Statistical Methods of Data Integration, Model Fusion, and Heterogeneity Detection in Big Biomedical Data Analysis

by

Lu Tang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2018

Doctoral Committee:

Professor Peter X.-K. Song, Chair
Associate Professor Jian Kang
Professor Karen E. Peterson
Associate Professor Brisa N. Sánchez

Lu Tang

lutang@umich.edu

ORCID iD: 0000-0001-6143-9314

To My Parents, Nalingna and Aiden

ACKNOWLEDGEMENTS

I feel so fortunate and grateful that I have been given the opportunity over the past five years in pursuit of statistical models in solving current and challenging research problems, and to turn them into my Ph.D. dissertation. I had one of the most enjoyable, enriching, and fulfilling Ph.D. experiences that anybody could hope for. It would not have been possible without the help of many people, including my advisor, dissertation committee, colleagues, friends, and family.

First, I am deeply appreciative of my advisor Dr. Peter X.-K. Song for his mentorship and support. He has taught me how to learn, to think, to question, and to innovate, with his profound knowledge and extensive experience. Besides everything he taught me in research, he has many memorable traits that I hope to cultivate as I begin my own career. He is my inspiration for the value of seeking simple yet clever solutions and focusing intensely on pursuing them. He embodies a high level of professional effectiveness which sets the standard I hope to reach. He is my role model for how to be a wonderful mentor to students.

I would also like to thank Drs. Karen E. Peterson, Brisa N. Sánchez, and Jian Kang for serving as members of my dissertation committee and providing me invaluable suggestions on my dissertation. Especially, I am grateful to Dr. Peterson for her domain expertise in nutritional and environmental health sciences, for leading me to the field of metabolomics in which I find immense motivation for methodology development that has shaped my dissertation, and for her continuous support in my work in the Children's Center as a data manager and analyst. I am also grateful to

Dr. Sánchez for her training and guidance on ELEMENT data management, and to both Drs. Kang and Sánchez for providing me many technical discussions and suggestions that have greatly improved the quality of my dissertation, for lending me great advices on my professional development, and for helping me improve the presentation of my work during my interview rehearsals.

My appreciation also goes to the fellow colleagues who directly helped me on the research projects that comprise this dissertation. Here is everyone listed in alphabetical order: Abraham Bagherjeiran, Mathieu Bray, Sougata Chaudhuri, Emily Hector, Jennifer LaBarre, Lan Luo, Wei Perng, Zhengling Qi, Yanyi Song, Fei Wang, Zhenzhen Zhang, and Ling Zhou. I am grateful to them for their generosity, especially, to Ling Zhou, who has helped me tremendously in developing the work presented in Chapter III. Additionally, I would like to thank my department, the faculty and staff, fellow students and friends, and the Song Lab for a supportive and encouraging environment for study, research and personal growth.

Finally, nothing would have meaning if I did not have the support of my family, especially my wife, Nalingna Yuan, and my parents, Xiaoping Tang and Huamin Shu, to whom I would like to dedicate this thesis, for their unconditional love. And special thanks to my son, Aiden, who has brought me endless joy and distraction, and without whom this thesis would have been completed sooner.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

Interesting and challenging methodological questions arise from the analysis of Big Biomedical Data, where viable solutions are sought with the help of modern computational tools. In this dissertation, I look at problems in biomedical studies related to data integration, data heterogeneity, and related statistical learning algorithms. The overarching strategy throughout the dissertation research is rooted in the treatment of individual datasets, but not individual subjects, as the elements of focus. Thus, I generalized some of the traditional subject-level methods to be tailored for the development of Big Data methodologies.

Following an introduction overview in the first chapter, Chapter II concerns the development of fusion learning of model heterogeneity in data integration via a regression coefficient clustering method. The statistical learning procedure is built for the generalized linear models, and enforces an adjacent fusion penalty on ordered parameters (*Wang et al.*, 2016). This is an adaptation of the fused lasso (*Tibshirani et al.*, 2005), and an extension to the homogeneity pursuit (*Ke et al.*, 2015) that only considers a single data set. Using this method, we can identify regression coefficient heterogeneity across sub-datasets and fuse homogeneous subsets to greatly simplify the regression model, so to improve statistical power. The proposed fusion learning algorithm (published as *Tang and Song* (2016)) allows the integration of a large number of sub-datasets, a clear advantage over the traditional methods with stratum-covariate interactions or random effects. This method is useful to cluster treatment effects, so some outlying studies may be detected. We demonstrate our method with datasets from the Panel Study of Income Dynamics and from the Early Life Exposures in Mexico to Environmental Toxicants study. This method has also been extended to the Cox proportional hazards model to handle time-to-event response.

Chapter III, under the assumption of homogeneous generalized linear model, focuses on the development of a divide-and-combine method for extremely large data that may be stored on distributed file systems. Using the means of confidence distribution (*Fisher*, 1956; *Efron*, 1993), I develop a procedure to combine results from different sub-datasets, where lasso is used to reduce model size in order to achieve numerical stability. The algorithm fits into the MapReduce paradigm and may be perfectly parallelized. To deal with estimation bias incurred by lasso regularization, a de-bias step is invoked so the proposed method can enjoy a valid inference. The method is conceptually simple, and computationally scalable and fast, with the numerical evidence illustrated in the comparison with the benchmark maximum likelihood estimator based on full data, and some other competing divide-and-combine-type methods. We apply the method to a large public dataset from the National Highway Traffic Safety Administration on identifying the risk factors of accident injury.

In Chapter IV, I generalize the fusion learning algorithm given in Chapter II and develop a coefficient clustering method for correlated data in the context of the generalized estimating equations. The motivation of this generalization is to assess model heterogeneity for the pattern mixture modeling approach (*Little*, 1993) where models are stratified by missing data patterns. This is one of primary strategies in the literature to deal with the informative missing data mechanism. My method aims to simplify the pattern mixture model by fusing some homogeneous parameters under the generalized estimating equations (GEE, *Liang and Zeger* (1986)) framework.

CHAPTER I

Introduction

1.1 Motivation

Massive amounts of data are being generated and processed every second at a speed we have never seen before. This is a collective effort of many new technologies that decrease the cost of data generation and storage, increase the speed of data transfer and sharing, and facilitate the use of scalable tools for data management and analysis. As *Fan et al.* (2014) precisely describes, we are in the era of Big Data where information explodes. Also as *Mayer-Schönberger and Cukier* (2013)'s analogy of Big Data interestingly says, it is a revolution that will transform the way we live, work and think, and furthermore, the way we conduct science, engineering and business. Although Big Data seems to be mostly related to the IT industry (e.g., Google, Facebook, etc.), it is also very common in biomedical areas. For example, as the cost of whole genome sequencing drops dramatically over years (*Stein*, 2010), we are seeing data not just of large p small n, but more and more with both large p and large n. Additionally, data being collected from study subjects are no longer constrained to vector forms, but also in the format of higher order arrays (i.e., tensors), such as biomedical imaging data of the human body.

Big Data brings both opportunities and challenges to many subject areas, including statistics, engineering and computer science, and often times requires interdisciplinary knowledge to adequately understand the problems and devise appropriate solutions. For statisticians, valid statistical analysis for Big Data is one of the most important concern. In this dissertation, I look at challenging problems in biomedical studies related to data integration, data heterogeneity, missing data, and their respective algorithms. I provide fast and accurate solutions to some of the most important problems related to big biomedical data. Yet, these solutions are general and also applicable to big data in other areas. Additionally, I emphasize on the deliverability of the proposed methods by providing ready-to-use software packages, and also consider the compatibility with modern big data infrastructures, such as distributed data storages (e.g., *Palankar et al.* (2008); *Shvachko et al.* (2010)) and cloud computing architectures (e.g., *Dean and Ghemawat* (2008); *Zaharia et al.* (2010)).

1.2 Statistical Challenges in Big Data

Under the Big Data setup, new issues arise when applying conventional statistical methods for doing data analysis. We describ in this section a list of issues that we should take into consideration when dealing with big data. Although this is by far the complete list of issues related to Big Data, I selected the most critical ones pertaining to the subject of statistics and biostatistics. The proposed methods in the following chapters are going to address the challenges listed below.

1.2.1 Heterogeneity

The large amounts of samples of Big Data are typically achieved by aggregating data from multiple studies, and/or at different time points, and/or using different technologies. For example, multiple clinical trial studies conducted at different hospitals can be combined as larger studies (*Lohmueller et al.*, 2003; *Sullivan et al.*, 2000). This poses an issue of data heterogeneity, and most of the time, due to factors that are not observed. Data heterogeneity will likely result in experimental variations and statistical biases, and requires us to develop more adaptive and robust procedures (*Fan et al.*, 2014). Previoius works can be found in the literature tackling this issue, such as *Shen and Huang* (2010), *Ke et al.* (2015), and *Wang et al.* (2016). Chapter II discusses the ideas and limitations of these work and proposes a new method that is dedicated to address the challenge of data heterogeneity.

1.2.2 Dimensionality

Due to the advance in data collection technologies, it is often the case that Big Data contains information collected that may or may not be related to study objectives, resulting in high-dimensionality in the covariates. Variable selection plays an important role in reducing the dimensionality of data and serves as a robustifier during numerical calculation. The most popular method is through regularization, with penalties including the lasso (*Tibshirani*, 1996), SCAD (*Fan and Li*, 2001), MCP (*Zhang*, 2010), and their extensions in various forms. Owning to their good numerical properties, regularization techniques play pivotal roles in the development of this dissertation.

1.2.3 Scalability

Big Data has motivated companies and research centers to develop compatible storage and computational infrastructures to efficiently handle its continuingly growing size. Distributed storage systems, giant clusters of computers collectively store huge data files, have become the state-of-the-art and give a solution to scalability. Examples include Amazon's Simple Cloud Storage Service(*Palankar et al.*, 2008) and Google's Cloud Bigtable (*Chang et al.*, 2008). Due to the nature that data are stored in different processors, algorithms that only make a linear pass of data are more preferable than those requiring iterative access of data. Based on such motivation, Chapter III proposes a divide-and-combine algorithm to regression modeling.

1.2.4 Missing Data

Different data availability and missing patterns can be regarded as special types of data heterogeneity. It is important to understand and acknowledge the differences in association due to missingness when data are not missing at random. The benefit of large sample sizes of Big Data allows us to more accurately study the nuance of missingness from a perspective different than traditional data analysis. In Chapter IV, I will extend the discussion in Chapter II to the problem of missing data under the framework of generalized estimating equations (*Liang and Zeger*, 1986) and propose a more flexible solution to longitudinal data than the classic pattern mixture models (*Little*, 1993).

1.3 Summary of Objectives

With a focus on the challenges presented above, I present in this dissertation methodologies that are aimed to achieve the following analytical objectives:

i To establish a data driven procedure to detect parameter homogeneity and heterogeneity under the scenario of data integration;

ii To devise a scalable divide-and-combine algorithm for the statistical inference of generalized linear models, under the modern distributed storage architecture;

iii To establish a heterogeneity detection procedure for longitudinal data with missingness across different types of missing patterns.

The three objectives are addressed in the proposed methods in Chapter II, Chapter III, and Chapter IV, respectively. More details on the background, literature, inspiration and methodology development, can be found in the introduction sections of each of the chapters.

CHAPTER II

Fused Lasso Approach in Regression Coefficients Clustering – Learning Parameter Heterogeneity in Data Integration

2.1 Introduction

Combining data sets collected from multiple studies is undertaken routinely in practice to achieve a larger sample size and higher statistical power. Such information integration is commonly seen in biomedical research, for example, the study of genetics or rare diseases where data repositories are available. The motivation of this chapter arises from the consideration of data heterogeneity during data integration. Although data integration has different meanings, in here, we consider the concatenation of data sets of similar studies over different subjects, where the number of integrated data sets can be very large.

Inter-study heterogeneity can result from the differences in study environment, population, design and protocols (*Leek and Storey*, 2007; *Sutton and Higgins*, 2008; *Liu et al.*, 2015). Data heterogeneity is likely attributed to population parameter heterogeneity, where the association of interest can differ across different study populations from which data sets are collected. Examples include multi-center clinical trials when participant data from different sites are combined (*Shekelle et al.*, 2003) and genetics studies when genomic data from multiple similar studies are combined (*Lohmueller et al.*, 2003; *Sullivan et al.*, 2000). Discrepancies in treatment effect or trait-gene association may arise due to the differences in facilities, practices and patient characteristics across studies, albeit the adjustment of confounding (*Leek and Storey*, 2007). The parameter heterogeneity introduced in data integration compromises the power of the larger sample size and may even lead to biased results and misleading scientific conclusions. Thus, counterintuitively, the model obtained from the combined studies may not serve as a proper prediction model for each individual study in the case of heterogeneous study populations.

Traditional treatments of parameter heterogeneity are not optimal. Meta-analysis methods such as combining summary statistics (*Glass*, 1976), estimating functions (*Hansen*, 1982; *Qin and Lawless*, 1994) or p-values functions (*Xie et al.*, 2012) are built upon the assumption of complete parameter homogeneity, as shown in the left panel of Figure 2.1. This assumption is hardly valid in practice. When individual participant data from multiple data sets are available, a retreat to the classical meta-analysis methods is necessary, because in this case assessing the assumption of inter-study homogeneity becomes possible. The two most common approaches to handling parameter heterogeneity include (i) specifying study-specific effects by including interaction terms between study indicator and covariates (e.g., *Lin et al.* (1998)), and (ii) utilizing random covariate effects by allowing variations across studies as random variables (e.g., *DerSimonian and Kacker* (2007)). Both approaches essentially assume fully heterogeneous covariate effects, namely, each study having its own set of regression coefficients, as shown in the right panel of Figure 2.1.

When study-specific effects are of interest, the interaction-based formulation may lead to over-parameterization, which impairs statistical power. The most straightforward way to reduce the number of parameters is to identify clusters of homogeneous parameters through exhaustive tests for the differences between every pair of study-

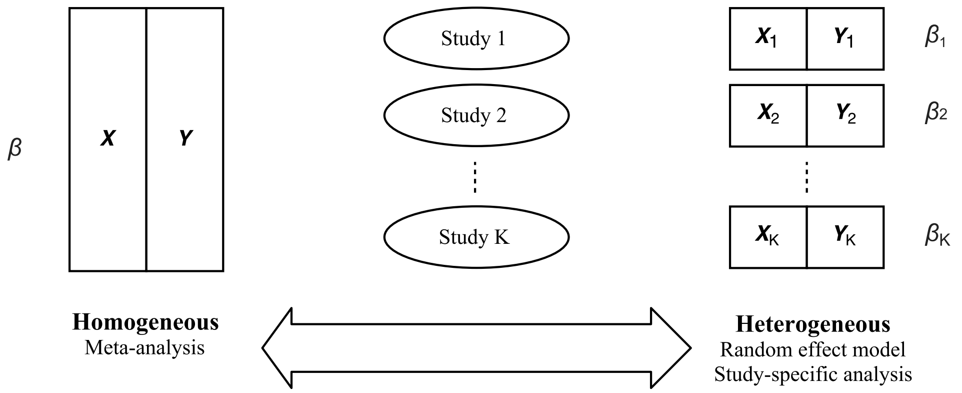


Figure 2.1: Homogeneous assumption (left) versus heterogeneous assumption (right).

specific coefficients. However, when the number of data sets is large, the use of hypothesis testing to determine parameter clusters becomes untrackable in addition to the multiple-testing problem. One may draw different or even conflicting conclusions due to different orders of hypotheses performed.

In reality, covariate effects from multiple studies are likely to form groups, a scenario falling in between the complete heterogeneity and the complete homogeneity. This leads to the following two essential yet related analytic tasks: (i) to assess the inter-study heterogeneity, so to determine an appropriate form of parsimonious parameterization in model specification; and (ii) to identify and merge groups of homogeneous parameters for better statistical power for parameter estimation and inference based on a more parsimonious model. Along the idea of lasso shrinkage estimator (*Tibshirani*, 1996), fused lasso methods (*Tibshirani et al.*, 2005; *Friedman et al.*, 2007; *Yang et al.*, 2012) have been introduced to achieve covariate grouping, where covariate adjacencies are naturally defined by a metric of time, location or network structure. In our problem of data integration, there does not exist a natural metric to define the ordering of regression coefficients from different studies. *Shen and Huang* (2010) proposed the grouping pursuit via penalization of all pairwise coefficient differences in a single study, where covariate orderings are not considered.

To reduce the computational burden in the all-pairs based regularization, *Wang et al.* (2016) and *Ke et al.* (2015) used the initial coefficient estimates to establish certain ordering and then to define parameter adjacencies. However, most of these studies have been entirely focusing on a single cohort of subjects from a single study. For example, *Shin et al.* (2016) proposed to fuse regression coefficients of different loss functions obtained from a single study, such as coefficients from different quantile regression models. Limited publication of fusion learning and grouping pursuit has been available in the literature, except *Wang et al.* (2016), to assess the differences and similarities among regression coefficients across multiple studies in the scenario of data integration.

In this chapter, we propose an agglomerative clustering method for regression coefficients in the context of data integration, named as the *Fused Lasso Approach in Regression Coefficients Clustering (FLARCC)*. FLARCC is proposed to identify heterogeneity patterns of regression coefficients across studies (or data sets) and to provide estimates of all regression coefficients simultaneously. It is interesting to draw a connection between our method and *Pan et al.* (2013) where they consider a classic clustering problem of individual responses by pairwise coefficient fusion via penalized regression. Their method aims at clustering subjects, while our method focuses on clustering regression coefficients across multiple data sets, and these two methods coincide only in a special case where each study is composed of only one subject. FLARCC achieves clustering of study-specific effects by penalizing the ℓ_1-norm differences of adjacent coefficients, with adjacency defined by the estimated ranks. Our method extends the bCARDS method in *Ke et al.* (2015) from one study to multiple studies as well as from the linear model to the generalized linear models, and focuses on simultaneous clustering of regression coefficients of individual covariates from multiple studies in data integration. An R package `metafuse` (current version 2.0-1) is created as part of our methodology development to perform the proposed integrated data analysis which can be downloaded from the Comprehensive R Archive Network (webpage link https://cran.r-project.org/web/packages/metafuse).

In the proposed method, tuning parameter is used to determine the clustering pattern of coefficients across data sets. Specifically, let λ be the tuning parameter of regularization. If $\lambda = 0$ (i.e., no penalty), FLARCC becomes a method under the setting of complete heterogeneity, so that study-specific regression coefficients for each covariate are assumed different across data sets. If λ is large enough that all differences of regression coefficients are shrunk to zero, FLARCC reduces to a homogeneous model in that a common regression coefficient for each covariate is assumed for all studies. In light of the hierarchical clustering scheme, these two extreme cases above correspond to the start and end of an agglomerative clustering, respectively; however, the reality is believed to reside in between. Analogous to dendrograms in the hierarchical clustering, we propose a new tree-type graphic display, named as *fusogram*, which presents tree-based coefficient clusters according to solution paths obtained from FLARCC. The selection of optimal λ pertains to pruning of clustering trees, which can be based on certain model selection criterion. We use the extended Bayesian information criterion (EBIC) proposed by *Chen and Chen* (2008) as our model selection criterion and show that EBIC exhibits better performance than BIC when the number of studies (or data sets) is large. In addition, we propose a scaling strategy to "harmonize" solution paths by covariate-wise adaptive weights to allow flexible tuning, which further improves the clustering performance.

The rest of this chapter is organized as follows. Section 2.2 describes FLARCC in detail under the generalized linear models (GLM) framework. Section 2.3 presents the theoretical properties of the proposed method. Section 2.4 discusses the interpretation and selection of the tuning parameter. In Section 2.5, we use simulation studies to evaluate the performance of our method. Real data analysis examples are given in Section 2.6 with interpretation of coefficient estimates and illustration of *fusograms.*

Discussion and concluding remarks are in Section 2.7.

2.2 Method of Parameter Fusion

In this section, we present the method and algorithm of FLARCC.

2.2.1 Notations and Method

We start by introducing necessary notations. Throughout this chapter, i, j and k are used to index subject, covariate and study, respectively. For instance, $X_{j,k}^{(i)}$ denotes the measurement of the jth covariate from the ith individual from study k, and $Y_k^{(i)}$ is the measurement of a response variable from the ith individual from study k. The total number of studies is denoted as K and the number of covariates involved is p. The sample size for study k is n_k, $k = 1, \ldots, K$, and the combined sample size is $N = \sum_{k=1}^{K} n_k$. The collection of all coefficients (covariates-wise) is denoted as $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,\bullet}^T, \boldsymbol{\beta}_{2,\bullet}^T, \ldots, \boldsymbol{\beta}_{p,\bullet}^T)^T$ with $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^T$ for $j = 1, \ldots, p$. An indicator vector $\boldsymbol{c} = (c_1, \ldots, c_p)^T$ is used to flag heterogeneous covariates, namely if the jth covariate is treated as heterogeneous (i.e., all different coefficients across K studies) then $c_j = 1$ and as homogeneous (a common coefficient across K studies) otherwise. Thus $c_j = 0$ for some $j \in \{1, \ldots, p\}$ implies that coefficient vector $\boldsymbol{\beta}_{j,\bullet}$ reduces to a common scalar parameter β_j for all K studies.

For illustration, let us consider a simple scenario of $\boldsymbol{c} = (1, 1, 0, \ldots, 0)^T$, in which the first two covariates are set as heterogeneous and the remaining $p-2$ covariates are set as homogeneous. The resulting coefficient vector is $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,\bullet}^T, \boldsymbol{\beta}_{2,\bullet}^T, \beta_3, \ldots, \beta_p)^T$. Then the corresponding design matrix $\boldsymbol{X}$ can be written as

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_{1,1} & & & \boldsymbol{X}_{2,1} & & & \boldsymbol{X}_{3,1} & \cdots & \boldsymbol{X}_{p,1} \\ & \ddots & & & \ddots & & \vdots & \ddots & \vdots \\ & & \boldsymbol{X}_{1,K} & & & \boldsymbol{X}_{2,K} & \boldsymbol{X}_{3,K} & \cdots & \boldsymbol{X}_{p,K} \end{pmatrix}_{N \times (2K+p-2)}$$

where $\boldsymbol{X}_{j,k} = (X_{j,k}^{(1)}, \ldots, X_{j,k}^{(n)})^T$, $j = 1, \ldots, p$, $k = 1, \ldots, K$. The specification of $\boldsymbol{c}$ is can be dependent on the study interest. For example, in a multi-center clinical trial where we believe that the differences between the services provided across centers are non-negligible, but the study participants are similar, we can specify the clinic-related variables (e.g., treatment and cost) to be heterogeneous and the patient-related variables (e.g., age and gender) to be homogeneous. In addition, the specification of $\boldsymbol{c}$ can be dependent on preliminary marginal analysis of the homogeneousness of each variable, such as tests for random effects. When the homogeneousness of a covariate is unclear, we suggest specifying it as heterogeneous rather than homogeneous.

Under the assumption that both within-study and between-study samples are independent, for any $\boldsymbol{c} = (c_1, \ldots, c_p)^T$ with $c_j \in \{0, 1\}$, $j = 1, \ldots, p$, the initial estimate of $\boldsymbol{\beta}$, which gives the starting level of clustering (i.e., $\lambda = 0$), can be consistently estimated by the maximum likelihood estimator

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{(K \times p)}}{\operatorname{argmax}} \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \log L_k(\boldsymbol{\beta}), \tag{2.1}$$

where $L_k(\boldsymbol{\beta}) = \prod_{i=1}^{n_k} L_k^{(i)}(\boldsymbol{\beta}), k = 1, \ldots, K$ are the study-specific likelihoods from the given GLMs. For the purpose of parameter grouping and fusion, we propose the regularized maximum likelihood estimation for $\boldsymbol{\beta}$ by minimizing the following objective function:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{(K \times p)}} \left(-\frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_k} \log L_k(\boldsymbol{\beta}) + P(\boldsymbol{\beta}) \right), \tag{2.2}$$

where $P(\boldsymbol{\beta})$ is a penalty function of certain form. Here we adopt weighting $\frac{1}{n_k}$ to balance the contribution from each study so to avoid the dominance of large studies. Other types of weighting schemes may be considered to serve for different purposes, such as the inverse of estimated variances of initial estimates, which helps to achieve better estimation precision.

To achieve parameter fusion, *Shen and Huang* (2010) proposed the grouping pursuit algorithm, which specifies the sum of ℓ_1-norm differences of all study-specific coefficient pairs among individual heterogeneous coefficient vectors $\boldsymbol{\beta}_{j,\bullet}$, where $c_j = 1$, as the penalty:

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} c_j \sum_{k=1}^{K-1} \sum_{k'>k}^{K} |\beta_{j,k} - \beta_{j,k'}|,$$

with $\lambda \geq 0$. In this penalty, there are $\binom{K}{2}$ terms of pairwise differences for each heterogeneous covariate and the total number of terms increases by an order of $O(K^2)$, given p fixed. This penalty contains many redundant constraints and imposes great computational challenges as pointed out in *Shen and Huang* (2010) and *Ke et al.* (2015).

Following arguments in *Wang et al.* (2016) and *Ke et al.* (2015), we develop the method of FLARCC by a simplified penalty function that uses the information on the ordering of coefficients. For the jth covariate, let $\boldsymbol{U}_j = (U_{j,1}, \ldots, U_{j,K})^T$ be the ranking with no ties of $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^T$, from the smallest to the largest. Specifically, $U_{j,k} = \sum_{k'=1}^{K} \mathbf{1}\{\beta_{j,k'} \leq \beta_{j,k}\}$ if there are no ties in $\boldsymbol{\beta}_{j,\bullet}$; otherwise, the ties in $\boldsymbol{U}_j$ are resolved by the first-occurrence-wins rule according to k to ensure rank uniqueness. Then, the fusion penalty in FLARCC with parameter orderings $\boldsymbol{U}_j$, $j = 1, \ldots, p$, takes the form:

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} c_j \nu_j \sum_{k=1}^{K-1} \sum_{k'>k}^{K} \mu_{j,k,k'} \mathbf{1}\{|U_{j,k} - U_{j,k'}| = 1\} |\beta_{j,k} - \beta_{j,k'}|, \tag{2.3}$$

where the constraints occur effectively only on adjacent ordered pairs. Clearly, the penalty in (2.3) only involves $K - 1$ terms for each case of $c_j = 1$, which is of an order $O(K)$, given p fixed. The ν_j's and $\mu_{j,k,k'}$'s in (2.3) are weights. Following *Zou* (2006), we choose adaptive weights $\hat{\mu}_{j,k,k'} = 1/|\hat{\beta}_{j,k} - \hat{\beta}_{j,k'}|^r$, $r > 0$, so that parameters with smaller difference will be penalized more than those with larger differences.

Similarly, for a group of parameters $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^T$, ν_j is an adaptive weight to characterize the degree of heterogeneousness of $\boldsymbol{\beta}_{j,\bullet}$. Specifically, in this chapter we let $\hat{\nu}_j = 1/|\hat{\beta}_{j,(K)} - \hat{\beta}_{j,(1)}|^s$, the inverse of the range of the estimates, with $s \geq 0$; when a covariate is homogeneous, the differences of study-specific coefficients will be penalized more than those that are heterogeneous. In this way, we can "harmonize" solution paths so to greatly improve the performance by a single tuning parameter. We compare $s = 0$ and $s = 1$ in the simulation experiments and show in Section 2.5 that the introduction of such group-wise weights ν_j, $j = 1, \ldots, p$, gives rise to improvement on the performance of identifying homogeneous covariates when K and p are large.

A sparse version of FLARCC can also be achieved by including the traditional lasso penalty in (2.3) for covariate selection. In order to minimize the interference between fusion and sparsity penalties, we only encourage sparsity for the coefficient closest to zero in each $\boldsymbol{\beta}_{j,\bullet} = (\beta_{j,1}, \ldots, \beta_{j,K})^T$, for $j = 1, \ldots, p$. Similar to the definition of $\boldsymbol{U}_j$, let $\boldsymbol{V}_j = (V_{j,1}, \ldots, V_{j,K})^T$ be the ranking with no ties, from the smallest to the largest, of the absolute values of $\boldsymbol{\beta}_{j,\bullet}$, i.e., $(|\beta_{j,1}|, \ldots, |\beta_{j,K}|)^T$. First we calculate $\boldsymbol{V}_j$ by $V_{j,k} = \sum_{k'=1}^{K} \mathbf{1}\{|\beta_{j,k'}| \leq |\beta_{j,k}|\}$, then we resolve the ties in $\boldsymbol{V}_j$ by the first-occurrence-wins rule according to k. Thus we can extend (2.3) to achieve variable selection by the following penalty function:

$$\begin{aligned} P_{\lambda,\alpha}(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} c_j \nu_j \sum_{k=1}^{K-1} \sum_{k'>k}^{K} \mu_{j,k,k'} \mathbf{1}\{|U_{j,k} - U_{j,k'}| = 1\} |\beta_{j,k} - \beta_{j,k'}| \\ + \alpha\lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \mu_{j,k} \mathbf{1}\{V_{j,k} = 1\} |\beta_{j,k}|, \end{aligned} \tag{2.4}$$

where $\alpha \geq 0$ is another tuning parameter that controls the relative ratio between fusion and sparsity penalties, and $\hat{\mu}_{j,k} = 1/|\hat{\beta}_{j,k}|^r$. The sparsity penalty, although only enforced on the smallest coefficient in absolute value of $\boldsymbol{\beta}_{j,\bullet}$, is capable of shrinking a group of coefficients to zero when combined with the fusion penalty.

In practice, the weights (ν_j, $\mu_{j,k,k'}$ and $\mu_{j,k}$) and the parameter orderings ($\boldsymbol{U}_j$ and $\boldsymbol{V}_j$) are unknown, for $j = 1, \ldots, p$. We replace them with their estimates based on root-n consistent estimates $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{1,\bullet}^T, \ldots, \hat{\boldsymbol{\beta}}_{p,\bullet}^T)^T$, such as those from (2.1). In the simulation experiments and the real data application of this chapter, we set $r = 1$ in $\hat{\mu}_{j,k,k'}$ and $\hat{\mu}_{j,k}$.

2.2.2 Algorithm

Optimization problem (2.2) with $P(\boldsymbol{\beta}) = P_{\lambda,\alpha}(\boldsymbol{\beta})$ given in (2.4) can be carried out by a lasso regression through suitable reparameterization. Let the ordered coefficients of $\boldsymbol{\beta}_{j,\bullet}$ in an ascending order based on ranking $\boldsymbol{U}_j$ be $(\beta_{j,(1)}, \ldots, \beta_{j,(K)})^T$, $j = 1, \ldots, p$. For the jth covariate, consider a set of transformed parameters $\boldsymbol{\theta}_{j,\bullet} = (\theta_{j,1}, \ldots, \theta_{j,K})^T$ defined by

$$\begin{aligned} \theta_{j,1} &= \beta_{j,k}, & &\text{for } k \text{ s.t. } V_{j,k} = 1; \\ \theta_{j,k} &= \beta_{j,(k)} - \beta_{j,(k-1)}, & &\text{for } k = 2, \ldots, K. \end{aligned}$$

Then the $P_{\lambda,\alpha}(\boldsymbol{\beta})$ in (2.4) can be rewritten as

$$P_{\lambda,\alpha}(\boldsymbol{\theta}) = \lambda \sum_{j=1}^{p} \sum_{k=1}^{K} \omega_{j,k} |\theta_{j,k}|, \tag{2.5}$$

where

$$\hat{\omega}_{j,k} = \begin{cases} \alpha \frac{1}{|\hat{\theta}_{j,1}|^r}, & \text{if } k = 1 \\ c_j \frac{1}{|\sum_{k'=2}^{K} \hat{\theta}_{j,k'}|^s} \frac{1}{|\hat{\theta}_{j,k}|^r}, & \text{if } k = 2, \ldots, K, \end{cases} \tag{2.6}$$

for $j = 1, \ldots, p$. Since no ties are allowed in the parameter ordering of FLARCC, one-to-one transformation exists between $\boldsymbol{\beta} = (\boldsymbol{\beta}_{1,\bullet}^T, \ldots, \boldsymbol{\beta}_{p,\bullet}^T)^T$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_{1,\bullet}^T, \ldots, \boldsymbol{\theta}_{p,\bullet}^T)^T$ by suitable sorting matrix $\boldsymbol{S}$ and reparameterization matrix $\boldsymbol{R}$; that is, $\boldsymbol{\theta} = \boldsymbol{RS\beta}$ and $\boldsymbol{\beta} = (\boldsymbol{RS})^{-1}\boldsymbol{\theta}$ with both $\boldsymbol{S}$ and $\boldsymbol{R}$ being full-rank square matrices. Thus, a solution to the fused lasso problem can be obtained equivalently by solving a routine lasso problem with respect to coefficient vector $\boldsymbol{\theta}$ and a transformed design matrix $\boldsymbol{X}(\boldsymbol{RS})^{-1}$. As aforementioned, the estimated parameter ordering is used to construct $\boldsymbol{S}$. It is obvious that the constraint in (2.5) is convex, thus FLARCC does not suffer from multiple local minimal issue. The optimization is done using R package `glmnet` (version 2.0-2) (*Friedman et al.*, 2010), which accommodates GLMs with Gaussian, binomial and Poisson distributions.

2.3 Large-sample Properties

First we present the oracle property of our method when the parameter ordering is known, then we prove that the same large-sample properties are preserved when consistently estimated parameter ordering is used. Here we assume K is fixed. Theorems will be stated under the setting of all coefficients being heterogeneous, i.e., $\boldsymbol{c} = (1, \ldots, 1)^T$. The large-sample theories for other specification of $\boldsymbol{c}$ can be established as a special case.

Denote the true parameter values as $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$. Let the collection of true parameter orderings of all covariates and their absolute values be $\boldsymbol{W} = \{\boldsymbol{U}_j, \boldsymbol{V}_j\}_{j=1}^p$, and the estimated orderings based on the root-n consistent estimator $\hat{\boldsymbol{\beta}}$ from (2.1) as $\hat{\boldsymbol{W}} = \{\hat{\boldsymbol{U}}_j, \hat{\boldsymbol{V}}_j\}_{j=1}^p$. Denote the FLARCC estimator of $\boldsymbol{\theta}^*$ as $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ when $\boldsymbol{W}$ is known, and $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$ when the estimated parameter ordering $\hat{\boldsymbol{W}}$ is used. Let $\mathcal{A} = \bigcup_{j=1}^p \{\mathcal{A}_j\}$ be the index set of nonzero values in $\boldsymbol{\theta}^*$, where $\mathcal{A}_j = \{(j,k) : \theta^*_{j,k} \neq 0\}$, and $\mathcal{A}^c$ be the complement of $\mathcal{A}$. Thus, $\boldsymbol{\theta}^*$ can be partitioned into two subsets, the true-zero set $\boldsymbol{\theta}^*_{\mathcal{A}^c}$ and the nonzero set $\boldsymbol{\theta}^*_{\mathcal{A}}$. Similarly, let $\hat{\mathcal{A}}^{\boldsymbol{W}}$ and $\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}}$ be the index sets of nonzero elements in $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ and $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$, respectively. Let $n = \min\limits_{1\leq k\leq K} n_k$, $N = \sum_{k=1}^K n_k$, and $\lambda_N = N\lambda$.

Theorem II.1. *Suppose that λ_N satisfies $\lambda_N/\sqrt{N} \to 0$ and $\lambda_N N^{(r-1)/2} \to \infty$. Then under some mild regularity conditions (see Appendix A), the FLARCC estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$*

based on the true parameter ordering $\boldsymbol{W}$ *satisfies*

(i) (Selection Consistency) $\lim_n P(\hat{\mathcal{A}}^{\boldsymbol{W}} = \mathcal{A}) = 1$*;*

(ii) (Asymptotic Normality) $\sqrt{N}(\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_{11}^{-1})$ *as* $n \to \infty$*, where* $\boldsymbol{I}_{11}$ *is the submatrix of Fisher information matrix* $\boldsymbol{I}$ *corresponding to set* $\mathcal{A}$*.*

Theorem II.1 states that when the coefficient orderings $\boldsymbol{W}$ of $\boldsymbol{\beta}$ is known, under mild regularity conditions, the FLARCC estimator $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ enjoys selection consistency and asymptotic normality. The proof of Theorem II.1 follows *Zou* (2006) and is given in Appendix A. Now we present Theorem II.3, which states that the same properties of Theorem II.1 hold for $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$, the FLARCC estimator of $\boldsymbol{\theta}^*$ based on the estimated parameter ordering $\hat{\boldsymbol{W}}$. In effect, Theorem II.3 is a consequence of the following lemma.

Lemma II.2. *If* $\hat{\boldsymbol{\beta}}$ *is a root-n consistent estimator of* $\boldsymbol{\beta}$*, then* $\lim_n P(\hat{\boldsymbol{U}}_j = \boldsymbol{U}_j) = 1$ *and* $\lim_n P(\hat{\boldsymbol{V}}_j = \boldsymbol{V}_j) = 1$ *for* $j = 1, \ldots, p$*.*

The proof of Lemma II.2 is given in Appendix A. Lemma II.2 implies that the parameter ordering can be consistently estimated. Using Lemma II.2, we are able to extend the properties of $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ in Theorem II.1 to the proposed FLARCC estimator $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$.

Theorem II.3. *Suppose that* $\lambda_N/\sqrt{N} \to 0$ *and* $\lambda_N N^{(r-1)/2} \to \infty$*. Let the estimated parameter ordering* $\hat{\boldsymbol{W}}$ *be the ranks from a root-n initial consistent estimator* $\hat{\boldsymbol{\beta}}$*. Under the same regularity conditions of Theorem II.1, the FLARCC estimator* $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$ *satisfies*

(i) (Selection Consistency) $\lim_n P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}) = 1$*;*

(ii) (Asymptotic Normality) $\sqrt{N}(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_{11}^{-1})$ *as* $n \to \infty$*, where* $\boldsymbol{I}_{11}$ *is the submatrix of Fisher information matrix* $\boldsymbol{I}$ *corresponding to set* $\mathcal{A}$*.*

The proof of Theorem II.3 is given in Appendix A. The asymptotic normality for $\hat{\boldsymbol{\beta}}$ can also be derived by a simple linear transformation.

2.4 Tuning Parameter

In this section, we provide interpretation of the tuning parameter λ and discuss the selection criteria used for selecting λ.

2.4.1 Interpretation of ν_j's

Intuitively speaking, the study-specific coefficients of a homogeneous covariate tend to be fused at a small λ value, say λ_1, but the fusion of a heterogeneous covariate requires another λ value, λ_2, assuming $\lambda_2 > \lambda_1$. The region to draw correct clustering conclusion is $[\lambda_1, \lambda_2]$, that is, any λ within this region will produce the correct clustering result. However, when the number of covariates p is large, the region that λ can take value from to ensure the correct clustering of all p coefficient vectors simultaneously becomes narrower and may even be empty. For example, when $\lambda_2 < \lambda_1$ in the above case, no single λ is able to correctly cluster both sets of parameters. The introduction of ν_j's in (2.4) creates larger separation between homogeneous and heterogeneous groups, so that the range for λ to identify the correct clustering pattern for all covariates is better established than the case with $s = 0$, namely no use of weighting ν_j's. When the number of covariates p is large, ν_j plays a more important role in harmonizing solution paths across covariates, and the performance will be greatly improved by simultaneous tuning via a single λ.

2.4.2 Model Selection

In the current literature, the tuning parameter λ may be selected by multiple model selection criteria, such as Bayesian information criterion (BIC) (*Schwarz*, 1978) and generalized cross-validation (GCV) (*Golub et al.*, 1979). In this chapter, we consider the widely used BIC and its modification, extended BIC, i.e., EBIC (*Chen and Chen*, 2008; *Gao and Song*, 2010), which has showed the benefit of achieving sparse solutions.

Following the derivation of BIC for weighted likelihoods in *Lumley and Scott* (2015), the conventional BIC for FLARCC is defined as follows:

$$BIC_\lambda = -2\sum_{k=1}^{K}\frac{\bar{n}}{n_k}\log L_k(\hat{\boldsymbol{\beta}}(\lambda)) + \mathrm{df}(\hat{\boldsymbol{\beta}}(\lambda))\log(N), \tag{2.7}$$

where $\bar{n} = N/K$ is the average sample size per study, $L_k(\boldsymbol{\beta})$ is the study-specific likelihood, $\hat{\boldsymbol{\beta}}(\lambda)$ is the estimation of $\boldsymbol{\beta}$ at tuning parameter value λ, and $\mathrm{df}(\hat{\boldsymbol{\beta}}(\lambda)) = \sum_{j=1}^{p}\mathrm{df}(\hat{\boldsymbol{\beta}}_{j,\boldsymbol{\cdot}}(\lambda))$ is the total number of distinct parameters in $\hat{\boldsymbol{\beta}}(\lambda)$. The study-specific log-likelihoods for three most common models are listed below:

$$\begin{aligned}
\text{Normal: } \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) &\propto -\frac{n_k}{2}\log\left\{\sum_{i=1}^{n_k}\left(Y_k^{(i)} - \boldsymbol{X}_k^{(i)T}\hat{\boldsymbol{\beta}}(\lambda)\right)^2/n_k\right\};\\
\text{Logistic: } \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) &\propto \sum_{i=1}^{n_k}\left\{Y_k^{(i)}\boldsymbol{X}_k^{(i)T}\hat{\boldsymbol{\beta}}(\lambda) - \log\left(1 + e^{\boldsymbol{X}_k^{(i)T}\hat{\boldsymbol{\beta}}(\lambda)}\right)\right\};\\
\text{Poisson: } \log L_k(\hat{\boldsymbol{\beta}}(\lambda)) &\propto \sum_{i=1}^{n_k}\left\{Y_k^{(i)}\boldsymbol{X}_k^{(i)T}\hat{\boldsymbol{\beta}}(\lambda) - e^{\boldsymbol{X}_k^{(i)T}\hat{\boldsymbol{\beta}}(\lambda)}\right\}.
\end{aligned}$$

To improve the BIC by further controlling model size and encouraging sparer models, we adapt the EBIC for FLARCC, which takes the following form:

$$EBIC_\lambda = -2\sum_{k=1}^{K}\frac{\bar{n}}{n_k}\log L_k(\hat{\boldsymbol{\beta}}(\lambda)) + \mathrm{df}(\hat{\boldsymbol{\beta}}(\lambda))\log(N) + 2\gamma\log\sum_{j=1}^{p}\binom{K}{\mathrm{df}(\hat{\boldsymbol{\beta}}_{j,\boldsymbol{\cdot}}(\lambda))}, \tag{2.8}$$

where $\gamma \in [0, 1]$ is a tuning parameter that is typically fixed at 1 as done in our numerical experiments. Note that EBIC reduces to BIC when $\gamma = 0$. The last term in (2.8) encourages a sparser solution in comparison to the conventional BIC. Simulation studies in Section 2.5 provide numerical evidence to elucidate the difference between

BIC and EBIC in terms of their performance on achieving sparsity.

In a view of hierarchical clustering, the solution path of each covariate can be thought of as a hierarchical clustering tree. For the jth covariate, $\lambda = 0$ corresponds to the bottom of the clustering tree; and $\lambda = \lambda_{Fuse,j}$, the smallest λ value to achieve complete parameter fusion, corresponds to the top of the clustering tree. The completely heterogeneous model corresponds to the position on the solution path at $\lambda = 0$ and the completely homogeneous model corresponds to the model at $\lambda = \lambda_{Fuse} := \max_{1 \le j \le p} \lambda_{Fuse,j}$.

2.5 Simulation Studies

This section presents results from two simulation experiments. The first simulation compares the performance of FLARCC under different GLM regression models. The second simulation is a more complicated scenario with large K and more non-important covariates, where covariate selection is also of interest.

2.5.1 Simulation Experiment 1

The first simulation study aims to assess the performance of our method for different GLM regression models. For this, we consider combining data sets from $K = 10$ different studies with, for simplicity, equal sample size $n_1 = \cdots = n_{10} = 100$. Data are simulated from the following mean regression model:

$$h\{E(Y_k^{(i)})\} = \beta_{1,k} X_{1,k}^{(i)} + \beta_{2,k} X_{2,k}^{(i)} + \beta_{3,k} X_{3,k}^{(i)},\ i = 1, \ldots, 100,\ k = 1, \ldots, 10,$$

where the true coefficient vectors have the following clustering structures:

$$\begin{aligned}
\boldsymbol{\beta}_{1,\bullet} &= (\underbrace{0,\ldots,0}_{10})^T;\\
\boldsymbol{\beta}_{2,\bullet} &= (\underbrace{0,\ldots,0}_{5},\underbrace{1,\ldots,1}_{5})^T;\\
\boldsymbol{\beta}_{3,\bullet} &= (\underbrace{-1,\ldots,-1}_{3},\underbrace{0,\ldots,0}_{4},\underbrace{1,\ldots,1}_{3})^T.
\end{aligned}$$

The true values in $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are heterogeneous, while the true values in $\boldsymbol{\beta}_1$ are homogeneous across studies. The three covariates are correlated with exchangeable correlation of 0.3 and marginally distributed according to the standard normal distributions, $\mathcal{N}(0,1)$. Three types of GLM regression models are considered: linear model for continuous normal outcomes (with errors simulated from $\mathcal{N}(0,1)$), logistic model for binary outcomes and Poisson model for count outcomes.

To evaluate the performance of FLARCC to correctly detect patterns of all covariates, we assume all covariates are heterogeneous across studies with no prior knowledge on clustering structure of any covariate. Intercept is fitted and assumed to be homogeneous. No sparsity penalty is applied on the covariates (i.e., $\alpha = 0$) in this simulation experiment. Coefficients of all three covariates are fused simultaneously, and the optimal tuning parameter λ_{opt} is selected by EBIC. We report sensitivity and specificity as metrics of the performance of FLARCC to identify similar and distinct coefficient pairs. Sensitivity measures the proportion of equal coefficient pairs that are correctly identified. Similarly, specificity measures the proportion of unequal coefficients pairs that are correctly identified; however, specificity is not defined for homogeneous covariates which have no unequal coefficient pairs. In addition, we calculate the mean squared error (MSE) for each $\hat{\boldsymbol{\beta}}_{j,\bullet}$ across all K studies, defined as $\text{MSE}_j = \sum_{k=1}^{K}(\hat{\beta}_{j,k} - \beta_{j,k})^2/K, j = 1,\ldots,p$, and compare with the MSE of each estimate based on homogeneous model ($\lambda = \lambda_{Fuse}$) and heterogeneous model ($\lambda = 0$).

Table 2.1 shows the results of simulation experiment 1 from 1,000 simulation repli-

Table 2.1: Results of simulation experiment 1 for FLARCC when scaling weight parameter $s = 0$ and $s = 1$ with λ selected by EBIC, for the linear, logistic and Poisson models. Tuning parameters are reported in log scale, i.e., $\tilde{\lambda} = \log_{10}(\lambda + 1)$. Results are summarized from 1,000 replications.

Method	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$ size	$\tilde{\lambda}_{Fuse,j}$	Sensi-	Speci-	MSE when $\lambda =$		
$(\tilde{\lambda}_{opt})$				tivity	ficity	λ_{opt}	λ_{Fuse}	0
			Linear: continuous response					
$s = 0$ (0.154)	$\boldsymbol{\beta}_1$	1.067	0.111	0.974	–	0.001	0.002	0.012
	$\boldsymbol{\beta}_2$	2.075	1.275	0.982	1.000	0.003	0.253	0.012
	$\boldsymbol{\beta}_3$	3.081	1.368	0.982	1.000	0.004	0.603	0.012
$s = 1$ (0.349)	$\boldsymbol{\beta}_1$	1.006	0.080	0.998	–	0.001	0.002	0.012
	$\boldsymbol{\beta}_2$	2.058	1.584	0.986	1.000	0.003	0.253	0.012
	$\boldsymbol{\beta}_3$	3.123	1.972	0.974	1.000	0.004	0.603	0.012
			Logistic: binary response					
$s = 0$ (0.066)	$\boldsymbol{\beta}_1$	1.270	0.064	0.898	–	0.010	0.005	0.070
	$\boldsymbol{\beta}_2$	2.572	0.318	0.819	0.963	0.047	0.268	0.087
	$\boldsymbol{\beta}_3$	3.682	0.437	0.784	0.964	0.069	0.607	0.091
$s = 1$ (0.112)	$\boldsymbol{\beta}_1$	1.075	0.050	0.972	–	0.007	0.005	0.069
	$\boldsymbol{\beta}_2$	2.478	0.414	0.837	0.952	0.052	0.268	0.088
	$\boldsymbol{\beta}_3$	3.912	0.711	0.749	0.971	0.064	0.607	0.091
			Poisson: count response					
$s = 0$ (0.187)	$\boldsymbol{\beta}_1$	1.087	0.129	0.976	–	0.001	0.005	0.008
	$\boldsymbol{\beta}_2$	2.084	1.751	0.984	1.000	0.001	0.271	0.008
	$\boldsymbol{\beta}_3$	3.076	1.885	0.986	1.000	0.002	0.659	0.008
$s = 1$ (0.433)	$\boldsymbol{\beta}_1$	1.047	0.087	0.992	–	0.001	0.005	0.008
	$\boldsymbol{\beta}_2$	2.088	2.060	0.984	1.000	0.002	0.271	0.008
	$\boldsymbol{\beta}_3$	3.111	2.536	0.978	1.000	0.002	0.657	0.008

cates. The MSE of all estimated covariates based on FLARCC ($\lambda = \lambda_{opt}$) are consistently and significantly smaller than those based on the homogeneous ($\lambda = \lambda_{Fuse}$) and heterogeneous ($\lambda = 0$) models, regardless of the model type. FLARCC performs very well in the linear and Poisson regressions in terms of identifying the correct clustering, with the sensitivity and specificity both above 95% for all covariates (specificity is not reported for $\boldsymbol{\beta}_1$ since there is no unequal pair within $\boldsymbol{\beta}_{1,\cdot}$). Sensitivity and specificity of FLARCC drop in the logistic regression, especially as the level of heterogeneity increases. One reason for the reduced performance of FLARCC in the logistic regression is that the estimated variances of regression coefficients in the logistic model are larger than in the linear and Poisson models, given the same coefficient setting. Therefore, the estimated parameter ordering for which our method is based on may be less accurate. For the logistic regression, increasing sample sizes is one of the possible ways to improve the performance. The performance difference between scaling weight parameter $s = 0$ and $s = 1$ in (2.4) is small in this case because of the relatively small number of covariates $p = 3$. Additionally, since K is small in this case, the optimal λ selected by BIC and EBIC are very close, thus we only display results based on EBIC. As p and K become larger, FLARCC will increasingly benefit from the additional weights ν_j (i.e., $s = 1$) and EBIC, as will be shown in Section 2.5.2. A sensitivity analysis to investigate how the initial ordering affect the performance of FLARCC is conducted, with results shown in Appendix B. We show that when the initial parameter ordering is slightly distorted, our method still achieves satisfactory performance.

2.5.2 Simulation Experiment 2

The second simulation study aims to evaluate the performance of FLARCC in a more challenging setting. More specifically, we consider data sets from $K = 100$ studies, each with a sample size 100, totaling 10,000 subject-level observations. Comparing to the previous setting, we increase the number of covariates and reduce the gaps between heterogeneous coefficients. For each study, we simulate data from the following linear regression model:

$$E(Y_k^{(i)}) = \sum_{j=1}^{8} \beta_{j,k} X_{j,k}^{(i)}, \ i = 1, \dots, 100, \ k = 1, \dots, 100.$$

The signals are set sparse; only the first four covariates with coefficient vectors, $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_4$, are influential to Y with the true clustered effect patterns given as follows:

$$
\begin{aligned}
\boldsymbol{\beta}_{1,\bullet} &= (\underbrace{0,\dots,0}_{50},\underbrace{0.5,\dots,0.5}_{50})^T,\\
\boldsymbol{\beta}_{2,\bullet} &= (\underbrace{-0.5,\dots,-0.5}_{30},\underbrace{0,\dots,0}_{40},\underbrace{0.5,\dots,0.5}_{30})^T,\\
\boldsymbol{\beta}_{3,\bullet} &= (\underbrace{-0.5,\dots,-0.5}_{25},\underbrace{0,\dots,0}_{25},\underbrace{0.5,\dots,0.5}_{25},\underbrace{1,\dots,1}_{25})^T,\\
\boldsymbol{\beta}_{4,\bullet} &= (\underbrace{-1,\dots,-1}_{20},\underbrace{-0.5,\dots,-0.5}_{20},\underbrace{0,\dots,0}_{20},\underbrace{0.5,\dots,0.5}_{20},\underbrace{1,\dots,1}_{20})^T,
\end{aligned}
$$

whereas $\boldsymbol{\beta}_5$ to $\boldsymbol{\beta}_8$ are all zero, i.e., $\boldsymbol{\beta}_{j,\bullet} = (0,\dots,0)^T$, for $j = 5,6,7,8$. All covariates are equally correlated with an exchangeable correlation of 0.3 and marginally distributed according to $\mathcal{N}(0,1)$. We set $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_8$ as being heterogeneous from the start and fuse all of them simultaneously. We apply the additional sparsity penalty to all covariates by setting $\alpha = 1$. The intercept is assumed to be homogeneous in the analysis.

Since K is large, we also present results from individual covariate K-means clustering. This is a two-step method where we first estimate regression coefficients within each study, and then separately for each covariate, we perform the K-means clustering on the estimated study-specific coefficients of each covariate. The number of clusters is selected by the generalized cross-validation criterion $\sum_{k=1}^{K}(\hat{\beta}_k - \hat{\beta}_{c(k)})^2/(K - \mathrm{GDF})^2$, with $\hat{\beta}_{c(k)}$ being the cluster center of $\hat{\beta}_k$ and GDF is the generalized degrees of freedom estimated according to *Ye* (1998), where purturbations are generated independently from $\mathcal{N}(0, 0.01)$. The cluster centroids are then used as the estimates of the group-level parameters.

Table 2.2 summarizes the simulation results for linear model where the errors are generated independently from $\mathcal{N}(0,1)$. Similar to simulation 1, FLARCC gives the smallest MSE for heterogeneous covariates, $\boldsymbol{\beta}_1$ to $\boldsymbol{\beta}_4$, among all three models,

Table 2.2: Result of simulation experiment 2 under the linear model. Scaling weight parameter is set at $s = 0$ and $s = 1$. Tuning parameters are reported in log scale, i.e., $\tilde{\lambda} = \log_{10}(\lambda + 1)$. Sparsity denotes the proportion of zero in estimation. Results are summarized from 1,000 replications.

Method	$\boldsymbol{\beta}$	$\hat{\boldsymbol{\beta}}$ size	$\tilde{\lambda}_{Fuse,j}$	Sensi-	Speci-	Spar-	MSE when $\lambda =$		
$(\tilde{\lambda}_{opt})$				tivity	ficity	sity	λ_{opt}	λ_{Fuse}	0
$s = 0$ BIC (0.143)	$\boldsymbol{\beta}_1$	8.115	1.517	0.373	0.995	0.199	0.006	0.063	0.014
	$\boldsymbol{\beta}_2$	10.689	1.551	0.401	0.996	0.166	0.008	0.150	0.014
	$\boldsymbol{\beta}_3$	13.107	1.962	0.443	0.997	0.113	0.008	0.313	0.014
	$\boldsymbol{\beta}_4$	15.178	1.984	0.461	0.997	0.091	0.009	0.500	0.014
	$\boldsymbol{\beta}_5$	4.818	0.301	0.322	–	0.338	0.003	0.000	0.014
	$\boldsymbol{\beta}_6$	4.860	0.305	0.322	–	0.339	0.003	0.000	0.014
	$\boldsymbol{\beta}_7$	4.860	0.302	0.321	–	0.330	0.003	0.000	0.014
	$\boldsymbol{\beta}_8$	4.820	0.301	0.319	–	0.336	0.003	0.000	0.014
$s = 0$ EBIC (0.159)	$\boldsymbol{\beta}_1$	7.538	1.509	0.417	0.994	0.224	0.006	0.063	0.014
	$\boldsymbol{\beta}_2$	9.975	1.546	0.441	0.995	0.182	0.007	0.150	0.014
	$\boldsymbol{\beta}_3$	12.212	1.953	0.483	0.996	0.124	0.008	0.313	0.014
	$\boldsymbol{\beta}_4$	14.096	1.978	0.503	0.997	0.101	0.008	0.500	0.014
	$\boldsymbol{\beta}_5$	4.388	0.298	0.377	–	0.394	0.003	0.000	0.014
	$\boldsymbol{\beta}_6$	4.413	0.303	0.379	–	0.397	0.003	0.000	0.014
	$\boldsymbol{\beta}_7$	4.408	0.299	0.374	–	0.385	0.003	0.000	0.014
	$\boldsymbol{\beta}_8$	4.385	0.298	0.374	–	0.392	0.003	0.000	0.014
$s = 1$ EBIC (0.492)	$\boldsymbol{\beta}_1$	3.563	1.589	0.800	0.981	0.422	0.006	0.063	0.014
	$\boldsymbol{\beta}_2$	6.111	1.810	0.708	0.989	0.291	0.007	0.150	0.014
	$\boldsymbol{\beta}_3$	8.843	2.388	0.667	0.994	0.168	0.007	0.313	0.014
	$\boldsymbol{\beta}_4$	11.358	2.521	0.635	0.995	0.128	0.008	0.500	0.014
	$\boldsymbol{\beta}_5$	1.329	0.275	0.928	–	0.932	0.000	0.000	0.014
	$\boldsymbol{\beta}_6$	1.321	0.280	0.933	–	0.937	0.000	0.000	0.014
	$\boldsymbol{\beta}_7$	1.311	0.274	0.934	–	0.935	0.000	0.000	0.014
	$\boldsymbol{\beta}_8$	1.321	0.273	0.929	–	0.936	0.000	0.000	0.014
							MSE from K-means		
K-means GCV (GDF)	$\boldsymbol{\beta}_1$	7.196	–	0.753	0.971	0.000			0.008
	$\boldsymbol{\beta}_2$	11.236	–	0.671	0.983	0.000			0.009
	$\boldsymbol{\beta}_3$	13.721	–	0.639	0.984	0.000			0.011
	$\boldsymbol{\beta}_4$	18.308	–	0.527	0.985	0.000			0.014
	$\boldsymbol{\beta}_5$	6.415	–	0.759	–	0.000			0.004
	$\boldsymbol{\beta}_6$	5.271	–	0.769	–	0.000			0.004
	$\boldsymbol{\beta}_7$	5.629	–	0.767	–	0.000			0.004
	$\boldsymbol{\beta}_8$	5.080	–	0.794	–	0.000			0.004

and has comparable MSE as the homogeneous model for homogeneous covariates, $\boldsymbol{\beta}_5$ to $\boldsymbol{\beta}_8$. More interestingly, when K is large, BIC does not provide satisfactory model selection, erring on the lack of parsimony, while EBIC encourages stronger fusion and improves the ability to detect equal coefficient pairs in all eight covariates, regardless of their levels of heterogeneity. In addition, EBIC improves the sparsity detection among both the important and nonimportant covariates. It is interesting to note that the choice between BIC and EBIC does not alter solution paths, but only model selection. FLARCC with scaling weight parameter $s = 1$ has the best clustering performance among all compared methods. The difference between the choices of $s = 0$ and $s = 1$ is substantial in simulation 2, in contrast to the results from simulation 1. This indicates that the covariate-specific weights for heterogeneity $\{\nu_j\}_{j=1}^{p}$ are very effective to improve the performance of the proposed fusion learning, especially when K and p are large. Sensitivity and specificity of the two-step K-means clustering method are higher than those of FLARCC with $s = 0$, but lower than those of FLARCC with $s = 1$. The two-step K-means has larger MSE than FLARCC because it does not consider the correlation between covariates. More importantly, the K-means clustering is a model-free method, so the results obtained from this method cannot be plugged in back to the model for prediction. As suggested from the empirical results of both simulation experiments, EBIC tends to provide better model selection for FLARCC than the conventional BIC.

2.6 Applications

2.6.1 Clustering of Regional Effects

In this data analysis example, we like to demonstrate the use of our method to derive clusters of regional effects. Here we consider the Panel Study of Income Dynamics (PSID), which is a household survey study following thousands of families

Table 2.3: Coefficient estimates of the homogeneous model ($\lambda = \lambda_{Fuse}$), the heterogeneous model ($\lambda = 0$) and the fused model using FLARCC with λ selected by EBIC, respectively.

Region	n	Intercept $\boldsymbol{\beta}_0$	Age $\boldsymbol{\beta}_1$	Sex $\boldsymbol{\beta}_2$	Birth Wt. $\boldsymbol{\beta}_3$	Income $\boldsymbol{\beta}_4$
(A) Homogeneous model – combine all regions						
All regions	1880	0.000	0.206	0.016	0.063	-0.096
(B) Heterogeneous model – region specific estimates						
1-Northeast	239	-0.133	0.228	-0.079	-0.003	0.004
2-Midwest	493	-0.054	0.229	0.017	0.124	-0.132
3-South	805	0.128	0.158	0.095	0.068	-0.071
4-West	343	-0.155	0.236	-0.083	0.057	-0.074
(C) Fused model using FLARCC						
1-Northeast	239	-0.093	0.201	-0.036	0.000	0.000
2-Midwest	493	-0.093	0.201	0.000	0.021	-0.047
3-South	805	0.075	0.201	0.000	0.021	-0.047
4-West	343	-0.093	0.201	-0.036	0.021	-0.047

across different states in the US. PSID collects information of employment, income, health, and so on. In this data analysis, we focus on the association of household income with body mass index (BMI) on school-aged children between age of 11 and 19, adjusted for age, gender and birth weight. Data of 1880 children were gathered from four census regions (1-Northeast, 2-Midwest, 3-South and 4-West), as defined by *U.S. Census Bureau* (2015). All variables are standardized before model fitting. We are interested in investigating if regional heterogeneity exists and if the effects of interest differ across regions with region-dependent patterns.

Table 2.3 shows the results of coefficient estimates obtained from three different models: (A) homogeneous model ($\lambda = \lambda_{Fuse}$), coefficients estimated by combining data sets from four regions, (B) heterogeneous model ($\lambda = 0$), coefficients estimated separately by region-specific data, and (C) FLARCC ($\lambda = \lambda_{EBIC}$). Model A suggests that age and birth weight are positively associated with BMI for the subjects, but income was negatively associated with BMI. The estimates from Model B suggest that heterogeneous coefficient patterns exist among these associations since conclusions

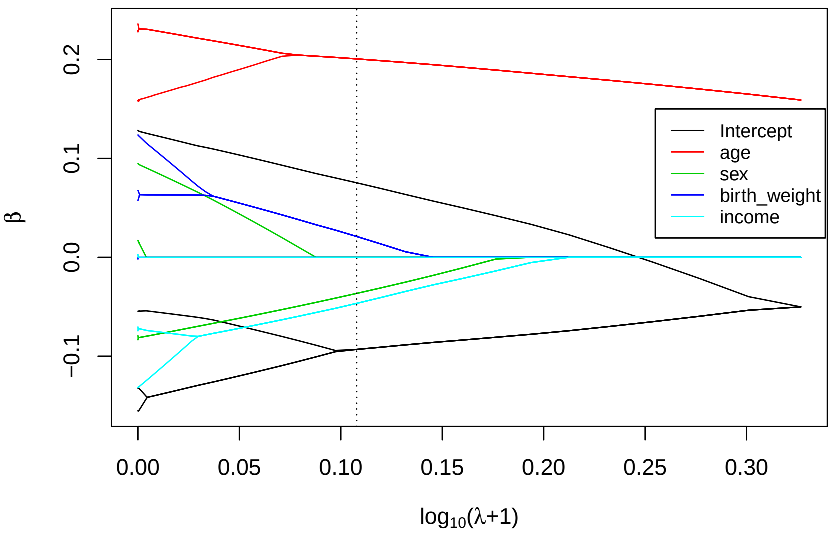
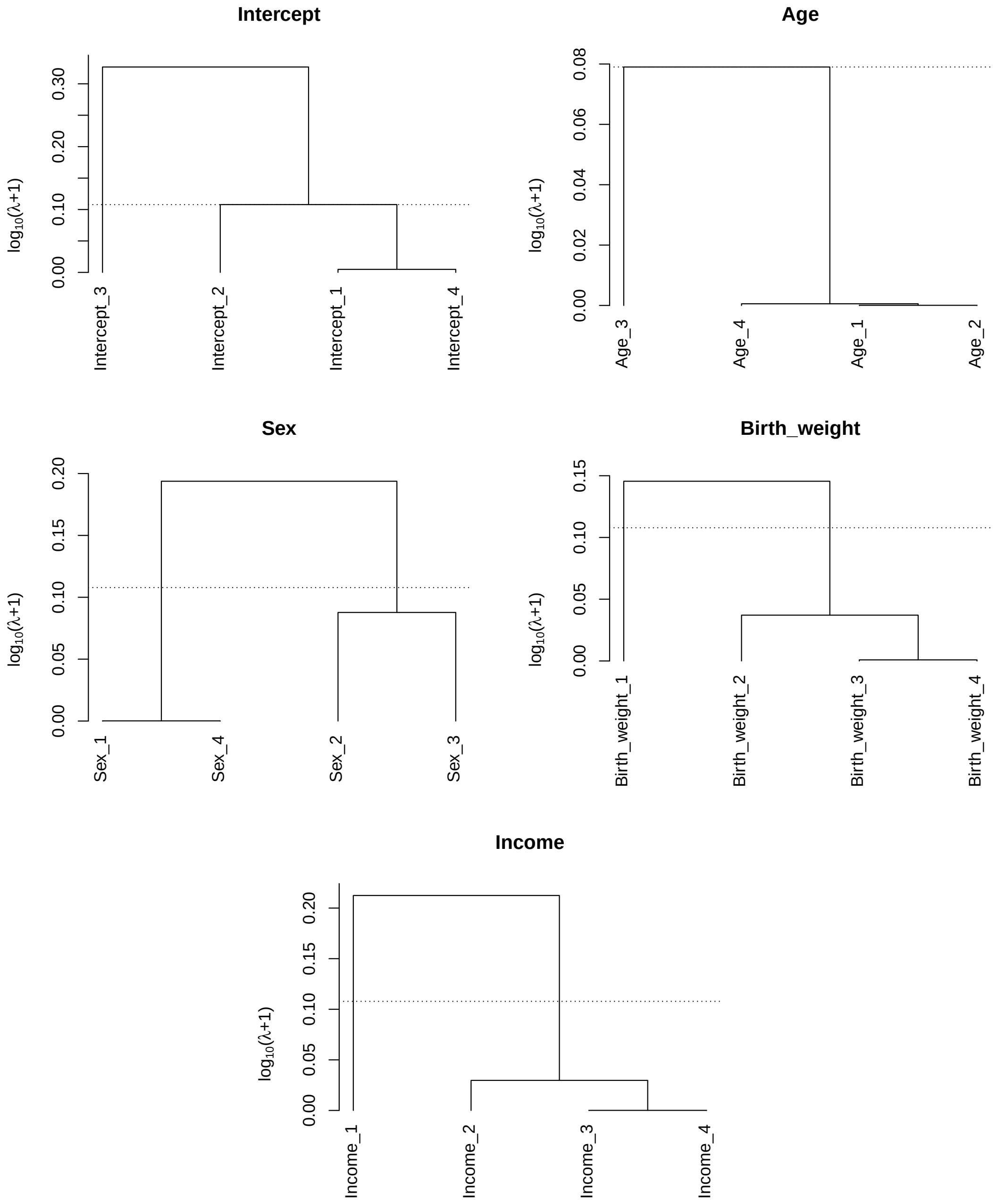


Figure 2.2: FLARCC solution paths of all covariates over the transformed tuning parameter $\tilde{\lambda} = \log_{10}(\lambda + 1)$, with $s = 0$. The vertical dotted line denotes the optimal tuning parameter value $\tilde{\lambda}_{EBIC}$.

differ between regions. Model C appears more sensible when regression coefficients are heterogeneous across these regions. Since K and p are small in this data application, we apply FLARCC with $s = 0$ on the PSID data, assuming effects of income, age, gender and birth weight are heterogeneous across regions, and set sparsity parameter $\alpha = 1$ for variable selection.

Based on the results from FLARCC, the estimated mean of standardized BMI in the South is 0.168 higher (or 0.97 higher in original scale of BMI) than that of the other three regions, which share the same mean. The effects of age are consistent across four regions. The effects of gender are classified into two clusters. The mean of standardized BMI of females is 0.036 lower (or 0.42 lower in original scale) than that of males in the Northeast and the West, but males and females have the same mean BMI in the Midwest and the South. Standardized BMI increases by 0.021 for every standard deviation increase of birth weight (or BMI increases by 0.19 for every unit increase of birth weight) in all regions except the Northeast. Similarly, standardized BMI decreases by 0.047 for every standard deviation increase of log income (or BMI



Intercept

Age

Sex

Birth_weight

Income

Figure 2.3: *Fusograms* of all covariates based on FLARCC solution paths. The horizontal dotted lines denote the optimal regression coefficient clustering determined by EBIC.

decreases by 0.27 for every unit increase of income) in all regions except the Northeast where BMI is not affected by income. The leave-one-out mean squared prediction errors for model A, B and C are 0.953, 0.945 and 0.950, respectively. The differences between the prediction errors are small because of the relatively small effect sizes of the heterogeneous covariates identified by FLARCC, i.e., sex, birth weight and income. The most significant covariate, age, is homogeneous thus it does not differentiate the prediction power among the three models. Solution paths and *fusograms* of all covariates are shown in Figure 2.2 and Figure 2.3, respectively, for illustration. In summary, FLARCC ensures parsimony where necessary to maximize the prediction power of the final model; and it provides more informative interpretation and better visualization than the other two traditional models.

2.6.2 Clustering of Cohort Effects

In the second data analysis example, we explore the heterogeneity across multiple cohorts base on an environmental study in Mexico City named Early Life Exposures in Mexico to ENvironmental Toxicants (ELEMENT). Since 1994, ELEMENT has recruited pregnant women and continuously followed their children during infancy, early childhood and adolescence. The ELEMENT study consists of three cohorts, each with slightly different objectives and study designs. Using ELEMENT, a previously study by *Watkins et al.* (2014) has found disruptive effect of prenatal exposures to phthalate, plasticizers that are added to daily plastic products to increase their flixibility, transparency and durability, on the timing of sexual maturation. Recently, we profiled the metabolome of the same group of $n = 242$ children at their adolescence, and investigate whether such disruptive effect exists in their metabolic profiles.

Univariate screening with false discovery rate controlled at 10% reveals that, in the female sample, maternal trimester 3 mono (2-ethyl-5-hydroxyhexyl) phthalate (MEHHP) is significantly associated with dodecenedioc acid – a median chain

Table 2.4: Coefficient estimates of the original homogeneous model ($\lambda = \lambda_{Fuse}$) (** indicates $p < 0.001$) and the fused model using FLARCC with λ selected by EBIC, respectively.

Cohort	n	Intercept $\boldsymbol{\beta}_0$	MEHHP $\boldsymbol{\beta}_1$	Age $\boldsymbol{\beta}_2$	BMI $\boldsymbol{\beta}_3$	Pubertal Onset $\boldsymbol{\beta}_4$
			Original model			
All females	128	-1.111 **	0.352 **	-0.191	-0.192 **	0.280
			Heterogeneous fusion model			
1-PL females	27	-0.808	0.264	0.147	-0.212	0.000
2-BI females	6	-0.808	0.000	0.147	0.000	0.000
3-SF females	95	-0.808	0.212	-0.506	-0.212	0.000

fatty acid implicated in metabolic risk in adult populations. Afterwards, we applied FLARCC on the female samples allowing for heterogeneous effects across the three cohorts. Results comparing the two different ways of modeling are shown in Table 2.4. As we can see, girls from the first cohort (PL) and the third cohort (SF) exhibit up-regulation in dodecenedioc acid due to MEHHP, whereas in the second cohort (BI) we fail to see any significance after regularization. Such heterogeneity may be due to the reason that PL and SF are pregnancy cohorts (mothers are admitted to the study at the beginning of pregnancy) thus the protocols for measuring prenatal exposures are better designed, but BI is a birth cohort (mothers are admitted close to delivery) thus the measurement at the 3rd trimester might not be as well planned and administered as PL and SF. Similar to many other papers that remove outliers to robustify regressions (e.g., *Hoaglin and Welsch* (1978); *Rousseeuw and Leroy* (2005)), we suggest treating BI as an “outlying cohort” and report separate results only based on PL and SF.

2.7 Concluding Remarks

The proposed method, published as *Tang and Song* (2016), brings a new perspective to model fitting when combining multiple data sets from different sources is of primary interest. As data volumes and data sources grow fast, more and more opportunities and demands emerge in practice to borrow strengths of combined data sets. In such case, traditional methods are challenged by the complex data structures and do not provide desirable treatments and meaningful interpretations to data heterogeneity, especially when the number of data sets is very large. FLARCC allows the flexibility to explore the heterogeneity pattern of parameters among large number of data sets by tuning the shrinkage parameter.

When K and p are small, weights $\{\nu_j\}_{j=1}^{p}$ do not contribute to much difference in terms of clustering and estimation. However, since only one tuning parameter is used to regularize the fusion of all covariates, when both K and p are large, we suggest letting $s > 0$ to allow covariate-specific weights adapting to the heterogeneousness of coefficients from individual covariates to achieve better results. In addition, the estimation consistency of rank estimator is a critical component needed to determine adjacent pairs. The current consistency is established under the case of K being fixed, and the validity of its property is unknown when K increases along the total sample size.

FLARCC can be applied to various scientific problems, such as the detection of outlying studies by singling out outlying coefficients; it can also be applied to the clustering of patient trajectories by viewing the time series data of patients as individual studies. Essentially, all study that are interested in the group-specific effects may be analyzed from the perspective of parameter fusion using the proposed method. The work in this chapter has been generalized to the partial likelihood framework for the Cox proportional-hazards model (*Cox*, 1972); details are presented in Appendix C.

CHAPTER III

Method of Divide-and-Combine in Regularized Generalized Linear Models for Big Data

3.1 Introduction

In this chapter, we consider the generalized linear model under the big data scenario that a dataset is too large to be centralized, thus has to be stored by the means of distributed computer clusters. It presents great challenges to statisticians to analyze such massive data because the entire dataset cannot be loaded to a single processor for computation (*Fan et al.*, 2014). Following the development of cloud storage and cloud computing, the method of divide-and-combine (*Aho and Hopcroft*, 1974) has become the state-of-the-art in big-data science to cope with the scalability issue. The most well-known example is MapReduce programming (*Dean and Ghemawat*, 2008), a divide-and-combine framework that executes on top of the Hadoop Distributed File System (*Shvachko et al.*, 2010). Divide-and-combine is a procedure to recursively subdivide the data into relatively independent batches, which are in turn processed in parallel. Next, the separate results are combined together in a way that algebra permits. Existing implementation of divide-and-combine is only available for a limited number of functions where parallelization is straightforward, such as the computation of mean, frequency and other summary statistics. Other more complicated methods require special treatment in order to be adapted to the parallel computing architecture, see for examples in *Guha et al.* (2009, 2012), *Mackey et al.* (2011) and *Madria and Hara* (2015). In this chapter, we consider the statistical inference problem for generalized linear models, using divide-and-combine, for extremely large datasets.

For a generalized linear model, the systematic component is specified by the mean of response y_i that is related to a p-dimensional vector of the covariates $\boldsymbol{x}_i$ by a known link function $g(\cdot)$ in the form $g\{E(y_i)\} = \boldsymbol{x}_i^T\boldsymbol{\beta}$, subject $i = 1, \ldots, N$. The random component is specified by the conditional density of $\boldsymbol{Y} = (y_1, \cdots, y_N)^T$ given $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^T$. The associated likelihood function is given by $\mathcal{L}_N(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^N \exp[\{y_i\theta_i - b(\theta_i)\}/\phi + c(y_i, \phi)]$, where the canonical parameters have the form $\theta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, $i = 1, \cdots, N$, with $\boldsymbol{\beta}$ being the p-element vector of regression parameters of interest and ϕ being the dispersion parameter. Both the sample size N and/or the number of covariates p may be large in practice. Due to the fact that the maximum likelihood estimator, $\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_N(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X})$, in general has no closed-form expression, existing methods often require iteratively accessing all sub-datasets repeatedly, resulting in high data communication cost. It is not trivial to formulate perfectly parallel algorithms that only require a single passing of each divided dataset (*Kleiner et al.*, 2014; *Song and Liang*, 2015) in the sense that they still provide numerically robust statistical inference as compared to using full data. By perfectly parallel, we mean that a big problem can be broken into small problems which can be executed in parallel and combined in the final step. In the development of the divide-and-combine strategy in the context of statistical inference, naturally one question arises: do the proposed estimator and the maximum likelihood estimator obtained from the entire data warrant asymptotic equivalence, leading to comparable statistical inferences?

Combining independent samples from different studies in the form of summary statistics has long been studied under the topic of meta-analysis (see for example

Sutton and Higgins (2008); *Stangl and Berry* (2000); *Hedges and Olkin* (2014)). The classical meta-analysis method uses inverse variance weighted average to combine separate point estimates from individual data batches. This is an efficient divide-and-combine solution to generalized linear regressions for data on distributed systems, because raw data can be processed locally and only summary information are sent between machines. *Lin and Zeng* (2010) showed that such meta estimator asymptotically achieves the Fisher's efficiency, in other words, it follows asymptotically the same distribution as the Fisher's maximum likelihood estimator directly obtained from the whole data. The Fisher's efficiency has been also established for a combined estimator by *Lin and Xi* (2011) through a meta-type method of aggregative estimating equations, under some relatively strong conditions, such as the number of sub-datasets K is of order $O(n^r)$ where $r < 1/3$ and n is the sample size of a sub-dataset. Recently, *Battey et al.* (2015) proposed test statistics and point estimators in the context of the divide and conquer algorithms, where the method of hypothesis testing is only developed for low dimensional parameters, and the combined estimator is given as an arithmetic average of sub-datasets. Under the Bayesian framework, similar procedures have also been developed focusing on the aggregation step of combining posterior distributions from divided sub-datasets, for example, *Minsker et al.* (2014) and *Srivastava et al.* (2015).

In this chapter, we adapt the confidence distribution approach (*Xie and Singh*, 2013) to combine sub-dataset results. The confidence distribution, originally proposed by *Fisher* (1956) and later formally formulated by *Efron* (1993), has recently attracted a surge of renewed attention in the statistical literature; see for example, *Singh et al.* (2005), *Xie and Singh* (2013) and references therein. An advantage of the confidence distribution approach is that it provides a unified framework for combining distributions of estimators, so statistical inference with the combined estimator can be established in a straightforward and mathematically rigorous fashion. Specifically related to divide-and-combine, *Xie et al.* (2012) developed a robust meta-analysis-type approach through confidence distribution, and *Liu et al.* (2015) proposed to combine the confidence density function in the same way as combining likelihood functions for inference that warrants the Fisher's efficiency.

If is often the case that big data in practice has extremely large sample size N and a relatively large number of covariates p, *say*, from hundreds to thousands. Although the overall variable-to-sample ratio p/N is typically small, in each sub-dataset such ratio becomes Kp/N when the full data are divided into K separate batches. Here, we consider p fixed whereas K can go to infinity. The sample size reduction due to the data division may cause numerical instability for the search for the maximum likelihood estimate. In addition, to deal with the case in that most of covariates are unimportant, which often occurs when hundreds to thousands covariates are included in the analysis, it is often preferable to invoke regularized or penalized methods for dimension reduction, such as lasso (*Tibshirani*, 1996; *Zou and Hastie*, 2005), SCAD (*Fan and Li*, 2001) and MCP (*Zhang*, 2010). For regularized estimation, constructing confidence density for penalized estimators is analytically challenging because: (i) sparse estimators such as lasso estimators do not have a tractable limiting distribution, and (ii) the oracle property such as the asymptotic normal distribution for estimators of the truly non-zero covariate effects is hardly used in practice because the truth of important or unimportant covariates is never known in advance. In the supplementary material, we provide additional theoretical results for $p \to \infty$.

When penalized regression is applied on each sub-dataset, variable selection procedure will choose different sets of important covariates by different tuning schemes. Such misaligned estimation results prohibit the meta-analysis approach from combining separate results; both dimensionality and meaning of the estimates across data batches may be very different. *Chen and Xie* (2014) proposed to use a majority-voting method to select the most frequently identified covariates by the lasso method across the sub-datasets to be combined in the final estimation. Unfortunately, not only is this method sensitive to the choice of inclusion criterion, but more critically it does not provide inference for the combined estimator. To overcome this problem, we propose a new approach along the lines of the post-selection inference developed for the penalized estimator by *van de Geer et al.* (2014) and *Zhang and Zhang* (2014), which allows us to combine lasso estimators obtained from sub-datasets. Our new contribution is two-fold: (i) the combined estimator achieves asymptotically the Fisher's efficiency; that is, it is asymptotically as efficient as the maximum likelihood estimator obtained from the direct analysis on the whole data; and (ii) the computation of searching for the maximum likelihood estimator is scalable and parallelized to address very large sample sizes through easy and fast parallel algorithmic implementation. The latter presents a desirable numerical recipe to handle the case when the whole data analysis is time consuming and CPU demanding, or even numerically prohibitive.

The remaining of this chapter is organized as follows. Section 3.2 focuses on the asymptotics of the debiased lasso estimator. Section 3.3 presents the confidence distribution method to combine results from multiple regularized regressions. Section 3.4 provides extensive simulation results, and Section 3.5 illustrates our method by a real data example. We conclude with a brief discussion in Section 3.6.

3.2 Regularized Regression

3.2.1 Lasso in Generalized Linear Models

This section focuses on the regularized estimation and confidence density for one sub-dataset of sample size n. We choose lasso, the least absolute shrinkage and selection operator (*Tibshirani*, 1996), as the method of penalized estimation in the development of a divide-and-combine procedure. With little effort, the other types of regularized estimating methods (e.g., SCAD and MCP) may be adopted in our proposed procedure. The lasso estimator is obtained by maximizing the following penalized log-likelihood function with respect to the regression parameters β, subject to a normalizing constant,

$$PL(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) \overset{\text{def}}{=} \tfrac{1}{n}\mathcal{L}_n(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) - \lambda\|\boldsymbol{\beta}\|_1 \propto \tfrac{1}{n}\textstyle\sum_{i=1}^n \left\{y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - b(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}/\phi - \lambda\|\boldsymbol{\beta}\|_1,$$

where λ is a nonnegative tuning parameter, and $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$ is the ℓ_1-norm of the regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$. Let $\hat{\boldsymbol{\beta}}_\lambda = \arg\max_\beta PL(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X})$ be a lasso estimator of $\boldsymbol{\beta}$ at a given tuning parameter $\lambda \geq 0$. Solving for $\hat{\boldsymbol{\beta}}_\lambda$ may be done by the coordinate descent algorithm via *Donoho and Johnstone* (1994)'s soft-thresholding approach, with the tuning parameter being determined by, *say*, multi-fold cross-validation (*Shao and Deng*, 2012), as is done by us using the R package `glmnet` with λ selected by 10-fold cross-validation.

3.2.2 Confidence Density for Bias-Corrected Lasso Estimator

To combine multiple lasso estimators obtained from separate sub-datasets, we need to overcome the issue of misalignment: the sets of selected covariates with non-zero estimates in the model are different across sub-datasets. Our solution is based on bias-corrected lasso estimators. The bias correction enables us not only to obtain non-zero estimates of all regression coefficients, but also, more importantly, to establish the distribution of regularized estimators. The latter is critical for us to utilize the confidence distribution to combine estimators. Denote the score function by $\boldsymbol{S}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{y_i - g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}\boldsymbol{x}_i/\phi$. It is known that the lasso estimator, $\hat{\boldsymbol{\beta}}_\lambda$, should satisfy the following Karush-Kuhn-Tucker condition:

$$\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda) - \lambda\hat{\boldsymbol{\kappa}} = 0, \tag{3.1}$$

where subdifferentials $\hat{\boldsymbol{\kappa}} = (\hat{\kappa}_1, \cdots, \hat{\kappa}_p)^T$ satisfy $\max_j |\hat{\kappa}_j| \leq 1$, and $\hat{\kappa}_j = \text{sign}(\hat{\beta}_{\lambda,j})$ if $\hat{\beta}_{\lambda,j} \neq 0$. The first-order Taylor expansion of $\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda)$ in (3.1) at the true value $\boldsymbol{\beta}_0$ leads to $-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0) + \lambda\hat{\boldsymbol{\kappa}} \approx \boldsymbol{S}_n(\boldsymbol{\beta}_0)$. It follows that $\hat{\boldsymbol{\beta}}_\lambda^c - \boldsymbol{\beta}_0 \approx \{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}^{-1}\boldsymbol{S}_n(\boldsymbol{\beta}_0)$, where $\hat{\boldsymbol{\beta}}_\lambda^c$ is a bias-corrected lasso estimator (*van de Geer et al.*, 2014) as follows,

$$\hat{\boldsymbol{\beta}}_\lambda^c \overset{\text{def}}{=} \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}^{-1}\lambda\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}^{-1}\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda). \tag{3.2}$$

The second equality in (3.2) follows directly from (3.1) and the sensitivity matrix $\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T / \left\{\phi\dot{g}(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}$, which is assumed to be negative-definitive. We show later in Theorem III.4 that under some regularity conditions, this bias-corrected estimator $\hat{\boldsymbol{\beta}}_\lambda^c$ is asymptotically normally distributed, namely

$$n^{1/2}(\hat{\boldsymbol{\beta}}_\lambda^c - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)), \text{ as } n \to \infty, \tag{3.3}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = [E\{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}]^{-1}$. Based on the asymptotic normality in (3.3), following *Xie and Singh* (2013), we form the asymptotic confidence density of $\boldsymbol{\beta}_0$ as $\hat{h}_n(\boldsymbol{\beta}_0) \propto \exp[-\frac{n}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_\lambda^c)^T \{\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\}^{-1} (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_\lambda^c)]$. Replacing $\boldsymbol{\beta}_0$ in the bias-correction term in (3.2) and the asymptotic variance $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0)$ in (3.3) by the sparse lasso estimator $\hat{\boldsymbol{\beta}}_\lambda$, we obtain

$$\hat{\boldsymbol{\beta}}_\lambda^c = \hat{\boldsymbol{\beta}}_\lambda + \hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda), \tag{3.4}$$

where the estimated variance covariance matrix is $\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_\lambda) = \{-\dot{\boldsymbol{S}}_n(\hat{\boldsymbol{\beta}}_\lambda)\}^{-1}$. Moreover, a "data-driven" version of the asymptotic confidence density is given by

$$\hat{h}_n(\boldsymbol{\beta}_0) \propto \exp\left[-\tfrac{n}{2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_\lambda^c)^T\{\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_\lambda)\}^{-1}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_\lambda^c)\right], \tag{3.5}$$

It is worth pointing out that this bias-corrected estimator in (3.4) is equivalent to a one-step Newton-Raphson updated estimator of the lasso estimator. In the generalized linear models framework, we have

$$\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_\lambda) = \{\tfrac{1}{n\phi}\boldsymbol{X}^T\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{X}\}^{-1}, \tag{3.6}$$

and the resulting confidence density may be expressed as

$$\hat{h}_n(\boldsymbol{\beta}_0) \propto \exp[-\tfrac{1}{2\phi}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^c_\lambda)^T\{\boldsymbol{X}^T\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{X}\}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^c_\lambda)], \tag{3.7}$$

where $\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda)$ is the diagonal weight matrix based on the variance function of a generalized linear model. In the case where the dispersion parameter ϕ is unknown, such as in the Gaussian linear regression, we use a root-n consistent estimator $\hat{\phi} = (n - |\hat{\boldsymbol{\beta}}_\lambda|_0)^{-1}\sum_{i=1}^n d(y_i, \hat{\mu}_i(\hat{\boldsymbol{\beta}}_\lambda))$, where $|\boldsymbol{x}|_0$ is the number of non-zero entries of vector $\boldsymbol{x}$, and $d(\cdot,\cdot)$ is the unit deviance function; refer to *Song* (2007) for details.

3.2.3 Examples

Example III.1. Gaussian linear model. Assume y_i follows a normal distribution with mean $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, variance ϕ and the canonical link function $g(x) = x$. The score function takes the form $\boldsymbol{S}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left\{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right\}\boldsymbol{x}_i/\phi$. From (3.4) and (3.7), we obtain the confidence density function $\hat{h}_n(\boldsymbol{\beta}_0)$ with the bias-corrected lasso estimator as $\hat{\boldsymbol{\beta}}^c_\lambda = \hat{\boldsymbol{\beta}}_\lambda + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda),$ and $\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda) = \boldsymbol{I}_n$.

Example III.2. Binomial logistic model. Assume y_i follows a Bernoulli distribution with probability of success π_i and the logit link function $g(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Similarly, from (3.4), the bias-corrected lasso estimator is given by $\hat{\boldsymbol{\beta}}^c_\lambda = \hat{\boldsymbol{\beta}}_\lambda + \{\boldsymbol{X}^T\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{X}\}^{-1}\boldsymbol{X}^T(\boldsymbol{Y} - \hat{\boldsymbol{\pi}})$, where $\hat{\pi}_i = \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)/\{1 + \exp(\boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)\}$ and $\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda) = \mathrm{diag}(\hat{v}_1, \ldots, \hat{v}_n)$ with $\hat{v}_i = \hat{\pi}_i(1 - \hat{\pi}_i)$.

Example III.3. Poisson log-linear model. Assume y_i follows a Poisson distribution with mean μ_i. The canonical link function is $g(\mu_i) = \log(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$. Equation (3.4) gives the bias-corrected lasso estimator of the form $\hat{\boldsymbol{\beta}}^c_\lambda = \hat{\boldsymbol{\beta}}_\lambda + \{\boldsymbol{X}^T\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{X}\}^{-1}\boldsymbol{X}^T$ $\{\boldsymbol{Y} - \hat{\boldsymbol{\mu}}\}$, where $\hat{\mu}_i = \hat{v}_i = \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_\lambda)$ and $\boldsymbol{P}_n(\hat{\boldsymbol{\beta}}_\lambda) = \text{diag}(\hat{v}_1, \ldots, \hat{v}_n)$.

3.2.4 Regularity Conditions

We list the regularity conditions used throughout this chapter. For convenience, we include subscript k to denote that a quantity is obtained from the k-th sub-dataset. In this section, we can simply omit k because we derive the asymptotic distribution for just one sub-dataset. In fact, these conditions are general, and are used in proofs of asympotic normality of estimators of individual sub-datasets as well as the final combined estimator based on full data.

(C1) For $k = 1, \ldots, K$, assume the same underlying true parameters $\boldsymbol{\beta}_{0,k} = \boldsymbol{\beta}_0$ across all sub-datasets. Let the score function satisfies $E[\{y_1 - g^{-1}(\boldsymbol{x}_1^T\boldsymbol{\beta}_0)\}\boldsymbol{x}_1/\phi] = 0$. Further, in a neighborhood around the true value $\boldsymbol{\beta}_0$, $\mathcal{N}_\delta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1 < \delta\}$ for some constant $\delta > 0$, it holds that for any $\boldsymbol{\beta} \in \mathcal{N}_\delta(\boldsymbol{\beta}_0)$, $|\{\dot{g}(\boldsymbol{x}^T\boldsymbol{\beta})\}^{-1}| > \phi$ for some $\phi > 0$ and that $\sup_{\boldsymbol{x}} |\{\dot{g}(\boldsymbol{x}^T\boldsymbol{\beta})\}^{-1}| = O(1)$. For any $\boldsymbol{\beta} \in \mathcal{N}_\delta(\boldsymbol{\beta}_0)$ and $\boldsymbol{x}$, $g(\boldsymbol{x}^T\boldsymbol{\beta})$ is continuous and twice differentiable, which is satisfied by the expotential family (*Song*, 2007).

(C2) Assume $\max_k \|\boldsymbol{X}_k\|_\infty = O_p(1)$, where $\|\boldsymbol{X}_k\|_\infty = \max_{i,j}|X_{k,ij}|$. Let $\underline{\Sigma}(\boldsymbol{M})$ and $\overline{\Sigma}(\boldsymbol{M})$ be the minimum and maximum singular values of a matrix $\boldsymbol{M}$, respectively. Assume $b \le \min_k\{\underline{\boldsymbol{\Sigma}}(n_k^{-1/2}\boldsymbol{X}_k)\} \le \max_k\{\overline{\boldsymbol{\Sigma}}(n_k^{-1/2}\boldsymbol{X}_k)\} \le B$, where b and B are two positive constants.

(C3) For some $\psi_0 > 0$, and for all $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta}_{S_0^c}\|_1 \le 3\|\boldsymbol{\beta}_{S_0}\|_1$, it holds that $\|\boldsymbol{\beta}_{S_0}\|_1^2 \le (\boldsymbol{\beta}^T\boldsymbol{X}_k^T\boldsymbol{X}_k\boldsymbol{\beta})s_0/n\psi_0^2$, for $k = 1, \ldots, K$, where s_0 is the number of true signals in $\boldsymbol{\beta}_0$. In addition, for $k = 1, \cdots, K$ and any p such that $0 < p < n_{min}$ with $n_{min} = \min_k n_k$, assume $\lambda_k = O\{(\log p/n_k)^{1/2}\}$ and $s_0 = o_p(n_{min}^{1/2}/\log p)$.

It is noteworthy that conditions (C1) and (C2) are two common regularity conditions; see for example *Liu et al.* (2015). Condition (C3) is the compatibility condition required by the lasso estimator $\hat{\boldsymbol{\beta}}_{\lambda,k}$, which is the same as condition (C2) given in *van de Geer et al.* (2014) for the asymptotic normality. Following Theorem 3 in *Zhang and Huang* (2008), applying conditions (C1) and (C2), we can show that the (adaptive) lasso estimator $\hat{\boldsymbol{\beta}}_{\lambda,k}$ satisfies condition (C3).

3.2.5 Large Sample Property

Theorem III.4. *Under conditions (C1)-(C3) in Section 3.2.4, the estimator* $\hat{\boldsymbol{\beta}}_{\lambda}^{c}$ *in* (3.4) *is consistent and asymptotically normally distributed, namely,* $n^{1/2}(\hat{\boldsymbol{\beta}}_{\lambda}^{c} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0))$, *as* $n \to \infty$, *where* $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}_0) = E\{\boldsymbol{x}_1\boldsymbol{x}_1^T/\phi\dot{g}(\boldsymbol{x}_1^T\boldsymbol{\beta}_0)\}$ *is the Fisher information matrix.*

Theorem III.4 establishes the consistency and asymptotic normality of the proposed estimator $\hat{\boldsymbol{\beta}}_{\lambda}^{c}$. Its proof is given in Appendix D. This theorem can be view as an extension of the element-wise asymptotic result in *van de Geer et al.* (2014) to the joint distribution of $\hat{\boldsymbol{\beta}}_{\lambda}^{c}$, but with an additional restriction that $p < n$. We emphasize on the joint distribution of $\hat{\boldsymbol{\beta}}_{\lambda}^{c}$ because it is necessary for the combination step using confidence distributions, to be described in Section 3.3. To the best of our knowledge, most existing work that allows $p \gg n$ only provides element-wise inference, or provides joint distribution on a subset of q covariates, where $p < n$, for example *van de Geer et al.* (2014) and *Javanmard and Montanari* (2014). From Theorem III.4, we construct the confidence density as expressed in (3.5). Theorem III.4 can be extended to the case when $p \to \infty$, which is presented as Theorem S1, along with its proof, available in the supplementary material.

Remark III.5. The procedure based on Theorem III.4 for the construction of the confidence density remains valid when the adaptive lasso estimator (*Zou*, 2006) is used to replace $\hat{\boldsymbol{\beta}}_{\lambda}$ in (3.4) and (3.7). An adaptive lasso estimator is obtained by $\check{\boldsymbol{\beta}}_{\lambda} = \arg\max_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}\{y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - b(\boldsymbol{x}_i^T\boldsymbol{\beta})\}/\phi - \lambda\sum_{j=1}^{p}\hat{w}_j|\boldsymbol{\beta}_{\boldsymbol{j}}|$, where the weights $\{\hat{w}_j\}_{j=1}^{p}$ are given by $\hat{w}_j = (|\hat{\beta}_j^{ini}|)^{-\gamma}$, with an initial root-$n$ consistent estimate $\hat{\boldsymbol{\beta}}^{ini}$ of $\boldsymbol{\beta}$ and some suitable constant $\gamma > 0$, which is typically set at 1.

Remark III.6. The collinearity problem is often encountered in high-dimensional data analysis where some of the covariates are highly correlated. One solution is to construct the confidence distribution in (3.7) by utilizing the Karush-Kuhn-Tucker condition of the elastic net estimator (*Zou and Hastie*, 2005). Another quick remedy is to apply ridge-type estimator to stabilize the matrix inverse and improve numerical stability through adding a ridge term $\tau \boldsymbol{I}_p$, $\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0) + \tau \boldsymbol{I}_p$, where $\tau > 0$, to $\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)$.

3.3 Combined Estimation and Inference

We now turn to the combined estimation and related inferences. We consider the full data of size N, which is randomly partitioned into K batches, each with size n_k, and $N = \sum_{k=1}^{K} n_k$. In reality, when we face extraordinarily large data where a direct analysis on the whole data is numerically impossible, we apply the strategy of divide-and-combine proposed by computer scientists. To proceed, we randomly partition the entire dataset into K sub-datasets, $\{(\boldsymbol{Y}_k, \boldsymbol{X}_k)\}_{k=1}^{K}$, each of which has n_k observations, namely $\boldsymbol{Y}_k$ is an $n_k \times 1$ vector and X_k is an $n_k \times p$ matrix. Here, K is not necessarily fixed. The choice of K in practice will be discussed in Section 3.6.

If there existed a "god-made" computer with unlimited computational capacity, all existing methods available in various statistical software could be applied directly to analyze the entire data regardless of the sample size and the resulting estimator. This estimator is denoted by $\hat{\boldsymbol{\beta}}_{full}$, which serves as the gold standard, and obtained by: $\hat{\boldsymbol{\beta}}_{full} = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_N(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = \arg\max_{\boldsymbol{\beta}} \sum_{k=1}^{K} \mathcal{L}_{n_k}(\boldsymbol{\beta}; \boldsymbol{Y}_k, \boldsymbol{X}_k)$, where $\mathcal{L}_N(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X})$ and $\mathcal{L}_{n_k}(\boldsymbol{\beta}; \boldsymbol{Y}_k, \boldsymbol{X}_k)$ are the log-likelihood functions of the full data $(\boldsymbol{Y}, \boldsymbol{X})$ and the k-th sub-dataset $(\boldsymbol{Y}_k, \boldsymbol{X}_k)$, respectively. There are many ways to combine results obtained from sub-datasets, *say*, $\hat{\boldsymbol{\beta}}_k = \arg\max_{\boldsymbol{\beta}} \mathcal{L}_{n_k}(\boldsymbol{\beta}; \boldsymbol{Y}_k, \boldsymbol{X}_k)$, $k = 1, \ldots, K$, in here, we consider using the confidence distribution due to its generalizability under unified objective functions and its ease to establish statistical inferences. For each sub-dataset $(\boldsymbol{Y}_k, \boldsymbol{X}_k)$, we first apply Theorem III.4 to construct the asymptotic confidence density $\hat{h}_{n_k}(\boldsymbol{\beta}_0)$, $k = 1, \ldots, K$. Then, in the same spirit to *Liu et al.* (2015), we may combine the K confidence densities to derive a combined estimator of $\boldsymbol{\beta}_0$. The combined estimator is denoted by $\hat{\boldsymbol{\beta}}_{dac}$, where *dac* refers to divide and combine, according to the following procedure:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{dac} &= \arg\max_{\boldsymbol{\beta}} \log \textstyle\prod_{k=1}^{K} \hat{h}_{n_k}(\boldsymbol{\beta}) \\ &= \arg\max_{\boldsymbol{\beta}} \textstyle\sum_{k=1}^{K} \frac{n_k}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^c_{\lambda_k,k})^T \{\hat{\boldsymbol{\Sigma}}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k})\}^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^c_{\lambda_k,k}), \end{aligned} \tag{3.8}$$

where $\hat{\boldsymbol{\beta}}^c_{\lambda_k,k}$ and $\hat{\boldsymbol{\Sigma}}_{n_k}(\beta)$ are estimates given in (3.4) and (3.6), respectively, with respect to the k-th sub-dataset $(\boldsymbol{Y}_k, \boldsymbol{X}_k)$. The key advantage of the confidence distribution approach is to allow us derive an inference procedure for the combined estimator $\hat{\boldsymbol{\beta}}_{dac}$, as stated in Theorem III.7. The key result established in Theorem III.7 is that the confidence density estimator $\hat{\boldsymbol{\beta}}_{dac}$ and the gold estimator $\hat{\boldsymbol{\beta}}_{full}$ are asymptotically equally efficient.

Theorem III.7. *Let $n_{min} = \min_k n_k$ and $K = O(N^{1/2-\delta})$ with constant $\delta \in (0, 1/2)$. Under conditions (C1)-(C3) stated in Section 3.2.4, the divide-and-combine estimator $\hat{\boldsymbol{\beta}}_{dac}$ obtained from (3.8) is consistent and asymptotically normally distributed, namely, $N^{1/2}(\hat{\boldsymbol{\beta}}_{dac} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma}_{dac}(\boldsymbol{\beta}_0))$ as $n_{min} \to \infty$. $\boldsymbol{\Sigma}^{-1}_{dac}(\boldsymbol{\beta}_0) = E\{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}$ is the Fisher information matrix of the full data when $K = 1$. That is, the estimator $\hat{\boldsymbol{\beta}}_{dac}$ is asymptotically as efficient as the gold estimator $\hat{\boldsymbol{\beta}}_{full}$.*

The proof of Theorem III.7 is given in Appendix D. It is worth noting that the conditions for the divide-and-combine estimator $\hat{\boldsymbol{\beta}}_{dac}$ is the same as those required for the regularized estimator in each sub-dataset, as long as the number of sub-datasets, K, is fixed. This is because in the procedure of constructing confidence densities, when the asymptotic normal distribution is used, conditions in the derivation of asymptotic distributions for the combined estimator are automatically satisfied. By some simple algebra, the solution to the optimization problem in (3.8), i.e., the proposed divide-and-combine estimator $\hat{\boldsymbol{\beta}}_{dac}$, can be expressed explicitly as a form of weighted average of $\hat{\boldsymbol{\beta}}^{c}_{\lambda,k}, k = 1, \ldots, K$, as

$$\hat{\boldsymbol{\beta}}_{dac} = \{\textstyle\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k})\}^{-1}\{\textstyle\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k})\hat{\boldsymbol{\beta}}^{c}_{\lambda_k,k}\} \tag{3.9}$$

where $\hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k}) = \frac{1}{n_k\hat{\phi}_k}\boldsymbol{X}_k^T\boldsymbol{P}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k})\boldsymbol{X}_k$. The practical implication of Theorem III.7 is that as long as the sample size of each sub-dataset is not small, the proposed $\hat{\boldsymbol{\beta}}_{dac}$ will have little loss of estimation efficiency, while enjoys fast computing and better numerical stability in the analysis of big data. It is interesting to note that in (3.9) matrix inverse is not required for the computation of $\hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k})$ since its inverse is the Fisher information matrix of the k-th sub-dataset. The only computation of matrix inversion is for the sum of the Fisher information matrices. The variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{dac}$ can be estimated by $\hat{\boldsymbol{\Sigma}}_{dac} = \{\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda_k,k})\}^{-1}$, from which confidence intervals can be derived.

Remark III.8. Note that when $\lambda = 0$, our proposed estimator $\hat{\boldsymbol{\beta}}_{dac}$ in (3.9) reduces to the meta estimator $\hat{\boldsymbol{\beta}}_{meta} = \{\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_k)\}^{-1}\{\sum_{k=1}^{K} n_k \hat{\boldsymbol{\Sigma}}^{-1}_{n_k}(\hat{\boldsymbol{\beta}}_k)\hat{\boldsymbol{\beta}}_k\}$, where $\hat{\boldsymbol{\beta}}_k$ is the estimates of effect sizes, for $k = 1, \ldots, K$. *Lin and Xi* (2011) found a similar result as a special case of the aggregated estimating equation estimator under the maximum likelihood estimation framework. However, the aggregated estimating equation estimator requires a strong assumption of $K = O(n^r_{min})$ $(r < 1/3)$, and it does not consider regularized estimation. Thus, it is a simpler framework to combine estimates that are not shrunk for the purpose of variable selection. In addition, regardless of different inputted estimators, the proposed estimator $\hat{\boldsymbol{\beta}}_{dac}$ and $\hat{\boldsymbol{\beta}}_{meta}$ take the same form for the combination of estimators. However, they are derived from different criteria with different purposes. Specifically, $\hat{\boldsymbol{\beta}}_{meta}$ aims at improving statistical power via weighted average, while $\hat{\boldsymbol{\beta}}_{dac}$ is obtained by minimizing the combined confidence densities for the interest of statistical inference theory. The flexibility of the confidence density approach allows to incorporate additional features in the combination; for example, the homogeneity may be relaxed by imposing a mixture of normals in (3.8), which is not the meta estimation method.

Remark III.9. The majority voting estimator proposed by *Chen and Xie* (2014) to combine lasso estimates from multiple sub-datasets is given as follows: $\hat{\boldsymbol{\beta}}_{mv} = \boldsymbol{A}\{\sum_{k=1}^{K} n_k \boldsymbol{A}^T \dot{\boldsymbol{S}}_{n_k}(\hat{\boldsymbol{\beta}}_k)\boldsymbol{A}\}^{-1}\{\sum_{k=1}^{K} n_k \boldsymbol{A}^T \dot{\boldsymbol{S}}_{n_k}(\hat{\boldsymbol{\beta}}_k)\boldsymbol{A}\hat{\boldsymbol{\beta}}_{k,\hat{\boldsymbol{A}}^{(v)}}\}$, where $\boldsymbol{A}$ is a $p \times |\hat{\boldsymbol{A}}^{(v)}|$ subsetting matrix corresponding to a majority voting set $\hat{\boldsymbol{A}}^{(v)} = \{j : \sum_{k=1}^{K} I(\hat{\boldsymbol{\beta}}_{k,j} \neq 0) > w\}$, which is a set of signals with votes higher than the prespecified threshold value $w \in [0, K)$, and $\hat{\boldsymbol{\beta}}_{k,\hat{\boldsymbol{A}}^{(v)}}$ denotes a corresponding sub-vector of $\hat{\boldsymbol{\beta}}_k$. The majority voting estimator $\hat{\boldsymbol{\beta}}_{mv}$ has been shown to have the oracle property (*Zou*, 2006) asymptotically, which however is not applicable to statistical inference in the sense given by *Fan and Li* (2001) that the truly non-zero coefficients are never known beforehand. Thus the oracle distribution cannot be used for inference on the entire coefficient vector.

3.4 Simulation Studies

In this section, we conduct extensive simulation experiments to demonstrate the numerical performance of the proposed method under Gaussian, logistic and Poisson regressions. Specifically, we compare across three divide-and-combine methods, including the meta-analysis method, the majority voting method (*Chen and Xie*, 2014), and our proposed method based on results of Theorem III.7. Note that when $K = 1$, under no data partitioning, meta-analysis is equivalent to generalized linear regression, the majority voting method is equivalent to lasso regression (*Tibshirani*, 1996), and our method is equivalent to lasso with post-selection inference from Theorem III.4 (*van de Geer et al.*, 2014).

All methods are compared thoroughly on the performance of variable selection, statistical inference and computation time. The evaluation metrics for variable selection include the sensitivity and specificity of correctly identifying non-zero coefficients. The evaluation metrics for statistical inference include mean squared error, absolute bias, coverage probability and asymptotic standard error of signal set $\mathcal{A}_0 = \{j : \beta_{0,j} \neq 0\}$ and non-signal set $\mathcal{A}_0^c = \{j : \beta_{0,j} = 0\}$, respectively, where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^\top$ denotes the vector of true coefficients. Coverage probability and standard error are not reported for the majority voting method since it does not provide inference. We use results from the conventional generalized linear regression estimator, $\hat{\boldsymbol{\beta}}_{full}$, as our golden standard during comparisons. In order to show the best variable selection results of the majority voting method, we carefully select ω in $\hat{\boldsymbol{\beta}}_{mv}$ such that the sum of sensitivity and specificity is maximized. The computation time of all methods includes the time of reading data from disks to memory and the time of numerical calculation. Under the divide-and-combine setting when $K > 1$, computation time is reported as the sum of the maximum time used among parallelized jobs and the time used to combine results. All simulation experiments are conducted on a standard Linux cluster with 16 GB of random-access memory per CPU.

Table 3.1 presents the simulation results from a moderate size dataset with $N = 50,000$ and $p = 300$ so that methods without data partition can be repeated in multiple rounds of simulations within a reasonable amount of time. Clearly, this is a typical regression data setting with $p \ll N$. We consider Gaussian, logistic and Poisson models, with responses generated from the mean model $g^{-1}\{E(Y_i)\} = \sum_{j=1}^{p} \beta_j X_j,\ i = 1, \ldots, N$, and covariates $\{X_j\}_{j=1}^{p}$ generated from the multivariate normal distribution with marginal mean of zero and variance of one, and with a compound symmetric covariance structure with correlation $\rho = 0.8$, a simulation setting similar to that provided by *van de Geer et al.* (2014). We report scenarios when the full dataset is randomly divide into $K = 25$ and 100 subsets of equal sizes, each with sample size $n_k = 2,000$ and $n_k = 500$, respectively. Results when $K = 1$, including those from the golden standard, are also reported. We randomly select $s_0 = 10$ coefficients from $\boldsymbol{\beta}_0$ to be set at non-zero. The non-zero coefficients are set at 0.3 for Gaussian models, 0.3 for logistic models, and 0.1 for Poisson models. Labels META, VOTING and MODAC are used to denote the meta-analysis method, the majority voting method and our method, respectively, whereas GLM, LASSO and LASSOINF correspond to META, VOTING and MODAC, respectively, when $K = 1$. The GLM column serves as the benchmark of all comparisons. Results are averaged across 500 replications.

In the results of Gaussian linear model in Table 3.1, reassuringly, all methods perform as good as the golden standard method. META and MODAC exhibit identical performances as that of GLM regardless of the choices of split K. Because under the Gaussian model, solutions to META and MODAC are exact. In this case, divide-and-combine methods gain significant computation time reduction without sacrificing statistical accuracy. Among all methods, VOTING has the highest sensitivity and specificity when $\omega = 12$ for $K = 25$ and $\omega = 50$ for $K = 100$. This shows the improvement of selection consistency when using divide-and-combine over LASSO, as well as other methods based on inference.

However, the merit of providing exact solutions for divide-and-combine methods under the Gaussian model does not carry forward to generalized linear models. It is worthwhile to note that although p is much smaller than N, data partition may result in p closer to n_k for each sub-dataset of size n_k. Regularization has been found to be an appealing step in this situation to reduce the dimension of the optimization so to achieve more stable numerical performance. The regularization is recommended to handle the situation where the Newton-Raphson iterative algorithm is needed in the search for the estimate, because the Hessian matrix may be poorly estimated with p being large and close to n_k. Especially, in the results of the logistic model

Table 3.1: Simulation results, summarized from 500 replications, under the setting of $N = 50,000$ and $p = 300$ for Gaussian, logistic and Poisson models. Methods with different K are compared. GLM denotes the conventional generalized linear regression method; META denotes the conventional meta-analysis method; LASSO denotes the conventional lasso method; VOTING denotes the majority voting method; LASSOINF denotes lasso method with inference; and MODAC denotes the proposed divide-and-combine method.

Gaussian Model

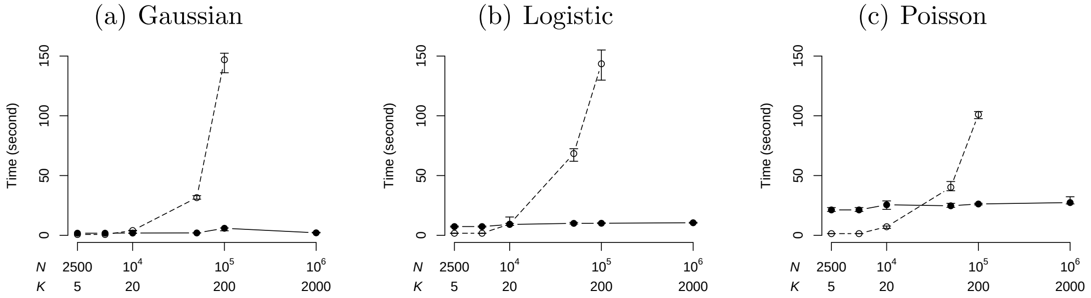
	GLM	META	META	LASSO	VOTING	VOTING	LASSOINF	MODAC	MODAC
	$(K = 1)$	$(K = 25)$	$(K = 100)$	$(K = 1)$	$(K = 25)$	$(K = 100)$	$(K = 1)$	$(K = 25)$	$(K = 100)$
					$(\omega = 12)$	$(\omega = 50)$			
Sensitivity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Specificity	0.95	0.95	0.95	0.91	1.00	1.00	0.95	0.95	0.95
100× MSE of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.01	0.01	0.01	0.03	0.01	0.01	0.01	0.01	0.01
100× MSE of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01
Absolute bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Absolute bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01
Cov. prob. of $\boldsymbol{\beta}_{\mathcal{A}_0}$	0.95	0.95	0.95	—	—	—	0.95	0.95	0.95
Cov. prob. of $\boldsymbol{\beta}_{\mathcal{A}_0^c}$	0.95	0.95	0.95	—	—	—	0.95	0.95	0.95
Asymp. st. err. of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.01	0.01	0.01	—	—	—	0.01	0.01	0.01
Asymp. st. err. of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.01	0.01	0.01	—	—	—	0.01	0.01	0.01
Computation time	34.85	0.62	0.20	31.50	2.16	2.08	36.61	2.28	2.14

Logistic Model

	GLM	META	META	LASSO	VOTING	VOTING	LASSOINF	MODAC	MODAC
	$(K = 1)$	$(K = 25)$	$(K = 100)$	$(K = 1)$	$(K = 25)$	$(K = 100)$	$(K = 1)$	$(K = 25)$	$(K = 100)$
					$(\omega = 7)$	$(\omega = 20)$			
Sensitivity	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
Specificity	0.95	1.00	1.00	0.89	1.00	1.00	0.95	0.95	0.96
100× MSE of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.08	0.57	189.38	0.23	0.20	0.29	0.08	0.09	0.10
100× MSE of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.08	0.05	4.15	0.00	0.00	0.00	0.08	0.08	0.07
Absolute bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.02	0.07	1.36	0.04	0.04	0.05	0.02	0.02	0.02
Absolute bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.02	0.02	0.16	0.00	0.00	0.00	0.02	0.02	0.02
Cov. prob. of $\boldsymbol{\beta}_{\mathcal{A}_0}$	0.95	0.36	1.00	—	—	—	0.95	0.94	0.92
Cov. prob. of $\boldsymbol{\beta}_{\mathcal{A}_0^c}$	0.95	1.00	1.00	—	—	—	0.95	0.95	0.96
Asymp. st. err. of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.03	0.03	1895.12	—	—	—	0.03	0.03	0.03
Asymp. st. err. of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.03	0.03	1893.23	—	—	—	0.03	0.03	0.03
Computation time	66.01	1.63	1.40	260.48	15.78	10.42	266.09	15.92	10.53

Poisson Model

	GLM	META	META	LASSO	VOTING	VOTING	LASSOINF	MODAC	MODAC
	$(K = 1)$	$(K = 25)$	$(K = 100)$	$(K = 1)$	$(K = 25)$	$(K = 100)$	$(K = 1)$	$(K = 25)$	$(K = 100)$
					$(\omega = 7)$	$(\omega = 26)$			
Sensitivity	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Specificity	0.95	0.94	0.91	0.91	1.00	1.00	0.95	0.95	0.95
100× MSE of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.70	0.80	0.90	1.70	0.80	0.50	0.70	0.70	0.70
100× MSE of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.70	0.70	0.80	0.00	0.10	0.00	0.70	0.70	0.70
Absolute bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01
Absolute bias of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.01	0.01	0.01	0.00	0.00	0.00	0.01	0.01	0.01
Cov. prob. of $\boldsymbol{\beta}_{\mathcal{A}_0}$	0.95	0.93	0.90	—	—	—	0.95	0.95	0.95
Cov. prob. of $\boldsymbol{\beta}_{\mathcal{A}_0^c}$	0.95	0.94	0.91	—	—	—	0.95	0.95	0.95
Asymp. st. err. of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$	0.01	0.01	0.01	—	—	—	0.01	0.01	0.01
Asymp. st. err. of $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0^c}$	0.01	0.01	0.01	—	—	—	0.01	0.01	0.01
Computation time	42.26	1.46	0.40	132.06	26.57	25.00	136.85	26.67	25.08

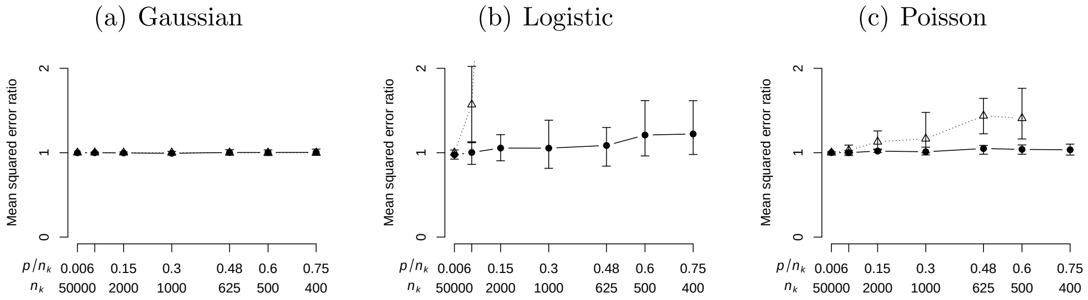


(a) Gaussian (b) Logistic (c) Poisson

Figure 3.1: Median computation time and interquartile range for conventional generalized linear regression (open dots) and our proposed method of divide-and-combine (solid dots) as N increases. The sample size of each sub-dataset n_k in our method is fixed at 500 by increasing K. Conventional generalized linear regression fails when $N = 10^6$ due to memory limitation.

presented in Table 3.1, META appears to be highly unstable within the numerical computation of each sub-dataset in both cases when $K = 25$ and $K = 100$. As K increases, the logistic regression in sub-datasets overestimates the bias and standard error, which are then carried over to the final META estimate. Specifically, the estimated mean of response $\hat{\pi}_i$ may get very close to the boundaries of 0 or 1, and consequently the estimated variance $\hat{\pi}_i(1 - \hat{\pi}_i)$ is close to 0, causing severe problems to the matrix inverse. On the other hand, the proposed MODAC exhibits robust performance similar to that of GLM. The bias of VOTING for $\mathcal{A}_0^c$ is higher than that of GLM as expected due to the ℓ_1 penalty.

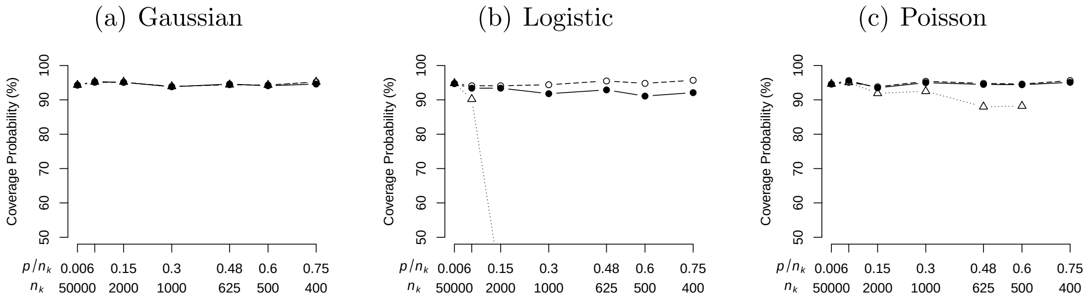
In the Poisson model section of Table 3.1, similar to our observation in the Gaussian and logistic models, MODAC again gives the most stable results among all divide-and-combine methods. On the other hand, META gives deviated coverage probability than the nominal rate of 95% as well as poorer selection accuracy than GLM. Although VOTING still gives the best variable selection results with ω carefully chosen, in practice, the choice of ω remains a challenging task when true signals are fully unknown. Additional simulation results are provided as supplementary material to show that the variable selection outcome is sensitive to the choice of ω in VOTING.


(a) Gaussian

(b) Logistic

(c) Poisson

Figure 3.2: The y-axis measures the ratio of mean squared error over that of the conventional generalized linear regression, for regression coefficients in set $\mathcal{A}_0$. Median and interqualtile range of the ratio for meta-analysis (triangles) and our proposed method of divide-and-combine (solid dots) are shown as the ratio p/n_k increases. We fix N at $50,000$ and p at 300. Conventional meta-analysis algorithm fails to converge for logistic and Poisson regressions when p/n_k is large.


(a) Gaussian

(b) Logistic

(c) Poisson

Figure 3.3: Coverage probability of regression coefficients in set $\mathcal{A}_0$ for conventional generalized linear regression (open dots), conventional meta-analysis (triangles) and our proposed method of divide-and-combine (solid dots) as the ratio of p and n_k increases. The total sample size N and number of covariates p are fixed at $50,000$ and 300, respectively, for all cases. Conventional meta-analysis algorithm fails to converge for logistic regression when $p/n_k \geq 0.3$.

In summary from Table 3.1, results under different generalize linear models are in favor of the invocation of regularization to achieve consistent and stable mean and variance estimation in the application of data partition to handle big data. We see that MODAC is the most stable method that produces the most comparable results to those of the golden standard, and is unaffected by the partition size K. In contrast, the performance of META and VOTING varies depending on K. A noticeable advantage of MODAC is that it requires less computation time than GLM due to the virtue of scalability. Despite the fact that META is the fastest as it does not involve the step of tuning parameter selection, its results are clearly unstable in both the logistic and Poisson models. Although VOTING provides better variable selection than MODAC, its results are sensitively dependent on the choice of the voting threshold ω, which may often be hard to determine in practice.

Under the same model settings as those in Tables 3.1, we conduct additional simulation experiments to explore the change of some important metrics in relation to the total sample size N, the number of division K and the sample size of sub-datasets n_k. We present the results in Figures 3.1-3.3, each based on 100 replication. Fig. 3.1 shows a comparison of computation time between MODAC and GLM as N increases, while holding n_k in MODAC fixed at 500. We also fix p at 300. We see that the computational burden increases sharply for GLM as N increases, whereas the computation time for MODAC remains almost the same in all three types of models due to its scalability. Computation time for GLM when $N = 10^6$ is not reported because the computation exceeds the maximum memory limit allowed on the Linux cluster. Fig. 3.2 shows the ratio of mean squared error over the benchmark value from GLM for $\hat{\boldsymbol{\beta}}_{\mathcal{A}_0}$ as K increases, while fixing $N = 50,000$ and $p = 300$, comparing between MODAC and META. We see that the mean squared error of MODAC is stable against the change of the ratio p/n_k. In contrast, the mean squared error of META quickly deviates from the mean squared error of GLM for both logistic and

Poisson models as p/n_k increases. In Fig. 3.3, we compare the coverage probability of $\boldsymbol{\beta}_{\mathcal{A}_0}$ between GLM, META and MODAC. The 95% confidence interval coverage probability by MODAC remains close to the nominal level, whereas the coverage probability of META deviates from 95% when p/n_k goes toward one, especially in the case of the logistic model.

3.5 Data Example

We illustrate a statistical inference application of our method using a publicly available dataset from the National Highway and National Automotive Sampling System (NASS) Crashworthiness Data System (CDS) between the years of 2009 and 2015. The NASS-CDS data contains detailed information of about 5,000 crashes each year sampled across the US. The response variable of interest is injury severity, which is dichotomized as 1 if a crash leads to suffer moderate or severer injury, and 0 if minor or no injury. Awaring of the high dependency of outcomes from the same vehicle, we only include drivers in our study. The full data consists of $N = 37,535$ samples, and 48 covariates. The dataset was randomly partitioned and stored as $K = 50$ sub-datasets, each with sample size of about 750. Table 3.2 shows the estimated coefficients, standard errors and p-values of candidate risk factors from logitc regressions based on GLM, META and MODAC. The computation time using MODAC is 0.66 second, one half of the time required by GLM, which is 1.17 seconds. MODAC produces consistent inference about the risk factors as that of GLM. Although META is super fast and finishes in 0.03 second, the inference results deviate from those of GLM and MODAC. Specifically, as inferred by the golden standard GLM method and MODAC, African American are less likely to suffer from modeterate to servere injury given in a car accident than White, and accidents are more likely to result in minor injuries on a Wednesday than a Sunday. On the other hand, META is not able to capture these two effects as its estimated bias and variance

Table 3.2: Estimation and inference results of association study between potential risk factors and binary injury outcome. Logistic model is fitted using the conventional generalized linear regression method (GLM), the meta-analysis method (META), and our proposed method (MODAC). Run time is presented in square brackets next to the method names.

	GLM (1.17s)			META (0.03s)			MODAC (0.62s)		
Variable	Estimate	Std. Error	*p*-value	Estimate	Std. Error	*p*-value	Estimate	Std. Error	*p*-value
AGE	0.08	0.01	0.00	0.08	0.01	0.00	0.08	0.01	0.00
OTHERPASS	-0.17	0.03	0.00	-0.16	0.03	0.00	-0.17	0.03	0.00
BELOW14	-0.31	0.06	0.00	-0.27	0.06	0.00	-0.26	0.05	0.00
FEMALE	-0.08	0.03	0.01	-0.08	0.03	0.02	-0.08	0.03	0.01
WEIGHT	0.10	0.01	0.00	0.09	0.01	0.00	0.10	0.01	0.00
HEIGHT	-0.09	0.02	0.00	-0.08	0.02	0.00	-0.09	0.02	0.00
PARUSE	-1.07	0.03	0.00	-1.00	0.03	0.00	-1.05	0.03	0.00
LANES	0.03	0.01	0.03	0.03	0.01	0.04	0.03	0.01	0.03
SPLIMIT	0.01	0.01	0.65	0.00	0.01	0.81	0.00	0.01	0.72
VEHAGE	0.01	0.01	0.43	0.01	0.01	0.40	0.01	0.01	0.40
DRINKING	0.00	0.04	0.90	0.01	0.05	0.78	0.00	0.04	0.90
DRGINV	0.03	0.04	0.51	0.03	0.05	0.49	0.03	0.04	0.54
HISPLAT	0.12	0.04	0.00	0.11	0.04	0.00	0.11	0.04	0.00
CURBWGT	-0.02	0.02	0.30	-0.01	0.02	0.48	-0.02	0.02	0.34
SURCOND	0.00	0.05	0.98	0.03	0.05	0.60	0.00	0.05	0.98
PREVACC	-0.11	0.03	0.00	-0.10	0.03	0.00	-0.10	0.03	0.00
FOURWHDR	0.01	0.04	0.69	0.02	0.04	0.67	0.01	0.04	0.70
OCCRACEblack	-0.07	0.03	0.03	-0.06	0.03	0.07	-0.07	0.03	0.03
OCCRACEasian	-0.08	0.07	0.23	-0.01	0.07	0.83	-0.08	0.07	0.23
CLIMATE21	-0.02	0.06	0.77	-0.03	0.06	0.58	-0.02	0.06	0.77
REGION2Mid_Atlantic	-0.16	0.04	0.00	-0.15	0.04	0.00	-0.15	0.04	0.00
REGION2Northeast	-0.07	0.06	0.22	-0.04	0.06	0.52	-0.07	0.06	0.23
REGION2Northwest	0.27	0.05	0.00	0.26	0.05	0.00	0.28	0.05	0.00
REGION2South	-0.29	0.05	0.00	-0.26	0.05	0.00	-0.27	0.04	0.00
REGION2Southeast	-0.29	0.06	0.00	-0.25	0.06	0.00	-0.26	0.06	0.00
REGION2Southwest	-0.13	0.04	0.00	-0.12	0.04	0.00	-0.12	0.04	0.00
LGTCOND2Dark	0.05	0.05	0.24	0.07	0.05	0.16	0.05	0.04	0.26
LGTCOND2DawnDusk	-0.02	0.06	0.76	0.03	0.07	0.71	-0.02	0.06	0.76
LGTCOND2Dk_Lighted	-0.03	0.03	0.33	-0.02	0.03	0.45	-0.03	0.03	0.33
MONTH2Fall	0.01	0.04	0.83	0.00	0.04	0.93	0.00	0.03	0.88
MONTH2Spring	0.12	0.03	0.00	0.11	0.04	0.00	0.11	0.03	0.00
MONTH2Winter	0.03	0.04	0.34	0.03	0.04	0.46	0.03	0.04	0.37
VEHTYPE2Truck	-0.05	0.04	0.19	-0.05	0.04	0.21	-0.05	0.04	0.18
TRAFFLOW2D_No_Bar	0.02	0.04	0.64	0.01	0.04	0.73	0.01	0.04	0.71
TRAFFLOW2No_Divid	-0.02	0.04	0.63	-0.03	0.04	0.49	-0.02	0.04	0.53
TRAFFLOW2One_Way	-0.19	0.06	0.00	-0.16	0.06	0.01	-0.17	0.06	0.00
DAYWEEK2Fri	-0.15	0.04	0.00	-0.15	0.04	0.00	-0.15	0.04	0.00
DAYWEEK2Mon	-0.21	0.05	0.00	-0.19	0.05	0.00	-0.21	0.04	0.00
DAYWEEK2Sat	-0.19	0.04	0.00	-0.18	0.04	0.00	-0.18	0.04	0.00
DAYWEEK2Thu	-0.17	0.04	0.00	-0.17	0.05	0.00	-0.17	0.04	0.00
DAYWEEK2Tue	-0.22	0.05	0.00	-0.21	0.05	0.00	-0.21	0.04	0.00
DAYWEEK2Wed	-0.09	0.04	0.03	-0.09	0.05	0.06	-0.09	0.04	0.03
YEAR2010	-0.06	0.04	0.15	-0.05	0.04	0.26	-0.05	0.04	0.18
YEAR2011	0.01	0.04	0.78	0.01	0.04	0.83	0.01	0.04	0.81
YEAR2012	0.11	0.04	0.01	0.11	0.04	0.01	0.10	0.04	0.01
YEAR2013	0.08	0.04	0.08	0.07	0.04	0.09	0.07	0.04	0.09
YEAR2014	0.04	0.05	0.32	0.06	0.05	0.22	0.04	0.04	0.34
YEAR2015	0.14	0.05	0.00	0.15	0.05	0.00	0.14	0.05	0.00

are inflated in sub-datasets as we have seen in Table 3.1.

3.6 Discussion

In this chapter, we proposed a scalable regression method in the context of generalized linear models with reliable statistical inference through the seminal work of confidence distribution. An earlier version of this work has been made available online (*Tang et al.*, 2016). Although the divide-and-combine idea has been widely adapted in practice to solve computational challenges arising from the analysis of big data, statistical inference has been little investigated and the conventional meta-analysis method has been taken for granted. We found in this chapter that regularization in the estimator is very appealing in the context of generalized linear models, especially in the logistic regression, because clearly this regularization enables to effectively increase the robustness of scalable regression analysis. Furthermore, the reliable statistical inference gives rise to great practical usefulness of the divide-and-combine strategy compared to many selection-only methodologies. Our method can be readily built-in into some of the most popular open source parallel computing libraries, such as MapReduce (*Dean and Ghemawat*, 2008) and Spark (*Zaharia et al.*, 2010). Source code to execute the proposed divide-and-combine method on distributed Hadoop clusters is made available as map and reduce functions, with additional instructions, available for download at http://www.umich.edu/∼songlab/software.html#MODAC. The work in this chapter has been extended to fit the ordinal logistic model by a simple data augmentation step for the application of ranking problems; see more details in Appendix F.

Under the small data situation, divide-and-combine may not be needed because it might slow down the computation. Nevertheless, it is always preferable to impose regularization to robustify the solution, as we see that numerical results may be unstable when n is close to p, specifically, for logistic and Poisson models. In other words, our method when n is small and with $K = 1$ and is still more robust than the maximum likelihood estimator, but requires some additional computation. Based on our simulation experiences, the partition size K should be chosen to ensure that n_{min} is reasonably larger than p, *say* $p/n_{min} \leq 0.5$. For example, if the number of available computer nodes is c, and if each node is capable of processing N/c amount of the samples all at once, we suggest setting $K = c$ so that all data partitions can be processed at once. If sample size N/c is beyond the capacity of a single node, then we select $K = mc$ where m is the smallest positive integer possible, such that each node will sequentially process m data partitions, each with size N/K.

CHAPTER IV

Homogeneity Pursuit in Pattern-Mixture Models by Penalized Generalize Estimating Equations

4.1 Introduction

The method of generalized estimating equations (GEE) by *Liang and Zeger* (1986) is widely used in many statistical problems, for example, to perform parameter estimation and inference in correlated data analysis (*Zeger et al.*, 1988; *Lipsitz et al.*, 1997; *Song*, 2007). It has been one of the standard methods of choice due to its minimal model specification of the first two moments, computational simplicity, and estimation consistency albeit misspecification of correlations structure. This chapter is motivated by a prospective cohort study in which an extension of the classic GEE method is proposed to handle nonignorable missing data in the framework of pattern-mixture models. We start by introducing the background of our motivating study data.

The Intern Health Study (IHS) is an NIH funded longitudinal cohort study that assesses stress and mood in medical interns at institutions around the US (*Sen et al.*, 2010). This study is motivated by the status quo that physicians are 2-3 times more likely to die by suicide (*Schernhammer and Colditz*, 2004), and the level of suicidal ideation is elevated in medical students and residents, and appears to increase with

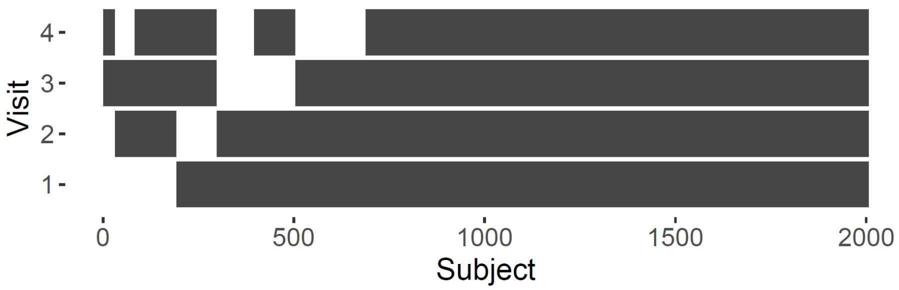


Figure 4.1: Response availability across the four longitudinal visits for participants in the IHS data set.

the onset of training (*Rotenstein et al.*, 2016). The overarching aim of IHS is to understand the factors involved in stress and depression among interns in order to foster a healthier, more educational environment for interns and safer care for the patients that they treat. Each year, IHS enrolls over 3,000 new interns into the study. Participating interns are recruited to the study before the start of their training for a baseline assessment. Then for every three months into their medical internship, data on mental well-being and other risk factors are collected through a mobile smart phone application. Specifically, participants will complete an initial 20-minute survey at baseline, and a short 5-minute follow-up survey at each of the four longitudinal visits. In this chapter, we use data collected between the years of 2012 and 2014 from IHS on participants who have responded to at least two consecutive screenings. This results in a pool of over 2,000 qualified subjects. Over 30% of the participants under consideration have at least one nonrespondent visit. Figure 4.1 shows response availability of subjects across all visits, sorted by data availability patterns. Is can be seen that missing data are pervasive, and none of the previous studies and publications on IHS data have systematically investigate the modeling bias caused by missing data. Thus, in this chapter, we introduce a new statistical methodology to systematically handle nonresponse missing data with flexible assumptions.

For longitudinal studies similar to IHS where it is often difficult to record the full response data for everyone and missing data are pervasive, statistical analyses should take into consideration the missing mechanism in order to avoid model estimation bias. Two missing data mechanisms commonly assumed in the statistical literature are missing completely at random (MCAR) and missing at random (MAR) (*Rubin*, 1976; *Little and Rubin*, 1987), or in other words, missing data are *ignorable* under the likelihood estimation and inference – the process of missing data is governed by observed data. Beyond the likelihood methodology, as noted in *Liang and Zeger* (1986), as a quasi-likelihood estimation and inference approach, GEE will yield biased estimators if missing data are not MCAR. To overcome, in the case of MAR missing data, inverse probability weighted GEE (*Robins et al.*, 1995) can be used to eliminate the systematic bias due to missingness. Nevertheless, MCAR and MAR may not be valid, and MAR assumption is generally not testable. There is very limited progress in the recent literature on the handling of *nonignorable* missing data, the case that the missing data mechanism is dependent on missing data themselves. The primary focus of this chapter is to develop a new method of longitudinal data analysis in the nonignorable missing data framework of pattern-mixture models.

Little (1993, 1994) as well as *Ekholm and Skinner* (1998) have considered pattern-mixture models in which incomplete data are stratified by patterns of missing values, and as a result, distinct models are specified within each missing data pattern stratum. Testing for MCAR hypothesis (*Chen and Little*, 1999; *Qu and Song*, 2002) may be used to determine if stratification is needed. If so, stratification has been an effective strategy to accommodate inhomogeneous missing patterns resulted from nonignorable missingness. A technical issue pertaining to the application of pattern-mixture modeling approach is over-stratification; that is, excessive stratification is imposed in the analysis. One consequence of this over-stratification is to unnecessarily increase the variance of estimates when missing data are indeed MCAR (*Chen and Little*, 1999). To systematically deal with this difficulty, we propose a penalized GEE method that enables us to fuse some "similar" stratum-specific parameters to reduce the number of strata but still achieve adequate stratification. The rationale of our approach is that similar strata should undergo merging so that resulting individual strata will end up with larger sample sizes.

Our approach is developed in the framework of penalized GEE. The penalized estimating equations is first studied by *Fu* (2003) to address collinearity issue, with the use of the bridge penalty. *Johnson et al.* (2008) extends the smoothness requirement by *Fu* (2003) to a more general discrete case with the consideration of variable selection, and establishes the oracle properties for a family of convex and nonconvex penalties. Later, *Wang et al.* (2012) considers multivariate correlated responses in that some key asymptotic oracle properties for variable selection are shown in the case of the number of covariates diverges. We adopt the penalized GEE with fused lasso penalty for the purpose of fusing similar strata.

The rest of this chapter is organized as follows. Section 4.2 revisits GEE in the setting of pattern-mixture models. Section 4.3 introduces our method of GEE with the fused penalties. Section 4.4 discusses an efficient algorithm for implementation. Section 4.5 presents some key asymptotic properties for the proposed estimator. We demonstrate our method with simulation experiments in Section 4.6 and apply it to the motivating IHS data in Section 4.7. Finally, we conclude in Section 4.8 with some discussion on the generalization of our method.

4.2 Pattern-Mixture Approach

We begin with some notation for GEE with missing values. Consider a longitudinal study of M designed visits for each of N individuals. If no observation missing, the design matrix of the ith individual is denoted as $\boldsymbol{X}_i^T = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iM})$, a $p \times M$ matrix, where $\boldsymbol{x}_{ij}$ is a p-element covariate vector measured at visit j, $j = 1, \ldots, M$. Similarly, the longitudinal response of subject i is $\boldsymbol{Y}_i^T = (Y_{i1}, \ldots, Y_{iM})$. Denote the first two marginal moments of Y_{ij} by $\mu_{ij} = E(Y_{ij})$ and $\sigma_{ij}^2 = \text{var}(Y_{ij})$. We assume in this chapter that the marginal density of Y_{ij} is in the family of exponential dispersion models (*Jorgensen*, 1997), where the mean μ_{ij} follows a generalized linear model, $g(\mu_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}$, and variance σ_{ij}^2 is given by $\sigma_{ij}^2 = \phi v(\mu_{ij})$, where $v(\cdot)$ is the unit variance function. In this chapter, we consider canonical link function g; that is, $\dot{g}(\mu) = v^{-1}(\mu)$. We also assume that outcome-covariate pairs $(Y_{ij}, \boldsymbol{x}_{ij})$ are simultaneously missing or observed, as is the case in the IHS study. Due to dropouts or intermittent missing visits, we may observe R distinct missing patterns. Stratify N subjects by the R missing patterns, each stratum has sample size n_k, $k = 1, \ldots, R$, and $N = \sum_{k=1}^R n_k$. Subjects in the kth pattern are observed at the set of visits $\mathcal{L}_k = \{j : \ell_{kj} = 1, j = 1, \ldots, M\}$ where $\ell_{kj} = 1$ if visit j is observed and 0 otherwise.

If missing observations are MCAR, according to *Liang and Zeger* (1986), the GEE estimator for $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_p)$ is the solution to the set of equations based on the observed cases, $\sum_{i=1}^N \frac{\partial \boldsymbol{\mu}_i^{obs}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}(\boldsymbol{V}_i^{obs})^{-1}(\boldsymbol{Y}_i^{obs} - \boldsymbol{\mu}_i^{obs}(\boldsymbol{\beta})) = \boldsymbol{0}$. We use subscript $\mathcal{L}_k$ to denote a subvector or submatrix corresponding to the indices of observed visits. Equivalently, the GEE can be written as $\sum_{k=1}^R \sum_{i=1}^{n_k} \frac{\partial \boldsymbol{\mu}_{ki,\mathcal{L}_k}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \boldsymbol{V}_{ki,\mathcal{L}_k}^{-1}(\boldsymbol{Y}_{ki,\mathcal{L}_k} - \boldsymbol{\mu}_{ki,\mathcal{L}_k}(\boldsymbol{\beta})) = \boldsymbol{0}$, where $\frac{\partial \boldsymbol{\mu}_{ki,\mathcal{L}_k}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{X}_{ki,\mathcal{L}_k}^T \boldsymbol{A}_{ki,\mathcal{L}_k}(\boldsymbol{\beta})$, and $\boldsymbol{V}_{ki,\mathcal{L}_k} = \boldsymbol{A}_{ki,\mathcal{L}_k}^{1/2}(\boldsymbol{\beta})\boldsymbol{R}_{\mathcal{L}_k}(\boldsymbol{\tau})\boldsymbol{A}_{ki,\mathcal{L}_k}^{1/2}(\boldsymbol{\beta})$ with $\boldsymbol{A}_{ki}(\boldsymbol{\beta}) = \mathrm{diag}(\sigma_{ki1}^2(\boldsymbol{\beta}), \ldots, \sigma_{kim}^2(\boldsymbol{\beta}))$ and $\boldsymbol{R}(\boldsymbol{\tau})$ a working correlation matrix whose structure is prespecified with correlation parameter $\boldsymbol{\tau}$. For notation simplicity, we suppress $\mathcal{L}_k$ in the remaining discussion. Then the estimating equations are expressed as

$$\sum_{k=1}^{R} \sum_{i=1}^{n_k} \boldsymbol{S}_{ki}(\boldsymbol{\beta}) = \boldsymbol{0} \tag{4.1}$$

where

$$\boldsymbol{S}_{ki}(\boldsymbol{\beta}) = \boldsymbol{X}_{ki}^T \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}) \boldsymbol{R}_k^{-1}(\boldsymbol{\tau}) \boldsymbol{A}_{ki}^{-1/2}(\boldsymbol{\beta})(\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta})).$$

Equation (4.1) is a GEE under the MCAR mechanism based on all available data.

When the MCAR assumption is violated, GEE is invalid due to biased sampling (*Song*, 2007). A popular method of choice to deal with nonignorable missingness is the pattern-mixture model proposed by *Little* (1993), in which the joint distribution of outcomes and covariates is assumed to be dependent on the missing patterns. Following the stratification principle in the pattern-mixture models, *Chen and Little* (1999) proposes to solve the following pattern-stratified GEE:

$$\sum_{k=1}^{R}\sum_{i=1}^{n_k}\boldsymbol{S}_{ki}(\boldsymbol{\beta}_k)=\mathbf{0}, \tag{4.2}$$

where regression coefficient vector $\boldsymbol{\beta}_k$ becomes stratum (or pattern) dependent. Equation (4.2) will be solved for $\boldsymbol{\beta}_k$ within each stratum: $\sum_{i=1}^{n_k}\boldsymbol{S}_{ki}(\boldsymbol{\beta}_k)=\mathbf{0}$, leading to stratified estimators for respective data patterns. A Wald-type test proposed by *Chen and Little* (1999) may be used to test the MCAR assumption and guide the decision between MCAR GEE (4.1) and pattern-mixture GEE (4.2). If the MCAR holds, there exists $\boldsymbol{\beta}_*$ such that $E\{\boldsymbol{S}_k(\boldsymbol{\beta}_*)\}=\mathbf{0}$ for all R patterns, and such $\boldsymbol{\beta}_*$ may be regarded as the common true value across all strata. On the other hand, let $\hat{\boldsymbol{\beta}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ be the GEE estimator of $\boldsymbol{\beta}_k$ and estimated sandwich variance of $\hat{\boldsymbol{\beta}}_k$ from stratum k. A meta estimator is given by $\hat{\boldsymbol{\beta}}_c=(\sum_{k=1}^{R}\boldsymbol{\Sigma}_k^{-1})^{-1}\sum_{k=1}^{R}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k$. Then, by *Chen and Little* (1999), a test statistics for the null hypothesis H_0: exist a $\boldsymbol{\beta}$ such that $E\{\boldsymbol{S}_k(\boldsymbol{\beta})\}=\mathbf{0}$ $(k=1,\ldots,R)$ is

$$d=\sum_{k=1}^{R}(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_c)^T\boldsymbol{\Sigma}_k^{-1}(\hat{\boldsymbol{\beta}}_k-\hat{\boldsymbol{\beta}}_c), \tag{4.3}$$

which is showed to follow asymptotically a χ_v^2 distribution with $v=R(p-1)$. Other tests for MCAR are also developed in the literature (see for examples *Diggle* (1989); *Qu and Song* (2002); *Qu et al.* (2011)).

4.3 Fusion Learning with Nonignorable Missing Values

In this section we develop a *fusion learning* approach to merging similar missing data patterns in longitudinal studies. This new method provides a generalization of the existing pattern-mixture models introduced in Section 4.2 in the sense that stratification may be relaxed to acquire a better GEE estimator. The proposed method takes initial estimates from individual GEE models stratified by the missing patterns and proceeds to fuse similar estimates across missing pattern strata.

Instead of estimating one common set of coefficients $(\beta_1, \ldots, \beta_p)$ as typically done under MCAR, we begin with estimates of R distinct sets, each set for one missing pattern. We denote the stratified parameters as $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T) = (\beta_{11}, \ldots, \beta_{1R}, \ldots, \beta_{p1}, \ldots, \beta_{pR})$, where β_{lk} is the regression coefficient for covariate l in stratum k. Different from the current approach to solving R separate GEEs in (4.2), we propose to solve these pR parameters jointly by the following penalized GEE:

$$\boldsymbol{U}(\boldsymbol{\beta}) = R^{-1} \begin{pmatrix} n_1^{-1}\boldsymbol{S}_1(\boldsymbol{E}_1\boldsymbol{\beta}) \\ \vdots \\ n_R^{-1}\boldsymbol{S}_R(\boldsymbol{E}_R\boldsymbol{\beta}) \end{pmatrix} - \boldsymbol{q}_\lambda(|\boldsymbol{D}\boldsymbol{\beta}|)\,\mathrm{sign}(\boldsymbol{D}\boldsymbol{\beta}) = \boldsymbol{0}, \tag{4.4}$$

where $\boldsymbol{S}_k(\boldsymbol{E}_k\boldsymbol{\beta}) = \sum_{i=1}^{n_k} \boldsymbol{S}_{ki}(\boldsymbol{E}_k\boldsymbol{\beta})$ is the GEE of the kth stratum, $\boldsymbol{E}_k\boldsymbol{\beta} = \boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{pk})^T$. Here $\boldsymbol{E}_k^T = (\boldsymbol{e}_k, \boldsymbol{e}_{R+k}, \ldots, \boldsymbol{e}_{(p-1)R+k})$ with $\boldsymbol{e}_i$ the ith unit vector of length pR. The penalty term $\boldsymbol{q}_\lambda(\boldsymbol{\beta})^T = (q_\lambda(\beta_{11}), \ldots, q_\lambda(\beta_{pR}))$ is a pR-dimensional function with $q_\lambda(x)$ defined on $x > 0$ and $\mathrm{sign}(\boldsymbol{\beta})^T = (\mathrm{sign}(\beta_{11}), \ldots, \mathrm{sign}(\beta_{pR}))$ is a pR-dimensional element-wise sign function. Equation (4.4) is similar to the penalized estimating equations studied in *Fu* (2003), *Johnson et al.* (2008) and *Wang et al.* (2012). Deviating from the focus of variable selection in the previous work, here we consider identifying homogeneous clustering structures of parameters in $\boldsymbol{\beta}$, via $\boldsymbol{D}\boldsymbol{\beta}$, where matrix $\boldsymbol{D}$ sets up contrasts between differences of elements in $\boldsymbol{\beta}$ to achieve fusion. This is the same idea of fused lasso (*Tibshirani et al.*, 2005), and the form of

$\boldsymbol{D}$ will be specified in Section 4.4. Stratum-specific weights $n_k^{-1}, k = 1, \ldots, R$, ensure all missing patterns are weighted equally so that estimates from smaller strata would not be discounted in comparison to those from larger strata. The tuning parameter λ is some nonnegative number that determines the weight of the penalty.

In the estimating equation (4.4), function q_λ is specified as subdifferentials of certain penalty functions. Some candidates for $q_\lambda(\cdot)$ include:

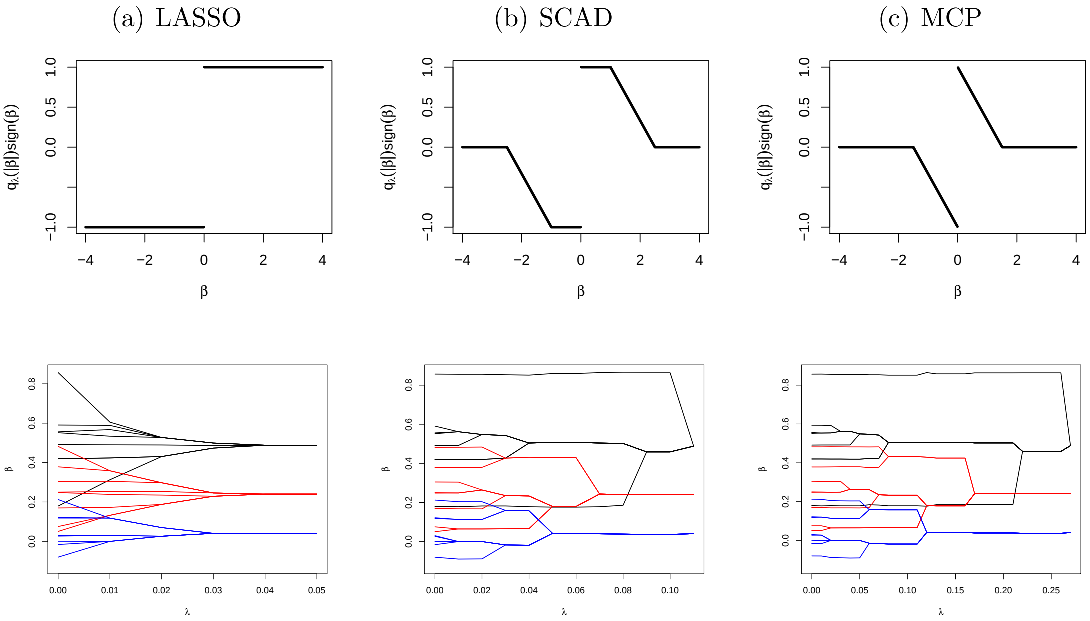
(i) Least absolute shrinkage and selection operator (LASSO) penalty (*Tibshirani*, 1996): $q_\lambda(x) = \lambda$, $x > 0$;

(ii) Smoothly clipped absolute deviation (SCAD) penalty (*Fan and Li*, 2001): $q_\lambda(x) = \lambda\{I(x \leq \lambda) + \frac{(a\lambda - x)_+}{(a-1)\lambda} I(x > \lambda)\}$, $x > 0$, $a > 2$;

(iii) Minimax concave penalty (MCP) (*Zhang*, 2010): $q_\lambda(x) = \lambda \frac{(a\lambda - x)_+}{a\lambda}$, $x > 0$, $a > 1$.

Although the LASSO penalty is convex with a guaranteed global optimal solution and computational ease, the resulting regularized solution tends to overshrink large coefficients and to produce too many subgroups (*Ma and Huang*, 2017). On the other hand, both SCAD and MCP penalties are nonconvex functions in that shrinkage tapers off for large coefficients. In this chapter, we choose the MCP type $q_\lambda(\cdot)$ function as it tends to provide more distinctive clustering in its solution paths than SCAD. This feature can also be seen in Figure 4.2, with plots of the different penalty functions and their respective solutions paths for fusion learning. Both SCAD and MPC penalties taper off quickly for large β values, leading to hierarchical clustering-like solution paths.

Since the function $\boldsymbol{q}_\lambda$ is discontinuous at $\mathbf{0}$, an exact solution to (4.4) might not exist. A merit of discontinuity is to create sparsity in the solution, namely, some roots are exactly zero. In this chapter, we define the solution $\hat{\boldsymbol{\beta}}$ to (4.4) as an approximate solution such that $\boldsymbol{U}(\hat{\boldsymbol{\beta}}) = o(\boldsymbol{a}_n)$ for a sequence $\boldsymbol{a}_n \to \mathbf{0}$ (*Wang et al.*, 2012).


(a) LASSO

(b) SCAD

(c) MCP

Figure 4.2: Plots of penalty functions (top) and solution paths (bottom) of fusion learning with LASSO, SCAD and MCP, respectively.

4.4 Penalized GEE Estimation

In this section, we present an efficient algorithm to obtain a solution for $\boldsymbol{\beta}$ to the penalized GEE in (4.4). We begin with specifying the form of contrast matrix $\boldsymbol{D}$. Let $\boldsymbol{D} = \text{block-diag}(\boldsymbol{D}_1, \ldots, \boldsymbol{D}_l, \ldots, \boldsymbol{D}_p)$ be an $Rp \times Rp$ dimension block-diagonal matrix for specification of contrasts between stratum-specific coefficients for individual covariates, $\boldsymbol{\beta}_l$, $l = 1, \ldots, p$. Following *Wang et al.* (2016), *Ke et al.* (2015) and *Tang and Song* (2016), we utilize the information of parameter ordering in the formulation of matrix $\boldsymbol{D}_l$ to remove redundant penalties. For an example of four missing patterns, $R = 4$, utilizing the known ordering of $\boldsymbol{\beta}_l$ for covariate l, say, $\beta_{l1} < \beta_{l4} < \beta_{l3} < \beta_{l2}$, we come up with matrix $\boldsymbol{D}_l$ of the following form:

$$\boldsymbol{D}_l = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix}.$$

The first row of $\boldsymbol{D}_l$ is the reference parameter that has the smallest amount in this group, which may be shrunk toward zero via variable selection. Clearly, $\boldsymbol{D}_l$ is of full rank, thus invertible.

Note that the stratum-specific weights n_k^{-1} in (4.4) may be absorbed into $\boldsymbol{S}_k$ as subject weights, that is,

$$\boldsymbol{S}_k(\boldsymbol{\beta}_k) = \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{\beta}_k) \boldsymbol{R}_k^{-1}(\boldsymbol{\tau}) \boldsymbol{w}_{ki} \boldsymbol{A}_{ki}^{-1/2}(\boldsymbol{\beta}_k)(\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_{ki}(\boldsymbol{\beta}_k)),$$

where $\boldsymbol{w}_{ki} = \bar{n}/n_k \boldsymbol{I}$ and $\bar{n} = \sum_{k=1}^{R} n_k/R$. As a result, solving (4.4) is equivalent to solving the following

$$\boldsymbol{S}(\boldsymbol{\beta}) - N\boldsymbol{q}_\lambda(|\boldsymbol{D}\boldsymbol{\beta}|)\,\mathrm{sign}(\boldsymbol{D}\boldsymbol{\beta}) = \boldsymbol{0}, \tag{4.5}$$

where $\boldsymbol{S}^T(\boldsymbol{\beta}) = (\boldsymbol{S}_1^T(\boldsymbol{\beta}_1), \ldots, \boldsymbol{S}_R^T(\boldsymbol{\beta}_R))$, with $\boldsymbol{\beta}_k = \boldsymbol{E}_k\boldsymbol{\beta}$. Denote $\boldsymbol{\theta} = \boldsymbol{D}\boldsymbol{\beta}$. Since $\boldsymbol{D}$ is invertible, we write the left hand side of (4.5) as a function of $\boldsymbol{\theta}$, and solve the following equation:

$$\boldsymbol{U}^{\boldsymbol{D}}(\boldsymbol{\theta}) = \boldsymbol{S}^{\boldsymbol{D}}(\boldsymbol{\theta}) - N\boldsymbol{q}_\lambda(|\boldsymbol{\theta}|)\,\mathrm{sign}(\boldsymbol{\theta}) = \boldsymbol{0}, \tag{4.6}$$

where $\boldsymbol{S}^{\boldsymbol{D}}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{D}^{-1}\boldsymbol{\theta}) = (\boldsymbol{S}_1^T(\boldsymbol{E}_1\boldsymbol{D}^{-1}\boldsymbol{\theta}), \ldots, \boldsymbol{S}_R^T(\boldsymbol{E}_R\boldsymbol{D}^{-1}\boldsymbol{\theta}))^T$. Consequently, $\boldsymbol{\beta} = \boldsymbol{D}^{-1}\boldsymbol{\theta}$, or $\boldsymbol{\beta}_l = \boldsymbol{D}_l^{-1}\boldsymbol{\theta}_l$, $l = 1, \ldots, p$, one-to-one correspondence between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. In fact, $\boldsymbol{U}^{\boldsymbol{D}}(\boldsymbol{\theta})$ gives a penalized GEE whose solution $\hat{\boldsymbol{\theta}}$ may be efficiently obtained by an iterative algorithm (*Wang et al.*, 2012). Subsequently, we obtain solution $\hat{\boldsymbol{\beta}}$ for (4.5) by transformation, $\hat{\boldsymbol{\beta}} = \boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}$.

Following *Wang et al.* (2012), we calculate estimate $\hat{\boldsymbol{\theta}}$ in (4.6) using a Newton-Raphson iterative algorithm for the GEE in combination with the minorization-maximization algorithm for nonconvex penalty (*Hunter and Li*, 2005). To proceed, for a small $\epsilon > 0$, we obtain the penalized GEE estimate $\hat{\boldsymbol{\theta}}^T$ for $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_p^T) = (\theta_{11}, \dots, \theta_{1R}, \dots, \theta_{p1}, \dots, \theta_{pR})$, with $\boldsymbol{\theta}_l = \boldsymbol{D}_l\boldsymbol{\beta}_l$, $l = 1, \dots, p$, as the solution that approximately satisfies

$$S_{kl}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta}) - Nq_\lambda(|\hat{\theta}_{lk}|)\,\mathrm{sign}(\hat{\theta}_{lk})\frac{|\hat{\theta}_{lk}|}{\epsilon + |\hat{\theta}_{lk}|},\ l = 1, \dots, p,\ k = 1, \dots, R,$$

where $S_{kl}(\cdot)$ denotes the l-th element of $\boldsymbol{S}_k(\cdot)$. The algorithm alternately updates $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\tau}}$ and $\hat{\phi}$. The updating step for $\hat{\boldsymbol{\theta}}$, at iteration b, is

$$\hat{\boldsymbol{\theta}}^b = \hat{\boldsymbol{\theta}}^{b-1} + \{\boldsymbol{H}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}^{b-1}) + N\boldsymbol{J}(\hat{\boldsymbol{\theta}}^{b-1})\}^{-1}\{\boldsymbol{S}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}^{b-1}) - N\boldsymbol{J}(\hat{\boldsymbol{\theta}}^{b-1})\hat{\boldsymbol{\theta}}^{b-1}\}$$

where $\boldsymbol{H}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}^{b-1}) = \mathrm{diag}\left\{\boldsymbol{H}_1(\boldsymbol{E}_1\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}^{b-1}), \dots, \boldsymbol{H}_R(\boldsymbol{E}_R\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}^{b-1})\right\}$ with

$$\boldsymbol{H}_k(\boldsymbol{E}_k\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}^{b-1}) = \sum_{i=1}^{n_k}\boldsymbol{X}_{ki}^T\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}^{b-1})\boldsymbol{R}_k^{-1}(\hat{\boldsymbol{\tau}}^{b-1})\boldsymbol{w}_{ki}\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}^{b-1})\boldsymbol{X}_{ki},$$

and

$$\boldsymbol{J}(\hat{\boldsymbol{\theta}}^{b-1}) = \mathrm{diag}\left\{\frac{q_\lambda(|\hat{\theta}_{11}^{b-1}|+)}{\epsilon + |\hat{\theta}_{11}^{b-1}|}, \dots, \frac{q_\lambda(|\hat{\theta}_{1R}^{b-1}|+)}{\epsilon + |\hat{\theta}_{1R}^{b-1}|}, \dots, \frac{q_\lambda(|\hat{\theta}_{p1}^{b-1}|+)}{\epsilon + |\hat{\theta}_{p1}^{b-1}|}, \dots, \frac{q_\lambda(|\hat{\theta}_{pR}^{b-1}|+)}{\epsilon + |\hat{\theta}_{pR}^{b-1}|}\right\}.$$

Both correlation and dispersion parameters $\hat{\boldsymbol{\tau}}$ and $\hat{\phi}$ can be estimated by the method of moments as suggested in the standard GEE, e.g., *Liang and Zeger* (1986), once $\hat{\boldsymbol{\beta}}^b$ is obtained by $\hat{\boldsymbol{\beta}}^b = \boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}}^b$. In this chapter, we choose $a = 1.5$ in the MCP penalty and $\epsilon = 10^{-6}$ in the above approximation as similar to *Wang et al.* (2012). It is worth noting that the developed algorithm is applicable to penalized GEE with other nonconvex penalties despite our choice of MCP in this chapter.

Since the true ordering of coefficients across missing patterns is unknown in practice, we propose to use the estimated ordering by the method of ranks. This method has been studied by *Wang et al.* (2016) and *Tang and Song* (2016). Thus, the contrast matrix $\boldsymbol{D}$ is in fact data-dependent, which we denote as $\boldsymbol{D}_{n_k}$. However, for simplicity, we suppress the index notation except in Section 4.5. We select the tuning value of λ based on the extended regularized information criterion (ERIC) (*Hui et al.*, 2015) of the following form

$$\text{ERIC}(\lambda) = \boldsymbol{S}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}_\lambda)^T \boldsymbol{C}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}_\lambda)^{-1} \boldsymbol{S}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}_\lambda) + 2v \log(N/\lambda) \sum_{j=1}^{p} \text{df}(\hat{\boldsymbol{\theta}}_{j\lambda})$$

where

$$\boldsymbol{C}^{\boldsymbol{D}}(\hat{\boldsymbol{\theta}}) = \text{block-diag}\{\boldsymbol{S}_1(\boldsymbol{E}_1\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}})\boldsymbol{S}_1(\boldsymbol{E}_1\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}})^T, \ldots, \boldsymbol{S}_R(\boldsymbol{E}_R\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}})\boldsymbol{S}_R(\boldsymbol{E}_R\boldsymbol{D}^{-1}\hat{\boldsymbol{\theta}})^T\},$$

v is some positive constant, and $\text{df}(\hat{\boldsymbol{\theta}}_j)$ denotes the number of nonzero values in $\hat{\boldsymbol{\theta}}_j$. The subscript λ in $\hat{\boldsymbol{\theta}}_\lambda$ is used to indicate its dependence on λ. The additional tuning parameter v provides flexibility to control the severity of penalization. According to *Hui et al.* (2015), a v smaller than 0.5 is better suited to high-dimensional data, thus we set $v = 0.4$ throughout this chapter. Starting from $\lambda = 0$, we fit a path of solutions $\hat{\boldsymbol{\theta}}_\lambda$ for a sequence of $\lambda \geq 0$. To accelerate computation, for the next value λ, we employ the warm-start technique and use $\hat{\boldsymbol{\theta}}$ from the current value of λ as the initial value of the iterative algorithm.

4.5 Theoretical Results

In this section, we consider the asymptotic properties for the estimator of the fusion GEE when the total number of visits M and the number of missing patterns R are finite and fixed, which is common for longitudinal studies, and we let $n = \min_k n_k \to \infty$, i.e., sample size of the smallest stratum goes to infinity. Then, without loss of generality, we let $n_k = n$ for $k = 1, \dots, R$.

The work by *Johnson et al.* (2008) has shown that the estimator for the penalized GEE in (4.6) behaves asymptotically as if the true model is known a priori, i.e., the so called *oracle* property, for a broad class of penalty functions $\boldsymbol{q}_\lambda(\cdot)$, including LASSO and SCAD. We show that this property also holds for our fusion estimator with MCP.

Let $\boldsymbol{\beta}_*^T = (\boldsymbol{\beta}_{*1}^T, \dots, \boldsymbol{\beta}_{*p}^T) = (\beta_{*11}, \dots, \beta_{*1R}, \dots, \beta_{*p1}, \dots, \beta_{*pR})$, denote the true values of $\boldsymbol{\beta}$ with some clustered structures, correspondingly, $\boldsymbol{\theta}_* = \boldsymbol{D}\boldsymbol{\beta}_*$ is sparse with some zero elements. Suppose that $\mathcal{A} = \{l : \theta_l \neq 0, l = 1, \dots, pR\}$. Following *Johnson et al.* (2008), we have the following theoretical results:

Theorem IV.1. *Under the regularity conditions listed in Appendix G, let $\boldsymbol{D}_N$ be the contrast matrix based on root-N consistent estimates such that $\lim_N \boldsymbol{D}_N = \boldsymbol{D}$, the approximate solution $\hat{\boldsymbol{\theta}}$ to the penalized GEE in* (4.6) *with MCP penalty satisfies:*

(i) (Selection Consistency) $\lim_N P(\hat{\theta}_l = 0 \textit{ for } l \notin \mathcal{A}) = 1$;

(ii) (Asymptotic Normality)

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{*\mathcal{A}}) \to_d \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\mathcal{A}})$$

where $\boldsymbol{\Sigma}_{\mathcal{A}} = \{\boldsymbol{H}_{\mathcal{A}}^{\boldsymbol{D}}(\boldsymbol{\theta}_*) + \boldsymbol{Q}_{\mathcal{A}}(\boldsymbol{\theta}_*)\}^{-1}\boldsymbol{V}_{\mathcal{A}}^{\boldsymbol{D}}(\boldsymbol{\theta}_*)\{\boldsymbol{H}_{\mathcal{A}}^{\boldsymbol{D}}(\boldsymbol{\theta}_*) + \boldsymbol{Q}_{\mathcal{A}}(\boldsymbol{\theta}_*)\}^{-1}$.

In Theorem IV.1, $\boldsymbol{H}^{\boldsymbol{D}}(\boldsymbol{\theta}_*) = \text{block-diag}\{\boldsymbol{H}_1^{\boldsymbol{D}}(\boldsymbol{\theta}_{*1}), \dots, \boldsymbol{H}_R^{\boldsymbol{D}}(\boldsymbol{\theta}_{*R})\}$, $\boldsymbol{V}^{\boldsymbol{D}}(\boldsymbol{\theta}_*) = \text{block-diag}\{\boldsymbol{V}_1^{\boldsymbol{D}}(\boldsymbol{\theta}_{*1}), \dots, \boldsymbol{V}_R^{\boldsymbol{D}}(\boldsymbol{\theta}_{*R})\}$, and $\boldsymbol{Q}(\boldsymbol{\theta}_*) = \text{diag}\{-\boldsymbol{q}'_{\lambda_N}(|\boldsymbol{\theta}_*|)\,\text{sign}(\boldsymbol{\theta}_*)\}$. More specifically, for k, $\boldsymbol{H}_k^{\boldsymbol{D}}(\boldsymbol{\theta}) = \sum_{i=1}^{n_k} \boldsymbol{X}_{ki}^T \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})\boldsymbol{R}_k^{-1}(\boldsymbol{\tau})\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})\boldsymbol{X}_{ki}$ and

$\boldsymbol{V}_k^{\boldsymbol{D}}(\boldsymbol{\theta}) = \sum_{i=1}^n \boldsymbol{X}_{ki}^T \boldsymbol{A}_{ki}^{1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})\boldsymbol{R}_k^{-1}(\boldsymbol{\tau})\boldsymbol{A}_{ki}^{-1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})\,\mathrm{var}(\boldsymbol{Y}_{ki})\boldsymbol{A}_{ki}^{-1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})$ $\boldsymbol{R}_k^{-1}(\boldsymbol{\tau})\boldsymbol{A}_{ki}^{1/2}(\boldsymbol{E}_k\boldsymbol{D}^{-1}\boldsymbol{\theta})\boldsymbol{X}_{ki}$. The proof of above theorem is presented in Appendix G. Subsequently, we can obtain similar asymptotic results for the estimator $\hat{\boldsymbol{\beta}}$ for (4.5). The theorem implies that with probability one, we are able to recover the true structure of the parameters, and that the nonzero differences of the adjacent parameters follow asymptotically a multivariate normal distribution.

4.6 Simulation study

In this section, we perform two simulation experiments to illustrate the performance of the proposed fusion learning method on the application of pattern-mixture models for nonignorable missingness. The assessment of our method includes the aspects of parameter estimation, model selection, and robustness against specification of working correlation structure. We consider both the linear model and the logistic model. For both cases, we simulate longitudinal data with $M = 4$ visits, where missing patterns are set as two or more consecutive measurements observed. In other words, the generated samples belong to one of the following $R = 8$ patterns $\mathcal{L}_k = \{j : \ell_j = 1, j = 1, \ldots, M\}$, $k = 1, \ldots, R$, where $\ell_1\ell_2\ell_3\ell_4 \in \{0011, 0110, 0111, 1011, 1101, 1100, 1110, 1111\}$. For simplicity, we consider equal sample sizes $n_k = n = 100$ for all the missing patterns, $k = 1, \ldots, R$.

For the linear model with continuous outcomes, the following true model is used to generate data:

$$Y_{kij} = \beta_{k1}X_{kij1} + \beta_{k2}X_{kij2} + \beta_{k3}X_{kij3} + \beta_{k4}X_{kij4} + \beta_{k5}X_{kij5} + \epsilon_{kij}$$

for $k = 1, \ldots, R$, $i = 1, \ldots, n$, and $j \in \mathcal{L}_k$, where $\beta_k^T = (\beta_{k1}, \ldots, \beta_{k5})$ is the true model coefficient vector under pattern k and $\boldsymbol{\epsilon}_{ki}$ is the marginal errors for subject i in pattern k derived from $(\epsilon_{ki1}, \epsilon_{ki2}, \epsilon_{ki3}, \epsilon_{ki4})^T \sim \mathcal{N}_4(\mathbf{0}, \phi\boldsymbol{R}(\tau))$. In our simulation experiments, data are simulated with exchangeable correlation where the true correlation parameter is $\tau = 0.6$. The covariates $\boldsymbol{X}_{kij}^T = (X_{kij1}, \ldots, X_{kij5})$, $k = 1, \ldots, R$, $i = 1, \ldots, n$, and $j \in \mathcal{L}_k$, are simulated independently from a standard multivariate normal distribution. We let the true coefficients of covariates $\boldsymbol{X}_1, \boldsymbol{X}_3$ and $\boldsymbol{X}_5$ to be homogeneous across missing patterns; they are, $\boldsymbol{\beta}_1^T = (\beta_{11}, \ldots, \beta_{1R}) = (0.5, \ldots, 0.5)$, $\boldsymbol{\beta}_3^T = (\beta_{31}, \ldots, \beta_{3R}) = (0.2, \ldots, 0.2)$, and $\boldsymbol{\beta}_5^T = (\beta_{51}, \ldots, \beta_{5R}) = (0.0, \ldots, 0.0)$. We let the true coefficients of $\boldsymbol{X}_2$ and $\boldsymbol{X}_4$ to be heterogeneous, each with K distinct groups, where δ denotes the gaps between distinct groups. We vary the value of K as $K = 1, 2, 3, 4$. For example, for the case of $K = 2$, we randomly partition elements in $\boldsymbol{\beta}_2^T = (\beta_{21}, \ldots, \beta_{2R})$ and $\boldsymbol{\beta}_4^T = (\beta_{41}, \ldots, \beta_{4R})$ into two groups, and the true values for one group are larger than the other group by δ. One example may look like the following vectors of coefficients: $\boldsymbol{\beta}_2$ has two groups with values 0.3 and $0.3+\delta$, while $\boldsymbol{\beta}_4$ has two groups with values 0 and δ, i.e., $\boldsymbol{\beta}_2^T = (0.3, 0.3+\delta, 0.3, 0.3, 0.3, 0.3+\delta, 0.3, 0.3+\delta)$ and $\boldsymbol{\beta}_4^T = (0, \delta, 0, 0, 0, \delta, 0, \delta)$. Likewise, we create heterogeneous groups for $K = 3$ and 4. Note that $K = 1$ corresponds to the homogeneous coefficients. To see the performance of fusion learning, we vary δ to demonstrate the ability of our method in detecting different levels of differences and recover the underlying group structures generated by the simulation models.

We compare our method with a two-step pattern-mixture modeling (denoted by PMM) approach given as follows: First, conduct a test for MCAR (or test for heterogeneity) by the Wald-type test statistic in *Chen and Little* (1999); then, fit either a common GEE if we fail to reject MCAR, or fit a pattern-stratified GEE if the test rejects MCAR. The metrics used to evaluate the two-step method include power of the MCAR test, mean squared error between the estimated and the true coefficients. To compare, for the fusion learning method, we report the number of times coefficients are not completely fused, denoted as sensitivity for heterogeneity, and mean squared error between the estimated and the true coefficients. Additionally, we report the number of groups estimated by the fusion approach. Although fusion learning method does not directly test for MCAR, the sensitivity is a comparable metric to power in the two-step PMM approach.

Table 4.1 summarizes the results for the linear model with independent working correlation for varying values of K and δ, each from 500 replications. The random error $\boldsymbol{\epsilon}_{k,i}$ is generated with correlation $\tau = 0.6$ but $\boldsymbol{R}(\cdot) = \boldsymbol{I}$ is used. It is interesting to see that the sensitivity of fusion learning is larger than the power of PMM, indicating that fusion learning is more often to call heterogeneity. In the case when $K = 1$, power of PMM corresponds to Type-1 error. We can see that even for the PMM method, Type-I error is not well controlled at 5% when the null is true. The MSEs of regression coefficients are almost consistently lower in the fusion learning method than the PMM approach, which reflects the advantage of our method in fully utilizing the underlying parameter structures to improve estimation. Especially, the MSEs for the homogeneous covariates are much smaller in the fusion approach. As for the detecting of grouping structures, we can see that as gap δ becomes larger, the estimated number of groups gets closer to the true number of groups for the fusion approach. Table 4.2 shows the results under the exact same setting as Table 4.1, but with correctly specified exchangeable working correlation during fitting. While we compare within each of the methods, the correct working correlation structure produces smaller MSE in estimation in both methods, indicating the benefit of accounting for within-subject correlations. Although correctly specifying the correlation structure improves parameter estimation, it does not improve the clustering performance of fusion learning. In other words, the clustering performance is quite insensitive to the working correlation structure.

For the logistic model with binary outcomes, the data are simulated by

$$\operatorname{logit}\{E(Y_{kij})\} = \beta_{k1}X_{kij1} + \beta_{k2}X_{kij2} + \beta_{k3}X_{kij3} + \beta_{k4}X_{kij4} + \beta_{k5}X_{kij5}$$

for $k = 1, \ldots, R$, $i = 1, \ldots, n$, and $j \in \mathcal{L}_k$. The covariates $\boldsymbol{X}_{kij}^T = (X_{kij1}, \ldots, X_{kij5})$ are simulated independently from a standard multivariate normal distribution. Similar to the linear model case, we simulate the responses with exchangeable correlation $\tau = 0.6$, and we consider homogeneous $\boldsymbol{\beta}_1^T = (0.5, \ldots, 0.5)$, $\boldsymbol{\beta}_3^T = (0.2, \ldots, 0.2)$ and $\boldsymbol{\beta}_5^T = (0.0, \ldots, 0.0)$, and heterogeneous $\boldsymbol{\beta}_2^T$ and $\boldsymbol{\beta}_4^T$ according to number of groups K and gap size δ. We summarize the results of the comparison metrics for the logistic model across 500 replications in Table 4.3 for independent working correlation and Table 4.4 for exchangeable working correlation. From both tables, we see that our proposed method is more sensitive to detecting deviation from MCAR than *Chen and Little* (1999)'s Wald-type test. In terms of MSE for $\boldsymbol{\beta}$, the two methods have comparable values when MCAR is true ($K = 1$), but the MSE for the two-step PMM method gradually increases when the violation of MCAR becomes more severe. On the other hand, our fusion approach gives stable MSE estimates for $\boldsymbol{\beta}$ regardless of the level of violation from MCAR, and provides satisfactory estimation of the grouping structures.

Table 4.1: Simulation results for the linear model with independent working correlation matrix when the true correlation is exchangeable ($\tau = 0.6$).

		PMM						Fusion										
			MSE ($\times 100$)						MSE ($\times 100$)					Average Size				
K	δ	Power	β_1	β_2	β_3	β_4	β_5	Sensitivity	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
										MCAR								
1	0	21.6	0.159	0.151	0.139	0.156	0.133	39.8	0.107	0.112	0.087	0.106	0.090	1.1(0.3)	1.1(0.3)	1.1(0.3)	1.1(0.3)	1.1(0.3)
										MNAR								
2	0.1	59.2	0.292	0.367	0.266	0.378	0.270	68.6	0.125	0.335	0.093	0.333	0.104	1.1(0.4)	1.4(0.5)	1.1(0.3)	1.4(0.5)	1.1(0.3)
2	0.2	98.4	0.403	0.408	0.377	0.415	0.386	98.4	0.120	0.388	0.090	0.439	0.092	1.1(0.4)	2.0(0.4)	1.1(0.3)	2.0(0.4)	1.1(0.3)
2	0.3	100.0	0.406	0.402	0.382	0.411	0.390	100.0	0.087	0.232	0.068	0.237	0.068	1.1(0.2)	2.0(0.2)	1.0(0.2)	2.0(0.2)	1.0(0.2)
2	0.4	100.0	0.406	0.402	0.382	0.411	0.390	100.0	0.076	0.128	0.058	0.149	0.056	1.0(0.2)	2.0(0.2)	1.0(0.2)	2.0(0.2)	1.0(0.2)
3	0.1	89.8	0.377	0.411	0.353	0.420	0.363	90.2	0.124	0.468	0.099	0.471	0.096	1.1(0.4)	1.8(0.5)	1.1(0.3)	1.8(0.5)	1.1(0.3)
3	0.2	99.8	0.405	0.403	0.381	0.412	0.390	99.6	0.101	0.567	0.081	0.576	0.075	1.1(0.3)	2.3(0.5)	1.1(0.3)	2.4(0.5)	1.1(0.3)
3	0.3	100.0	0.406	0.402	0.382	0.411	0.390	100.0	0.086	0.368	0.067	0.372	0.066	1.1(0.2)	2.8(0.5)	1.0(0.2)	2.8(0.5)	1.0(0.2)
3	0.4	100.0	0.406	0.402	0.382	0.411	0.390	100.0	0.070	0.187	0.059	0.226	0.053	1.0(0.2)	2.9(0.4)	1.0(0.2)	2.9(0.4)	1.0(0.1)

Table 4.2: Simulation results for the linear model with exchangeable working correlation matrix when the true correlation is exchangeable ($\tau = 0.6$).

		PMM						Fusion										
			MSE ($\times 100$)						MSE ($\times 100$)					Average Size				
K	δ	Power	β_1	β_2	β_3	β_4	β_5	Sensitivity	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
										MCAR								
1	0	26.4	0.100	0.096	0.098	0.098	0.096	7.8	0.031	0.032	0.031	0.029	0.029	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)
										MNAR								
2	0.1	82.8	0.214	0.239	0.203	0.247	0.214	32.4	0.034	0.238	0.030	0.240	0.031	1.0(0.1)	1.2(0.4)	1.0(0.1)	1.2(0.4)	1.0(0.2)
2	0.2	99.8	0.244	0.234	0.229	0.243	0.242	95.0	0.033	0.217	0.031	0.232	0.028	1.0(0.1)	1.9(0.3)	1.0(0.1)	1.9(0.3)	1.0(0.1)
2	0.3	100.0	0.244	0.233	0.229	0.243	0.242	99.8	0.031	0.114	0.028	0.108	0.025	1.0(0.1)	2.0(0.1)	1.0(0.1)	2.0(0.1)	1.0(0.1)
2	0.4	100.0	0.244	0.233	0.229	0.243	0.242	99.8	0.029	0.071	0.026	0.075	0.025	1.0(0.0)	2.0(0.1)	1.0(0.1)	2.0(0.1)	1.0(0.1)
3	0.1	98.6	0.242	0.234	0.227	0.243	0.240	74.8	0.034	0.363	0.031	0.367	0.031	1.0(0.1)	1.7(0.5)	1.0(0.1)	1.6(0.5)	1.0(0.2)
3	0.2	100.0	0.244	0.233	0.229	0.243	0.242	99.0	0.032	0.420	0.028	0.449	0.031	1.0(0.1)	2.3(0.5)	1.0(0.1)	2.3(0.5)	1.0(0.1)
3	0.3	100.0	0.244	0.233	0.229	0.243	0.242	100.0	0.030	0.223	0.027	0.212	0.025	1.0(0.1)	2.8(0.4)	1.0(0.1)	2.8(0.4)	1.0(0.1)
3	0.4	100.0	0.244	0.233	0.229	0.243	0.242	100.0	0.029	0.109	0.027	0.112	0.025	1.0(0.0)	2.9(0.3)	1.0(0.1)	2.9(0.3)	1.0(0.1)

Table 4.3: Simulation results for the logistic model with independent working correlation matrix when the true correlation is exchangeable ($\tau = 0.6$).

		PMM						Fusion										
			MSE (×100)						MSE (×100)					Average Size				
K	δ	Power	β_1	β_2	β_3	β_4	β_5	Sensitivity	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
										MCAR								
1	0	15.0	0.673	0.548	0.540	0.561	0.567	51.8	0.756	0.559	0.559	0.576	0.56	1.2(0.4)	1.2(0.4)	1.2(0.4)	1.2(0.4)	1.2(0.4)
										MNAR								
2	0.1	19.2	0.776	0.834	0.652	0.819	0.632	61.0	0.806	0.889	0.592	0.833	0.556	1.2(0.4)	1.3(0.4)	1.2(0.4)	1.2(0.4)	1.2(0.4)
2	0.2	37.2	1.193	1.565	0.945	1.504	0.958	74.0	0.807	1.555	0.559	1.451	0.586	1.2(0.4)	1.5(0.5)	1.2(0.4)	1.5(0.5)	1.2(0.4)
2	0.3	71.0	1.850	2.190	1.519	2.006	1.480	92.4	0.854	2.117	0.598	1.915	0.567	1.3(0.4)	1.7(0.5)	1.2(0.4)	1.7(0.5)	1.2(0.4)
2	0.4	91.2	2.183	2.306	1.834	2.084	1.807	98.8	0.773	2.098	0.631	1.932	0.566	1.2(0.4)	2(0.4)	1.2(0.4)	2(0.4)	1.2(0.4)
3	0.1	29.8	0.999	1.237	0.849	1.209	0.822	72.2	0.784	1.227	0.564	1.229	0.535	1.2(0.4)	1.4(0.5)	1.2(0.4)	1.4(0.5)	1.2(0.4)
3	0.2	77.2	1.894	2.038	1.609	1.956	1.571	93.2	0.762	2.099	0.630	2.000	0.570	1.2(0.4)	1.8(0.5)	1.2(0.4)	1.8(0.5)	1.2(0.4)
3	0.3	95.2	2.231	2.034	1.874	1.964	1.813	99.0	0.759	2.466	0.605	2.376	0.500	1.2(0.4)	2.1(0.5)	1.2(0.4)	2.1(0.5)	1.1(0.3)
3	0.4	99.4	2.294	2.064	1.946	2.011	1.896	100.0	0.660	2.648	0.508	2.622	0.526	1.2(0.4)	2.4(0.5)	1.1(0.3)	2.4(0.5)	1.1(0.3)

Table 4.4: Simulation results for the logistic model with exchangeable working correlation matrix when the true correlation is exchangeable ($\tau = 0.6$).

		PMM						Fusion										
			MSE (×100)						MSE (×100)					Average Size				
K	δ	Power	β_1	β_2	β_3	β_4	β_5	Sensitivity	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
										MCAR								
1	0	16.6	0.588	0.472	0.478	0.482	0.478	35.6	0.448	0.342	0.379	0.365	0.328	1.1(0.3)	1.1(0.3)	1.1(0.3)	1.1(0.3)	1.1(0.3)
										MNAR								
2	0.1	25.0	0.748	0.835	0.626	0.768	0.589	42.2	0.478	0.649	0.362	0.610	0.312	1.1(0.3)	1.2(0.4)	1.1(0.3)	1.1(0.4)	1.1(0.3)
2	0.2	49.6	1.191	1.504	0.975	1.355	0.936	62.2	0.518	1.313	0.360	1.240	0.350	1.1(0.4)	1.4(0.5)	1.1(0.3)	1.4(0.5)	1.1(0.3)
2	0.3	79.6	1.623	1.928	1.358	1.684	1.318	84.4	0.592	1.888	0.402	1.688	0.361	1.2(0.4)	1.6(0.6)	1.1(0.3)	1.7(0.5)	1.1(0.3)
2	0.4	94.2	1.838	1.961	1.547	1.711	1.522	98.2	0.540	1.847	0.39	1.587	0.392	1.1(0.3)	1.9(0.4)	1.1(0.3)	2(0.4)	1.1(0.3)
3	0.1	37.6	0.937	1.180	0.817	1.122	0.772	53.8	0.485	1.036	0.373	0.978	0.323	1.1(0.4)	1.3(0.5)	1.1(0.3)	1.3(0.4)	1.1(0.3)
3	0.2	84.8	1.676	1.739	1.424	1.571	1.334	87.8	0.529	1.956	0.410	1.804	0.357	1.2(0.4)	1.7(0.5)	1.1(0.3)	1.8(0.5)	1.1(0.3)
3	0.3	98.0	1.864	1.708	1.578	1.566	1.494	98.4	0.470	2.269	0.343	2.195	0.317	1.1(0.3)	2.1(0.5)	1.1(0.3)	2.1(0.5)	1.1(0.3)
3	0.4	99.2	1.885	1.757	1.609	1.658	1.537	100.0	0.433	2.479	0.344	2.301	0.304	1.1(0.3)	2.3(0.5)	1.1(0.3)	2.4(0.5)	1.1(0.2)

4.7 Application: Intern Health Study

We apply the proposed fusion learning to analyze the suicidal ideation data from the Intern Health Study introduced in Section 1. The data was collected from longitudinal suicidal ideation screening on over 2400 medical interns in their first year of residency from hospitals across the US, along with the collection of various measurements on psychiatric health assessment and other potential risk factors from baseline and each of the four visits. The baseline visit occurs in April before the onset of medical training in June, and the four scheduled visits occur every three month throughout the first year of internship, in September, December, March and June, respectively. The four baseline covariates of interest are age, gender, baseline suicidal ideation (SI) and score of psychological health from Patient Health Questionnaire (PHQ) (*Kroenke et al.*, 2001), and the other four time dependent risk factors are PHQ score, anxiety score from General Anxiety Disorder questionnaire (GAD) (*Spitzer et al.*, 2006), binary indicator of whether conducted medical error in the past three month (MEDERR) and average work hours in the past three month (HOUR). During data preprocessing, we keep subjects who has at least attended two consecutive visits, resulting in $R = 8$ distinct missing data patterns. See Table 4.5 for a summary of the variables. Continuous covariates are standardized before data analysis. Our goal is go identify predictors of suicidal behavior in order to implement early intervention measures. Before we proceed with fitting GEE, we apply the test for MCAR versus nonignorable missingness using *Chen and Little* (1999)'s method. This test rejects the hypothesis of MCAR with significance ($p < 10^{-9}$). Thus, the test warrants the application of PMM that allows to analyze effects of the risk factors according to heterogeneous missing patterns.

Table 4.5: Summary statistics for suicidal ideation data. Means (and standard deviations) are reported for continuous covariates and percentages are reported for binary covariates.

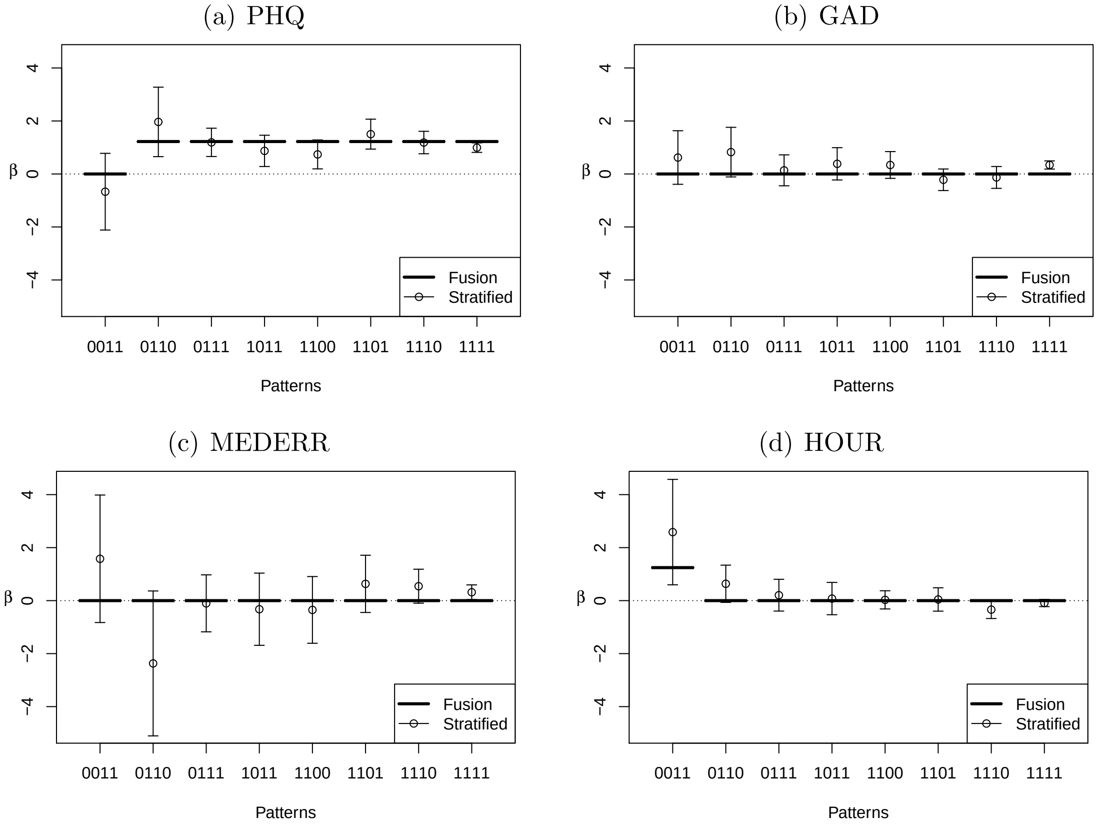
Missing Pattern	0011	0110	0111	1011	1100	1101	1110	1111
Sample size	41	68	120	128	150	141	249	1570
Baseline								
Age	27.9(3.5)	27.2(2.1)	27.6(3)	27.8(3.5)	27.6(3)	27.7(3)	27.3(2.3)	27.4(2.5)
Female (%)	56.1	47.1	38.3	54.7	54.0	49.6	47.0	51.3
Baseline SI (%)	4.9	4.4	4.2	4.7	3.3	3.5	3.6	2.7
Baseline PHQ score	2.6(3.1)	1.8(2.3)	2.8(3.5)	2.7(3.1)	2.5(2.8)	3.1(3.4)	2.6(3)	2.4(2.7)
Time Dependent								
PHQ score	5.6(4.8)	6.1(4.3)	5.1(4.1)	6.4(4.9)	6.4(4.6)	6.4(4.7)	5.6(4.5)	5.3(4.2)
GAD score	4.1(4.4)	4.8(4)	3.9(4.2)	5(4.5)	5.2(4.5)	5.7(4.9)	4.5(4.4)	4.4(4.2)
MEDERR (%)	19.5	21.3	20.0	15.9	26.3	22.5	19.0	18.2
HOUR	64.7(17)	64.7(18.1)	66.3(17.5)	63.6(17.5)	65.5(18.6)	65.3(20)	63.8(19.2)	63.8(18.5)
SI (%)	8.5	11.0	8.1	9.6	10.3	10.9	9.2	6.8

We invoke the PMM for the binary suicidal ideation outcome $E(SI_{k,ij}) = \mu_{k,ij}$:

$$\begin{aligned} \text{logit}(\mu_{kij}) &= \beta_0 + \beta_1 age_i + \beta_2 sex_i + \beta_3 SI_{i0} + \beta_4 PHQ_{i0} \\ & \quad \beta_{5k} PHQ_{kij} + \beta_{6k} GAD_{kij} + \beta_{7k} MEDERR_{kij} + \beta_{8k} HOUR_{kij}, \end{aligned} \tag{4.7}$$

$k = 1, \ldots, 8$, where the baseline covariates are assumed to be homogeneous, and the effects of time-dependent risk factors are set as being heterogeneous according to the 8 missing patterns. We postulate that heterogeneous risk groups may be smaller than 8, the number of missing patterns; in other words, there may exist some common grouping patterns within effects of the risk factors, $\boldsymbol{\beta}_5, \ldots, \boldsymbol{\beta}_8$. To deal with potential estimation bias due to sample size differences across missing pattern strata, as pointed out earlier in this chapter, we use the inverse of sample sizes as sample weights in the penalized GEE. The smallest ERIC (*Hui et al.*, 2015) is the optimal criterion for the tuning.

We compare the fusion learning estimates with the pattern-mixture modeling approach, which fits stratified GEE models separately, each for one of $R = 8$ missing patterns. Figure 4.3 overlays the coefficient estimates from the fusion learning and the coefficient estimates and 95% confidence intervals from the stratified GEE approach for the four time-dependent covariates of interest, PHQ, GAD, MEDERR and HOUR, respectively. Due to the use of sparsity penalty in the penalized GEE, some of the coefficients are estimated exactly as zero. Comparison of these two methods indicates that the largest contrast lies in the evidence that the fusion learning clusters coefficients across missing patterns to form some lower dimensional grouping structures. On the other hand, the values of our estimates are strikingly consistent with the coverage of zero for the 95% confidence intervals from stratified analysis. For example, the 95% confidence intervals of PHQ effects do not cover zero, except for the "0011" pattern, which are in agreement with the findings from the fusion estimates.


(a) PHQ

(b) GAD

(c) MEDERR

(d) HOUR

Figure 4.3: Coefficient estimates from fusion learning, and coefficient estimates and 95% confidence intervals for stratified GEE across eight missing patterns for time dependent covariates PHQ, GAD, MEDERR and HOUR, respectively.

4.8 Concluding Remarks

In this chapter, we adopt the MCP penalty for fusion learning and extend the likelihood based fusion learning method in Chapter II to the generalized estimating equations framework. Such extension allows us to model correlated data, and only requires specification of the first two moments. Therefore, the method in this chapter can be readily applied to much more general cases, such as quantile regression, survival models, and missing data problems.

We would also like to draw comparison between the fusion learning method and modeling with interactions, because they are similar in some sense. When the number of missing pattern strata is small, we may often investigate the pattern-mixtures by including interaction terms of covariates and the strata using the nonpenalized GEE. Such model allow us to conduct hypothesis testing to determine whether the covariate effects in any strata is different from that of the baseline strata. However, in order to recover the underlying grouping structure, additional testing is required. When the number of strata is large, for example, in the IHS study, pairwise testing requires testing of 8 choose 2 differences for each of the 4 variables, which is a total of 112 tests. Such testing raises the concern of the multiple testing issue, and likely lead to conflicting conclusions. Therefore, fusion learning is much superior and requires less manual effort.

Here, we assume the number of missing pattern strata is fixed, which may be the case for traditional longitudinal studies with a fixed number of design visits. However, when the measurement is more dense, such as in tracking data or accelerometry data, the number of missing patterns will explode, posing a challenge to the existing fusion learning framework. In those settings, extension of the current method is needed since stratum-specific estimates might be infeasible to obtain.

CHAPTER V

Summary and Future Work

Motivated by the studies of large scale data sets, this dissertation has focused on the extensions of classic regression methodologies to big data scenarios. Under the big data setting, some conventional assumptions and beliefs are no longer satisfied. Traditional statistics assumes carefully collected random samples represent some larger populations, and the main objective is to generalize conclusions within the random samples to the populations. On the other hand for big data, we when have rich amount of information on a large percentage of population, interest shifts toward subgroup analysis within big data in search of commonality between subjects. This idea, along with real data, has motivated the development of methods in Chapter II and IV. Besides, big data often pose computationally challenging to standard machines, and due to privacy protection, many biomedical data are not centrally stored, but stored in parts with different level of protection protocols. Driven by this concern, we developed the divide-and-combine method in Chapter III. The three methods in this dissertation respectively address different statistical challenges, which includes: data integration, dimension reduction, inference, clustering, and optimization. Each of the methods can be further extended and improved along in their own framework and settings as have been discussed in each of the chapters. But more importantly, it is of great interest to borrow the strengths of each of the proposed methods, pairing with new methodologies developed by others, to further generalize the big data analytical toolbox, and construct a standard framework that is applicable to a broader range of problems pertaining to modern big data sets. Therefore, we conclude this chapter by pointing out potential directions for future research of modeling for big data.

One promising direction is to incorporate fusion learning methodology into the confidence distribution framework by directly combine the methods in Chapter II and III. This allows confidence distributions from various data sets or studies to be heterogeneous, while borrowing the strength of computational simplicity. The confidence distribution can be traced back to Fisher's fiducial argument (*Fisher*, 1930), and was later formulated by *Efron* (1993). This can also be related to the least squares approximation described in *Wang and Leng* (2007). Using this framework, we can reduce the representation of likelihoods or estimating functions to their minimal form which only describes the parameters of interests. By doing so, we reduce the amount of data used to represent the desired knowledge. And in this way, we allow existing methods to become more scalable.

Driven by the recent initiatives of precision medicine, another promising direction and natural extension of current work is to study the heterogeneity within subjects, instead of data sets as considered in this dissertation. Subject-level heterogeneity may be of more interest in medical studies or in online advertisement, because knowing the exact behavior of individuals by drawing similarity with comparable subjects can help better target medical treatment or recommendation of products. To generalize further along this line, a more ambitious goal is develop multi-resolution analysis tools, where we can zoom-in to study the heterogeneity and homogeneity between subjects, and we can also zoom-out to any level to study the heterogeneity and homogeneity between any grouping of our choices, such as by city, state, country, gender, ethnicity, disease type, etc. The technical challenges involved include the development of more advanced algorithm to handle individual level complexity, and more flexible assumption so that the methods can be generalizable to real scenarios. We may develop methodologies within the scheme of hierarchical clustering (*Johnson*, 1967).

Last but not least, we would like to extend the consideration of data integration to more complicated settings than currently assumed. Especially, we may no longer assume responses and covariates collected are the same across different data sets, we may no longer assume the measurements have the same granularity across all sources, we may no longer assume data are homogeneous in the sense that the means and variances of the parameters of interest are consistent across partitions of the data, and we may no longer assume that the methods used to analyze each data sets to be the same. Although the above mentioned cases are very common in real life settings, there has not been many research in the literature in this area. We would like to address these important issues by leveraging the knowledge accumulated through the development of this dissertation.

APPENDICES

APPENDIX A

Chapter II: Proofs

Proof of Theorem II.1

Proof. The proof of Theorem II.1 closely follows arguments given in *Zou* (2006). Without loss of generality, we assume $n_1 = \cdots = n_K = n$ and $N = Kn$. As K is fixed, $n \to \infty$ implies $N \to \infty$ in the same order. We assume the following regularity conditions:

(i) The Fisher information matrix is finite and positive definite,

$$\boldsymbol{I}(\boldsymbol{\theta}^*) = E\left[\phi''\left(\boldsymbol{X}^T\boldsymbol{\theta}^*\right)\boldsymbol{X}\boldsymbol{X}^T\right].$$

Here, $\boldsymbol{\theta}^*_{(Kp\times 1)}$ is the true parameters, $\boldsymbol{X}_{(N\times Kp)}$ is the design matrix corresponding to $\boldsymbol{\theta}$ and ϕ is the link function (i.e., $\phi' = h^{-1}$) defined in the following optimization problem

$$\hat{\boldsymbol{\theta}}^{\boldsymbol{W}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\left\{-\frac{1}{K}\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n_k}\left(Y_k^{(i)}\boldsymbol{X}_k^{(i)T}\boldsymbol{\theta}(\lambda) - \phi\left(\boldsymbol{X}_k^{(i)T}\boldsymbol{\theta}(\lambda)\right)\right) + P_{\lambda,\alpha}(\boldsymbol{\theta})\right\}$$

with $P_{\lambda,\alpha}(\boldsymbol{\theta})$ as defined in (2.4), and $\hat{\boldsymbol{\theta}}^{\boldsymbol{W}}$ is the estimator with true ordering $\boldsymbol{W}$ given.

(ii) There is a sufficiently large open set $\mathcal{O}$ that contains $\boldsymbol{\theta}^*$ such that $\forall\boldsymbol{\theta} \in \mathcal{O}$,

$$|\phi'''(\boldsymbol{X}^T\boldsymbol{\theta})| \leq M(\boldsymbol{X}^T) < \infty, \text{ and}$$
$$E\left[M(\boldsymbol{X})|x_j x_k x_l|\right] < \infty$$

for a suitable function M and all $1 \leq j,k,l \leq Kp$.

First we prove asymptotic normality. For $\forall s \geq 0$ and $r > 0$, let $\boldsymbol{\theta} = \boldsymbol{\theta}^* + \boldsymbol{u}/\sqrt{N}$. Define

$$\Gamma_N(\boldsymbol{u}) = -\sum_{k=1}^{K}\sum_{i=1}^{n}\left(Y_k^{(i)}\boldsymbol{X}_k^{(i)T}\left(\boldsymbol{\theta}^* + \frac{\boldsymbol{u}}{\sqrt{N}}\right) - \phi\left(\boldsymbol{X}_k^{(i)T}\left(\boldsymbol{\theta}^* + \frac{\boldsymbol{u}}{\sqrt{N}}\right)\right)\right)$$
$$+\lambda_N\sum_{j=1}^{p}\sum_{k=1}^{K}\hat{\omega}_{j,k}\left|\theta_{j,k}^* + \frac{u_{j,k}}{\sqrt{N}}\right|$$

where $\hat{\omega}_{j,k}$ is specified in (2.6). Let $\hat{\boldsymbol{u}}^{(N)} = \arg\min_{\boldsymbol{u}}\Gamma_N(\boldsymbol{u})$; then $\hat{\boldsymbol{u}}^{(N)} = \sqrt{N}(\hat{\boldsymbol{\theta}}^{\boldsymbol{W}} - \boldsymbol{\theta}^*)$. By Taylor expansion, we have $\Gamma_N(\boldsymbol{u}) - \Gamma_N(\boldsymbol{0}) = H^{(N)}(\boldsymbol{u})$, where

$$H^{(N)}(\boldsymbol{u}) \equiv A_1^{(N)} + A_2^{(N)} + A_3^{(N)} + A_4^{(N)},$$

with

$$A_1^{(N)} = -\sum_{k=1}^{K}\sum_{i=1}^{n}\left[Y_k^{(i)} - \phi'(\boldsymbol{X}_k^{(i)T}\boldsymbol{\theta}^*)\right]\frac{\boldsymbol{X}_k^{(i)T}\boldsymbol{u}}{\sqrt{N}},$$
$$A_2^{(N)} = \sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{2}\phi''(\boldsymbol{X}_k^{(i)T}\boldsymbol{\theta}^*)\boldsymbol{u}^T\frac{\boldsymbol{X}_k^{(i)}\boldsymbol{X}_k^{(i)T}}{\sqrt{N}}\boldsymbol{u},$$
$$A_3^{(N)} = \frac{\lambda_N}{\sqrt{N}}\sum_{j=1}^{p}\sum_{k=1}^{K}\hat{\omega}_{j,k}\sqrt{N}\left(\left|\theta_{j,k}^* + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta_{j,k}^*\right|\right),$$
$$\text{and } A_4^{(N)} = N^{-3/2}\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{1}{6}\phi'''\left(\boldsymbol{X}_k^{(i)T}\tilde{\boldsymbol{\theta}}_*\right)\left(\boldsymbol{X}_k^{(i)T}\boldsymbol{u}\right)^3,$$

where $\tilde{\boldsymbol{\theta}}_*$ is between $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^* + \boldsymbol{u}/\sqrt{N}$. The asymptotic limits of $A_1^{(N)}$, $A_2^{(N)}$ and $A_4^{(N)}$ is exactly the same as those in the proof in (*Zou*, 2006, Theorem 4). It suffice to show that $A_3^{(N)}$ has the same asymptotic limit. If $\theta^*_{j,k} \neq 0$, $\hat{\omega}_{j,1} \to_p \alpha|\theta^*_{j,1}|^{-r}$, $\hat{\omega}_{j,k} \to_p |\sum_{k'=2}^{K} \theta^*_{j,k'}|^{-s}|\theta^*_{j,k}|^{-r}$ for $k = 2, \ldots, K$, and $\sqrt{N}\left(\left|\theta^*_{j,k} + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta^*_{j,k}\right|\right) \to u_{j,k}\,\mathrm{sgn}(\theta^*_{j,k})$. Thus by Slutsky's theorem, $A_3^{(N)} \to 0$. If $\theta^*_{j,k} = 0$, for $k = 1$, since $\sqrt{N}\hat{\theta}_{j,1} = O_p(1)$, $\frac{\lambda_N}{\sqrt{N}} N^{r-2}\alpha(|\sqrt{N}\hat{\theta}_{j,1}|)^{-r} \to \infty$; for $k = 2, \ldots, K$, if $\sum_{k'=2}^{K} \theta^*_{j,k'} = 0$ (i.e., homogeneous), $\sqrt{N}\sum_{k'=2}^{K} \hat{\theta}_{j,k'} = O_p(1)$, thus $\frac{\lambda_N}{\sqrt{N}} N^{\frac{s+r}{2}}(|\sqrt{N}\sum_{k'=2}^{K} \hat{\theta}_{j,k'}|)^{-s}$ $(|\sqrt{N}\hat{\theta}_{j,k}|)^{-r} \to \infty$; similarly, if $\sum_{k'=2}^{K} \theta^*_{j,k'} \neq 0$ (i.e., heterogeneous), $\sum_{k'=2}^{K} \hat{\theta}_{j,k'} \to_p \sum_{k'=2}^{K} \theta^*_{j,k'}$, $\frac{\lambda_N}{\sqrt{N}}\hat{\omega}_{j,k} \to \infty$ still holds. And since $\sqrt{N}\left(\left|\theta^*_{j,k} + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta^*_{j,k}\right|\right) \to |u_{j,k}|$, we have the following result summary:

$$\frac{\lambda_N}{\sqrt{N}}\hat{\omega}_{j,k}\sqrt{N}\left(\left|\theta^*_{j,k} + \frac{u_{j,k}}{\sqrt{N}}\right| - \left|\theta^*_{j,k}\right|\right) \to_p \begin{cases} 0 & \text{if } \theta^*_{j,k} \neq 0 \\ 0 & \text{if } \theta^*_{j,k} = 0 \text{ and } u_{j,k} = 0 \\ \infty & \text{if } \theta^*_{j,k} = 0 \text{ and } u_{j,k} \neq 0. \end{cases}$$

Following same arguments in (*Zou*, 2006, Theorem 4), we have $\hat{\boldsymbol{u}}_{\mathcal{A}}^{(N)} \to_d \mathcal{N}(0, \boldsymbol{I}_{11}^{-1})$ and $\hat{\boldsymbol{u}}_{\mathcal{A}^c}^{(N)} \to_d \mathbf{0}$. The proof of the consistency part is similar and thus omitted. □

Proof of Lemma II.2

Proof. The estimated ordering $\hat{\boldsymbol{U}}_j$ of $\boldsymbol{\beta}^*_{j,\bullet}$ is only determined by the differences between distinct parameter groups within $\boldsymbol{\beta}^*_{j,\bullet}$. First note that for any $0 < \epsilon < 1$, if two parameters $\beta^*_{j,k}$ and $\beta^*_{j,k'}$ are in the same parameter group (i.e., $\beta^*_{j,k} = \beta^*_{j,k'}$), assigning arbitrary ordering between them will not affect the estimated ordering of the parameters between groups, because the ordering within the same parameter group is exchangeable. On the other hand, when two parameters $\beta^*_{j,k}$ and $\beta^*_{j,k'}$ are from different parameter groups, without loss of generality, let $\beta^*_{j,k} > \beta^*_{j,k'}$, the probability of estimating a wrong ordering

$$\begin{aligned}P\left(\mathbf{1}\{\hat{\beta}_{j,k'} \geq \hat{\beta}_{j,k}\} > \epsilon\right) &= P\left(\hat{\beta}_{j,k'} \geq \hat{\beta}_{j,k}\right) \\ &= P\left(\hat{\beta}_{j,k'} - \hat{\beta}_{j,k} + \beta^*_{j,k} - \beta^*_{j,k'} \geq \beta^*_{j,k} - \beta^*_{j,k'}\right) \\ &\leq P\left(|\hat{\beta}_{j,k'} - \beta^*_{j,k'}| + |\hat{\beta}_{j,k} - \beta^*_{j,k}| > 0\right) \\ &= 1 - P\left(\hat{\beta}_{j,k'} = \beta^*_{j,k'}\right) P\left(\hat{\beta}_{j,k} = \beta^*_{j,k}\right) \to 0\end{aligned}$$

as $n \to \infty$ since $\hat{\beta}_{j,k'}$ and $\hat{\beta}_{j,k}$ are independent and consistent estimators. Similarly, the consistency of the estimated ordering $\hat{\boldsymbol{V}}_j$ of the absolute values in vector $\boldsymbol{\beta}^*_{j,\bullet}$ can be derived by taking the square of the absolute values and following the same argument as for $\hat{\boldsymbol{U}}_j$. □

Proof of Theorem II.3

Proof. Here we assume the same regularity condition as in Theorem II.1. To complete this proof, we first define the event $\mathcal{W}$ when the orderings of all p covariates are correctly assigned as

$$\mathcal{W} = \bigcap_{j=1}^{p} \left(\{\hat{\boldsymbol{U}}_j = \boldsymbol{U}_j\} \cap \{\hat{\boldsymbol{V}}_j = \boldsymbol{V}_j\}\right).$$

Let $\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}}$ be $\hat{\boldsymbol{\theta}}_{\mathcal{W}}$ when $\mathcal{W}$ occurs; otherwise, denote it as $\hat{\boldsymbol{\theta}}_{\mathcal{W}^c}$. Then, the estimator can be rewritten as

$$\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}} = \hat{\boldsymbol{\theta}}_{\mathcal{W}}\mathbf{1}\{\mathcal{W}\} + \hat{\boldsymbol{\theta}}_{\mathcal{W}^c}\mathbf{1}\{\mathcal{W}^c\}$$

and therefore

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{W}}} - \boldsymbol{\theta}^*\right) = \sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}} - \boldsymbol{\theta}^*\right)\mathbf{1}\{\mathcal{W}\} + \sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}^c} - \boldsymbol{\theta}^*\right)\mathbf{1}\{\mathcal{W}^c\}. \tag{A.1}$$

By Theorem II.1, we have $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}} - \boldsymbol{\theta}^*\right) = O(1)$ and $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}^c} - \boldsymbol{\theta}^*\right) = O(1)$ as $n \to \infty$. By Lemma II.2, we have $P(\mathcal{W}) \to 1$ and $P(\mathcal{W}^c) \to 0$ as $n \to \infty$. Therefore, by Slutsky's Theorem, (A.1) converge to the same distribution as $\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathcal{W}} - \boldsymbol{\theta}^*\right)$. Similarly, by results from Theorem II.1 and Lemma II.2, we have selection consistency

$$P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}) = P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}|\mathcal{W})P(\mathcal{W}) + P(\hat{\mathcal{A}}^{\hat{\boldsymbol{W}}} = \mathcal{A}|\mathcal{W}^c)P(\mathcal{W}^c) \to 1$$

as $n \to \infty$. This completes the proof of the Theorem II.3. □

APPENDIX B

Chapter II: Additional Simulation

Performance with Distorted Parameter Ordering

Under the same setting as simulation experiment 1 in Section 2.5.1 with $\alpha = 0$ and $s = 0$, we conduct a sensitivity analysis to evaluate the performance of FLARCC when parameter ordering is incorrectly specified. Specifically, we report results of sensitivity, specificity and MSE for the linear regression model when the coefficient ordering is determined from the initial estimate with distortion through an added disturbance ϵ, $\hat{\boldsymbol{\beta}}+\epsilon$, where $\hat{\boldsymbol{\beta}}$ from (2.1) and $\epsilon \sim \mathcal{N}(0, v^2)$. As v^2 increases, the percent of order switching in initial estimates increases. Sensitivity, specificity and MSE in relation to the percentage of wrongly ordered parameters are displayed in Figure B.1 for the two heterogeneous effects $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, and the homogeneous parameter $\boldsymbol{\beta}_1$ is not included in the comparison because of no effect from the distortion on its performance. As the percentage of wrongly ordered parameters increases, as expected, sensitivity becomes lower and MSE becomes larger. However, specificity remains unaffected. When the distortion of ordering is mild ($\leq$ 10%), the performance of FLARCC appears satisfactory in this simulation setting.

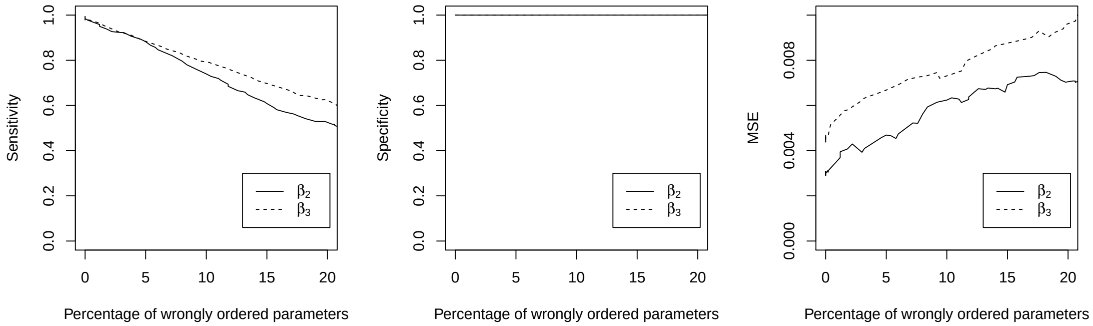


Figure B.1: Clustering sensitivity and mean squared error of two heterogeneous slope parameters $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ based on FLARCC with λ selected by EBIC, as the percent of distorted ordering increases. Results are summarized from 100 replications.

APPENDIX C

Chapter II: Extension to the Cox Proportional Hazards Model

In addition to the original notations defined in Chapter II, we include notations for the formulation of Cox model. The K datasets are denoted as $\{y_{k,i}, \boldsymbol{x}_{k,i}, \delta_{k,i}\}_{i=1}^{n_k}$ for $k = 1, \ldots, K$, with individual's covariates $\boldsymbol{x}_{k,i} = (x_{1,k,i}, \ldots, x_{p,k,i})^T$, observed time $y_{k,i} \triangleq \min(T_{k,i}, C_{k,i})$ and event indicator $\delta_{k,i} \triangleq I(T_{k,i} \leq C_{k,i})$, where $T_{k,i}$ and $C_{k,i}$ are the failure time and the censoring time, respectively.

We assume that $T_{k,i}$ and $C_{k,i}$ are conditionally independent given $\boldsymbol{x}_{k,i}$, and that the censoring mechanism is noninformative. Additionally, we assume that all datasets have a common baseline hazard $\lambda_0(t)$. We start from the complete heterogeneous Cox proportional hazards model specification that takes the form:

$$\lambda_k(t|\boldsymbol{x}) = \lambda_0(t) \exp\left(\boldsymbol{x}_k^T \boldsymbol{\beta}_k\right), \quad k = 1, \ldots, K. \tag{C.1}$$

The partial likelihood for model (C.1) on the combined data can be written as

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} L_k(\boldsymbol{\beta}) = \prod_{k=1}^{K} \prod_{i=1}^{n_k} \left[\frac{\exp\left(\boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_k\right)}{\sum_{kk=1}^{K} \sum_{ii=1}^{n_{kk}} I(y_{kk,ii} \geq y_{k,i}) \exp\left(\boldsymbol{x}_{kk,ii}^T \boldsymbol{\beta}_{kk}\right)} \right]^{\delta_{k,i}}, \tag{C.2}$$

where $L_k(\boldsymbol{\beta})$ is the likelihood piece corresponding to the kth dataset. We specify the risk set (i.e., the denominator) in (C.2) based on all subjects from the combined dataset to assess the overall risk in this data integration problem. Using the fused lasso penalty with parameter ordering on the partial likelihood (C.2), we solve the same optimization problem as in (2.2), where $\ell_k(\boldsymbol{\beta})$ is the log partial likelihood of the kth dataset given by

$$\ell_k(\boldsymbol{\beta}) = \log L_k(\boldsymbol{\beta}) = \sum_{i=1}^{n_k} \delta_{k,i} \left\{ \boldsymbol{x}_{k,i}^T \boldsymbol{\beta}_k - \log \left[\sum_{kk=1}^{K} \sum_{ii=1}^{n_{kk}} I(y_{kk,ii} \geq y_{k,i}) \exp\left(\boldsymbol{x}_{kk,ii}^T \boldsymbol{\beta}_{kk} \right) \right] \right\}, \tag{C.3}$$

and penalty $P_\lambda(\boldsymbol{\beta})$ is the fused lasso penalty defined by (2.4). The implementation is similar to that of Section 2.2. R package `glmnet` is used for the lasso optimization problem of the Cox proportional hazards model.

APPENDIX D

Chapter III: Proofs

Proof of Theorem III.4

Proof. For the sake of notation consistency in the context of divide-and-combine, we explicitly write both subscripts k and n_k in all terms in the proof, and show the consistency and asymptotic normality of the bias-corrected estimator $\hat{\boldsymbol{\beta}}^c_{\lambda,k}$ of the k-th dataset. Here, n_k denotes the sample size of the k-th dataset.

Following the Karush-Kuhn-Tucker condition in (D.2) and condition (C3) that $\lambda_k = (\log p/n_k)^{1/2}$ and $p < n_k$, we have $\boldsymbol{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k}) = o_p(1)$. Under conditions (C1) and (C2), for any $\boldsymbol{\beta} \in \mathcal{N}_\delta(\boldsymbol{\beta}_0)$,

$$
\begin{aligned}
c_1 \leq \underline{\sigma}\left(\boldsymbol{P}^{1/2}_{n_k}(\boldsymbol{\beta})\right) \underline{\sigma}(n_k^{-1/2}\boldsymbol{X}_k) \leq \underline{\sigma}\left(n_k^{-1/2}\boldsymbol{P}^{1/2}_{n_k}(\boldsymbol{\beta})\boldsymbol{X}_k\right) \\
\leq \overline{\sigma}\left(n_k^{-1/2}\boldsymbol{P}^{1/2}_{n_k}(\boldsymbol{\beta})\boldsymbol{X}_k\right) \leq \overline{\sigma}\left(\boldsymbol{P}^{1/2}_{n_k}(\boldsymbol{\beta})\right) \overline{\sigma}(n_k^{-1/2}\boldsymbol{X}_k) \leq C_1,
\end{aligned}
$$

where $\boldsymbol{P}_{n_k}(\boldsymbol{\beta}) = \operatorname{diag}\{v_{k,1}, \ldots, v_{k,n_k}\}$, $v_{k,i}$ is the variance function under the canonical link functions, c_1 and C_1 are two positive constants, and $\underline{\sigma}(\cdot)$ and $\overline{\sigma}(\cdot)$ are defined in condition (C2). Hence, $-\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)$ is positive definite and $\|\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\|_2 = O_p(1)$, where $\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0) \overset{def}{=} \{-\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\}^{-1}$. On the other hand, by the law of large numbers, $\boldsymbol{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k}) \to E\left\{\boldsymbol{S}_{n_k}(\boldsymbol{\beta})\right\}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\lambda,k}} \to 0$. Combining this with condition (C1) that $E\left\{\boldsymbol{S}_{n_k}(\boldsymbol{\beta}_0)\right\} = 0$ and the negative definite property of $\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)$, we have $\hat{\boldsymbol{\beta}}_{\lambda,k} \to \boldsymbol{\beta}_0$. Then, the consistency of $\hat{\boldsymbol{\beta}}^c_{\lambda,k}$ follows from its definition in (D.3) and $\|\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\|_2 = O_p(1)$.

Next, we show the asymptotic normality of $\hat{\boldsymbol{\beta}}^c_{\lambda,k}$. Again, following the Karush-Kuhn-Tucker condition in (D.2), by the first-order Taylor expansion and conditions (C1)-(C3), we have

$$\hat{\boldsymbol{\beta}}^c_{\lambda,k} - \boldsymbol{\beta}_0 = \hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\boldsymbol{S}_{n_k}(\boldsymbol{\beta}_0) + \|\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\|_2 O_p(s_0\lambda_k^2). \tag{D.1}$$

Furthermore, using the central limit theorem and Slutsky's theorem, the first term in (D.1) $n_k^{1/2}\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\boldsymbol{S}_{n_k}(\boldsymbol{\beta}_0) \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}(\boldsymbol{\beta}_0)\right)$ asymptotically as $n_k \to \infty$. On the other hand, by the condition (C3), we can show that the second term $\|\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\|_2 O_p(s_0\lambda_k^2) = O_p(s_0\lambda_k^2) = o_p(n_k^{-1/2})$. In summary, the proof of Theorem III.4 is completed. □

Proof of Theorem III.7

Proof. Denote $r_N(\boldsymbol{\beta}) = \frac{1}{N}\sum_{k=1}^K \partial \log \hat{h}_{n_k}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ and $r(\boldsymbol{\beta}) = \lim_{n_{\min}\to\infty} r_N(\boldsymbol{\beta})$. It is easy to see $r(\hat{\boldsymbol{\beta}}_{dac}) = 0$. On the other hand,

$$\begin{aligned} r_N(\boldsymbol{\beta}_0) &= -\tfrac{1}{N}\textstyle\sum_{k=1}^K n_k \left\{\hat{\boldsymbol{\Sigma}}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k})\right\}^{-1}\left\{\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{\lambda,k} - \hat{\boldsymbol{\Sigma}}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k})\boldsymbol{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k})\right\} \\ &= \tfrac{1}{N}\textstyle\sum_{k=1}^K n_k \left\{\boldsymbol{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k}) + \dot{\boldsymbol{S}}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{\lambda,k})\right\} \\ &= \tfrac{1}{N}\textstyle\sum_{k=1}^K n_k \boldsymbol{S}_{n_k}(\boldsymbol{\beta}_0) + O_p\left(N^{-1}K\right), \end{aligned}$$

where the second equality holds by conditions (C1)-(C3). Then, by the law of large numbers, $r(\boldsymbol{\beta}_0) = E\left\{\boldsymbol{S}_N(\boldsymbol{\beta}_0)\right\} = 0$, where the second equation follows from condition (C1). Furthermore, we have $\dot{r}(\boldsymbol{\beta}_0) = -\Sigma(\boldsymbol{\beta}_0)$, which is a positive definite matrix. Combining this with $r(\boldsymbol{\beta}_0) = r(\hat{\boldsymbol{\beta}}_{dac}) = 0$, the consistency of $\hat{\boldsymbol{\beta}}_{dac}$ follows.

Next we prove the asymptotic normality of $\hat{\boldsymbol{\beta}}_{dac}$. By some simple algebra, we can obtain that

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{dac} &= \left[\sum_{k=1}^{K} n_k \left\{\hat{\boldsymbol{\Sigma}}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k})\right\}^{-1}\right]^{-1} \left[\sum_{k=1}^{K} n_k \left\{\hat{\boldsymbol{\Sigma}}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k})\right\}^{-1} \hat{\boldsymbol{\beta}}^{c}_{\lambda,k}\right] \\ &= \left\{\frac{1}{N}\sum_{k=1}^{K} n_k \dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1} \left\{\frac{1}{N}\sum_{k=1}^{K} n_k \dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\hat{\boldsymbol{\beta}}^{c}_{\lambda,k}\right\} + O_p(N^{-1}K) + o_p(N^{-1/2}),\end{aligned}$$

and $\text{var}(\hat{\boldsymbol{\beta}}_{dac}) = N^{-1}\boldsymbol{\Sigma}_{dac}$. Combining with the condition that $K = O(N^{1/2-\delta})$ with $\delta \in (0, 1/2)$ being a constant and the central limit theorem, the asymptotic normal distribution in Theorem III.7 follows.

Finally, it suffices to show the gold estimator $\hat{\boldsymbol{\beta}}_{full}$ has the same asymptotic distribution as $\hat{\boldsymbol{\beta}}_{dac}$. By the definition of $\hat{\boldsymbol{\beta}}_{full}$ in Theorem III.7, we have $\hat{\boldsymbol{\beta}}_{full} - \boldsymbol{\beta}_0 = \{\dot{\boldsymbol{S}}_N(\boldsymbol{\beta}_0)\}^{-1}\boldsymbol{S}_N(\boldsymbol{\beta}_0) + o_p(N^{-1/2})$. The asymptotically equivalent efficiency claimed in Theorem III.7 follows by the central limit theorem. □

An Extension of Theorem III.4 with $p \to \infty$

For any fixed q, denote

$$\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda) - \lambda\hat{\boldsymbol{\kappa}} = 0, \tag{D.2}$$

$$\hat{\boldsymbol{\beta}}^{c}_{\lambda} \overset{\text{def}}{=} \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}^{-1}\lambda\hat{\boldsymbol{\kappa}} = \hat{\boldsymbol{\beta}}_\lambda + \{-\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_0)\}^{-1}\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda), \tag{D.3}$$

$$\hat{\boldsymbol{\gamma}}_\lambda = \boldsymbol{H}\hat{\boldsymbol{\beta}}_\lambda + \boldsymbol{H}\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}}_\lambda)\boldsymbol{S}_n(\hat{\boldsymbol{\beta}}_\lambda), \tag{D.4}$$

where $\boldsymbol{H}$ is a rank q matrix of dimension $q \times p$ with the (i, j)-th element denoted as h_{ij}. $\boldsymbol{H}$ can increase in dimension, but in practice, $\boldsymbol{H}$ is fixed. For $1 \le i, k \le p$, let σ_{ik} satisfies, $\sum_{k=1}^{p} \sigma_{ik} E[\{\phi\dot{g}(\boldsymbol{x}_1^T\boldsymbol{\beta}_0)\}^{-1}x_{1k}x_{1j}] = \delta_{ij}$, where $\delta_{ij} = 1$ for $i = j$, and $\delta_{ij} = 0$, for $i \ne j$.

Theorem D.1. *Under conditions (C1)-(C3) in Section 3.2.4, the estimator $\hat{\boldsymbol{\gamma}}_\lambda$ given in* (D.4) *is consistent and asymptotically normally distributed, namely, $n^{1/2}(\hat{\boldsymbol{\gamma}}_\lambda - \boldsymbol{\gamma}_0) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{A}_{\boldsymbol{\gamma}})$, as $n \to \infty$ where $\boldsymbol{A}_{\boldsymbol{\gamma}}$ is a matrix of dimension $q \times q$ with the*

(i, j)-th element $a_{ij} = \sum_{k_1,k_2=1}^{p} h_{ik_1}\sigma_{k_1k_2}h_{k_2j}$.

Proof. For the sake of notation consistency in the context of divide-and-combine, we explicitly write both subscripts k and n_k in all terms in the proof, and show the consistency and asymptotic normality of the bias-corrected estimator $\hat{\boldsymbol{\gamma}}_{\lambda,k}$ of the k-th dataset. Here, n_k denotes the sample size of the k-th dataset.

Following the Karush-Kuhn-Tucker condition in (D.2) and condition (C3) that $\lambda_k = (\log p/n_k)^{1/2}$ and $p < n_k$, we have $\boldsymbol{S}_{n_k}(\hat{\boldsymbol{\beta}}_{\lambda,k}) = o_p(1)\mathbf{1}_p$, where $\mathbf{1}_p$ is a p-dimensional vector with values all at 1. Under conditions (C1) and (C2), for any $\boldsymbol{\beta} \in \mathcal{N}_\delta(\boldsymbol{\beta}_0)$,

$$\begin{aligned} c_1 \le \underline{\sigma}\left(\boldsymbol{P}_{n_k}^{1/2}(\boldsymbol{\beta})\right)\underline{\sigma}(n_k^{-1/2}\boldsymbol{X}_k) &\le \underline{\sigma}\left(n_k^{-1/2}\boldsymbol{P}_{n_k}^{1/2}(\boldsymbol{\beta})\boldsymbol{X}_k\right) \\ &\le \overline{\sigma}\left(n_k^{-1/2}\boldsymbol{P}_{n_k}^{1/2}(\boldsymbol{\beta})\boldsymbol{X}_k\right) \le \overline{\sigma}\left(\boldsymbol{P}_{n_k}^{1/2}(\boldsymbol{\beta})\right)\overline{\sigma}(n_k^{-1/2}\boldsymbol{X}_k) \le C_1, \end{aligned}$$

where $\boldsymbol{P}_{n_k}(\boldsymbol{\beta}) = \operatorname{diag}\{v_{k,1}, \ldots, v_{k,n_k}\}$, $v_{k,i}$ is the variance function under the canonical link function, c_1 and C_1 are two positive constants, and $\underline{\sigma}(\cdot)$ and $\overline{\sigma}(\cdot)$ are defined in condition (C2). Hence, $-\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)$ is positive definite and $\|\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\|_\infty = O_p(1)$, where $\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0) \overset{def}{=} \{-\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\}^{-1}$. On the other hand, by using $PL(\hat{\boldsymbol{\beta}}_{\lambda,k}; \boldsymbol{Y}, \boldsymbol{X}) \ge PL(\boldsymbol{\beta}_0; \boldsymbol{Y}, \boldsymbol{X})$, we have that

$$\begin{aligned} \lambda\|\boldsymbol{\beta}_0\|_1 &\ge \frac{1}{n}\left\{\mathcal{L}_n(\boldsymbol{\beta}_0; \boldsymbol{Y}, \boldsymbol{X}) - \mathcal{L}_n(\hat{\boldsymbol{\beta}}_{\lambda,k}; \boldsymbol{Y}, \boldsymbol{X})\right\} + \lambda\|\hat{\boldsymbol{\beta}}_{\lambda,k}\|_1 \\ &= -\boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0) + \frac{1}{2n\phi}(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)^T\left\{\boldsymbol{X}_k^T\boldsymbol{P}_n(\tilde{\boldsymbol{\beta}}_k)\boldsymbol{X}_k\right\}(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0) \\ &\quad +\lambda\|\hat{\boldsymbol{\beta}}_{\lambda,k}\|_1, \end{aligned}$$

which indicates

$$\|\boldsymbol{P}_n^{1/2}(\tilde{\boldsymbol{\beta}}_k)\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)\|_2^2/(n_k\phi) + 2\lambda_k\|\hat{\boldsymbol{\beta}}_{\lambda,k}\|_1 \le 2\boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0) + 2\lambda\|\boldsymbol{\beta}_0\|_1.$$

By using Corollary 6.2 in (*Bühlmann and Van De Geer*, 2011) and conditions (C1)-

(C3), it is easy to get that

$$\|\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}_{\lambda_k} - \boldsymbol{\beta}_0)\|_2^2/n_k + \lambda_k\|\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0\|_1 \le C\lambda_k^2 s_0 \tag{D.5}$$

Next, we show the consistency and asymptotic normality of $\hat{\boldsymbol{\gamma}}_{\lambda,k}$. Again, following the Karush-Kuhn-Tucker condition in (D.2), by the first-order Taylor expansion and conditions (C1)-(C3), we have

$$\hat{\boldsymbol{\gamma}}_{\lambda,k} - \boldsymbol{\gamma}_0 = \boldsymbol{H}_k\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\boldsymbol{S}_{n_k}(\boldsymbol{\beta}_0) + \boldsymbol{R}_{n_k}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0), \tag{D.6}$$

where

$$\begin{aligned}
\boldsymbol{R}_{n_k}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) &= \boldsymbol{H}_k\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\left\{\dot{\boldsymbol{S}}_{n_k}(\tilde{\boldsymbol{\beta}}) - \dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)\\
&= \boldsymbol{H}_k\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\frac{1}{n_k\phi}\boldsymbol{X}_k^T\left\{\boldsymbol{P}_{n_k}(\tilde{\boldsymbol{\beta}}) - \boldsymbol{P}_{n_k}(\boldsymbol{\beta}_0)\right\}\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)\\
&= \boldsymbol{H}_k\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\frac{1}{n_k\phi}\boldsymbol{X}_k^T\boldsymbol{Z}_k,
\end{aligned}$$

where $\boldsymbol{Z}_k$ is a n_k dimensional vector with $Z_{ki} = \boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)(\dot{g}^{-1}(\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) - \dot{g}^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0))$. Note that

$$\begin{aligned}
&\|\boldsymbol{R}_{n_k}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0; \boldsymbol{H}_k)\|_2^2\\
=\ & \mathrm{tr}\left[\boldsymbol{H}_k\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\tfrac{1}{n_k\phi}\boldsymbol{X}_k^T\boldsymbol{Z}_k\boldsymbol{Z}_k^T\boldsymbol{X}_k\tfrac{1}{n_k\phi}\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\boldsymbol{H}_k^T\right]\\
\le\ & \phi^{-2}\mathrm{tr}\left(\tfrac{1}{n_k}\boldsymbol{Z}_k\boldsymbol{Z}_k^T\right)\mathrm{tr}\left[\boldsymbol{X}_k\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\boldsymbol{H}_k^T\boldsymbol{H}_k\left\{\dot{\boldsymbol{S}}_{n_k}(\boldsymbol{\beta}_0)\right\}^{-1}\tfrac{1}{n_k}\boldsymbol{X}_k^T\right]\\
\le\ & c_1^{-1}\phi^{-2}\left[\tfrac{1}{n_k}\textstyle\sum_{i=1}^{n_k}\left\{\boldsymbol{x}_i^T(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)\right\}^2\left\{\dot{g}^{-1}(\boldsymbol{x}_i^T\tilde{\boldsymbol{\beta}}) - \dot{g}^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)\right\}^2\right]\\
&\qquad\qquad\times\mathrm{tr}\left[\boldsymbol{H}_k n_k\left\{\boldsymbol{X}_k^T\boldsymbol{P}_{n_k}(\boldsymbol{\beta}_0)\boldsymbol{X}_k\right\}^{-1}\boldsymbol{H}_k^T\right]\\
\le\ & c_1^{-2}\phi^{-2}\|\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}_{\lambda,k} - \boldsymbol{\beta}_0)\|_2^2/n_k\mathrm{tr}\left[\boldsymbol{H}_k n_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^{-1}\boldsymbol{H}_k^T\right].
\end{aligned} \tag{D.7}$$

Furthermore, using the central limit theorem and Slutsky's theorem, the first term in

(D.6) $n_k^{1/2}\boldsymbol{H}_k\hat{\boldsymbol{\Sigma}}_{n_k}(\boldsymbol{\beta}_0)\boldsymbol{S}_{n_k}(\boldsymbol{\beta}_0) \sim \mathcal{N}\left(0, \boldsymbol{A}_{\boldsymbol{\gamma}}\right)$ asymptotically as $n_k \to \infty$. On the other hand, combining condition (C2) with inequalities (D.5) and (D.7), it is easy to show that the second term $\|\boldsymbol{R}_{n_k}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0; \boldsymbol{H}_k)\|_2^2 = O_p(s_0\lambda_k^2) = o_p(n_k^{-1/2})$. In summary, the proof of Theorem D.1 is completed. □

Tuning Parameter λ Selected by Cross-Validation

Let D_{train} and D_{val} denote a training and validating split of data D. Denote $\hat{\boldsymbol{\beta}}_{\lambda_f}$ the estimator based on the tunning parameter λ_{cv} obtained from cross validation, and $\hat{\boldsymbol{\beta}}_{\lambda}$ the estimator based on λ satisfying condition (C3), that is, $\lambda = O(\log p/n_k)^{1/2}$, both obtained from D_{train}. Following the definition of cross validation, we have that

$$
\begin{aligned}
0 \ &\leq \ D_N(D_{val}; \hat{\boldsymbol{\beta}}(\lambda)) - D_N(D_{val}; \hat{\boldsymbol{\beta}}(\lambda_f)) \\
&= \ 2\phi\boldsymbol{\mathcal{L}}_N(\hat{\boldsymbol{\beta}}(\lambda_f); D_{val}) - 2\phi\boldsymbol{\mathcal{L}}_N(\hat{\boldsymbol{\beta}}(\lambda); D_{val}) \\
&= \ 2\phi\boldsymbol{\mathcal{L}}_N(\boldsymbol{\beta}_0; D_{val}) + 2\phi\boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0) \\
&\qquad\qquad +\phi(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0)^T\left\{\dot{\boldsymbol{S}}_N(\boldsymbol{\beta}_{m1})\right\}(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0) \\
&\quad -2\phi\boldsymbol{\mathcal{L}}_N(\boldsymbol{\beta}_0; D_{val}) - 2\phi\boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0) \\
&\qquad\qquad -\phi(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0)^T\left\{\dot{\boldsymbol{S}}_N(\boldsymbol{\beta}_{m2})\right\}(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0),
\end{aligned}
$$

where $D_N(\cdot)$, $\mathcal{L}_N(\cdot)$, $\boldsymbol{S}_N(\cdot)$ are the deviance function, log-likelihood function, and score function based on the test data, respectively, $\hat{\boldsymbol{\beta}}(\lambda)$ and $\hat{\boldsymbol{\beta}}(\lambda_f)$ are the corresponding estimators based on the training data. $\boldsymbol{\beta}_{m1}$ is a value between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}(\lambda_f)$, $\boldsymbol{\beta}_{m2}$ is a value between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}(\lambda)$.

Based on the fact that $\boldsymbol{S}_n(\boldsymbol{\beta}_0)$ and $\hat{\boldsymbol{\beta}}(\lambda)$ are evaluated using independent datasets D_{val} and D_{train}, respectively, the expected values of $2\phi\boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0)$ and

$2\phi \boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0)$ are both zero. Hence, we have that

$$\begin{aligned} & 2\phi \boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}(\lambda) - \boldsymbol{\beta}_0) - 2\phi \boldsymbol{S}_n(\boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0) \\ & \qquad -\phi(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0)^T \left\{ \dot{\boldsymbol{S}}_N(\boldsymbol{\beta}_{m1}) \right\} (\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0) \\ = \;& 2\phi \boldsymbol{S}_n(\boldsymbol{\beta}_0)^T \left[\left\{ -\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_{m4}) \right\}^{-1} - \left\{ -\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_{m3}) \right\}^{-1} \right] \boldsymbol{S}_n(\boldsymbol{\beta}_0) \\ & \qquad -\phi(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0)^T \left\{ \dot{\boldsymbol{S}}_N(\boldsymbol{\beta}_{m1}) \right\} (\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0) \\ & \qquad +2\phi \boldsymbol{S}_n(\boldsymbol{\beta}_0)^T \left\{ -\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_{m3}) \right\}^{-1} \lambda_f \hat{\boldsymbol{\kappa}}_f - 2\phi \boldsymbol{S}_n(\boldsymbol{\beta}_0)^T \left\{ -\dot{\boldsymbol{S}}_n(\boldsymbol{\beta}_{m4}) \right\}^{-1} \lambda \hat{\boldsymbol{\kappa}} \\ \leq \;& C\lambda^2 s_0. \end{aligned}$$

where the last inequality holds by using expressions (D.5) and condition (C3). Combining with conditions (C1) and (C2), it follows that $\|\boldsymbol{X}\left(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0\right)\|_2^2/n_k \leq C\lambda^2 s_0$.

On the other hand, it is easy to get that $\|n_k^{-1/2}\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}_k(\lambda_f) - \boldsymbol{\beta}_0)\|_1 \leq \|\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0)\|_2/\sqrt{n_k} \times \sqrt{n_k}$. That is, $\|n_k^{-1}\boldsymbol{X}_k\left(\hat{\boldsymbol{\beta}}_k(\lambda_f) - \boldsymbol{\beta}_0\right)\|_1 \leq C\lambda\sqrt{s_0}$. By using condition (C2) that $\max_k \|\boldsymbol{X}\|_\infty = O(1)$, it follows that, $\|\hat{\boldsymbol{\beta}}_k(\lambda_f) - \boldsymbol{\beta}_0\|_1 \leq C\lambda\sqrt{s_0}$. Thus,

$$\|\boldsymbol{X}_k(\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0)\|_2^2/n_k + \lambda\|\hat{\boldsymbol{\beta}}(\lambda_f) - \boldsymbol{\beta}_0\|_1 \leq C\lambda^2 s_0.$$

Then following similar proof of Theorem III.4, we can get that Theorem D.1 holds for $\hat{\boldsymbol{\gamma}}_{\lambda,k}$ with λ being selected via the cross validation.

APPENDIX E

Chapter III: Additional Simulation

Covariate Correlation Versus Coverage Probability

To establish some guidelines about how to select n_k, we consider an additional simulation in which the correlation between covariates varies in terms of correlation coefficients ρ, and evaluate the performance of MODAC and META under different choices of K. Table E.1 provides statistical inference results. The asymptotic confidence intervals of $\boldsymbol{\beta}_{\mathcal{A}_0}$ of MODAC achieve the 95% nominal coverage in most scenarios, except for the logistic regression with ρ being small. Clearly, better performance of coverage occurs with bigger sub-dataset sizes. It is interesting to see that the performance gets better when the correlation ρ goes higher. The poorer performance of MODAC in the logistic regression with a small ρ may be due to the curse of dimensionality. As pointed out by *Hall et al.* (2005), data tend to lie deterministically at the vertices of a regular simplex when the number of independent covariates goes to infinity and sample size is fixed. In other words, a limited amount of data would be problematic to make a valid statistical inference. On the other hand, larger correlation ρ reduces an effective degree of freedom which makes statistical inference a relatively easier task. Overall, the coverage probabilities of MODAC is uniformly

Table E.1: Simulation results when $N = 10,000$ and $p = 300$ for Gaussian, logistic and Poisson models. Methods with different size of partition K and and compound symetric correlation ρ are compared. $\mathcal{A}_0$ and $\mathcal{A}_0^c$ denote the set of non-zero and zero coefficients in $\boldsymbol{\beta}_0$, respectively. Results are from an average of 500 replications

				MODAC	MODAC	MODAC	MODAC	META	META	META	META
K	n_k	Type	Set	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.8$
20	500	Gaussian	$\mathcal{A}_0$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
20	500	Binomial	$\mathcal{A}_0$	0.75	0.85	0.91	0.95	1.00	1.00	1.00	1.00
20	500	Poisson	$\mathcal{A}_0$	0.94	0.95	0.95	0.95	0.85	0.89	0.90	0.92
10	1000	Gaussian	$\mathcal{A}_0$	0.95	0.94	0.94	0.95	0.95	0.95	0.94	0.95
10	1000	Binomial	$\mathcal{A}_0$	0.87	0.89	0.92	0.95	0.01	0.01	0.15	0.07
10	1000	Poisson	$\mathcal{A}_0$	0.95	0.95	0.95	0.95	0.89	0.92	0.92	0.94
2	5000	Gaussian	$\mathcal{A}_0$	0.95	0.95	0.96	0.95	0.95	0.95	0.95	0.95
2	5000	Binomial	$\mathcal{A}_0$	0.94	0.94	0.94	0.95	0.96	0.96	0.96	0.96
2	5000	Poisson	$\mathcal{A}_0$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.94
20	500	Gaussian	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
20	500	Binomial	$\mathcal{A}_0^c$	0.96	0.96	0.96	0.96	1.00	1.00	1.00	1.00
20	500	Poisson	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.90	0.91	0.91	0.92
10	1000	Gaussian	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
10	1000	Binomial	$\mathcal{A}_0^c$	0.96	0.96	0.96	0.96	1.00	1.00	0.88	0.18
10	1000	Poisson	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.92	0.93	0.93	0.94
2	5000	Gaussian	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
2	5000	Binomial	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96
2	5000	Poisson	$\mathcal{A}_0^c$	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95

MODAC denotes our proposed divide-and-combine method and META denotes the meta-analysis method.

more consistent than those of META. Based on the empirical results of MODAC, in practice, we suggest choosing a reasonably large n_k in the logistic regression when covariates have weak dependence.

Sensitivity of ω in Majority Voting

Figure E.1 presents a sensitivity analysis of variable selection performance of the VOTING method by *Chen and Xie* (2014) with respect to the choice of ω under three models, Gaussian, logistic and Poisson. We let $N = 10,000$, $p = 300$, $s_0 = 10$ and we vary the number of split K and the correlation coefficient ρ from a compound symmetric structure. The non-zero coefficients are set at 0.3 for Gaussian models, 0.3 for logistic models, and 0.1 for Poisson models. As shown, clearly the Gaussian model is much more robust that the other two models by allowing a much wider range of ω to achieve the highest sensitivity and specificity. However, for logistic and Poisson

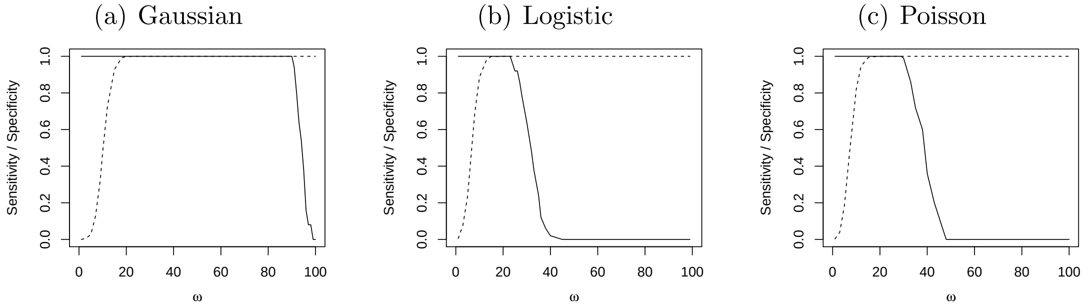


Figure E.1: Sensitivity (solid) and specificity (doted) of VOTING as voting threshold ω varies from 0 to 100. The total sample size $N = 50,000$, the number of split $K = 100$, and the number of covariates $p = 300$.

models, only a very small range of ω around 20 is optimal to variable selection. The performance out of such ranges drops quickly. This poses a potential issue to real data analysis when the best range of ω is unknown.

APPENDIX F

Chapter III: Application to the Ordinal Logistic Model

In addition to the linear, logistic and Poisson models presented in Chapter III, we outline the procedure for our method to be applied to fit the ordinal logistic regression model for large scale ranking problems, such as in online advertisement. This method application has been published in *Tang et al.* (2018), with more details provided within.

We adopt the method of reduction from ordinal ranking to binary classification given by a previous work (*Li and Lin*, 2007). Ordinal outcomes naturally inspire a binary classification approach for training models. As an example, consider the satisfaction level of a user for a product, with five possible levels. By asking the question "is the satisfaction level for the user greater than level k", one can get a binary classification problem for a fixed k, since the answer would be *yes* or *no* (1 or 0). By varying $k = 1, 2, 3, 4$, for each user, one can get 4 different binary classification problems. The approach then reduces to a question of how the classification models be trained and combined to obtain an ordinal ranking model. The main advantage of reducing ordinal ranking problem into binary classification problem is that it facilitates the usage of well-tuned binary classifiers available with standard libraries.

For a binary logistic regression, with instance $x \in \mathbb{R}^D$ and label $y \in \{0, 1\}$, the binary classifier $f(x)$ to be learnt is parameterized by $\beta \in \mathbb{R}^D$, i.e., $f(x) = x^T\beta$. The loss (or the negative log likelihood) function of a training dataset is

$$\textstyle\sum_{i=1}^{N} \left\{\log\left(1 + e^{f(x_i)}\right) - y_i f(x_i)\right\} \tag{F.1}$$

where N is the training sample size, and the estimated coefficient vector $\hat{\beta}$ is the minimizer to (F.1). A K class ordinal ranking problem is defined by an instance $x \in \mathcal{X} \subseteq \mathbb{R}^D$ and label $y \in \mathcal{Y} = \{1, 2, \ldots, K\}$, where $1 \leq 2 \leq \ldots \leq K$. The objective is to learn a ranking rule $r : \mathcal{X} \mapsto \mathcal{Y}$, which will minimize a weighted point-wise loss function with weights defined by some cost $C_{y,r(x)}$. Each instance and label pair (x_i, y_i) is reduced to a binary classification pair (along with introduction of a weight) by the following technique:

$$\begin{aligned} x_i^k =& (x_i^T, e_k^T)^T \in \mathbb{R}^{D+K-1}, \\ y_i^k =& 1[k < y], \\ w_i^k =& |C_{y_i,k} - C_{y_i,k+1}|, \end{aligned} \tag{F.2}$$

for $k = 1, \ldots, K-1$, where $C_{y,k}$ is the cost for assigning an outcome of k when the actually value is y, and e_k is the standard basis vector in dimension $K-1$. *As a result, the original sample size expands from N to $(K-1)N$.* Then, a logistic classifier $f(\cdot)$ can be trained on the expanded training set by minimizing the new loss function

$$\textstyle\sum_{i=1}^{N} \sum_{k=1}^{K-1} w_i^k \left\{\log\left(1 + e^{f(x_i^k)}\right) - y_i^k f(x_i^k)\right\}. \tag{F.3}$$

This can be viewed as the loss (negative log likelihood) of a training data with sample size $\tilde{N} = (K-1)N$, feature dimension $\tilde{D} = D + K - 1$, and sample weights specified by w_i^k. The solution to (F.3) would lead to a classifier $f(\cdot)$ of the form $f(\cdot) =$

$(g(\cdot), b_1, b_2, \ldots, b_{K-1})$, where g is defined by a parameter vector $\beta \in \mathbb{R}^D$ ($g(x) = x^T\beta \mapsto \mathbb{R}$) and $\{b_1, \ldots, b_{K-1}\}$ are bias terms. Thus, $f(\cdot)$ can be represented as a linear function with parameter $\theta \in \mathbb{R}^{\tilde{D}}$ as $\theta = [\beta, b_1, \ldots, b_{K-1}]^T$, with $f(x^k) = x^{kT}\theta = x^T\beta + b_k$. The authors guaranteed (Thm.2, *Li and Lin* (2007)) when $C_{y,r(x)}$ is convex, the bias terms are *rank monotone* such that $b_1 \geq b_2 \geq \cdots \geq b_{K-1}$, therefore $f(x^1) \geq f(x^2) \geq \cdots \geq f(x^{K-1})$. This justifies the ranking rule of predicting the ordinal class of a new instance $x_* \in \mathbb{R}^D$ by

$$r(x_*) = 1 + \textstyle\sum_{k=1}^{K-1} 1[f(x_*^k) > 0]. \tag{F.4}$$

Here, we consider the convex absolute loss $C_{y,r(x)} = |y - r(x)|$ in the reduction to binary classification to ensure the biases to be rank monotone as described by the authors. As a result, $w_i^k = 1$ for all i, k. As a result of the reduction, we have $K - 1$ times the sample size as before, leading to a massive amount of data. To speed up computation, we can now fit the logistic regression with parallelized MODAC algorithm to obtain the coefficients and biases, and easily convert them back to the corresponding parameters in the ordinal logistic model. This facilitates the usage of many readily available and powerful logistic classifiers, avoiding the trouble to directly parallelize ordinal logistic regression algorithms.

APPENDIX G

Chapter IV: Proofs

Regularity Conditions

We impose the following conditions:

C1 There exists a nonsingular matrix $\boldsymbol{A}$ such that for any given constant M,

$$\sup_{|\boldsymbol{D}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_*)|\leq MN^{-1/2}} |N^{-1/2}\boldsymbol{S}(\boldsymbol{D}^{-1}\boldsymbol{\theta})-N^{-1/2}\boldsymbol{S}(\boldsymbol{D}^{-1}\boldsymbol{\theta}_*)-N^{1/2}\boldsymbol{A}\boldsymbol{D}^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_*)| = o_p(1).$$

Furthermore, $N^{-1/2}\boldsymbol{S}(\boldsymbol{D}^{-1}\boldsymbol{\theta}_*) \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{V})$ for $\boldsymbol{V}$ a $pR \times pR$ matrix.

C2 The penalty function $q_{\lambda_N}(\cdot)$ has the following properties:

a. For nonzero fixed θ, $\lim N^{1/2}q_{\lambda_N}(|\theta|) = 0$ and $\lim q'_{\lambda_N}(|\theta|) = 0$.

b. For any $M > 0$, $\lim \sqrt{N} \inf_{|\theta|\leq MN^{-1/2}} q_{\lambda_N}(|\theta|) \to \infty$.

Condition C1 is satisfied by many commonly used estimating functions. Condition C2 is satisfied by several commonly used penalties with proper choices of the regularization parameters λ_N. Under the MCP penalty, that is, $q_{\lambda_N}(x) = \lambda_N \frac{(a\lambda_N - x)_+}{a\lambda_N}$, with $a > 1$, it is easy to see that if we choose $\lambda_N \to 0$ and $\sqrt{N}\lambda_N \to \infty$, then condition C2 holds.

Proof of Theorem IV.1

Proof. We provide a sketch for the proof. First, we assume $\boldsymbol{D}$ is known. Based on previous results, selection consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ in (4.6) can be derived following the similar arguments in Theorem 1 of *Johnson et al.* (2008) for $\hat{\boldsymbol{\theta}}$ estimated based on $\boldsymbol{D}$. Since $\lim_{n\to\infty} \boldsymbol{D}_N = \boldsymbol{D}$, following the proof of Theorem II.3, it can be shown that selection consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$ holds for $\hat{\boldsymbol{\theta}}$ estimated based on $\boldsymbol{D}_N$. Thus the results in Theorem IV.1 follows.

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Aho, A. V., and J. E. Hopcroft (1974), *Design & Analysis of Computer Algorithms*, Pearson Education India.

Battey, H., J. Fan, H. Liu, J. Lu, and Z. Zhu (2015), Distributed estimation and inference with statistical guarantees, *arXiv preprint arXiv:1509.05457*.

Bühlmann, P., and S. Van De Geer (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chang, F., J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber (2008), Bigtable: A distributed storage system for structured data, *ACM Transactions on Computer Systems (TOCS)*, *26*(2), 4.

Chen, H. Y., and R. Little (1999), A test of missing completely at random for generalised estimating equations with missing data, *Biometrika*, *86*(1), 1–13.

Chen, J., and Z. Chen (2008), Extended bayesian information criteria for model selection with large model spaces, *Biometrika*, *95*(3), 759–771.

Chen, X., and M. Xie (2014), A split-and-conquer approach for analysis of extraordinarily large data, *Statistica Sinica*, *24*, 1655–1684.

Cox, D. R. (1972), Regression models and life tables (with discussion), *Journal of the Royal Statistical Society*, *34*(2), 187–220.

Dean, J., and S. Ghemawat (2008), Mapreduce: simplified data processing on large clusters, *Communications of the ACM*, *51*(1), 107–113.

DerSimonian, R., and R. Kacker (2007), Random-effects model for meta-analysis of clinical trials: an update, *Contemporary Clinical Trials*, *28*(2), 105–114.

Diggle, P. J. (1989), Testing for random dropouts in repeated measurement data, *Biometrics*, *45*(4), 1255–1258.

Donoho, D. L., and J. M. Johnstone (1994), Ideal spatial adaptation by wavelet shrinkage, *Biometrika*, *81*(3), 425–455.

Efron, B. (1993), Bayes and likelihood calculations from confidence intervals, *Biometrika*, *80*(1), 3–26.

Ekholm, A., and C. Skinner (1998), The muscatine childrens obesity data reanalysed using pattern mixture models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *47*(2), 251–263.

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

Fan, J., F. Han, and H. Liu (2014), Challenges of big data analysis, *National science review*, *1*(2), 293–314.

Fisher, R. A. (1930), Inverse probability, in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 26, pp. 528–535, Cambridge University Press.

Fisher, R. A. (1956), *Statistical methods and scientific inference.*, Oxford, England: Hafner Publishing Co.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani (2007), Pathwise coordinate optimization, *The Annals of Applied Statistics*, *1*(2), 302–332.

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

Fu, W. J. (2003), Penalized estimating equations, *Biometrics*, *59*(1), 126–132.

Gao, X., and P. X.-K. Song (2010), Composite likelihood bayesian information criteria for model selection in high-dimensional data, *Journal of the American Statistical Association*, *105*(492), 1531–1540.

Glass, G. V. (1976), Primary, secondary, and meta-analysis of research, *Educational Researcher*, *5*(10), 3–8.

Golub, G. H., M. Heath, and G. Wahba (1979), Generalized cross-validation as a method for choosing a good ridge parameter, *Technometrics*, *21*(2), 215–223.

Guha, S., R. Hafen, P. Kidwell, and W. S. Cleveland (2009), Visualization databases for the analysis of large complex datasets, *Journal of Machine Learning Research*, *5*, 193–200.

Guha, S., R. Hafen, J. Rounds, J. Xia, J. Li, B. Xi, and W. S. Cleveland (2012), Large complex data: divide and recombine (d&r) with rhipe, *Stat*, *1*(1), 53–67.

Hall, P., J. S. Marron, and A. Neeman (2005), Geometric representation of high dimension, low sample size data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(3), 427–444.

Hansen, L. P. (1982), Large sample properties of generalized method of moments estimators, *Econometrica*, *50*(4), 1029–1054.

Hedges, L. V., and I. Olkin (2014), *Statistical methods for meta-analysis*, Academic press.

Hoaglin, D. C., and R. E. Welsch (1978), The hat matrix in regression and anova, *The American Statistician*, *32*(1), 17–22.

Hui, F. K., D. I. Warton, and S. D. Foster (2015), Tuning parameter selection for the adaptive lasso using eric, *Journal of the American Statistical Association*, *110*(509), 262–269.

Hunter, D. R., and R. Li (2005), Variable selection using mm algorithms, *The Annals of Statistics*, *33*(4), 1617.

Javanmard, A., and A. Montanari (2014), Confidence intervals and hypothesis testing for high-dimensional regression., *Journal of Machine Learning Research*, *15*(1), 2869–2909.

Johnson, B. A., D. Lin, and D. Zeng (2008), Penalized estimating functions and variable selection in semiparametric regression models, *Journal of the American Statistical Association*, *103*(482), 672–680.

Johnson, S. C. (1967), Hierarchical clustering schemes, *Psychometrika*, *32*(3), 241–254.

Jorgensen, B. (1997), *The theory of dispersion models*, CRC Press.

Ke, Z. T., J. Fan, and Y. Wu (2015), Homogeneity pursuit, *Journal of the American Statistical Association*, *110*(509), 175–194.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2014), A scalable bootstrap for massive data, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(4), 795–816.

Kroenke, K., R. L. Spitzer, and J. B. Williams (2001), The phq-9, *Journal of General Internal Medicine*, *16*(9), 606–613.

Leek, J. T., and J. D. Storey (2007), Capturing heterogeneity in gene expression studies by surrogate variable analysis, *PLoS Genetics*, *3*(9), 1724–1735.

Li, L., and H.-T. Lin (2007), Ordinal regression by extended binary classification, *Advances in neural information processing systems*, *19*, 865.

Liang, K.-Y., and S. L. Zeger (1986), Longitudinal data analysis using generalized linear models, *Biometrika*, *73*(1), 13–22.

Lin, D., and D. Zeng (2010), On the relative efficiency of using summary statistics versus individual-level data in meta-analysis, *Biometrika*, p. asq006.

Lin, H.-M., et al. (1998), Center-specific graft and patient survival rates: 1997 united network for organ sharing (unos) report, *Journal of the American Medical Association*, *280*(13), 1153–1160.

Lin, N., and R. Xi (2011), Aggregated estimating equation estimation, *Statistics and Its Interface*, *4*(1), 73–83.

Lipsitz, S. R., G. M. Fitzmaurice, G. Molenberghs, and L. P. Zhao (1997), Quantile regression methods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *46*(4), 463–476.

Little, R. J. (1993), Pattern-mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, *88*(421), 125–134.

Little, R. J. (1994), A class of pattern-mixture models for normal incomplete data, *Biometrika*, *81*(3), 471–483.

Little, R. J., and D. B. Rubin (1987), *Statistical analysis with missing data*, New York: Wiley.

Liu, D., R. Y. Liu, and M. Xie (2015), Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness, *Journal of the American Statistical Association*, *110*(509), 326–340.

Lohmueller, K. E., C. L. Pearce, M. Pike, E. S. Lander, and J. N. Hirschhorn (2003), Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease, *Nature Genetics*, *33*(2), 177–182.

Lumley, T., and A. Scott (2015), AIC and BIC for modeling with complex survey data, *Journal of Survey Statistics and Methodology*, *3*(1), 1–18.

Ma, S., and J. Huang (2017), A concave pairwise fusion approach to subgroup analysis, *Journal of the American Statistical Association*, *112*(517), 410–423.

Mackey, L. W., M. I. Jordan, and A. Talwalkar (2011), Divide-and-conquer matrix factorization, in *Advances in Neural Information Processing Systems*, pp. 1134–1142.

Madria, S., and T. Hara (2015), Big data analytics and knowledge discovery.

Mayer-Schönberger, V., and K. Cukier (2013), *Big data: A revolution that will transform how we live, work, and think*, Houghton Mifflin Harcourt.

Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2014), Scalable and robust bayesian inference via the median posterior., in *ICML*, pp. 1656–1664.

Palankar, M. R., A. Iamnitchi, M. Ripeanu, and S. Garfinkel (2008), Amazon s3 for science grids: a viable solution?, in *Proceedings of the 2008 international workshop on Data-aware distributed computing*, pp. 55–64, ACM.

Pan, W., X. Shen, and B. Liu (2013), Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty, *The Journal of Machine Learning Research*, *14*(1), 1865–1889.

Qin, J., and J. Lawless (1994), Empirical likelihood and general estimating equations, *The Annals of Statistics*, *22*(1), 300–325.

Qu, A., and P. X.-K. Song (2002), Testing ignorable missingness in estimating equation approaches for longitudinal data, *Biometrika*, *89*(4), 841–850.

Qu, A., G. Yi, P. X.-K. Song, and P. Wang (2011), Assessing the validity of weighted generalized estimating equations, *Biometrika*, *98*(1), 215–224.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995), Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, *90*(429), 106–121.

Rotenstein, L. S., M. A. Ramos, M. Torre, J. B. Segal, M. J. Peluso, C. Guille, S. Sen, and D. A. Mata (2016), Prevalence of depression, depressive symptoms, and suicidal ideation among medical students: a systematic review and meta-analysis, *Journal of the American Medical Association*, *316*(21), 2214–2236.

Rousseeuw, P. J., and A. M. Leroy (2005), *Robust regression and outlier detection*, vol. 589, John wiley & sons.

Rubin, D. B. (1976), Inference and missing data, *Biometrika*, *63*(3), 581–592.

Schernhammer, E. S., and G. A. Colditz (2004), Suicide rates among physicians: a quantitative and gender assessment (meta-analysis), *American Journal of Psychiatry*, *161*(12), 2295–2302.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*(2), 461–464.

Sen, S., H. R. Kranzler, J. H. Krystal, H. Speller, G. Chan, J. Gelernter, and C. Guille (2010), A prospective cohort study investigating factors associated with depression during medical internship, *Archives of General Psychiatry*, *67*(6), 557–565.

Shao, J., and X. Deng (2012), Estimation in high-dimensional linear models with deterministic design matrices, *The Annals of Statistics*, *40*(2), 812–831.

Shekelle, P. G., M. L. Hardy, S. C. Morton, M. Maglione, W. A. Mojica, M. J. Suttorp, S. L. Rhodes, L. Jungvig, and J. Gagné (2003), Efficacy and safety of ephedra and ephedrine for weight loss and athletic performance: a meta-analysis, *Journal of the American Medical Association*, *289*(12), 1537–1545.

Shen, X., and H.-C. Huang (2010), Grouping pursuit through a regularization solution surface, *Journal of the American Statistical Association*, *105*(490), 727–739.

Shin, S., J. Fine, and Y. Liu (2016), Adaptive estimation with partially overlapping models, *Statistica Sinica*, *26*(1), 235–253.

Shvachko, K., H. Kuang, S. Radia, and R. Chansler (2010), The hadoop distributed file system, in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*, pp. 1–10, IEEE.

Singh, K., M. Xie, and W. E. Strawderman (2005), Combining information from independent sources through confidence distributions, *The Annals of Statistics*, *33*(1), 159–183.

Song, P. X.-K. (2007), *Correlated data analysis: modeling, analytics, and applications*, New York: Springer.

Song, Q., and F. Liang (2015), A split-and-merge bayesian variable selection approach for ultrahigh dimensional regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *77*(5), 947–972.

Spitzer, R. L., K. Kroenke, J. B. Williams, and B. Löwe (2006), A brief measure for assessing generalized anxiety disorder: the gad-7, *Archives of Internal Medicine*, *166*(10), 1092–1097.

Srivastava, S., V. Cevher, Q. Tran-Dinh, and D. B. Dunson (2015), Wasp: Scalable bayes via barycenters of subset posteriors., in *AISTATS*.

Stangl, D., and D. A. Berry (2000), *Meta-analysis in medicine and health policy*, CRC Press.

Stein, L. D. (2010), The case for cloud computing in genome informatics, *Genome Biology*, *11*(5), 207.

Sullivan, P. F., M. C. Neale, and K. S. Kendler (2000), Genetic epidemiology of major depression: review and meta-analysis, *American Journal of Psychiatry*, *157*(10), 1552–1562.

Sutton, A. J., and J. Higgins (2008), Recent developments in meta-analysis, *Statistics in Medicine*, *27*(5), 625–650.

Tang, L., and P. X.-K. Song (2016), Fused lasso approach in regression coefficients clustering: learning parameter heterogeneity in data integration, *The Journal of Machine Learning Research*, *17*(1), 3915–3937.

Tang, L., L. Zhou, and P. X.-K. Song (2016), Method of divide-and-combine in regularised generalised linear models for big data, *arXiv preprint arXiv:1611.06208*.

Tang, L., S. Chaudhuri, A. Bagherjeiran, and L. Zhou (2018), Learning large scale ordinal ranking model via divide-and-conquer technique, in *Companion Proceedings of the The Web Conference 2018*, pp. 1901–1909, International World Wide Web Conferences Steering Committee.

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

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005), Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

U.S. Census Bureau (2015), Regions and divisions.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014), On asymptotically optimal confidence regions and tests for high-dimensional models, *The Annals of Statistics*, *42*(3), 1166–1202.

Wang, F., L. Wang, and P. X.-K. Song (2016), Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements, *Biometrics*, *72*(4), 1184–1193.

Wang, H., and C. Leng (2007), Unified lasso estimation by least squares approximation, *Journal of the American Statistical Association*, *102*(479), 1039–1048.

Wang, L., J. Zhou, and A. Qu (2012), Penalized generalized estimating equations for high-dimensional longitudinal data analysis, *Biometrics*, *68*(2), 353–360.

Watkins, D. J., M. M. Téllez-Rojo, K. K. Ferguson, J. M. Lee, M. Solano-Gonzalez, C. Blank-Goldenberg, K. E. Peterson, and J. D. Meeker (2014), In utero and peripubertal exposure to phthalates and bpa in relation to female sexual maturation, *Environmental Research*, *134*, 233–241.

Xie, M., and K. Singh (2013), Confidence distribution, the frequentist distribution estimator of a parameter: a review, *International Statistical Review*, *81*(1), 3–39.

Xie, M., K. Singh, and W. E. Strawderman (2012), Confidence distributions and a unifying framework for meta-analysis, *Journal of the American Statistical Association*, *106*(493), 320–333.

Yang, S., L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye (2012), Feature grouping and selection over an undirected graph, in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 922–930, ACM.

Ye, J. (1998), On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association*, *93*(441), 120–131.

Zaharia, M., M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica (2010), Spark: Cluster computing with working sets., *HotCloud*, *10*(10-10), 95.

Zeger, S. L., K.-Y. Liang, and P. S. Albert (1988), Models for longitudinal data: a generalized estimating equation approach, *Biometrics*, *44*(4), 1049–1060.

Zhang, C.-H. (2010), Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, *38*(2), 894–942.

Zhang, C.-H., and J. Huang (2008), The sparsity and bias of the lasso selection in high-dimensional linear regression, *The Annals of Statistics*, *36*(4), 1567–1594.

Zhang, C.-H., and S. S. Zhang (2014), Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(1), 217–242.

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

Zou, H., and T. Hastie (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.