L_0 Constraint Optimization, Homogeneity Fusion, and Mediation Analyses

by

Wen Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2020

Doctoral Committee:

Professor Peter X.K. Song, Chair
Professor John D. Kalbfleisch
Assistant Professor Gongjun Xu
Assistant Professor Ziwei Zhu

Wen Wang

wangwen@umich.edu

ORCID iD: 0000-0001-9509-7383

DEDICATION

To Shasta and Hendrix.

ACKNOWLEDGEMENTS

Foremost, I want to thank Department of Biostatistics of University of Michigan for giving me the opportunity and support to pursue doctoral degree. And I would like to express my sincere gratitude to my advisor Prof. Peter X.K. Song for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to express my deepest thanks to Prof. John D. Kalbfleisch for invaluable help, guidance and offering me the opportunity to work as a research assistant.

In addition, I would like to thank Prof. Alan Leichtman, Prof. Michael A. Rees, Prof. Ling Zhou, Prof. Gongjun Xu and Prof. Ziwei Zhu for their great help, insightful comments and encouragement.

Furthermore, I would love to thank Dr. Mathieu Bray, Dr. Yan Zhou, Wei Hao, Margaret Banker, Prof. Emily Hector, Yiwang Zhou, Prof. Luo Lan for their help and advice.

Last but not least, I would like to thank my family and my friend Wenjia Zhang for supporting me immensely.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

The focus of this dissertation is to develop a framework of L_0 regularized statistical procedures to identify subgroups among regression coefficients and estimation of subgroup-specific parameters. The proposed constrained discrete optimization methodology estimates group labels by solving mixed integer programming problem (MIP) via efficient algorithms. I develop key large-sample theory for the proposed methods, including subgroup selection consistency and estimation consistency using some new non-asymptotic bounds. Also, the R statistical software is made available to the public for the proposed methods.

In the first project presented in Chapter II, I consider a high-dimensional regression setting. The primary objective is to develop a dimension reduction method that can identify homogeneous subgroups among regression coefficients and sparse feature selection simultaneously. The resulting estimates of regression coefficients in each subgroup share the same value. To encourage sparsity, a large subgroup of coefficients is allowed to be estimated exactly as zero. To achieve this objective, I propose a new L_0 constrained optimization method, which is formulated as a MIP problem. To implement this MIP method, I develop a novel algorithm with warm start via both a discrete first-order method and segment neighborhood method, and establish its convergence properties. This new approach is able to solve the MIP problem with satisfactory accuracy in short time. To attain global optimality of the MIP method, I reformulate the constrained optimality as another MIP problem that can then be solved efficiently by Kelley's cutting plane method. A sufficient condition for

consistent estimation of group labels is affirmed, which is proved to be the necessary condition under which any method attains consistency of subgroup clustering up to a constant factor. Surprisingly, to achieve the clustering consistency, the sample size only needs to grow at the same rate as the sum of logarithm of the number of regression coefficients and the logarithm of the true number of subgroups. A real data analysis is used to illustrate the performance of the proposed method and algorithms. In the second project presented in Chapter III, I consider a structural equation model, and aim to estimate model parameters in causal mediation pathways in the presence of high-dimensional potential mediators. I develop statistical procedures to select sparse important mediators and to identify sparse causal pathways simultaneously. To address the technical challenge, I propose a new L_0 constrained optimization method, which leads to an MIP formulation. To solve this MIP problem, I develop a new warm start algorithm by using the discrete first-order method and establish its convergence properties. This new algorithm is able to quickly attain a near-optimal solution. To achieve the global optimality of the MIP problem, I reformulate it, so that I can solve this MIP problem efficiently using Kelley's cutting plane method. I present a sufficient condition for the proposed method for the selection consistency of causal pathways, which is proved as the necessary condition under which any method can achieve the causal pathway selection consistency up to a constant factor. Simulation studies and real world data analyses are used to demonstrate the performance of the proposed method and algorithms.

CHAPTER I

Introduction

1.1 Motivation

Constrained maximum likelihood (CML) is a methodology to enforce constraints on the parameter space when finding maximum likelihood estimators (*Schoenberg*, 1997). It utilizes prior knowledge as constraints to get more sensible estimators and generates data driven hypotheses for further testing. In CML, continuous constraints have been thoroughly studied and widely applied to a variety of fields (*Hathaway et al.*, 1985; *Molenberghs and Verbeke*, 2007). In contrast, discrete constraints are circumvented by continuous surrogates due to its NP-complete computational nature. Some famous continuous surrogates are Lasso (*Tibshirani*, 2011), truncated L_1 (*Shen et al.*, 2012) and SCAD (*Tibshirani*, 2011). Although the continuous surrogates are popular in practice, they face a lack of expressiveness when the discrete constraints they try to approximate become complicated. For example, to approximate a discrete constraint involving the basic "and" boolean operation or counting the number of discrete values in estimators, the continuous surrogates either exert strong assumptions (*Tibshirani et al.*, 2005; *Friedman et al.*, 2010; *Ke et al.*, 2015) or introduce too many penalty terms and tuning parameters (*Shen and Huang*, 2010). Fortunately, recent developments in optimization suggest that directly solving CML with discrete constraints is not as formidable as statistics community thinks (*Bertsimas et al.*, 2016, 2020).

In fact, the discrete constraint formulation of a high dimensional feature selection problem in (*Bertsimas et al.*, 2020) can often be solved faster than its Lasso counterpart. Inspired by this computational advance, my dissertation aims to solve CML with complicated discrete constraints.

1.2 Statistical Challenges

Complicated discrete constraints give rise to new difficulties in CML. This section makes a list of the complicated discrete constraints of interest and the entailed technical difficulties.

1.2.1 Homogeneity Pursuit in Coefficients in Regression

In high-dimensional regression, homogeneity in coefficients is pursued in addition to sparse feature selection to achieve dimension reduction. Homogeneity emerges when coefficients are divided into disjoint groups such that coefficients in each group are approximately the same. As a special case, sparsity enforces a large group fixed at zero. Moreover, homogeneity pursuit in coefficients is helpful in scientific discovery and gives rise to higher predictive performance (*Bondell and Reich*, 2008; *Shen and Huang*, 2010; *Ke et al.*, 2015; *Zhu et al.*, 2013; *Jeon et al.*, 2017). In the current literature, discretely constrained CML estimators for homogeneity pursuit are approximated by continuous surrogates of discrete constraints. The continuous surrogates are constructed by either penalizing distance between any two coefficients or penalizing distance between neighbouring coefficients based on additional assumptions. The former over-penalize the distance between different groups especially for large groups. The latter relies on high quality prior knowledge. In my dissertation, I proposed a new L_0 constraint formulation and analyze a new algorithm via modern optimization technique and segment neighborhood method to explore homogeneity in Chapter II.

1.2.2 Limit Number of Mediators in Exploratory Mediator Analysis

Structural equation models (SEM) play a central role in modeling causal pathways in the literature (*Hernán and Robins*; *Fritz and MacKinnon*, 2007; *Preacher*, 2015). As a special case, SEMs are applied in exploratory mediation analysis to identify true mediators among high-dimensional potential mediators, whose adjacent causal paths directed from exposure and to outcome are both associated with non-zero coefficients. Sparsity of selected mediators in estimator not only serves as a prior knowledge but also is crucial to generate data-driven hypotheses to test whether the sparse mediator assumption is true. In current literature, little attention is given to limit number of mediators in exploratory mediation. In *Serang et al.* (2017), exploratory mediation analysis is treated as feature selection and it limits the number of selected causal paths from exposure to potential mediators and from potential mediators to outcome in two separate steps. In *Derkach et al.* (2019), exposure is assumed to directly influence a group of latent factors, which are associated with both the outcome and a sparse subset of the potential mediators. The number of latent factors and the number of their associated potential mediators are limited by Lasso penalty. However, this method introduces strong assumptions on existence of latent factors. In my dissertation, I proposed a new L_0 constraint optimization method and analyze a new algorithm via modern optimization technique to do exploratory mediator analysis in Chapter III.

1.3 Summary of Objectives

With a focus on the key challenges presented above, I organized this dissertation as follows.

Aim 1: To develop an L_0 constraint formulation, analyze a new algorithm via modern optimization technique and derive conditions for grouping consistency to explore

homogeneity.

Aim 2: To devise a new L_0 constraint optimization method, create a new algorithm to solve the optimization problem and obtain conditions for mediator selection consistency to do exploratory mediator analysis.

Two projects are presented to address the above challenges, respectively, in Chapter II and Chapter III. More details on backgrounds, literature review, existing methodology and numerical illustrations can be found in the respective introduction sections of the two chapters.

CHAPTER II

Supervised Homogeneity Fusion in Regression Analysis

2.1 Introduction

Identifying homogeneous subgroups of regression coefficients sharing the same value has received increasing attention due to its flexibility of integrating with present biological knowledge for data analysis and its higher predictive performance. Moreover, the homogeneous groups naturally provides a structure that can be helpful in scientific discoveries. Regression under the homogeneity setting can be summarized into two types: subgroup analysis and grouping pursuit. For the former, homogeneity assumption is crucial to explore the individual attributes and account for the similarity among some individuals at the same time. See, *Ke et al.* (2016); *Shen and He* (2015); *Ma and Huang* (2017); *Lian et al.* (2017), among others; For the latter, the homogeneity lies among covariates. In this sense, covariates having similar effects are aggregated together to reduce estimation error and improve interpretation, especially in high-dimensional analysis. Related literature include *Bondell and Reich* (2008); *Shen and Huang* (2010); *Zhu et al.* (2013); *Ke et al.* (2015); *Jeon et al.* (2017). It is grouping pursuit that we shall focus on in this paper.

In environmental health sciences, forming exposure mixtures of toxic agents such as

endocrine disrupting compounds (EDCs) (e.g. PBA, phthalates and heavy metals) remains an unsolved problem. A analytic task of interest is to identify forms of toxic agents and evaluate their effects on human health outcomes. Consider a set of p toxicants denoted by, *say*, $X_1, \ldots, X_p$. We may consider a linear regression analysis $Y \sim \sum_{j=1}^{p} \beta_j X_j + Z\boldsymbol{\alpha}$ to evaluate effect of a mixture (i.e. a predictor) $A = \sum_{j=1}^{p} \beta_j X_j$ on outcome Y, where Z is a set of covariates. In this model, this mixture $\sum_{j=1}^{p} \beta_j X_j$ serves as a predictor, where coefficients β_j may take zero, so only a subset of the p toxicants is used in the configuration of X. In the current literature, principal component (PC) analysis approach has been the default choice of the method to derive A. This however, has never gained popularity in environmental health sciences due to an implicit cancellation among toxic agents. This is because a PC type mixture has both positive and negative loading coefficients, which are determined in an unsupervised learning fashion with no use of outcome Y. Thus, such PC-type mixtures often lack meaningful scientific interpretations. Alternatively, in practice scientists often consider a cumulative exposure by a sum of certain selected toxicants, called a sum-mixture. For example, SumDEHP is a sum of four phthalates, MECPP, MEOHP, MEHHP, and MEHP, which quantifies DEHP exposure from products such as PVC plastics used in food processing/packaging materials as well as building materials and medical devices *Schettler* (2006); *Kobrosly et al.* (2012); *Braun et al.* (2012); also see *Marsee et al.* (2006); *Marie et al.* (2015) for another sum-mixture SumAA that adds three extra phthalates MBP, MiBP, and MBzP on the sum-mixture SumDEHP. Technically, a sum-mixture takes a form of $\sum_{j=1}^{p} \beta_j X_j$, where coefficients β_j are binary, taking values of 0 or 1. The objective is to make an optimal selection of a subset of X_j's in the regression model. This paper is motivated by this scientific problem, and plans to solve it using an L_0 sparsity penalty on β_j's.

The new statistical method developed in this project will be applicable not only in our motivating example but in many other practical studies to answer important scientific

questions, which otherwise cannot be answered with existing methods. The salient feature of grouping pursuit is simultaneous estimation of grouping and sparseness structures. Now consider a linear model with homogeneity assumption as follows,

$$Y = \sum_{j=1}^{p} X_j \beta_j + \boldsymbol{Z}^T \boldsymbol{\alpha} + \varepsilon, \quad \beta_j \in \{0, \gamma_1, \gamma_2, \ldots, \gamma_K\}, j = 1, \ldots, p, \tag{2.1.1}$$

where ε follows from a normal distribution with mean 0 and variance σ^2, and the coefficient $\beta_j, j = 1, \ldots, p$ belongs to a set including 0 and K unknown different nonzero values $\gamma_k, k = 1, \ldots, K$. Here both $\boldsymbol{\alpha} = \{\alpha_j, j = 1, \ldots, q\}$, $\boldsymbol{\beta} = \{\beta_j, j = 1, \ldots, p\}$ and $\boldsymbol{\gamma} = \{\gamma_k, k = 1, \ldots, K\}$ are unknown parameters. The goal of this paper is to develop an approach to estimate $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ simultaneously. Clearly, a straightforward optimization problem can be written as follows,

$$\begin{aligned} &\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}} \sum_{i=1}^{n} (Y_i - X_{ij}\beta_j - \boldsymbol{Z}_i^T \boldsymbol{\alpha})^2, \\ \text{subject to:} \quad &\beta_j \in \{0, \gamma_1, \gamma_2, \ldots, \gamma_K\}, j = 1, \ldots, p, \\ &\|\boldsymbol{\beta}\|_0 \stackrel{def}{=} \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s, \end{aligned} \tag{2.1.2}$$

where the first constraint is used to control the degree of grouping and the second constraint is for the sparsity. Here, $K, s \geq 0$ are two tuning parameters. The optimizing problem without the grouping constraint degenerates to the well known best subset problem (*Miller*, 2002) with subset size s. The cardinality constraint makes it widely dismissed as being intractable as NP-hard (*Natarajan*, 1995). Instead of the best subset problem, the regularized counterpart (2.1.3), which is,

$$\sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta} - \boldsymbol{Z}_i^T \boldsymbol{\alpha})^2 + \lambda \|\boldsymbol{\beta}\|_0, \tag{2.1.3}$$

obtain considerable interests due to its computational merits. Different alternatives

have been proposed to overcome the computational problem, like LASSO (*Tibshirani*, 1996), SCAD (*Fan and Li*, 2001), MCP (*Zhang et al.*, 2010), among others. Recently, *Liu and Li* (2016) propose an efficient EM algorithm to approximate problem (2.1.3). However, as pointed out in (*Shen et al.*, 2013), the best subset problem and its regularized counterpart (2.1.3) may not be equivalent in their global minimizers due to the non-convex property. Moreover, tuning involves discrete parameters K in the best subset problem, which is easier than that for its regularized counterpart (2.1.3) with a continuous parameter $\lambda > 0$. This phenomenon has been also observed in (*Gu*, 1998). *Shen et al.* (2012) proposed to use a truncated L_1 function as a computational surrogate to approximate the L_0 function in the best subset problem, which however involves extra tuning parameters to control the approximate error. *Bertsimas et al.* (2016) first demonstrate that Mixed Integer Optimization (MIO) could be a tractable solution method for the best subset problem.

The best subset problem is only a simplified version of our problem (2.1.2). The solution of the optimization problem (2.1.2) gives us all distinctive values, $\boldsymbol{\gamma}$, sparseness structure, as well as corresponding subgroups of homogeneous predictors. As will be shown in both the theoretical results and the numerical results, recovering oracle estimator in the sense of simultaneous grouping pursuit and feature selection is more difficult than that of feature selection alone. There is a paucity of literature for guiding practice. *Shen and Huang* (2010) rewrote the problem as minimizing $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, where

$$S(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} X_{ij} \beta_j \right)^2 + \lambda_1 \sum_{j<j'} J_\tau(|\beta_j - \beta_{j'}|),$$

where λ_1 is a tuning parameter corresponding to K in problem (2.1.2) and $J_\tau(z) = \min(z\tau^{-1}, 1)$ is a surrogate of the L_0-function, with $\tau > 0$ being another tuning parameter. It is easy to see that sparsity was not considered in their work. In addition,

an additional tuning parameter τ was included to evaluate the approximation error between the truncated penalty with the exact condition in problem (2.1.2). Besides, the pairwise differences penalties usually lead to redundant comparisons and extra computational complexity. As an extension, *Zhu et al.* (2013) considered simultaneous grouping pursuit and feature selection by adding additional penalties on the individual coefficient, that is,

$$S(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j\right)^2 + \lambda_1 \sum_{(j,j')\in\mathcal{E}} J_\tau(||\beta_j| - |\beta_{j'}||) + \lambda_2 \sum_{j=1}^{p} J_\tau(|\beta_j|).$$

Zhu et al. (2013) extended the method proposed by (*Shen et al.*, 2012) to do pursuit grouping and feature selection simultaneously under a much stronger condition with a prior knowledge about the $\mathcal{E}$-net. Proper prior knowledge no doubt will reduce the computation burden and improve the estimation efficiency. However, as will shown later in the numerical studies, inappropriate knowledge of the $\mathcal{E}$-net usually causes biased estimation and then leads to wrong group structures. It is always challenging in practice to seek for a proper $\mathcal{E}$-net, which makes the method less appealing. To avoid exhaustive pairwise searching, *Ke et al.* (2015) proposed clustering algorithm in regression via data-driven segmentaion (CARDS). They used a preliminary estimate to determine "adjacent" coefficient pairs, thus their estimators depend on the initial ordering of the coefficients, which could be not reliable when the value of coefficients is small.

This article proposes a homogeneity fusion method to solve the problem (2.1.2) directly through MIO. Our main contributions are summarized in four folds: (a) To the best of our knowledge, it is the first time that we formulate the group pursuit as an MIO problem, which reduces the computation complexity to $O(KP)$ from $O(P^2)$ for the pairwise searching. This also provides a new framework for similar problems, like parameter merging in meta-analysis. (b)The proposed method shares the merit

of integer programming in the way that it obtains a global solution instead of a local solution. As we all known, those approximation surrogates, like CARDS in (*Ke et al.*, 2015), are usually only able to find a local solution. (c) Theoretically we establish both finite-sample misselection error bounds for homogeneity pursuit problem (2.1.2) and asymptotic normality for the estimates of parameters. *Zhu et al.* (2013) also considered the theoretical investigation of constrained L_0-version with strong conditions on the prior knowledge $\mathcal{E}$-net. Moreover, we give the necessary condition for the structure consistency. (d) Nevertheless, the global solution is paid at the price of more computation time. However, the difference convex (DC) programming used in (*Zhu et al.*, 2013) involves additional tuning parameters which requires extra workload. In our implementation procedure, following a similar spirit of (*Bertsimas et al.*, 2016), we provide a warm start algorithm to reduce the computational time significantly. Further, the convergence of the algorithm is theoretically assured.

2.2 Formulations for the Homogeneity Fusion

We present a brief overview of MIO, including the simply astonishing advances it has enjoyed in the last twenty-five years. We then introduced the proposed MIO formulations for the homogeneity fusion problem.

2.2.1 Brief Background on MIO

The general form of a Mixed Integer Quadratic Optimization (MIO) problem is given as follows:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^T \boldsymbol{Q} \boldsymbol{\alpha} + \boldsymbol{\alpha}^T \boldsymbol{a} \\ &\text{s.t. } \boldsymbol{A}\boldsymbol{\alpha} \leq \boldsymbol{b}, \\ &\qquad \alpha_i \in \{0,1\}, \quad i \in \mathcal{I}, \\ &\qquad \alpha_j \geq 0, \qquad j \notin \mathcal{I}, \end{aligned} \tag{2.2.1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m$, $\boldsymbol{A} \in \mathbb{R}^{k \times m}$, $\boldsymbol{b} \in \mathbb{R}^k$ and $\boldsymbol{Q} \in \mathbb{R}^{m \times m}$ (positive semidefinite) are the given parameters of the problem: the symbol "$\leq$" denotes element-wise inequalities and we optimize over $\boldsymbol{\alpha} \in \mathbb{R}^m$ containing both discrete ($\alpha_i, i \in \mathcal{I}$) and continuous ($\alpha_i, i \notin \mathcal{I}$) variables, with $\mathcal{I} \subset \{1, \cdots, m\}$.For background on MIO, see (*Bertsimas and Weismantel*, 2005; *Jünger and Reinelt*, 2013). Subclasses of MIO problems include convex quadratic optimization problems ($\mathcal{I} = \emptyset$), mixed integer ($\boldsymbol{Q} = \boldsymbol{0}_{m \times m}$) and linear optimization problems ($\mathcal{I} = \emptyset, \boldsymbol{Q} = \boldsymbol{0}_{m \times m}$). Some examples of modern integer optimization solvers include CPLEX, GLPK, MOSEK and GUROBI.

Cutting plane theory (*Dantzig et al.*, 1954; *Gouonr*, 1958; *Gomory*, 1960), disjunctive programming for branching rules (*Markowitz and Manne*, 1957; *Eastman*, 1958; *Land and Doig*, 1960), improved heuristic methods (*Berthold*, 2006), techniques for preprocessing MIOs (*Savelsbergh*, 1994), using linear optimization methods have all contributed greatly to the speed imptovements in MIO solvers. Branch-and-cut search is a complete procedure designed to find the optimal solution of a given problem instance or prove infeasibility thereof. See (*Cook et al.*, 1995) for a review on the history of branch-and-bound. Preprocessing, or presolving, means to transform a given problem instance into a different but equivalent problem instance that is hopefully easier to solve by the subsequently invoked solution algorithm. In contrast, the goal of

primal heuristics is to find good feasible solutions quickly. It is often sufficient in practice to provide a good solution whereas a proof of optimality may not even be computationally tractable. MIO solvers provide both feasible solutions as well as lower bounds to the optimal value. As the MIO solver progresses toward the optimal solution, the lower bounds improve and provide an increasingly better guarantee of suboptimality, which is especially useful if the MIO solver is stopped before reaching the global optimum. In contrast, heuristic methods do not prove such a certificate of suboptimality. The detailed introduction about the foregoing algorithms can be found in (*Wolsey*, 2008).

2.2.2 MIO Formulations For the Homogeneity Fusion

We first present a simple reformulation to problem (2.1.2) as a MIO problem:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta}} \sum_{i=1}^{n}(Y_i - X_{ij}\beta_j - \boldsymbol{Z}_i^T\boldsymbol{\alpha})^2, \qquad (2.2.2)\\
\text{subject to:}\quad &\eta_{jk} \in \{0,1\}, k = 0,\ldots,K; j = 1,\ldots,p,\\
&\eta_{jk}(\beta_j - \gamma_k) = 0, k = 0,\ldots,K; j = 1,\ldots,p,\\
&\gamma_k < \gamma_{k+1}, k = 1,\ldots,K-1,\ \ \gamma_0 = 0,\\
&\sum_{k=0}^{K}\eta_{jk} = 1, j = 1,\ldots,p,\\
&\sum_{j=1}^{p}\eta_{j0} \geq p - s,
\end{aligned}$$

where the number of groups K and the degree of sparsity s are two predetermined tuning parameters. $\{\eta_{jk}, j = 1,\ldots,p, k = 1,\ldots,K\}$ are 0/1 binary variables. With constraint $\eta_{jk}(\beta_j - \gamma_k) = 0$, η_{jk} actually is the group membership indicator in the sense that $\eta_{jk} = 1$ ($\eta_{jk} = 0$) represents the j-th covariate (not) belongs to the k-th group, that is, $\beta_j = \gamma_k$ ($\beta_j \neq \gamma_k$). The constraint $\eta_{jk}(\beta_j - \gamma_k) = 0$ is also called as specially ordered sets of type 1 (SOS-1) in (*Bertsimas and Weismantel*, 2005). The constraint

$\gamma_k < \gamma_{k+1}$ is for the identification issue such that $\{\gamma_k, k = 1, \ldots, K\}$ could be uniquely determined as long as the number of the groups K is given. $\gamma_0 = 0$ indicates the 0-th group in which all the members have exactly zero effects on the response. Each covariate X_j if and only if belongs to one group according to the condition $\sum_{k=1}^{K} \eta_{jk} = 1$. The constraint $\sum_{j=1}^{P} \eta_{j0} \geq p - s$ is for the sparsity assumption, which ensures the size of the 0-th group must be bigger than $p - s$. It is worthwhile to note that the final solution of the problem (2.2.2) could have K_f groups with $K_f \leq K$.

Any standard software can be used to solve the problem, such as, GUROBI. This problem can be easily extended to accommodate more reliable practical situations by incorporating biology prior informations. For example,

- Some covariates are known in advance that belong to the same group. Denote the index set to be $\mathcal{J}$, that is, $\beta_{j\in\mathcal{J}}$ are equal. Then this information can be formulated as using $\sum_{j\in\mathcal{J}} X_j$ instead of $\{X_j, j \in \mathcal{J}\}$ in the design matrix.

- According to their practical meanings, some covariates $\{X_j, j \in \mathcal{J}\}$ should not belong to the same group. Then the information can be written as $\sum_{j\in\mathcal{J}} \eta_{jk} \leq 1, k = 0, \ldots, K$. Further, if those covariates could only have the same zero effects, this information can be simply formulated as $\sum_{j\in\mathcal{J}} \eta_{jk} \leq 1, k = 1, \ldots, K$.

To build a connection to the conventional pairwise search approach, we transform

problem (2.2.2) into the following big-M formulation:

$$\begin{aligned}
\mathcal{O}_1 = \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\eta}} & \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j - \sum_{d=1}^{q} Z_{ij}^T\alpha_d)^2, \qquad (2.2.3)\\
\text{subject to:} \quad & \eta_{jk} \in \{0,1\}, k = 0,\ldots,K; j = 1,\ldots,p,\\
& \eta_{jk}(\beta_j - \gamma_k) = 0, k = 0,\ldots,K; j = 1,\ldots,p,\\
& \gamma_k < \gamma_{k+1}, k = 1,\ldots,K-1, \;\; \gamma_0 = 0,\\
& \underline{-M \le \gamma_k \le M, k = 1,\ldots,K,}\\
& \underline{-M \le \alpha_d \le M, d = 1,\ldots,q,}\\
& \sum_{k=0}^{K} \eta_{jk} = 1, j = 1,\ldots,p,\\
& \sum_{j=1}^{p} \eta_{j0} \ge p - s,
\end{aligned}$$

where M is a constant such that if $\boldsymbol{\theta}$ is a minimizer of problem (2.2.3), then $M \ge \|\hat{\boldsymbol{\theta}}\|_\infty$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, and $\|\boldsymbol{\theta}\|_\infty = \max\{\theta_j : 1 \le j \le q\}$ for $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_{p+q})^T$. Provided that M is choose to be sufficiently large, a solution to problem (2.2.3) will be a solution to problem (2.2.2). Clearly, M is not known a prior, and a small value of M might lead to a solution different from (2.2.2). The choice of M affects the strength of the formulation and is critical for obtaining solution quickly in practice. For formulation (2.2.3), the structure of the convex hull of its constraints is:

$$\begin{aligned}
&\text{Conv}\left(\left\{\boldsymbol{\theta} : \beta_j = \boldsymbol{\eta}_{j\cdot}^T\boldsymbol{\gamma}, |\gamma_j| \le M, \boldsymbol{\eta}_{j\cdot}^T\mathbf{1} = 1, j = 1,\ldots,p, |\alpha_d| \le M, d = 1,\ldots,q, \sum_{j=1}^{p}\eta_{j0} \ge p - s\right\}\right)\\
&\quad \subset \text{Conv}\left(\left\{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty \le M, \sum_{i\ne j}|\beta_i - \beta_j| \le C_1 M, \|\boldsymbol{\beta}\|_1 \le sM\right\}\right)\\
&\quad \subset \left\{\boldsymbol{\theta} : \sum_{i\ne j}|\beta_i - \beta_j| \le C_1 M, \|\boldsymbol{\beta}\|_1 \le sM\right\},
\end{aligned}$$

where $C_1 = [s(p-s)I(s < p/2) + p^2/4I(s \ge p/2)] + 2(1 - 1/K)s^2$. Thus, the mini-

mum of problem (2.2.3) is lower-bounded by the optimum objective value of both the following convex optimization problems:

$$\begin{aligned}
\mathcal{O}_2 = \min_{\boldsymbol{\theta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha}\|_2^2, \\
\text{subject to } \|\boldsymbol{\theta}\|_\infty \leq M, \sum_{i \neq j} |\beta_i - \beta_j| \leq C_1 M, \|\boldsymbol{\beta}\|_1 \leq sM, \qquad (2.2.4) \\
\mathcal{O}_3 = \min_{\boldsymbol{\theta}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha}\|_2^2, \\
\text{subject to} \sum_{i \neq j} |\beta_i - \beta_j| \leq C_1 M, \|\boldsymbol{\beta}\|_1 \leq sM, \qquad (2.2.5)
\end{aligned}$$

where (2.2.5) is the conventional pairwise fusion and LASSO in constrained form. This is a weaker relaxation than formula (2.2.4), which in addition to the pure ℓ_1 and pairwise ℓ_1 constraints on $\boldsymbol{\beta}$, has box constraints controlling the values of the θ_i's. Obviously, the following ordering exists: $\mathcal{O}_3 \leq \mathcal{O}_2 \leq \mathcal{O}_1$, with the inequalities being strict in most instances. In terms of approximating the optimal solution to problem (2.2.3), the MIO solver begins by first solving a continuous relaxation of problem (2.2.3). The pairwise fusion (2.2.5) is weaker than this root node relaxation. Moreover, MIO is typically able to significantly improve the quality of the root node solution as the MIO solver progresses toward the optimal solution.

2.2.3 Warm Start

As we said earlier, the constant bound M is not necessarily required, but if it is provided, it improves the strength of the MIO formulation. In other words, formulations with tightly specified bounds provide better lower bounds to the global optimization problem in a specified amount of time, when compared to a MIO formulation with loose bound specification. Following (*Bertsimas et al.*, 2016), we now describe a similar discrete first-order method to provide good upper bounds to problem (2.1.1). The solutions when supplied as a warm-start to the MIO formulation (2.2.2) are often improved by MIO, thereby leading to high quality solutions to problem (2.1.1) within

several minutes.

Consider the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\theta},\boldsymbol{\gamma}} g(\boldsymbol{\theta}), \\ &\text{s.t. } \beta_j \in \{\gamma_0, \gamma_1, \ldots, \gamma_K\}, j = 1, \ldots, p \\ &\|\boldsymbol{\beta}\|_0 \leq s, \end{aligned} \tag{2.2.6}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_K)^T$, and $g(\boldsymbol{\theta}) \geq C_2$ is convex and has Lipschitz continuous gradient: $\|\nabla g(\boldsymbol{\theta}) - \nabla g(\tilde{\boldsymbol{\theta}})\|_2 \leq l\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2$ with C_2 and l being some finite number and positive constant respectively. Denote the feasible region of $\boldsymbol{\theta}$ in Problem (2.2.6) as

$$\boldsymbol{\Theta}(K, s) = \{\boldsymbol{\theta} : \beta_j \in \{0, \gamma_1, \ldots, \gamma_K\}, j = 1, \ldots, p; \|\boldsymbol{\beta}\|_0 \leq s\}.$$

Obviously, $\boldsymbol{\Theta}(K, s)$ is a closed set. Let $H_{K,s}(\boldsymbol{c})$ be the set of optimal solutions to problem (2.2.6) with $g(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{c}\|_2^2$ and some constant vector $\boldsymbol{c} = (c_1, \ldots, c_{q+p})^T$. Clearly, for any solution, say $(\hat{\boldsymbol{\theta}}^T, \hat{\boldsymbol{\gamma}}^T)^T \in H_{K,s}(\boldsymbol{c})$, we have $\hat{\boldsymbol{\gamma}} = \{\text{all different values in } \hat{\boldsymbol{\beta}}\}$. That is, $\hat{\boldsymbol{\gamma}}$ could be uniquely determined by the values of $\hat{\boldsymbol{\beta}}$. For ease of presentation, we generally omit parameters $\hat{\gamma}$, and say $\hat{\boldsymbol{\theta}} \in H_{K,s}(c)$ if $\hat{\boldsymbol{\theta}}$ is the solution minimizing the problem (2.2.6), except where necessary. Denote the index operator for the elements of $\boldsymbol{\beta}$ with value r by $\mathcal{G}(\boldsymbol{\beta}; r) = \{j : \beta_j = r, j = 1, \ldots, p\}$ and the grouping operator by $\mathcal{G}(\boldsymbol{\beta}) = \{\mathcal{G}(\boldsymbol{\beta}; r) \,|\, r \neq 0, \mathcal{G}(\boldsymbol{\beta}; r) \neq \emptyset\}$. Let $|\mathcal{G}(\boldsymbol{\beta}; r)|$ and $|\mathcal{G}(\boldsymbol{\beta})|$ be the cardinality of $\mathcal{G}(\boldsymbol{\beta}; r)$ and $\mathcal{G}(\boldsymbol{\beta})$, respectively.

Definition II.1. (first-order stationary point). Given problem (2.2.6) and certain positive constant $L \geq l$, a vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}(K, s)$ is said to be a first-order stationary point if $\boldsymbol{\theta} \in H_{K,s}(\boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta}))$.

We first show that the optimal solution set has only one element $\boldsymbol{\theta}$ who satisfies a

first-order stationary point.

Proposition II.2. *Suppose a positive constant $L > l$,*

1. *if $\boldsymbol{\theta}$ is a solution to problem* (2.2.6)*, then it is a first-order stationary point.*
2. *if $\boldsymbol{\theta}$ satisfies a first-order stationary point, then the set $H_{K,s}(\boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta}))$ has exactly one element $\boldsymbol{\theta}$.*

Next we present Algorithm 1 to find a feasible point whose objective function value is the same as a first-order stationary point of problem (2.2.6).

Algorithm 1.

Input: $g(\boldsymbol{\theta})$, number of groups: K, sparsity constraint: s, parameter: L and convergence tolerance: ε.

Output: A feasible point $\boldsymbol{\theta}^*$ such that $g(\boldsymbol{\theta}^*) = g(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a first-order stationary point.

1. Initialize with $\boldsymbol{\theta}_1 \in \mathbb{R}^{p+q}$.
2. For $m \geq 1$, $\boldsymbol{\theta}_{m+1} \in H_{K,s}(\boldsymbol{\theta}_m - \frac{1}{L}\nabla g(\boldsymbol{\theta}_m))$.
3. Repeat Step 2, until $g(\boldsymbol{\theta}_m) - g(\boldsymbol{\theta}_{m+1}) \leq \varepsilon$.
4. Return $\boldsymbol{\theta}_{m+1}$.

To obtain an element in $H_{K,s}(c)$ in Step 2, we give the subroutine Algorithm 0 in Appendix A.1.1, which is a generalization of segment neighbourhood method (*Auger and Lawrence*, 1989).

Now we describe the asymptotic convergence property and convergence rate of Algorithm 1 through Proposition II.3 and Theorem (2.2.1), respectively.

Proposition II.3. *For problem* (2.2.6) *and some positive constant $L > l$, let $\boldsymbol{\theta}_m, m \geq 1$ be the sequence generated by Algorithm 1, we have*

1. $g(\boldsymbol{\theta}_m) - g(\boldsymbol{\theta}_{m+1}) \geq \frac{L-l}{2}\|\boldsymbol{\theta}_m - \boldsymbol{\theta}_{m+1}\|_2^2$.

2. $\|\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m\|_2 \to 0$ *as* $m \to \infty$.

Theorem 2.2.1. *Let some constant* $L > l$, *and* $\boldsymbol{\theta}^*$ *denote a first-order stationary point of Algorithm 1. After M iterations, Algorithm 1 satisfies*

$$\min_{m=1,\ldots,M} \|\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m\|_2^2 \leq \frac{2(g(\boldsymbol{\theta}_1) - g(\boldsymbol{\theta}^*)}{M(L-l)},$$

where $g(\boldsymbol{\theta}_m) \downarrow g(\boldsymbol{\theta}^*)$ *as* $m \to \infty$.

Finally, we show that Algorithm 1 outputs a point whose objective value is the same as some first-order stationary point under mild conditions:

Proposition II.4. *Consider problem* (2.2.6) *and some constant* $L > l$, *let* $\boldsymbol{\theta}_m, m \geq 1$ *be the sequence generated by Algorithm 1. If*

1. g *has second order derivative and,*

2. *there exists* $l' > 0$ *such that* $l' \left\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\right\|_2 \leq \left\|\nabla g(\boldsymbol{\theta}) - \nabla g(\tilde{\boldsymbol{\theta}})\right\|_2$ *for any* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta(K, s)$ *satisfying* $\mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}(\tilde{\boldsymbol{\beta}})$ *and,*

3. $\{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K, s) \,|\, g(\boldsymbol{\theta}) \leq C\}$ *is bounded for any* $C \in \mathbb{R}$,

then $g(\boldsymbol{\theta}_m)$ *converges to* $g(\boldsymbol{\theta})$ *where* $\boldsymbol{\theta}$ *is a first-order stationary point.*

The detailed proof of Proposition II.4 is in Proposition A.1 and Remark A.1.1 in Appendix A.1.2.

2.3 Theoretical Investigation

With regard to simultaneous grouping pursuit and feature selection, in this section we will prove the global minimizers of problem (2.1.2) reconstruct the ideal "oracle estimator " as if the true grouping were available in advance, under a "degree-of-separation" condition. To understand how the proposed method performs in a

high-dimensional situation, it is imperative that we study necessary and sufficient conditions for achieving grouping pursuit consistency as well as feature selection consistency.

2.3.1 A "Degree of Separation" Measure

Throughout this section, we write the $n \times p$ design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ and the $n \times q$ matrix $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_q)$, where $\boldsymbol{x}_j$ and $\boldsymbol{z}_d$ are the jth and dth columns of $\boldsymbol{X}$ and $\boldsymbol{Z}$, respectively. Define a_0 to be the true value of a. For example, $\boldsymbol{\theta}_0$ is the true parameter of $\boldsymbol{\theta}$.

We first define a distance between two groupings corresponding to estimators $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ denoted by $d(\boldsymbol{\beta}, \boldsymbol{\beta}')$:

Definition II.5. (distance between groupings)

$$d(\boldsymbol{\beta}, \boldsymbol{\beta}') = \max\left\{ \min_{f \in \mathcal{F}} \left| \cup_{w \in \mathcal{G}(\boldsymbol{\beta}')} w \setminus \cup_{w \in \mathcal{G}(\boldsymbol{\beta})} \left(w \cap f(w) \right) \right|, 1 \right\}, \tag{2.3.1}$$

where $|\mathcal{G}(\boldsymbol{\beta})| \leq |\mathcal{G}(\boldsymbol{\beta}')|$, $\mathcal{F} = \{\text{all injective functions } f : \mathcal{G}(\boldsymbol{\beta}) \to \mathcal{G}(\boldsymbol{\beta}')\}$, and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}(K, s)$ with $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ and $\boldsymbol{\theta}' = (\boldsymbol{\alpha}'^T, \boldsymbol{\beta}'^T)^T$.

This distance is the least number of modifications to the selected features in $\boldsymbol{\beta}'$ involved in changing $\mathcal{G}(\boldsymbol{\beta}')$ into $\mathcal{G}(\boldsymbol{\beta})$, where each modification is changing the membership of some $\boldsymbol{\beta}_j$ to some other group.

Then we can define a measure of easiness level for simultaneous grouping pursuit and feature selection as follows:

Definition II.6. (degree of separation)

$$C_{\min} \equiv C_{\min}(\boldsymbol{\theta}_0, \boldsymbol{X}, \boldsymbol{Z}, s) = \min_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K_0, s) \\ \mathcal{G}(\boldsymbol{\beta}) \neq \mathcal{G}(\boldsymbol{\beta}_0)}} \frac{\|\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \boldsymbol{Z}(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\|_2^2}{n \max\left(d(\boldsymbol{\beta}, \boldsymbol{\beta}_0), 1\right)}. \tag{2.3.2}$$

Here, $C_{\min}$ measures the degree of separation between the true signal and the estimated signals based on wrong groupings. More specifically, it is the least difference between the true signal and an estimated signal based on a wrong grouping per distance between the true grouping and the wrong grouping. If $C_{\min}$ is small, then estimating the true grouping is difficult due to the estimated signals based on some wrong grouping is very similar to the true signal. Thus $C_{\min}$ characterizes the easiness level of the underlying problem.

2.3.2 Necessary Condition

We first characterize consistent grouping pursuit and feature selection for any method through one simple necessary condition in the L_2-metric, which is sufficient up to a constant factor. By the necessary condition in Theorem 1 of (*Shen et al.*, 2013), the necessary condition for feature selection alone requires that

$$C_{\min} \geq d_1 \frac{\log p}{n} \sigma^2, \text{ as } n, p \to \infty,$$

for some positive constant $d_1 \leq 1/4$ that may dependent on $\boldsymbol{X}$. In short, the minimal degree of separation is required for correct identification of informative features, translating to an upper bound on p that is in an order of $\exp(nC_{\min}/(d_1\sigma^2))$, for any method and $(\boldsymbol{\beta}_0, \boldsymbol{X})$.

As pointed out in (*Zhu et al.*, 2013), the problem of recovering oracle estimator in the sense of simultaneous grouping pursuit and feature selection is more difficult than that of feature selection alone. To derive a lower bound requirement for $C_{\min}$, we construct an approximate least favorable situation under P, over $B_0(K, s, \ell) = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \boldsymbol{\Theta}(K, s), C_{\min}(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, s) \geq \ell\}$ to avoid super-efficiency.

Theorem 2.3.1. *For any $K \geq 2$, $p \geq s \geq K - 1, \ell > 0$, for any estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$,*

we have

$$\sup_{\boldsymbol{\theta}_0 \in B_0(K,s,\ell)} \mathbb{P}(\mathcal{G}(\hat{\boldsymbol{\beta}}) \neq \mathcal{G}(\boldsymbol{\beta}_0)) \to 0, \ \ as \ \ n, p \to \infty,$$

implying that

$$\ell \geq \frac{\sigma^2}{r(\boldsymbol{X}, \boldsymbol{Z}, K, s)} \frac{\log(pK)}{2n},$$

where $r(\boldsymbol{X}, \boldsymbol{Z}, K, s) = \frac{\max_{1 \leq j \leq p} n^{-1} \|\boldsymbol{x}_j\|_2^2}{\min\limits_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K,s) \\ |\beta_j - \beta_{j'}| \geq 1/K, \forall \beta_j \neq \beta_{j'}}} C_{\min}(\boldsymbol{\theta}, \boldsymbol{X}, \boldsymbol{Z}, s)}$.

The detailed proof of Theorem III.16 is in Appendix A.1.3. Theorem III.16 shows that the necessary condition of uniformly attaining grouping consistency in simultaneous homogeneity pursuit and feature selection for a collection of easy problems requires a lower bound of the level of easiness of those problems. More specifically, it requires

$$C_{\min} \geq d_2 \frac{\log(pK)}{n} \sigma^2, \tag{2.3.3}$$

for some constant d_2 that may dependent on $(\boldsymbol{X}, \boldsymbol{Z})$.

2.3.3 Sufficient Condition

Given $\mathcal{G}(\boldsymbol{\beta})$, define $\boldsymbol{X}_{\mathcal{G}(\boldsymbol{\beta})}$ as $(\sum_{k \in \mathcal{G}(\boldsymbol{\beta}; \gamma_1)} x_k, \ldots, \sum_{k \in \mathcal{G}(\boldsymbol{\beta}; \gamma_K)} x_k)$ to be a collapsed matrix by summing columns of $\boldsymbol{X}$ according to $\mathcal{G}(\boldsymbol{\beta})$. Given $B = (i_1, \ldots, i_{|B|}) \in \mathcal{I}$, where $i_1 < \cdots < i_{|B|}$, define $\boldsymbol{X}_B$ as $(X_{i_1}, \ldots, X_{i_{|B|}})$ to be a submatrix of $\boldsymbol{X}$ and $\boldsymbol{\beta}_B$ to be vector $(\beta_{i_1}, \ldots, \beta_{i_{|B|}})$ for any $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T \in \boldsymbol{\Theta}(K, s)$.

Definition II.7. (Oracle estimator). Given the true coefficient $\boldsymbol{\beta}_0$, the oracle estimator $\hat{\boldsymbol{\theta}}^{ol} = (\hat{\boldsymbol{\alpha}}^{ol,T}, \hat{\boldsymbol{\beta}}^{ol,T})^T$ is defined as

$$\underset{\mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}(\boldsymbol{\beta}_0)}{\arg\min} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha} \right\|_2^2.$$

More specifically, in $\hat{\boldsymbol{\beta}}^{ol} = (\hat{\beta}_1^{ol}, \ldots, \hat{\beta}_p^{ol})^T$, $\hat{\beta}_j^{ol}$ is $\hat{\gamma}_k$ if $j \in \mathcal{G}(\boldsymbol{\beta}_0; \gamma_{0,k})$; $k = 1, \ldots, K_0$,

and $\hat{\beta}_j^{ol}$ is 0 if $j \in \mathcal{G}(\boldsymbol{\beta}_0; 0)$, where

$$(\hat{\boldsymbol{\gamma}}^T, \hat{\boldsymbol{\alpha}}^T) = (\hat{\gamma}_1, \ldots, \hat{\gamma}_{K_0}, \hat{\boldsymbol{\alpha}}^T) = \underset{(\boldsymbol{\gamma}^T, \boldsymbol{\alpha}^T) \in \mathbb{R}^{K_0+q}}{\arg\min} \left\| \boldsymbol{Y} - \boldsymbol{X}_{\mathcal{G}(\boldsymbol{\beta}_0)} \boldsymbol{\gamma} - \boldsymbol{Z}^T \boldsymbol{\alpha} \right\|_2^2.$$

We now derive a nonasymptotic probability error bound for simultaneous grouping pursuit and feature selection, based on which we prove the oracle estimator. The next theorem says that a global minimizer of problem (2.2.2) consistently reconstructs the oracle estimator at a degree of separation level that is slightly higher than the minimal in Theorem III.16. Without loss of generality, assume that a global minimizer of (2.2.2) exists.

Denote the solution to problem (2.2.2) as $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T$.

Theorem 2.3.2. *In 2.2.2, when $K = K_0$ and $s = s_0$, we have that*

$$\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \leq \left(\frac{2}{1 - e^{-2/3}} + 1 \right) \exp\left\{ -\frac{n}{18\sigma^2} \left(C_{\min} - 36\sigma^2 \frac{\log(pK_0) + 1 - \frac{\log 4}{2}}{n} \right) \right\},$$

which implies that when $C_{\min} > 36\sigma^2 \frac{\log(pK_0)+1-\frac{\log 4}{2}}{n}$, $\hat{\boldsymbol{\theta}}$ consistently reconstructs $\hat{\boldsymbol{\theta}}^{ol}$, i.e., as $n, p, K_0 \to \infty$, $\mathbb{P}(\mathcal{G}(\hat{\boldsymbol{\theta}}) \neq \mathcal{G}(\hat{\boldsymbol{\theta}}^{ol})) \to 0$.

Theorem III.18 says that $\hat{\boldsymbol{\theta}}$ consistently reconstructs the oracle estimator $\hat{\boldsymbol{\theta}}^{ol}$ as long as the degree-of-separation condition is satisfied, which is,

$$C_{\min} \geq d_3 \frac{\log(pK_0) + 1 - \frac{\log 4}{2}}{n} \sigma^2, \tag{2.3.4}$$

where $d_3 > 36$ is a constant. The lower bound of $C_{\min}$ in necessary condition (3.4.1) and that in sufficient condition (3.4.2) they are at the same order.

2.4 Real Data Analysis

The proposed method is used to analyze data from the Early Life Exposure in Mexico to ENvironmental Toxicants (ELEMENT) project. The ELEMENT project consists of four mother-child pregnancy and birth cohorts originally initiated in the mid-1990s to explore early life chemical exposures and developmental outcomes. In this analysis, we focus on two cohorts of infants whose mothers' exposure to phthaltes has been measured. We are interested in investigating how prenatal exposures to mixtures (or groups) of phthalates, which are known endocrine disruptive compounds measured by mother's blood biospecimen during pregnancy, may affect infant's length at birth. It is know that phthalates are naturally grouped according to types of chemicals (i.e. plasticizers) used as additives in plastics to make products more resilient, cosmetic products (e.g. lotions and perfumes) and other products (e.g. food processing equipment, adhesives, and rainwear). There are 191 mother-child pairs in the data and 10 phthalates measured for each mother across three trimesters, respectively, which are log-transformed. Additionally, we standardize log-transformed phthalates, and adjust for confounding factors including a binary cohort indicator, the total years of school in mother's education, gender of child and mother's gestational age in the model. Based on 10-fold cross validation as shown in Figure 2.1 (the lightest cell with $\lambda = 4$), the chosen model contains two mixtures (or groups) of 4 phthalates measured during the first trimester of pregnancy, each with two phthalates; see Table 2.1. One phthalate mixture consists of MBP and MEHHP with associated coefficient 0.23, and the other mixture contains MBzP and MEOHP with coefficient -0.28. The first mixture does not appear to be statistically significant, while the second mixture has a negative association with birth length. These p-values are obtained by refitting the regression model after these mixtures have been identified. It suggests that higher exposure to the second mixture the shorter the length an infant has at birth. In addition, both sex and gestational age are both statistically significant; girls tend to be taller at

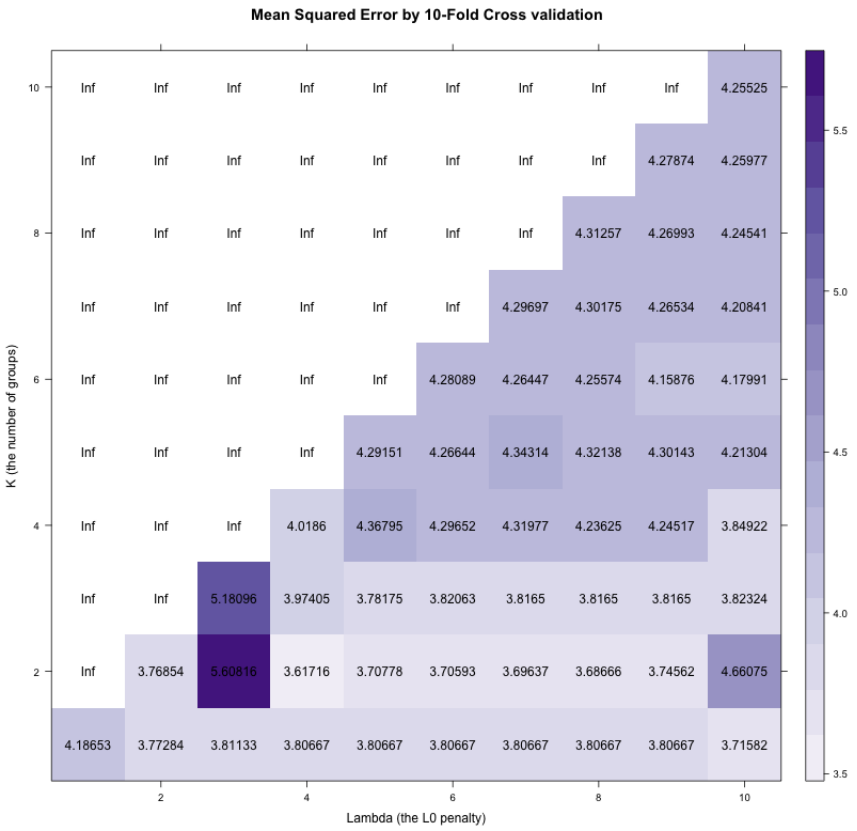

Figure 2.1: 10-fold cross-validation suggested two mixtures ($K = 2$) determined by the lightest cell with $\lambda = 4$.

birth, and the longer the pregnancy duration the talker an infant at birth.

2.5 Discussion

This paper developed a dimension reduction method that can identify homogeneous subgroups among regression coefficients and sparse feature selection simultaneously in the high-dimensional regression setting. We invoked the L_0 constrained optimization MIP method whose implementation was carried out by well-chosen initial values and then the Kelley's cutting plane method to search for optimal solutions. We established both algorithmic convergence properties as well as clustering and estimation consistency. We showed that to achieve the clustering consistency, the sample size

Table 2.1: The estimation results from the association study of birth length and prenatal exposure to PBA and phthalates during the first trimester of pregnancy.

	estimate	s.e.	lower	upper	p-value	mixture member
intercept	35.40	3.43	28.63	42.17	0.00	
cohort	-0.71	0.39	-1.48	0.07	0.07	
sex	0.82	0.27	0.28	1.35	0.00	
school year	0.05	0.05	-0.04	0.15	0.26	
gestage	0.36	0.09	0.18	0.53	0.00	
mixture1	0.23	0.13	-0.04	0.49	0.09	MBP,MEHHP
mixture2	-0.28	0.13	-0.54	-0.02	0.04	MBzP,MEOHP

only needs to grow at the same rate as the sum of logarithmic of the number of regression coefficients and logarithmic of the true number of subgroups. Simulation studies are used to illustrate the performance of the proposed method and algorithms.

The proposed upper bound in the search of warm start may be generalized to a general convex objective function with little effort. However, generalization of the lower bound to a setting beyond the least square objective function is not that straightforward due to the complexity of the kernel function used in the formulation. This is worth a future exploration. Also, due to the use of the ridge penalty in the formation of the lower bound problem, the resulting lower bound solution is just an approximate to the original optimization problem, but this approximation can be asymptotically diminished if the tuning parameter is chosen in the order of $o(1/n)$, under which the clustering consistency is warrant.

Obviously this new group fusion method may be extended to the framework of generalized linear models where iterative procedures used in the parameter estimation are essentially relied on weighted least squares objective functions. Thus, this extension is technically manageable but it may require substantial computational effort. Also, we would consider an extension of this method to the setting of estimating equations, which could cover a broad range of important statistical models, such as GEE regression, Cox regression and quantile regression.

CHAPTER III

L_0 Regularized Selection and Estimation of High-dimensional Mediators in Structural Equation Models

3.1 Introduction

In this paper, we consider the structural equation models (SEM) that have played a central role in modeling causal pathways in the literature (*Hernán and Robins*; *Fritz and MacKinnon*, 2007; *Preacher*, 2015). This is because such model provides a representation of a causal graphical model and natural interpretation of both direct and indirect effects of exposure variables on outcomes, which are the key estimates to explain causality (*MacKinnon and Dwyer*, 1993; *MacKinnon et al.*, 1995). However, this methodological framework is greatly challenged by high-dimensional mediators, especially in the case where the sample size n is smaller than the number of mediators p. For simplicity, we assume the dimension of exposure variables is high. Thus, new statistical methods are called for to solve this large p small n problem. Technically speaking, in order to develop a viable solution to this challenge, it is necessary to invoke regularization methods that enable us to identify a handful of important mediators predominantly influencing the underlying causal pathways.

For ease of exposition, let us first introduce the SEM whose graphic representation is

shown in Figure 3.1.

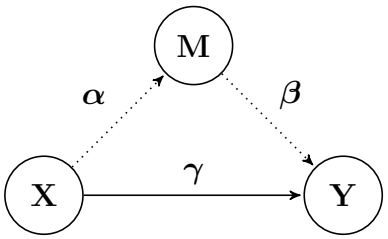

Figure 3.1: A graphic illustration of the causal pathways among exposure X, outcome Y and mediators M. Vertices indicate variables and edges indicate causal paths directed from causes to effects. Dotted edges indicates possible causal paths whose existence is to be determined in the analysis.

In the context of a mediation analysis, a primary task is to assess which mediators from the collection $\mathbf{M} = (\mathbf{M}_1, \ldots, \mathbf{M}_p)^\top$ alter the causal relationships between exposure variables $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_q)^\top$ and outcome variables $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_m)^\top$. In this paper, we focus on a setting with a large number of potential mediators or a large p, but only a handful of them are the true mediators. Analytically such a graphical model in Figure 3.1 may be formulated by a structural equation model of the following form:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{M} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\alpha} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\gamma} & \boldsymbol{\beta} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{M} \\ \mathbf{Y} \end{pmatrix} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is a $(q + p + m) \times 1$-dimensional multivariate normal (MVN) random vector following $\text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_p^\top)^\top, \boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_m^\top)^\top$ are $p \times q$, $m \times p$ and $m \times q$ unknown parameter matrices, respectively. Interpretations of these model parameters have been discussed extensively in the literature. For example, a nonzero coefficient $\boldsymbol{\alpha}_i$ indicates a causal relationship from exposure variables $\mathbf{X}$ to the mediator $\mathbf{M}_i$, while a nonzero coefficient $\boldsymbol{\beta}_i$ suggests a causal link from the mediator $\mathbf{M}_i$ to outcome Y. More importantly, if

both $\boldsymbol{\alpha}_i \neq \mathbf{0}$ and $\boldsymbol{\beta}_i \neq \mathbf{0}$ hold simultaneously, then $\mathbf{M}_i$ is a mediator of interest responsible for a so-called causal pathway. Configurations with the parameter sparsity define causal subgroups of mediators, leading to different scientific understandings of causality and interpretations. This subgroup topology is the core of knowledge that we aim to attain from the available data. To achieve this, we need to overcome the key technical difficulty that pertains to the need of two simultaneous regularization procedures, one concerning the mediator selection and the other relating to estimation for causal effects of important mediators.

To facilitate the subsequent discussion, we rewrite the above model as follows via the operation of Kronecker's matrix product:

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{M} \\ \mathbf{Y} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{I}_p \otimes \mathbf{X}^\top & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{M}^\top \otimes \boldsymbol{I}_m & \boldsymbol{I}_m \otimes \mathbf{X}^\top \end{pmatrix} (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)^\top + \boldsymbol{\varepsilon}, \tag{3.1.1}$$

where $\otimes$ denotes the Kronecker's matrix product (*Van Loan*, 2000). It is worth mentioning that in this paper, without loss of generality, we assume the covariance matrix $\boldsymbol{\Sigma}$ to be known; otherwise, we may use a consistently estimated version of the matrix from the residuals obtained by the proposed method with independent errors. Then, we may apply a de-association transformation by multiplying $\boldsymbol{\Sigma}^{-1/2}$ to the left of both sides of the above model, we obtain a transformed model:

$$\mathbf{V} = \mathcal{D}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{3.1.2}$$

where $\mathbf{V}$ is a $(q+p+m) \times 1$ response vector, $\mathcal{D}$ is a $(q+p+m) \times (qp+pm+qm)$ design matrix, $\boldsymbol{\theta}$ is $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p, \boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_p^\top, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_m)^\top$ and $\boldsymbol{\epsilon}$ is a $(q+p+m) \times 1$ noise vector following MVN$(\mathbf{0}, \boldsymbol{I})$. This is the SEM model that will be used in the

rest of this paper for the proposed regularization method and algorithms. With a slight abuse of notation, we may refer to the forms of $\mathbf{V}$, $\mathcal{D}$ and $\boldsymbol{\epsilon}$ of equation (3.1.2) as those corresponding terms in equation (3.1.1). Given of a dataset of n independent samples, we stack n SEMs together, each from one sample, and obtain the following SEM regression model:

$$\mathbb{V} = \mathbb{D}\boldsymbol{\theta} + (\boldsymbol{\epsilon}_1^\top, \ldots, \boldsymbol{\epsilon}_n^\top)^\top. \tag{3.1.3}$$

We propose to perform a constrained statistical analysis by minimize the objective function $(\mathbb{V} - \mathbb{D}\boldsymbol{\theta})^\top(\mathbb{V} - \mathbb{D}\boldsymbol{\theta})$ over the feasible region $\Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ defined as $\{\boldsymbol{\theta} \mid \sum_{i=1}^p I(\boldsymbol{\alpha}_i \neq \mathbf{0} \text{ and } \boldsymbol{\beta}_i \neq \mathbf{0}) \leq U^{\alpha\beta}, \sum_{i=1}^p I(\boldsymbol{\alpha}_i \neq \mathbf{0}) + I(\boldsymbol{\beta}_i \neq \mathbf{0}) \leq U^{\alpha+\beta}\}$. In other words, we optimize

$$\min_{\boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})} (\mathbb{V} - \mathbb{D}\boldsymbol{\theta})^\top(\mathbb{V} - \mathbb{D}\boldsymbol{\theta}), \tag{3.1.4}$$

where the constraint $\boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ is imposed to achieve two goals simultaneously. They are, (i) to identify subgroup labels for each of the p mediators in one of the four subgroups: $\mathcal{G}^\alpha = \{i : \boldsymbol{\alpha}_i \neq \mathbf{0}, \boldsymbol{\beta}_i = \mathbf{0}\}$, $\mathcal{G}^\beta = \{i : \boldsymbol{\alpha}_i = \mathbf{0}, \boldsymbol{\beta}_i \neq \mathbf{0}\}$, $\mathcal{G}^{\alpha\beta} = \{i : \boldsymbol{\alpha}_i \neq \mathbf{0}, \boldsymbol{\beta}_i \neq \mathbf{0}\}$, and $\bar{\mathcal{G}} = \{i : \boldsymbol{\alpha}_i = \mathbf{0}, \boldsymbol{\beta}_i = \mathbf{0}\}$, and (ii) to estimate $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$'s that are not zero. Unlike the existing group lasso approach (*Yuan and Lin*, 2006; *Simon et al.*, 2013), in our optimization problem in (3.1.4) the group memberships are unknown, and will be estimated by the proposed L_0 regularization method described in section 3.2.

We make two new contributions to the literature. First, we formulate the above L_0 penalization problem in the form of a mixed integer programming (MIP) optimization problem, which enables us to estimate group memberships as part of mediation analysis for causal pathway identification. Such characterization of mediator subgroups is not only useful to identify causal pathways, but also helpful to improve the power

of existing methods of hypothesis testing for causal effects, such as the Sobel test (*Sobel*, 1982), the joint significant method (*MacKinnon et al.*, 2002), the bootstrap method (*Fritz and MacKinnon*, 2007) and the difference method (*MacKinnon et al.*, 1995) in which test statistics follow different distributions over the four subgroups . The power improvement is due to the fact that our method provides estimated subgroup sizes that are of critical importance to weigh subgroup-specific test statistics in the construction of an overall test. Second, we develop a fast algorithm with both computationally efficient upper and lower bounds to solve the L_0 optimization problem. In this way, we come up with an appealing approximation to the solution of the NP-hard problem. Our approach is indeed new in the literature of SEM-based causal analyses since only L_1 penalty has been previously considered (*Serang et al.*, 2017; *Derkach et al.*, 2019). As shown analytically in the paper, the proposed L_0 regularized approach produces consistently estimated group labels and asymptotically normally distributed estimators of the model parameters, with asymptotically ignorable estimation bias. These properties, unfortunately, cannot be directly obtained from the L_1 penalty without further work on bias correction.

This paper is organized as follows. Section 3.2 discusses an MIP formulation for the regularized estimation defined in equation (3.1.4). Section 3.3 presents our algorithms to obtain an approximate solution of the MIP optimization, with both upper and lower bounds, in which we established algorithmic convergence. We also established both estimation and selection consistency for the proposed method in Section 3.4. Some numerical illustrations, including both simulation studies and data analysis examples, are given in Sections 3.5 and 3.6. We make some concluding remarks in Section 3.7. Some technical details such as proofs of propositions and theorems are included in the appendix.

3.2 Notation and Optimization Formulation

As mentioned above, we propose an L_0 regularization to identify and estimate important mediators in the SEM. To proceed, in a similar spirit to (*Bertsimas et al.*, 2016), we want to find the sparse solution of the parameters in equation 3.1.3 by the following mixed integer programming (MIP) problem. First, we introduce latent indicator variables, denoted by the binary variable vectors $\boldsymbol{\eta}^{\alpha}$, $\boldsymbol{\eta}^{\beta}$ and $\boldsymbol{\eta}^{\alpha\beta}$. They are $\boldsymbol{\eta}^{\alpha} = (\boldsymbol{\eta}_1^{\alpha}, \ldots, \boldsymbol{\eta}_p^{\alpha})^{\top}$, $\boldsymbol{\eta}^{\beta} = (\boldsymbol{\eta}_1^{\beta}, \ldots, \boldsymbol{\eta}_p^{\beta})^{\top}$ and $\boldsymbol{\eta}^{\alpha\beta} = (\boldsymbol{\eta}_1^{\alpha\beta}, \ldots, \boldsymbol{\eta}_p^{\alpha\beta})^{\top}$, respectively. Each pair of binary variables can uniquely determine the subgroup membership of a mediator. For example, when both $\boldsymbol{\eta}_i^{\alpha}$ and $\boldsymbol{\eta}_i^{\beta}$ are 1 if $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are non-zero vectors, mediator i belongs to subgroup $\mathcal{G}^{\alpha\beta}$. Likewise, we have

$$\mathcal{G}^{\alpha} = \{i : \boldsymbol{\eta}_i^{\alpha} = 1, \boldsymbol{\eta}_i^{\beta} = 0\},\ \mathcal{G}^{\beta} = \{i : \boldsymbol{\eta}_i^{\alpha} = 0, \boldsymbol{\eta}_i^{\beta} = 1\},\ \bar{\mathcal{G}} = \{i : \boldsymbol{\eta}_i^{\alpha} = 0, \boldsymbol{\eta}_i^{\beta} = 0\}.$$

The presence and absence of a pathway may be characterized by $\boldsymbol{\eta}_i^{\alpha\beta} = \boldsymbol{\eta}_i^{\alpha}\boldsymbol{\eta}_i^{\beta}$. Clearly, $\mathcal{G}^{\alpha\beta} = \{i : \boldsymbol{\eta}_i^{\alpha\beta} = 1\}$ corresponding to the collection of mediators with causal pathways present, and the union $\mathcal{G}^{\alpha} \cup \mathcal{G}^{\beta} \cup \bar{\mathcal{G}}$ is the collection of absent causal pathways. This dual-task optimization problem takes the following form:

Problem III.1.

$$
\begin{aligned}
&\min_{\boldsymbol{\theta},\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta},\boldsymbol{\eta}^{\alpha\beta}} (\mathbb{V}-\mathbb{D}\boldsymbol{\theta})^{\top}(\mathbb{V}-\mathbb{D}\boldsymbol{\theta});\\
&\boldsymbol{\eta}_i^{\alpha},\boldsymbol{\eta}_i^{\beta} \in \{0,1\}; i=1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha\beta} \in [0,1]; i=1,\ldots,p;\\
&\textit{SOS-1: } (1-\boldsymbol{\eta}_i^{\alpha})\boldsymbol{\alpha}_i = \mathbf{0}; i=1,\ldots,p;\\
&\textit{SOS-1: } (1-\boldsymbol{\eta}_i^{\beta})\boldsymbol{\beta}_i = \mathbf{0}; i=1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha\beta} \leq \boldsymbol{\eta}_i^{\alpha}; i=1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha\beta} \leq \boldsymbol{\eta}_i^{\beta}; i=1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha}+\boldsymbol{\eta}_i^{\beta}-1 \leq \boldsymbol{\eta}_i^{\alpha\beta}; i=1,\ldots,p;\\
&\sum_{i=1}^{p} \boldsymbol{\eta}_i^{\alpha\beta} \leq U^{\alpha\beta};\\
&\sum_{i=1}^{p} \boldsymbol{\eta}_i^{\alpha}+\boldsymbol{\eta}_i^{\beta} \leq U^{\alpha+\beta}.
\end{aligned}
$$

In Problem III.1 above , the objective function is the sum of squares, which is known to be a convex function. The model parameter vector $\boldsymbol{\theta}$ defined as in equation (3.1.3) are continuous, while the latent vectors of indicators $\boldsymbol{\eta} = \left((\boldsymbol{\eta}^{\alpha})^{\top},(\boldsymbol{\eta}^{\beta})^{\top},(\boldsymbol{\eta}^{\alpha\beta})^{\top}\right)^{\top}$ are binary. $(\boldsymbol{\eta}_1^{\alpha},\ldots,\boldsymbol{\eta}_p^{\alpha})^{\top}$, $(\boldsymbol{\eta}_1^{\beta},\ldots,\boldsymbol{\eta}_p^{\beta})^{\top}$ and $(\boldsymbol{\eta}_1^{\alpha\beta},\ldots,\boldsymbol{\eta}_p^{\alpha\beta})^{\top}$, respectively. Binary variables $\boldsymbol{\eta}_i^{\alpha}$ and $\boldsymbol{\eta}_i^{\beta}$ are 1 if $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are non-zero vectors, respectively. In addition, there are two tuning parameters $U^{\alpha\beta}$ and $U^{\alpha+\beta}$ to control the sparsity of the solution. We let $\boldsymbol{\theta}[\alpha_i]$ denote the subvector of $\boldsymbol{\theta}$ with elements selected according to $\boldsymbol{\alpha}_i$. Let $\boldsymbol{\theta}[\eta_i^{\alpha}]$, $\boldsymbol{\theta}[\eta_i^{\beta}]$ and $\boldsymbol{\theta}[\eta_i^{\alpha\beta}]$ denote, respectively, the nonzero elements of $\boldsymbol{\theta}$ according to nonzero subvector $\boldsymbol{\theta}[\alpha_i] \neq \mathbf{0}$, $\boldsymbol{\theta}[\beta_i] \neq \mathbf{0}$ and $\boldsymbol{\theta}[\alpha_i] \neq \mathbf{0}, \boldsymbol{\theta}[\beta_i] \neq \mathbf{0}$, respectively. Let $\Theta(U^{\alpha\beta},U^{\alpha+\beta})$ denote the feasible parameter space, $\{(\boldsymbol{\theta},\boldsymbol{\theta}[\eta^{\alpha}],\boldsymbol{\theta}[\eta^{\beta}],\boldsymbol{\theta}[\eta^{\alpha\beta}])\}$ in Problem III.1. By feasibility, we mean that given the sparsity $U^{\alpha\beta},U^{\alpha+\beta}$, a $\boldsymbol{\theta} \in \Theta(U^{\alpha\beta},U^{\alpha+\beta})$ is a viable solution of Problem III.1.

3.3 Algorithm for L_0 Regularized Estimation

Due to the NP-hard nature of Problem III.1, there is little hope for any efficient algorithm to yield the global optimal solution. As a comprise widely adopted in practice, one attempts to pursue a solution of near-optimality in a reasonable amount of computation time. For example, in the literature general solvers such as Gurobi (*Optimization*, 2019b) and CPlex (*Optimization*, 2019a) are developed by the branch and bound algorithm (*Morrison et al.*, 2016) to handle integer programming problems. To handle the high-dimensional mediators in the causal pathway analysis, in this section we want to develop a highly scalable algorithm that enables us to solve Problem III.1 efficiently. Our improvement is achieved by two sharp and computationally easy bounds that can fast find a warm start near the solution to Problem III.1. It is nontrivial to develop such good upper and lower bounds of the optimal objective value, which is one of the key technical gap to be filled in this paper. In short, our algorithmic strategy to solve Problem III.1 consists of two steps: First, we develop two bounds to generate near-optimal starting values, and then, we apply the Kelley-Cheney-Goldstein method algorithm (*Kelley*, 1960) to push the search result as close to the global optimality as possible.

3.3.1 Upper Bound

In this subsection, the solution point corresponding to a near-optimal upper bound of the optimal objective value is pursued to provide a warm start in solving Problem III.1. In fact, such upper bound may be derived from a more general class of objective functions with the same feasible region $\Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ than that given in Problem III.1. This leads to Problem III.2, of which Problem III.1 is a special case.

Problem III.2. *Suppose function $g(\boldsymbol{\theta})$ is convex, and has a finite lower bound, and satisfies the condition of Lipchitz continuous gradient, $\|\nabla g(\boldsymbol{\theta}_1) - \nabla g(\boldsymbol{\theta}_2)\|_2 \leq$*

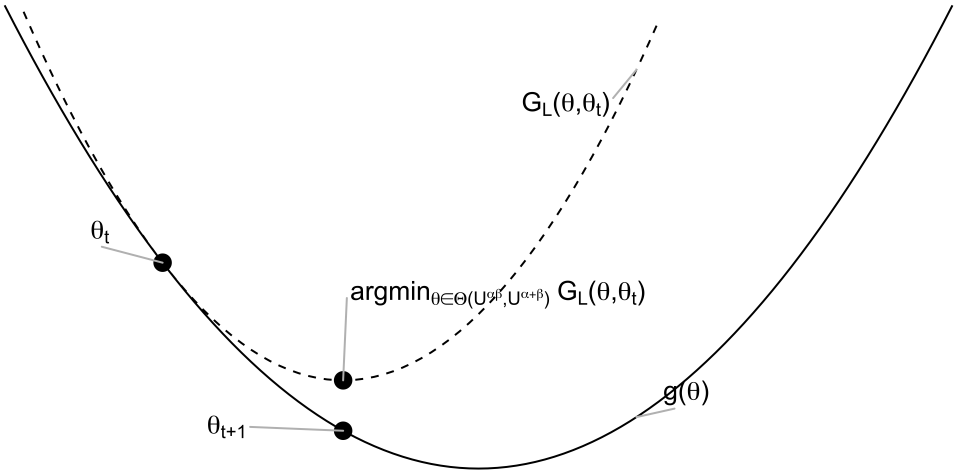

Figure 3.2: An illustration of Algorithm 2 searching for a near-optimal upper bound of the optimal objective value in Problem III.1. Solid line is $g(\boldsymbol{\theta})$. Starting from $\boldsymbol{\theta}_t \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})$, we construct $G_L(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ tangent to $g(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_t$ (the dotted line) . The minimizer of $G_L(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ in the feasible region $\Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ is $\boldsymbol{\theta}_{t+1}$. Then $g(\boldsymbol{\theta}_{t+1}) \leq g(\boldsymbol{\theta}_t)$.

$l \left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|_2$ *where* l *is a constant. Solve the following minimization:*

$$\min_{\boldsymbol{\theta}\in\Theta(U^{\alpha\beta},U^{\alpha+\beta})} g(\boldsymbol{\theta}).$$

As illustrated in Figure 3.2, to attain a good upper bound, we start at $\boldsymbol{\theta}_t \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ and create a quadratic curve $G_L(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ above $g(\boldsymbol{\theta})$ and tangent to $g(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_t$. A form of such quadratice function is given in Proposition III.3. An optimal solution $\boldsymbol{\theta}_{t+1}$ to the quadratic function $G_L(\boldsymbol{\theta}, \boldsymbol{\theta}_t)$ in the feasible region $\Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ is very easy to get numerically. This leads to a better upper bound because of the descending property, $g(\boldsymbol{\theta}_{t+1}) \leq g(\boldsymbol{\theta}_t)$. We keep iterating this search until $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2$ shrinks to 0. At the convergence, we hope to obtain a value near the global minimum.

Proposition III.3. *For a convex function* $g(\boldsymbol{\theta})$ *having Lipschitz continuous gradient*

and for any $L \geq l$, we have

$$g(\boldsymbol{\theta}) \leq G_L(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}) := g(\tilde{\boldsymbol{\theta}}) + \nabla g(\tilde{\boldsymbol{\theta}})^\top (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) + \frac{L}{2}\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_2^2$$
$$= \frac{L}{2}\|\boldsymbol{\theta} - (\tilde{\boldsymbol{\theta}} - \frac{1}{L}\nabla g(\tilde{\boldsymbol{\theta}}))\|_2^2 - \frac{1}{2L}\|\nabla g(\tilde{\boldsymbol{\theta}})\|_2^2 + g(\tilde{\boldsymbol{\theta}})$$

for all $\boldsymbol{\theta}$, $\tilde{\boldsymbol{\theta}}$ with equality holding at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

To get $\boldsymbol{\theta}_{t+1}$, the following algorithm will return an element in $H_{U^{\alpha\beta}, U^{\alpha+\beta}}(\boldsymbol{c})$, which denotes the set of optimal points of Problem III.1 when $g(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{c}\|_2^2$.

Algorithm 1 ($\Omega(p \log p)$)**.**

Input: $\boldsymbol{c} \in \mathbb{R}^{qp+qm+pm}$, $U^{\alpha\beta}$ *and* $U^{\alpha+\beta}$.

Output: $\boldsymbol{\theta} \in H_{U^{\alpha\beta}, U^{\alpha+\beta}}(\boldsymbol{c})$.

1. $\boldsymbol{\theta}[\gamma] = \boldsymbol{c}[\gamma]$.
2. *Let δ be a bijection from $\{1, \ldots, 2p\}$ to $\{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ such that $\|\boldsymbol{c}[\delta(1)]\|_2 \geq \ldots \geq \|\boldsymbol{c}[\delta(2p)]\|_2$.*
3. *Let Γ be a bijection from $\{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ to itself such that $\Gamma(\alpha_i) = \beta_i$ and $\Gamma(\beta_i) = \alpha_i$ for $i = 1, \ldots, p$.*
4. $\boldsymbol{\theta}[\alpha] = \mathbf{0}, \boldsymbol{\theta}[\beta] = \mathbf{0}, u^{\alpha\beta} = 0, u^{\alpha+\beta} = 0$.
5. *For i from 1 to $2p$:*

    If $u^{\alpha+\beta} < U^{\alpha+\beta}$:

        If $\boldsymbol{\theta}[(\Gamma \circ \delta)(i)] = \mathbf{0}$:

            $\boldsymbol{\theta}[\delta(i)] = \boldsymbol{c}[\delta(i)]$.

            set $u^{\alpha+\beta}$ to $u^{\alpha+\beta} + 1$.

        Else if $\boldsymbol{\theta}[(\Gamma \circ \delta)(i)] \neq \mathbf{0}$ and $u^{\alpha\beta} < U^{\alpha\beta}$:

            $\boldsymbol{\theta}[\delta(i)] = \boldsymbol{c}[\delta(i)]$.

            set $u^{\alpha\beta}$ to $u^{\alpha\beta} + 1$, $u^{\alpha+\beta}$ to $u^{\alpha+\beta} + 1$.

6. *Return* $\boldsymbol{\theta}$.

To describe the value $g(\boldsymbol{\theta}_{t+1})$ converges to, we define the first-order stationary point in Definition III.4.

Definition III.4 (first-order stationary point)**.** Given Problem III.1 and $L \geq l$, the vector $\boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ is said to be a first-order stationary point if $\boldsymbol{\theta} \in H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta}))$.

The following proposition presents the relation between a first-order stationary point and a solution to Problem III.1:

Proposition III.5. *Suppose $L > l$. We have the following:*

1. *If $\boldsymbol{\theta}$ is a first-order stationary point, then the set $H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta}))$ has only one element: $\boldsymbol{\theta}$.*

2. *If $\boldsymbol{\theta}$ is a solution to Problem III.1, then it is a first-order stationary point.*

3. *Consider a first-order stationary point $\boldsymbol{\theta}$ of Problem III.1. If θ satisfies the following two conditions:*

   (a) $\sum_{i=1}^{p} \boldsymbol{\theta}[\boldsymbol{\eta}_i^{\alpha\beta}] < U^{\alpha\beta}$;

   (b) $\sum_{i=1}^{p} \boldsymbol{\theta}[\boldsymbol{\eta}_i^{\alpha}] + \boldsymbol{\theta}[\boldsymbol{\eta}_i^{\beta}] < U^{\alpha+\beta}$.

   then $\boldsymbol{\theta}$ is a solution to Problem III.1.

We formally present the following algorithm to search for objective value of a first-order stationary point as a good upper bound:

Algorithm 2.

Input: $g(\boldsymbol{\theta})$, $U^{\alpha\beta}$, $U^{\alpha+\beta}$, L such that $L > l$, convergence tolerance ϵ.

Output: A feasible point $\boldsymbol{\theta}^*$ such that $g(\boldsymbol{\theta}^*) = g(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is some first-order stationary point.

1. *Randomly draw* $\boldsymbol{\theta}_1 \in \mathbb{R}^{qp+qm+pm}$.

2. *For* $t \geq 1$, $\boldsymbol{\theta}_{t+1} \in H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{\theta}_t - \frac{1}{L}\nabla g(\boldsymbol{\theta}_t))$.

3. *Repeat Step 2, until* $\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t+1}\|_2 \leq \epsilon$.

4. *Return* $\boldsymbol{\theta}_{t+1}$.

The following shows that Algorithm 2 terminates after finite iterations:

Proposition III.6. *Consider Problem III.1. Let* $\boldsymbol{\theta}_t, t \geq 1$ *be the sequence generated by Algorithm 2. Then we have:*

1. *For any* $L \geq l$, *the sequence* $g(\boldsymbol{\theta}_t)$ *is decreasing, converges and satisfies*

$$g(\boldsymbol{\theta}_t) - g(\boldsymbol{\theta}_{t+1}) \geq \frac{L-l}{2}\left\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\right\|_2^2.$$

2. *If* $L > l$, *then* $\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_2 \to 0$ *as* $t \to \infty$.

Given a feasible $\boldsymbol{\theta}$, the following definition introduces two values: $\boldsymbol{\theta}[\tau]$ which is the norm of the weakest pathway selected in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}[\rho]$ which is the norm of the weakest pathway among pathways directed from or to the selected mediators in $\boldsymbol{\theta}$.

Definition III.7. Given $\boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})$, we define

$$\boldsymbol{\theta}[\tau] = \begin{cases} \min\Big(\min_{i:\boldsymbol{\theta}[\eta_i^\alpha]=1}\big\|\boldsymbol{\theta}[\alpha_i]\big\|_2, \min_{i:\boldsymbol{\theta}[\eta_i^\beta]=1}\big\|\boldsymbol{\theta}[\beta_i]\big\|_2\Big), & \text{if } \sum_{i=1}^p \boldsymbol{\theta}[\eta_i^\alpha] + \boldsymbol{\theta}[\eta_i^\beta] > 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$\boldsymbol{\theta}[\rho] = \begin{cases} \min_{i:\boldsymbol{\theta}[\eta_i^{\alpha\beta}]=1}\Big\{\min\Big(\big\|\boldsymbol{\theta}[\alpha_i]\big\|_2, \big\|\boldsymbol{\theta}[\beta_i]\big\|_2\Big)\Big\}, & \text{if } \sum_{i=1}^p \boldsymbol{\theta}[\eta_i^{\alpha\beta}] > 1, \\ 0, & \text{otherwise.} \end{cases}$$

The following shows that under mild conditions Algorithm 2 outputs a feasible point whose objective function value is the same with some first-order stationary point:

Proposition III.8. *Consider Problem III.1. Let* $\{\boldsymbol{\theta}_t\}_{t>1}$ *be the sequence generated by Algorithm 2 and* $L > l$*. Then we have:*

1. *If* $\liminf_{t\to\infty} \boldsymbol{\theta}_t[\tau] > 0$*, then* $\boldsymbol{\theta}_t[\eta^\alpha]$ *and* $\boldsymbol{\theta}_t[\eta^\beta]$ *converge for* $i = 1, \ldots, p$*.*

2. *In addition to the condition in Statement 1a, if there exists convergent subsequence* $\boldsymbol{\theta}_{f(t)}$*, then* $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}$ *is a first-order stationary point.*

3. *If there exists a subsequence* $\boldsymbol{\theta}_{f(t)}$ *such that* $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}[\rho] = 0$ *and* $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}[\tau] = 0$*, then* $\lim_{t\to\infty} \nabla g(\boldsymbol{\theta}_{f(t)}) = \mathbf{0}$*.*

4. *If there exists a convergent subsequence* $\boldsymbol{\theta}_{f(t)}$ *such that* $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}[\rho] = 0$ *and* $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}[\tau] = 0$*, then* $\lim_{t\to\infty} g(\boldsymbol{\theta}_t) = \min_{\boldsymbol{\theta}\in\mathbb{R}^{qp+qm+pm}} g(\theta)$*.*

5. *If* $\liminf_{t\to\infty} \boldsymbol{\theta}_t[\rho] > 0$*, then* $\boldsymbol{\theta}_t[\eta^{\alpha\beta}]$ *converges. And if* $\boldsymbol{\theta}_{f(t)}$ *satisfies* $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}[\tau] = 0$*, then for any* $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ *we have:*

$$\lim_{t\to\infty} \Big\{ 1 - I\big(\boldsymbol{\theta}_{f(t)}[s] = \mathbf{0}, \boldsymbol{\theta}_{f(t)}[\Gamma(s)] \neq \mathbf{0}\big) \Big\} \big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s] = \mathbf{0}.$$

6. *If* $\liminf_{t\to\infty} \boldsymbol{\theta}_t[\tau] = 0$*,* $\liminf_{t\to\infty} \boldsymbol{\theta}_t[\rho] > 0$ *and there exists a convergent subsequence* $\boldsymbol{\theta}_{f(m)}$ *such that* $\lim_{m\to\infty} \boldsymbol{\theta}_{f(t)}[\tau] = 0$ *and* $\liminf_{m\to\infty} \boldsymbol{\theta}_{f(t)}[\rho] > 0$*, then* $\boldsymbol{\theta}_{f(m)}$ *converges to a first-order stationary point.*

Remark III.9. In Proposition III.8 Statement 1a, the convergence of $\boldsymbol{\theta}_t[\eta^\alpha]$ and $\boldsymbol{\theta}_t[\eta^\beta]$ implies that the selected causal pathways remain after finite iterations.

Remark III.10. The convergent subsequence conditions in Proposition III.8 Statement 1b, 2b and 3b are satisfied under fairly weak condition, such as $\{\boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta}) \,|\, g(\boldsymbol{\theta}) \leq a\}$ is bounded for any $a \in \mathbb{R}$.

Remark III.11. In Proposition III.8 Statement 1b and 3b, a subsequence's convergence to a first order stationary point $\boldsymbol{\theta}$ is adequate since our aim is to find a feasible point, not necessarily a first-order stationary point, attaining $g(\boldsymbol{\theta})$.

Remark III.12. In summary of Proposition III.8, under some mild condition as stated in Remark III.10, Algorithm 2 will always find a feasible point attaining $g(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is some first-order stationary point.

3.3.2 Lower Bound

In this subsection, a near-optimal lower bound of the optimal objective value is pursued for Problem III.13. Problem III.13 has an additional ridge penalty term with tuning paremeter Δ in the objective function and an equivalent feasible region but with a different formulation compared to Problem III.1. More specifically, $\boldsymbol{\eta}$'s control over mediator sparsity and causal pathway sparsity in $\boldsymbol{\theta}$ are expressed as SOS-1's in Problem III.1 while expressed as $\mathbf{B}(\boldsymbol{\eta}^{\alpha}, \boldsymbol{\eta}^{\beta})\boldsymbol{\theta}$ by direct multiplication in Problem III.13, where "diag$(\cdot)$" is a diagonal matrix whose diagonal entries is the variable. So $\mathbf{B}(\boldsymbol{\eta}^{\alpha}, \boldsymbol{\eta}^{\beta})\boldsymbol{\theta}$ in Problem III.13 has the same feasible region as $\boldsymbol{\theta}$ in Problem III.1.

Problem III.13.

$$
\begin{aligned}
&\min_{\boldsymbol{\theta},\boldsymbol{\eta}} \frac{1}{\Delta}\left\|\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})\boldsymbol{\theta}\right\|_2^2 + \left\|\mathbb{V} - \mathbb{D}\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})\boldsymbol{\theta}\right\|_2^2;\\
&\boldsymbol{\eta}_i^{\alpha},\boldsymbol{\eta}_i^{\beta} \in \{0,1\}; i = 1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha\beta} \in [0,1]; i = 1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha\beta} \leq \boldsymbol{\eta}_i^{\alpha}; i = 1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha\beta} \leq \boldsymbol{\eta}_i^{\beta}; i = 1,\ldots,p;\\
&\boldsymbol{\eta}_i^{\alpha} + \boldsymbol{\eta}_i^{\beta} - 1 \leq \boldsymbol{\eta}_i^{\alpha\beta}; i = 1,\ldots,p;\\
&\sum_{i=1}^{p} \boldsymbol{\eta}_i^{\alpha\beta} \leq U^{\alpha\beta};\\
&\sum_{i=1}^{p} \boldsymbol{\eta}_i^{\alpha} + \boldsymbol{\eta}_i^{\beta} \leq U^{\alpha+\beta};\\
&\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta}) = \begin{pmatrix} diag(\boldsymbol{\eta}^{\alpha}) \otimes I_q & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & diag(\boldsymbol{\eta}^{\beta}) \otimes I_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_p \otimes I_q \end{pmatrix}.
\end{aligned}
$$

Ξ denotes the feasible region of $\boldsymbol{\eta}$ described by all the constraints in Problem III.13, and $\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]$ denotes the matrix whose columns are the non-zero columns in $\mathbb{D}\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})$. By fixing $\boldsymbol{\eta}^{\alpha}$ and $\boldsymbol{\eta}^{\beta}$, we can solve $\boldsymbol{\theta}$ directly. Then we can simplify Problem III.13 as the following by Woodbury matrix identity (*Higham*, 2002):

$$
\begin{aligned}
& \min_{\boldsymbol{\eta}\in\Xi} \mathbb{V}^{\top}\left\{I - \mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]\Big(I/\Delta + \mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]^{\top}\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]\Big)^{-1}\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]^{\top}\right\}\mathbb{V}\\
= \quad & \min_{\boldsymbol{\eta}\in\Xi} \mathbb{V}^{\top}\left\{I + \Delta\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^{\alpha},\boldsymbol{\eta}^{\beta})]^{\top}\right\}^{-1}\mathbb{V}\\
= \quad & \min_{\boldsymbol{\eta}\in\Xi} \mathbb{V}^{\top}\left(I + \Delta\sum_{i=1}^{p}\boldsymbol{\eta}_i^{\alpha}\mathbb{D}[\alpha_i]\mathbb{D}[\alpha_i]^{\top} + \Delta\sum_{i=1}^{p}\boldsymbol{\eta}_i^{\beta}\mathbb{D}[\beta_i]\mathbb{D}[\beta_i]^{\top} + \Delta\mathbb{D}[\gamma]\mathbb{D}[\gamma]^{\top}\right)^{-1}\mathbb{V} \quad (3.3.1)
\end{aligned}
$$

By the Schur complement condition in (*Zhang*, 2006), the epigraph of the objective function in Equation 3.3.1 of $\left((\boldsymbol{\eta}^{\alpha})^{\top},(\boldsymbol{\eta}^{\beta})^{\top}\right)^{\top}$ on domain $\mathbb{R}_+^{2p}$ can be written as the

following :

$$\left\{ \begin{pmatrix} \boldsymbol{\eta}^{\alpha} \\ \boldsymbol{\eta}^{\beta} \\ x \end{pmatrix} \in \mathbb{R}_{+}^{2p+1} \;\middle|\; \begin{pmatrix} x & \mathbb{V}^{\top} \\ \mathbb{V} & \mathbf{Q} \end{pmatrix} \in S_{+}^{n(q+p+m)+1} \right\},$$

where

$$\mathbf{Q} = I + \Delta \sum_{i=1}^{p} \boldsymbol{\eta}_i^{\alpha} \mathbb{D}[\alpha_i]\mathbb{D}[\alpha_i]^{\top} + \Delta \sum_{i=1}^{p} \boldsymbol{\eta}_i^{\beta} \mathbb{D}[\beta_i]\mathbb{D}[\beta_i]^{\top} + \Delta \mathbb{D}[\gamma]\mathbb{D}[\gamma]^{\top}.$$

where $\mathbb{R}_{+}^{2p+1}$ is the set of all $2p+1$ dimensional non-negative vectors and $S_{+}^{n(q+p+m)+1}$ is the convex set of all $n(q+p+m)+1$ dimensional positive semi-definite matrices. It is easy to check this epigraph is convex. Then the objective function in Equation 3.3.1 is a convex function of $\left((\boldsymbol{\eta}^{\alpha})^{\top}, (\boldsymbol{\eta}^{\beta})^{\top}\right)^{\top}$ on domain $\mathbb{R}_{+}^{2p}$.

To further simplify Equation 3.3.1, by the following Theorem III.14, we can reformulate Problem III.13 as its dual problem Equation 3.3.2.

Theorem III.14. *(Vapnik, 1998) The problem*

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^{qp+qm+pm}} \frac{1}{\Delta} \|\boldsymbol{\theta}\|_2^2 + \|\mathbb{V} - \mathbb{D}\boldsymbol{\theta}\|_2^2$$

can equivalently be formulated as the unconstrained maximization problem

$$\max_{\mathbf{Z} \in \mathbb{R}^{n(q+p+m)}} -\Delta \mathbf{Z}^{\top}\mathbb{D}\mathbb{D}^{\top}\mathbf{Z} - \mathbf{Z}^{\top}\mathbf{Z} + 2\mathbb{V}^{\top}\mathbf{Z}.$$

$$\begin{aligned} \min_{\boldsymbol{\eta} \in \Xi} \max_{\mathbf{Z} \in \mathbb{R}^{n(q+p+m)}} & -\Delta \sum_{i=1}^{p} \boldsymbol{\eta}_i^{\alpha} \mathbf{Z}^{\top}\mathbb{D}[\alpha_i]\mathbb{D}[\alpha_i]^{\top}\mathbf{Z} - \Delta \sum_{i=1}^{p} \boldsymbol{\eta}_i^{\beta} \mathbf{Z}^{\top}\mathbb{D}[\beta_i]\mathbb{D}[\beta_i]^{\top}\mathbf{Z} \\ & -\Delta \mathbf{Z}^{\top}\mathbb{D}[\gamma]\mathbb{D}[\gamma]^{\top}\mathbf{Z} - \mathbf{Z}^{\top}\mathbf{Z} + 2\mathbb{V}^{\top}\mathbf{Z} \end{aligned} \tag{3.3.2}$$

To reduce the dimension of $\mathbf{Z}$ in Equation 3.3.2, we utilize the fact that the optimal $\mathbf{Z}$

of the inner maximization question is always in the linear space generated by columns in $(\mathbb{D}, \mathbb{V})$ and have the following equivalent formulation:

$$\min_{\boldsymbol{\eta}\in\Xi} \max_{\mathbf{U}\in\mathbb{R}^{qp+\min(p,n)m+qm+1}} -\Delta\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^\alpha,\boldsymbol{\eta}^\beta)]\mathbb{D}[\mathbf{B}(\boldsymbol{\eta}^\alpha,\boldsymbol{\eta}^\beta)]^\top\mathbb{B}\mathbf{U}$$
$$-\mathbf{U}^\top\mathbb{B}^\top\mathbb{B}\mathbf{U}+2\mathbb{V}^\top\mathbb{B}\mathbf{U}, \tag{3.3.3}$$

where $\mathbb{B}$ is a matrix whose columns are a set of maximal linearly independent columns in $(\mathbb{D}, \mathbb{V})$.

To get a lower bound of the optimal objective value of Equation 3.3.3, we relax Equation 3.3.3 by dropping the integer constraints in Ξ. It is easy to check that the resulting relaxed Ξ is $\mathrm{Conv}(\Xi)$. By the minimax theorem in (*Sion et al.*, 1958), the convex relaxation of Equation 3.3.3 shares the optimal objective value with the following:

$$\max_{\mathbf{U}\in\mathbb{R}^{qp+\min(p,n)m+qm+1}} \min_{\boldsymbol{\eta}\in\mathrm{Conv}(\Xi)} -\Delta\sum_{i=1}^{p}\boldsymbol{\eta}_i^\alpha\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\alpha_i]\mathbb{D}[\alpha_i]^\top\mathbb{B}\mathbf{U}-\Delta\sum_{i=1}^{p}\boldsymbol{\eta}_i^\beta\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\beta_i]\mathbb{D}[\beta_i]^\top\mathbb{B}\mathbf{U}$$
$$-\Delta\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\gamma]\mathbb{D}[\gamma]^\top\mathbb{B}\mathbf{U}-\mathbf{U}^\top\mathbb{B}^\top\mathbb{B}\mathbf{U}+2\mathbb{V}^\top\mathbb{B}\mathbf{U}. \tag{3.3.4}$$

For fixed $\mathbf{U}$, the optimal $\boldsymbol{\eta}$ of the inner minimization problem in Equation 3.3.4 is $\boldsymbol{\theta}[\boldsymbol{\eta}]$ where $\boldsymbol{\theta}$ is the output of Algorithm 1 with $\mathbb{D}^\top\mathbb{B}\mathbf{U}$ as input. Then Equation 3.3.4 has an optimal point with integer $\boldsymbol{\eta}$. Note that the objective function of the outer maximization problem in Equation 3.3.4 is a non-differentiable concave function of U. Then Equation 3.3.4 can be solved by a subgradient method (*Shor*, 1985). However, the subgradient method converges slowly compared to an interior point method, since it is a first order method (*Boyd et al.*, 2003). In order to apply an interior point method, we further reformulate Equation 3.3.4 into a differentiable concave function maximization problem by converting the inner minimization problem, which is a linear programming problem, to its dual. As a result, we get the following Second Order

Cone Programming (SOCP) problem equivalent to Equation 3.3.4:

$$\max_{\substack{\mathbf{U}\in\mathbb{R}^{qp+\min(p,n)m+qm+1}\\ \boldsymbol{\xi}\in\mathbb{R}^{6p+2}}} \mathbf{C}^\top\boldsymbol{\xi} - \Delta\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\gamma]\mathbb{D}[\gamma]^\top\mathbb{B}\mathbf{U} - \mathbf{U}^\top\mathbb{B}^\top\mathbb{B}\mathbf{U} + 2\mathbb{V}^\top\mathbb{B}\mathbf{U} \tag{3.3.5}$$

$$\mathbf{A}_{\cdot i}{}^\top\boldsymbol{\xi} \le -\Delta\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\alpha_i]\mathbb{D}[\alpha_i]^\top\mathbb{B}\mathbf{U}, i = 1,\dots,p,$$

$$\mathbf{A}_{\cdot p+i}{}^\top\boldsymbol{\xi} \le -\Delta\mathbf{U}^\top\mathbb{B}^\top\mathbb{D}[\beta_i]\mathbb{D}[\beta_i]^\top\mathbb{B}\mathbf{U}, i = 1,\dots,p,$$

$$\mathbf{A}_{\cdot 2p+i}{}^\top\boldsymbol{\xi} \le 0, i = 1,\dots,p,$$

$$\boldsymbol{\xi} \ge \mathbf{0},$$

where the linear constraints equivalent to $\boldsymbol{\eta} \in \mathrm{Conv}(\Xi)$ are

$$\mathbf{A}((\boldsymbol{\eta}^\alpha)^\top, (\boldsymbol{\eta}^\beta)^\top, (\boldsymbol{\eta}^{\alpha\beta})^\top)^\top \ge \mathbf{C},$$
$$((\boldsymbol{\eta}^\alpha)^\top, (\boldsymbol{\eta}^\beta)^\top, (\boldsymbol{\eta}^{\alpha\beta})^\top)^\top \ge \mathbf{0}.$$

Then we solve the SOCP problem in Equation 3.3.5 by interior point method and use its optimal point and objective value as a warm start and lower bound of Problem III.13 respectively.

3.3.3 Global Optimality

After getting an upper bound and a lower bound of Problem III.13 in Subsection 3.3.1 and 3.3.2 respectively, we finally try to get global solution to Problem III.13. Note that the formulation of Equation 3.3.1 is the same as the one in Theorem 1 in (*Bertsimas et al.*, 2020) except a minor difference that Equation 3.3.1 has extra linear constraints for $\boldsymbol{\eta}^{\alpha\beta}$. So we can apply the Kelley's cutting plane method (*Kelley*, 1960; *Duran and Grossmann*, 1986) described in Section 3 in (*Bertsimas et al.*, 2020) to get a global solution to Equation 3.3.1.

3.4 Theoretical Guarantees

We now present the sufficient and necessary conditions for the selection and estimation consistency for the proposed method.

3.4.1 A "degree of separation" measure

Throughout this section, define a_0 to be the true value of a. For example, $\boldsymbol{\theta}_0$ is the true parameter of $\boldsymbol{\theta}$.

We define a measure of easiness for feature selection as follows:

Definition III.15. (degree of separation)

$$C_{\min} \equiv C_{\min}(\boldsymbol{\theta}_0, \mathbb{D}, U^{\alpha\beta}, U^{\alpha+\beta})$$

$$= \min_{\substack{\boldsymbol{\theta}\in\Theta(U^{\alpha\beta},U^{\alpha+\beta}) \\ \boldsymbol{\theta}[\eta^\alpha]\neq\boldsymbol{\theta}_0[\eta^\alpha] \text{ or } \boldsymbol{\theta}\neq\boldsymbol{\theta}_0}} \frac{\left\|\mathbb{D}\boldsymbol{\theta} - \mathbb{D}\boldsymbol{\theta}_0\right\|_2^2}{n \max\left(\sum_{i=1}^{p} I(\boldsymbol{\theta}[\eta_i^\alpha] = 0, \boldsymbol{\theta}_0[\eta_i^\alpha] = 1) + I(\boldsymbol{\theta}[\eta_i^\beta] = 0, \boldsymbol{\theta}_0[\eta_i^\beta] = 1), 1\right)}.$$

Here, $C_{\min}$ measures the degree of separation between the true signal and the estimated true signals based on wrong feature selections. More specifically, it is the least difference between the true signal and an estimated true signal based on a wrong feature selection per number of false negative features. If $C_{\min}$ is small, then recovery of the true feature selection is difficult due to the estimated true signal based on some wrong feature selection is very similar to the true signal. Thus $C_{\min}$ characterizes the easiness level of the underlying problem.

3.4.2 Necessary Condition

To derive a necessary condition for selection consistency, we first define the set of all easy problems with $C_{\min} \geq \ell$ as

$$B_0(U^{\alpha\beta}, U^{\alpha+\beta}, \ell) = \left\{\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta}), C_{\min}(\boldsymbol{\theta}, \mathbb{D}, U^{\alpha\beta}, U^{\alpha+\beta}) \geq \ell\right\}.$$

The following theorem gives a necessary condition for all problems in $B_0(U^{\alpha\beta}, U^{\alpha+\beta}, \ell)$ uniformly attaining selection consistency:

Theorem III.16. *(Shen et al., 2013) In Problem III.1, for any $U^{\alpha\beta} \geq 0, 1 \leq U^{\alpha+\beta} \leq 2p$ and $\ell > 0$, for any estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$, we have*

$$\sup_{\boldsymbol{\theta}_0 \in B_0(U^{\alpha\beta}, U^{\alpha+\beta}, \ell)} P(\hat{\boldsymbol{\theta}}[\boldsymbol{\eta}] \neq \boldsymbol{\theta}_0[\boldsymbol{\eta}]) \to 0, \quad as \ n, p \to \infty,$$

implying that

$$\ell \geq \frac{1}{r(\mathbb{D}, U^{\alpha\beta}, U^{\alpha+\beta})} \frac{\log(2p)}{4n(q+p+m)},$$

where $r(\mathbb{D}, U^{\alpha\beta}, U^{\alpha+\beta}) = \frac{\max_{1 \leq j \leq p} n^{-1} \|\mathbb{D}_{\cdot j}\|_2^2}{\min\limits_{\substack{\boldsymbol{\theta} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta}) \\ \|\boldsymbol{\theta}[\alpha_j]\|_\infty \geq \boldsymbol{\theta}[\eta_j^\alpha], \forall 1 \leq j \leq p \\ \|\boldsymbol{\theta}[\beta_j]\|_\infty \geq \boldsymbol{\theta}[\eta_j^\beta], \forall 1 \leq j \leq p}} C_{\min}(\boldsymbol{\theta}, \mathbb{D}, U^{\alpha\beta}, U^{\alpha+\beta})}$.

The proof of Theorem III.16 is applying the proof of the necessary condition in Theorem 1 of (*Shen et al.*, 2013) to Problem III.1.

Theorem III.16 shows that the necessary condition of uniformly attaining selection consistency for a collection of easy problems requires a lower bound of the level of easiness of those problems. More specifically, it requires

$$C_{\min} \geq d_2 \frac{\log(2p)}{n(q+p+m)}, \tag{3.4.1}$$

for some constant d_2 that may dependent on $\mathbb{D}$.

3.4.3 Sufficient Condition

Definition III.17. (Oracle estimator). Given the true coefficient $\boldsymbol{\theta}_0$, the oracle estimator $\hat{\boldsymbol{\theta}}^{ol}$ is defined as

$$\underset{\boldsymbol{\theta}[\eta] = \boldsymbol{\theta}_0[\eta]}{\arg\min} \left\| \mathbb{V} - \mathbb{D}\boldsymbol{\theta} \right\|_2^2.$$

We now derive a nonasymptotic probability error bound for feature selection in mediation analysis. Based on this, we prove the oracle property. The next theorem

says that a global minimizer of Problem III.1 consistently reconstructs the oracle estimator at a degree of separation level that is slightly higher than the minimal in Theorem III.16. Without loss of generality, assume that a global minimizer of Problem III.1 exists. Denote the solution to Problem III.1 as $\hat{\boldsymbol{\theta}}$.

Theorem III.18. *(Shen et al., 2013) In Problem III.1, when $U^{\alpha\beta} = U_0^{\alpha\beta}$ and $U^{\alpha+\beta} = U_0^{\alpha+\beta}$, we have that*

$$P(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \leq \frac{e+1}{e-1} \exp\left\{ -\frac{n(q+p+m)}{18} \left(C_{\min} - 36 \frac{\log(2p)}{n(q+p+m)} \right) \right\},$$

which implies that when $C_{\min} > 36\frac{\log(2p)}{n(q+p+m)}$, $\hat{\boldsymbol{\theta}}$ consistently reconstructs $\hat{\boldsymbol{\theta}}^{ol}$, i.e., as $n, p \to \infty$, $P(\hat{\boldsymbol{\theta}}[\eta] \neq \boldsymbol{\theta}_0[\eta]^{ol}) \to 0$.

Theorem III.18 says that $\hat{\boldsymbol{\theta}}$ consistently reconstructs the oracle estimator $\hat{\boldsymbol{\theta}}^{ol}$ as long as the degree-of-separation condition is satisfied, which is,

$$C_{\min} \geq d_3 \frac{\log(2p)}{n(q+p+m)}, \tag{3.4.2}$$

where $d_3 > 36$ is a constant. The lower bound of $C_{\min}$ in the necessary condition (3.4.1) and in the sufficient condition (3.4.2) they are of the same order.

3.5 Simulation Studies

3.5.1 Small-Scale L_0 method Simulation Experiment

We begin with a small-scale simulation study to numerically illustrate the performance of the proposed L_0 regularized procedures for estimation of both subgroup labels and subgroup effects. This type of small-scale problem is often seen in practical studies such as omics' causal pathway analysis with a specific group of omic variants (e.g. lipids). We want to demonstrate numerically two types of consistency, namely in

subgroup identification and parameter estimation. We consider a one-dimensional continuous exposure variable $q = 1$ and a one-dimensional continuous outcome with $m = 1$, and set the dimension of mediators p at 100 or 200, which is of central interest in environmental health science real data analysis. To check the grouping consistency, we vary sample size n from 500, 1000 to 2000. We draw the summary statistics from the simulations with 200 replicates. Table 3.1 lists the results, including estimation bias, mean squared error (MSE), sensitivity and specificity for both variable selection and subgroup identification, and warm start gap as an indicator of algorithmic convergence.

We design the following structural equation model for data simulation. The p-dimensional exposure vector $\mathbf{X}$ is comprised of n *iid* draws from Bernoulli(0.5) and each entry in the m-dimensional parameter vector $\boldsymbol{\alpha}$ is assigned the value 0 or 0.2. The same procedure is used on the specification of the m-dimensional parameter vector $\boldsymbol{\beta} \in \{0, 0.2\}$. The one-dimensional parameter of the direct effect $\boldsymbol{\gamma}$ is set at 1. We consider a block-diagonal covariance matrix $\boldsymbol{\Sigma} = \text{diag}(0, I_p, 2)$. Subsequently, we simulate the data of mediators $\mathbf{M}$, and outcome $\mathbf{Y}$ from the structural equation model (3.1.1).

Among the p potential mediators, the following sparsity scenarios are considered: 10% being the true mediators ($\alpha = \beta = 0.2$), 10% being not associated with either $\mathbf{X}$ or $\mathbf{Y}$ ($\alpha = \beta = 0$), 40% being only associated with $\mathbf{X}$ ($\alpha = 0.2, \beta = 0$), and 40% being only associated with $\mathbf{Y}$ ($\alpha = 0, \beta = 0.2$).

In each of 200 replicates, we use 5-fold cross-validation to choose the tuning parameters $U^{\alpha\beta}$ in $\{5, 10, 15\}$ and $U^{\alpha+\beta}$ in $\{95, 100, 105\}$. When solving the MIP problem, we ran the first-order discrete algorithm with 1000 random start points chosen randomly from a uniform distribution on interval $(-2, 2)$, which helped to attain a warm start point and a upper bound of the optimal value. Using the warm start point, we then ran the integer programming software Gurobi for 10 minutes to search for better

parameter values that would be much closer to the global optimal solution than those obtained from the acceleration algorithm.

Table 3.1: Simulation results for $q = 1$, $p \in \{100, 200\}$, $m = 1$, $n \in \{500, 1000, 2000\}$ across 200 replicates. "$\boldsymbol{\alpha}$ MSE is the average MSE over all entries in $\boldsymbol{\alpha}$. "$\boldsymbol{\alpha}$ Bias" is the average absolute value of bias over all entries in $\boldsymbol{\alpha}$. "Warm start gap" means (warm start algorithm's upper bound-Gurobi's lower bound)/(Gurobi's lower bound).

n		500	1000	2000	500	1000	2000
p		100	100	100	200	200	200
$\boldsymbol{\alpha}_j$	MSE	3.51e-03	1.31e-03	5.35e-04	3.43e-03	1.31e-03	5.54e-04
$\boldsymbol{\beta}_j$	MSE	4.43e-03	1.48e-03	5.83e-04	5.82e-03	1.63e-03	6.16e-04
$\boldsymbol{\gamma}$	MSE	1.52e-02	6.62e-03	2.93e-03	3.25e-02	1.11e-02	3.02e-03
$\boldsymbol{\alpha}_j$	Bias	3.39e-02	2.02e-02	1.30e-02	3.32e-02	1.99e-02	1.30e-02
$\boldsymbol{\beta}_j$	Bias	3.78e-02	2.13e-02	1.31e-02	4.40e-02	2.20e-02	1.36e-02
$\boldsymbol{\gamma}$	Bias	9.43e-02	6.50e-02	4.25e-02	1.42e-01	8.21e-02	4.47e-02
$\boldsymbol{\alpha}$	true positive	47.130	49.615	49.955	94.650	99.005	99.800
$\boldsymbol{\beta}$	true positive	45.705	49.360	49.845	87.080	98.130	99.670
$\boldsymbol{\alpha}$	true negative	44.540	47.745	48.940	90.300	97.005	99.470
$\boldsymbol{\beta}$	true negative	45.220	47.830	49.785	90.205	96.305	98.850
$\boldsymbol{\alpha}_j \neq 0, \boldsymbol{\beta}_j \neq 0$	true positive	6.445	9.400	9.905	13.105	18.535	19.730
$\boldsymbol{\alpha}_j \neq 0, \boldsymbol{\beta}_j = 0$	true positive	37.175	39.065	39.885	73.340	78.080	79.475
$\boldsymbol{\alpha}_j = 0, \boldsymbol{\beta}_j \neq 0$	true positive	36.370	38.725	39.765	69.120	77.840	79.525
$\boldsymbol{\alpha}_j = 0, \boldsymbol{\beta}_j = 0$	true positive	4.565	7.460	9.010	11.110	16.265	19.070
warm start gap		0.300%	0.196%	0.119%	0.462%	0.357%	0.263%

From Table 3.1 it is evident that the proposed L_0 regularization method and algorithms worked very well. Estimation bias is small, indicating that the L_0 regularization method is clearly advantageous over the popular L_1 penalty. More importantly, both sensitivity and specificity of parameter sparsity and subgroup identification are very satisfactory, and as the sample size increase the selection accuracy tends to the true proportions.

3.5.2 Large-Scale $L_0 + L_2$ Simulation Experiment

3.5.2.1 Computational Efficiency

The above small-scale simulation experiment has shown the desirable approximate solution to the MIP problem. Gurobi may fail to deliver solutions due to excessive

computational burden. Instead, we propose to use the $L_0 + L_2$ method to get solution using the cutting plane algorithm that requires good warm starting values. Here we consider three algorithms to deliver warm starting values: Subgradient Method, Second-Order Cone Programming (SOCP), and Discrete First-Order Method. All these algorithms find approximate solutions close to the global optimal. In other words, these three algorithms will first send the search result near the orbit of the true values, and then the cutting algorithm will follow to refine the search, leading to solutions much closer to the true values. The goal of this simulation study is twofold: (i) to demonstrate the performance of the proposed $L_0 + L_2$ regularization in large-scale setting, and (ii) to show the computational efficiency of the proposed two-stage cutting plane method with a comparison to the commercial package Gurobi (MIO).

To fulfill such objectives, we utilize the same SEM design as that given in the small-scale simulation experiment above, except setting the dimension of mediators $p \in \{500, 1000, 5000, 10000, 50000\}$ and the sample size $n \in \{500, 1000, 5000, 10000, 500000\}$. The variance of the error term, Σ, is $\text{diag}(1, CS(0.1), 1)$, where $CS(0.1)$ is a compound symmetry correlation matrix with correlation set as 0.1. True number of mediators $U_0^{\alpha\beta}$ is set as $\left\lfloor U_0^{\alpha+\beta}/4 \right\rfloor$. Both $U_0^{\alpha\beta}$ and $U_0^{\alpha+\beta}$ are assumed known in each method. The ridge penalty coefficient Δ in the lower bounder objective function is set as $1/n$. For each method, computation is terminated when runtime exceeds 1200 seconds or relative gap, defined as $\frac{\text{upper bound}-\text{lower bound}}{\text{upper bound}}$, is less than 1‰.

Let us first investigate the computational efficiency in terms of average runtime. Figure 3.3 displays the log-mean runtime in seconds of 10 replicates for the four methods in the comparison under large-scale problems $p \in \{500, 1000, 5000\}$ and the sample size $n \in \{500, 1000, 5000\}$.. The white number on each bar indicates the mean relative gap (‰) attained by the time of the corresponding method's termination. "NA" means that the method fails to generate relative gap by the time of termination. Clearly, the fourth method "discrete first-order method+cutting plane method" is the

winner. Runtime is measured on one core in a Dell PowerEdge R430 with Intel Xeon E5-2690 v4 @2.60GHz and 384GB of memory.

Figure 3.4 shows the log-mean runtime in seconds of 10 replicates of "discrete first-order method+cutting plane method" for the ultra large-scale problems with $p \in \{5000, 10000, 50000\}$ and the sample size $n \in \{5000, 10000, 500000\}$. The white number on each bar indicates the mean relative gap (‰) attained by the time of termination. "NA" means that the method fails to generate relative gap by the time of termination. Runtime is measured on one core and three cores for $p = 20000$ cases and $p = 50000$ cases respectively in a Dell PowerEdge R430 with Intel Xeon E5-2690 v4 @2.60GHz and 384GB of memory.

3.5.2.2 Selection Efficiency

Now we assess the performance of the proposed $L_0 + L_2$ method for its estimation and identification using the following metrics: the average number of true positive and average false positive among 100 replicates. We assess the sample size needed to achieve selection consistency and its change when number of potential mediators p, the true number of non-zero coefficients $U_0^{\alpha+\beta}$, correlation of potential mediators ρ and variance of error term σ^2 change separately. We use 10-fold cross validation to choose $U^{\alpha\beta}$ and $U^{\alpha+\beta}$ in $[U_0^{\alpha\beta} - 2, U_0^{\alpha\beta} + 2]$ and $[U_0^{\alpha+\beta} - 2, U_0^{\alpha+\beta} + 2]$ respectively. Table 3.2 shows the parameters and their values used in this section. Note that the value with asterisk is the default value when simulations focus on other parameters. We choose "SOCP + cutting plane" method when $p = 500$ and "discrete first-order method + cutting plane" method in other cases. Computation time limit is 10 minutes for each fold of each tuning parameter combination as well as the final fit.

First, we focus on different dimensions of mediators p. Figure 3.5 shows the average number of true positive and false positive over 100 replicates when varying p and n. In all three p values, sample size 500 is able to perfectly detect all signals and

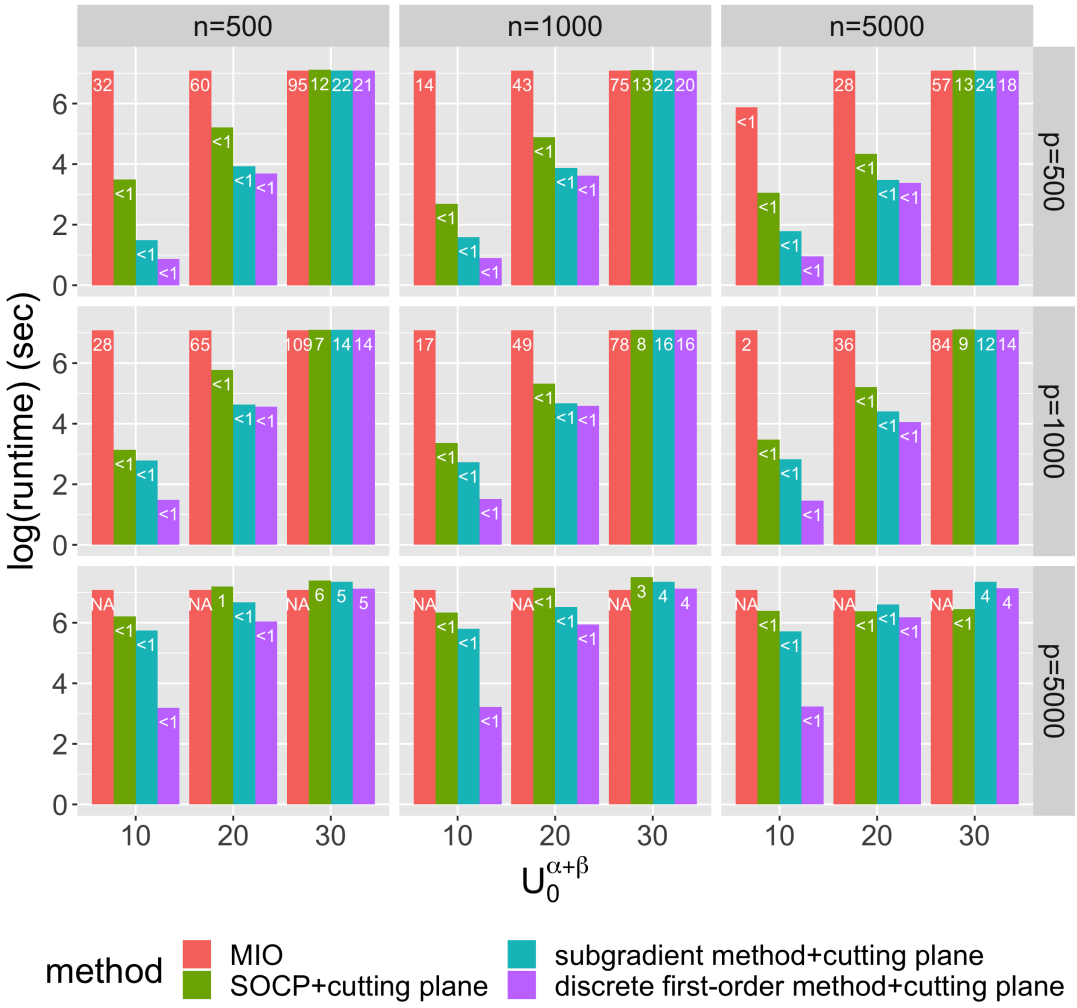

Figure 3.3: Log mean runtime in seconds of 10 replicates for various method, sample size n, number of potential mediators p and true number of non-zero signals $U_0^{\alpha+\beta}$.

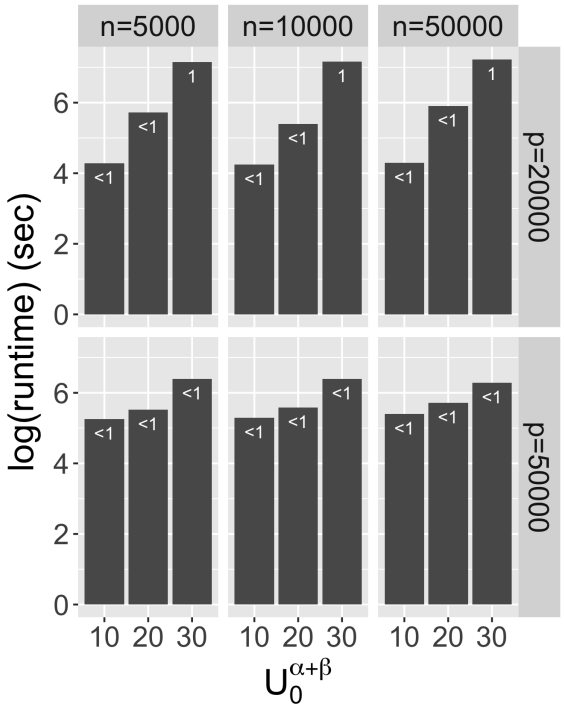

Figure 3.4: Log mean runtime in seconds of 10 replicates of "discrete first-order method+cutting plane method" for various sample size n, number of potential mediators p and true number of non-zero signals $U_0^{\alpha+\beta}$. Runtime is measured on one core and three cores for $p = 20000$ cases and $p = 50000$ cases respectively in a Dell PowerEdge R430 with Intel Xeon E5-2690 v4 @2.60GHz and 384GB of memory.

Table 3.2: Parameter values used in simulations in Section 3.5.2.2. The value with asterisk is the default value when simulations focus on other parameters. $CS(\rho)$ is a compound symmetry correlation matrix with correlation set as ρ.

parameter	values
n	$50, 100, 200, 500, 1000$
p	$500^*, 1000, 5000$
m	1^*
q	1^*
Σ	$\sigma^2 \text{diag}(1, CS(\rho), 1)^*$
ρ	$0.1^*, 0.3, 0.5$
σ^2	$1^*, 2, 5$
$U_0^{\alpha+\beta}$	$10^*, 20, 30$
$U_0^{\alpha\beta}$	$\left\lfloor U_0^{\alpha+\beta}/4 \right\rfloor^*$
$\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\gamma}_i$	0 or 1
Δ	$1/n^*$
$U^{\alpha+\beta}$ tuning range	$[U_0^{\alpha+\beta} - 2, U_0^{\alpha+\beta} + 2]^*$
$U^{\alpha\beta}$ tuning range	$[U_0^{\alpha\beta} - 2, U_0^{\alpha\beta} + 2]^*$
cross validation fold number	10^*
replicate	100^*
algorithm termination time limit	10min^*

nearly perfectly avoid false positive. The sample size needed to achieve selection consistency is larger when the number of potential mediators grows. Second, the focus is varying the true number of signals $U_0^{\alpha+\beta}$. Figure 3.6 shows that sample size 500 is adequate to attain selection consistency. When $U_0^{\alpha+\beta}$ increases, the selection accuracy increase more slowly as sample size increases. Third, with different inter-mediator correlations as the focus, Figure 3.7 shows that sample size 500 is enough to get selection consistency. When inter-mediator correlations ρ rises, selection quality is getting better more slowly as sample size is increasing.

Fourth, the parameter of focus is the variance of the error term σ^2. Once again, Figure 3.8 suggests 500 samples is enough to attain perfect true positive detection and very small false positive. The selection performance of the proposed method is insensitive to the varying variances.

In summary, the above simulation results suggest that the sample size 500, which is one tenth of the largest dimension of mediators (5000), seems to be sufficient to

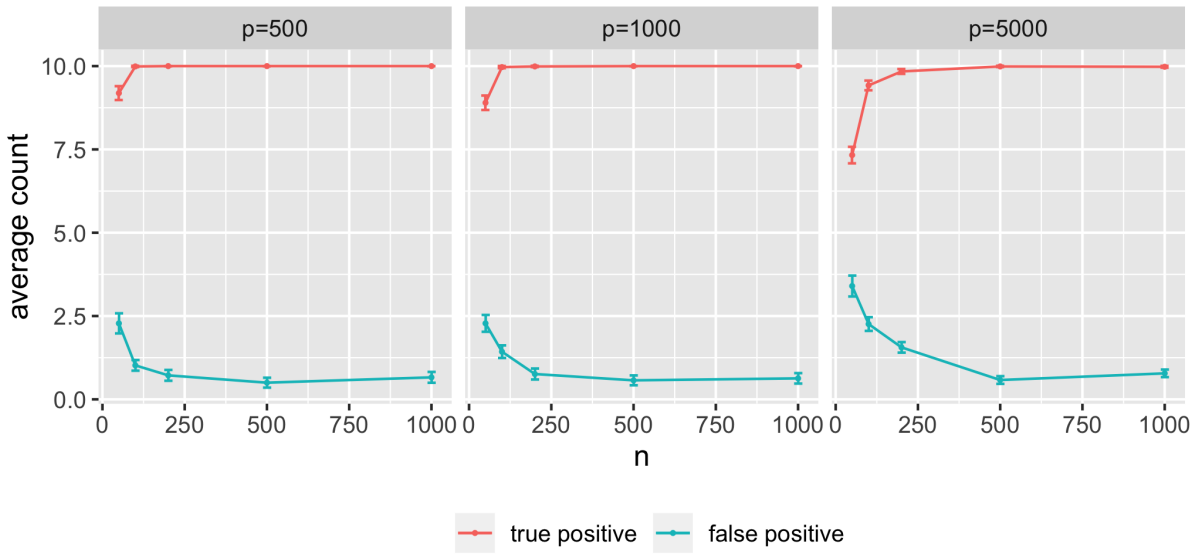

Figure 3.5: Average number of true positive and false positive of 100 replicates of the proposed method for various sample size n and number of potential mediators p.

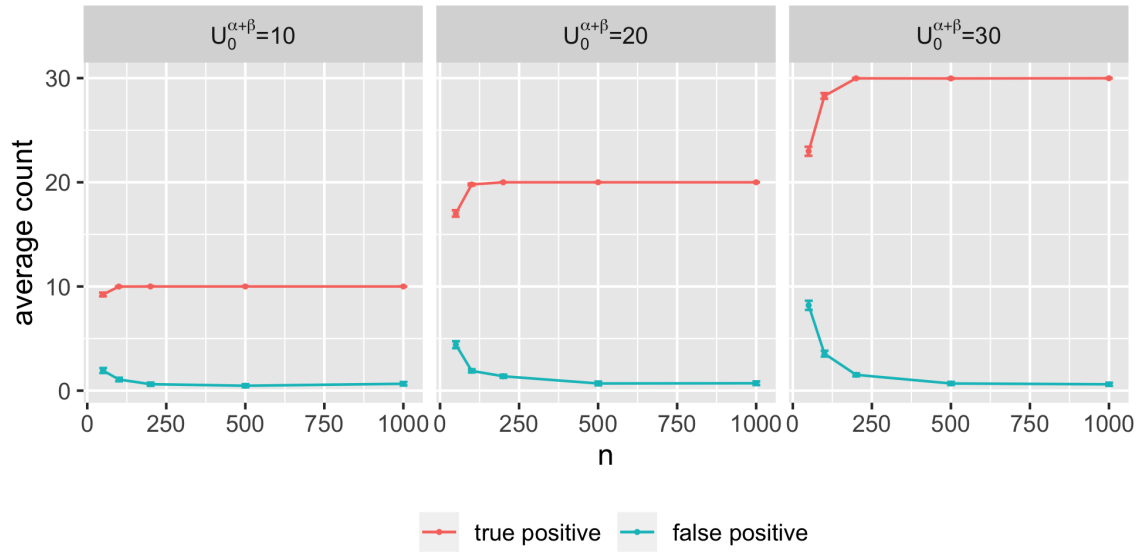

Figure 3.6: Mean number of true positive and false positive of 100 replicates of the proposed method for various sample size n and number of true signals $U_0^{\alpha+\beta}$.

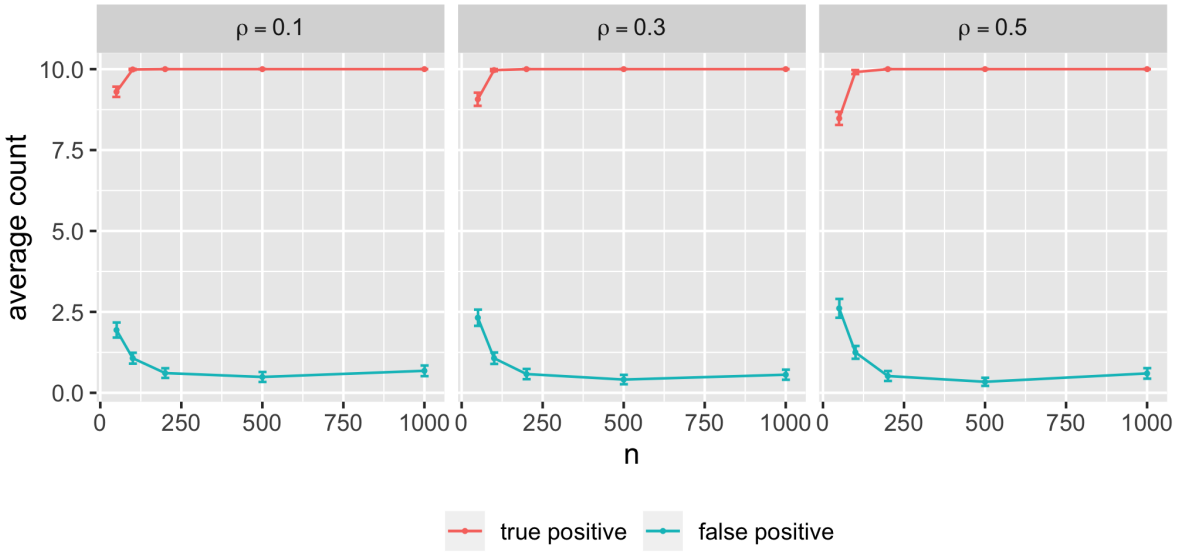

Figure 3.7: Mean number of true positive and false positive of 100 replicates of the proposed method for various sample size n and inter-mediator correlations ρ.

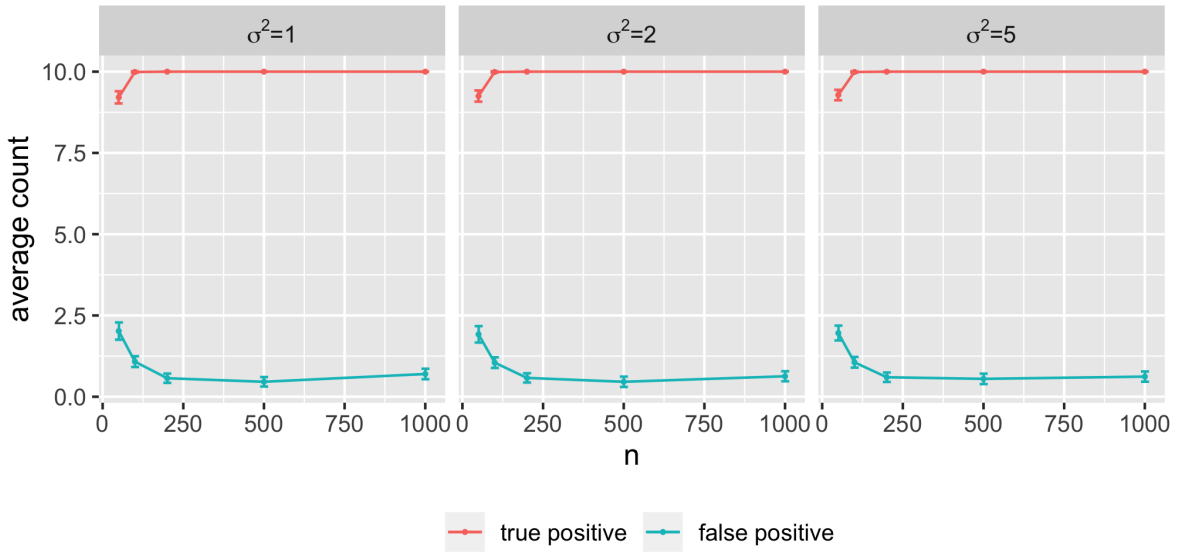

Figure 3.8: Mean number of true positive and false positive of 100 replicates of the proposed method for various sample size n and different variances of the error term σ^2.

ensure a stable and desirable performance of the proposed method over different types of model parameters and data generation scenarios.

3.6 Data Analysis

The proposed method is used to analyze the ELEMENT project data to study the causal mediation pathways with the group of 149 lipids measured from blood samples using the technique of mass spectrometry, the largest metabolites. Metabolomics pertain to a key component of system biology, and are essentially responsible for cellular energy via substances produced during metabolism such as digestion and other bodily chemical processes. As a matter of fact, metabolomics may be altered by environmental factors such as nutrients or toxicants (e.g. phthalates). Many lipids are fatty acids that carry a special kind of energy needed in various cellular operations. Excessive expressions of certain lipids may lead to high body mass index (BMI) or even obesity. Thus, it is of great interest to study how the association of exposure to phthalates with BMI may be mediated by some of lipids. Here, individual phthalates are used in the detection of causal mediation pathway with the BMI outcome. Age and gender are two confounding factors are used to adjust both direct and indirect effects. This analysis assumes no unmeasured confounders for the causal relations from phthalates to lipids, from lipids to BMI and from phthalates to BMI.

The data contains 381 adolescents aged 8-18, out of which 191 subjects are boys and 190 subjects are girls. We use 12 phthalates measured from blood samples of mothers, and take a log-transformation in the data preprocessing to correct the skewness of the exposure variables. We process the data and fit the structural equation model for each phthalate separately by the following steps:

1. Impute the missing phthalate data by the fully conditional specification method using the `MICE` R package.

2. Use inverse normal transformation to make all variables normally distributed other than gender.

3. Using the WHO BMI chart , we adjust BMI for age and gender to produce a BMI z-score.

4. Begin with a working independence covariance $\Sigma = I$ and tune $U^{\alpha\beta}$ and $U^{\alpha+\beta}$ using cross validation in the model.

5. Fit the model with the selected sparsity tuning parameters $U^{\alpha\beta}$ and $U^{\alpha+\beta}$, and use residuals to estimate the covariance of the errors $\hat{\Sigma}$ which allows to perform a deassociation transformation of the data.

6. After the deassociation transformation, and retune $U^{\alpha\beta}$ and $U^{\alpha+\beta}$ using cross validation.

7. Refit the model with the updated sparsity tuning parameters $U^{\alpha\beta}$ and $U^{\alpha+\beta}$.

In addition to the analysis with all subjects, we further stratify the data by gender where the above steps are repeated for boys and girls separately. All causal mediation pathways from phthalate through lipids to BMI discovered in both the analysis of all subjects and the stratified analyses of boys and girls, respectively, are shown in Table 3.6. We have found many causal pathways through fatty acids (the metabolites starting with "FA"). In particular, the pathway: MEOEP $\rightarrow$ lipid "FA.5.0.OH" $\rightarrow$ BMI occurs in both the combined analysis and the stratified analysis of boys. FA.5.0.OH is a fatty acid with 5 carbons, no double bonds, and a hydroxy group, is of special interest. Based on the results, We came up with the following conjecture of how phthalate MEOHP affects obesity through FA.5.0.OH. In the literature, obesity and insulin resistance are related to elevation of leucine in plasma, where leucine is an essential amino acid whose primary metabolic end product is acetoacetate (an energy source in blood) (*She et al.*, 2007). As shown in Figure 3.6, the elevation of

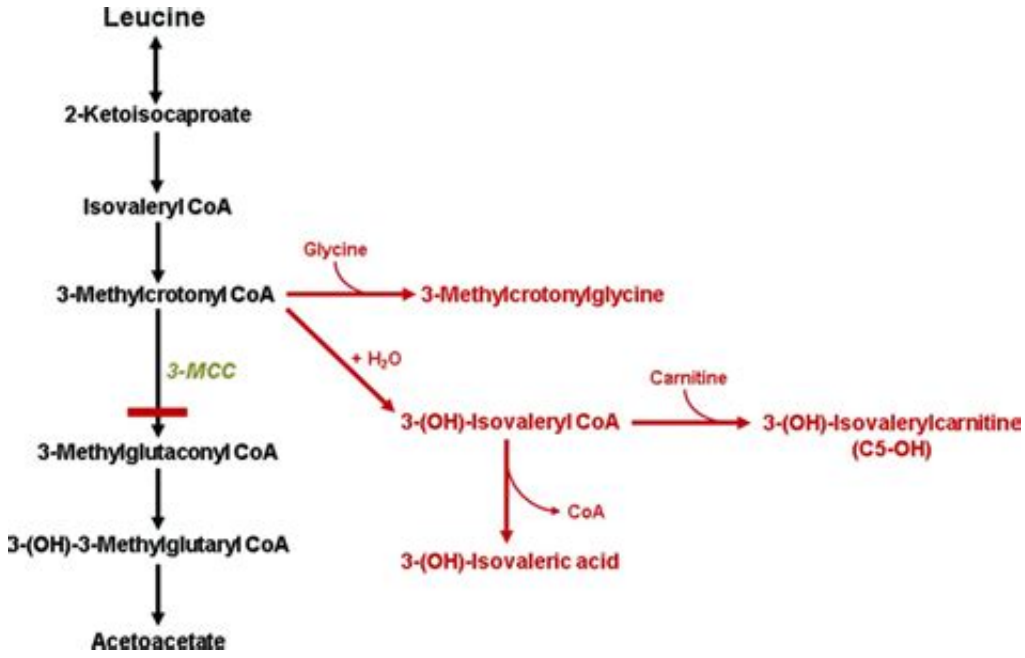

Figure 3.9: Blockage of the chain of metabolism from leucine to its end product acetoacetate results in byproduct 3-(OH)-Isovalerylcarnitine, which is FA.5.0.OH in Table 3.6.

leucine is caused by blockage of the chain of metabolism from leucine to acetoacetate. And FA.5.0.OH is a byproduct of this blockage. In summary, exposure to phthalate MEOHP removes blockage of leucine's metabolism and causes reduction of FA.5.0.OH, then reduces obesity. Note that this conjecture is based on the "no unmeasured confounders" assumption.

3.7 Concluding Remarks

This project developed a fast L_0 regularized estimation method to detect causal mediation pathways in the context of structural equation models. Although this project is motivated from mediation analysis of metabolomic biomarkers, the entire framework is general and it can be applied to analyze other structural equation models with high-dimensional correlated mediators. Using extensive simulation studies, we showed that the proposed methods are numerically flexible and stable to search for the exact solution of the constrained nonconvex objective functions. We compare

Table 3.3: Causal pathways from phthalates through metabolites to BMI found in combined analysis and gender stratified analysis. α, β and γ indicate the coefficients associated with phthalate $\rightarrow$ metabolite, metabolite $\rightarrow$ BMI and phthalate $\rightarrow$ BMI, respectively. The bold lines are related to FA.5.0.OH, which is of special interest.

strata	phthalate	lipid	$\alpha \times 10^{-2}$	$\beta \times 10^{-2}$	$\gamma \times 10^{-2}$
combined	**MEOHP**	**FA.5.0.OH**	**-3.29**	**9.12**	**-1.55**
combined	**MEHP**	**FA.5.0.OH**	**-5.35**	**1.17**	**-1.14**
boy	MCPP	FA.18.0.OH_1	6.34	8.31	-2.65
boy	MECPP	FA.8.0.NH2.2.aminooctanoate_2	6.67	-7.74	-1.71
boy	MECPP	FA.9.0.OH	-7.28	6.61	-1.71
boy	MECPP	MG.0.0.14.0.0.0	8.99	8.81	-1.71
boy	MEHHP	FA.8.0.NH2.2.aminooctanoate_2	5.72	-7.31	-1.67
boy	MEHHP	FA.20.2.0.072688._22.9894	-3.95	7.68	-1.67
boy	MEHHP	FA.9.0.OH	-3.12	6.29	-1.67
boy	MEHHP	MG.0.0.14.0.0.0	9.42	7.72	-1.67
boy	MIBP	X1.OLEOYL.RAC.GLYCEROL	6.15	8.13	-3.66
boy	MIBP	FA.20.2.0.072688._22.9894	6.97	7.98	-3.66
boy	**MEOHP**	**FA.5.0.OH**	**-4.25**	**8.78**	**-7.29**
boy	MEOHP	MG.0.0.14.0.0.0	3.27	7.71	-7.29
boy	MBP	FA.18.0.OH_1	4.41	9.19	-4.45
boy	MBP	FA.20.2_0.072688._22.9894	1.51	8.41	-0.445
boy	MBP	FA.27.5_0.112886._22.761	-2.33	5.63	-0.445
boy	MBP	MG.0.0.14.0.0.0	4.70	7.38	-0.445
boy	MBzP	Keto.18.0_2	3.73	7.19	-2.28
boy	MNP	FA.27.5.0.112886._22.761	-5.47	6.73	-3.12
boy	MCOP	FA.27.5.0.112886._22.761	-6.30	6.61	4.91
girl	MCPP	X1.OLEOYL.RAC.GLYCEROL	2.52	5.92	-1.23
girl	MEOHP	FA.12.0.OH_1	4.95	-8.28	0.895
girl	MEOHP	FA.18.0.OH_4	4.90	7.41	0.895
girl	MEOHP	Keto.14.0_1	5.15	-8.49	0.895
girl	MEOHP	FA.14.0.OH.0.068835._20.7795	-1.86	-7.42	0.895
girl	MCOP	FA.14.0.OH.0.068835._20.7795	-5.34	-7.80	0.647

different algorithms to generate warm starting values in both small-scale, large-scale and ultra large scale settings. These proposed algorithms can be very appealing to handle a broad range of practical problems that cannot be easily solved using existing toolboxes.

This project established a rigorous theoretical framework, including both sufficient and necessary conditions for the subgroup selection consistency. Such conditions are all related to the concept of easiness of an optimization problem, which defines the boundary for viable optimal solutions. All the theoretical results are given under the condition of high-dimensional mediators, namely, large p small n, where p is the number of potential mediators under screening and n is the sample size.

From both theoretical and numerical work, we showed that the proposed algorithm can handle a large-scale problem of causal mediation pathway detection. In the simulation study, we demonstrated success in running our method with 50,000 mediators, which is very challenging to any existing MIP type solvers.

Finally, the selected subgroups are very useful to perform hypothesis testing in causal mediation analysis because it is known that test statistics are different under different types of scenarios according to α and/or β being zero or not. The application of the established selection consistency to develop better hypothesis testing methods is worth a serious investigation.

CHAPTER IV

Summary and Future Work

4.1 Summary

This dissertation developed a rigorous statistical methodology based on the L_0 regularization to carry out both subgroup label detection and group-specific parameter estimation. This proposed methodology has been established in the context of linear regression models and structural equation models. The proposed optimization algorithms have justified for their large-sample properties such as selection consistency in which both sufficient and necessary conditions are given. In addition, the proposed algorithms have been examined using extensive simulation studies, and illustrated by real world data analyses.

The main contribution in the first project is the methodology of homogeneity pursuit that allows us to identify a group of covariates with the same effect size. This is well motivated by the open problem of developing a mixture of toxic agents in environmental health sciences. As shown in the data analysis, we are able to form interpretable mixtures using our proposed model formulation and L_0 solver.

The main contribution in the second project is the methodology of identifying causal mediation pathways in the presence of high-dimensional potential mediators. We developed a two-stage search algorithm; in stage I, using the upper bound of the objective function, we can generate high-quality warm starting values that are shown

to be near the true values; and in stage II, we use cutting plane algorithm to deliver a finer solution using the lower bound of the constrained objective function. We tested the computational efficiency of the proposal algorithm, and demonstrated that our proposed algorithm is able to handle as many as 50,000 mediators in the optimization. This presents a useful toolbox that can be handed into the hands of practitioners to solve a broad range of problems that currently cannot be solved by existing algorithms.

4.2 Future Work

There are many possible future directions that can advance the proposed methodology. Below are a few problems of interest to us.

- In the methodology of homogeneity pursuit, it is of great interest to generate the linear model considered in this dissertation to generalized linear models for nonnormal outcomes, such as logistic regression model for binary outcome. In addition, given the importance of the Cox proportional hazards model, a generalization of the proposed methodology in survival analysis would be valuable. As far as the ELEMENT study concerns, the participants are also measures with their timing of sexual maturation, an interesting setting where the association of hazard for sexual maturation with mixtures of phthalates may be examined.

- Another direction of future work for the homogeneity pursuit is to consider longitudinal outcomes, either in generalized estimating equations (GEE) or mixed-effects models. The ELEMENT cohort study contains repeated measurements of somatic growth variables during age 0 to 5 years. Identifying mixtures affecting somatic growth trajectories is of great interest.

- In the methodology for the identification of causal mediation pathway, it is of great interest to establish statistical inference such as hypothesis testing in

addition to estimation. The proposed L_0 regularized solution provides an ideal setting to further develop this needed inference theory, methods, and software packages.

- Similar to those generalizations considered in Project I, we may consider the structural equation models for nonnormal variables, time-to-event outcomes, and repeated measurements. These generalizations shall make the proposed methodology even broader impacts in practice.

- For both homogeneity pursuit and mediation pathway identification, tuning the L_0 penalized model is difficult compared to Lasso methodology. It is of great interest to study the counterpart of solution path in Lasso methodology.

- Note that the structural equation model is the simplest directed acyclic graph (DAG), which may be further extended to account for more complex DAGs. This seems to be a long-term future work given that the statistical theory and algorithms for the SEM with high-dimensional mediators has not been fully known yet.

APPENDICES

APPENDIX A

Appendices for Chapter II

A.1 Appendices for Chapter II

A.1.1 Algorithm 0

Algorithm 0 ($\Omega(Ks^2 + p\log p + q)$)**.**

Input: $\boldsymbol{c} \in \mathbb{R}^{q+p}$*, the number of groups* K *and the sparsity restriction* s*;*

Output: a member in $H_{K,s}(c)$*.*

1. $\hat{\boldsymbol{\alpha}} = (c_1, c_2, \ldots, c_q)^T$.

2. *Let* δ *be a bijection on* $\{q{+}1, \ldots, q{+}p\}$ *such that* $c_{\delta(q+1)} \leq c_{\delta(q+2)} \leq \ldots \leq c_{\delta(q+p)}$.

3. *set* x_{lk}, y_{lk}, x'_{lk} *and* y'_{lk} *to* 0 *for* $l = 0, \ldots, p+1$ *and* $k = 0, \ldots, K$.

4. *For* l *from* 1 *to* s*:*

   For k *from* 1 *to* K*:*

   $x_{lk} = \arg\max_{1 \leq i \leq l}\{y_{i-1,k-1} + \frac{(\sum_{j=i}^{l} c_{\delta(q+j)})^2}{l-i+1}\}$.

   $y_{lk} = \max_{1 \leq i \leq l}\{y_{i-1,k-1} + \frac{(\sum_{j=i}^{l} c_{\delta(q+j)})^2}{l-i+1}\}$.

5. *If* $s < p$*:*

   For l *from* p *to* $p - s + 1$*:*

   For k *from* 1 *to* K*:*

   $x'_{lk} = \arg\max_{l \le i \le p} \{y'_{i+1,k-1} + \frac{(\sum_{j=l}^{i} c_{\delta(q+j)})^2}{i-l+1}\}$.

   $y'_{lk} = \max_{l \le i \le p} \{y'_{i+1,k-1} + \frac{(\sum_{j=l}^{i} c_{\delta(q+j)})^2}{i-l+1}\}$.

6.

$$(k^*, l^*, m^*) = \underset{\substack{0 \le k \le K \\ 0 \le l \le s \\ p - min(s,p) + l + 1 \le m \le p}}{\arg\max} y_{kl} + y'_{K-k,m}$$

7. *For* l *from* $l^* + 1$ *to* $m^* - 1$*:*

   $\hat{\beta}_{\delta(q+l)} = 0$.

8. *Set* $t = l^*$

   For k *from* k^* *to* 1*:*

   For l *from* t *to* $x_{t,k}$*:*

   $\hat{\beta}_{\delta(q+l)} = \frac{\sum_{j=t}^{x_{tk}} c_{\delta(q+j)}}{t - x_{tk} + 1}$.

   $t = x_{tk} - 1$.

9. *Set* $t = m^*$

   For k *from* $K - k^*$ *to* 1*:*

   For l *from* t *to* x'_{tk}*:*

   $\hat{\beta}_{\delta(q+l)} = \frac{\sum_{j=t}^{x'_{tk}} c_{\delta(q+j)}}{x'_{tk} - t + 1}$.

   $t = x'_{tk} + 1$.

10. *Return* $(\hat{\boldsymbol{\alpha}}^T, \hat{\boldsymbol{\beta}}^T)^T$.

A.1.2 Proof of Proposition II.4

Proposition A.1. *Consider problem* (2.2.6) *and some constant* $L > l$, *let* $\boldsymbol{\theta}_m, m \geq 1$ *be the sequence generated by Algorithm 1. Define*

$$\rho_m = \begin{cases} \min\limits_{\substack{\beta_{m,j} \neq \beta_{m,j'} \\ \beta_{m,j}, \beta_{m,j'} \neq 0}} |\beta_{m,j} - \beta_{m,j'}|, & \text{if there are } K \text{ distinct non-zero values in } \boldsymbol{\beta}_m, \\ 0, & \text{otherwise}; \end{cases}$$

$$\tau_m = \begin{cases} \min_{\beta_{m,j} \neq 0} |\beta_{m,j}|, & \text{if there is some non-zero value in } \boldsymbol{\beta}_m, \\ 0, & \text{otherwise.} \end{cases}$$

1. *When* $\liminf_{m\to\infty} \rho_m > 0$ *and* $\liminf_{m\to\infty} \tau_m > 0$*:*

   (a) $\mathcal{G}(\boldsymbol{\beta}_m)$ *converges.*

   (b) *If* g *has second order derivative and there exists* $l' > 0$ *such that* $l' \left\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\right\|_2 \leq \left\|\nabla g(\boldsymbol{\theta}) - \nabla g(\tilde{\boldsymbol{\theta}})\right\|_2$ *for any* $\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}} \in \Theta(K, s)$ *satisfying* $\mathcal{G}(\boldsymbol{\beta}) = \mathcal{G}(\tilde{\boldsymbol{\beta}})$*, then the sequence* $\boldsymbol{\theta}_m$ *is bounded and converges to a first-order stationary point.*

2. *When* $\liminf_{m\to\infty} \tau_m = 0$*:*

   (a) $\liminf_{m\to\infty} \|\nabla g(\boldsymbol{\theta}_m)\|_\infty = 0$.

   (b) *If there exists a convergent subsequence* $\{\boldsymbol{\theta}_{f(m)}\}$ *such that* $\lim_{m\to\infty} \tau_{f(m)} = 0$*, then* $\lim_{m\to\infty} g(\boldsymbol{\theta}_m) = \min_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$.

3. *When* $\liminf_{m\to\infty} \rho_m = 0$ *and* $\liminf_{m\to\infty} \tau_m > 0$*:*

   (a) $\mathcal{G}(\boldsymbol{\beta}_m; 0)$ *converges and* $\liminf_{m\to\infty} \left\|(\nabla g(\boldsymbol{\theta}_m))_{\mathcal{G}^c(\boldsymbol{\beta}_m;0)}\right\|_\infty = 0$.

   (b) *If there exists a convergent subsequence* $\{\boldsymbol{\theta}_{f(m)}\}$ *such that* $\lim_{m\to\infty} \rho_{f(m)} = 0$*, then* $\boldsymbol{\theta}_{f(m)}$ *converges to a first-order stationary point.*

Remark A.1.1. *The convergent subsequence condition could be satisfied under some weak conditions, like,* $\{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K, s) | g(\boldsymbol{\theta}) \leq C\}$ *is bounded for any* $C \in \mathbb{R}$.

Proof. 1. (a) For large enough m, if $\mathcal{G}(\boldsymbol{\beta}_m) \neq \mathcal{G}(\boldsymbol{\beta}_{m+1})$, then $\|\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m\|_2 > \min(\liminf_{m\to\infty} \rho_m, \liminf_{m\to\infty} \tau_m)/\sqrt{2}$, in contradiction to Proposition II.3 Statement 2.

(b) Due to Statement 1a, there exists M such that for any $m \geq M$, $\mathcal{G}(\boldsymbol{\beta}_m)$ are the same. Then for any $m > M$:

$$\|\boldsymbol{\theta}_{m+2} - \boldsymbol{\theta}_{m+1}\|_2 = \left\| H_{K,s}(\boldsymbol{\theta}_{m+1} - \frac{1}{L}\nabla g(\boldsymbol{\theta}_{m+1})) - H_{K,s}(\boldsymbol{\theta}_m - \frac{1}{L}\nabla g(\boldsymbol{\theta}_m)) \right\|_2$$

$$= \left\| \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{I}_q \end{pmatrix} [(\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m) - \frac{1}{L}(\nabla g(\boldsymbol{\theta}_{m+1}) - \nabla g(\boldsymbol{\theta}_m))] \right\|_2$$

where $\mathbf{A}_{p\times p}$ is an idempotent matrix $\left(\frac{I(\beta_{m,p'} = \beta_{m,p''})}{|\mathcal{G}(\boldsymbol{\beta}_m; \boldsymbol{\beta}_{m,p'})|} \right)_{p',p''\in\{1,\ldots,p\}}$

$$= \left\| \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{I} \end{pmatrix} (I - \frac{1}{L}\nabla^2 g(\boldsymbol{\theta}'))(\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m) \right\|_2 \leq \sqrt{1 - \frac{l'^2}{L^2}} \|\theta_{m+1} - \theta_m\|_2$$

Since $0 < \frac{l'}{L} \leq 1$, θ_m converges to a first order stationary point.

2. (a) Since $\boldsymbol{\theta}_{m+1} - \boldsymbol{\theta}_m$ converges, we have $\lim_{m\to\infty} \left\| \frac{\partial g(\boldsymbol{\theta}_m)}{\partial \boldsymbol{\alpha}} \right\|_\infty = 0$. There exists a subsequence $\{\boldsymbol{\theta}_{f(m)}\}$ that $\lim_{m\to\infty} \tau_{f(m)} = 0$. Without loss of generality, we assume $|\mathcal{G}(\boldsymbol{\beta}_{f(m)}; \tau_{f(m)})| = t > 0$. Let us use $\boldsymbol{c}_m$ to denote $\boldsymbol{\theta}_m - \frac{1}{L}\nabla g(\boldsymbol{\theta}_m)$. Fix m, for any $p' \in \{1, \ldots, p\}$ such that $|\mathcal{G}(\boldsymbol{\beta}_{f(m)}; \beta_{f(m),p'})| = t' > 1$, we create $\tilde{\boldsymbol{\theta}}$ whose grouping is the same as $\boldsymbol{\theta}_{f(m)}$ except that the 0-group and $\tau_{f(m)}$-group in $\boldsymbol{\theta}_{f(m)}$ are merged as the new 0-group and that $\beta_{f(m),p'}$ is singled out as a new group. Then

$$G_L(\theta_{f(m)}, \theta_{f(m)-1}) - G_L(\tilde{\theta}, \theta_{f(m)-1}) =$$

$$\begin{cases} -t\tau_{f(m)}^2 + c_{f(m)-1,p'}^2 \leq 0, \text{ if } \beta_{f(m),p'} = 0 \text{ or } \tau_{f(m)}; \\ -t\tau_{f(m)}^2 + (1 + \frac{1}{t'-1})(\beta_{f(m),p'} - c_{f(m)-1,p'})^2 \leq 0, \text{ otherwise.} \end{cases}$$

So for any $p' \in \{1, \ldots, p\}$, we have $\frac{1}{L}|\frac{\partial g(\theta_{f(m)-1})}{\partial \beta_{p'}}| = |\beta_{f(m)-1,p'} - c_{f(m)-1,p'}| \leq |\beta_{f(m)-1,p'} - \beta_{f(m),p'}| + |\beta_{f(m),p'} - c_{f(m)-1,p'}| \leq \left\|\boldsymbol{\theta}_{f(m)} - \boldsymbol{\theta}_{f(m)-1}\right\|_2 + (\sqrt{s} + 1)\tau_{f(m)}$. Thus $\lim_{m\to\infty} \left\|\frac{\partial g(\boldsymbol{\theta}_{f(m)-1})}{\partial \boldsymbol{\beta}}\right\|_\infty = 0$.

(b) Due to Statement 2a, we have $\lim_{m\to\infty} \left\|\nabla g(\boldsymbol{\theta}_{f(m)-1})\right\|_\infty = 0$. Since $\lim_{m\to\infty} \boldsymbol{\theta}_{f(m)-1} = \boldsymbol{\theta}'$, we have $g(\boldsymbol{\theta}') = \min_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$. Since $g(\boldsymbol{\theta}_m)$ converges, we have $\lim_{m\to\infty} g(\boldsymbol{\theta}_m) = \min_{\boldsymbol{\theta}} g(\boldsymbol{\theta})$.

3. (a) Due to the proof of Statement 1a, if $\liminf_{m\to\infty} \rho_m = 0$, then $\mathcal{G}(\boldsymbol{\beta}_m; 0)$ converges. There exists sequences $\{\boldsymbol{\theta}_{f(m)}\}$, $\{p_m\}$ and $\{p'_m\}$ such that for any $m > 0$ we have $\beta_{f(m),p_m} \neq 0$, $\beta_{f(m),p'_m} \neq 0$, $\beta_{f(m),p_m} \neq \beta_{f(m),p'_m}$ and $\lim_{m\to\infty} \beta_{f(m),p_m} - \beta_{f(m),p'_m} = 0$. Fix m, let t and t' denote $|\mathcal{G}(\boldsymbol{\beta}_{f(m)}; \beta_{f(m),p_m})|$ and $|\mathcal{G}(\boldsymbol{\beta}_{f(m)}; \beta_{f(m),p'_m})|$. For any $p'' \in \{1, \ldots, p\}$ such that $\beta_{f(m),p''} \neq 0$ and $|\mathcal{G}(\boldsymbol{\beta}_{f(m)}; \beta_{f(m),p''})| = t'' > 1$, we create $\tilde{\boldsymbol{\theta}}$ whose grouping is the same as $\boldsymbol{\theta}_{f(m)}$ except that the $\beta_{f(m),p_m}$-group and $\beta_{f(m),p'_m}$-group in $\boldsymbol{\theta}_{f(m)}$ are merged as a new group and that $\beta_{f(m),p''}$ is singled out as a new group. Then

$$
\begin{aligned}
&G_L(\boldsymbol{\theta}_{f(m)}, \boldsymbol{\theta}_{f(m)-1}) - G_L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}_{f(m)-1}) \\
&= \begin{cases}
-(\frac{1}{t} + \frac{1}{t'})^{-1}(\beta_{f(m),p_m} - \beta_{f(m),p'_m})^2 + \frac{t+t'}{t+t'-1}(\frac{t\beta_{f(m),p_m} + t'\beta_{f(m),p'_m}}{t+t'} - c_{f(m)-1,p''})^2 \leq 0, \\
\text{if } \beta_{f(m),p''} = \beta_{f(m),p_m} \text{ or } \beta_{f(m),p'_m}; \\
-(\frac{1}{t} + \frac{1}{t'})^{-1}(\beta_{f(m),p_m} - \beta_{f(m),p'_m})^2 + \frac{t''}{t''-1}(\beta_{f(m),p''} - c_{f(m)-1,p''})^2 \leq 0, \text{otherwise.}
\end{cases}
\end{aligned}
$$

So for any $p'' \in \{1, \ldots, p\}$ such that $\beta_{f(m),p''} \neq 0$, we have $\frac{1}{L}|\frac{\partial g(\boldsymbol{\theta}_{f(m)-1})}{\partial \beta_{p''}}| = |\beta_{f(m)-1,p''} - c_{f(m)-1,p''}| \leq |\beta_{f(m)-1,p''} - \beta_{f(m),p''}| + |\beta_{f(m),p''} - c_{f(m)-1,p''}| \leq \left\|\boldsymbol{\theta}_{f(m)-1} - \boldsymbol{\theta}_{f(m)}\right\|_2 + (s+1)|\beta_{f(m),p_m} - \beta_{f(m),p'_m}|$.

(b) On top of the proof of Statement 3a, for fixed m and any $p', p'' \in \{1, \ldots, p\}$ such that $\beta_{f(m),p'} = 0$ and $\beta_{f(m),p''} \neq 0$, we create $\tilde{\boldsymbol{\theta}}$ whose grouping is the same with $\boldsymbol{\theta}_{f(m)}$ except that the $\beta_{f(m),p_m}$-group and $\beta_{f(m),p'_m}$-group in $\theta_{f(m)}$

are merged as a new group and that $\beta_{p'}$ is singled out as a new group and $\beta_{p''}$ is put in 0-group. Let t'' denote $|\mathcal{G}(\boldsymbol{\beta}_{f(m)}; \beta_{f(m),p''})|$. Then

and

$$
\begin{aligned}
&G_L(\theta_{f(m)}, \theta_{f(m)-1}) - G_L(\tilde{\theta}, \theta_{f(m)-1}) \\
&= \begin{cases}
-(\frac{1}{t} + \frac{1}{t'})^{-1}(\beta_{f(m),p_m} - \beta_{f(m),p'_m})^2 + \frac{t+t'}{t+t'-1}(\frac{t\beta_{f(m),p_m} + t'\beta_{f(m),p''_m}}{t+t'} - c_{f(m)-1,p'})^2 \\
-c^2_{f(m)-1,p''} + c^2_{f(m)-1,p'} \leq 0, \text{if } \beta_{f(m),p''} = \beta_{f(m),p_m} \text{ or } \beta_{f(m),p'_m}; \\
-(\frac{1}{t} + \frac{1}{t'})^{-1}(\beta_{f(m),p_m} - \beta_{f(m),p'_m})^2 + \frac{I(t''>1)t''}{t''-1}(\beta_{f(m),p''} - c_{f(m)-1,p''})^2 \\
-c^2_{f(m)-1,p''} + c^2_{f(m)-1,p'} \leq 0, \text{otherwise.}
\end{cases}
\end{aligned}
$$

Since $\boldsymbol{\theta}_{f(m)}$ converges to $\boldsymbol{\theta}'$, we have $\lim_{m\to\infty} \boldsymbol{\theta}_{f(m)-1} = \boldsymbol{\theta}'$ and $\nabla g(\boldsymbol{\theta}_{f(m)-1})$ and $c^2_{f(m)-1,p'} - c^2_{f(m)-1,p''}$ converge. So $\lim_{m\to\infty} c^2_{f(m)-1,p'} - c^2_{f(m)-1,p''} \leq 0$. So $\boldsymbol{\theta}$ is a first-order stationary point.

□

A.1.3 Proof of Theorem III.16

Proof. By Fano's lemma in (*Ibragimov and Has'minskii*, 1981), for all sequences of $t \geq 2$ probability distributions $\{P_1, \ldots, P_t\}$ on the same measurable space, and events $A_1, \ldots, A_t$ that form a partition of the space, we have that

$$
t^{-1} \sum_{j=1}^{t} P_j(A_j) \leq \sum_{1\leq j,k\leq t} \frac{K(P_j, P_k) + \log 2}{t^2 \log(t-1)},
$$

where $K(P_j, P_k)$ is the Kullback-Leibler information for distributions P_j versus P_k. Let $S = \{\boldsymbol{\beta}^{(j)}\}_{j=0}^{Kp}$ be a collection of parameters of distinct groupings with components belonging to $V \in \{\frac{1}{K}\gamma_{\min}, \frac{2}{K}\gamma_{\min}, \ldots, \frac{K}{K}\gamma_{\min}, 0\}$ such that for any $0 \leq j, j' \leq Kp$, we have $\left\|\boldsymbol{\beta}^{(j)} - \boldsymbol{\beta}^{(j')}\right\|_2^2 \leq 4\gamma^2_{\min}$. For example, we can set $\boldsymbol{\beta}^{(0)}$ as any parameter with components belonging to V such that $|\mathcal{G}(\beta^{(0)}; 0)| = p-s+1$ and has at least $K-1$ non-

empty non-zero groups. at most one empty non-zero group. Then we can get elements in S by switching each component in $\boldsymbol{\beta}^{(0)}$ to other $K-1$ values in V. There are two possible situations where this switching will generate duplicate groupings. One is that switching the only component in a non-zero group to another one-component non-zero group. We fix it by switching both one-component groups to zero. The other is that switching the only component in a non-zero group to an empty non-zero group. We fix it by switching a component in zero group to the empty group additionally. Note that

$$\begin{aligned} K(N(\boldsymbol{X}\boldsymbol{\beta}^{(j)},\sigma^2 I_n), N(\boldsymbol{X}\boldsymbol{\beta}^{(j')},\sigma^2 I_n)) &= \frac{1}{2\sigma^2}\left\|\boldsymbol{X}(\boldsymbol{\beta}^{(j)}-\boldsymbol{\beta}^{(j')})\right\|_2^2 \\ &\leq \frac{2\max_{1\leq j\leq p}\|\boldsymbol{x}_j\|_2^2\,\gamma_{\min}^2}{\sigma^2} \leq \frac{2nr(\boldsymbol{X},\boldsymbol{Z},s)\max_{\boldsymbol{\theta}:\boldsymbol{\beta}\in S} C_{\min}(\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{Z},s)}{\sigma^2}. \end{aligned}$$

Fano's lemma with probability distributions $\{N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2 I_n)\,|\,\boldsymbol{\beta}\in S\}$ shows that for any estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$ we have

$$(Kp+1)^{-1}\sum_{j\in S} P_j\big(\mathcal{G}(\hat{\boldsymbol{\beta}})=\mathcal{G}(\boldsymbol{\beta}^{(j)})\big) \leq \frac{2nr(\boldsymbol{X},\boldsymbol{Z},s)\max_{\boldsymbol{\theta}:\boldsymbol{\beta}\in S} C_{\min}(\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{Z},s)+\sigma^2\log 2}{\sigma^2\log(Kp)}.$$

Thus

$$\sup_{\substack{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}(K,s),\\ C_{\min}(\boldsymbol{\theta}_0,\boldsymbol{X},\boldsymbol{Z},s)\leq\\ \max_{\boldsymbol{\theta}:\boldsymbol{\beta}\in S} C_{\min}(\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{Z},s)}} \mathbb{P}(\mathcal{G}(\hat{\boldsymbol{\beta}})\neq\mathcal{G}(\boldsymbol{\beta}_0)) \geq 1-\frac{2nr(\boldsymbol{X},\boldsymbol{Z},s)\max_{\boldsymbol{\theta}:\boldsymbol{\beta}\in S} C_{\min}(\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{Z},s)+\sigma^2\log 2}{\sigma^2\log(Kp)}.$$

When $\gamma_{\min}$ varies from 0 to ∞, $\max_{\boldsymbol{\theta}\in S} C_{\min}(\boldsymbol{\theta},\boldsymbol{X},\boldsymbol{Z},s)$ varies from 0 to ∞, too. Then for any $L>0$ we have:

$$\sup_{\substack{\boldsymbol{\theta}_0\in\boldsymbol{\Theta}(K,s),\\ C_{\min}(\boldsymbol{\theta}_0,\boldsymbol{X},\boldsymbol{Z},s)\leq L}} \mathbb{P}\big(\mathcal{G}(\hat{\boldsymbol{\beta}})\neq\mathcal{G}(\boldsymbol{\beta}_0)\big) \geq 1-\frac{2nr(\boldsymbol{X},\boldsymbol{Z},s)L+\sigma^2\log 2}{\sigma^2\log(Kp)}.$$

When $L \leq \frac{(1-c)\sigma^2 \log(Kp)}{2nr(\boldsymbol{X},\boldsymbol{Z},s)}$, we have $\sup_{\substack{\boldsymbol{\theta}_0 \in (K,s), \\ C_{\min}(\boldsymbol{\theta}_0,\boldsymbol{X},\boldsymbol{Z},s) \leq L}} \mathbb{P}\big(\mathcal{G}(\hat{\boldsymbol{\beta}}) \neq \mathcal{G}(\boldsymbol{\beta}_0)\big) \geq c$. Then $\sup_{\boldsymbol{\theta}_0 \in B_0(K,s,\ell)} \mathbb{P}\big(\mathcal{G}(\hat{\boldsymbol{\beta}}) \neq \mathcal{G}(\boldsymbol{\beta}_0)\big) \to 0$, as $n, p \to \infty$ implies $l \geq \frac{\sigma^2 \log(Kp)}{2nr(\boldsymbol{X},\boldsymbol{Z},s)}$. □

A.1.4 Proof of Theorem III.18

Proof. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}(K_0, s_0)$, define $\boldsymbol{P}_{\mathcal{G}(\boldsymbol{\theta})}$ as the projection matrix of $(\boldsymbol{X}_{\mathcal{G}(\boldsymbol{\theta})}, \boldsymbol{Z})$.

$$
\begin{aligned}
&\mathbb{P}\left(\min_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K_0,s_0) \\ \mathcal{G}(\boldsymbol{\beta}) \neq \mathcal{G}(\boldsymbol{\beta}_0)}} \big\|\boldsymbol{Y} - (\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}\big\|_2^2 < \big\|\boldsymbol{Y} - (\boldsymbol{X},\boldsymbol{Z})\hat{\boldsymbol{\theta}}^{ol}\big\|_2^2\right) \\
=&\mathbb{P}\left(2\boldsymbol{\varepsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0 + \big\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0\big\|_2^2 - \boldsymbol{\varepsilon}^T(\boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta}_0)})\boldsymbol{\varepsilon} < 0\right) \\
\leq&\mathbb{P}\left(2\boldsymbol{\varepsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0 + \delta\big\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0\big\|_2^2 < 0\right) + \\
&\mathbb{P}\left((1-\delta)\big\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0\big\|_2^2 - \boldsymbol{\varepsilon}^T(\boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta}_0)})\boldsymbol{\varepsilon} < 0\right), \text{ for any } 0 < \delta < 1 \\
\leq&\mathbb{E}\left[\exp\left\{-\frac{2t_1\boldsymbol{\varepsilon}^T(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0}{\sigma^2}\right\}\right]\exp\left\{-\frac{t_1\delta\big\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0\big\|_2^2}{\sigma^2}\right\} + \\
&\mathbb{E}\left[\exp\left\{\frac{t_2\boldsymbol{\varepsilon}^T(\boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta}_0)})\boldsymbol{\varepsilon}}{\sigma^2}\right\}\right]\exp\left\{-\frac{t_2(1-\delta)\big\|(\boldsymbol{I} - \boldsymbol{P}_{\mathcal{G}(\boldsymbol{\beta})})(\boldsymbol{X},\boldsymbol{Z})\boldsymbol{\theta}_0\big\|_2^2}{\sigma^2}\right\}, \\
&\text{for any } 0 < t_1, t_2 < 1/2 \text{ by Markov's inequality} \\
\leq&\exp\left\{\frac{2t_1^2 - t_1\delta}{\sigma^2} nd(\boldsymbol{\beta},\boldsymbol{\beta}_0)C_{\min}\right\} + \\
&\exp\left\{-\frac{(1-\delta)t_2 nd(\boldsymbol{\beta},\boldsymbol{\beta}_0)C_{\min}}{\sigma^2} + 2t_2|\mathcal{G}(\boldsymbol{\beta})\backslash\mathcal{G}(\boldsymbol{\beta}_0)|\right\}, \\
&\text{when } 2t_1 < \delta \text{ due to Lemma 4 in (\textit{Shen et al.}, 2013) and proof of Theorem 2 in (\textit{Shen et al.}, 201} \\
\leq&2\exp\left\{-\frac{nd(\boldsymbol{\beta},\boldsymbol{\beta}_0)C_{\min}}{18\sigma^2} + \frac{2}{3}|\mathcal{G}(\boldsymbol{\beta})\backslash\mathcal{G}(\boldsymbol{\beta}_0)|\right\}, \text{ when } t_1 = t_2 = \frac{1}{3} \text{ and } \delta = \frac{5}{6}.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \\
\leq & \sum_{\omega \in \{\mathcal{G}(\boldsymbol{\beta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}(K_0, s_0), \mathcal{G}(\boldsymbol{\beta}) \neq \mathcal{G}(\boldsymbol{\beta}_0)\}} \mathbb{P}\bigg(\min_{\substack{\boldsymbol{\theta} \in \boldsymbol{\Theta}(K_0, s_0) \\ \mathcal{G}(\boldsymbol{\beta}) = \omega}} \big\| \boldsymbol{Y} - (\boldsymbol{X}, \boldsymbol{Z}) \boldsymbol{\theta} \big\|_2^2 < \big\| \boldsymbol{Y} - (\boldsymbol{X}, \boldsymbol{Z}) \hat{\boldsymbol{\theta}}^{ol} \big\|_2^2 \bigg) \\
\leq & \sum_{i=1}^{s_0} \sum_{j=0}^{i} \binom{s_0}{i} K_0^i \binom{p - s_0}{j} K_0^j \, 2 \exp\bigg(- \frac{niC_{\min}}{18\sigma^2} + \frac{2}{3}(2i + j) \bigg) \\
\leq & \sum_{i=1}^{s_0} 2 \exp\bigg(- \frac{niC_{\min}}{18\sigma^2} + \frac{4}{3} i + i \log(K_0 s_0) \bigg) \sum_{j=0}^{i} \exp\Big[j \big\{ \frac{2}{3} + \log\big(K_0 (p - s_0) \big) \big\} \Big] \\
\leq & \frac{2}{1 - e^{-2/3}} \sum_{i=1}^{s_0} \exp\bigg\{ - \frac{niC_{\min}}{18\sigma^2} + \frac{4}{3} i + i \log(K_0 s_0) + \frac{2}{3} i + i \log\big(K_0 (p - s_0) \big) \bigg\} \\
\leq & \frac{2}{1 - e^{-2/3}} \sum_{i=1}^{s_0} \exp\bigg\{ - \frac{niC_{\min}}{18\sigma^2} + (2 - \log(4)) i + i \log\big(K_0 p \big) \bigg\}, \\
& \text{due to } \log\big(K_0 (p - s_0) \big) + \log(K_0 s_0) \leq \log\big(\frac{K_0^2 p^2}{4} \big) \leq 2 \log(K_0 p) - \log(4) \\
\leq & \frac{2}{1 - e^{-2/3}} \frac{\exp\Big\{ - \frac{n}{18\sigma^2} \big(C_{\min} - 36\sigma^2 \log(K_0 p)/n - 18(2 - \log 4)\sigma^2/n \big) \Big\}}{1 - \exp\Big\{ - \frac{n}{18\sigma^2} \big(C_{\min} - 36\sigma^2 \log(K_0 p)/n - 18(2 - \log 4)\sigma^2/n \big) \Big\}} \\
& \text{when } C_{\min} \geq 36\sigma^2 \frac{\log(pK_0) + 1 - \frac{\log 4}{2}}{n}.
\end{aligned}
$$

Due to $\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \leq 1$, we have

$$
\mathbb{P}(\hat{\boldsymbol{\theta}} \neq \hat{\boldsymbol{\theta}}^{ol}) \leq \Big(\frac{2}{1 - e^{-2/3}} + 1 \Big) \exp\bigg\{ - \frac{n}{18\sigma^2} \bigg(C_{\min} - 36\sigma^2 \frac{\log(pK_0) + 1 - \frac{\log 4}{2}}{n} \bigg) \bigg\}.
$$

□

APPENDIX B

Appendices for Chapter III

B.1 Proofs for Chapter III

B.1.1 Proof of Proposition III.3

Given $\tilde{\boldsymbol{\theta}} \in H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{c})$, obviously we have $\tilde{\boldsymbol{\theta}}[\gamma] = \boldsymbol{c}[\gamma]$, $\tilde{\boldsymbol{\theta}}[\alpha_i] = \boldsymbol{c}[\alpha_i]$ or $\mathbf{0}$ and $\tilde{\boldsymbol{\theta}}[\beta_i] = \boldsymbol{c}[\beta_i]$ or $\mathbf{0}$ for $i = 1, \ldots, p$. Let $\boldsymbol{\theta}$ denote the output of Algorithm 1. If $g(\boldsymbol{\theta}) > g(\tilde{\boldsymbol{\theta}})$, then there exists a smallest $j \in \{1, \ldots, 2p\}$ s.t. $\boldsymbol{\theta}[\delta(j)] = \mathbf{0}$ and $\tilde{\boldsymbol{\theta}}[\delta(j)] = \boldsymbol{c}[\delta(j)] \neq \mathbf{0}$. There are two possible situations in the jth for loop of Step 5. One is $u^{\alpha+\beta} = U^{\alpha+\beta}$. Since $\boldsymbol{c}[\delta(j)] \neq \mathbf{0}$, then $\sum_{i=1}^{p} \boldsymbol{\theta}[\eta_i^{\alpha}] + \boldsymbol{\theta}[\eta_i^{\beta}] = U^{\alpha+\beta}$. This implies $\boldsymbol{\theta}[\delta(k)] = \mathbf{0}$ for $k \geq j$. Thus if $\boldsymbol{\theta}[\delta(k)] = \mathbf{0}$ and $\tilde{\boldsymbol{\theta}}[\delta(k)] = \boldsymbol{c}[\delta(k)] \neq \mathbf{0}$ then $k \geq j$. If $\boldsymbol{\theta}[\delta(k)] = \boldsymbol{c}[\delta(k)] \neq \mathbf{0}$ and $\tilde{\boldsymbol{\theta}}[\delta(k)] = \mathbf{0}$ then $k < j$. Due to $\tilde{\boldsymbol{\theta}} \in H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{c})$, we have $g(\boldsymbol{\theta}) \leq g(\tilde{\boldsymbol{\theta}})$. Contradiction. The other is $u^{\alpha+\beta} < U^{\alpha+\beta}$ and $\boldsymbol{\theta}[(\Gamma \circ \gamma)(j)] \neq \mathbf{0}$ and $u^{\alpha\beta} = U^{\alpha\beta}$. This implies $\sum_{i=1}^{p} \boldsymbol{\theta}[\eta_i^{\alpha\beta}] = U^{\alpha\beta}$. Since $\tilde{\boldsymbol{\theta}} \in H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{c})$, there exists $k < j$ such that $\boldsymbol{\theta}[\delta(k)] \neq \mathbf{0}$, $\boldsymbol{\theta}[(\Gamma \circ \delta)(k)] \neq \mathbf{0}$ and $\tilde{\boldsymbol{\theta}}[\delta(k)] = \mathbf{0}$. Then we have $g(\tilde{\boldsymbol{\theta}}') \leq g(\tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}}' \in H_{U^{\alpha\beta},U^{\alpha+\beta}}(\boldsymbol{c})$ equals to $\tilde{\boldsymbol{\theta}}$ except that $\tilde{\boldsymbol{\theta}}'[\delta(j)] = \mathbf{0}$, $\tilde{\boldsymbol{\theta}}'[\delta(k)] = \boldsymbol{c}[\delta(k)]$. We restart the above discussion for $\tilde{\boldsymbol{\theta}}'$ instead of $\tilde{\boldsymbol{\theta}}$. Note that in this new discussion the j is larger. Repeat this routine until $j > 2p$. Contradiction.

B.1.2 Proof of Proposition III.5

1. It follows by Proposition III.6 Statement 1.

2. If $\boldsymbol{\theta}$ is a solution to Problem III.1 but not a first-order stationary point, then exists $\tilde{\boldsymbol{\theta}} \in \Theta(U^{\alpha\beta}, U^{\alpha+\beta})$ s.t. $g(\tilde{\boldsymbol{\theta}}) \leq G_L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) < G_L(\boldsymbol{\theta}, \boldsymbol{\theta}) = g(\boldsymbol{\theta})$. Contradiction.

3. If the first-order stationary point $\boldsymbol{\theta}$ satisfies the two conditions, then $\boldsymbol{\theta} = \boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta})$, due to Algorithm 1. Thus $\nabla g(\boldsymbol{\theta}) = 0$. Since g is convex, then θ is a solution to the problem.

B.1.3 Proof of Proposition III.6

1. By Proposition III.3, we have:

$$\begin{aligned} g(\boldsymbol{\theta}_t) = G_L(\boldsymbol{\theta}_t, \boldsymbol{\theta}_t) &\geq G_L(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) = G_l(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) + \frac{L-l}{2}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_2^2 \\ &\geq g(\boldsymbol{\theta}_{t+1}) + \frac{L-l}{2}\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\|_2^2. \end{aligned}$$

   The above inequality implies $g(\boldsymbol{\theta}_t)$ is decreasing. Since $g(\boldsymbol{\theta})$ has finite lower bound, $g(\boldsymbol{\theta}_t)$ is convergent.

2. Obvious due to Proposition III.6 Statement 1.

B.1.4 Proof of Proposition III.8

1. If $\boldsymbol{\theta}_t[\eta_i^\alpha]$ does not converge and $\liminf_{t\to\infty} \boldsymbol{\theta}_t[\tau] > 0$, then there exists subsequence $\boldsymbol{\theta}_{f(t)}$ s.t. $\liminf_{t\to\infty} \left\|\boldsymbol{\theta}_{f(t)+1} - \boldsymbol{\theta}_{f(t)}\right\|_2 \geq \liminf_{t\to\infty} \boldsymbol{\theta}_t[\tau] > 0$. Contradiction to Proposition III.6 Statement 2.

2. Let $\boldsymbol{\theta} = \lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}$, $\boldsymbol{c}_t = \boldsymbol{\theta}_t - \frac{1}{L}\nabla g(\boldsymbol{\theta}_t)$, $\boldsymbol{c} = \boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}}$ denote output of Algorithm 1 with $\boldsymbol{c}$ as input. From Proposition III.8 Statement 1a, we have

$\boldsymbol{\theta}[\eta^\alpha] = \lim_{t\to\infty} \boldsymbol{\theta}_t[\eta^\alpha]$ and $\boldsymbol{\theta}[\eta^\beta] = \lim_{t\to\infty} \boldsymbol{\theta}_t[\eta^\beta]$.

For any $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ s.t. $\boldsymbol{\theta}[s] = \mathbf{0}$, there are two possible situations. One is $\boldsymbol{\theta}[\Gamma(s)] = \mathbf{0}$, then for any $s' \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ s.t. $\boldsymbol{\theta}[s'] \neq \mathbf{0}$, we have $\big\|\boldsymbol{c}[s']\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s']\big\|_2 \geq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s]\big\|_2 = \big\|\boldsymbol{c}[s]\big\|_2$. If $\sum_{i=1}^p \boldsymbol{\theta}[\eta_i^\alpha] + \boldsymbol{\theta}[\eta_i^\beta] = U^{\alpha+\beta}$, by Algorithm 1, then $\tilde{\boldsymbol{\theta}}[s] = \mathbf{0}$. If $\sum_{i=1}^p \boldsymbol{\theta}[\eta_i^\alpha] + \boldsymbol{\theta}[\eta_i^\beta] < U^{\alpha+\beta}$, then $\boldsymbol{c}[s] = \lim_{t\to\infty} \boldsymbol{c}_{f(t)}[s] = \mathbf{0}$, thus $\tilde{\boldsymbol{\theta}}[s] = \mathbf{0}$. The other is $\boldsymbol{\theta}[\Gamma(s)] \neq \mathbf{0}$, then $\boldsymbol{c}[s] \leq \boldsymbol{c}[\Gamma(s)]$ and for any $j \in \{1, \ldots, p\}$ s.t. $\boldsymbol{\theta}[\alpha_j] \neq \mathbf{0}$ and $\boldsymbol{\theta}[\beta_j] \neq \mathbf{0}$, we have $\big\|\boldsymbol{c}[\alpha_j]\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[\alpha_j]\big\|_2 \geq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s]\big\|_2 = \big\|\boldsymbol{c}[s]\big\|_2$ and $\big\|\boldsymbol{c}[\beta_j]\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[\beta_j]\big\|_2 \geq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s]\big\|_2 = \big\|\boldsymbol{c}[s]\big\|_2$. If $\sum_{i=1}^p \boldsymbol{\theta}[\eta_i^{\alpha\beta}] = U^{\alpha\beta}$, by Algorithm 1, then $\tilde{\boldsymbol{\theta}}[s] = \mathbf{0}$. If $\sum_{i=1}^p \boldsymbol{\theta}[\eta_i^{\alpha\beta}] < U^{\alpha\beta}$, then for any $s' \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ s.t. $\boldsymbol{\theta}[s'] \neq \mathbf{0}$, we have $\big\|\boldsymbol{c}[s']\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s']\big\|_2 \geq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s]\big\|_2 = \big\|\boldsymbol{c}[s]\big\|_2$, thus $\tilde{\boldsymbol{\theta}}[s] = \mathbf{0}$.

For any $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ s.t. $\boldsymbol{\theta}[s] \neq \mathbf{0}$, by Algorithm 1, we have $\tilde{\boldsymbol{\theta}}[s] = \boldsymbol{c}[s]$. From Proposition III.8 Statement 1a, we have $\big(\nabla g(\boldsymbol{\theta})\big)[s] = \lim_{t\to\infty}\big(\nabla g(\boldsymbol{\theta}_{f(t)})\big)[s] = 0$. Then $\tilde{\boldsymbol{\theta}}[s] = \boldsymbol{\theta}[s]$.

So $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$.

3. Let $\boldsymbol{c}_t = \boldsymbol{\theta}_t - \frac{1}{L}\nabla g(\boldsymbol{\theta}_t)$. Given $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ and t, there are three situations. First, if $\boldsymbol{\theta}_{f(t)}[s] \neq \mathbf{0}$, then

$$\big\|\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s]\big\|_2 = L\big\|\boldsymbol{\theta}_{f(t)}[s] - \boldsymbol{\theta}_{f(t)-1}[s]\big\|_2.$$

Second, if $\boldsymbol{\theta}_{f(t)}[s] = \boldsymbol{\theta}_{f(t)-1}[s] = \mathbf{0}$, then

$$\big\|\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s]\big\|_2 \leq L\boldsymbol{\theta}_{f(t)}[\rho] + L\boldsymbol{\theta}_{f(t)}[\tau].$$

Third, if $\boldsymbol{\theta}_{f(t)}[s] = \mathbf{0}$ and $\boldsymbol{\theta}_{f(t)-1}[s] \neq \mathbf{0}$, then

$$\begin{aligned}\big\|\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s]\big\|_2 &\leq L\big\|\boldsymbol{c}_{f(t)-1}[s]\big\|_2 + L\big\|\boldsymbol{\theta}_{f(t)}[s] - \boldsymbol{\theta}_{f(t)-1}[s]\big\|_2 \\ &\leq L\boldsymbol{\theta}_{f(t)}[\rho] + L\boldsymbol{\theta}_{f(t)}[\tau] + L\big\|\boldsymbol{\theta}_{f(t)}[s] - \boldsymbol{\theta}_{f(t)-1}[s]\big\|_2.\end{aligned}$$

Due to Proposition III.8 Statement 1b, we have $\lim_{t\to\infty} \nabla g(\boldsymbol{\theta}_{f(t)-1}) = \lim_{t\to\infty} \nabla g(\boldsymbol{\theta}_{f(t)}) = \mathbf{0}$.

4. Obvious due to Proposition III.8 Statement 2a.

5. If $\boldsymbol{\theta}_t[\eta^{\alpha\beta}]$ does not converge and $\liminf_{t\to\infty} \boldsymbol{\theta}_t[\rho] > 0$, then there must exists subsequence $\boldsymbol{\theta}_{h(t)}$ s.t. $\big\|\boldsymbol{\theta}_{h(t)} - \boldsymbol{\theta}_{h(t)-1}\big\|_2 \geq \boldsymbol{\theta}_{h(t)}[\rho] > 0$. Contradiction to Proposition III.6 Statement 2.
   Let $\boldsymbol{c}_t = \boldsymbol{\theta}_t - \frac{1}{L}\nabla g(\boldsymbol{\theta}_t)$ and $\lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}[\tau] = 0$. Given $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$ and t, consider two situations. One is that if $\boldsymbol{\theta}_{f(t)}[s] \neq \mathbf{0}$, then

$$\big\|\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s]\big\|_2 = L\big\|\boldsymbol{\theta}_{f(t)}[s] - \boldsymbol{\theta}_{f(t)-1}[s]\big\|_2.$$

   The other is that if $\boldsymbol{\theta}_{f(t)}[s] = \boldsymbol{\theta}_{f(t)}[\Gamma(s)] = \mathbf{0}$, then

$$\begin{aligned}\big\|\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s]\big\|_2 &\leq L\big\|\boldsymbol{c}_{f(t)-1}[s]\big\|_2 + L\big\|\boldsymbol{\theta}_{f(t)}[s] - \boldsymbol{\theta}_{f(t)-1}[s]\big\|_2 \\ &\leq L\boldsymbol{\theta}_{f(t)}[\tau] + L\big\|\boldsymbol{\theta}_{f(t)}[s] - \boldsymbol{\theta}_{f(t)-1}[s]\big\|_2.\end{aligned}$$

   Due to Proposition III.8 Statement 1b, we have

$$\lim_{t\to\infty}\Big\{1 - I\big(\boldsymbol{\theta}_{f(t)}[s] = \mathbf{0}, \boldsymbol{\theta}_{f(t)}[\Gamma(s)] \neq \mathbf{0}\big)\Big\}\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s] = \mathbf{0}.$$

6. Let $\boldsymbol{\theta} = \lim_{t\to\infty} \boldsymbol{\theta}_{f(t)}$, $\boldsymbol{c}_t = \boldsymbol{\theta}_t - \frac{1}{L}\nabla g(\boldsymbol{\theta}_t)$, $\boldsymbol{c} = \boldsymbol{\theta} - \frac{1}{L}\nabla g(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}}$ denote output of Algorithm 1 with $\boldsymbol{c}$ as input.

For any $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$, if $I\big(\boldsymbol{\theta}_{f(t)}[s] = \mathbf{0}, \boldsymbol{\theta}_{f(t)}[\Gamma(s)] \neq \mathbf{0}\big)$ does not converge to 1, then from Proposition III.8 Statement 3a there exists a subsequence $\boldsymbol{\theta}_{h(t)}$ of $\boldsymbol{\theta}_{f(t)}$ such that $\lim_{t\to\infty}\big(\nabla g(\boldsymbol{\theta}_{h(t)-1})\big)[s] = \mathbf{0}$. Since $\boldsymbol{\theta}_{f(t)}$ converges and Proposition III.6 Statement 2,

$$\big(\nabla g(\boldsymbol{\theta})\big)[s] = \lim_{t\to\infty}\big(\nabla g(\boldsymbol{\theta}_{f(t)})\big)[s] = \lim_{t\to\infty}\big(\nabla g(\boldsymbol{\theta}_{f(t)-1})\big)[s] = \lim_{t\to\infty}\big(\nabla g(\boldsymbol{\theta}_{h(t)-1})\big)[s] = \mathbf{0}.$$

For any $s \in \{\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p\}$, if $\lim_{t\to\infty} I\big(\boldsymbol{\theta}_{f(t)}[s] = \mathbf{0}, \boldsymbol{\theta}_{f(t)}[\Gamma(s)] \neq \mathbf{0}\big) = 1$, then

$$\lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[s]\big\|_2 \leq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[\Gamma(s)]\big\|_2$$

and there are two situations. One is $\sum_{i=1}^{p}\boldsymbol{\theta}[\eta_i^{\alpha\beta}] = U^{\alpha\beta}$, then for any $i \in \{1, \ldots, p\}$ such that $\lim_{t\to\infty}\boldsymbol{\theta}_{f(t)}[\eta_i^{\alpha\beta}] = 1$ we have

$$\begin{aligned}
\big\|\boldsymbol{c}[s]\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[s]\big\|_2 \leq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[\alpha_i]\big\|_2 = \big\|\boldsymbol{c}[\alpha_i]\big\|_2,\\
\big\|\boldsymbol{c}[s]\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[s]\big\|_2 \leq \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[\beta_i]\big\|_2 = \big\|\boldsymbol{c}[\beta_i]\big\|_2.
\end{aligned}$$

Thus $\tilde{\boldsymbol{\theta}}[s] = \mathbf{0}$ by Algorithm 1. The Other is $\sum_{i=1}^{p}\boldsymbol{\theta}[\eta_i^{\alpha\beta}] < U^{\alpha\beta}$, then

$$\big\|\boldsymbol{c}[s]\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)}[s]\big\|_2 = \lim_{t\to\infty}\big\|\boldsymbol{c}_{f(t)-1}[s]\big\|_2 \leq \lim_{t\to\infty}\boldsymbol{\theta}_{f(t)}[\tau] = 0.$$

Thus $\tilde{\boldsymbol{\theta}}[s] = \mathbf{0}$ by Algorithm 1. So $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}$.

BIBLIOGRAPHY

Auger, I. E., and C. E. Lawrence (1989), Algorithms for the optimal identification of segment neighborhoods, *Bulletin of mathematical biology*, *51*(1), 39–54.

Berthold, T. (2006), Primal heuristics for mixed integer programs.

Bertsimas, D., and R. Weismantel (2005), *Optimization over integers*, vol. 13, Dynamic Ideas Belmont.

Bertsimas, D., A. King, and R. Mazumder (2016), Best subset selection via a modern optimization lens, *The annals of statistics*, pp. 813–852.

Bertsimas, D., B. Van Parys, et al. (2020), Sparse high-dimensional regression: Exact scalable algorithms and phase transitions, *The Annals of Statistics*, *48*(1), 300–323.

Bondell, H. D., and B. J. Reich (2008), Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar, *Biometrics*, *64*(1), 115–123.

Boyd, S., L. Xiao, and A. Mutapcic (2003), Subgradient methods, *lecture notes of EE392o, Stanford University, Autumn Quarter*, *2004*, 2004–2005.

Braun, J. M., et al. (2012), Assessing windows of susceptibility to lead-induced cognitive deficits in mexican children, *Neurotoxicolog*, *33*(5), 1040–1047.

Cook, W., L. Lovász, P. D. Seymour, et al. (1995), *Combinatorial optimization: papers from the DIMACS Special Year*, vol. 20, American Mathematical Soc.

Dantzig, G., R. Fulkerson, and S. Johnson (1954), Solution of a large-scale traveling-salesman problem, *Journal of the operations research society of America*, *2*(4), 393–410.

Derkach, A., R. M. Pfeiffer, T.-H. Chen, and J. N. Sampson (2019), High dimensional mediation analysis with latent variables, *Biometrics*, *75*(3), 745–756.

Duran, M. A., and I. E. Grossmann (1986), An outer-approximation algorithm for a class of mixed-integer nonlinear programs, *Mathematical programming*, *36*(3), 307–339.

Eastman, W. L. (1958), Linear programming with pattern constraints: a thesis, Ph.D. thesis.

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.

Friedman, J., T. Hastie, and R. Tibshirani (2010), A note on the group lasso and a sparse group lasso, *arXiv preprint arXiv:1001.0736*.

Fritz, M. S., and D. P. MacKinnon (2007), Required sample size to detect the mediated effect, *Psychological science*, *18*(3), 233–239.

Gomory, R. E. (1960), Solving linear programming problems in integers, *Combinatorial Analysis*, *10*, 211–215.

Gouonr, R. (1958), Outline of an algorithm for integer solutions to linear programs, *Bull. Am. Math. Soc*, *64*, 3.

Gu, C. (1998), Model indexing and smoothing parameter selection in nonparametric function estimation, *Statistica Sinica*, pp. 607–623.

Hathaway, R. J., et al. (1985), A constrained formulation of maximum-likelihood estimation for normal mixture distributions, *The Annals of Statistics*, *13*(2), 795–800.

Hernán, M., and J. Robins (), *Causal Inference: What If,*, Boca Raton, FL. Chapman & Hall/CRC, unpublished.

Higham, N. J. (2002), Accuracy and stability of numerical algorithms, 2nd edition, p. 258, Siam.

Ibragimov, I., and R. Has'minskii (1981), *Statistical Estimation: Asymptotic Theory*, New York: Springer.

Jeon, J.-J., S. Kwon, and H. Choi (2017), Homogeneity detection for the high-dimensional generalized linear model, *Computational Statistics & Data Analysis*, *114*, 61–74.

Jünger, M., and G. Reinelt (2013), *Facets of Combinatorial Optimization*, Springer.

Ke, Y., J. Li, W. Zhang, et al. (2016), Structure identification in panel data analysis, *The Annals of Statistics*, *44*(3), 1193–1233.

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

Kelley, J. E., Jr (1960), The cutting-plane method for solving convex programs, *Journal of the society for Industrial and Applied Mathematics*, *8*(4), 703–712.

Kobrosly, R. W., L. E. Parlett, R. W. Stahlhut, E. S. Barrett, and S. H. Swan (2012), Socioeconomic factors and phthalate metabolite concentrations among united states women of reproductive age, *Environmental Research*, *115*, 11–17.

Land, A. H., and A. G. Doig (1960), An automatic method of solving discrete programming problems, *Econometrica: Journal of the Econometric Society*, pp. 497–520.

Lian, H., X. Qiao, and W. Zhang (2017), Homogeneity pursuit in single index models based panel data analysis, *arXiv preprint arXiv:1706.00857*.

Liu, Z., and G. Li (2016), Efficient regularized regression with penalty for variable selection and network construction, *Computational and mathematical methods in medicine*, *2016*.

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

MacKinnon, D. P., and J. H. Dwyer (1993), Estimating mediated effects in prevention studies, *Evaluation review*, *17*(2), 144–158.

MacKinnon, D. P., G. Warsi, and J. H. Dwyer (1995), A simulation study of mediated effect measures, *Multivariate behavioral research*, *30*(1), 41–62.

MacKinnon, D. P., C. M. Lockwood, J. M. Hoffman, S. G. West, and V. Sheets (2002), A comparison of methods to test mediation and other intervening variable effects., *Psychological methods*, *7*(1), 83.

Marie, C., F. Vendittelli, and M. P. Sauvant-Rochat (2015), Obstetrical outcomes and biomarkers toassess exposure to phthalates: A review, *Environment International*, *83*, 116–136.

Markowitz, H. M., and A. S. Manne (1957), On the solution of discrete programming problems, *Econometrica: journal of the Econometric Society*, pp. 84–110.

Marsee, K., T. J. Woodruff, D. A. Axelrad, A. M. Calafat, and S. H. Swan (2006), Estimated dailyphthalate exposures in a population of mothers of male infants exhibiting reduced anogenital distance, *Environmental Health Perspectives*, *114*, 805–809.

Miller, A. (2002), *Subset selection in regression*, CRC Press.

Molenberghs, G., and G. Verbeke (2007), Likelihood ratio, score, and wald tests in a constrained parameter space, *The American Statistician*, *61*(1), 22–27.

Morrison, D. R., S. H. Jacobson, J. J. Sauppe, and E. C. Sewell (2016), Branch-and-bound algorithms: A survey of recent advances in searching, branching, and pruning, *Discrete Optimization*, *19*, 79–102.

Natarajan, B. K. (1995), Sparse approximate solutions to linear systems, *SIAM journal on computing*, *24*(2), 227–234.

Optimization, C. (2019a), *Cplex Optimizer 12.9*.

Optimization, G. (2019b), *Gurobi Optimizer 8.1*.

Preacher, K. J. (2015), Advances in mediation analysis: A survey and synthesis of new developments, *Annual review of psychology*, *66*, 825–852.

Savelsbergh, M. W. (1994), Preprocessing and probing techniques for mixed integer programming problems, *ORSA Journal on Computing*, *6*(4), 445–454.

Schettler, T. (2006), Human exposure to phthalates via consumer products, *Journal of Andrology*, *29*, 134–139.

Schoenberg, R. (1997), Constrained maximum likelihood, *Computational Economics*, *10*(3), 251–266.

Serang, S., R. Jacobucci, K. C. Brimhall, and K. J. Grimm (2017), Exploratory mediation analysis via regularization, *Structural equation modeling: a multidisciplinary journal*, *24*(5), 733–744.

She, P., C. Van Horn, T. Reid, S. M. Hutson, R. N. Cooney, and C. J. Lynch (2007), Obesity-related elevations in plasma leucine are associated with alterations in enzymes involved in branched-chain amino acid metabolism, *American Journal of Physiology-Endocrinology and Metabolism*, *293*(6), E1552–E1563.

Shen, J., and X. He (2015), Inference for subgroup analysis with a structured logistic-normal mixture model, *Journal of the American Statistical Association*, *110*(509), 303–312.

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

Shen, X., W. Pan, and Y. Zhu (2012), Likelihood-based selection and sharp parameter estimation, *Journal of the American Statistical Association*, *107*(497), 223–232.

Shen, X., W. Pan, Y. Zhu, and H. Zhou (2013), On constrained and regularized high-dimensional regression, *Annals of the Institute of Statistical Mathematics*, *65*(5), 807–832.

Shor, N. Z. (1985), The subgradient method, in *Minimization methods for non-differentiable functions*, pp. 22–47, Springer.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013), A sparse-group lasso, *Journal of computational and graphical statistics*, *22*(2), 231–245.

Sion, M., et al. (1958), On general minimax theorems., *Pacific Journal of mathematics*, *8*(1), 171–176.

Sobel, M. E. (1982), Asymptotic confidence intervals for indirect effects in structural equation models, *Sociological methodology*, *13*, 290–312.

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

Tibshirani, R. (2011), Regression shrinkage and selection via the lasso: a retrospective, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282.

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.

Van Loan, C. F. (2000), The ubiquitous kronecker product, *Journal of computational and applied mathematics*, *123*(1-2), 85–100.

Vapnik, V. (1998), The support vector method of function estimation, in *Nonlinear Modeling*, pp. 55–85, Springer.

Wolsey, L. A. (2008), Mixed integer programming, *Wiley Encyclopedia of Computer Science and Engineering*.

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

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

Zhang, F. (2006), *The Schur complement and its applications*, vol. 4, Springer Science & Business Media.

Zhu, Y., X. Shen, and W. Pan (2013), Simultaneous grouping pursuit and feature selection over an undirected graph, *Journal of the American Statistical Association*, *108*(502), 713–725.