_\infty \leq p^{-1}r$, there exists a minimizer $\hat{\boldsymbol{\gamma}}$ of $S_{\lambda_{N,m}}(\boldsymbol{\gamma}|\hat{\boldsymbol{\beta}})$, such that $\hat{\boldsymbol{\gamma}} \in B(\boldsymbol{\gamma}^0, r)$. The proof of Lemma 5 is completed.

Next we show that the objective function $Q_{N,m}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ is convex at $\{\boldsymbol{\theta}^* \in B(\boldsymbol{\theta}^0, (\tau^s_{N,m})^{-1/2}d)\} \cap \{\boldsymbol{\gamma}^* \in B(\boldsymbol{\gamma}^0, (\tau^s_{N,m})^{-1/2}d)\}$ when m is sufficiently large. Note that, if $\beta^0_{ik} = \gamma^0_k$, we have

$$\begin{aligned}
\sup_{\beta^*_{ik}\in B(\beta^0_{ik}), \gamma^*_k \in B(\gamma^0_k)} |\beta^*_{ik} - \gamma^*_k| &\leq \sup_{\beta^*_{ik}\in B(\beta^0_{ik})} |\beta^*_{ik} - \beta^0_{ik}| + \sup_{\gamma^*_k \in B(\gamma^0_k)} |\gamma^*_k - \gamma^0_k| + |\beta^0_{ik} - \gamma^0_k| \\
&\leq 2(\tau^s_{N,m})^{-1/2}d + |\beta^0_{ik} - \gamma^0_k| \to 0,
\end{aligned}$$

and $\inf_{\beta_{ik}^* \in B(\beta_{ik}^0)} |\beta_{ik}^*| \geq (|\gamma_k^0| - (\tau_{N,m}^s)^{-1/2} d)_+ \to |\gamma_k^0|$. It follows

$$P\bigg(\sup_{\beta_{ik}^* \in B(\beta_{ik}^0), \gamma_k^* \in B(\gamma_k^0)} |\beta_{ik}^* - \gamma_k^*| \leq \inf_{\beta_{ik}^* \in B(\beta_{ik}^0)} |\beta_{ik}^*| \bigg) \to 1.$$

Define

$$\tilde{S}_{\lambda_{N,m}}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) = \lambda_{N,m} \sum_{k=1}^{p} \bigg\{ \sum_{i \in \mathcal{G}_k^c} |\beta_{ik}^*| + \sum_{i \in \mathcal{G}_k} |\beta_{ik}^* - \gamma_k^*| \bigg\},$$

and $\tilde{Q}_{N,m} = L_{N,m} + \tilde{S}_{\lambda_{N,m}}$. We have $Q_{N,m}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*) = \tilde{Q}_{N,m}(\boldsymbol{\theta}^*, \boldsymbol{\gamma}^*)$ at $\{\boldsymbol{\theta}^* \in B(\boldsymbol{\theta}^0, (\tau_{N,m}^s)^{-1/2} d)\} \cap \{\boldsymbol{\gamma}^* \in B(\boldsymbol{\gamma}^0, (\tau_{N,m}^s)^{-1/2} d)\}$ when $\tau_{N,m}^s$ is sufficiently large, and thus $\operatorname{argmin} Q_{N,m} = \operatorname{argmin} \tilde{Q}_{N,m}$.

Let $\boldsymbol{\theta}^{**} = \boldsymbol{\theta}^0 + \lambda_{N,m}^{-1} \boldsymbol{u}$ and $\boldsymbol{\gamma}^{**} = \boldsymbol{\gamma}^0 + \lambda_{N,m}^{-1} \boldsymbol{v}$, similarly it follows that

$$\begin{aligned} D_{N,m}(\boldsymbol{u}, \boldsymbol{v}) &= \tilde{Q}_{N,m}(\boldsymbol{\theta}^{**}, \boldsymbol{\gamma}^{**}) - \tilde{Q}_{N,m}(\boldsymbol{\theta}^0, \boldsymbol{\gamma}^0) = L_{N,m}(\boldsymbol{\theta}^{**}) - L_{N,m}(\boldsymbol{\theta}^0) + \tilde{S}_{\lambda_{N,m}}(\boldsymbol{\beta}^{**}, \boldsymbol{\gamma}^{**}) \\ &= \frac{(\tau_{N,m}^s)^{1/2}}{\lambda_{N,m}} (\tau_{N,m}^s)^{-1/2} \dot{L}_{N,m}^T(\boldsymbol{\theta}^0) \boldsymbol{u} + \frac{\tau_{N,m}^s}{\lambda_{N,m}^2} \frac{1}{2} (\tau_{N,m}^s)^{-1} \boldsymbol{u}^T \ddot{L}_{N,m}(\boldsymbol{\theta}^0) \boldsymbol{u} \\ &\quad + \lambda_{N,m} \sum_{k=1}^{p} \bigg\{ \sum_{i \in \mathcal{G}_k^c} \lambda_{N,m}^{-1} |u_{ik}| + \sum_{i \in \mathcal{G}_k} \lambda_{N,m}^{-1} |u_{ik} - v_k| \bigg\}. \end{aligned}$$

Since $\frac{\lambda_{N,m}}{(\tau_{N,m}^s)^{1/2}} \to \infty$, hence $D_{N,m}(\boldsymbol{u}, \boldsymbol{v}) \to_p D(\boldsymbol{u}, \boldsymbol{v})$, where

$$D(\boldsymbol{u}, \boldsymbol{v}) = \sum_{k=1}^{p} \bigg\{ \sum_{i \in \mathcal{G}_k^c} |u_{ik}| + \sum_{i \in \mathcal{G}_k} |u_{ik} - v_k| \bigg\},$$

which is minimized at $\{u_{ik} = 0 | i \in \mathcal{G}_k^c; \quad u_{ik} = v_k | i \in \mathcal{G}_k\}$. Because $D_{N,m}(\boldsymbol{u}, \boldsymbol{v})$ is a convex function, it follows ([21]) that $\operatorname{argmin} D_{N,m} \to \operatorname{argmin} D$, and thus $\operatorname{argmin} Q_{N,m} \to \operatorname{argmin} D$. This implies that $P(\hat{\beta}_{ik} = 0 | i \in \mathcal{G}_k^c) \to 1$ and $P(\hat{\beta}_{ik} = \hat{\gamma}_k | i \in \mathcal{G}_k) \to 1$. The proof of Theorem 5 is completed.

3.8.8 Proof of Theorem 6

For proposed objective function $Q_{N,m}$, at mth iteration in Algorithm 2, it is obvious that

$$\begin{aligned} L_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}) + S_{N,m}^{(n-1)}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n-1)}) &\leq L_{N,m}(\hat{\boldsymbol{\beta}}^{(n-1)}, \hat{\boldsymbol{\alpha}}^{(n-1)}) + S_{N,m}^{(n-1)}(\hat{\boldsymbol{\beta}}^{(n-1)}, \hat{\boldsymbol{\gamma}}^{(n-1)}) \\ &= L_{N,m}(\hat{\boldsymbol{\beta}}^{(n-1)}, \hat{\boldsymbol{\alpha}}^{(n-1)}) + S_{N,m}(\hat{\boldsymbol{\beta}}^{(n-1)}, \hat{\boldsymbol{\gamma}}^{(n-1)}) \end{aligned}$$

holds after Step 3, where $L_{N,m}(\cdot)$ is the squared-loss function.

Next, after Step 4 (grouping) in Algorithm 2, we have

$$S_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)}) \leq S_{N,m}^{(n)}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n-1)}) \leq S_{N,m}^{(n-1)}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n-1)}).$$

This follows

$$L_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}) + S_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)}) \leq L_{N,m}(\hat{\boldsymbol{\beta}}^{(n-1)}, \hat{\boldsymbol{\alpha}}^{(n-1)}) + S_{N,m}(\hat{\boldsymbol{\beta}}^{(n-1)}, \hat{\boldsymbol{\gamma}}^{(n-1)}),$$

which implies a non-increasing sequence of $Q_{N,m}(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)})$ obtained by Algorithm 2. Note that $Q_{N,m}$ is non-negative, thus the obtained iterations $(\hat{\boldsymbol{\beta}}^{(n)}, \hat{\boldsymbol{\alpha}}^{(n)}, \hat{\boldsymbol{\gamma}}^{(n)})$ would converge to a local minimizer.

3.9 Tables and Figures

Table 3.1: The average root mean square error (RMSE) of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 1, 2$, where Sub, Homo, FusedL, Lasso, AdapL, SCAD and MCP stand for subject-wise model, homogeneous model, the fused Lasso ([83]), the Lasso ([81]), the adaptive Lasso ([100]), the SCAD([17]) and the MCP ([95]) regularization models, respectively. The number of subgroups (two) is correctly specified in the proposed model.

Sample	Cluster				Methods				
Size (N)	Size(m)	**MDSP**	Sub	Homo	FusedL	Lasso	AdapL	SCAD	MCP
					$\gamma = 1$				
40	10	0.267	0.349	0.504	0.323	0.439	0.339	0.344	0.350
	20	**0.120**	0.232	0.502	0.206	0.298	0.207	0.201	0.201
100	10	0.262	0.350	0.501	0.319	0.394	0.334	0.335	0.345
	20	**0.119**	0.233	0.501	0.210	0.271	0.208	0.205	0.206
					$\gamma = 2$				
40	10	**0.122**	0.349	1.004	0.317	0.408	0.309	0.311	0.309
	20	**0.048**	0.232	1.002	0.204	0.293	0.181	0.168	0.167
100	10	**0.113**	0.350	1.001	0.318	0.387	0.305	0.300	0.299
	20	**0.037**	0.233	1.001	0.210	0.274	0.208	0.206	0.206

Table 3.2: The average RMSE of the estimated subgroup homogeneous effect $\hat{\gamma}$ from the proposed model based on 100 simulations (empirical standard errors in parenthesis), with sample size $N = 40, 100$, cluster size $m = 10, 20$.

Homogeneous	N=40		N=100	
Effect	$T = 10$	$T = 20$	$T = 10$	$T = 20$
$\gamma = 1$	1.03(0.08)	1.00(0.05)	1.02(0.05)	1.00(0.03)
$\gamma = 2$	2.01(0.07)	2.00(0.05)	2.00(0.05)	2.00(0.03)

Table 3.3: The average correct variable selection rate (CVSR), sensitivity and specificity of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, and subgroup homogeneous effect $\gamma = 1, 2$, where Sub, Homo, FusedL, Lasso, AdapL, SCAD and MCP stand for subject-wise model, homogeneous model, the fused Lasso ([83]), the Lasso ([81]), the adaptive Lasso ([100]), the SCAD([17]) and the MCP ([95]) regularization models, respectively. The number of subgroups (two) is correctly specified in the proposed model.

Variable Selection	Sample Size (N)	Cluster Size(m)	Methods					
			MDSP	FusedL	Lasso	AdapL	SCAD	MCP
			$\gamma = 1$					
CVSR	40	10	0.916	0.692	0.876	0.820	0.717	0.741
		20	0.970	0.678	0.924	0.869	0.778	0.829
	100	10	0.909	0.673	0.862	0.840	0.718	0.754
		20	0.963	0.682	0.890	0.888	0.773	0.833
Sensitivity	40	10	0.942	0.978	0.898	0.943	0.975	0.966
		20	0.985	1.000	0.990	0.997	0.999	0.999
	100	10	0.946	0.986	0.917	0.941	0.974	0.967
		20	0.990	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.909	0.406	0.853	0.696	0.460	0.517
		20	0.956	0.356	0.857	0.742	0.557	0.659
	100	10	0.886	0.360	0.807	0.739	0.462	0.542
		20	0.942	0.364	0.787	0.782	0.547	0.669
			$\gamma = 2$					
CVSR	40	10	0.959	0.639	0.886	0.884	0.800	0.852
		20	0.972	0.670	0.928	0.940	0.908	0.953
	100	10	0.940	0.648	0.868	0.898	0.809	0.871
		20	0.965	0.682	0.890	0.888	0.773	0.832
Sensitivity	40	10	0.997	0.996	0.997	0.998	1.000	0.998
		20	1.000	1.000	1.000	1.000	1.000	1.000
	100	10	0.998	0.997	0.998	0.998	0.999	0.999
		20	1.000	0.999	0.993	0.994	0.999	0.997
Specificity	40	10	0.922	0.282	0.774	0.771	0.602	0.705
		20	0.945	0.340	0.856	0.880	0.816	0.906
	100	10	0.882	0.299	0.738	0.797	0.620	0.744
		20	0.930	0.365	0.787	0.782	0.546	0.668

Table 3.4: The average root mean square error (RMSE) of the proposed MDSP model with different working correlation structures based on 100 simulations, including AR-1 ($\boldsymbol{\beta}_{AR1}$), exchangeable ($\boldsymbol{\beta}_{Ex}$) and independent ($\boldsymbol{\beta}_{Ind}$) models. The true structures for the within-subject serial correlation are AR-1 or exchangeable, and correlation parameter $\rho = 0.5$, sample size $N = 20, 80$, cluster size $m = 10, 20$.

True Correlation	Cluster size (m)	$N = 20$			$N = 80$		
		$\boldsymbol{\beta}_{AR1}$	$\boldsymbol{\beta}_{Ex}$	$\boldsymbol{\beta}_{Ind}$	$\boldsymbol{\beta}_{AR1}$	$\boldsymbol{\beta}_{Ex}$	$\boldsymbol{\beta}_{Ind}$
Exch	10	0.209	**0.165**	0.265	0.193	**0.110**	0.258
	20	0.072	**0.053**	0.078	0.067	**0.051**	0.076
AR-1	10	**0.182**	0.230	0.258	**0.183**	0.205	0.256
	20	**0.091**	0.121	0.132	**0.089**	0.112	0.130

Table 3.5: The mean of identified subgroup numbers of the proposed model compared with the two-stage OLSK method based on 100 simulations, with sample size $N = 60, 120$, cluster size $m = 5, 10, 20$. The first three scenarios contain one individualized predictor ($p = 1$) of one, two and three groups, respectively. The last scenario contains two individualized predictors ($p = 2$), one with two groups and the other with three groups. The subgroup sizes are equal in each scenario. The subgroup homogeneous effects are listed as possible values for β_i in the table.

Number of individualized variables		$p = 1$						$p = 2$			
Sample Size (N)	Cluster Size(m)	$\beta_i = 0$		$\beta_i = 0, 1$		$\beta_i = 0, 2, 5$		$\beta_{1i} = 0, 2$		$\beta_{2i} = -2, 0, 1$	
		MDSP	OLSK	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK	MDSP	OLSK
60	5	1.0(100)	1.0(100)	2.0(95)	1.0(2)	2.9(88)	2.5(68)	2.0(100)	1.5(52)	3.2(85)	1.2(0)
	10	1.0(100)	1.0(100)	2.0(100)	1.3(26)	3.1(90)	2.7(74)	2.0(100)	2.0(100)	3.1(90)	2.4(44)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.1(92)	2.8(78)	2.0(100)	2.0(100)	3.0(100)	2.8(80)
120	5	1.0(100)	1.0(100)	2.0(96)	1.0(2)	3.2(86)	2.8(82)	2.0(100)	1.7(72)	3.1(90)	1.4(0)
	10	1.0(100)	1.0(100)	2.0(100)	1.2(24)	3.1(92)	2.9(86)	2.0(100)	2.0(100)	3.1(90)	2.6(64)
	20	1.0(100)	1.0(100)	2.0(100)	2.0(100)	3.0(98)	2.9(96)	2.0(100)	2.0(100)	3.1(92)	2.78(78)

Table 3.6: The average RMSE and CVSR of the proposed MDSP model compared to the subject-wise model (Sub), the fused Lasso (FusedL), the Lasso, the adaptive Lasso (Adapl), the SCAD and the MCP penalization models, based on 100 simulations with sample size $N = 60$ and cluster size $m = 10$. The first case contains a population homogeneous effect ($G_k = 1$) and the second case contains an individualized predictor of three subgroups ($G_k = 3$) with equal subgroup size. In both cases the proposed model assumes two subgroups. The estimated subgroup homogeneous effects from the proposed model are $\hat{\gamma} = 2.01(0.06)$ and $\hat{\gamma} = -2.99(0.06)$ in these two cases (with empirical standard errors in parenthesis), respectively.

Case		**MDSP**	Sub	FusedL	Lasso	AdapL	SCAD	MCP
$G_k = 1$	RMSE	0.115	0.346	0.319	0.414	0.373	0.346	0.345
$(\beta_i = 2)$	CVSR	0.996	-	0.993	0.994	0.992	0.995	0.996
$G_k = 3$	RMSE	0.277	0.349	0.315	0.410	0.335	0.337	0.338
$(\beta_i = -3, 0, 1)$	CVSR	0.901	-	0.748	0.877	0.902	0.816	0.817

Table 3.7: The estimated coefficients of the population model, the random-effects model, the L_1-penalty model and the proposed model with corresponding median prediction errors (MPE) for the ACTG data. The individualized coefficient estimators $\hat{\beta}_{izt}$'s in the Lasso model, the fused Lasso (fusedL) model and the proposed (MDSP) model are not listed.

Model	$\hat{\beta}_0$	$\hat{\beta}_t$	$\hat{\beta}_z$	$\hat{\beta}_a$	$\hat{\beta}_g$	$\hat{\beta}_{zt}$	$\hat{\gamma}^+$	$\hat{\gamma}^-$	MPE
Population	3.09	-0.68	-0.54	0.01	-0.01	-0.24	-	-	1.67
Random-effects	2.56	-0.68	-0.57	0.02	-0.01	-0.29	-	-	1.70
Lasso	3.09	-0.76	-0.54	0.01	-0.01	-	-	-	1.64
fusedL	3.05	-0.72	-0.52	0.01	-0.01	-	-	-	1.62
MDSP	3.10	-0.68	-0.56	0.01	-0.01	-	0.62	-0.60	**1.44**

Table 3.8: The treatment effect estimators within each subgroup model (zero-effect group: β_{zt}^0, negative-effect group: β_{zt}^- and positive-effect group β_{zt}^+) as well as the standard errors (s.e.) and the p-values. Each subgroup consists of the corresponding individuals in the treatment group identified by the Lasso model or the proposed model (MDSP) as well as all the individuals in the control group. The proportion of individuals with the treatment classified into each subgroup is provided.

Model		Estimates	s.e.	p-value	Proportion
Lasso	$\hat{\beta}_{zt}^0$	-0.24	0.17	0.14	0.75
	$\hat{\beta}_{zt}^-$	-0.73	0.31	0.02	0.18
	$\hat{\beta}_{zt}^+$	0.82	0.48	0.10	0.07
MDSP	$\hat{\beta}_{zt}^0$	-0.04	0.30	0.89	0.20
	$\hat{\beta}_{zt}^-$	-0.68	0.08	$\mathbf{0.00}$	0.64
	$\hat{\beta}_{zt}^+$	0.72	0.33	$\mathbf{0.02}$	0.16

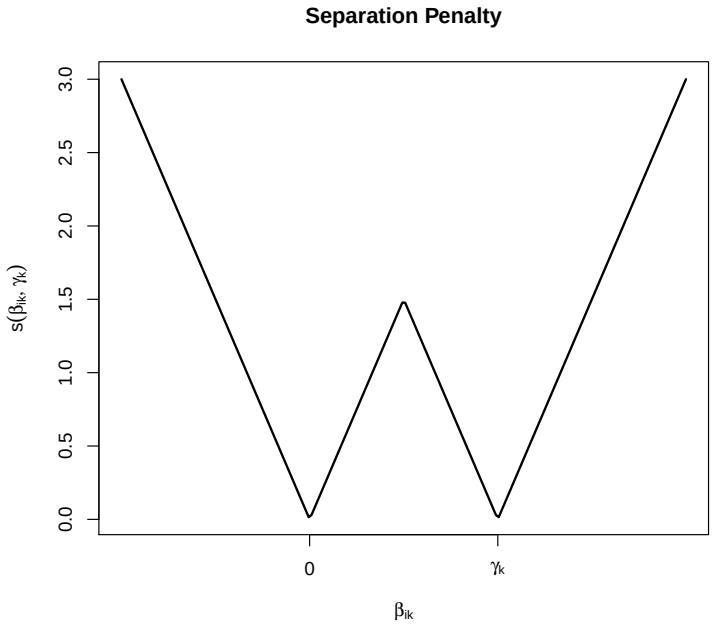

Figure 3.1: The separation penalty $s(\beta_{ik}, \gamma_k)$ given γ_k.

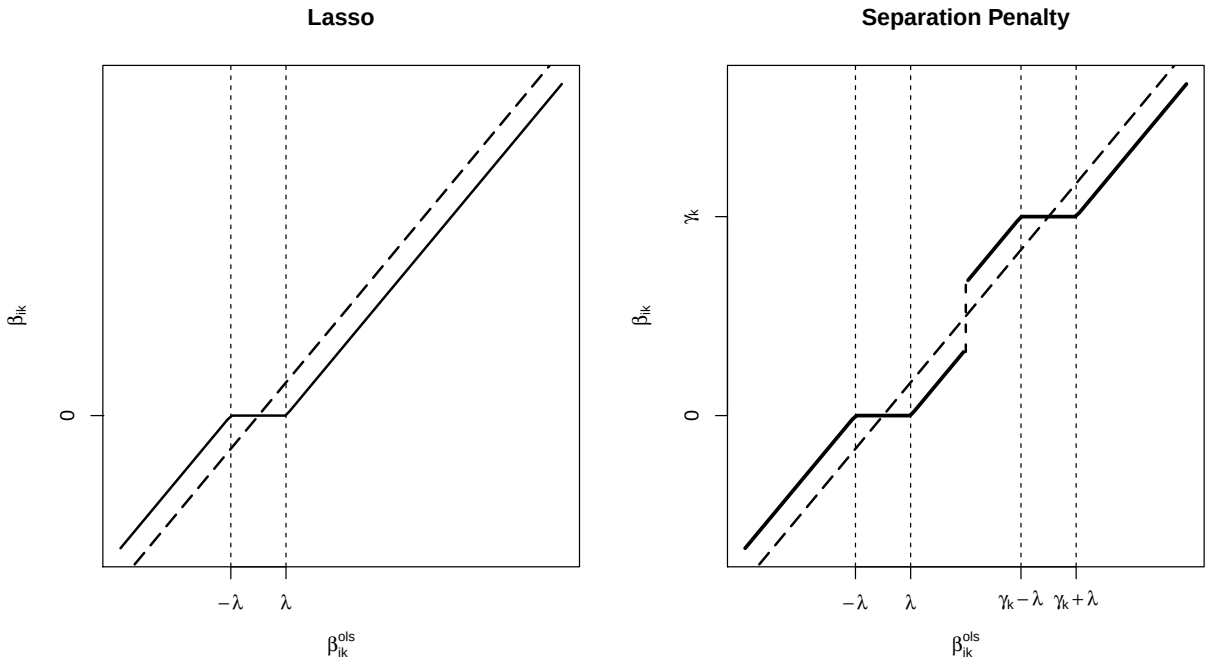

Figure 3.2: Thresholding functions for the Lasso and the proposed separation penalty.

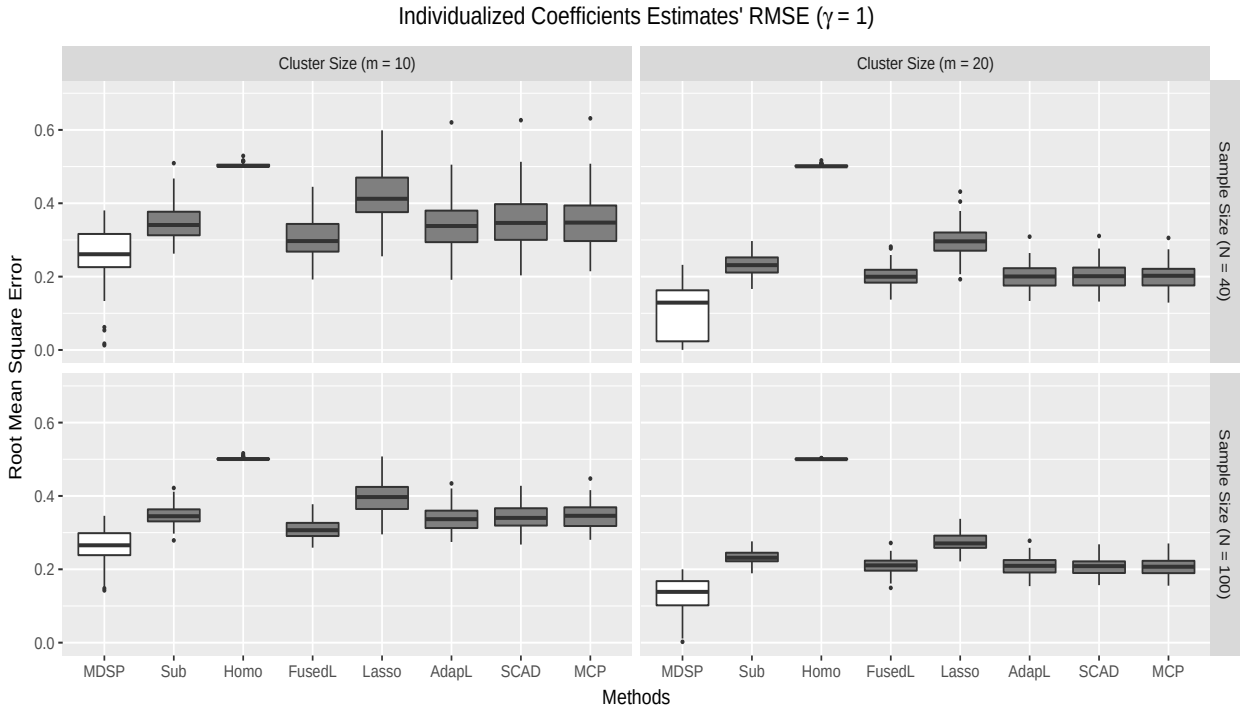

Figure 3.3: The boxplot of RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, where homogeneous effect $\gamma = 1$.

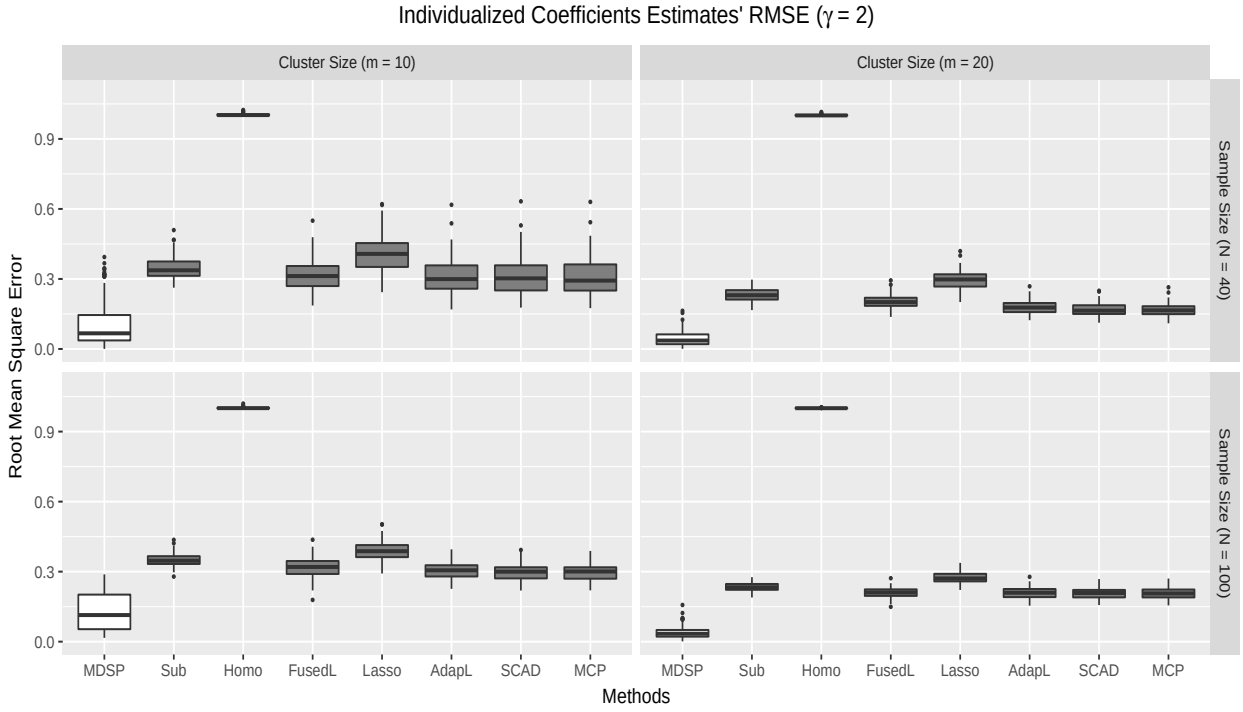

Figure 3.4: The boxplot of RMSE of the proposed MDSP model compared with other approaches based on 100 simulations, with sample size $N = 40, 100$, cluster size $m = 10, 20$, where homogeneous effect $\gamma = 2$.

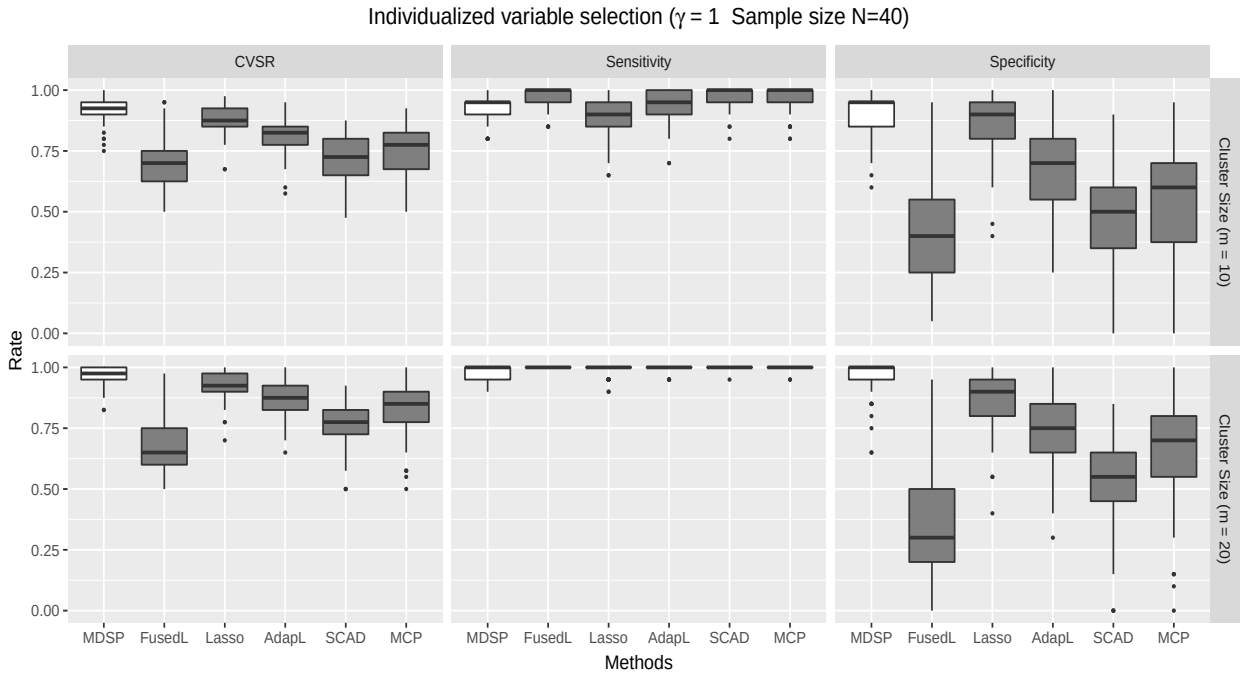

Figure 3.5: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 1$ and sample size $N = 40$.

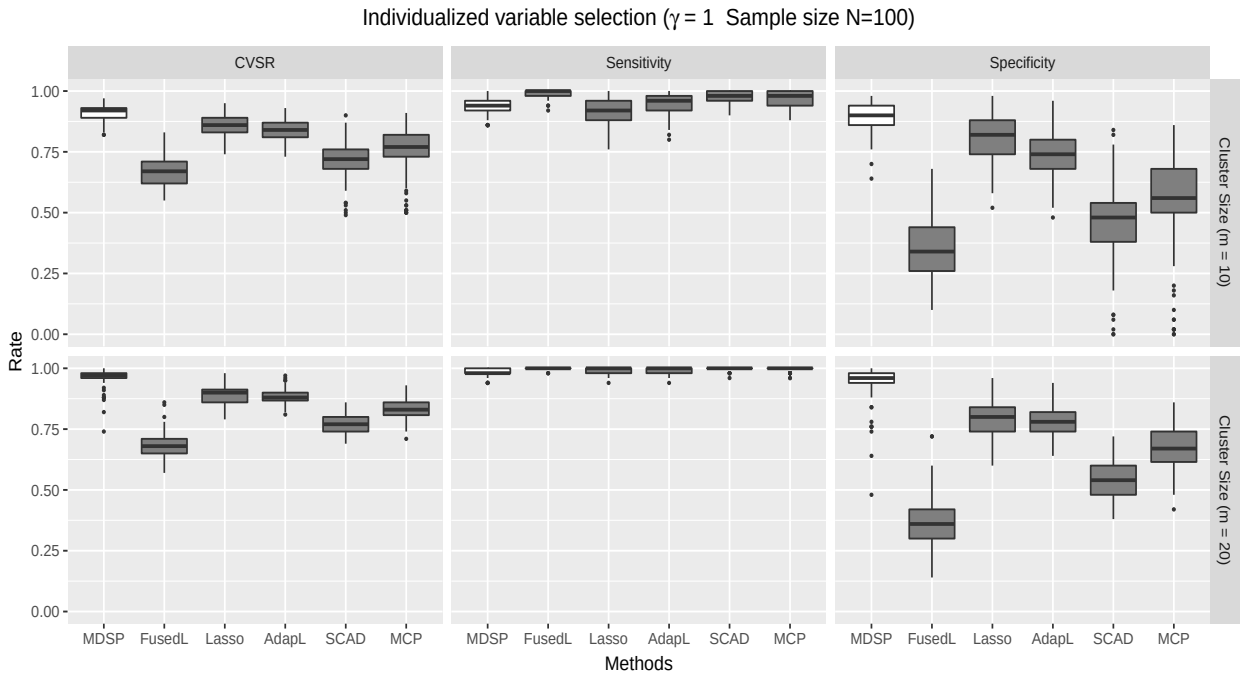

Figure 3.6: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 1$ and sample size $N = 100$.

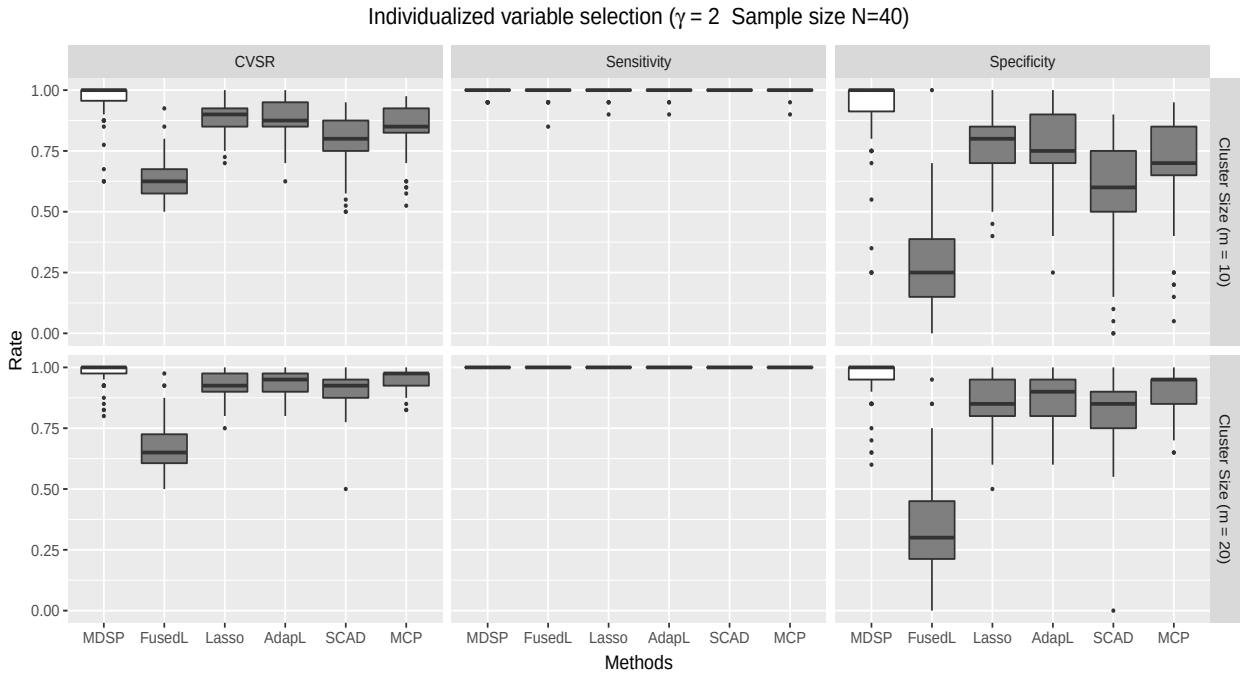

Figure 3.7: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 2$ and sample size $N = 40$.

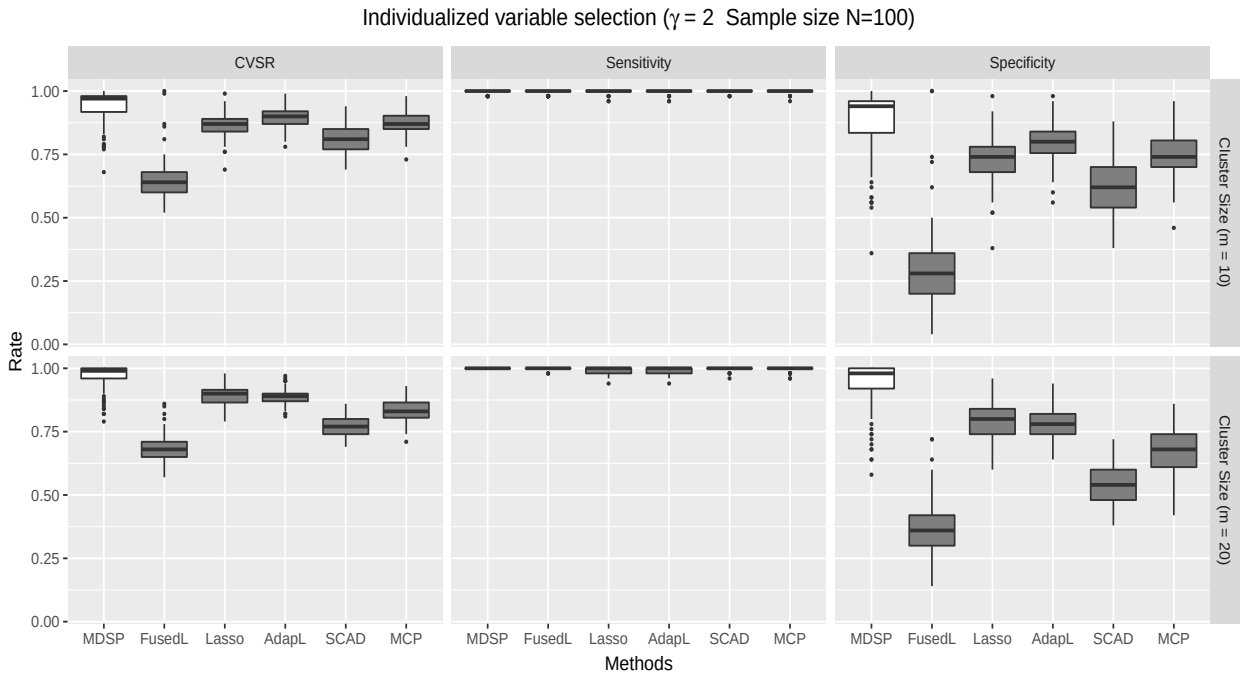

Figure 3.8: The boxplots of CVSR, sensitivity and specificity for all regularization approaches based on 100 simulations, with cluster size $m = 10, 20$, where homogeneous effect $\gamma = 2$ and sample size $N = 100$.

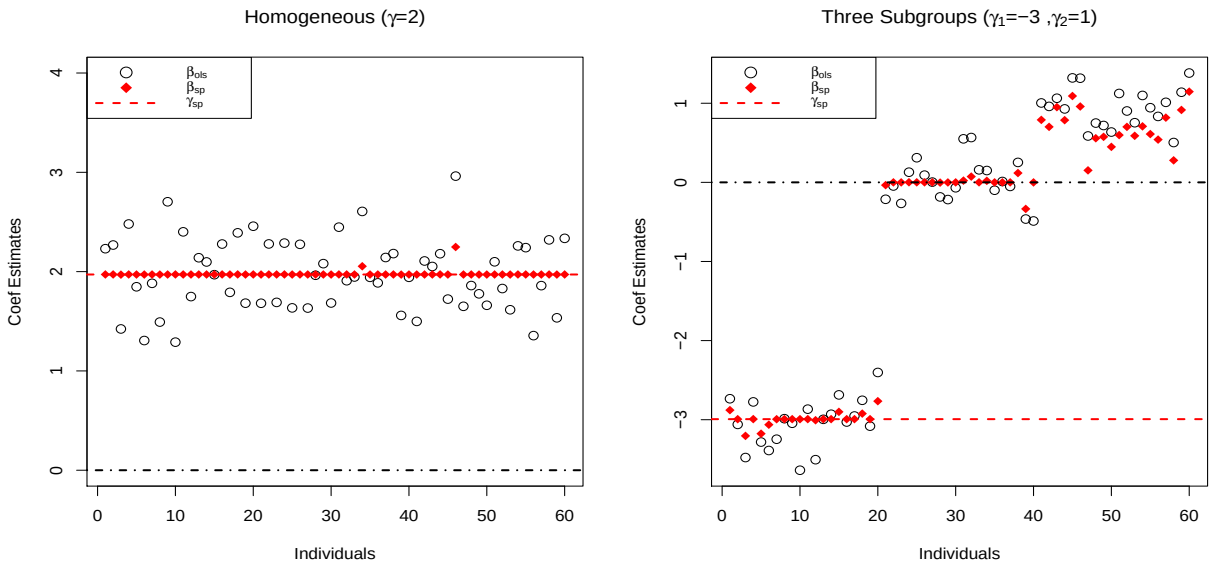

Figure 3.9: The subject-wise least squares estimator and the proposed estimator assuming two subgroups (including a zero group) for individualized parameters in two scenarios: a homogeneous group, and three subgroups, where the sample size $N = 60$ and cluster size $m = 10$.

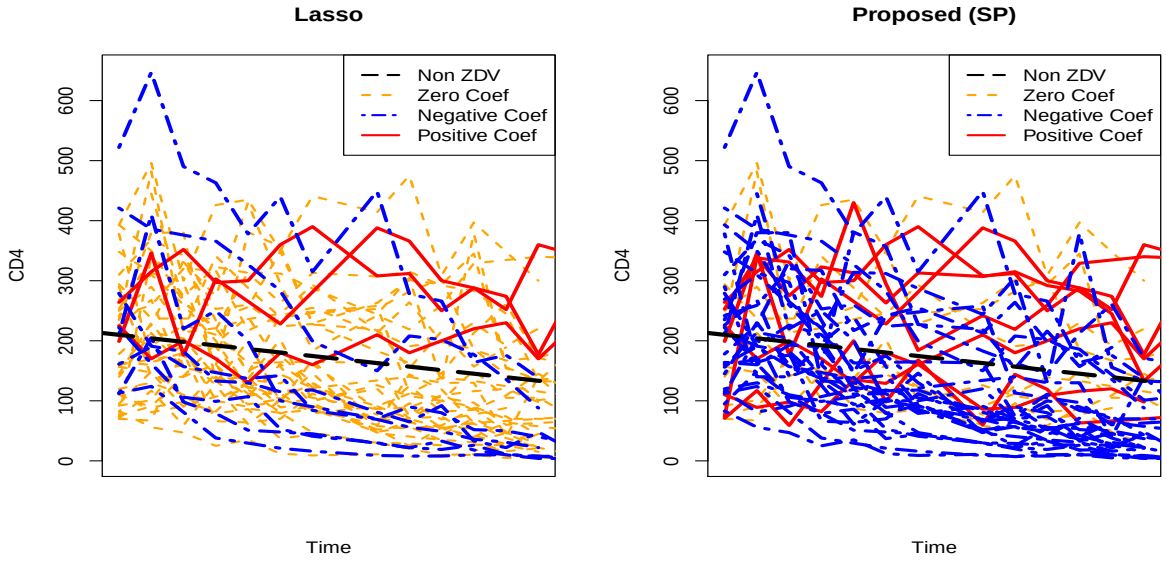

Figure 3.10: The different individuals corresponding to no effect, positive effect and negative effect in the treatment group selected by the Lasso model and the proposed method.

Chapter 4

Individualized Multi-layer Tensor Learning

4.1 Introduction

In recent years, imaging analysis has experienced an explosive growth due to high demand and applications in biomedical imaging for diagnosing disease status and assessing treatment outcomes. The biomedical applications of imaging analyses are especially powerful in cancer radiotherapy and neuroimaging [18, 50, 9, 14, 98, 42]. In addition, advanced technology developing in imaging analysis can be utilized to better understand structures of the human body associated with biological, psychological and clinical traits, [46, 41, 20, 27, 34, 71], so cancer and chronic diseases can be diagnosed earlier, and intervention treatments can be implemented.

The imaging data are usually stored and formatted as a multi-dimensional array (also known as a multi-order tensor), which is quite different from standard vector data since the tensor data have higher-order structures and contain rich spatial information. Moreover, the tensor covariates could contain other dimensions of information in addition to imaging, for example, a dimension of time in time-course imaging or a dimension of multimodality imaging. This imposes great challenges in developing new statistical tools to effectively extract essential imaging features associated with disease outcomes and to utilize information from multiple sources jointly.

Traditional regression methods treat covariates as vectors. However, transforming an imaging array to a vector becomes infeasible when an array size is large. For instance, a four-modality breast cancer imaging size of $4,000 \times 4,000$ for each modality implicitly requires $4 \times 4,000^2 = 6.4 \times 10^7$ regression parameters. This leads to an ultra-high dimensional problem which could be computationally infeasible and non-scalable. Most critically, vectorization of an array is not

capable of preserving the spatial structure of imaging, and therefore spatial dynamics information is not captured entirely.

Under the vectorization framework, Bayesian variable selection approaches are developed for high-dimensional imaging regression models to identify important regions by applying Markov random field priors to account for the spatial correlation between voxels, e.g., [58, 6, 7, 13, 42]. However, this increases complexity of the prior rapidly when the tensor order and dimension increases, and could be computationally infeasible if the tensor order is high. Alternatively, functional data analysis can be adopted to construct a two-dimensional image predictor [65] in a functional regression model. Nevertheless, the extension to three dimensions or beyond could be impractical due to high-dimensional parameters arising from higher-order imaging data [98].

Recent major developments of imaging analyses employ a two-stage strategy, that is, perform a dimension reduction such as principle component analysis (PCA) in the first step, and fit a regression model based on the extracted principal components in the second step [9]. [96] propose a two-stage multi-modal and multi-task learning method by support vector machine (SVM). One critical issue is that the PCA is an unsupervised dimension reduction technique, that is, the extracted principal components or features might not be relevant to the responses. In addition, [98, 43] propose a tensor regression model, where the coefficients associated to imaging voxels are formulated as a tensor and assumed to be low-rank. However, they require locations of imaging voxels to be fixed for different individuals, which are not applicable for breast cancer imaging.

One unique aspect of breast cancer imaging is that different individuals might have breast imaging at very different locations, which is quite different from brain imaging where the target locations of the brain are typically known and fixed. This creates a technical difficulty in that the imaging background and the signal location could vary for different individuals. In fact, the random-located-signal structure is equivalent to the case with very weak signals as it cannot cumulate enough information if the sample size is not large.

Another prevalent tool for imaging analysis, especially on image recognition and classification, is the deep learning method such as the convolutional neural network (CNN). In addition to input

and output layers, a CNN contains multiple hidden layers which are either convolutional, pooling or fully connected, which employs filters to reduce parameter dimension and extract local features. However, without massive data, the deep learning approach cannot be well trained as it does not pre-specify any structures. Thus a CNN could be lack of prediction power when training sample size is not sufficiently large or imaging signals are not strong enough which is common at early stage of disease.

In this chapter, we develop an individualized multilayer model utilizing an additional layer of individual imaging structure and construct a high-order tensor decomposition to reduce the dimensionality of features shared by populations. In addition, we propose an even higher order tensor representation to integrate information from multimodality imaging data, which enables us to capture the spatial structures of microvesicles associated with an early cancer stage.

The strength of the proposed approach is that we are able to capture the locations of microvesicles more accurately when the locations of signals vary for different individuals, which is quite common among breast cancer patients. In addition, the individualized multilayer tensor model is able to identify the magnitude of microvesicle density more effectively compared to existing methods. This also implies medical and clinical significance when there is an indication of association between tumor-associated microvesicles (TMVs) and development of carcinogenesis.

The conventional tensor decomposition methods for feature extraction [49, 3] such as the CANDECOMP/PARAFAC (CP) decomposition assume common-shared parallel factors for any dimension of the tensor, which could be inefficient due to the complex tensor data structure. For instance, in a higher-order multimodality imaging tensor, each modality of imaging could have its unique modality-specific structure, which varies significantly for different modalities. In contrast, the proposed method allows different structures (layers) along different dimensions, e.g., different dimensions for individuals and modalities. Consequently, the proposed method is capable of capturing important tensor features from different layers effectively and achieving dimension reduction more efficiently, thus enhancing the prediction power.

Another advantage of the proposed method is that the extracted individuated layer is able to

effectively integrate multi-modality information as it is shared by different image modalities from the same subject. Although many work have been made to deal with multimodal images [96], to best of our knowledge, most of existing methods combine the information extracted independently from each image modality in a prediction stage, which could be inefficient in our situation that the signal strength is very weak within each modality compared to the modality-specific background noise.

This chapter is organized as follows. Section 4.2 introduces the notation and background model framework. Section 4.3 presents the proposed method and establishes theoretical results. Section 4.4 proposes a scalable parallel algorithm. Section 4.5 provides simulation studies. Section 4.6 illustrates an application for breast cancer optical imaging data. The last section provides concluding remarks and discussion.

4.2 Background and Framework

4.2.1 Notation

In this section, we provide the notation and formulations of tensor. A Dth-order (or D-way) tensor is a D-dimensional array $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$, where the order of a tensor is defined as the number of dimensions, and is also known as ways or modes [37]. In this chapter, we use the terms ways, modes and order interchangeably. For example, a vector is a one-way (first-order or one-mode) tensor and a matrix is a two-way tensor. In addition, a fiber is defined as a vector by fixing every index but one in a tensor [38], which is analogue to a matrix row or a column. Moreover, p_d ($d = 1, \ldots, D$) is the marginal dimension of each mode, or the length of the corresponding fiber.

In the following, we introduce tensor operations. We denote $\text{vec}(\cdot)$ as a vectorizing operation which converts a tensor to a vector, where the element $x_{i_1,\ldots,i_D}$ in a D-way tensor $\boldsymbol{X}$ is turned to be the $\left(i_1 + \sum_{d=2}^{D}[(i_d - 1)\prod_{j=1}^{d-1} p_j]\right)$th element in the long vector $\text{vec}(\boldsymbol{X})$. In addition, the inner

product $\langle \cdot, \cdot \rangle$ of two tensors with the same dimension is defined as

$$\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \langle \text{vec}\boldsymbol{A}, \text{vec}\boldsymbol{B} \rangle = \sum_{i_1=1}^{p_1} \sum_{i_2=1}^{p_2} \cdots \sum_{i_D=1}^{p_D} a_{i_1 i_2 \ldots i_D} b_{i_1 i_2 \ldots i_D}.$$

It follows immediately that $\langle \boldsymbol{A}, \boldsymbol{A} \rangle = \| \boldsymbol{A} \|_F^2$, where $\| \cdot \|_F$ is the Frobenius norm defined as the square root of the sum of the squares of all elements. Moreover, an outer product "$\circ$" operating on multiple vectors $\boldsymbol{b}^{(1)} \in \mathbb{R}^{p_1}, \ldots, \boldsymbol{b}^{(D)} \in \mathbb{R}^{p_D}$ creates a D-way tensor

$$\boldsymbol{X} = \boldsymbol{b}^{(1)} \circ \boldsymbol{b}^{(2)} \circ \ldots \circ \boldsymbol{b}^{(D)},$$

where the $(i_1, i_2, \ldots, i_D)$th element of $\boldsymbol{X}$ is defined as $x_{i_1,\ldots,i_D} = b_{i_1}^{(1)} b_{i_2}^{(2)} \ldots b_{i_D}^{(D)}$, and $\boldsymbol{b}^{(1)}, \ldots \boldsymbol{b}^{(D)}$ do not need to have the same dimension.

Consequently, a D-way tensor $\boldsymbol{X} \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$ is rank K if it can be represented as $\boldsymbol{X} = \sum_{k=1}^{K} \boldsymbol{b}_k^{(1)} \circ \boldsymbol{b}_k^{(2)} \circ \ldots \circ \boldsymbol{b}_k^{(D)}$, where $\boldsymbol{b}_k^d$'s ($k = 1, \ldots, K$) are p_d-dimensional vectors ($d = 1, \ldots, D$). We denote $\boldsymbol{B}^d = [\boldsymbol{b}_1^d, \boldsymbol{b}_2^d, \ldots, \boldsymbol{b}_K^d] \in \mathbb{R}^{p_d \times K}$ as rank-K bases on mode d.

4.2.2 Background of the Two-Stage Model

One way to model an association between the image tensor predictors and the outcome response is through a generalized linear model [51]. That is,

$$g(\mu_i) = \boldsymbol{\alpha}^T \boldsymbol{Z}_i + \boldsymbol{\beta}^T f(\boldsymbol{X}_i), \tag{4.1}$$

where $g(\cdot)$ is a link function, $\mu_i = \mathbf{E}(y_i)$, and $\boldsymbol{Z}_i$ is a common vector covariate. Alternatively, we can employ a machine learning model such as the support vector machine (SVM) for binary outcomes:

$$\min \sum_{i=1}^{n} \Big(1 - y_i(\boldsymbol{\alpha}^T \boldsymbol{Z}_i + \boldsymbol{\beta}^T f(\boldsymbol{X}_i)) \Big)_+. \tag{4.2}$$

For both models above, the D-way tensor predictor $\boldsymbol{X}_i$ is incorporated through an appropriate feature extraction or transformation procedure $f(\cdot)$. One naive transformation method is to convert

$\boldsymbol{X}_i$ to a vector via $f(\boldsymbol{X}_i) = \text{vec}(\boldsymbol{X}_i)$. However, the number of unknown parameters using a vector is $p_{\boldsymbol{\alpha}} + \prod_{d=1}^{D} p_d$, which is ultra-high dimensional and leads to an estimable model.

A natural solution to solve this problem is to employ a dimension reduction technique to extract important features from the tensor predictor at the first stage, and then fit the model in either (4.1) or (4.2) based on the extracted information $f(\boldsymbol{X}_i)$. In the first dimension reduction step, we consider a low-rank approximation for tensor $\boldsymbol{X}_i$ as

$$\boldsymbol{X}_i = \sum_{k=1}^{K} w_{ik} \boldsymbol{B}_k, \quad i = 1, \ldots, N,$$

where N is the sample size, and $\boldsymbol{B}_k$'s ($k = 1, \ldots, K$) are regularized D-way tensor bases shared by populations. Given a rank K, the tensor bases can be obtained by minimizing the difference between the observed imaging and approximated values:

$$\min_{\{\boldsymbol{B}_k, \boldsymbol{w}_k\}_{k=1}^{K}} \sum_{i=1}^{N} \parallel \boldsymbol{X}_i - \sum_{k=1}^{K} w_{ik} \boldsymbol{B}_k \parallel_F^2, \quad \text{s.t.} \quad \parallel \boldsymbol{B}_k \parallel_F = 1, \quad k = 1, \ldots, K, \tag{4.3}$$

where $\boldsymbol{w}_k = (w_{1k}, \ldots, w_{Nk})'$ is the loading vector. Note that $\parallel \boldsymbol{X}_i \parallel_F = \parallel \text{vec}(\boldsymbol{X}_i) \parallel_F$. The number of unknown parameters to be estimated in (4.3) is $K(N + \prod_{d=1}^{D} p_d - 1)$, which is still ultra-large if the order of the tensor is high.

To overcome the computational difficulty, [9] propose a dimension reduction technique for fMRI image data, which converts fMRI images to a matrix with one dimension as a vectorizing MRI image and the second dimension as time. By mimicking [9]'s approach, we obtain a set of common bases for individual tensor predictors by using a marginal principal component analysis (MPCA) technique.

Specifically, for a two-way individual tensor (matrix), we first apply singular value decomposition on each individual matrix such that $\boldsymbol{X}_i = \boldsymbol{U}_i \boldsymbol{W}_i \boldsymbol{V}_i^T$, and select the first K_s component vectors $[\boldsymbol{u}_{i1}, \ldots, \boldsymbol{u}_{iK_s}]$ and $[\boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{iK_s}]$ from $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$, respectively. Then we apply principal component analysis on each set of combined individual component vectors $\{\boldsymbol{u}_{ik}\}_{i=1,\ldots,N;k=1,\ldots,K}$ and $\{\boldsymbol{v}_{ik}\}_{i=1,\ldots,N;k=1,\ldots,K}$, and generate two sets of eigen-bases $[\boldsymbol{e}_1^u, \ldots, \boldsymbol{e}_{K_p}^u]$ and $[\boldsymbol{e}_1^v, \ldots, \boldsymbol{e}_{K_p}^v]$. Con-

sequently, the common bases for two-way tensors are obtained by $\boldsymbol{B}_k = \boldsymbol{e}_k^u \circ \boldsymbol{e}_k^v$, $k = 1, \ldots, K_p$.

Note that any higher-order tensor can be transformed to a two-way matrix, and the MPCA is applicable for the transformed two-way matrix. However, this kind of transformation is usually arbitrary without knowing which part of the tensor should be converted to a vector, and thus making the MPCA rather limited.

4.3 Proposed Method

4.3.1 Individualized Multilayer Model

In this section, we propose a novel individualized multilayer method for tensor feature extraction based on high-order decomposition. Here we assume that each individual has a D-way tensor covariate $\boldsymbol{X}_i \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$. We integrate all individual tensors to a higher-order $(D+1)$-way grand tensor $\boldsymbol{X} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N] \in \mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D \times N}$. Analogous to singular value decomposition, we allow the $(D+1)$-way grand tensor $\boldsymbol{X}$ to follow a low-rank structure as

$$\boldsymbol{X} = \sum_{k=1}^{K} \boldsymbol{b}_k^{(1)} \circ \boldsymbol{b}_k^{(2)} \circ \ldots \circ \boldsymbol{b}_k^{(D)} \circ \boldsymbol{w}_k, \tag{4.4}$$

where $\boldsymbol{b}_k^{(d)}$'s ($k = 1, \ldots, K$) are p_d-dimensional vectors ($d = 1, \ldots, D$), and $\boldsymbol{w}_k$'s are N-dimensional vectors representing the dimension of sample individuals. This tensor rank decomposition is also known as CANDECOMP/PARAFAC decomposition, or Kruskal decomposition [40]. To ensure identifiability, we require $\| \boldsymbol{b}_k^{(d)} \|_F = 1$ for all k's and d's. Note that the CP decomposition of a tensor is not always unique, except under certain conditions [77].

An alternative tensor decomposition is high-order singular value decomposition, also called Tucker decomposition [85], which decomposes a tensor into a D-way core tensor associated with D orthonormal bases matrices. However, the core tensor in Tucker decomposition is not guaranteed to be diagonal, and thus the tensor rank is not estimable. Therefore we adopt the CP decomposition

here. The rank-K CP decomposition in (4.4) is obtained by minimizing

$$\min_{\{\boldsymbol{b}_k^{(1)},\ldots,\boldsymbol{b}_k^{(D)},\boldsymbol{w}_k\}_{k=1}^K} \| \boldsymbol{X} - \sum_{k=1}^{K} \boldsymbol{b}_k^{(1)} \circ \boldsymbol{b}_k^{(2)} \circ \ldots \circ \boldsymbol{b}_k^{(D)} \circ \boldsymbol{w}_k \|_F^2 . \tag{4.5}$$

Since $\| \boldsymbol{b}_k^{(1)} \circ \boldsymbol{b}_k^{(2)} \circ \ldots \circ \boldsymbol{b}_k^{(D)} \|_F = \prod_{d=1}^{D} \| \boldsymbol{b}_k^{(d)} \|_F = 1$, it can be shown that (4.3) is equivalent to (4.5) if the basis tensor $\boldsymbol{B}_k$ in (4.3) admits a rank-one decomposition as $\boldsymbol{B}_k = \boldsymbol{b}_k^{(1)} \circ \boldsymbol{b}_k^{(2)} \circ \ldots \circ \boldsymbol{b}_k^{(D)}$. However, the number of unknown parameters in (4.5) is $K\{N + \sum_{d=1}^{D}(p_d - 1)\}$, which is much smaller than that in (4.3) when the order D and the marginal dimensions p_d's are large. Consequently, the tensor covariate $\boldsymbol{X}_i$ is reduced to $f(\boldsymbol{X}_i) = \boldsymbol{w}_i = (w_{i1}, \ldots, w_{iK})'$, which can be fitted in a prediction model as (4.1) or (4.2).

The higher-order CP decomposition (HOCPD) described in (4.5) is powerful for reducing the tensor covariates' dimensionality; however, it depends on the low-rank assumption in order to achieve an integrated high-order tensor. The high-order tensor decomposition methods could fail to capture complex tensor data information if there is significant heterogeneous variation arising from different individuals. In the following, we propose an individualized multilayer tensor learning (IMTL) method. For the ith individual, we assume

$$\boldsymbol{X}_i = \sum_{k=1}^{K} w_{ik}\boldsymbol{B}_k + u_i\boldsymbol{S}_i, \text{ s.t. } \quad \langle \boldsymbol{B}_k, \boldsymbol{S}_i \rangle = 0, \quad k = 1, \ldots, K, \tag{4.6}$$

where $\boldsymbol{B}_k = \boldsymbol{b}_k^{(1)} \circ \boldsymbol{b}_k^{(2)} \circ \ldots \circ \boldsymbol{b}_k^{(D)}$ is the population-shared basis, and $\boldsymbol{S}_i = \boldsymbol{s}_i^{(1)} \circ \boldsymbol{s}_i^{(2)} \circ \ldots \circ \boldsymbol{s}_i^{(D)}$ is an individualized rank-1 basis with $\boldsymbol{s}_i^{(d)} \in \mathbb{R}^{p_d}$ $(d = 1, \ldots, D)$. An orthogonal constraint is imposed between homogeneous bases in (4.4) and heterogeneous bases in (4.6) to guarantee the identifiability between these two different layers. Therefore the extracted information for tensor covariates based on the IMTL method is $f(\boldsymbol{X}_i) = (\boldsymbol{w}_i, u_i, \text{vec}(\boldsymbol{S}_i)')'$.

Through (4.6), each tensor covariate $\boldsymbol{X}_i$ can be represented by two different layers, one layer consisting of a linear combination of homogeneous structures ($\boldsymbol{B}_k$'s), and the other layer containing an individualized structure ($\boldsymbol{S}_i$) capturing the heterogeneity of individual features. In addition,

the individualized basis $\boldsymbol{S}_i$ is likely to contain rich spatial imaging information for different individuals.

We can further fit a generalized linear model or a SVM with imaging information from both layers. For example, the fitted generalized linear model is

$$g(\mathbb{E}[y_i]) = \boldsymbol{\alpha}^T \boldsymbol{Z}_i + \boldsymbol{\beta}^T f(\boldsymbol{X}_i) = \boldsymbol{\alpha}^T \boldsymbol{Z}_i + \boldsymbol{\beta}_w^T \boldsymbol{w}_i + \boldsymbol{\beta}_u^T \boldsymbol{u}_i, \tag{4.7}$$

where $\boldsymbol{u}_i = (u_i, \text{vec}(\boldsymbol{S}_i)')'$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_w', \boldsymbol{\beta}_u')'$. In addition, we can impose additional penalties for feature selection when the sample size N is much smaller than the total number of the parameters $(p_\alpha + K + \prod_{d=1}^{D} p_d + 1)$, where p_α is the dimension of the vector-based predictor $\boldsymbol{Z}_i$.

4.3.2 Generalization for Multimodality Tensor

Multimodality imaging produces multiple images using different wavelengths of light from a single examination and is widely adopted in optical imaging. In particular, the multimodality multiphoton imaging technique [84] is capable of generating imaging in different modalities at tissue, cellular and molecular scales. Although a single modality model is applicable, there is a critical need to combine all information collected from multimodality imaging, so important features associated with disease status and clinical outcomes can be extracted effectively. In addition, through combining multimodal data, we can capture the spatial information shared by different modalities from the same individual. In the following, we develop multimodality imaging tensor model which extends the individualized multilayer tensor model to incorporate different sources of modality information.

We consider a M-modality tensor covariate $\boldsymbol{X}_i = [\boldsymbol{X}_i^1, \ldots, \boldsymbol{X}_i^M]$, where each signal-modality tensor $\boldsymbol{X}_i^m$ $(m = 1, \ldots, M)$ has the same size $\mathbb{R}^{p_1 \times p_2 \times \ldots \times p_D}$. For the purpose of feature extraction,

we propose an individualized multilayer model for the mth modality:

$$\boldsymbol{X}_i^m = \sum_{k=1}^{K_m} w_{ik}^m \boldsymbol{b}_k^{m,(1)} \circ \ldots \circ \boldsymbol{b}_k^{m,(D)} + u_i^m \boldsymbol{s}_i^{(1)} \circ \ldots \circ \boldsymbol{s}_i^{(D)} = \sum_{k=1}^{K_m} w_{ik}^m \boldsymbol{B}_k^m + u_i^m \boldsymbol{S}_i, \quad \text{s.t.} \quad \langle \boldsymbol{B}_k^m, \boldsymbol{S}_i \rangle = 0, \tag{4.8}$$

where $\{\boldsymbol{B}_k^m = \boldsymbol{b}_k^{m,(1)} \circ \ldots \circ \boldsymbol{b}_k^{m,(D)}\}_{k=1}^{K_m}$ is the set of bases for the mth modality, $\boldsymbol{S}_i = \boldsymbol{s}_i^{(1)} \circ \ldots \circ \boldsymbol{s}_i^{(D)}$ is a rank-1 individualized basis shared within the ith subject, and K_m is the rank of the mth modality's common bases. The proposed model is estimated by minimizing a sum of squared loss

$$\min_{\{\boldsymbol{w}_i^m, \boldsymbol{B}_k^m, \boldsymbol{u}_i, \boldsymbol{S}_i\}_{i,k,m}} \sum_{i=1}^{N} \sum_{m=1}^{M} \| \boldsymbol{X}_i^m - \sum_{k=1}^{K_m} w_{ik}^m \boldsymbol{B}_k^m - u_i^m \boldsymbol{S}_i \|_F^2, \text{ s.t. } \langle \boldsymbol{B}_k^m, \boldsymbol{S}_i \rangle = 0, \tag{4.9}$$

where $\boldsymbol{w}_i^m = (w_{i1}^m, \ldots, w_{iK_m}^m)'$, and $\boldsymbol{u}_i = (u_i^1, \ldots, u_i^M)'$. Consequently, for this M-modality tensor, the extracted information is $\hat{f}(\boldsymbol{X}_i) = (\boldsymbol{w}_i^{1'}, \ldots, \boldsymbol{w}_i^{M'}, \boldsymbol{u}_i', \text{vec}(\boldsymbol{S}_i)')'$. Figure 4.1 provides an illustration of the individualized layers and the modality-specific layers on the four-modality optical breast cancer images.

In the following we consider a new subject with the M-modality tensor covariate $\boldsymbol{X}_j^* = [\boldsymbol{X}_j^{*1}, \ldots, \boldsymbol{X}_j^{*M}]$. In order to make a prediction of outcome y_j^*, first we obtain its extracted tensor covariate's information following the training model, denoted as $\hat{f}(\boldsymbol{X}_j^*)$. Note that the modality-specific loading $\boldsymbol{w}_j^{*m} = (w_{j1}^{*m}, \ldots, w_{jK_m}^{*m})' \in \mathbb{R}^{K_m}$ for the new subject's tensor predictor can be calculated as a projection corresponding to the estimated modality layers

$$\boldsymbol{w}_j^{*m} = (\hat{\boldsymbol{A}}'^m \hat{\boldsymbol{A}}^m)^{-1} \hat{\boldsymbol{A}}'^m \text{vec}(\boldsymbol{X}_j^{*m}),$$

where $\hat{\boldsymbol{A}}^m = [\text{vec}(\boldsymbol{B}_1^m), \ldots, \text{vec}(\boldsymbol{B}_{K_m}^m)] \in \mathbb{R}^{\prod_{d=1}^{D} p_d \times K_M}$ is the extracted tensor basis matrix for the mth modality from the training model. Next we obtain the individualized layer for the new subject by

$$\min_{u_j^*, \boldsymbol{S}_j^*} \sum_{m=1}^{M} \| \boldsymbol{X}_j^m - \sum_{k=1}^{K_m} w_{jk}^{*m} \boldsymbol{B}_k^m - u_j^{*m} \boldsymbol{S}_j^* \|_F^2,$$

where $\boldsymbol{S}_j^* = \boldsymbol{s}_j^{*(1)} \circ \ldots \circ \boldsymbol{s}_j^{*(D)} \in \mathbb{R}^{p_1 \times \ldots \times p_D}$ has a rank-1 structure. In addition, we have $\hat{f}(\boldsymbol{X}_j^*) =$

$(\boldsymbol{w}_j^{*1'}, \ldots, \boldsymbol{w}_j^{*M'}, \boldsymbol{u}_j^{*'}, \text{vec}(\boldsymbol{S}_j^*)')'$, and the prediction for the new subject provided by the generalized linear model is

$$\mu_j^* = g^{-1}\left(\hat{\boldsymbol{\alpha}}^T \boldsymbol{Z}_j^* + \hat{\boldsymbol{\beta}}^T \hat{f}(\boldsymbol{X}_j^*)\right),$$

where $\mu_j^* = \mathbf{E}[y_j^*]$, $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^{p_\alpha}$ and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{\sum_{m=1}^M (K_m+1)+\prod_{d=1}^D p_d}$ are the estimated coefficients from the training model.

In addition, the individualized layer $\boldsymbol{S}_i$ usually contains important spatial information for the image tensor. However, the traditional tensor decomposition techniques, e.g., the CP decomposition and the Tucker decomposition, are not effective to capture the spatial information, which is a remaining challenge in imaging analysis. To solve this problem, we propose a penalized decomposition model based on (4.9) as follows

$$\min_{\{\boldsymbol{w}_i^m, \boldsymbol{B}_k^m, \boldsymbol{u}_i, \boldsymbol{S}_i\}_{i,k,m}} \sum_{i=1}^N \left(\sum_{m=1}^M \parallel \boldsymbol{X}_i^m - \sum_{k=1}^{K_m} w_{ik}^m \boldsymbol{B}_k^m - u_i^m \boldsymbol{S}_i \parallel_F^2 + \lambda \rho_f(\boldsymbol{S}_i) \right), \text{ s.t. } \quad \langle \boldsymbol{B}_k^m, \boldsymbol{S}_i \rangle = 0, \tag{4.10}$$

where λ is a tuning parameter and $\rho_f(\cdot)$ adopts a fusion-type penalty:

$$\rho_f(\boldsymbol{S}_i) = \sum_{d=1}^D \sum_{j_d=1}^{p_d-1} |s_{ij_d}^{(d)} - s_{i(j_d+1)}^{(d)}|,$$

and $\boldsymbol{s}_i^{(d)} = (s_{i1}^{(d)}, \ldots, s_{ip_d}^{(d)})'$ is the p_d-dim vector for the dth mode of the CP decomposition for $\boldsymbol{S}_i$.

By imposing a fusion penalty on the adjacent components of each mode's decomposition vector $(\boldsymbol{s}_i^{(d)}, d = 1, \ldots, D)$, the proposed penalized tensor decomposition method takes neighboring correlation information into account and thus utilizes the individualized spatial information on multimodal images. Note that the proposed penalized tensor decomposition model is a general method to incorporate spatial information in a tensor decomposition, which is not limited to only individualized layer $\boldsymbol{S}_i$ in this case.

4.3.3 Theoretical Results

In this section, we develop the theoretical foundation for the proposed model. First, we provide necessary and sufficient conditions to achieve the model identifiability. In addition, we show that the proposed method provides an estimated tensor converging to the hypothetical true tensor under regularity conditions, which lays the foundation of estimation consistency.

Before establishing the statistical property, it is important and crucial to deal with the model identifiability issue, which is always challenging for tensor-structure framework. Here we provide some necessary and sufficient conditions to achieve identifiable layers in the proposed multi-layer tensor model, which are also easy to check in practice. The following discussion focuses on the single modality model while the presented conditions could be easily extended to the multi-modality model.

In the proposed framework, the unidentifiability lies in the multi-layer CP decomposition of tensor predictors in 4.6, which is attributed to three aspects. The first two aspects are indeterminacies of scaling and permutation, The last aspect is the non-uniqueness of the CP decomposition for a tensor, which means that there might be more than one possible combination of population layers and individualized layers that sum to the underlying true image tensor.

The proposed single-modality individual tensor consists of $R+1$ layers: $\boldsymbol{X}_i = \sum_{r=1}^{R} w_{ir}\boldsymbol{B}_r + u_i\boldsymbol{S}_i$, where $\boldsymbol{B}_r = \boldsymbol{b}_r^{(1)} \circ \cdots \circ \boldsymbol{b}_r^{(D)}$ and $\boldsymbol{S}_i = \boldsymbol{s}_i^{(1)} \circ \cdots \circ \boldsymbol{s}_i^{(D)}$. The scaling indeterminacy refers to the fact that we can arbitrarily scale the factor vectors of different modes of each layer since

$$\boldsymbol{B}_r = a^{(1)}\boldsymbol{b}_r^{(1)} \circ \cdots \circ a^{(D)}\boldsymbol{b}_r^{(D)}$$

holds as long as $\prod_{d=1}^{D} a^{(d)} = 1$. Therefore we impose a unit-norm constraint on parameterization, that is, $\|\boldsymbol{b}_r^{(d)}\|_2 = 1$ ($d = 1, \ldots, D$; $r = 1, \ldots, R$) and $\|\boldsymbol{s}_i^{(d)}\|_2 = 1$ ($d = 1, \ldots, D$; $i = 1, \ldots, N$), to eliminate the indeterminacy from the scaling issue. In addition, the permutation indeterminacy comes from the arbitrary reordering of the population layers. To treat the permutation indeterminacy, we could align the population layers according to a descending order of the first element of

mode-1 factor vectors, that is, requiring $\boldsymbol{b}_1^{(1)}[1] \geq \boldsymbol{b}_2^{(1)}[1] \geq \ldots \geq \boldsymbol{b}_R^{(1)}[1]$.

After controlling the scaling and the permutation, in general, the CP decomposition for a tensor may still not be unique. It is possible to have more than one combination of rank-one tensors that sums to $\boldsymbol{X}_i$. Note that the individualized layer is of rank one. The following lemma guarantees the uniqueness of the CP decomposition of any rank-1 tensor, which implies the identifiability of the individualized layer given the population layers.

Lemma 7. *Any rank-1 CP decomposition for a D-way ($D \geq 2$) tensor is unique up to only scaling indeterminacy.*

Next we introduce the concept of k-rank of a matrix, which is introduced by [40]. Specifically, the k-rank of a matrix $\boldsymbol{A}$, denoted as $\mathcal{K}_{\boldsymbol{A}}$ is defined as

$$\mathcal{K}_{\boldsymbol{A}} = \max\{k : \text{ any } k \text{ columns of } \boldsymbol{A} \text{ are linearly independent}\}.$$

Let $\tilde{\boldsymbol{B}}_i^{(d)} = [\boldsymbol{b}_1^{(d)} \ \cdots \ \boldsymbol{b}_R^{(d)} \ \boldsymbol{s}_i^{(d)}]$ denote the mode-d factor matrix of individual tensor predictor $\boldsymbol{X}_i$. The next proposition provides a necessary and a sufficient condition on model identifiability following the standard results from [40, 77].

Proposition 2. *i) (sufficient condition) For the D-way tensor $\boldsymbol{\mathcal{X}}_i$ in 4.6 ($1 \leq i \leq N$), the multi-layer CP decomposition is unique up to scaling and permutation if*

$$\sum_{d=1}^{D} \mathcal{K}_{\tilde{\boldsymbol{B}}_i^{(d)}} \geq 2R + D + 1;$$

ii) (necessary condition) $R \leq \min_{1 \leq d \leq D} (\prod_{m \neq d} \mathcal{K}_{\boldsymbol{B}^{(m)}} + 1) - 1$.

The conditions above are easy-to-check in numerical studies. Combined with Lemma 1, the following corollary provides a more straightforward condition.

Corollary 5. *For any D-way tensor $\boldsymbol{X}_i$ ($1 \leq i \leq N$), where $D \geq 3$, if the factor matrix $\tilde{\boldsymbol{B}}_i^{(d)}$ has a full k-rank at each mode, that is, $\mathcal{K}_{\tilde{\boldsymbol{B}}_i^{(d)}} = R+1$ for $1 \leq d \leq D$; then the the decomposed layers in 4.6 are unique up to scaling and permutation.*

In the case of $D = 2$, where the individual predictor is of a matrix structure, the above condition does not hold. One way to solve the identifiability issue is to impose additional orthogonal constraints between factor vectors within each mode, which is analogous to the singular value decomposition. However, this brings additional computation cost. Moreover, Proposition 1 has to check all individual tensors separately, which is not effective in practice, especially when sample size N is increasing. Next we provide a much weaker sufficient condition based on the integrated higher-order tensor without imposing any additional constraints on parameterization.

Let $\boldsymbol{X}_{[1:n]}$ denote an integrated $(D+1)$-way tensor combining n individual tensors. Without loss of generality, we assume $\boldsymbol{X}_{[1:n]} = [\boldsymbol{X}_1 \ \cdots \ \boldsymbol{X}_n]$. Note that there is a $(R+n)$-rank representation for the integrated tensor

$$\boldsymbol{X}_{[1:n]} = \sum_{r=1}^{R} \boldsymbol{w}_r^{[1:n]} \circ \boldsymbol{b}_r^{(1)} \circ \cdots \circ \boldsymbol{b}_r^{(D)} + \sum_{i=1}^{n} u_i^{[1:n]} \circ \boldsymbol{s}_i^{(1)} \circ \cdots \circ \boldsymbol{s}_i^{(D)},$$

where $\boldsymbol{w}_r^{[1:n]} = (w_{1r}, \ldots, w_{nr})'$ and $\boldsymbol{u}_i^{[1:n]} = (\underbrace{0, \ldots, 0}_{1,\ldots,i-1}, \underbrace{1}_{i}, \underbrace{0, \ldots, 0}_{i+1,\ldots,n})'$. Similarly, we denote $\tilde{\boldsymbol{B}}_{[1:n]}^{(d)} = [\boldsymbol{b}_1^{(d)} \ \cdots \ \boldsymbol{b}_R^{(d)} \ \boldsymbol{s}_1^{(d)} \ \cdots \ \boldsymbol{s}_n^{(d)}]$ as the mode-d factor matrix for then tensor $\boldsymbol{\mathcal{X}}_{[1:n]}$ for $1 \leq d \leq D$. We have the following result providing a sufficient condition for the identifiability of the multi-layer tensor decomposition in 4.9.

Proposition 3. *If there exists n individual tensors ($2 \leq n \leq N$), such that for the integrated high-order tenor $\boldsymbol{X}_{[1:n]}$,*

$$\sum_{d=1}^{D} \mathcal{K}_{\tilde{\boldsymbol{B}}_{[1:n]}^{(d)}} \geq 2R + n + D$$

holds, then the multi-layer decomposition in 4.6 is unique up to scaling and permutation.

The above sufficient condition is weak and easy-to-check as it requires holding only for an arbitrary n. For example, let $n = 2$ and assume that the factor matrices are of full rank, that is,

$\mathcal{K}_{\tilde{\boldsymbol{B}}^{(d)}_{[1:n]}} = R + n$, then the condition in Proposition 2 reduces to $D(R+n) \geq 2R + n + D$, which holds as long as $D \geq 2$.

Next we establish the statistical properties for the proposed estimator. We denote $\boldsymbol{\gamma}$ as the vector of all latent variable parameters. It is straightforward that $\dim(\boldsymbol{\gamma}) = (K+1)(\sum_{d=1}^{D} p_d + NM)$. Let $\boldsymbol{\Theta} = (\theta^m_{i,j_1\cdots j_D})$ and $\theta^m_{i,j_1\cdots j_D} = \sum_{k=1}^{K_m}(w^m_{ik} b^{m,(1)}_{kj_1} \cdots b^{m,(D)}_{kj_D}) + u^m_i s^{(1)}_{ij_1} \cdots s^{(D)}_{ij_D}$. In the proposed model, we assume that $\mathbf{E}[x^m_{i,j_1\ldots j_D}] = \theta^m_{i,j_1\ldots j_D}$, where $x^m_{i,j_1\ldots j_D}$ denotes an element of the observed tensor, for example, a pixel in an image.

In practice, each pixel of a tensor image can only range from white to black and is usually normalized. Hence, it is sensible to assume that $\|\boldsymbol{\Theta}\|_\infty \leq C_0$ and $\|\boldsymbol{\gamma}\|_\infty \leq C_1$ for large constants $C_0 \geq 0$ and $C_1 \geq 0$. One challenge to our theory derivation is that the proposed individualized layer of tensor recovery implicitly assumes that the number of parameters grows as the number of subjects increases. Therefore we impose a condition such that the parameter space is restricted based on the regularization function. Specifically, let $p(\gamma)$ be a positive penalty function. As the dimension of the parameter space increases as N and M increases, we assume that $p(\gamma) \leq r^2$ and $r = O(\sqrt{(K+1)(\sum_{d=1}^{D} p_d + NM)})$. Then we define the vector parameter space

$$S_\Theta(r) = \{\theta : \|\boldsymbol{\Theta}\|_\infty \leq C_0, p(\gamma) \leq r^2\}$$

and

$$S_\gamma(r) = \{\gamma : \|\boldsymbol{\gamma}\|_\infty \leq C_1, p(\gamma) \leq r^2\}.$$

This controls the overall degree of freedom for parameters. In addition, other regularizations include the orthogonality requirement $\langle \boldsymbol{B}^m_k, \boldsymbol{S}_i \rangle = 0$ imposed in Section 3.2, the L_1 and L_2 penalty to control weight decay, penalty functions to ensure identifiability, or any combinations of these.

For the $(j_1, \ldots, j_D)$th element of $\mathbf{X}_i^m$, we define the loss function

$$l(\mathbf{\Theta}, x^m_{i,j_1\cdots j_D}) = (x^m_{i,j_1\cdots j_D} - \theta^m_{i,j_1\cdots j_D})^2 \tag{4.11}$$

$$= (x^m_{i,j_1\cdots j_D} - \sum_{k=1}^{K_m} w^m_{ik} b^{m,(1)}_{j_1 k} \cdots b^{m,(D)}_{k j_D} - u^m_i s^{(1)}_{ij_1} \cdots s^{(D)}_{ij_D})^2. \tag{4.12}$$

In the following, we assume that the overall criterion function is an additive form of the loss function and the penalty function, that is,

$$L(\mathbf{\Theta}) = \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{j_1} \cdots \sum_{j_D} l(\mathbf{\Theta}, x^m_{i,j_1\cdots j_D}) + \lambda p(\boldsymbol{\gamma}),$$

where λ is a penalization coefficient. Suppose that $\mathcal{S}$ is the parameter space of $\mathbf{\Theta}$, and that

$$\hat{\mathbf{\Theta}}_N = \arg\min L(\mathbf{\Theta}), \tag{4.13}$$

then the following theory provides the consistency of the parameter estimation.

Theorem 7. *For the sample minimizer $\hat{\mathbf{\Theta}}_N$, we have*

$$P(\frac{1}{N}\|\hat{\mathbf{\Theta}}_N - \mathbf{\Theta}_0\|_F \geq \eta_N) \leq 7\exp(-c_1 N\eta_N^2),$$

where $c_1 \geq 0$ is a constant, $\eta_N = \max(\varepsilon_N, \lambda_N^{1/2})$, and $\varepsilon_N = \frac{1}{N^{1/2}}$ is the best possible rate achieved when $\lambda_N \sim \varepsilon_N^2$.

Theorem 7 indicates that the minimizer obtained in (4.13) converges to the true parameter when the sample size goes to infinity and the penalization coefficient goes to zero faster than the best convergence rate. In other words, each element of the estimated tensor converges to the corresponding element of the true tensor. In addition, this theorem is established under the tensor CP decomposition framework which requires a smaller number of parameters, assuming a low rank of the true tensor. Furthermore, we can show a model collapsing multimodality imaging into vectors or matrices leads to a larger number of parameters in order to preserve the true tensor structure,

and hence entails a slower convergence rate than the rate provided in Theorem 7. Furthermore, given the identifiable conditions, Theorem 7 also implies the consistency of parameter estimator $\hat{\boldsymbol{\gamma}}$, that is, $\hat{\boldsymbol{\gamma}} \to \boldsymbol{\gamma}^0$ as $N \to \infty$.

4.4 Implementation

In this section, we propose a two-step alternating least-squares (ALS) algorithm to solve the estimation problem for the proposed IMTL method in (4.9). In contrast to traditional CP decomposition, incorporating individual layers of the proposed method significantly increases the computation cost, and algorithms feasible for tensor decomposition are not necessarily scalable in our situation. In this section, we provide the algorithm for the multimodality data case; the estimation for single-modality data is just a special case with the number of modalities $M = 1$.

The proposed algorithm first estimates the modality-specific layers by minimizing the within-modality loss

$$\min \parallel \boldsymbol{X}^m - \hat{\boldsymbol{X}}^m \parallel_F^2 \quad \text{with} \quad \hat{\boldsymbol{X}}^m = \sum_{k=1}^{K_m} \boldsymbol{b}_k^{m,(1)} \circ \ldots \circ \boldsymbol{b}_k^{m,(D)} \circ \boldsymbol{w}_k^m, \tag{4.14}$$

where $\boldsymbol{X}^m$ is the $(D+1)$-way tensor for the mth modality combing all individuals' tensors, and the $(D+1)$th mode of $\boldsymbol{X}^m$ denotes the dimension of individuals and $\boldsymbol{w}_k^m \in \mathbb{R}^N$. Let $\sum_{k=1}^{K_m} \boldsymbol{b}_k^{m,(1)} \circ \ldots \circ \boldsymbol{b}_k^{m,(D)} \circ \boldsymbol{w}_k^m = \lfloor \boldsymbol{B}^{m,(1)}, \ldots, \boldsymbol{B}^{m,(D)}; \boldsymbol{w}^m \rfloor$, where $\boldsymbol{B}^{m,(d)} = \{\boldsymbol{b}_1^{m,(d)}, \ldots, \boldsymbol{b}_{K_m}^{m,(d)}\} \in \mathbb{R}^{p_d \times K_m}$ is the basis matrix of the dth mode for the mth modality, and satisfies the constraint of $\|\boldsymbol{b}_k^{m,(d)}\|_F = 1$ $(k = 1, \ldots, K_m)$.

To solve the optimization in (4.14), we update one mode's basis matrix by fixing the other modes' parameters, which reduces the problem to a least-squares type of problem at each iteration. For example, to update $\boldsymbol{B}^{m,(1)}$ at the tth iteration $\boldsymbol{B}^{m,(1),[t]}$, the above minimization problem becomes

$$\min_{\boldsymbol{B}^{m,(1)}} \sum_{n=1}^{p_1} \left\| \boldsymbol{X}^m[n,:,\ldots,:] - \sum_{k=1}^{K_m} b_{nk}^{m,(1)} \left(\boldsymbol{b}_k^{m,(2),[t-1]} \circ \ldots \circ \boldsymbol{b}_k^{m,(D),[t-1]} \circ \boldsymbol{w}_k^{m,[t-1]} \right) \right\|_F^2,$$

where $\boldsymbol{X}^m[n,:,\ldots,:]$ is a D-way tensor fixing the index of the first mode, $b_{nk}^{m,(1)}$ is the (n,k)th element of the unknown $\boldsymbol{B}^{m,(1)}$, $\boldsymbol{w}_k^{m,[t-1]}$ and $\boldsymbol{b}_k^{m,(d),[t-1]}$'s are estimated factors of the other modes from the $(t-1)$th iteration. Let $\boldsymbol{B}^{m,(1)}[n,:]$ denote the nth row of $\boldsymbol{B}^{m,(1)}$, $\tilde{\boldsymbol{B}}_k^{m,[t-1]}[-1] = \boldsymbol{b}_k^{m,(2),[t-1]} \circ \ldots \circ \boldsymbol{b}_k^{m,(D),[t-1]} \circ \boldsymbol{w}_k^{m,[t-1]}$, and $\tilde{\boldsymbol{V}}^{m,[t-1]}[-1] = \left[\text{vec}(\tilde{\boldsymbol{B}}_1^{m,[t-1]}[-1]), \ldots, \text{vec}(\tilde{\boldsymbol{B}}_{K_m}^{m,[t-1]}[-1])\right] \in \mathbb{R}^{N\prod_{d=2}^{D} p_d \times K_m}$. Then the minimization is solved by

$$(\tilde{\boldsymbol{B}}^{m,(1),[t]}[n,:])^T = \left((\tilde{\boldsymbol{V}}^{m,[t-1]}[-1])^T \tilde{\boldsymbol{V}}^{m,[t-1]}[-1]\right)^{-1} (\tilde{\boldsymbol{V}}^{m,[t-1]}[-1])^T \text{vec}(\boldsymbol{X}^m[n,:,\ldots,:]) \tag{4.15}$$

for $n = 1, \ldots, p_1$, which only involves the inverse calculation of a $K_m \times K_m$ matrix. The estimations of factors $\tilde{\boldsymbol{B}}^{m,(d)}$'s $(d = 1, \ldots, D)$ corresponding to the other modes follow similarly. Finally, we normalize the columns of $\tilde{\boldsymbol{B}}^{m,(d),[t-1]}$ to obtain $\hat{\boldsymbol{B}}^{m,(d),[t-1]}$; that is, let $\hat{\boldsymbol{b}}_k^{m,(d),[t-1]} = \frac{\tilde{\boldsymbol{b}}_k^{m,(d),[t-1]}}{\|\tilde{\boldsymbol{b}}_k^{m,(d),[t-1]}\|_F}$ $(k = 1, \ldots, K_m)$ and $\hat{\boldsymbol{w}}_k^{m,[t-1]}$ is updated by $\hat{\boldsymbol{w}}_k^{m,[t-1]} \|\tilde{\boldsymbol{b}}_k^{m,(d),[t-1]}\|_F$.

Next we estimate the individualized layers by minimizing the within-subject loss while fixing the modality-specific layers ($\hat{w}_{ik}^{m,[t-1]}$'s and $\hat{\boldsymbol{B}}_k^{m,[t-1]}$'s) estimated from the first step, that is,

$$\min_{\{u_i^m\}_{m=1}^M, \{\boldsymbol{s}_i^{(d)}\}_{d=1}^D} \sum_{m=1}^{M} \|\boldsymbol{X}_i^m - \sum_{k=1}^{K_m} \hat{w}_{ik}^m \hat{\boldsymbol{B}}_k^m - u_i^m \boldsymbol{S}_i\|_F^2, \quad \text{with} \quad \boldsymbol{S}_i = \boldsymbol{s}_i^{(1)} \circ \ldots \circ \boldsymbol{s}_i^{(D)}. \tag{4.16}$$

The individualized parameters u_i^m's $(m = 1, \ldots, M)$ and $\boldsymbol{s}_i^{(d)}$'s $(d = 1, \ldots, D)$ are estimated by employing the ALS algorithm, as in the procedure in estimating modality-specific layers above. The proposed two-step algorithm can be summarized in Algorithm **??**.

The proposed algorithm makes parallel computing feasible in estimating different modality-specific bases at Step 2 and individualized layers at Step 3. In addition, parallel computing can also be applied in calculating different rows of modality-specific factors $\tilde{\boldsymbol{B}}^{m,(d)}$'s in (4.15). This ensures that the computation of the proposed method is highly scalable and efficient.

Algorithm 3 A Two-Step ALS Algorithm with Parallel Computing

1. (*Initialization*) Input all observed $\boldsymbol{X}_i$'s, the rank for each modality K_m's, initial values for $\boldsymbol{B}^{m,(d),[0]}$'s ($m = 1, \ldots, M; d = 1, \ldots, D$). Set $\boldsymbol{S}_i^{(d),[0]} = \mathbf{0}$ and $\boldsymbol{u}_i^{[0]} = \mathbf{0}$ for $i = 1, \ldots, N$, and a stopping criterion $\varepsilon = 10^{-4}$.

2. (*Modality-specific layers*) At the tth iteration, for each modality m ($m = 1, \ldots, M$),

   1. Let $\boldsymbol{X}_i^{m,[t]} = \boldsymbol{X}_i^m - \hat{u}_i^{m,[t-1]} \hat{\boldsymbol{S}}_i^{m,[t-1]}$ and replace $\boldsymbol{X}_i^m$ with $\boldsymbol{X}_i^{m,[t]}$ in (4.15);

   2. Update $\hat{\boldsymbol{B}}^{m,(d),[t]}$ through (4.15) given $\left(\hat{\boldsymbol{B}}^{m,(1),[t]}, \ldots, \hat{\boldsymbol{B}}^{m,(d-1),[t]}, \hat{\boldsymbol{B}}^{m,(d+1),[t-1]}, \ldots, \hat{\boldsymbol{B}}^{m,(D),[t-1]} \right)$ sequentially.

3. (*Individualized layers*). At the tth iteration, for each subject i, update $u_i^{m,[t]}$ and $\boldsymbol{S}_i^{[t]}$ through (4.16) for $i = 1, \ldots, N$.

4. (*Stopping Criterion*). Calculate the fitted tensor $\hat{\boldsymbol{X}}_i^{m,[t]} = \sum_{k=1}^{K_m} \hat{w}_{ik}^{m,[t]} \hat{\boldsymbol{B}}_k^{m,[t]} + \hat{u}_i^{m,[t]} \hat{\boldsymbol{S}}_i^{[t]}$, and the difference between the two latest fitted tensors $\sum_{i=1}^{N} \sum_{m=1}^{M} \| \hat{\boldsymbol{X}}_i^{m,[t]} - \hat{\boldsymbol{X}}_i^{m,[t-1]} \|_F$. Continue the iteration processes in Step 2 and 3 until the difference of the two latest fitted tensors is smaller than the stopping criterion ε.

4.5 Numerical Studies

In this section, we provide simulation studies to illustrate the numerical performance of the proposed method compared with other competing methods. Specifically, we first consider two simulation studies for single modality data in Sections 4.5.1 and 4.5.2, and investigate multimodality imaging in Section 4.5.3.

4.5.1 Simulation A: Random Signal Area

In this study, we simulate random-signal-area imaging data which frequently arise in clinical diagnosis. In the real case, disease is detected through identifying an unusual area (signal area) by medical imaging. The location of the signal area could be restricted within a subregion, but it is typically not fixed.

The $D \times D$ two-way image predictor $\boldsymbol{X}_i$ is generated as $\boldsymbol{X}_i = \boldsymbol{B}_i + \boldsymbol{N}_i$, where $\boldsymbol{B}_i$ is a $D \times D$ sparse feature matrix and $\boldsymbol{N}_i$ is a noise matrix with each component generated from $N(0, 0.1)$.

Specifically, the entries of $\boldsymbol{B}_i$ within a subregion (rows 8-13 and columns 8-13) are independently generated from a Bernoulli distribution with a success probability 0.01 of being 1. The response y_i is set as 1 (disease) if the number of non-zero entries in $\boldsymbol{B}_i$ is greater than zero, or 0 (normal) otherwise, Figure 4.2 illustrates cancer and normal imaging where the location of signals (white color) could randomly appear in a subregion containing 36 pixels compared to the whole imaging with 400 pixels. The outcome rate of cancer is $1 - 0.99^{36} = 0.30$, and the probability of having more than one signal is 0.05.

We set the sample size for the training set to be $N_{tr} = 40$ and the testing set to be $N_{ts} = 100$. The marginal imaging dimension D is set as 20. The simulation results are summarized based on 100 replications.

We compare the proposed individualized multilayer tensor learning model with the higher-order CP decomposition method in (4.5), the marginal principal component analysis method described in Section 4.2, the vectorizing L_1-penalized logistic regression model (VPL), the mean-distance classification method (MeanDist), and the tensor regression (TR) model [98].

The vectorizing L_1-penalized logistic regression model converts the imaging covariate $\boldsymbol{X}_i$ to a $D^2 \times 1$ vector predictor and then fits an L_1-penalized logistic regression model for the binary response, which is implemented by the R package *"penalized."* The mean-distance model calculates the sample mean images $M_1 = \frac{1}{|\mathcal{I}_1|}\sum_{i\in\mathcal{I}_1}\boldsymbol{X}_i$ and $M_0 = \frac{1}{|\mathcal{I}_0|}\sum_{i\in\mathcal{I}_0}\boldsymbol{X}_i$ for the disease group and the control group in the training set, respectively, where $\mathcal{I}_1 = \{i : y_i = 1\}$ and $\mathcal{I}_0 = \{i : y_i = 0\}$, and $|\cdot|$ denotes the set size. For a new subject with an imaging predictor $\boldsymbol{X}_j^*$, the MeanDist model assigns a predicted label as $\hat{y}_j^* = \mathbf{1}_{\{\|\boldsymbol{X}_j^* - M_1\|_F < \|\boldsymbol{X}_j^* - M_0\|_F\}}$. In addition, the tensor regression method is implemented by [97]'s Matlab toolbox *"TensorReg."*

To evaluate the prediction performance, we calculate the overall prediction accuracy rate ($\mathbf{P}[\hat{y} = y^0]$), the sensitivity ($\mathbf{P}[\hat{y} = y^0|y^0 = 1]$) and the specificity ($\mathbf{P}[\hat{y} = y^0|y^0 = 0]$) on the testing test, where y^0 denotes the true label. The tuning parameters associated with the latent rank for the IMTL, the TR, the HOCPD and the MPCA are selected through minimizing the prediction error rates ($1 - \mathbf{P}[\hat{y} = y^0]$) in the validation set.

Table 4.1 provides prediction performance comparisons for various methods. It shows that the proposed IMTL model outperforms all the other methods in prediction. In general, the vectorizing penalization logistic model and the tensor regression model perform inadequately with an overall prediction accuracy below 65%. These two approaches assume that the signal pixels occur at the same locations in different images and share the same association strength with the response through a population coefficient model. However, in this simulation setting, the signal location is random for different images. In addition, Table 4.1 shows that HOCPD, MCPA and MeanDist perform similarly and have better overall accuracy than VPL and TR. This indicates that the dimension-reduction methods are more accurate in capturing relevant features associated with responses in this setting. The proposed IMTL method outperforms the three competitive dimension-reduction methods with about 20% improvement in prediction accuracy. Moreover, the proposed IMTL model achieves 97% overall accuracy, indicating that heterogeneous variations from the individual images indeed contain essential information in predicting disease outcomes.

4.5.2 Simulation B: Multiple Random Weak Signals

In this simulation study, we consider a multiple-random-weak-signal problem motivated by a real data example, which is also quite common in practice. The disease imaging contains a large number of signal pixels with random locations, while the normal imaging has a much smaller number of signal pixels with random locations. This setting mimics breast cancer imaging where cancerous tissue has a lot more tumor-associated microvesicles compared to normal tissue.

Similarly to Simulation A, we generate the $D \times D$ two-way image predictor from $\boldsymbol{X}_i = \boldsymbol{B}_i + \boldsymbol{N}_i$, For each $\boldsymbol{B}_i$, we randomly select S_i entries to be 1 (signal pixels) while the other entries are 0. We generate the response label y_i from a Bernoulli distribution with a probability 0.5. The number of the signal pixels S_i is generated from a Poisson distribution with means μ_C and μ_N for the cancer group ($y_i = 1$) and the normal group ($y_i = 0$), respectively. Figure 4.3 illustrates the cancer and normal images.

We set the training and testing set sizes to be $N_{tr} = 50$ and $N_{ts} = 100$ respectively, the

marginal imaging dimension to be $D = 32$, and $\mu_C = 20, 30$ and $\mu_N = 5$. The prediction results from various methods are summarized based on 100 simulations.

We compare the proposed IMTL method with the five competing methods described in Section 4.5.1. Table 4.2 provides the prediction results, indicating the superior performance of the proposed method in terms of the highest overall accuracy, sensitivity and specificity. Figure 4.3 illustrates that the signals occur throughout the entire region randomly, but each pixel could be very weakly associated with the response outcome. Therefore the VPL method and the TR method perform poorly. In addition, if the disease imaging has more signal quantity than a normal image with $\mu_C = 30$, then the MPCA and the HOCPD methods perform reasonably well with an overall prediction accuracy of 83.4% and 99.2%, respectively. However, if the difference between the disease and normal signal patterns is relatively smaller with $\mu_C = 20$, then the MPCA and the HOCPD methods lose prediction power more rapidly, while the proposed IMTL method still achieves over 96% overall accuracy.

4.5.3 Simulation C: Multimodality Data

In this simulation study, we simulate four-modality imaging data. Four modalities imaged from the same individual share multi-random-weak-signal features as described in Section 4.5.2, where the same modality imaging from different individuals contains its unique background bases. Figure 4.4 illustrates control and disease imaging for each modality.

We simulate the mth-modal image for the ith subject $\boldsymbol{X}_{mi}$ ($D \times D$-dimensional) as $\boldsymbol{X}_{mi} = \boldsymbol{A}_{mi} + \boldsymbol{B}_i + \boldsymbol{N}_i$, $m = 1, \ldots, 4$, where the feature image $\boldsymbol{B}_i$, the noise image $\boldsymbol{N}_i$ and the response label y_i are generated similarly as those in Section 4.5.2. The mean number of signals for cancer imaging μ_C is chosen as 30. The first modality $\boldsymbol{A}_{1i}$ is generated as a full-rank random noise matrix with elements generated from $N(0, 0.5^2)$; the second modality imaging has a uniform background with $\boldsymbol{A}_{2i} = w_{2i}\mathbf{1}_D\mathbf{1}_D^T$, where $\mathbf{1}_D$ is a $D \times 1$ vector of 1's and w_{2i} is generated from an absolute value of $N(0, 0.5^2)$; and both the third and fourth modality imaging have low-rank structures with $\boldsymbol{A}_{mi} = \sum_{k=1}^{5} w_{mik}\boldsymbol{a}_{mk}^{(1)} \circ \boldsymbol{a}_{mk}^{(2)}$ ($m = 3, 4$), where w_{mik}'s are generated from $N(0, 0.5^2)$, $\boldsymbol{a}_{mk}^{(1)}$'s and

$\boldsymbol{a}_{mk}^{(2)}$'s are generated from $N(\mathbf{0}, 0.5^2\mathbf{I}_D)$, and $\mathbf{I}_D$ is the D-dimensional identity matrix. We set the training and testing set sizes as $N_{tr} = 100$ and $N_{ts} = 100$ respectively, and the marginal imaging dimension is $D = 64$.

We compare the proposed individualized multilayer tensor learning method to the VPL method, the MPCA method, the TR method and the HOCPD method. The VPL is applied on a $16,384$-dimensional vector predictor by vectorizing all four modalities. In addition, the MPCA is applied on each individual modality at the first stage and then fits a logistic model with extracted features from all modalities at the second stage. We do not apply the MPCA on the integrated imaging since the multimodality dimension ($D_m = 4$) is not comparable to the marginal imaging dimension ($D = 64$). The HOCPD is applied on the integrated multimodality image (third-order tensor predictor).

Table 4.3 provides the prediction results on the testing set. The proposed method (IMTL) outperforms other methods with the highest prediction accuracy (84.2%) and sensitivity (80.5%). Moreover, both the proposed IMTL method and the HOPCD method have significant advantage over the VPL method, the TR method and the MPCA method which assume that the four modalities of imaging are independent. This indicates that integrating different modalities' information enhances the prediction power. In addition, the proposed IMTL method achieves more than 12.6% improvement in prediction accuracy than the HOCPD, indicating that the proposed method is more effective in utilizing correlation information among different modalities. Note that the basis structures of the imaging tensor ($\boldsymbol{A}_{mi}$) vary significantly among four modalities, and the corresponding ranks are $4,096$, 1, 5 and 5, respectively. Consequently, the HOCPD method assuming a low-rank structure with common bases shared by different modalities might not have sufficient power to extract heterogeneous signal features. In contrast, the proposed IMTL method is able to capture within-subject homogeneous features by utilizing an individualized layer in addition to modality-specific layers.

4.6 Real Data: Multiphoton Imaging Data for Breast Cancer

Our research problems are motivated by multimodality breast cancer imaging data produced by Boppart Lab [84] at University of Illinois at Urbana-Champaign. We applied the proposed method to multiphoton imaging data for breast cancer diagnosis. To better visualize the biological tissue at cellular and molecular levels, [84]'s multiphoton microscope generates multimodal images using two-photon auto-fluorescence (2PAF), three-photon auto-fluorescence (3PAF), second-harmonic generation (SHG) and third-harmonic generation (THG). Two-photon-fluorescence microscopy is commonly used to visualize tissue morphology and physiology at a cellular level, and three-photon-fluorescence with longer wavelength can reach deeper levels of the tissue and thus provide higher resolution [11, 28]. Figure 4.5 illustrates the four modalities of 2PAF and 3PAF, SHG and THG for normal rat's breast tissue and cancerous rat's breast tissue. The new technique is able to identify cancer cell clusters in a specimen which are not easily identified by histology imaging.

Figure 4.5 provides the multiple contrast mechanisms produced by four-modality microscope imaging, which highlight the structural components of tissues. In contrast to the normal rat's tissue, multiple modalities clearly indicate a large number of biological tumor-associated micro-vesicles (circled in Figure 4.5) which appear spatially aligned in a tubular formation on the cancer rat's tissue, particularly in the 3PAF image. In addition, Figure 4.5 also shows that the microvesicles are visible in the THG and 3PAF images, but are not obvious in the SHG and 2PAF images through visualization, indicating that there is a critical need to integrate all modalities for more efficient detection of TMVs using novel statistics and machine learning tools.

Furthermore, although different individuals have imaging at very different locations, different modalities from the same individual are observed from the same tissue and thus share some common structures, which could be informative for capturing the spatial locations and formations of TMVs. Therefore, it is crucial to utilize homogeneous information from multimodality imaging within individuals.

Prior knowledge in cancer detection shows that TMVs are frequently observed at the lipid

boundary area, therefore we study a segmented imaging of 150×150 pixels more closely at the boundary area for both the normal rat and the cancerous rat (see Figure 4.6). In addition, due to the limited sample size at the current experimental stage, we generate more sample images using a resampling technique. Specifically, every original image is segmented into nine subregions with no overlapping, and each subregion has a size of 50×50 pixels. Additional sample images are generated by randomly sampling from the original subregions with replacement as well as adding certain noise to the subregions samples. The noise added to each pixel is generated from $N(0, \sigma^2)$, where σ is set as $\hat{\sigma}_m/5$, and $\hat{\sigma}_m$ ($m = 1, \ldots, 4$) is the sample standard deviation for the mth-modality imaging. Consequently, we generate a training data set and a testing data set with a total of 40 subjects for each data set, where each subject has four-modality images taken in the same subregion.

We compare the proposed IMTL method to the four models described in Section 4.3. The tuning parameters associated with the latent rank for the TR, the MPCA, the HOCPD and the IMTL are selected in order to minimize the prediction error rates in the validation set, which is generated following the same resampling procedure as described above.

Table 4.4 provides the prediction results on the testing set, which illustrates that the proposed method outperforms the other methods significantly in terms of achieving the highest overall prediction accuracy rate, sensitivity and specificity. In addition, both the proposed IMTL and the HOCPD utilizing all four-modality imaging outperform all other methods, showing the prediction power improved by integrating multimodality information. Figure 4.6 displays several common tubular spatial structures of TMVs shared by the tumor imaging on modalities 1, 2 and 4. The proposed IMTL method applying an individualized layer to different imaging modalities from the same subject performs the best for capturing important heterogeneous TMVs' patterns and thus enhances the prediction power for cancer detection. The VPL method and the tensor regression model perform inadequately with prediction accuracy below 55%. This is because the locations of the TMV's vary heterogeneously for different subjects and the signals of the TMV's are weak compared to the modality background (e.g., third modality). Therefore the VPL and the TR are

not powerful at capturing the TMV's effects in predicting disease outcomes.

4.7 Discussion

In this article, we propose an individualized multilayer tensor learning model incorporating imaging covariates to predict targeted responses. In the proposed two-stage model, we first extract important features from tensor covariates incorporating different layers to achieve dimension reduction through tensor decomposition techniques, and then fit a prediction model with the extracted features. We illustrate the proposed method through numerical studies and data application on both single-modality and multimodality imaging data.

A major contribution of the proposed method is that we achieve tensor decomposition through utilizing an additional layer of individual structure in addition to population-shared modality-specific structure following the CANDECOMP/PARAFAC decomposition. Our method is motivated by a multiphoton multimodality imaging study for breast cancer diagnosis, where tumor locations of imaging can vary for different individuals, yet the multimodality images from the same individual share important spatial information. Most existing methods assuming fixed signal locations are either infeasible or inefficient in our setting. In contrast, the proposed individualized layer is capable of capturing within-subject spatial features through integrating different modalities' imaging information for the same individual. Both simulation studies and real data analyses demonstrate that the proposed method can achieve higher diagnostic accuracy compared to other competing methods.

In the proposed method, we only consider a linear transformation for dimension reduction on the tensor data, e.g., the CP decomposition. Due to the complex nature of imaging data, it will be our next step to employ nonlinear transformation techniques such as manifold dimension reduction. Moreover, it is worth future research to develop supervised feature extraction through constructing a constraint tensor decomposition conditional on outcome responses.

4.8 Proof of Theoretical Results

Proof of Theorem 7

The proof can follow Corollary 2 of [74]. For each $X^m_{i,j_1\cdots j_D}$, let $l_d(\mathbf{\Theta}|X^m_{i,j_1\cdots j_D}) = l(\mathbf{\Theta}, X^m_{i,j_1\cdots j_D}) - l(\mathbf{\Theta}_0, X^m_{i,j_1\cdots j_D})$ be the loss difference, where $\mathbf{\Theta}_0$ corresponds to the unique true parameter. We first define:

$$K(\mathbf{\Theta}, \mathbf{\Theta}_0) = \frac{1}{NMp_1\cdots p_D}\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{j_1}\cdots\sum_{j_D}\mathrm{E}\{l_d(\mathbf{\Theta}|X^m_{i,j_1\cdots j_D})\},$$

which is the expected loss difference. Since $\mathbf{\Theta}_0$ is the unique true parameter, we have $K(\mathbf{\Theta}, \mathbf{\Theta}_0) \geq 0$ for all $\mathbf{\Theta} \in \mathcal{S}$ and $K = 0$ if and only if $\mathbf{\Theta} = \mathbf{\Theta}_0$. Therefore, we define the distance between $\mathbf{\Theta}$ and $\mathbf{\Theta}_0$ as $\rho(\mathbf{\Theta}, \mathbf{\Theta}_0) = K^{1/2}(\mathbf{\Theta}, \mathbf{\Theta}_0)$. We also define the variance of the loss difference as follows:

$$V(\mathbf{\Theta}, \mathbf{\Theta}_0) = \frac{1}{NMp_1\cdots p_D}\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{j_1}\cdots\sum_{j_D}\mathrm{Var}\{l_d(\mathbf{\Theta}|X^m_{i,j_1\cdots j_D})\}.$$

Under the L_2-loss, it is expected that $K(\mathbf{\Theta}, \mathbf{\Theta}_0) = \frac{1}{NMp_1\cdots p_D}\|\mathbf{\Theta} - \mathbf{\Theta}_0\|_2^2$, and that $V(\mathbf{\Theta}, \mathbf{\Theta}_0) = \frac{4\sigma^2}{NMp_1\cdots p_D}\|\mathbf{\Theta} - \mathbf{\Theta}_0\|_2^2$, where σ^2 is assumed to be the same variance of each element of the tensor, and $\|\cdot\|_2$ is the L_2-norm of a vectoring tensor.

We consider the parameter in a small and restricted space

$$\mathcal{S}_s(M_l, M_p) = \{\mathbf{\Theta} \in \mathcal{S} : M_l \leq \|\mathbf{\Theta} - \mathbf{\Theta}_0\|_2 \leq 2M_l, p(\mathbf{\Theta}) \leq M_p\},$$

and let $\mathcal{F}(M_l, M_p) = \{l_d(\mathbf{\Theta}|\cdot) : \mathbf{\Theta} \in \mathcal{S}_s\}$ be the range of $l_d(\mathbf{\Theta}|\cdot)$ that corresponds to $\mathcal{S}_s$.

To verify several conditions of Corollary 2 in [74], first, it is apparent to show that

$$\sup_{\mathcal{S}_s(M_l,M_p)} V(\mathbf{\Theta}, \mathbf{\Theta}_0) \leq c_1 M_l^2\{1 + (M_l^2 + M_p)^{\beta_1}\}$$

for a constant $c_1 \geq 0$ and a constant β_1 on the restricted space $\mathcal{S}_s(M_l, M_p)$.

Second, we verify that $\sup_{\mathcal{S}_s(M_l,M_p)} \|\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\|_{\sup} \leq c_2(M_l^2+M_p)^{\beta_2}$ for a constant $c_2 \geq 0$ and $\beta_2 \in [0,1)$ by applying Lemma 2 of [74]. Define $f_0 = \boldsymbol{\Theta}-\boldsymbol{\Theta}_0$. Recall that

$$\max_m \|(\boldsymbol{w}^m, \boldsymbol{B}^{m,(1)}, \ldots, \boldsymbol{B}^{m,(D)})\|_\infty \leq C_2,$$

and $\max_i \|(u_i, \boldsymbol{s}_i^{(1)}, \ldots, \boldsymbol{s}_i^{(D)})\|_\infty \leq C_2$. Denote $\gamma = \sum_{m=1}^{M}(\sum_{d=1}^{D} p_d+N)K_m + \sum_{i=1}^{N}(\sum_{d=1}^{D} p_d+1)$ as the total number of parameters. Since f_0 is a $(D+1)$-degree polynomial function of elements of $\boldsymbol{w}^m$'s, $\boldsymbol{B}^{m,(d)}$'s, u_i's and $\boldsymbol{s}_i^{(d)}$'s, we have $f_0 \in W_2^\infty[-C_2,C_2]^\gamma$ where W_2^∞ is a Sobolev space, and $\|f_0\|_2 = \rho(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0) \leq c_3$ for a constant $c_3 > 0$. In addition, we have $f_0^{(\alpha)} = 0$ for $\alpha = \infty$. Following Lemma 2 of Shen (1998), we have

$$\|f_0\|_\infty = \|\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\|_\infty \leq 2c_3.$$

Therefore, the required conditions are fulfilled by defining $c_2 = 2c_3$ and $\beta_2 = 0$.

Next, we define the Hellinger metric entropy with L_2 bracketing. Let

$$N(\varepsilon, q) = \{f_1^l, f_1^u, \ldots, f_q^l, f_q^u\}$$

be a set of functions from the L_2 space satisfy that $\max_{j=1,\ldots,m} \|f_j^u - f_j^l\| \leq \varepsilon$ and for any $l_d \in \mathcal{F}(M_l, M_p)$, there exists $j \in \{1,\ldots,q\}$, such that $f_j^l \leq l_d \leq f_j^u$ almost surely. Then we define the Hellinger distance as $H(\varepsilon,\mathcal{F}) = \log\{q : \min N(\varepsilon, q)\}$. Let

$$\psi(M_l, M_p) = \frac{1}{L}\int_L^U H^{1/2}(\delta,\mathcal{F})d\delta$$

where $L = c_4\lambda(M_l^2+M_p)$ and $U = c_5\varepsilon_N(M_l^2+M_p)^{(1+\max\{\beta_1,\beta_2\})/2}$, $\varepsilon_N = N^{-1/2}$ and c_4 and c_5 are two non-negative constants.

Based on Theorem 5.2 of [4], the Hellinger metric entropy can be controlled by

$$H(\varepsilon,\mathcal{F}) \leq c_6\epsilon_N^\omega$$

for a constant $c_6 \geq 0$ and a constant rate ω.

The result of Theorem 7 then follows by applying Corollary 2 of [74]. This completes the proof.

4.9 Tables and Figures

Table 4.1: Prediction performance of different methods on the testing set for the random-signal-area study based on 100 replications.

Model	VPL	TR	MeanDist	MPCA	HOCPD	IMTL
Overall Accuracy	0.641	0.531	0.753	0.710	0.774	**0.966**
Sensitivity	0.873	0.525	0.680	0.753	0.782	0.981
Specificity	0.249	0.541	0.877	0.633	0.763	0.939

Table 4.2: The prediction results of different methods on the testing set for multi-random-weak-signal study based on 100 replications.

	$\mu_C = 30$			$\mu_C = 20$		
	Overall Accuracy	Sensitivity	Specificity	Overall Accuracy	Sensitivity	Specificity
VPL	0.723	0.045	0.998	0.714	0.030	0.998
TR	0.767	0.585	0.852	0.667	0.505	0.740
MeanDist	0.732	0.072	1.000	0.717	0.032	1.000
MPCA	0.834	0.574	0.933	0.698	0.368	0.837
HOCPD	0.992	0.973	1.000	0.889	0.676	0.977
IMTL	**0.998**	**0.994**	**1.000**	**0.964**	**0.924**	**0.980**

Table 4.3: The prediction results on the testing data set of different models for the four-modality imaging study based on 100 replications.

Model	VPL	TR	MPCA	HOCPD	IMTL
Overall Accuracy	0.598	0.554	0.571	0.748	**0.842**
Sensitivity	0.105	0.544	0.314	0.638	0.805
Specificity	0.901	0.577	0.752	0.868	0.875

Table 4.4: The prediction results on the testing data set for different models, including overall prediction accuracy rate (OPAR), sensitivity and specificity.

Model	VPL	TR	MPCA	HOCPD	IMTL
Overall Accuracy	0.539	0.526	0.766	0.803	**0.854**
Sensitivity	0.604	0.538	0.823	0.813	0.832
Specificity	0.471	0.524	0.705	0.786	0.869

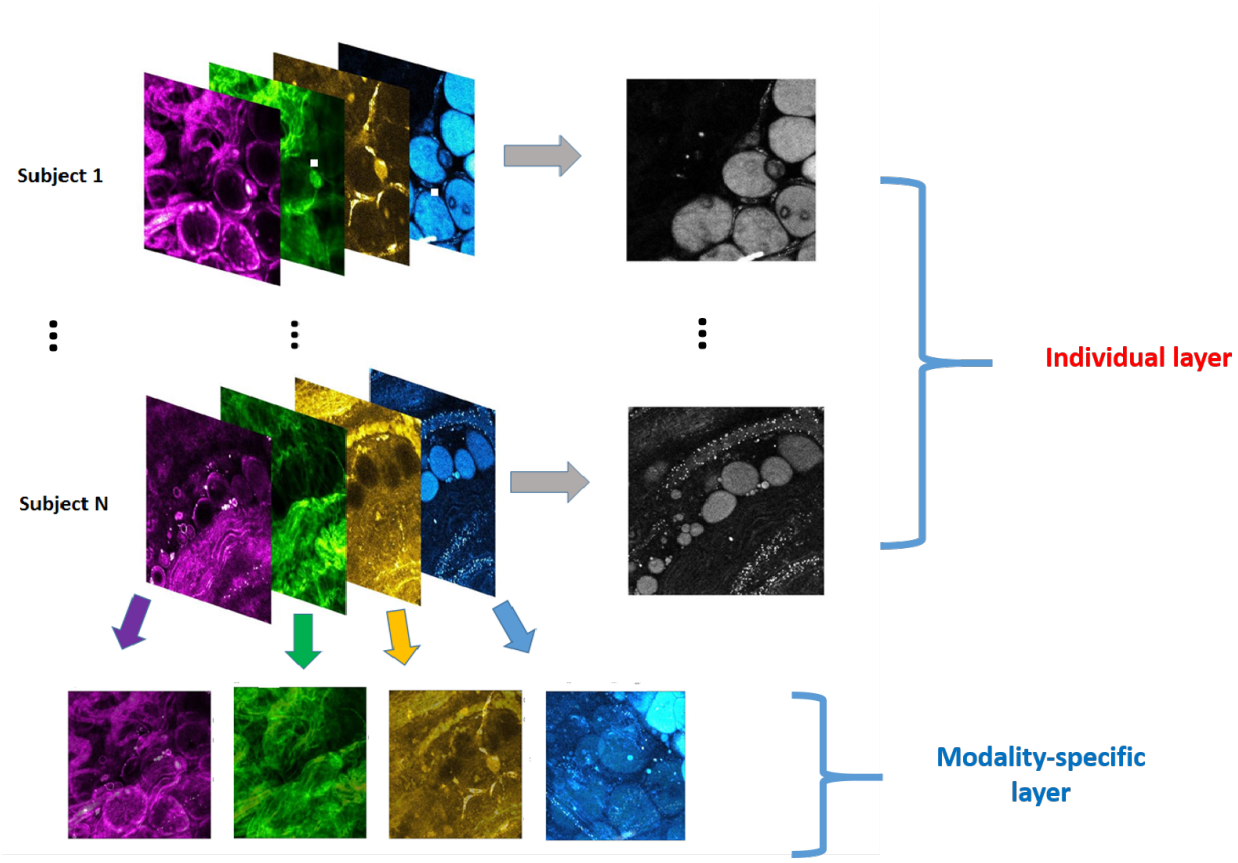

Figure 4.1: An illustration of the individualized layers and the modality-specific layers for four-modality optical images of the breast cancer tissues.

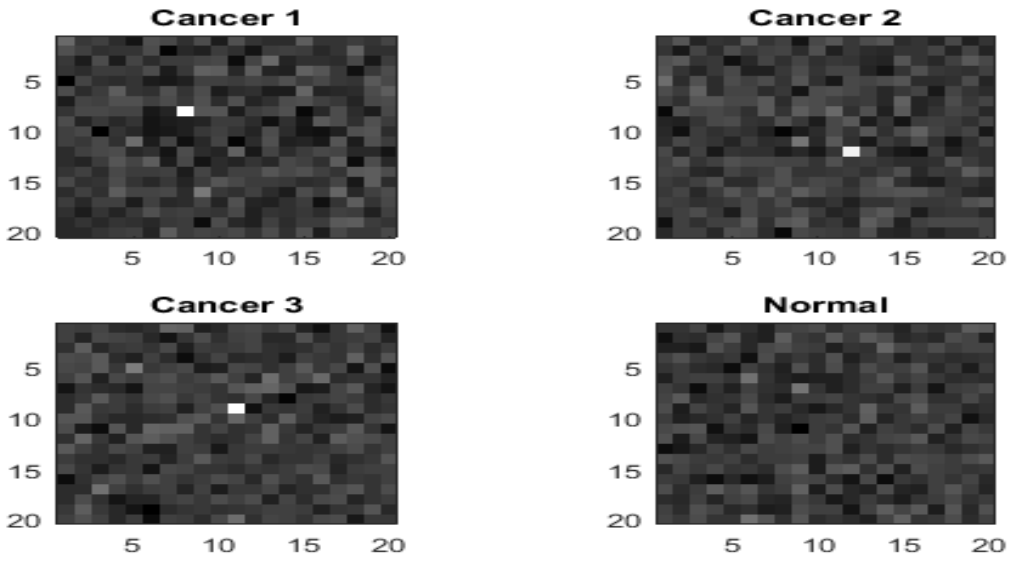

Figure 4.2: Simulated cancer images and normal image with random signal area in a subregion.

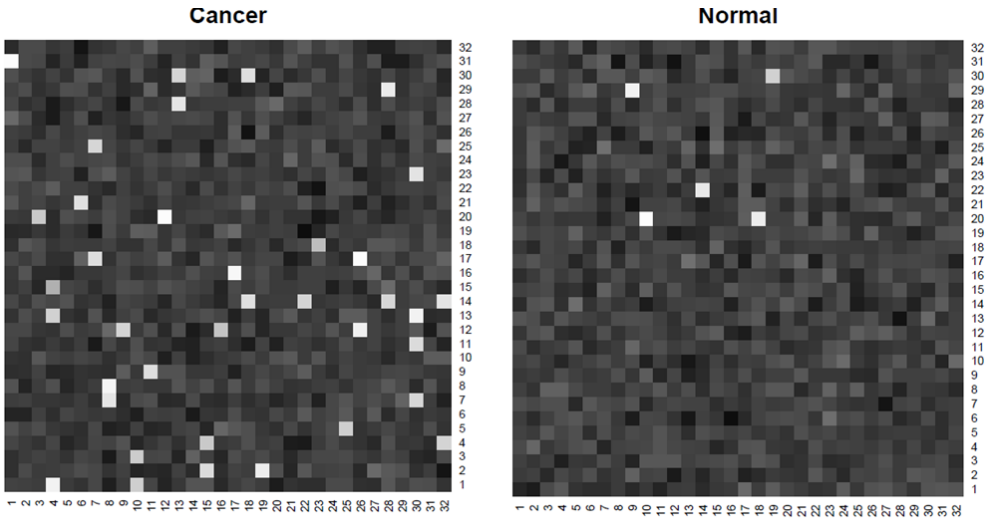

Figure 4.3: A simulated cancer image ($\mu_C = 30$) and a normal image. White spots are the signals.

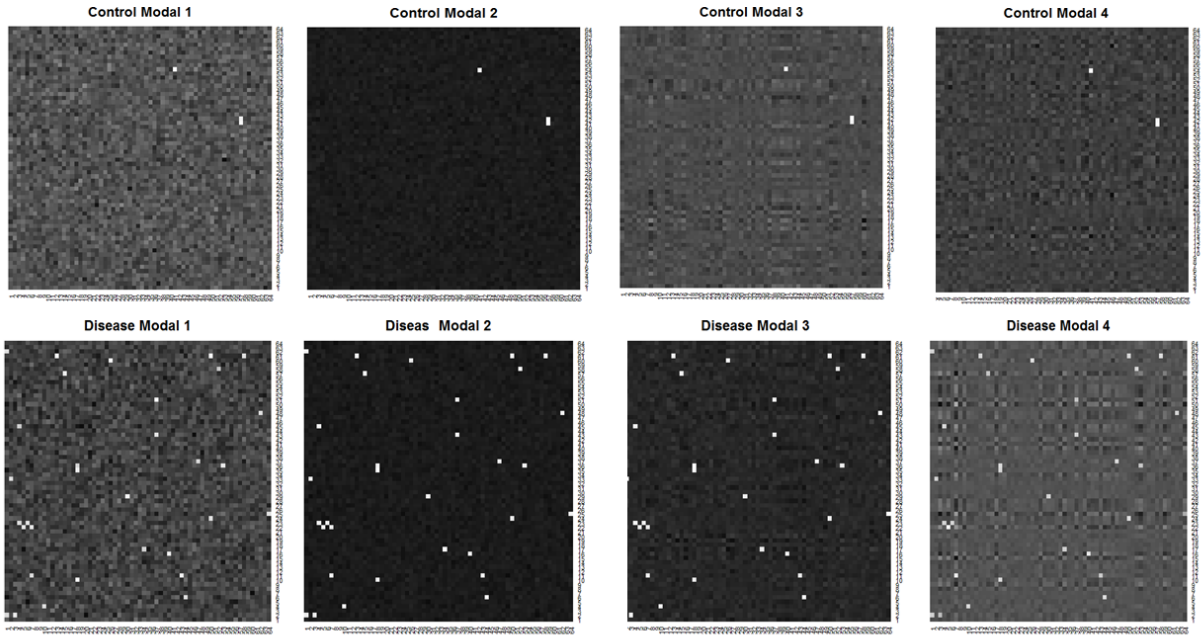

Figure 4.4: The simulated four-modality images for a cancer image and a normal image.

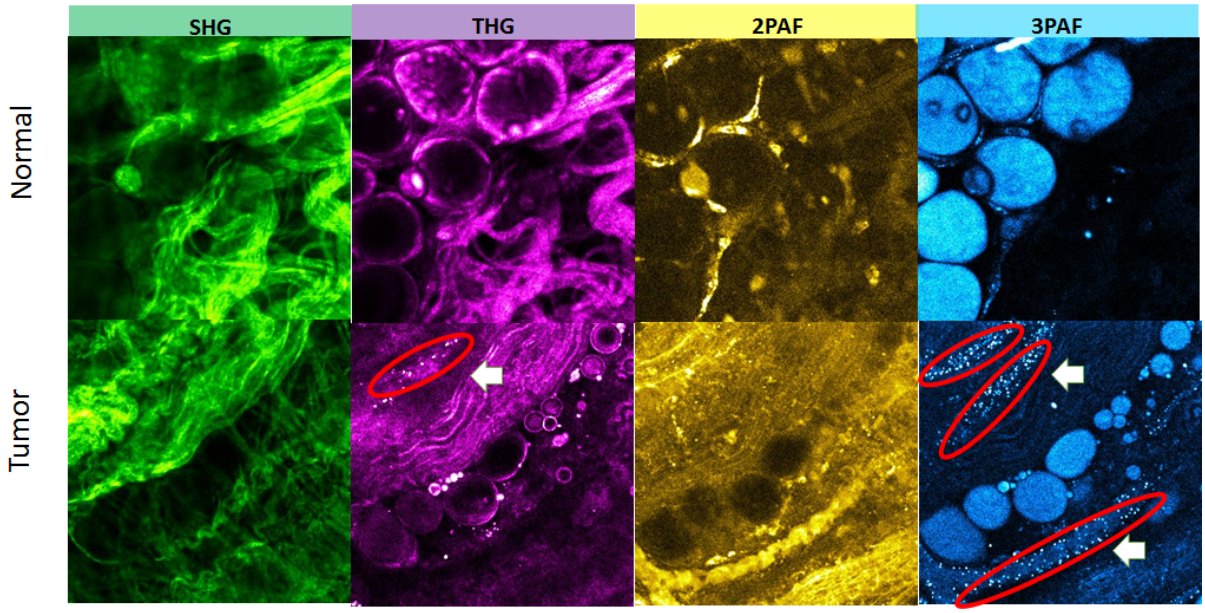

Figure 4.5: Four-modality microscope images for a normal rat's tissue and a cancerous rat's tissue.

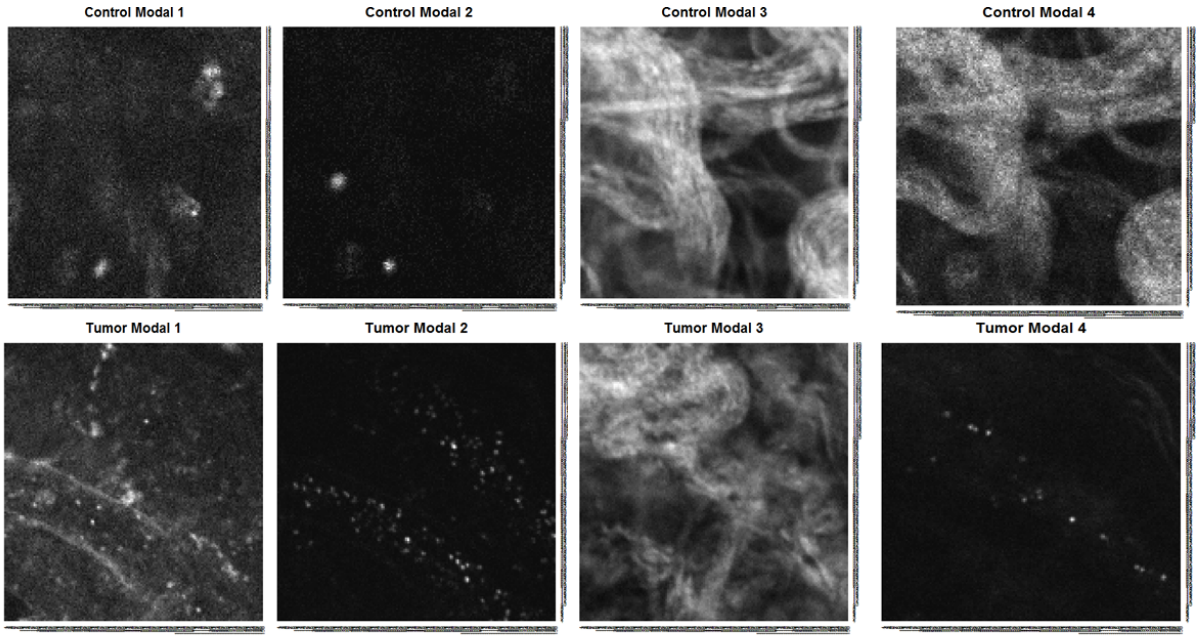

Figure 4.6: Four-modality images of a normal rat's tissue and a cancerous rat's tumor tissue at the lipid boundary area.

References

[1] Bahadur, R. R. (1961) A representation of the joint distribution of responses to n dichotomous items. *In Studies on Item Analysis and Prediction*, 158-68. California: Stanford University Press.

[2] Balan, R. M. and Schiopu-Kratina, I. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics 32*, 522-541.

[3] Beckmann, C. F. and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *NeuroImage 25(1)*, 294-311.

[4] Birman, M. Š., and M. Z. Solomjak. (1967) Piecewise-polynimial approximations of functions of the classes W_p^α. *Sbornik: Mathematics 2*, 295-317.

[5] Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics 64*, 115-123.

[6] Bowman, F. D. (2007). Spatiotemporal models for region of interest analyses of functional neuroimaging data. *Journal of American Statistics Association 102*, 442-453.

[7] Bowman, F. D., Caffo, B., Bassett, S. S. and Kilts, C. (2008). A Bayesian hierarchical framework for spatial modeling of fMRI data. *NeuroImage 39*, 146-156.

[8] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression with applications to biological feature selection. *Annals of Applied Statistics 5*, 232-253.

[9] Caffo, B., Crainiceanu, C., Verduzco, G., Joel, S., Mostofsky, S. H., Bassett, S. and Pekar, J. (2010). Two-stage decompositions for the analysis of functional connectivity for fMRI with application to Alzheimer's disease risk. *NeuroImage 51 (3)*, 1140-1149.

[10] Casey, B., Soliman, F., Bath, K. G. and Glatt, C. E. (2010). Imaging genetics and development: Challenges and promises. *Human Brain Mapping 31*, 838-851.

[11] Chu, S., Tai, S., Ho, C., Lin, C. and Sun, C. (2005). High-resolution simultaneous three-photon fluorescence and third-harmonic-generation microscopy *Microscopy Reasearch and Techinques 66*, 193-197.

[12] Dempster, A. P., Laird, N. M., and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (B) 39*, 1-47.

[13] Derado, G., Bowman, F. D. and Kilts, C. D. (2010). Modeling the spatial and temporal dependence in fMRI data. *Biometrics 66*, 949-957.

[14] Dhawan, A. P., D'Alessandro, B. and Fu, X. (2010). Optical imaging modalities for biomedical applications. *IEEE Reviews in Biomedical Engineering 3*, 69-92.

[15] Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of longitudinal data.* New York: Oxford University Press.

[16] Dolin, R., Amato, D. A., Fischl, M. A. et al. (1995). Zidovudine compared with Didanosine in patients with advanced HIV type 1 infection and little or no previous experience with Zidovudine. *Archives of Internal Medicine 155*, 961-74.

[17] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348-1360.

[18] Fass, L. (2008). Imaging and cancer: A review. *Molecular Oncology 2*, 115-152.

[19] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33(1)*, 1-22.

[20] Friston, K. J. (2009). Modalities, modes, and models in functional neuroimaging. *Science 326*, 399-403.

[21] Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. Unpublished manuscript.

[22] Goeman, J., Meijer, R., Chaturvedi, N. and Lueder, M. (2017). Penalized: L1 (Lasso and fused Lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model. R package version 0.9-50. https://cran.r-project.org/web/packages/penalized/index.html

[23] Guo, F. J., Levina, E., Michailidis, G. and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics 66*, 793-804.

[24] Han, F. and Liu, H. (2017). Statistical analysis of latent generalized correlation matrix estimation in transelliptical distribution. *Bernoulli 23(1)*, 23-57.

[25] Heyde, C. C. (1997). *Quasi-likelihood and its application.* New York: Springer.

[26] Hocking, T., Joulin, A., Bach, F. and Vert, J.-P. (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In L. Getoor and T. Scheffer (Eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, 745-752.

[27] Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., Johnson, S. C. and ADNI (2009). Spatially augmented lpboosting for ad classification with evaluations on the adni dataset. *NeuroImage 48*, 138-149.

[28] Horton, N. G., Wang, K., Kobat, D., Clark, C. G., Wise, F. W., Schaffer, C. B. and Xu, C. (2013). In vivo three-photon microscopy of subcortical structures within an intact mouse brain. *Nature Photonics 7*, 205-209.

[29] Huang, M., Li, R., Wang, H., and Yao, W. (2014). Estimating mixture of Gaussian processes by kernel smoothing. *Journal of Business and Economic Statistcs 32*, 259-270.

[30] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Comp. 3*, 79-87.

[31] Jiang, W. and Tanner, M. A. (1999b). On the identifiability of mixtures-of-experts. *Neural Networks. 12*, 1253-1258.

[32] Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comp. 214*, 181-214.

[33] Jung, S. and Marron, J. S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics 37*, 4104-4130.

[34] Kang, H., Ombao, H., Linkletter, C., Long, N., and Badre, D. (2012). Spatio-spectral mixed effects model for functional magnetic resonance imaging data. *Journal of American Statistical Association 107*, 568-577.

[35] Ke, T., Fan, J. and Wu, Y. (2010). Homogeneity in regression. *Journal of the American Statistical Association 110*, 175-194.

[36] Knight, K. and Fu, W. (2000). Asymptotic for Lasso-type estimators. *The Annals of Statistics 28*, 1356-1378.

[37] Kolda, T. G. (2006). Multilinear operators for higher-order decompositions. Technical report, Sandia National Laboratories.

[38] Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Rev. 51 (3)*, 455-500.

[39] Kraemer, N., Schaefer, J. and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene regulatory networks with Gaussian graphical models. *BMC Bioinformatics 10*, 384

[40] Kruskal, J. B. (1989). Rank, decomposition, and uniqueness for 3-way and n-way arrays. In R. Coppi and S. Bolasco, editors, *Multiway Data Analysis*, 7-18, Amsterdam.

[41] Lazar, N. A. (2008). *The statistical analysis of functional MRI Data.* New York: Springer.

[42] Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L. and Coan, J. A. (2015).Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *Annals of Applied Statistics 9*, 687-713.

[43] Li, X., Zhou, H. and Li, L. (2013). Tucker tensor regression and neuroimaging analysis. *arXiv:1304.5637*

[44] Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13-22.

[45] Lindsten, F., Ohlsson, H. and Ljung, L. (2011). Clustering using sum-of-norms regularization: with application to particle filter output computation. *2011 IEEE Statistical Signal Processing Workshop (SSP)*, 201-204.

[46] Lindquist, M. (2008). The statistical analysis of fmri data. *Statistical Science 23*, 439-464.

[47] Ma, S., and Huang, J. (2016) A Concave Pairwise Fusion Approach to Subgroup Analysis. *Journal of the American Statistical Association*, in press.

[48] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P. and Gonzalez, J. (2016). Cluster: Finding Groups in Data. R package version 2.0.5. https://cran.r-project.org/web/packages/cluster/index.html

[49] Martinez, E., Valdes, P., Miwakeichi, F., Goldman, R. I., and Cohen, M. S. (2004). Concurrent EEG/fMRI analysis by multiway partial least squares. *NeuroImage 22(3)*, 1023-1034.

[50] Martino, F. D., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage 43*, 44-58.

[51] McCullagh, P. (1983). Quasi-likelihood functions. *Annals of Statistics 11*, 59-67.

[52] McLachlan, G. J. and Peel, D. (2000). *Finite mixture models.* New York: John Wiley & Sons.

[53] Muthén, B. and Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research 24*, 882-891.

[54] Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*, 1-32.

[55] Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. New York: Academic Press.

[56] Pan, W. and Shen, X. (2006). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research 8*, 1145-1164.

[57] Pan, W., Shen, X. and Liu, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research 14*, 1865-1889.

[58] Penny, W. D., Trujillo-Barreto, N. J. and Friston, K. J. (2005). Bayesian fMRI time series analysis with spatial priors. *NeuroImage 24*, 350-362.

[59] Pew Research Center (2010). Four years later Republicans faring better with men, whites, independents and seniors (press release). Available at http://www.people-press.org/files/legacy-pdf/643.pdf

[60] Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*, 186-199.

[61] Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika 90*, 455-463.

[62] Qian, W. and Titterington, D. M. (1992) Stochastic relaxations and EM algorithms for Markov random fields. *Journal of Statistical Computation and Simulation 40*, 55-69.

[63] Qu, A., Lindsay, G. B., and Lu, L. (2010). Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random. *Journal of the American Statistical Association 105*, 194-204.

[64] Raftery, A. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Asscociation 101*, 168-178.

[65] Reiss, P., and Ogden, R. (2010). Functional generalized linear models with images as predictors. *Biometrics 66*, 61-69.

[66] Robins J. M., Rotnitzky A., and Zhao L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*, 106-121.

[67] Rosen, O., Jiang, W., and Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika 87*, 391-404.

[68] Rotnitzky A., Robins J. M., and Zhao L. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association 93*, 1321-1339.

[69] Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*, 581-592.

[70] Rubin, D. B. and Wu, Y. (1997). Modeling schizophrenic behavior using general mixture components. *Biometrics 53*, 243-261.

[71] Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole-brain classification of fMRI Data. *NeuroImage 51*, 752-764.

[72] Schwarz, C. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461-464.

[73] Seo, B. and Kim, D. (2012). Root Selection in Normal Mixture Models. *Computational Statistics & Data Analysis 56*, 2454-2470.

[74] Shen, X. (1998). On the method of penalization. *Statistica Sinica 8*, 337-357.

[75] Shen, X. and Huang, H.-C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association 105*, 727-739.

[76] Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association 107*, 223-232.

[77] Sidiropoulos, N. D. and Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics 14 (3)*, 229-239.

[78] Small, G. C., Wang, J., and Yang, Z. (2000). Eliminating multiple root problems. *Statistical Science 15*, 313-332.

[79] Sugar, C. A. and James, G. M. (2003). Finding the number of clusters in a data set: an information theoretic approach. *Journal of the American Statistical Association 98*, 750-763.

[80] Sun, Z., Rosen, O., and Sampson, A. R. (2007). Multivariate Bernoulli mixture models with application to postmortem tissue studies in schizophrenia. *Biometrics 63*, 901-909.

[81] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Ser. B 58*, 267-288.

[82] Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the gap statistic. *Journal of the Royal Statistical Society: Ser. B 63*, 411ÍC423.

[83] Tibshirani, S., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of Royal Statistical Society: Ser. B 67*, 91-108.

[84] Tu, H., Liu, Y., Turchinovich, D., Marjanovic, M., Lyngs?, J. K., L?gsgaard, J., Chaney, E. J., Zhao, Y., You, S., Wilson, W., Xu, B., Dantus, M. and Boppart, S. A. (2016). Stain-free histopathology by programmable supercontinuum pulses. *Nature Photonics 10*, 534-540.

[85] Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279-311.

[86] Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Ser.B 71*, 671-683.

[87] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*, 553-568.

[88] Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics 68(2)*, 353-360.

[89] Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika 61*, 439-447.

[90] Xie, B., Pan, W., and Shen, X. (2010). Penalized mixture of factor analyzers with application to variable selection in clustering high-dimensional data. *Bioinformatics 26*, 501-508.

[91] Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *The Annals of Statistics 31*, 310-347.

[92] Yao, W. (2012). Model based labeling for mixture models. *Statistics and Computing 22*, 337-347.

[93] Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association 104*, 758-767.

[94] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Ser. B 68*, 49-67.

[95] Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*, 894-942.

[96] Zhang, D. and Shen, D. (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage 59*, 895-907.

[97] Zhou, H. Matlab TensorReg toolbox version 0.0.2, available online, July 2013.

[98] Zhou, H., Li, L. and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of American Statistics Association 108*, 229-239.

[99] Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Ser. B 67*, 301-320.

[100] Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association 101*, 1418-1429.