The Pennsylvania State University
The Graduate School

STATISTICAL INFERENCE FOR HIGH-DIMENSIONAL MEDIATION MODEL

A Dissertation in
Statistics
by
Mudong Zeng

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2021

The dissertation of Mudong Zeng was reviewed and approved by the following:

Runze Li
Eberly Family Chair Professor
Dissertation Advisor, Chair of Committee

Bing Li
Verne M. Willaman Professor of Statistics

Zhibiao Zhao
Associate Professor of Statistics

Hong Ma
Huck Distinguished Research Professor of Plant Molecular Biology

Ephraim Hanks
Associate Professor of Statistics
Chair of Graduate Studies

Abstract

High-dimensional mediation models (HDMM) draw increasing attention in many scientific areas such as genetic study, clinical trial, and internet analysis. This dissertation is concerned with a new statistical estimation and inference procedure for HDMM, and consists of three projects. In the first project, we propose an estimation procedure for the indirect effects of the models via a partial penalized least squares method, and further establish its theoretical properties. We propose a partial penalized Wald test on the indirect effects and prove that the proposed test has a χ^2 limiting null distribution. We also propose an F-type test for direct effects and show that the proposed test asymptotically follows a χ^2-distribution under the null hypothesis and a noncentral χ^2-distribution under local alternatives. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed tests and to compare its performance with existing tests. To illustrate the proposed methodology, we applied the newly proposed statistical inference procedures to study stock reaction to COVID-19 pandemic via an empirical analysis of studying the mediation effects of financial metrics that bridge the company's sector and stock return. In the second project, we conduct a case study by applying high-dimensional linear mediation models. We study the mediating role of DNA methylation relating childhood trauma and cortisol stress reactivity, with several clinical variables involved as confounders. We develop relevant tests for the direct and indirect effects of the early life trauma on cortisol stress reactivity. In the third project, as a natural extension of the HDMM, we propose high-dimensional

generalized mediation model (HDGMM) to deal with binary or count responses. We propose an estimation procedure for the indirect effects of the models via a partial penalized likelihood method, and further establish its theoretical properties. We also propose a partial penalized Wald test and a partial penalized likelihood ratio test for indirect and direct effects respectively. We show that the proposed test asymptotical distribution under the null hypothesis and local alternatives. Simulation studies are conducted to assess the performance of the estimation and inference procedures for HDGMM. We use HDGMM to classify valuable stocks in COVID-19 pandemic via exploring the underlying mechanism of the relationship between stock market sectors and stock returns.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am wholeheartedly grateful to Professor Runze Li's devotion in the guidance to me. When I encounter difficulties, Professor Li always patiently gives me hints and carefully enlightens me with a lot of illumination. I remember the time when I anxiously approached to him for help because I failed to estimate a mediation model after several attempts. He first calmed down me and then introduced a new algorithm with patience. Next, he put forward local linear approximation algorithm and ask me to try by myself. With his help, I learned quickly and gradually gain confidence in my research. Besides, he cares about my life and provides emotional supports. When I had no idea to teach undergraduate courses, he audited my class and gave me several useful suggestions that brought the lessons with vivid atmosphere. It is my great honor to follow and work with him.

I would also like to thank Professor Bing Li, Prof. Dr. Zhibiao Zhao and Professor Hong Ma for their extremely useful comments to this proposal. I took many courses from Professor Bing Li, and I really learned a lot. He explained profound theories in an easy-to-understand approach. I learned happily and deeply. I can quickly recall the knowledge from his classes when I work on research. Thanks to Dr. Zhibiao Zhao, I had a chance to teach a course that I am familiar with and bring out my potential to give a speech in front of a crowd. I regard this as a very precious opportunity. Professor Hong Ma gave me lots of guidance on biology knowledge and Intellectual Property transfer when I interned at office of innovation. Since then I was appealed to his research. I admire his rich knowledge in plant biology and appreciate his patient for imparting the biotechnologies in a lay-man language. Thanks to his help, I can closely work with his lab and contribute mediation model to their work. Finally, none of this would have been possible without my parents to whom I want to dedicate this dissertation. I have learned so much from my parents and I appreciate their unconditional love and constant support through the ups and downs of my life.

This research was partially supported by National Science Foundation grants DMS 1820702, DMS 1953196 and DMS 2015539.

Chapter 1

Introduction

Since the seminal work of Baron and Kenny (1986), mediation analysis has been used in various scientific research, such as economics, psychology, pedagogy, and behavioral science (MacKinnon, 2012; Conti et al., 2016; Hayes, 2017). It is designed to investigate the mechanisms whereby exposure variables affect an outcome through intermediate variables, which are termed as mediators (MacKinnon, 2012; Hayes, 2017; VanderWeele, 2015). For instance, in the field of policy evaluation, while there certainly is no shortage of techniques assessing effects of policies or other treatments on an outcome (Donald and Hsu, 2014; Athey et al., 2018; Ai et al., 2021), mediation analyses move a step further to disentangle such effect into indirect effects through mediators, such as certain economic indices, and direct effects. There are well-developed statistical methods for mediation model with low-dimensional mediators (VanderWeele and Vansteelandt, 2014; VanderWeele, 2015; Nguyen et al., 2015; Huang and Yang, 2017), but challenges arise when the dimension of potential mediators is high.

On account of modern data-collecting technology, mediation analysis extends its territory to quantitative finance, genomics, internet analysis, biomedical research, among other data-intensive fields (Ottaviani and Vandone, 2018; Peng et al., 2019; Chandrima et al., 2020; Esubalew and Raghurama, 2020; Song et al., 2020; Xu et al., 2021). This brings in high-dimensional mediators and requires attention on high-dimensional mediation model (HDMM), where the number of potential mediators is much larger than the sample size. For example, in stock performance analysis, hundreds of financial statements items and metrics are potential

covariates; in human body hormone regulation analysis using DNA methylation data, tens of thousands of expression of loci are potential predictors. The high-dimensionality, on every account, poses both computational and statistical challenges for carrying out efficient mediation analysis. For instance, the traditional structural equation modeling fails due to the rank-deficiency of the observed covariance matrix. However, notwithstanding the high dimensional mediation structure, the number of truly active mediators is typically assumed small and less than the sample size. This is referred to as the sparsity assumption in the literature, although the sparsity pattern is unknown and thus to be recovered. See, for example, Fan et al. (2020a) and references therein. Many existing methods in literature break through such obstacle by utilizing the dimension reduction techniques in regular linear models. For example, Huang and Pan (2016) and Chén et al. (2018) adopted principal components analysis to compress the dimensionality of mediators, and applied bootstrap for inference. These methods are intuitive and simple to implement, but lack theoretical justification about asymptotic distributions of the test statistics. As an extension of Huang and Pan (2016), Zhao et al. (2020) further introduced sparse principal component analysis to mediation models. Zhang et al. (2016) used a two-stage technique with (a) first screening out "unimportant" mediators, and then (b) applying existing procedures for the post-screened outcome model. Zhou et al. (2020) introduced debiased penalized estimators for the direct and indirect effects, with theoretical guarantees of the related tests. However, their method involves estimating high dimensional matrices, leading to potentially unstable estimates and expensive computation. Furthermore, imposing penalization on all parameters reduces the efficiency of estimators, and hence tests. There are many developments on this topic in the recent literature (Chakrabortty et al., 2018; Derkach et al., 2019; Song et al., 2020). In this dissertation, we propose new statistical inference procedures for HDMM. This dissertation consists of three projects, and their contributions are introduced below.

1.1 Linear mediation models with high dimensional mediators

In the first project, we propose statistical inference procedures for linear mediation models with high dimensional mediators. Statistical inference for high-dimensional data has been an active research topic in the literature (Belloni et al., 2014; Zhang and Zhang, 2014; VanderWeele and Vansteelandt, 2014; Javanmard and Montanari, 2014; Shi et al., 2019; Fan et al., 2020b, 2019). However, there are much less work on statistical inference for HDMM. To our best knowledge, Zhou et al. (2020) is the only one on testing hypothesis on indirect effect with solid theoretical analysis. Our inference procedure on indirect effect is distinguished from Zhou et al. (2020) in that we observe the indirect effect in HDMM indeed is a low dimensional parameter and is the difference between the total effect and the direct effect in the HDMM. This motivates us to develop a new statistical inference procedure for high dimensional linear mediation models.

We propose to estimate the total effect via least squares method and the direct effect by partial penalized least squares method, and then estimate the indirect effect by the difference between the estimates of the total effect and the direct effect. We establish the asymptotical normality of the indirect effect estimate and further develop a Wald test for the indirect effect. We estimate the direct effect in the HDMM by partial penalized least squares method, and propose an F-type test for it. The statistical inference on the direct effect essentially is the same as statistical inference on low dimensional coefficients in high-dimensional linear models. This topic has been studied under the setting in which the covariate vector in the high-dimensional linear models is fixed design (Zhang and Zhang, 2014; Van de Geer et al., 2014; Shi et al., 2019). Due to the nature of HDMM, the design matrix in HDMM must be random rather than fixed since mediators are random. Thus, the statistical setting studied in this dissertation is different from the one in Shi et al. (2019), in which the covariate vector is assumed to be fixed design. We study the asymptotical property of the proposed estimator in the random-design setting. The random design imposes challenges in deriving the rate of convergence and asymptotical normality of the partial penalized least squares estimates. Under mild regularity conditions, we prove the sparsity and

establish the rate of convergence of the partial penalized least squares estimate. We further establish an asymptotical representation of the estimate. Based on the asymptotical representation, we can easily derive the asymptotical normality of the estimate and derive the asymptotical distributions of the proposed test for the direct effect under null hypothesis and under local alternative.

We show that the proposed estimate of indirect effect is asymptotically more efficient than the one proposed in Zhou et al. (2020), and indeed is asymptotically efficient under normality assumption. This is because the debias step of debiased Lasso inflates the asymptotical variance of the resulting estimate. We conduct Monte Carlo simulation studies to assess the finite sample performance of the proposed estimate in terms of bias and variance and to examine Type I error and power of the proposed test. We also conduct numerical comparisons among the proposed estimate, the oracle estimate and the estimate proposed in Zhou et al. (2020). Our numerical comparison indicates that the proposed estimate performs as well as the oracle one, and outperforms the estimate proposed by Zhou et al. (2020).

We utilize the proposed method to study the mediator role of financial metrics that bridge company's sector and stock return. We select six financial metrics out of all the 550 that indeed mediate the pathways linking company sector and stock return, with interestingly and informatively financial interpretations. We also compare the metrics selected using our data during the COVID-19 period and those classical findings in existing works, including Fama and French (2015), Edirisinghe and Zhang (2007), among others. We indeed discover some unique patterns and features due to the pandemic. Moreover, according to the proposed tests for effects of sector, both its direct effect and indirect effect via financial metrics are statistically significant. Therefore, evaluating the selected financial metrics, as well as the sector information, might help investors to make wiser investment decisions and choose stocks especially during the pandemic.

1.2 A case study: an application of high dimensional linear mediation models.

In the second project, we apply the proposed procedures to study the mediation effect of DNA methylation in the relation between childhood trauma and cortisol stress reactivity. Childhood trauma plays a pivotal role in the development of psychiatric disorders across life span (Burke et al., 2005; Petrowski et al., 2013). Its persistently detrimental influence is typically realized via altering neuroendocrine substances like cortisol (Carpenter et al., 2007; Heim et al., 2000). Ever since the pilot study conducted by Luecken (1998), researchers thereof have sought for the mechanism relating cortisol change to various circumstances of childhood trauma, such as maltreatment (Carpenter et al., 2007), physical abuse (Bremner et al., 2003; Heim et al., 2000; Elzinga et al., 2010; Carpenter et al., 2011), early parental loss (Luecken, 1998; Kraft and Luecken, 2009), separation experience (Pesonen et al., 2010), among others.

On finding such relations, the aforementioned works nevertheless have not reached a concordant solution. This pushes through deeper exploration towards epigenetic alteration involved in the traumatic stress. Convincingly demonstrated by preclinical studies, childhood trauma tends to influence neuroendocrine system in adulthood via altering DNA methylation patterns. McGowan et al. (2009) studied the epigenetic regulation of glucocorticoid receptor (NR3C1) in human brain associated with childhood abuse. Perroud et al. (2011) showed that early life adverse events may permanently impact on the Hypothalamus-Pituitary-Adrenal axis (HPAA) though epigenetic modifications of NR3C1. Edelman et al. (2012) demonstrated epigenetic changes at the GR exon 1F correlate with HPAA reactivity measured by total cortisol (area under curve). See Vinkers et al. (2015) for a comparative review of literature regarding trauma-induced changes in DNA methylation in humans.

These works mainly concentrated on single-layer linear models, where effects of early life trauma on DNA methylation and effects of DNA methylation on HPAA or cortisol alteration are separately evaluated. However, DNA methylation ought to play a bridging role in the relation between childhood trauma and cortisol stress reactivity. In addition, all of their scientific findings are based on epigenetic

modifications of a single gene. In theory, this is unlikely to be the case and would result in estimation bias. In sight of such issues, Houtepen et al. (2016) conducted a genome-wide mediation analysis and identified a locus on the KITLG gene that mediates the relationship between childhood trauma and cortisol stress reactivity. Although starting at 385,882 DNA methylation loci, only the top three loci were selected for further investigation by the QQ plot of the p-values obtained from individual significance tests, with a total discard of the dependence structure and joint effects of DNA methylation. To account for the between-loci dependency, van Kesteren and Oberski (2019) proposed an embedding algorithm called coordinate-wise mediation filter (CMF), which consists of an inner loop and outer loop. A key strategy of CMF to address dependency is the use of residuals and projection when detecting loci in the inner loop. This CMF algorithm targets dichotomous decisions - whether each of the DNA methylation locus should be recognized as a mediator, while offers no guarantee of either statistical significance or model fits. Interestingly, van Kesteren and Oberski (2019) identified completely different DNA methylation loci from Houtepen et al. (2016), based on the same data set but using the CMF algorithm. In response to this contradiction, we carry out an in-depth mediation analysis for a thorough understanding of how early life trauma affects cortisol stress reactivity in adulthood via DNA methylation.

From a statistical point of view, this is a high dimensional mediation problem, with DNA methylation loci being potential mediators, the vast majority of which though are supposed to be inactive. Nothwithstanding no shortage of strategies dealing with high dimensional mediators, including those in Houtepen et al. (2016) and van Kesteren and Oberski (2019), most existing literature rely on the marginal screening or penalized regression for sparse estimation. See for instance Preacher and Hayes (2008), Zhang et al. (2016), Serang et al. (2017), and so forth. A pitfall of using these dimension reduction techniques in each or either layer of mediation models lies in the pertinent difference between penalizing paths and finding actual mediators. That is, they choose paths instead of mediators. As a potential insight to break through this obstacle, Zhou et al. (2020) proposed a debiased Lasso method that can integrate the two layers of high dimensional mediation models, and they also developed significance tests for both direct and indirect effects. However, the method proposed in Zhou et al. (2020) involves high dimensional matrix

estimation and operation, which might bring about a huge computational burden. In addition, the procedure penalizes all parameters, and the debiased step relies on the entire covariance matrix. This leads to inevitable efficiency loss of the estimators. More recently, Chapter 3 observed that despite of high dimensional mediators, the direct and indirect effects are both low dimensional, with sum being the total effect. They thereby proposed a partial penalized approach for estimating the direct effects, which avoids high dimensional matrix estimation and the debiased step, and thus enhances efficiency of proposed estimators. In spite of the plausible theory and efficient algorithms, Chapter 3 have not yet explicitly elucidated the method with potential confounders, which typically should be considered when studying traumatic effects on cortisol alteration via DNA methylation, as in the literature (Houtepen et al., 2016; van Kesteren and Oberski, 2019). Therefore, we extend the work of Chapter 3 to the models with confounders. Then we utilize our new procedure to study the mediating role of DNA methylation relating childhood trauma and cortisol stress reactivity, with several clinical variables involved as confounders. We further develop relevant tests for the direct and indirect effects of the early life trauma on cortisol stress reactivity.

Aside from the eight DNA methylation loci detected by Houtepen et al. (2016) and van Kesteren and Oberski (2019), we identify three additional loci on the RAB5IF gene (cg19230917), the CPQ gene (cg06422529) and the AGPAT1 gene (cg03199124) as mediators. We look through existing literature, and find reasonably neurobiological interpretations towards these three genes. Thus, our findings point out a potential direction for deeper neurobiological and epigenetic investigation of the connection between traumatic stress and cortisol alteration. From statistical point of view, we perform several statistical tests, and the results are also in support of the selected genes. According to our newly proposed tests for the direct and indirect effects, the childhood trauma influences cortisol reactivity only through DNA methylation, since the indirect effect is negatively significant, yet the direct effect is not significant. In the full model with all detected loci, those from the newly identified genes are all significant, while the KITLG gene (cg27512205) selected by Houtepen et al. (2016), the HNRNPF gene (cg12500973) and the ZSCAN30 gene (cg16657538) selected by van Kesteren and Oberski (2019) are no longer significant. However, models with only the genes in Houtepen et al.

(2016) yield a contradictory conclusion that KITLG is significant.

1.3 Generalized mediation models with high dimensional mediators

In the third project, we study the mediation model under the framework of generalized linear models whose response being either binary, count or categorical variable. Currently there are some established theories about statistical inference for generalized linear models, but there is few literature about high-dimensional generalized mediation model (HDGMM). Zhang et al. (2016) propose maximum likelihood with minimax concave penalty (MCP) (Zhang, 2010) for variable selection and joint significance test for testing the indirect effect. This procedure seems intuitive but is lack of theoretical guarantee of asymptotic distribution of the test statistics.

The third part contributions of the dissertation aims to develop statistical estimation and inference procedures for the indirect effect and direct effect in HDGMM. The indirect effect of HDGMM is defined as the product of a high-dimensional matrix and a high-dimensional vector. This makes the statistical inference on the indirect effect very challenging. To tackle this challenge, we propose a partial penalized maximum likelihood method, which may effectively deal with the high dimensionality and estimate related parameters. We establish the asymptotical normality of the proposed estimator of indirect effect, and develop a Wald test for the indirect effect. We show that under the null hypothesis, the proposed Wald test asymptotically follows a χ^2 distribution. We propose a partial penalized likelihood ratio test for direct effect based on constrained and unconstrained penalized maximum likelihood estimates. We show that the proposed likelihood ratio test follows χ^2 limiting null distribution. When the direct effect is absent, a more efficient estimate of the indirect effect is also developed.

We conduct Monte Carlo simulation studies to assess the finite sample performance of the proposed estimate in terms of bias and variance and to examine Type I error and power of the proposed test. We also conduct numerical comparisons among the proposed procedure, the oracle procedure and the procedure

proposed in Djordjilović et al. (2019). Our numerical comparison implies that the proposed procedure performs as well as the oracle one, and outperforms the procedure proposed by Djordjilović et al. (2019). Moreover, we illustrate the proposed methodology via an empirical analysis of U.S. stock data during early stage of the COVID-19 pandemic to gain insight of portfolio construction.

1.4 Organization

The rest of the proposal is organized as follows. Chapter 2 provides an overview of penalized least squares, and reviews several essential algorithms for estimation and inference.

Chapter 3 develops the estimation procedure of the total effect, direct effect and indirect effect in HDMM. In section 3.1, we propose a new statistical inference procedure for the indirect effect and establish its theoretical properties. We also construct an F-type test for the direct effect. Section 3.2 presents numerical studies and a real data example. Conclusion and discussion are given in section 3.3. All proofs about HDMM are presented in section 3.4.

Chapter 4 is a case study by applying HDMM to a real data: childhood trauma tends to influence cortisol stress reactivity through the mediating effects of DNA methylation. In section 4.2, we construct a partial penalized likelihood ratio test for the direct effect. Section 4.3 presents numerical studies. Conclusion and discussion are given in section 4.4. All proofs about HDMM with confounding variables are presented in section 4.5.

Chapter 5 develops the estimation procedure of direct and indirect effect for HDGMM. In section 5.1, we introduce a new statistical inference procedure for the indirect effect and establish its theoretical properties for HDGMM. In section 5.2, we construct a partial penalized likelihood ratio test for the direct effect. section 5.3 presents numerical studies. Conclusion and discussion are given in section 5.4. All proofs about HDGMM are presented in section 5.5.

Chapter 2

Literature Review

This chapter reviews the previous work that pertains to the penalized linear model and extension to generalized linear models. Starting from the challenging of high dimensional data analysis, the section 2.1 introduces penalized least-squares approaches to variable selection problems with focus on folded concave penalty function and discusses the potential algorithms for solving the folded concave penalty function. Section 2.2 extends the techniques pertain to penalized least squares to generalized linear model when the response variable is binary, categorical and count. Besides high-dimensional data analysis, this chapter also introduces mediation model and extends from low dimensional mediators setting to high-dimensional one. In addition, section 2.3 presents a review of statistical test on direct effect and indirect effect of mediation model.

2.1 Penalized least squares

High-dimensional data analysis is challenging when the number of independent variables p exceeds the sample size n, known as the curse of dimensionality. Variable selection is essential in high dimensional data analysis. This is because excluding unnecessary variables from the model can avoid over-fitting and reduce the prediction variance. It is reasonable to assume that only a few independent variables should be in the model. That is, many independent variables should have zero coefficients. This is the sparsity assumption, under which penalized least squares has been used for variable selection and coefficient estimation.

A classical variable selection technique is the best subset selection. Among all the models with the same number of variables, the best subset selection picks the model with smallest residual sum of squares. Intuitively, one prefers the better fit model among the models with same complexity. However, the computation cost of the best subset selection is expensive since it grows exponentially with data dimension. Therefore, it is impossible to conduct the best subset selection when the dimensionality is high. One solution is to formulate variable selection as penalized least squares. Tibshirani (1996) developed the least absolute shrinkage and selection operator (LASSO) and Frank and Friedman (1993) developed ℓ_2 ridge regression are examples of penalty function in penalized least squares. Both LASSO and ridge regression have huge computation advantage over the best subset selection.

Consider linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.1.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector, $\mathbf{X}$ is an $n \times d$ matrix and $\boldsymbol{\beta}$ is the parameter to estimate. Fan and Li (2001) proposed a unified framework for variable selection using penalized loss function.

2.1.1 Penalized least squares for variable selection

For linear model, the loss function is the least squares and the form of penalized least squares, the target function, is

$$Q(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{d} p_\lambda \left(|\beta_j|\right), \tag{2.1.2}$$

where $\mathbf{y}$ is the vector of responses, design matrix $\mathbf{X}$ is an $n \times p$ matrix of covariates, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$ is the vector of unknown regression coefficients. $\|\cdot\|$ denotes the L_2 norm, and $p_\lambda(\cdot)$ is a penalty function depending on tuning parameter $\lambda > 0$. The penalty function $p_\lambda(\cdot)$ in (2.1.2) is symmetric around the origin and depend only on tuning parameter λ. When the design matrix is orthonormal, minimizing problem of (2.1.2) is equivalent to solve component-wise.

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta), \tag{2.1.3}$$

where $\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{y}$. The best subset selection can fit into this framework and its penalty function is

$$p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda), \tag{2.1.4}$$

and the solution to (2.1.2) is $\widehat{\theta} = zI(|z| > \lambda)$ which means that the model selects significant variables and does not penalize them. For LASSO, the penalty function is

$$p_\lambda(|\theta|) = \lambda|\theta|, \tag{2.1.5}$$

and the solution to 2.1.2 is $\widehat{\theta} = sgn(z)(|z| - \lambda)_+$ where $x_+ = x$ if $x > 0$ otherwise equal to 0. For ridge regression, the penalty function is

$$p_\lambda(|\theta|) = \lambda|\theta|^2. \tag{2.1.6}$$

Ridge regression does not have explicit form of thresholding rule but it is useful to scarifies estimation bias to trade for small estimation variance. To evaluate variable selection, Fan and Li (2001) advocates three properties for a good penalty function:

1. Unbiasedness: the estimated result is nearly unbiased for large true value.

2. Sparsity: The threshold will set small coefficient to zero for variable selection.

3. Continuity: The estimator is continuous in value and no jump.

These groundbreaking properties have become the well-recognized criteria to evaluate penalty function. Unbiasedness avoids model bias when the coefficient is large because the variable is significant and no need to be penalized. Taking the first order derivative of (2.1.3) with respect to θ: $\text{sgn}(\theta)\{|\theta| + p'_\lambda(|\theta|)\} - z$. A sufficient condition is that if $p'_\lambda(|\theta|) = 0$ for large θ, then $\widehat{\theta} = z$ meaning that the estimator is unbiased. Sparsity reduces the model complexity by throw away small coefficients. A sufficient condition is $\min\{|\theta| + p'(|\theta|)\} > 0$ because when $|z| < \min\{|\theta| + p'(|\theta|)\}$, the estimator $\widehat{\theta} = 0$ which means that the coefficient is set to 0. Continuity contributes to stable estimation and prediction because the

estimator can take any possible values that results in the most suitable model. The sufficient and necessary condition is the minimum value of $|\theta| + p'(|\theta|)$ is attained at 0 so it is non-decreasing at $\theta > 0$. LASSO estimator satisfies sparsity and continuity properties but is biased for large true value. Ridge regression satisfies continuity but is biased for large coefficient and cannot select variables since coefficients won't be set to zero. Previous penalty functions do not simultaneously satisfy above three properties.

Fan and Li (2001) proposed Smoothly Clipped Absolute Deviation Penalty (SCAD) that improve L_1 penalty and the hard thresholding penalty. SCAD satisfies these three properties simultaneously. The first order derivative is defined as below:

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}. \tag{2.1.7}$$

It is a non-convex penalty function and consist of a spline with node at λ and $a\lambda$. The solution of SCAD is

$$\widehat{\theta} = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda, \\ \{(a-1)z - \text{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda < |z| \leq a\lambda, \\ z, & \text{when } |z| > a\lambda \end{cases} \tag{2.1.8}$$

Parameter a is determined by Bayesian risk assuming the prior distribution of θ is a normal distribution with mean 0 and variance $a\lambda$ given a and λ. Simulation on different dimension shows that minimum of the Bayesian risk is attain around $a = 3.7$ and they further used general cross validation that yield similar result.

Zhang (2010) proposed minimax concave penalty (MCP) from minimize the maximum of concavity to construct penalty function. MCP does not satisfies continuity property. The first order derivative is defined as below:

$$p'_\lambda(t) = (\lambda - t/a)_+. \tag{2.1.9}$$

The solution of MCP when $a \geq 1$ is

$$\theta_{\text{MCP}}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+/(1 - 1/a), & \text{when } |z| < a\lambda, \\ z, & \text{when } |z| \geq a\lambda. \end{cases} \tag{2.1.10}$$

The solution has discontinuos points at $|z| = \lambda$ resulting in model instability.

An estimator that enjoys oracle property means that it can estimate zero coefficients as exactly zero with probability approaching to one and nonzero coefficients as efficiently as if knowing sparsity pattern neforehand. Both SCAD and MCP penalty possess oracle property. In particular, the SCAD penalty has oracle property that similar to Hodges phenomenon (Lehmann and Casella, 2006). The estimator correctly set coefficient at point 0 without sacrificing estimating coefficient elsewhere. The authors prove that for SCAD the penalized likelihood estimator is root-n consistent if $\lambda_n \to 0$. By choosing proper λ, SCAD penalized regression exist root-n estimator. Furthermore, if $\sqrt{n}\lambda_n \to 0$, with probability tending to 1, the performance of the root-n consistent estimators show that true zero coefficients exactly equal to 0 and asymptotic normality of true nonzero coefficient. That is, if $n\lambda \to 0$, the SCAD penalized likelihood estimator possess oracle property and performs no worse than estimating nonzero coefficient by knowing the true zero covariates.

2.1.2 Numerical algorithms

LASSO penalty is computational efficient but has no guarantee of variable selection consistent. For nonconvex penalty SCAD proposed by Fan and Li (2001) and MCP proposed by Zhang (2010) can achieve oracle property under certain condition. That is, the minimum value of objective function (2.1.2) is achieved when they estimate zero coefficients as exact zero with probability approaching one and estimate the nonzero coefficients as efficiently as if the true sparsity pattern is known in advance. It is easy to find the local minima but it is challenging to find the global solution for (2.1.2) since it is nonconvex. Folded concave penalty may enjoy oracle property but it is computational challenging to maximize because of multiple local solution issue and singularity. Summing the convex loss function and folded concave penalty function, the target function might have many local solution making minimization difficult. Penalty functions that have sparse property are non-differentiable at origin.

This chapter introduces algorithms for optimizing the folded concave penalized least sqaures functions. These algorithms are useful for computing LASSO and

can be extended to more general folded concave penalized least squares functions such as SCAD. In addition to the procedure for penalized least squares, these algorithms reveal the statistical insights on the methods.

2.1.2.1 Local quadratic approximation

Fan and Li (2001) suggest using quadratic function to locally approximate the penalty function and iteratively solving it to obtain the result. This method is called local quadratic approximation (LQA).

Suppose that given an initial value $\widehat{\boldsymbol{\beta}}^{(0)}$ that is close to the true value $\boldsymbol{\beta}$, we approximate the first order derivative of penalty function by a linear function:

$$\left[p_\lambda\left(|\beta_j|\right)\right]' = p'_\lambda\left(|\beta_j|\right)\operatorname{sign}\left(\beta_j\right) \approx \left\{p'_\lambda\left(\left|\beta_j^{(0)}\right|\right) / \left|\beta_j^{(0)}\right| |\right\}\beta_j, \tag{2.1.11}$$

(2.1.11) can also be written in terms of penalty function which represents local quadratic approximation of penalty function.

$$p_\lambda\left(|\beta_j|\right) \approx p_\lambda\left(\left|\beta_j^{(0)}\right|\right) + \frac{1}{2}\left\{p'_\lambda\left(\left|\beta_j^{(0)}\right|\right) / \left|\beta_j^{(0)}\right|\right\}\left(\beta_j^2 - \beta_j^{(0)2}\right) \triangleq q(\boldsymbol{\beta}|\boldsymbol{\beta}_0). \tag{2.1.12}$$

Denote the approximation in (2.1.12) as $q(\boldsymbol{\beta}|\boldsymbol{\beta}_0)$, then the objective function can be written as $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)})$. With a good initial estimate, penalized estimator can be obtained by iterative Newton-Raphson algorithm. That is, updating $\boldsymbol{\beta}$ by

$$\boldsymbol{\beta}^{(k+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\left\{\sum_{i=1}^{n}\ell_i(\boldsymbol{\beta}) - n\sum_{j=1}^{p}\frac{p'_\lambda\left(\left|\beta_j^{(k)}\right|\right)}{2\left|\beta_j^{(k)}\right|}\beta_j^2\right\}, \tag{2.1.13}$$

for $k = 1, 2, \cdots$. And at k-th iteration, set j-th component $\beta_j^{(k)} = 0$ when $\beta_j^{(k)}$ is smaller than a predetermined threshold $|\beta_j^{(k)}| < \varepsilon_0$. Stop the iteration if the sequence of $\{\boldsymbol{\beta}^{(k)}\}$ converges. So if a covariate is deleted in any LQA iteration, it will not be selected in the final model.

Hunter and Li (2005) show that LQA algorithm is one specific case of minorize–maximize (MM) algorithm. Local quadratic approximation has properties that

$$q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)}) \geq p_\lambda(\boldsymbol{\beta}) \text{ and } q(\boldsymbol{\beta}^{(0)}|\boldsymbol{\beta}^{(0)}) = p_\lambda(\boldsymbol{\beta}^{(0)}),$$

which means that $q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)})$ is a convex majorant of $p_\lambda(\boldsymbol{\beta})$. So objective function

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)}) \geq Q(\boldsymbol{\beta}^{(0)}|\boldsymbol{\beta}^{(0)}).$$

By the definition of minimization, we have

$$Q\left(\boldsymbol{\beta}^{(k+1)}\right) \leq Q\left(\boldsymbol{\beta}^{(k+1)}|\boldsymbol{\beta}_0\right) \leq Q\left(\boldsymbol{\beta}_0|\boldsymbol{\beta}_0\right) = Q\left(\boldsymbol{\beta}_0\right).$$

Therefore, LQA algorithm has monotonic decreasing property. That is, the target function decreases after each iteration and will converge.

The LQA is an efficient algorithm but has drawbacks in backward stepwise variable selection and choosing the cutoff threshold. If the algorithm delete a covariate at any step, it will no longer include the covariate in the final selected model. This is one drawback of LQA algorithm and another disadvantage is choosing ε_0, which will be another tuning parameter in practice. To address these two drawbacks, Hunter and Li (2005) suggest that adding small noise on penalty function in (2.1.13) can avoid numerical instability. The modified LQA is

$$\boldsymbol{\beta}^{(k+1)} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) - n \sum_{j=1}^{p} \frac{p'_\lambda\left(\left|\beta_j^{(k)}\right|\right)}{2\left\{\left|\beta_j^{(k)}\right| + \tau_0\right\}} \right\}, \tag{2.1.14}$$

where τ_0 is a predetermined noise. Stop the iteration when $\{\boldsymbol{\beta}^{(k)}\}$ converges. In practice, it is difficult to determine the noise level τ_0 because it is a trade off between the speed of convergence and sparsity level.

2.1.2.2 Local linear approximation

The LQA algorithm approximates penalty function using quadratic function resulting in unstable sparse estimation and speed of convergence. Since the LQA is similar to ridge regression, unless directly set to zero, the LQA will not have sparse result. Zou and Li (2008) suggest that non-convex penalty function can be approximated by linear function. This algorithm is called local linear approximation (LLA). They prove that linear approximation can automatically produce sparsity result without setting small value coefficients to zero at any iteration.

Moreover, LLA converge very fast. It inherits LASSO's computation efficiency advantage because penalty is ℓ_1 norm. Penalty function can be approximated at the neighborhood of $\boldsymbol{\beta}^{(0)}$ by linear function,

$$p_\lambda\left(|\beta_j|\right) \approx p_\lambda\left(\left|\beta_j^{(0)}\right|\right) + p'_\lambda\left(\left|\beta_j^{(0)}\right|\right)\left(|\beta_j| - \left|\beta_j^{(0)}\right|\right). \tag{2.1.15}$$

Then the original objective function $Q(\boldsymbol{\beta})$, the sum of likelihood function and penalty function, can be approximated by

$$G\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\beta}) - n\sum_{j=1}^{p}\left[p_\lambda\left(\left|\beta_j^{(0)}\right|\right) + p'_\lambda\left(\left|\beta_j^{(k)}\right|\right)\left(|\beta_j| - \left|\beta_j^{(k)}\right|\right)\right]. \tag{2.1.16}$$

$G\left(\boldsymbol{\beta}|\boldsymbol{\beta}^{(k)}\right)$ is a minorization of $Q(\boldsymbol{\beta})$ and solving for $\boldsymbol{\beta}^{(k+1)}$ is the maximization step. Thus LLA belongs to MM algorithm because $Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)}) \geq Q(\boldsymbol{\beta}^{(0)}|\boldsymbol{\beta}^{(0)})$ The solution of (2.1.16) can be obtained by

$$\boldsymbol{\beta}^{(k+1)} = \arg\max_{\boldsymbol{\beta}}\left\{\sum_{i=1}^{n}\ell_i(\boldsymbol{\beta}) - n\sum_{j=1}^{p}p'_\lambda\left(\left|\beta_j^{(k)}\right|\right)|\beta_j|\right\}. \tag{2.1.17}$$

(2.1.17) is similar to LASSO but different in tuning parameter for every covariate. Zou and Li (2008) prove that the linear approximation is always smaller or equal to original objective function. The equality holds only at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$. Thus, LLA has ascent property. Furthermore, they prove that if initialized by a good estimator, LLA can converge in one step and additional iteration will still have same result.

For linear regression, the initial estimator can be ordinary least squares estimator (when $p < n$). If the penalty is SCAD, $\sqrt{n}\lambda_n \to \infty$ and $\lambda_n \to 0$, then one step LLA can estimate truely zero coefficient as zero with probability tending to one and truly non-zero coefficients as good as when the true model is known.

2.1.2.3 Calibrated concave convex procedure

Under the unified penalized regression, we are interested in finding the solution to penalized least squares from a range of values of tuning parameter λ. Efron et al. (2004) discuss the solution path that is for different values of λ, the solutions $\widehat{\boldsymbol{\beta}}(\lambda)$ to the penalized least squares depends on λ. Solution path is useful to reveal how

an coefficient is included in the final result as λ decrease towards zero.

For nonconvex penalty, it is easy to find the local minima but it is challenging to find the global solution for (2.1.2). Multiple local minima result in multiple solution paths, not all of which are guaranteed to contain oracle solution. To address this problem, proposed calibrated concave convex procedure (CCCP) algorithm to solve high-dimensional nonconvex penalized regression. The CCCP is an easy-to-implement algorithm that works by decompose nonconvex penalty function $C(\boldsymbol{\beta})$ as the sum of a convex function $C_{vex}(\boldsymbol{\beta})$ and a concave function $C_{cav}(\boldsymbol{\beta})$. Given current solution $\beta^{(k)}$, for each iteration minimize the tight convex upper bound of $C(\boldsymbol{\beta})$ which is $Q(\boldsymbol{\beta}) = C_{vex}(\boldsymbol{\beta}) + \nabla C_{cav}(\boldsymbol{\beta}^{(k)})'\boldsymbol{\beta}$ where $\nabla C_{cav}(\boldsymbol{\beta}^{(k)}) = \partial C_{cav}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. The CCCP algorithm has monotonic decreasing property that is the objective function decrease after each iteration but there is no guarantee that the solution converge to global minimum.

Because multiple local minima provide not unique solution path, there is no guarantee for the solution containing the oracle estimator. Wang et al. (2013) prove that a calibration of the CCCP algorithm for non-convex penalized regression provide consistent solution path containing the oracle estimator with probability approaching one in two steps under following mild conditions:

A1 The minimum absolute signal value should exceed certain threshold.

A2 Random errors are i.i.d. mean zero sub-Gaussian random variables with bounded variance.

A3 The penalty function $p_\lambda(t)$ is monotonic increasing and concave for $t \in [0, +\infty)$ with a continuous derivative on $(0, +\infty)$. The penalty function is assumed to not penalize large coefficients.

A4 A restraint to design matrix $\boldsymbol{X}$ so that LASSO can be model selection consistent. Because the first step of calibrated CCCP is Lasso estimator, the asymptotic properties of CCCP and Lasso estimators are related.

A5 The value of tuning parameter is in smaller order of the minimum absolute signal value.

The calibrated algorithm consist of two steps:

1 Let $\widehat{\boldsymbol{\beta}}^{(1)}(\lambda) = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)}, \lambda)$,

2 Let $\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}|\boldsymbol{\beta}^{(1)}, \lambda)$.

Considering a sequence of tuning parameter values, the algorithm yields a solution path $\{\widehat{\boldsymbol{\beta}}(\lambda) : \lambda > 0\}$. A solution path is called *path consistent* if it contains the oracle estimator. Under conditions A1 - A5, calibrated CCCP algorithm is proved to produce a consistent solution path that contains oracle estimator.

2.1.2.4 ADMM

The alternating direction method of multipliers (ADMM) was proposed originally in Glowinski and Marroco (1975) and Gabay and Mercier (1976). Boyd et al. (2011) gives a systematic review of ADMM. The ADMM algorithm is an optimization algorithm suitable for non-convex optimization problem. The ADMM solves the problem of following general form,

$$\begin{aligned} &\min_{x_1, x_2} f_1(x_1) + f_2(x_2), \\ &\text{s.t. } A_1 x_1 + A_2 x_2 = c. \end{aligned} \tag{2.1.18}$$

$x_1 \in \mathcal{R}^{p_1}$ and $x_2 \in \mathcal{R}^{p_2}$ are two variables to estimate and $A_1 \in \mathcal{R}^{n \times p_1}$, $A_2 \in \mathcal{R}^{n \times p_2}$, $c \in \mathcal{R}^n$ are constants. ADMM works by breaking the problem into two small parts and iteratively minimizing each part, which is easy to work with. ADMM begins by rewriting the optimization with constraint as augmented Lagrangian function,

$$L_\phi(x_1, x_2; \gamma) = f_1(x_1) + f_2(x_2) + \langle \gamma, A_1 x_1 + A_2 x_2 - c \rangle + \frac{\phi}{2} \|A_1 x_1 + A_2 x_2 - c\|_2^2, \tag{2.1.19}$$

where $\gamma \in \mathcal{R}^n$ is the dual variable and $\phi > 0$ is prespecified parameter. For initial value of (x_1^0, x_2^0, γ^0), instead of updating x_1 and x_2 at the same time, the x_1 and x_2 are updated alternatively and the iterative update of each part proceeds as follow:

$$\begin{aligned} x_1^{k+1} &= \arg\min_{x_1} L_\phi\left(x_1, x_2^k; \gamma^k\right), \\ x_2^{k+1} &= \arg\min_{x_2} L_\phi\left(x_1^{k+1}, x_2; \gamma^k\right), \\ \gamma^{k+1} &= \gamma^k + \phi\left(A_1 x_1^{k+1} + A_2 x_2^{k+1} - c\right). \end{aligned} \tag{2.1.20}$$

This scheme is account for alternative direction. Gradient decent method is used to update the dual γ.

In statistical machine learning, optimizing least squares with nonconvex penalty function is an example of non-convex problem and can use ADMM algorithm. Consider penalized least squares with form

$$\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\sum_j P_\lambda(|\beta_j|). \tag{2.1.21}$$

(2.1.21) can be written as an optimization problem with constraints and further write as augmented Lagrange problem:

$$L_\phi(\boldsymbol{\beta};\gamma)=\frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2+\sum_j P_\lambda(|z_j|)+\gamma^T(\boldsymbol{\beta}-\boldsymbol{z})+\frac{\phi}{2}\|\boldsymbol{\beta}-\boldsymbol{z}\|_2^2, \tag{2.1.22}$$

where $\gamma\in\mathcal{R}^p$ and ϕ usually set to 1. The solution of (2.1.22) iteratively update of each part proceeds as follow:

$$\begin{aligned}
&\boldsymbol{\beta}^{k+1}\arg\min_{\boldsymbol{\beta}} L_\phi\left(\boldsymbol{\beta},\boldsymbol{z}^k;\gamma^k\right),\\
&z_j^{k+1}=\arg\min_{z_j} L_\phi\left(\boldsymbol{\beta}_1^{k+1},z_j;\gamma^k\right),\text{for } j=1,\cdots,p,\\
&\gamma^{k+1}=\gamma^k+\left(\boldsymbol{\beta}^{k+1}-\boldsymbol{z}^{k+1}\right),
\end{aligned} \tag{2.1.23}$$

for the $k+1$ iteration. Since (2.1.22) is twice differentiable, $\boldsymbol{\beta}^{k+1}$ can be obtained by the Newton–Raphson algorithm. z_j^{k+1} may have a closed form for some popular penalties such as the Lasso, SCAD or MCP penalty. In our implementation, we use the SCAD penalty (2.1.7). When z_j^{k+1} is smaller than certain predetermined threshold, directly set it to zero, which is helpful to the algorithm convergence.

ADMM has advantage of breaking complex optimization problem into two small sub-problems and alternative solving these two problem. It is easy to implement, but it converge slowly and may by influenced by local minimum.

2.1.3 Tuning parameter selection

In application for penalized least squares, one needs to select proper tuning parameter λ. The larger the parameter λ is, the sparser the selected model will be.

In some scenario, one can determine the value of λ through the desired number of independent variables. There are some R packages can help to achieve this goal by visualizing the relationship between λ and the number of final coefficients in the model. In most cases, one usually has no idea of the number of coefficients in the final model. Thus, it is desirable to have a data-driven method to select λ.

2.1.3.1 Generalized cross validation

Using data-driven approach, the common idea is to choose λ that can minimize prediction error. k-fold cross validation, a popular method for tuning parameter selection in machine learning, is based on this idea.

The k-fold cross validation begins by separating the data into k subsets. It uses one of the subset as the testing set and remaining $k-1$ subsets of data as training set to train a model and then calculate the prediction error rate on the testing set. Repeat this process for k times on different testing sets and average the prediction error. Usually, one set k equal to 5 or 10 to have trade-off between computation time and selection property.

One special case is n-fold cross validation, where n is the sample size, which is also called leave-one-out cross validation.

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}^{(-i)}\left(\mathrm{x}_i\right)\right)^2, \tag{2.1.24}$$

where $\widehat{f}^{(-i)}\left(\mathrm{x}_i\right)$ is the predicted value at $\boldsymbol{x}_i$ computed by using all the data except the ith observation. This requires huge computation cost because one needs to repeat n times to compute. Golub, Heath and Wahba (1979) suggest generalize cross validation as an approximation of n-fold cross validation using linear smoother's property. A fitted model on data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ is a linear smoother if it can be written as

$$\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}, \tag{2.1.25}$$

where $\mathbf{S}$ is a $n \times n$ matrix only depends on $\boldsymbol{X}$. Linear regression and generalized linear regression are belong to linear smoother. For example, $\mathbf{S} = \boldsymbol{X}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}$ and $tr(\mathbf{S}) = p$. Linear smoother has self-stable property that is the fit based on a new data point is the same as the fit without it. With the self-stable property of

linear smoother, it can be show that leave-one-out cross validation error is equal to $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\widehat{y}_i}{1-S_{ii}}\right)^2$.

Golub, Heath and Wahba (1979) suggest to approximate each diagonal elements of $\mathbf{S}$ by their average which is $\frac{\mathrm{tr}\mathbf{S}}{n}$ and the GCV is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\widehat{y}_i}{1-S_{ii}}\right)^2 \approx \frac{1}{n}\frac{\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2}{\left(1-\frac{1}{n}\,\mathrm{tr}\,\mathrm{S}\right)^2} := \mathrm{GCV}.$$

GCV value depends on λ, so we can choose the λ that produce the minimal GCV. GCV is computation effecient compare to the cross validation method. One disadvantage of generalized cross validation is that it is often result in overfitting. When the number of dimension is fixed, Wang, Li and Tsai (2007) rigorously proved that generalized cross-validation will produce an overfitted model with a positive probability for penalized regression with SCAD penalty.

2.1.3.2 High-dimensional Bayesian information criteria

Other criteria such as BIC are also studied for choosing tuning parameter λ. For fixed p, effective BIC-type criterion for nonconvex penalized regression has been investigated in Wang, Li and Tsai (2007) and Zhang, Li and Tsai (2010); and for diverging p with $p < n$, Wang, Li and Leng (2009) provided through study. But they do not perform well in ultra-high dimensional case. Wang et al. (2013) extend BIC to high-dimensional case and use it to determine the tuning parameter value λ.

$$\mathrm{HBIC}(\lambda) = \log\left(\widehat{\sigma}_\lambda^2\right) + |M_\lambda|\,\frac{C_n\log(p)}{n}, \tag{2.1.26}$$

where $M_\lambda = \left\{j : \widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0\right\}$, $|M_\lambda|$ is the cardinality of M_λ, $\widehat{\sigma}_\lambda^2 = n^{-1}\mathrm{SSE}_\lambda$ with $\mathrm{SSE}_\lambda = \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda)\|^2$, and C_n is a sequence of numbers that diverges to ∞. The optimal λ is chosen where the minimum value of HBIC is achieved. Resorting to HBIC, it avoids the drawbacks of theoretical optimal tuning parameter, which depends on unknown variance of unobserved random noise, and cross-validation, which often results in overfitting.

2.2 High-dimensional Generalized Linear Model Review

There are many works about variable selection in high dimensional regression in recent years (Sun and Zhang, 2010). However testing high dimensional estimator can be challenging. Tibshirani et al. (2016) and Lee et al. (2016) proposed testing method for LASSO estimators but can only inference nonzero coefficients in LASSO regression. Moreover, these methods is designed for linear model and they may fail to work under generalized linear model.

This chapter extends the methods in variable selection, testing and inference from linear model to generalized linear models with binary, categorical and count response. The review begins with introducing generalized linear models and then discusses unified framework for penalized likelihood estimation method, numerical optimization method. Finally, asymptotic properties, testing and inference for generalized linear model will be covered.

2.2.1 Penalized generalized linear model for variable selection

Generalized linear models is very important in fields of medicine, psychology, engineering and so on. McCullagh and Nelder (1989) systematically introduce the generalized linear models. Let Y be the response and $\boldsymbol{X}$ consist of covariates. Assuming that the distribution of Y belongs to exponential family, generalized linear model with canonical link can be expressed as

$$\exp\left(\frac{Y\boldsymbol{\beta}^T\boldsymbol{X} - b\left(\boldsymbol{\beta}^T\boldsymbol{X}\right)}{\phi}\right)c(Y), \tag{2.2.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients and ϕ is some positive nuisance parameter, $b(\cdot)$ is assumed to be twice continuously differentiable and $b''(\cdot) > 0$. The penalized likelihood function can be written as

$$\min_{\beta} \ell_n(\boldsymbol{\beta}) + P_\lambda(|\boldsymbol{\beta}|), \tag{2.2.2}$$

where the first term is a convex function and the second term is a penalty function. We assumed that the penalty function is folded concave. For linear model and generalized linear model, $\ell_n(\boldsymbol{\beta})$ can be log-likelihood function. With this general form, the estimation procedure can be carried out in the similar matter as linear model as discussed in chapter 2.2.

2.2.2 Strong oracle optimality of folded concave penalized estimation

Folded concave penalty methods such as SCAD and MCP are proved to have strong oracle property in high-dimensional model selection. But a folded concave penalty has multiple local solutions, it is uncertain whether the local solution obtained by a given optimization algorithm enjoys the oracle property. Fan et al. (2014) provide a unified theory to obtain local solution based on local linear approximation algorithm (LLA) and derive a probability lower bound that this solution exactly equals the oracle estimator. LLA algorithm makes use of special folded concave penalty structure and majorization-minimization (MM) principle to transform the concave regularization problem into an adaptive LASSO problem. Note that this local solution may not be global solution but it possesses the desired properties.

The target of estimation of (2.2.2) is $\boldsymbol{\beta}^* = (\beta_1^*, \cdots, \beta_p^*)'$, where p is the dimension that is larger than sample size n. And oracle estimator is the estimated result by knowing the true non-zero coefficients

$$\widehat{\boldsymbol{\beta}}^{\text{oracle}} = \left(\widehat{\boldsymbol{\beta}}_{\mathcal{A}}^{\text{oracle}}, \mathbf{0}\right) = \arg \min_{\boldsymbol{\beta}:\boldsymbol{\beta}_{\mathcal{A}^c}=\mathbf{0}} \ell_n(\boldsymbol{\beta}), \tag{2.2.3}$$

where $\mathcal{A} = \left\{j : \beta_j^\star \neq 0\right\}$.

Suppose the minimal signal strength of $\boldsymbol{\beta}^*$ satisfies $\|\boldsymbol{\beta}_{\mathcal{A}}^\star\|_{\min} > (a+1)\lambda$. Recall that a general folded concave penalty function $P_\lambda(|t|)$, $t \in (-\infty, \infty)$ satisfying:

(i) $P_\lambda(t)$ is monotonely increasing and concave in $t \in [0, \infty)$ with $P_\lambda(0) = 0$

(ii) $P_\lambda(t)$ is diferentiable in $t \in (0, \infty)$ with $P'_\lambda(0) := P'_\lambda(0+) \geq a_1\lambda$

(iii) $P'_\lambda(t) \geq a_1\lambda$ for $t \in (0, a_2\lambda]$

(iv) $P'_\lambda(t) = 0$ for $t \in [a\lambda, \infty)$ with prespecified constant $a > a_2$.

The authors proved that LLA algorithm can find oracle estimator in no more than two steps with probability greater than $1-\delta_0-\delta_1-\delta_2$ if the parameter is initialized by some appropriate initial estimator $\widehat{\boldsymbol{\beta}}^{\text{initial}}$, $a_0 = \min\{1, a_2\}$.

$$\begin{aligned}
\delta_0 &= \Pr\left(\left\|\widehat{\boldsymbol{\beta}}^{\text{initial}} - \boldsymbol{\beta}^\star\right\|_{\max} > a_0\lambda\right), \\
\delta_1 &= \Pr\left(\left\|\nabla_{\mathcal{A}^c}\ell_n\left(\widehat{\boldsymbol{\beta}}^{\text{oracle}}\right)\right\|_{\max} \geq a_1\lambda\right), \\
\delta_2 &= \Pr\left(\left\|\widehat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}}\right\|_{\min} \leq a\lambda\right) \leq \Pr\left(\left\|\widehat{\boldsymbol{\beta}}^{\text{oracle}}_{\mathcal{A}} - \boldsymbol{\beta}^\star_{\mathcal{A}}\right\|_{\max} > \lambda\right).
\end{aligned} \tag{2.2.4}$$

The δ_0 measures the closeness of the initial estimator to the true signal. Provided that reasonable initial estimator contains all the non-zero true value, δ_0 goes to zero. If δ_0 cannot go to zero, then it means that the model is extremely hard to estimate no matter how clever an estimator is. The chosen of initial estimator will be discussed in the following two examples. δ_1 measures the closeness of score function between true zero coefficients and their oracle estimates. δ_1 is usually small. δ_2 measure the closeness of closeness of oracle and true nonzero coefficients. Oracle estimator uniformly converges to true estimator due to intrinsic sparse dimension. δ_2 is usually small. Both δ_1 and δ_2 are the condition for oracle estimator so are referred to as oracle regularization condition.

Then the authors provide examples to illustrate the choosing initial estimator and how to apply LLA for linear regression, logistic regression.

For sparse linear regression, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \varepsilon$, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)$ being i.i.d sub-Gaussian (σ) for some fixed constant $\sigma > 0$ and $E\{\exp(t\varepsilon_i^2)\} < \exp(\sigma^2 t^2/2)$. The folded concave penalized least squares estimation,

$$\min_{\beta} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \sum_j P_\lambda\left(|\beta_j|\right). \tag{2.2.5}$$

The authors suggest to initiate LLA algorithm using LASSO estimator:

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\ell_2}^2 + \lambda_{\text{lasso}}\|\boldsymbol{\beta}\|_{\ell_1}. \tag{2.2.6}$$

Under the condition of restricted eigenvalue condition of LASSO estimator proposed by Bickel et al. (2009), the LLA algorithm converges to oracle estimator in two steps with probability at lease $1 - \delta_0 - \delta_1 - \delta_2$ and $\delta_0, \delta_1, \delta_2$ converge to zero

quickly.

For sparse logistic regression, $\Pr\left(y_i = 1|\mathbf{x}_i, \boldsymbol{\beta}^\star\right) = \exp\left(\mathbf{x}_i'\boldsymbol{\beta}^\star\right) / \left(1 + \exp\left(\mathbf{x}_i'\boldsymbol{\beta}^\star\right)\right)$, the penalized log-likelihood is

$$\min_{\beta} \frac{1}{n} \sum_i \left\{-y_i \mathbf{x}_i'\boldsymbol{\beta} + b\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right\} + \sum_j P_\lambda\left(|\beta_j|\right), \tag{2.2.7}$$

where canonical link $b(t) = \log\{1 + \exp(t)\}$.

Similarly, the authors suggest that using LASSO estimator as initial value, the LLA algorithm can converge in two steps with overwhelming probability.

$$\widehat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_i \left\{-y_i \mathbf{x}_i'\boldsymbol{\beta} + \psi\left(\mathbf{x}_i'\boldsymbol{\beta}\right)\right\} + \lambda_{\text{lasso}} \|\boldsymbol{\beta}\|_{\ell_1}. \tag{2.2.8}$$

The simulation study shows that local optimal solved by LLA algorithm have different solution path compare to coordinate decent and different initial value of LLA produce different converged solution. LLA initialized by LASSO estimator is cheap to compute and fully converge to the true sparsity model with overwhelming probability. Note that this method exerts penalty on all predictors, and minimum signal strength is a key assumption in order to not missing important predictors. Penalized on all the coefficients may not be suitable, and partial penalized problem deserves careful investigation.

2.2.3 Linear hypothesis testing for high-dimensional generalized linear model

In many scenarios, one may not want to penalize all the predictors because there are some predictors known important and of interest but their signal may weak. Traditional estimation and testing impose minimum signal on the all nonzero coefficients (Fan and Peng, 2004; Fan and Lv, 2011) and small signal signal making the test do not have any power for local alternative. To address this problem, Shi et al.(2019) start with considering a more general form of hypothesis testing, then propose partial penalized test statistics based on partial penalized likelihood function that does not penalize important predictors. The partial penalized likelihood

function:

$$Q_n(\boldsymbol{\beta}, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \boldsymbol{\beta}^T \boldsymbol{X}_i - b\left(\boldsymbol{\beta}^T \boldsymbol{X}_i\right) \right\} - \sum_{j \notin \mathcal{M}} p_\lambda \left(|\beta_j| \right), \tag{2.2.9}$$

where $\mathcal{M} \subseteq [1, \ldots, p]$ is a subset of coefficients that do not penalized. This is useful because the signal in $\mathcal{M}$ can be very small but it is of interest, so we do not penalize it in the model. Partial penalize can avoid imposing minimal signal condition on true value $\boldsymbol{\beta}_{0,\mathcal{M}}$, which can evaluate the small signal evaluation uncertainty. $p_\lambda(\cdot)$ is penalty function with a tuning parameter λ. Define

$$\widehat{\boldsymbol{\beta}}_0 = \arg\max_{\boldsymbol{\beta}} Q_n \left(\boldsymbol{\beta}, \lambda_{n,0} \right) \text{ subject to } \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t},$$

$$\widehat{\boldsymbol{\beta}}_a = \arg\max_{\boldsymbol{\beta}} Q_n \left(\boldsymbol{\beta}, \lambda_{n,a} \right).$$

Assuming $\log p = O(n^\alpha)$ for some $0 < \alpha < 1$, the linear hypothesis

$$H_0 : \boldsymbol{C}\boldsymbol{\beta}_{0,\mathcal{M}} = \boldsymbol{t}. \tag{2.2.10}$$

for a given $\mathcal{M} \subseteq [1, \ldots, p]$, $\boldsymbol{C}_{r \times |\mathcal{M}|}$ is a full row rank matrix and $t_{r \times 1}$ is a vector. Let m be the cardinal of $\mathcal{M}$ and $r \leq m$. If setting $\boldsymbol{C} = \mathbf{I}_r$ identity matrix and $\mathbf{t}$ as a vector of zero, then the test (2.2.10) reduce to an important class of hypothesis $\beta_{0,\mathcal{M}} = \mathbf{0}$.

Under mild assumption, (A1) the covariates vectors $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ are uniformly bounded, (A2) minimum nonzero signal in $\mathcal{M}^c$ is greater than certain threshold, (A3) some common conditions for general linear model, (A4) the sequence of penalty tuning parameter is bounded by constant, for folded concave penalty function the authors proved that each coefficients in $\widehat{\boldsymbol{\beta}}_{0,S}$ and $\widehat{\boldsymbol{\beta}}_{a,S}$ is nonzero, where $S = \{j \in \mathcal{M}^c : \beta_{0,j} \neq 0\}$. And $\widehat{\boldsymbol{\beta}}_{0,\mathcal{M}^c}$ has the same sign as $\widehat{\boldsymbol{\beta}}_{a,\mathcal{M}^c}$. They also implies that if have more information about $\boldsymbol{\beta}_0$, $\widehat{\boldsymbol{\beta}}_0$ will converge faster than $\widehat{\boldsymbol{\beta}}_a$ because the estimator will be more accurate. And under certain condition, $\widehat{\boldsymbol{\beta}}_{0,\mathcal{M}}$, $\widehat{\boldsymbol{\beta}}_{0,S}$, $\widehat{\boldsymbol{\beta}}_{a,\mathcal{M}}$, $\widehat{\boldsymbol{\beta}}_{a,S}$ achieve the root-n consistent estimator of $\boldsymbol{\beta}_{0,\mathcal{M}}$, $\boldsymbol{\beta}_{0,\mathcal{S}}$, $\boldsymbol{\beta}_{0,\mathcal{M}}$, $\boldsymbol{\beta}_{0,\mathcal{S}}$ respectively.

Partial penalized likelihood ratio test statistic is defined to be

$$T_L = 2n \left\{ L_n \left(\widehat{\boldsymbol{\beta}}_a \right) - L_n \left(\widehat{\boldsymbol{\beta}}_0 \right) \right\} / \widehat{\phi}, \tag{2.2.11}$$

where $L_n(\boldsymbol{\beta}) = \sum_i \left\{ Y_i \boldsymbol{\beta}^T \boldsymbol{X}_i - b\left(\boldsymbol{\beta}^T \boldsymbol{X}_i \right) \right\} / n$.

Partial penalized Wald statistic is defined to be

$$T_W = \left(\boldsymbol{C} \widehat{\boldsymbol{\beta}}_{a,\mathcal{M}} - t \right)^T \left(\boldsymbol{C} \widehat{\boldsymbol{\Omega}}_{a,mm} \boldsymbol{C}^T \right)^{-1} \left(\boldsymbol{C} \widehat{\boldsymbol{\beta}}_{a,\mathcal{M}} - \boldsymbol{t} \right) / \widehat{\phi}, \tag{2.2.12}$$

where $\widehat{\boldsymbol{\Omega}}_a = n \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_a \right) \boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_a \right) \boldsymbol{X}_{\widehat{S}_a} \\ \boldsymbol{X}_{\widehat{S}_a}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_a \right) X_{\mathcal{M}} & \boldsymbol{X}_{\widehat{\Omega}_a}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_a \right) X_{\widehat{S}_a} \end{pmatrix}^{-1}$, and $\widehat{\boldsymbol{\Omega}}_{a,mm}$ is its first m rows and columns submatrix.

Partial penalized score statistic is defined to be

$$T_S = \left\{ \boldsymbol{Y} - \boldsymbol{\mu} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_0 \right) \right\}^T \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}} \\ \boldsymbol{X}_{\widehat{S}_0} \end{pmatrix} \widehat{\boldsymbol{\Omega}}_0 \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}} \\ \boldsymbol{X}_{\widehat{S}_0} \end{pmatrix}^T \left\{ \boldsymbol{Y} - \boldsymbol{\mu} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_0 \right) \right\} / \widehat{\phi}, \tag{2.2.13}$$

where $\widehat{S}_0 = \{ j \in \mathcal{M}^c : \widehat{\beta}_{0,j} \neq 0 \}$, and

$$\widehat{\boldsymbol{\Omega}}_0 = n \begin{pmatrix} \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_0 \right) \boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\mathcal{M}}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_0 \right) \boldsymbol{X}_{\widehat{S}_0} \\ \boldsymbol{X}_{\widehat{S}_0}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_0 \right) \boldsymbol{X}_{\mathcal{M}} & \boldsymbol{X}_{\widehat{S}_0}^T \boldsymbol{\Sigma} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}_0 \right) \boldsymbol{X}_{\widehat{S}_0} \end{pmatrix}^{-1}.$$

If T is greater than $\chi^2_\alpha(r)$, we reject the null hypothesis under α significance level for $T = T_W, T_s$ or T_L, where r is the degree of freedom and also the number of constraints. When $r = O(1)$, according to Lyaponuv condition assure the asymptotic normality of $\widehat{\beta}_0$.

To compute test statistics, Shi et al. (2019) suggest using ADMM to compute partial penalized estimator $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_a$. Below present the constraint estimation of $\widehat{\boldsymbol{\beta}}_0$ and $\widehat{\boldsymbol{\beta}}_a$ can be computed in the similar method.

$$\widehat{\boldsymbol{\beta}}_0^\lambda = \arg\min_\beta \left(-L_n(\boldsymbol{\beta}) + \sum_{j \in \mathcal{M}^c} p_\lambda \left(|\beta_j| \right) \right) \text{ subject to } \boldsymbol{C} \boldsymbol{\beta}_{\mathcal{M}} = \boldsymbol{t}.$$

Write the above constraints into augmented Lagrange:

$$\begin{aligned} L_\rho(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{v}) = &- L_n(\boldsymbol{\beta}) + \sum_{j=1}^{p-m} p_\lambda\left(|\theta_j|\right) + \boldsymbol{v}^T \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t}, \\ \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta} \end{pmatrix} \\ &+ \frac{\rho}{2} \left\| \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t} \right\|_2^2 + \frac{\rho}{2} \left\| \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta} \right\|_2^2, \end{aligned} \tag{2.2.14}$$

for predetermined $\rho > 0$ and applying dual descent for $\boldsymbol{\beta}$ and θ until converge:

$$\begin{aligned} \boldsymbol{\beta}^{k+1} &= \arg\min_{\boldsymbol{\beta}} \left\{ \left(\boldsymbol{v}^k\right)^T \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t}, \\ \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta}^k \end{pmatrix} + \tfrac{\rho}{2} \| \begin{matrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}} - \boldsymbol{t} \|^2 \\ \boldsymbol{\beta}_{\mathcal{M}^c} - \boldsymbol{\theta}^k \|_2^2 \end{matrix} \right\}, \\ \boldsymbol{\theta}^{k+1} &= \arg\min_{\boldsymbol{\theta}} \left\{ \textstyle\sum_{j=1}^{p-m} p_\lambda\left(|\theta_j|\right) + \frac{\rho}{2} \left\| \boldsymbol{\beta}_{\mathcal{M}^c}^{k+1} - \boldsymbol{\theta} \right\|_2^2 + \left(\boldsymbol{v}^k\right)^T \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}}^{k+1} - \boldsymbol{t}, \\ \boldsymbol{\beta}_{\mathcal{M}^c}^{k+1} - \boldsymbol{\theta} \end{pmatrix} \right\}, \\ \boldsymbol{v}^{k+1} &= \boldsymbol{v}^k + \rho \begin{pmatrix} \boldsymbol{C}\boldsymbol{\beta}_{\mathcal{M}}^{k+1} - \boldsymbol{t} \\ \boldsymbol{\beta}_{\mathcal{M}^c}^{k+1} - \boldsymbol{\theta}^{k+1} \end{pmatrix}. \end{aligned} \tag{2.2.15}$$

The tuning parameter λ and be choosen through the minimum HBIC as discussed in (2.1.26).

Numerical study shows that under these null hypotheses, Type I error rates of the three tests are well controlled and close to the nominal level. Under the alternative hypotheses, the powers of these three test statistics increase as true nonzero value increases, showing the consistency of our testing procedure. Moreover, the empirical rejection rates between these three test statistics are very close and robust for different simulation settings.

2.3 Mediation model

Consider common linear model, an exposure variable has an effect on the outcome variable Y. In causal analysis, X is called causal variable and the variable that results in is the outcome. In Figure 2.1 (a), the simple linear regression can only explain the relation between the causal variable and the outcome. The coefficient c can interpret as the total effect of causal variable on the outcome variable.

Figure 2.1 (b) reflects the mediation model. It shows the total effect of an exposure on the outcome variable can be divided into two effects called direct effect and indirect effect. Direct effect means that exposure has direct impact on

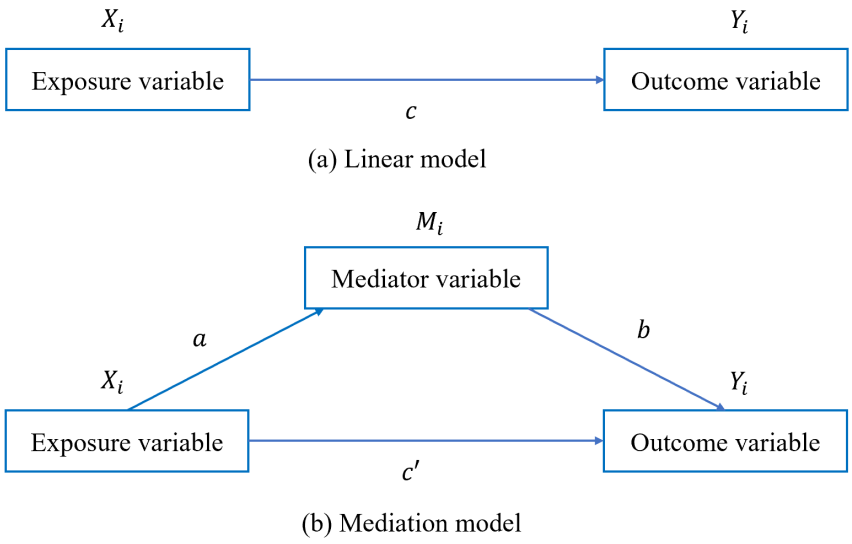

Figure 2.1. (a) Depicts the linear model that shows the causal relationship between the exposure variable and outcome variable. (b) depicts the mediation model that shows the causal relationship between the exposure variable and outcome variable which can decompose as direct effect and indirect effect that pass through the mediator.

the outcome variable. Indirect effect means that the exposure variable has impact on third observed variable, a mediator, and then the mediator exerts impact on the outcome variable. So part of total effect is mediated by mediator. Mediation analysis aims to reveal and understand the underlying mechanism through which the exposure variable affects the outcome variable. It attracts a lot of attention in wide research areas, psychology, genomics and epidemiology (MacKinnon, 2012) The simplest case of mediation analysis with one mediator is shown in Figure 2.1 (b) with three variables: exposure or independent variable X_i, mediator variable M_i, and outcome or dependent variable Y_i, for $i = 1, \cdots, n$, let n be the sample size. Let a be the effect of independent variable on mediator variable, b be the effect of mediator on the outcome variable so the indirect effect is the product of a and b. Denote c' to be the direct effect that exposure have on outcome variable.

There are two mediation model settings complete mediation and incomplete mediation. Complete mediation refers to the case when an exposure variable does not have direct effect on exposure on the outcome variable. That is, parameter c' is zero. This setting is common for genetic variants located in noncoding regions of the genome which can only exert their effect on a phenotype through changing

gene expression. Incomplete mediation model means that both direct and indirect effect of an exposure variable on the outcome variable exist. That is, c' is not zero. This is a common scenario for example, in DNA methylation studies that study the effect of environmental exposure on phenotype by changing DNA methylation pattern, which is indirect effect. In the meanwhile, the environmental exposure can also have direct effect on the response variable without passing through DNA methylation process.

The mediator can be regarded as intervening or process variable because it stays in the path way of the exposure variable to have effect on outcome variable. Mediation model has two different settings, complete mediation and incomplete mediation. Complete mediation is the setting in which variable X affects Y only through M, making path c' zero. Incomplete mediation or partial mediation is the setting in which X has both direct effect and indirect effect on Y. That is, from statistical model perspective, the path from X to Y is reduced in absolute size but is still different from zero when the mediator is introduced. Mediation model is a causal model. One key assumption is that the mediator causes the outcome and not vice versa. Otherwise, the results from the mediation analysis are likely of little value because it cannot explain interrelationship. Overall, a variable can be regarded as mediator if it can account for the relationship between exposure variable and the outcome variable (Baron and Kenny, 1986).

There is a long history in the study of mediation and is a very popular topic (Hyman, 1955; MacCorquodale and Meehl, 1948). Testing mediation is beneficial to understand the mechanism through which the expose variable affects the outcome variable. Mediation and moderation analyses are a key part of what has been called process analysis, but mediation analyses tend to be more powerful than moderation analyses (Baron and Kenny, 1986).

MacKinnon (2012) summarizes the application of mediation analysis in fields of psychology, genomics and epidemiology. In psychology, stimulus-organism-response is a mediation model. In the past, scientists analyzed stimulus-response model to understand the outside stimulus' effect on the behavior of a living thing on macro level. However, with the advancement of technology, scientists can measure and observe micro level reaction and propose that organism act as an intervening variable of stimulus' effect on the response. Organism explains how the relationship

between the stimulus and response by postulating different mediating mechanisms operating in it. Mediating mechanisms are what determines how an organism responds to a stimulus. For example, a stimulus may trigger a memory mechanism that identifies the stimulus as a threat that leads to an avoidance response, or a stimulus may trigger an attraction process that leads to a physiological response such as a male bird sing songs that is heard by a female birds. Then hormones is triggered inside the female bird resulting in approaching response to the male bird. Moreover, mediation analysis can be used to analyze the way to train animals. For example, a good behavior from a training animal is rewarded and a bad behavior is penalized when facing with different intervening factors.

In epidemiology fields, mediation analysis was used to determine the spread of yellow fever in Latin American area. The yellow fever prevention illustrates an intervention that changing mediators to affect the outcome variable during Panama canal construction. Before the Panama canal was built, there were two popular theory about the cause of yellow fever. The first one is person-to-person contact because people get infected after contact with yellow fever patient. Another one is that mosquito carried malaria and transmitted the disease. At that time, to reduce the human exposure of mosquito, the epidemiologists increased the mosquito predators and the local government improved drainage system, because standing water was the ideal place to breed mosquito. These interventions did take an effect to reduce infected people as the number of mosquito significantly decreased. This example shows that through intervention the mediator that is human exposure to mosquito, the result of infection can be changed. In practice, public health promotion and disease prevention methods uses mediation model to find essential mediators and design experiments to alter the mediators resulting in changing the outcome.

2.3.1 Estimation and inference of low dimensional mediation model

In low dimensional mediation model, the number of mediators is fixed and finite. If the mediation model is correctly specified, the direct and indirect effects can be estimated by multiple regression using ordinary least squares for linear model

and maximum likelihood for generalized linear model. That is, parameters a, b, c and c' can be estimated using multiple regression. Regardless of which data analytic method is used, the steps necessary for testing mediation are the same. Baron and Kenny (1986) suggest four steps for estimation and hypothesis testing for mediation analysis:

- Step 1: Find whether the direct effect is zero or not. Fit a regression of Y as the response variable on X as predictors and obtain the estimated coefficient $\widehat{c}$ of total effect. This step provides whether there is potential mediators can mediate the direct effect. If c is not close to zero, then the exposure variable is correlated with the outcome variable so one needs conducting following steps for mediation analysis.

- Step 2: Find whether X is correlated with M or not. Fit a regression of M as the response variable on X as predictors and obtain estimated coefficient $\widehat{a}$. If the regression coefficient a is not close to zero, then mediator can be treated as if it were the response of X.

- Step 3: Find whether M is correlated with Y or not. Fit a regression of Y as the response variable on M as predictors to obtain estimated coefficient $\widehat{b}$. Finding the correlation of mediator with outcome variable is not enough because they are both the result of exposure variable X. Therefore, X must be controlled when examining the effect of the mediator.

- Step 4: Find whether is complete mediation case. Fit a regression on Y as response and X as predictor by controlling M. If $\widehat{c}'$ is close to zero, then it is a complete mediation model.

If the answer to above four steps are positive, then mediator M completely mediates the relationship of exposure X and the outcome Y. If only the first three steps are positive while the step 4 is negative, this means that incomplete mediation model is suitable. Note that steps are stated in terms of zero and nonzero coefficients, not in terms of statistical significance, because large coefficient can be insignificant if sample size is small while small coefficient can be significant if sample size is large. Baron and Kenny (1986) want to study the proper model

without the discussion of sample size. Moreover, exposure variable may be highly correlated with mediators which is common in real dataset and may make direct effect insignificant. That is, when a is very large and b is zero. There is not mediation effect so intuitively $c = c'$. But multicolinearity issue arise in Step 4 when a is large, resulting in c' is insignificant different from zero, which is counterintuitive.

Baron and Kenny (1986) method lacks of the support of statistical theory. Hayes (2017) makes a further step on one dimensional mediator case and suggests using simple linear regression to estimate direct effect and indirect effect for simple mediation model. Suppose one dimensional mediation is denoted as:

$$\begin{aligned} M &= \beta_{01} + aX + e_1, \\ Y &= \beta_{02} + c'X + bM + e_2, \end{aligned} \tag{2.3.1}$$

where $M_{n\times 1}$ is one-dimensional mediator, $X_{n\times 1}$ is one-dimensional exposure variable, $Y_{n\times 1}$ is one-dimensional response variable, n is the sample size. β_{01} and β_{02} are the intercept of regression of M on X and Y on (M, X) respectively. e_1 and e_2 are the error terms in estimation of M and Y respectively. Coefficients are estimated as if the two equations are independent using odinary least squares (OLS). To be specific:

$$\widehat{a} = (X^T X)^{-1}(X^T M),$$

$$\begin{pmatrix} \widehat{c'} \\ \widehat{b} \end{pmatrix} = [(X, M)^T (X, M)]^{-1}[(X, M)^T Y].$$

Direct effect is c' which means that given $M = m$, how much variation of Y can be explained by X. If c' is positive, higher on X is expected to have higher on Y. If c' is negative, higher on X is expected to have lower on Y. If X is binary taking value only 0 or 1, then c' shows the group difference holding M as constant.

Indirect effect is the product of a and b, $\widehat{\gamma} = \widehat{a}\widehat{b}$. a shows how much the variation in M is the result of X. If a is positive, higher on X will expect to have higher on M and vice versa. b shows how much the variation in Y is the result of M holding X as a constant. If b is positive, higher on X will expect to have higher on M and vice versa. If a and b are same sign, then the interpretation of mediator is simple. If a and b are both positive, then higher on X is expected to have higher on M resulting in higher on Y. Similar for the case when both a and b are negative. If

the sign of a and b are different, then indirect effect is negative meaning that higher on X is expected to have lower on Y. That mediator is a resistance that resist the change in X resulting in reverse change in Y is an example of this case. Total effect $c = c' + ab$ which is also equal to fit a regression of Y on X: $Y = \beta_0 + cX + e$ because direct effect and indirect effect partition the variation of Y that can be explained by X. Expressing Y as only funtion of X:

$$Y = \beta_{02} + b\,(\beta_{01} + aX + e_1) + c'X + e_2 = (\beta_{02} + b\beta_{01}) + (ab + c')X + (be_1 + e_2),$$

For hypothesis testing of mediation model, testing indirect effect attracts lots of attention (Sobel, 1982; Bollen and Stine, 1990). Sobel (1982) developed Sobel test to test indirect effect, which is asymptotically normal but it needs assumption that a and b is independent. Bollen and Stine (1990) developed bootstrapping to test indirect effect. Bootstrapping is a non-paramatric based method that resampling with replacement to calculate the indirect effect $\gamma = ab$. By repeating this process for several times, bootstrapping gives a distribution of $\widehat{\gamma}$ which can be used to calculate confidence interval based on certain significance level. If zero is in the confidence interval, one can conclude that the indirect effect does not exist otherwise indirect effect exists.

As real world problems are complex, one mediator model is not sufficient for analysis. This leads to extend the model to include more than one intervening variable that stay in the pathway of an exposure variable's effect on the outcome variable. This brings in low dimensional mediation model that has multiple mediators and the number of mediators is fixed and finite. In low dimensional mediation model, mediators can be tested simultaneously or separately. Testing mediators simultaneously is helpful to learn the joint effect of mediators. This approach may fail to identify indirect effect if different mediators are highly correlated and loss information of the intervening effect of individual mediators. In contrast, testing mediators separately is useful to understand the intervening effect of each mediators. Suppose predictors may have opposite effect that cancels out the indirect effect, then the indirect effect γ may close to zero. Then testing the mediators separately is useful to discover mediators that take effects. However, testing separately incur false discovery and spurious correlation issue casting doubts on the

inference. Thus, choosing the testing multiple mediators procedure should align with research goal and data driven approach.

2.3.2 Estimating and testing high-dimensional mediation effects

On account of modern data-collecting technology, mediation analysis has extended its territory to genomics, internet analysis, biomedical research, among other data-intensive fields. This brings in high-dimensional mediators and requires attention of high-dimensional mediation model (HDMM). In HDMM, the number of mediator diverges as sample size increases, meaning that the number of potential mediators is much larger than the sample size. Assuming that only a small proportion of actually functioning mediators, the sparsity pattern is typically unknown. An illustrative example is that millions of DNA methylation markers might mediate pathways linking childhood trauma and cortisol stress reactivity (Houtepen et al., 2016). The high-dimensionality, however, poses both computational and statistical challenges for carrying out efficient mediation analysis. For instance, the traditional structural equation modeling fails due to the rank-deficiency of the observed covariance matrix.

High dimensional mediation model can be presented as:

$$\begin{aligned} Y &= \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \varepsilon_1, \\ \boldsymbol{m} &= \Gamma^T \boldsymbol{x} + \boldsymbol{\varepsilon}, \end{aligned} \tag{2.3.2}$$

where Y is the outcome, $\boldsymbol{m}$ is the p-dimensional mediator, $\boldsymbol{x}$ is the q-dimensional exposure variable, and a^T denotes transpose of a. We assume p is high dimensional, while q is fixed and finite. Correspondingly, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are p- and q-dimensional regression coefficient vectors, and Γ is a $q \times p$ coefficient matrix. Linear model should have intercept, for matrix representation we can assume that the first column of bx is a vector of one. So the first element in Γ is the intercept for model of expose variable and mediator variable and the first element of α_1 is the intercept of total model. Zhang et al. (2016) proposed three steps for estimating and testing high dimensional mediation model.

Step 1 Screen potential mediators from ultra-high dimension to a moderate number that is smaller than the sample size. Use Sure Independence Screening (SIS) (Fan and Lv, 2008) to select a subset $\mathcal{I} = \{1 \leq k \leq p : M_k$ is among the top $d = [2n/\log(n)]$ largest effects for the response $Y\}$. That is, we select top d-th absolute value largest marginal indirect effects assuming all the mediators are independent with each other.

Step 2 Select mediators from the result of Step 1. Use folded concave penalty function for variable selection because it can select correct variables with probability tending to 1. Obtain $\widehat{\alpha}_{0k}, k \in \mathcal{I}$ by minimizing

$$Q = \sum_{i=1}^{n}(Y_i - \sum_{k\in\mathcal{I}} \alpha_{0k}M_{ik} - \alpha_1 X_i)^2 + \sum_{k\in\mathcal{I}} p_\lambda(|\alpha_{0k}|),$$

where $p_\lambda(|\alpha_{0k}|) = \lambda\left[|\beta_k| - \frac{|\beta_k|^2}{2\lambda}\right] I\{0 \leq |\beta_k| < \lambda\} + \frac{\lambda^2}{2} I\{|\beta_k| \geq \lambda\}$ is minimax concave penalty (MCP).

Step 3 Carrying out joint significance testing for mediation effect. Testing $H_0 : \alpha_{0k} = 0$. The raw p-value is $P_{raw,1k} = 2[1 - \Phi(\frac{|\alpha_{0k}|}{\widehat{\sigma}_{1k}})]$, where $\Phi(\cdot)$ is the cumulative distribution of standard normal. Since comparing multiple test simultaneously, Bonferroni correction is applied to adjust raw p-value:

$$P_{\text{corr, }1k} = \min\left(P_{\text{raw ,}1k} \cdot |\mathcal{S}|, 1\right)$$

where $\mathcal{S} = \{k : \widehat{\alpha}_{0k} \neq 0\}$. The validity of the joint significance hypothesis testing is ensured because the MCP has the model selection consistency Zhang (2010).

The above three steps is to test indirect effect of relationship between Y and $\boldsymbol{M}$. Similarly, the raw P-value for testing $H_0 : \Gamma_k = 0$:

$$P_{\text{raw ,}2k} = 2\left\{1 - \Phi\left(\frac{|\widehat{\Gamma_k}|}{\widehat{\sigma}_{2k}}\right)\right\}.$$

$\widehat{\Gamma}_k$ is the ordinary least squares estimator for Γ_k and σ_{2k} is the corresponding

estimated standard error. Similar to step 3, the Bonferroni corrected p-value is

$$P_{\text{corr, },2k} = \min\left(P_{\text{raw, }2k} \cdot |\mathcal{S}|, 1\right).$$

When both Γ_k and α_{0k} are significant, the corresponding M_k will be selected as significant mediator. Therefore, the Bonferroni correction for indirect effect is

$$P_{\text{corr, }k} = \max\left(P_{\text{corr },1k}, P_{\text{corr },2k}\right).$$

Simulation study shows that the test is powerful when mediators value is close to zero and power may decrease as the number potential non-zero mediators increases (Zhou et al., 2020).

Zhang et al. (2016) apply this methodology to genomic data to analyze the relationship between smoking and lung function mediated by DNA methylation markers. Since the number of DNA methylation markers is much larger than the sample size, SIS is introduced to quickly reduce the mediators dimension. Then use penalized likelihood for variable selection. Since smoking will reduce lung function therefore positive indirect effect mediators are removed. They found several methylation markers that are associated with lung function are all well supported by current literature study of their role in lung functions. This application shows that high dimensional mediation model has promising future for genomic data analysis.

Zhang et al. (2016) method is easy to implement and useful for ultra-high dimensional mediator scenario. Note that they do not provide theory about the asymptotic distribution of test statistics making the inference without the support of certain distribution theory. In addition, the method use Bonferroni correction for multiple testing, and the test will be too rigid to show significance when the number of potential mediator is large.

2.3.3 Debiased method for estimation and testing

Zhou et al. (2020) propose an inference procedure to testing indirect effect in high-dimensional mediator scenario and develop asymptotic distribution for estimation as well as test statistics. Considering the estimation and inference for incomplete

mediation, the model is setup as following. For the i-th subject, Y_i be the outcome M_i be vector of p mediators and X_i be a vector of q exposures, assuming p to be larger than the sample size n but X_i is low-dimensional

$$\begin{aligned} Y_i &= M_i^\top \alpha_0 + X_i^\top \alpha_1 + \epsilon_{1i}, \\ M_i^\top \alpha_0 &= X_i^\top \beta_0 + \epsilon_{2i}. \end{aligned} \tag{2.3.3}$$

where $\epsilon_1 \sim N(0, \sigma_1^2)$ and $\epsilon_2 \sim N(0, \sigma_2^2)$ are normal error terms independent of M_i and X_i. β_0 denotes indirect effect and α_1 denote direct effect. Note that the second regression model contains unknown parameter α_0 in the response, which is not a standard regression. However, the least squares estimator would be the best estimator for β_0 because inference for indirect effect β_0 is the main goal, that is, $\left(\sum_i X_i X_i^\top\right)^{-1}\left(\sum_i X_i M_i^\top \alpha_0\right)$. However, the result is biased if high-dimensional α_0 is estimated by LASSO. The authors propose debias quantity for $E(X_i M_i^\top \alpha_0)$ which is derived from Karush-Kuhn-Tucker conditions and constrained ℓ_1 optimization. Z is the $n \times (p+q)$ design matrix (M, X) and the estimate $\widehat{\alpha} = (\widehat{\alpha}_0^\top, \widehat{\alpha}_1^\top)^\top$ is obtained by the scaled LASSO that satisfies Karush-Kuhn-Tucker condition:

$$\lambda_n \widehat{\kappa} = \frac{1}{n} Z^\top (Y - Z\widehat{\alpha}) = \frac{1}{n} Z^\top \left\{ Z(\alpha - \widehat{\alpha}) + \epsilon_1 \right\} = \widehat{\Sigma}_{ZZ}(\alpha - \widehat{\alpha}) + \frac{1}{n} Z^\top \epsilon_1, \tag{2.3.4}$$

where λ_n is the tuning parameter of scaled LASSO and $\widehat{\Sigma}_{ZZ}$ is the covariance matrix of Z. If there is a matrix Ω_{ZZ} such that

$$\widehat{\Omega}_{ZZ}\widehat{\Sigma}_{ZZ} \approx \begin{pmatrix} \widehat{\Sigma}_{XM} & 0 \\ 0 & \widehat{\Sigma}_{XX} \end{pmatrix} \equiv \widehat{D}, \tag{2.3.5}$$

where $\widehat{\Sigma}_{XM} = n^{-1}\sum_i X_i M_i^\top$ and $\widehat{\Sigma}_{XX} = n^{-1}\sum_i X_i X_i^\top$ then plug in 2.3.4

$$\widehat{D}(\widehat{\alpha} - \alpha) + \lambda_n \widehat{\Omega}_{ZZ}\widehat{\kappa}\widehat{\kappa} = \widehat{D}(\widehat{\alpha} - \alpha) + {}^{-1} Z^\top (Y - Z\widehat{\alpha} + Z\alpha - Z\alpha) = n^{-1}\widehat{\Omega}_{ZZ}\widehat{\kappa} Z^\top \epsilon_1 + \widehat{\Delta}, \tag{2.3.6}$$

where $\widehat{\Delta} = (\widehat{D} - \widehat{\omega}_{ZZ}\widehat{\Sigma}_{ZZ})(\widehat{\alpha} - \alpha)$. Therefore, Zhou et al. (2020) proposed $\widehat{b}$ and $\widehat{a}$ as debiased estimator of β_0 and α_0, respectively. The estimators is asymptotically

normal if $\left(I_2 \otimes \Sigma_{XX}^{-1}\right) \widehat{\Delta}$ is small,

$$\begin{pmatrix} \widehat{b} \\ \widehat{a} \end{pmatrix} = \begin{pmatrix} \widehat{\Sigma}_{XX}^{-1} \widehat{\Sigma}_{XM} \widehat{\alpha}_0 \\ \widehat{\alpha}_1 \end{pmatrix} + I_2 \otimes \widehat{\Sigma}_{XX}^{-1} \lambda_n \widehat{\Omega}_{ZZ} \widehat{\kappa}, \tag{2.3.7}$$

where I_2 is the 2×2 identity matrix. .

It remains to find $\widehat{\Omega}$, which is difficult because $(p+q) > n$. So the constrained ℓ_1 optimization, which is similar to the precision matrix estimation procedure of Cai et al. (2011), is used for estimating $\widehat{\Omega}$.

$$\min |\Omega|_1 \text{ subject to } \left| \widehat{D} - \Omega \widehat{\Sigma}_{ZZ} \right|_\infty \leq \tau_n,$$

where τ_n is a tuning parameter. It is proved that under certain condition $\widehat{\Sigma}_{ZZ}$ converge to true $\Omega = E(\widehat{D})\Sigma_{ZZ}^{-1}$, which facilitates the proof of convergence and asymptotic variance of $\widehat{b}$ and $\widehat{a}$. Upon obtaining $\widehat{\Omega}$, it natural to derive the asymptotic distribution of the estimators

$$n^{1/2} \begin{pmatrix} \widehat{b} - \beta_0 \\ \widehat{a} - \alpha_1 \end{pmatrix} \to N(0, V), \text{ where } V = \begin{pmatrix} \sigma_1^2 \Gamma + \sigma_2^2 \Sigma_{XX}^{-1} & -\sigma_1^2 \Gamma \\ -\sigma_1^2 \Gamma & \sigma_1^2 \left(\Gamma + \Sigma_{XX}^{-1}\right) \end{pmatrix} \tag{2.3.8}$$

where $\Gamma = \Sigma_{XX}^{-1} \Sigma_{XM} (\Sigma_{MM} - \Sigma_{MX} \Sigma_{XX}^{-1} \Sigma_{MX}) \Sigma_{MX} \Sigma_{XX}^{-1}$.

Based on the asymptotic distribution, Zhou et al. (2020) propose using wald test for indirect denoted as Sn,

$$S_n = n \frac{\widehat{b}^2}{\sigma_1^2 \Gamma + \sigma_2^2 \Sigma_{XX}^{-1}}.$$

Zhou et al. (2020) work is ground breaking to propose theory for high-dimension mediation model. If the direction of mediation paper is correctly specified, the model's estimated coefficients are root-n consistent and asymptotically normal. Indirect effect test is established and simulation result shows the nice property of it. But the property of direct effect is not established. And using the direct effect $\widehat{a}$ for testing will have low power. So a more powerful test of direct effect when the indirect effect is present requires further study.

It is challenging to obtain $\widehat{\Omega}$ and Cai et al. (2011) propose CLIME to solve it.

For p-variate random vector $\mathbf{X} = (X_1, \cdots, X_p)^T$, its covariance matrix is denoted as $\mathbf{\Sigma}_0$ and precision matrix is denoted as $\mathbf{\Omega}_0$. For samples $\mathbf{X}_1, \cdots, \mathbf{X}_n$ the best estimator for $\mathbf{\Sigma}_0$ is

$$\mathbf{\Sigma_n} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbf{X}_k - \overline{\mathbf{X}}\right)\left(\mathbf{X}_k - \overline{\mathbf{X}}\right)^T.$$

where $\overline{\mathbf{X}} = n^{-1}\sum_{k=1}^{n}\mathbf{X}_k$. But $\mathbf{\Sigma}_n$ is singular if $n < p$ and the inverse of $\mathbf{\Sigma}_n$ does not exist. To overcome this difficulty, previous work impose certain structure assumption on the $\mathbf{\Sigma}_n$ for example Bickel and Levina (2008) studied banded co-variance structure commonly seen in time series dataset. Estimation of the precision matrix $\mathbf{\Omega}_0$ is more challenging because this process needs unknown pivotal estimator while estimating $\mathbf{\Sigma}_0$ already has.

Motivated by graphical LASSO that the difference between the inverse of solution matrix and covariance matrix is smaller than λ_n, Cai et al. (2011) introduce a new method of constrained ℓ_1-minimization for inverse matrix estimation (CLIME) and prove that convergence rate is faster than maximum likelihood with LASSO penalty. Let $\{\widehat{\mathbf{\Omega}}_1\}$ be the solution set of following problem

$\min \|\mathbf{\Omega}\|_1 \quad$ subject to:

$$\left|\widehat{\mathbf{\Sigma}}_n\widehat{\mathbf{\Omega}} - \mathbf{I}\right|_\infty \leq \lambda_n, \quad \mathbf{\Omega} \in \mathbb{R}^{p\times p}, \tag{2.3.9}$$

where λ_n is tuning parameter. $\widehat{\mathbf{\Omega}}_1 = \left(\widehat{\omega}_{ij}^1\right) = \left(\widehat{\omega}_1^1, \ldots, \widehat{\omega}_p^1\right)$. CLIME estimator $\widehat{\mathbf{\Omega}}$ of $\mathbf{\Omega}_0$ is

$$\begin{aligned} &\widehat{\mathbf{\Omega}} = \left(\widehat{\omega}_{ij}\right), \quad \text{where} \\ &\widehat{\omega}_{ij} = \widehat{\omega}_{ji} = \widehat{\omega}_{ij}^1 I\left\{\left|\widehat{\omega}_{ij}^1\right| \leq \left|\widehat{\omega}_{ji}^1\right|\right\} + \widehat{\omega}_{ji}^1 I\left\{\left|\widehat{\omega}_{ij}^1\right| > \left|\widehat{\omega}_{ji}^1\right|\right\}. \end{aligned} \tag{2.3.10}$$

$\widehat{\mathbf{\Omega}}$ is symmetric and prove to be positive definite with high probability. Convex optimization problem 2.3.9 can be solved component-wise. Cai et al. (2011) prove that $\widehat{\mathbf{\Omega}}_1 = \{\widehat{\beta}_1, \cdots, \widehat{\beta}_p\}$, where the right hand side is the solution of problem (2.3.11)

$$\min |\boldsymbol{\beta}|_1 \quad \text{subject to} \quad |\mathbf{\Sigma}_n\boldsymbol{\beta} - \mathbf{e}_i|_\infty \leq \lambda_n,, \tag{2.3.11}$$

where $\beta \in \mathbb{R}^P$, $\mathbf{e}_i$ be a standard unit vector in $\mathbb{R}^P$ with 1 in the ith coordinate and 0 in all other coordinates.

CLIME has computational advantage in high-dimensional setting because it can be obtained one column at a time through linear programming and then combine the vectors to get final precision matrix. But the solution is biased for large true coefficient.

Chapter 3

Linear Mediation Models with High-dimensional Mediators

Mediation analysis draws increasing attention in many scientific areas such as genomics, epidemiology and finance (MacKinnon, 2012; Conti et al., 2016; Barfield et al., 2017). This requires considering mediation model with potential high dimensional mediators. Traditional estimation and inference approaches for mediation analysis cannot be used to make statistical inference for high dimensional linear mediation models (HDMM) due to high-dimensionality of the mediators. To overcome the challenge, many existing methods utilize the dimension reduction techniques in regular linear models. For example, Huang and Pan (2016) and Chén et al. (2018) adopted principal components analysis to compress the dimensionality of mediators, and applied bootstrap for inference, which is lack of proof for asymptotic distributions of the test statistics. To improve the work by Huang and Pan (2016), Zhao et al. (2020) adapted sparse principal component analysis to HDMM. Zhou et al. (2020) implemented debiased penalized estimators for the indirect effects, and provided theoretical justifications. But their method requires heavy computation cost and suffers numerical instability because of estimating high dimensional matrices.

In this chapter, we propose new statistical estimation procedures for HDMM, via a partial penalized least squares method. Next, we develop a partial penalized Wald test on the indirect effects, and prove that the proposed test has a χ^2 limiting null distribution. We also propose an F-type test for direct effects and show that

the proposed test asymptotically follows a χ^2-distribution under null hypothesis and a noncentral χ^2-distribution under local alternatives. Monte Carlo simulations are conducted to examine the finite sample performance of the proposed tests and compare their performance with existing ones. Then, we apply the newly proposed statistical inference procedures to study stock reaction to COVID-19 pandemic via an empirical analysis of studying the mediation effects of financial metrics that bridge company's sector and stock return.

In section 3.1, we propose a new statistical inference procedure for the indirect effect and establish its theoretical properties. We also construct an F-type test for the direct effect. Section 3.2 presents numerical studies and a real data example. Conclusion and discussion are given in section 3.3. All proofs about HDMM are presented in section 3.4.

3.1 Tests of hypotheses on indirect and direct effects

Consider the mediation models

$$y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \varepsilon_1, \tag{3.1.1}$$

$$\boldsymbol{m} = \Gamma^T \boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{3.1.2}$$

where y is the outcome, $\boldsymbol{m}$ is the p-dimensional mediator, $\boldsymbol{x}$ is the q-dimensional exposure variable, and a^T denotes transpose of a. We in this paper assume p is high dimensional, while q is fixed and finite. Correspondingly, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are p- and q-dimensional regression coefficient vectors, and Γ is a $q \times p$ coefficient matrix. Following the literature on high-dimensional mediation model (Zhang et al., 2016; van Kesteren and Oberski, 2019; Zhou et al., 2020), we impose a sparsity assumption that only a small proportion of entries in $\boldsymbol{\alpha}_0$ are nonzero. This implies that the corresponding variables in $\boldsymbol{m}$ are actually relevant to y. Notably, from equation (3.1.2), $\boldsymbol{m}$ must be random. We further assume that ε_1 and $\boldsymbol{\varepsilon}$ are independent random errors with $\text{var}(\varepsilon_1) = \sigma_1^2$ and $\text{cov}(\boldsymbol{\varepsilon}) = \Sigma^*$; ε_1 is independent of $\boldsymbol{m}, \boldsymbol{x}$, and $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{x}$.

Plugging (3.1.2) into (3.1.1) yields

$$y = (\boldsymbol{\beta} + \boldsymbol{\alpha}_1)^T \boldsymbol{x} + \varepsilon_1 + \varepsilon_2 = \boldsymbol{\gamma}^T \boldsymbol{x} + \varepsilon_3, \tag{3.1.3}$$

where $\boldsymbol{\beta} = \Gamma\alpha_0, \varepsilon_2 = \boldsymbol{\alpha}_0^T\boldsymbol{\varepsilon}$ with $\text{var}(\varepsilon_2) = \sigma_2^2 = \boldsymbol{\alpha}_0^T\Sigma^*\boldsymbol{\alpha}_0, \boldsymbol{\gamma} = \boldsymbol{\beta} + \boldsymbol{\alpha}_1$, and $\varepsilon_3 = \varepsilon_1 + \varepsilon_2$ is the total random error. Following the literature (Imai et al., 2010; VanderWeele and Vansteelandt, 2014), we refer $\boldsymbol{\beta}$ to the indirect effect of $\boldsymbol{x}$ on y mediated by $\boldsymbol{m}$, $\boldsymbol{\alpha}_1$ to the direct effect, and $\boldsymbol{\gamma} = \boldsymbol{\alpha}_1 + \boldsymbol{\beta}$ to the total effect. A causal interpretation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}_1$ is briefly discussed in the Appendix.

3.1.1 Estimating indirect and direct effects

In practice, of interest is to test whether there exists significant (joint) indirect effect or not. This can be formulated as the following hypothesis testing problem

$$H_0 : \boldsymbol{\beta} = 0 \text{ versus } H_1 : \boldsymbol{\beta} \neq 0. \tag{3.1.4}$$

When both p and q are finite-dimensional, $\boldsymbol{\beta}$ can be estimated through $\widehat{\boldsymbol{\beta}} = \widehat{\Gamma}\widehat{\boldsymbol{\alpha}}_0$, where $\widehat{\Gamma}$ and $\widehat{\boldsymbol{\alpha}}_0$ are $\sqrt{n}$-consistently estimated from models (3.1.1) and (3.1.2). That is, $\widehat{\Gamma} = \Gamma + \boldsymbol{E}_\gamma$ and $\widehat{\boldsymbol{\alpha}}_0 = \boldsymbol{\alpha}_0 + \boldsymbol{e}_\alpha$, where $\boldsymbol{E}_\gamma = O_P(1/\sqrt{n})$ and $\boldsymbol{e}_\alpha = O_P(1/\sqrt{n})$ are estimation errors. Then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| \leq \|\Gamma\boldsymbol{e}_\alpha\| + \|\boldsymbol{E}_\gamma\boldsymbol{\alpha}_0\| + \|\boldsymbol{E}_\gamma\boldsymbol{e}_\alpha\| = O_P(1/\sqrt{n}), \tag{3.1.5}$$

where $\|\cdot\|$ stands for the Euclidean norm.

When p is high-dimensional, however, the right-hand side of (3.1.5) is no longer $O_P(1/\sqrt{n})$. This results in potentially non-ignorable estimation error of $\widehat{\boldsymbol{\beta}}$. Moreover, $\boldsymbol{\beta}$ is challenging to be estimated through $\Gamma\boldsymbol{\alpha}_0$ as it involves estimation of a high-dimensional matrix and a high-dimensional vector, though, interestingly, $\boldsymbol{\beta} = \Gamma\boldsymbol{\alpha}_0$ is q-dimensional, fixed and finite.

As a key observation from (3.1.3), the indirect effect $\boldsymbol{\beta} = \boldsymbol{\gamma} - \boldsymbol{\alpha}_1$, is the difference between the total effect and direct effect. This motivates us to estimate $\boldsymbol{\beta}$ by separately estimating $\boldsymbol{\gamma}$ via (3.1.3) and $\boldsymbol{\alpha}_1$ via (3.1.1), respectively, rather than estimating the high-dimensional Γ and $\boldsymbol{\alpha}_0$.

Suppose that $\{\boldsymbol{m}_i, \boldsymbol{x}_i, y_i\}$, $i = 1, \cdots, n$ is a random sample from (3.1.1) and (3.1.2). Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$. Then we estimate $\boldsymbol{\gamma}$ by its least squares estimate

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{3.1.6}$$

While for the estimator of $\boldsymbol{\alpha}_1$, due to the high-dimensionality of $\boldsymbol{\alpha}_0$, we propose the following partial penalized least squares method:

$$(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_0) = \arg\min_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|), \tag{3.1.7}$$

where $\boldsymbol{M} = (\boldsymbol{m}_1, \cdots, \boldsymbol{m}_n)^T$ and $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. The regularization is only applied to the high-dimensional yet sparse $\boldsymbol{\alpha}_0$. We opt not penalize $\boldsymbol{\alpha}_1$ to achieve local power on the direct effect $\boldsymbol{\alpha}_1$ and the indirect effect $\boldsymbol{\beta}$ under local alternatives. See Theorem 2 and Corollary 1 below for more details. Thus, our proposal is different from Zhou et al. (2020), in which the central idea is to develop a debiased estimator not of $\boldsymbol{\alpha}_0$ or $\boldsymbol{\beta}$, but of $\tilde{\Sigma}_{XM}\boldsymbol{\alpha}_0$ with $\tilde{\Sigma}_{XM} = E[\boldsymbol{x}\boldsymbol{m}^T]$. This may lead to less efficient estimators due to debiasing, as discussed in the next subsection.

3.1.2 Theoretical results

In this section, we investigate statistical properties of the estimators. We first present some notations and assumptions. For the penalty function, it is assumed that $p_\lambda(t_0)$ is increasing and concave in $t_0 \in [0, \infty)$, and has a continuous derivative $p'_\lambda(t_0)$ with $p'_\lambda(0+) > 0$. Denote $\rho(t_0, \lambda) = p_\lambda(t_0)/\lambda$ for $\lambda > 0$. Further, $\rho'(t_0, \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+, \lambda)$ does not depend on λ. Define $\bar{\rho}(\boldsymbol{v}, \lambda) = \{\text{sgn}(v_1)\rho'(|v_1|, \lambda), \cdots, \text{sgn}(v_l)\rho'(|v_l|, \lambda)\}^T$ for any vector $\boldsymbol{v} = (v_1, \cdots, v_l)^T$, where $\text{sgn}(\cdot)$ is the sign function. Define the local concavity of $\rho(\cdot)$ at $\boldsymbol{v}$ as

$$\kappa(\rho, \boldsymbol{v}, \lambda) = \lim_{\epsilon \to 0^+} \max_{1 \le j \le l} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\rho'(t_2, \lambda) - \rho'(t_1, \lambda)}{t_2 - t_1}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_0^T)^T$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_1^{\star T}, \boldsymbol{\alpha}_0^{\star T})^T$, the true value of $\boldsymbol{\theta}$. Further let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}_1^T, \widehat{\boldsymbol{\alpha}}_0^T)$ be the estimator of $\boldsymbol{\theta}_0$. Denote $\mathcal{A} = \{j : \alpha_{0j}^\star \neq 0\}$, and $s = |\mathcal{A}|$ is

the number of elements in $\mathcal{A}$. Moreover, $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_{0,\mathcal{A}}^T)^T$. And $\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\vartheta}}$ are similarly defined. Let $\boldsymbol{M}^j$ denote the jth column of $\boldsymbol{M}$. Let $\boldsymbol{M}_{\mathcal{A}}$ be the submatrix of $\boldsymbol{M}$ formed by columns in $\mathcal{A}$. $\boldsymbol{m}_{i,\mathcal{A}}$ is the ith column of the matrix $\boldsymbol{M}_{\mathcal{A}}^T$. Similarly, let $\boldsymbol{\alpha}_{0,\mathcal{A}}^\star$ be the subvector of $\boldsymbol{\alpha}_0^\star$ formed by elements in $\mathcal{A}$. Define $\mathcal{A}^c = [1, \cdots, p] - \mathcal{A}$ as the complement set of $\mathcal{A}$. Define $\mathcal{N}_0 = \{\boldsymbol{\delta} \in R^s : \|\boldsymbol{\delta} - \boldsymbol{\alpha}_{0,\mathcal{A}}^\star\|_2 \leq d_n\}$. Let $\Sigma_{MM} = E[\boldsymbol{m}_{\mathcal{A}}\boldsymbol{m}_{\mathcal{A}}^T]$, $\Sigma_{MX} = E[\boldsymbol{m}_{\mathcal{A}}\boldsymbol{x}^T]$, and $\Sigma_{XX} = E[\boldsymbol{x}\boldsymbol{x}^T]$. Denote

$$\Sigma \;=\; \begin{pmatrix} \Sigma_{XX} & \Sigma_{XM} \\ \Sigma_{MX} & \Sigma_{MM} \end{pmatrix}.$$

In this paper, for any vector $\boldsymbol{v} = (v_1, \cdots, v_l)^T$, $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$ and $\|\boldsymbol{v}\|_2 = (\boldsymbol{v}^T\boldsymbol{v})^{1/2}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denotes the minimum and maximum eigenvalues of the matrix A, respectively. $\|A\|_{2,\infty} = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_\infty$. Further $a \gg b$ means $\lim_{n\to\infty} a/b = \infty$. We impose the following conditions:

A1. $\lambda_{\min}(\Sigma) \geq c > 0$, $\lambda_{\max}(\Sigma) = O(1)$, and $\|\boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})\|_{2,\infty} = O_P(n)$.

A2. Let d_n be the half minimum signal of $\boldsymbol{\alpha}_{0,\mathcal{A}}^\star$, i.e. $d_n = \min_{j\in\mathcal{A}} |\alpha_{0j}^\star|/2$. Assume that $d_n \gg \lambda_n \gg \max\{\sqrt{s/n}, \sqrt{\log p/n}\}$, $p'_{\lambda_n}(d_n) = o((ns)^{-1/2})$, $\lambda_n\kappa_0 = o(1)$ where $\kappa_0 = \max_{\delta\in\mathcal{N}_0} \kappa(\rho, \delta, \lambda_n)$.

A3. For some $\varpi > 2$, there exists a positive sequence K_n such that

$E[\|\boldsymbol{m}_{\mathcal{A}^c}\varepsilon_1\|_\infty^\varpi] \leq K_n^\varpi$ and $K_n^2 \log p/n^{1-2/\varpi-\varsigma} \to 0$ for some arbitrary small $\varsigma > 0$. Further assume that $\max_{1\leq j\leq p+q} E(z_j^4) < C < \infty$, here $\boldsymbol{z} = (\boldsymbol{m}, \boldsymbol{x})$, z_j is the j-th component of $\boldsymbol{z}$.

To emphasize the dependence on the sample size, in the above conditions and the Appendix, we use λ_n to denote the tuning parameter. The first two conditions are mild and commonly assumed. See for instance Fan and Lv (2011). Condition A2 imposes a minimal signal condition on nonzero elements in $\boldsymbol{\alpha}_0$, but not on $\boldsymbol{\alpha}_1$. Since our primary interest is to make statistical inference on direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta} = \boldsymbol{\gamma} - \boldsymbol{\alpha}_1$, and $\boldsymbol{\alpha}_0$ may be treated as a nuisance parameter in this model. Thus, Condition A2 is reasonable in practice. Condition A3 is imposed for establishing sparsity result. Compared with existing literature, A3 is very mild. In fact, to simplify the proof, some papers assume that all covariates are

uniformly bounded - see for instance Wang et al. (2012). Under bounded covariates condition, A3 reduces to $E(|\varepsilon_1|^{\varpi}) \leq C$ by taking K_n as a constant. Furthermore, the dimension of p is allowed to be an exponential order of the sample size n according to conditions A2 and A3.

Theorem 1. *Suppose that Conditions (A1)-(A3) hold, and $s = o(n^{1/2})$, then with probability tending to 1, $\widehat{\boldsymbol{\alpha}}_0$ must satisfy (i) $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$. (ii) $\|\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}^{\star}_{0,\mathcal{A}}\|_2 = O_P(\sqrt{s/n})$. Let $\epsilon_1 = (\varepsilon_{11}, \cdots, \varepsilon_{n1})^T$. If further $s = o(n^{1/3})$, we obtain that*

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \;=\; \frac{1}{\sqrt{n}}\Sigma^{-1}\begin{pmatrix} \boldsymbol{X}^T\epsilon_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\epsilon_1 \end{pmatrix} + o_P(1).$$

The above results provide the sparsity of $\widehat{\boldsymbol{\alpha}}_0$, the convergence rate of $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}$ and the asymptotic representation of $\widehat{\boldsymbol{\vartheta}}$, respectively.

Based on the results in Theorem 1, we further obtain the following corollary:

Corollary 1. *Suppose that Conditions (A1)-(A3) hold, and $s = o(n^{1/3})$,*

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^{\star}) \to N(0, \sigma_1^2(\Sigma_{XX}^{-1} + B)), \text{ and}$$
$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\star}) \to N(0, \sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2 B),$$

where $B = \Sigma_{XX}^{-1}\Sigma_{XM}(\Sigma_{MM} - \Sigma_{MX}\Sigma_{XX}^{-1}\Sigma_{XM})^{-1}\Sigma_{MX}\Sigma_{XX}^{-1}$, and $\boldsymbol{\beta}^{\star}$ is the true value of $\boldsymbol{\beta}$.

This corollary presents the asymptotic normalities of the estimators $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$. We next make theoretical comparison with the estimators in Zhou et al. (2020). Note that the asymptotic variance matrices of $\widehat{\boldsymbol{\alpha}}_1^Z$ and $\widehat{\boldsymbol{\beta}}^Z$ in Zhou et al. (2020) are $\sigma_1^2(\Sigma_{XX}^{-1} + \tilde{B})$ and $\sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2\tilde{B}$, respectively, where $\tilde{\Sigma}_{MM} = E[\boldsymbol{m}\boldsymbol{m}^T], \tilde{\Sigma}_{MX} = E[\boldsymbol{m}\boldsymbol{x}^T]$, $\Sigma_{XX} = E[\boldsymbol{x}\boldsymbol{x}^T]$, and $\tilde{B} = \Sigma_{XX}^{-1}\tilde{\Sigma}_{XM}(\tilde{\Sigma}_{MM} - \tilde{\Sigma}_{MX}\Sigma_{XX}^{-1}\tilde{\Sigma}_{XM})^{-1}\tilde{\Sigma}_{MX}\Sigma_{XX}^{-1}$. To show our proposed estimators are more efficient than those proposed in Zhou et al. (2020), it suffices to show that $\tilde{B} > B$. Note that $\Sigma_{XX}^{-1} + B = (I_q, 0_{q\times s})\Sigma^{-1}(I_q, 0_{q\times s})^T$, and

$$\begin{aligned}\Sigma_{XX}^{-1} + \tilde{B} &= (I_q, 0_{q\times p})\begin{pmatrix} E[\boldsymbol{x}\boldsymbol{x}^T] & E[\boldsymbol{x}\boldsymbol{m}^T] \\ E[\boldsymbol{m}\boldsymbol{x}^T] & E[\boldsymbol{m}\boldsymbol{m}^T] \end{pmatrix}^{-1}(I_q, 0_{q\times p})^T \\ &= (I_q, 0_{q\times s})(\Sigma - E[\boldsymbol{x}\boldsymbol{m}_{\mathcal{A}^c}^T]E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{m}_{\mathcal{A}^c}^T]^{-1}E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{x}^T])^{-1}(I_q, 0_{q\times s})^T.\end{aligned}$$

Thus, $\tilde{B} > B$ since $(\Sigma - E[\boldsymbol{x}\boldsymbol{m}_{\mathcal{A}^c}^T]E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{m}_{\mathcal{A}^c}^T]^{-1}E[\boldsymbol{m}_{\mathcal{A}^c}\boldsymbol{x}^T])^{-1} > \Sigma^{-1}$. Hence our proposed estimators are more efficient than those proposed in Zhou et al. (2020). This should not be surprised because the debias Lasso inflates its asymptotical variance in the debiased step for high-dimensional linear model (Van de Geer et al., 2014). The proposed partial penalized least squares method does not penalize $\boldsymbol{\alpha}_1$, and hence the debiased step becomes unnecessary.

Under normality assumption that $\varepsilon_1 \sim N(0, \sigma_1^2)$ and $\boldsymbol{\varepsilon} \sim N(0, \Sigma^*)$, it can be shown that our proposed estimators are indeed asymptotically efficient. Under the normality assumption, the maximum likelihood estimator (MLE) of $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_{0,\mathcal{A}}$ in the oracle model knowing $\boldsymbol{\alpha}_{0,\mathcal{A}^c} = 0$ satisfies

$$\begin{pmatrix} \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^M - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1^M) \\ \boldsymbol{M}_{\mathcal{A}}^T(\boldsymbol{y} - \boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^M - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1^M) \end{pmatrix} = 0. \tag{3.1.8}$$

This implies that $\widehat{\boldsymbol{\vartheta}}^M = (\widehat{\boldsymbol{\alpha}}_1^M, \widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^M)$ has the same asymptotic distribution as $\widehat{\boldsymbol{\vartheta}}$.

Since the MLE of $\Gamma_{\mathcal{A}}$ is $\widehat{\Gamma}_{\mathcal{A}}^M = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}}$, the MLE of $\boldsymbol{\beta}$ can be written as

$$\widehat{\boldsymbol{\beta}}^M = \widehat{\Gamma}_{\mathcal{A}}^M\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^M = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^M. \tag{3.1.9}$$

By the definition of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}_1$, it follows that

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\alpha}}_1 \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{M}\widehat{\boldsymbol{\alpha}}_0) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\widehat{\boldsymbol{\alpha}}_0. \end{aligned} \tag{3.1.10}$$

Recall that Theorem 1 indicates that with probability tending to 1, $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$, and hence

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}. \tag{3.1.11}$$

Note that $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}$ and $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^M$ have the same asymptotic distribution. Consequently, $\widehat{\boldsymbol{\beta}}$ has the same asymptotic distribution as $\widehat{\boldsymbol{\beta}}^M$. Thus it is asymptotically efficient.

3.1.3 Test for indirect effect

To form the test statistic for the indirect effect $\boldsymbol{\beta}$, we first study its asymptotic variance matrix. Let $\widehat{\mathcal{A}} = \{j : \widehat{\alpha}_{0j} \neq 0\}$. With probability tending to 1, $\widehat{\mathcal{A}} = \mathcal{A}$. Then the variance matrix Σ and σ_1^2 can be estimated by the estimated sample version and the mean squared errors, respectively.

$$\widehat{\Sigma} = \frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{M}_{\widehat{\mathcal{A}}} \\ \boldsymbol{M}_{\widehat{\mathcal{A}}}^T\boldsymbol{X} & \boldsymbol{M}_{\widehat{\mathcal{A}}}^T\boldsymbol{M}_{\widehat{\mathcal{A}}} \end{pmatrix}, \quad \text{and} \quad \widehat{\sigma}_1^2 = \frac{1}{n-\widehat{s}-q}\|\boldsymbol{y} - \boldsymbol{M}\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1\|^2,$$

where $\widehat{s} = |\widehat{\mathcal{A}}|$. As is shown, $\widehat{\sigma}_1^2 = \sigma_1^2 + o_P(1)$. In fact, when $s = o(n^{1/2})$, $\widehat{\sigma}_1^2 = \sigma_1^2 + O_P(n^{-1/2})$. Alternatively, we can estimate σ_1^2 using refitted cross-validation (Fan et al., 2012) or the scaled lasso (Sun and Zhang, 2012).

As to σ_2^2, we first estimate $\sigma^2 = \text{var}(\varepsilon_3) = \sigma_1^2 + \sigma_2^2$ by the classic least squares residual variance estimator $\widehat{\sigma}^2$ based on model (3.1.3). Thus $\widehat{\sigma}_2^2 = \widehat{\sigma}^2 - \widehat{\sigma}_1^2$. In practice, $\widehat{\sigma}_1^2$ may sometimes be larger than $\widehat{\sigma}^2$, where we would simply set $\widehat{\sigma}_2^2 = 0$. This is possible when no mediators are relevant. That is, $\boldsymbol{\alpha}_0 = 0$, and hence σ_2^2 indeed equals zero.

According to Corollary 1, the asymptotic variance matrices of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ can be consistently estimated by:

$$\widehat{\sigma}_1^2(I_q, 0_{q\times\widehat{s}})\widehat{\Sigma}^{-1}(I_q, 0_{q\times\widehat{s}})^T;\ \widehat{\sigma}_2^2\widehat{\Sigma}_{XX}^{-1} + \widehat{\sigma}_1^2[(I_q, 0_{q\times\widehat{s}})\widehat{\Sigma}^{-1}(I_q, 0_{q\times\widehat{s}})^T - \widehat{\Sigma}_{XX}^{-1}], \quad (3.1.12)$$

where $\widehat{\Sigma}_{XX} = \boldsymbol{X}^T\boldsymbol{X}/n$. Then Wald test statistic for the hypotheses in (3.1.4) can be derived as

$$S_n = n\widehat{\boldsymbol{\beta}}^T\left\{\widehat{\sigma}_2^2\widehat{\Sigma}_{XX}^{-1} + \widehat{\sigma}_1^2[(I_q, 0_{q\times\widehat{s}})\widehat{\Sigma}^{-1}(I_q, 0_{q\times\widehat{s}})^T - \widehat{\Sigma}_{XX}^{-1}]\right\}^{-1}\widehat{\boldsymbol{\beta}}.$$

Clearly, under H_0, $S_n \to \chi_q^2$, a chi-square random variable with q degrees of freedom.

To investigate the local power of S_n, we consider the local alternative hypotheses $H_{1n} : \boldsymbol{\beta} = \boldsymbol{\delta}/\sqrt{n}$, where $\boldsymbol{\delta}$ is a constant vector. From Corollary 1, under such local alternative hypotheses, $S_n \to \chi_q^2(\boldsymbol{\delta}^T(\sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2 B)^{-1}\boldsymbol{\delta})$, a chi-square random variable with q degrees of freedom and noncentrality parameter $\boldsymbol{\delta}^T(\sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2 B)^{-1}\boldsymbol{\delta}$. Thus, S_n can detect local effects that converge to 0 at

root-n rate.

3.1.4 F-type Test on direct effect

It is of interest to test the following hypothesis

$$H_{02} : \boldsymbol{\alpha}_1 = 0 \text{ versus } H_{12} : \boldsymbol{\alpha}_1 \neq 0. \tag{3.1.13}$$

(3.1.1) and (3.1.2) are called complete or full mediation models under H_{02}, while incomplete or partial mediation models under H_{12}.

Testing the hypothesis in (3.1.13) essentially is to test low dimensional regression coefficients in linear regression model (3.1.1). This has been studied when the covariates in (3.1.1) are fixed design (Zhang and Zhang, 2014; Van de Geer et al., 2014; Shi et al., 2019). Due to the nature of mediation model, the covariates in (3.1.1) are random design. The fixed-design assumption on $\boldsymbol{m}$ is inappropriate in mediation models.

We will propose an F-type test for (3.1.13), and further show that the proposed F-test asymptotically has a chi-square distribution with q degrees of freedom under H_{02}, and a noncentral chi-square distribution with q degrees of freedom under H_{12}. Similar to F-test, we need to calculate the residual sum of squares (RSS) under the null and alternative hypotheses. Under H_{02}, the penalized least squares function for model (3.1.1) becomes

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0\|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|). \tag{3.1.14}$$

Denote by $\tilde{\boldsymbol{\alpha}}_0$ the resulting penalized least squares estimator. Then the RSS under H_{02} is $\mathrm{RSS}_0 = \|\boldsymbol{y} - \boldsymbol{M}\tilde{\boldsymbol{\alpha}}_0\|^2$. Under H_{12}, we can estimate $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ by the partial penalized least squares method in (3.1.7). Then we calculate $\mathrm{RSS}_1 = \|\boldsymbol{y} - \boldsymbol{M}\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1\|^2$, the RSS under H_{12}.

The F-type test for hypothesis (3.1.13) is defined to be

$$T_n = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)}{\mathrm{RSS}_1/(n-q)}. \tag{3.1.15}$$

Theorem 2 below shows that the asymptotical null distribution of T_n is a chi-square

distribution with q degrees of freedom. To evaluate the local power of T_n under local alternative hypotheses, we impose the following assumption.

A4. Consider local alternative hypotheses $H_{1n} : \boldsymbol{\alpha}_1 = \boldsymbol{h}_n$. Assume that $\|\boldsymbol{h}_n\|_2 = O(\sqrt{1/n})$.

Theorem 2. *Suppose that Conditions (A1)-(A4) hold, and $s = o(n^{1/3})$. It follows that*

$$\sup_x |P(T_n \leq x) - P(\chi_q^2(n\boldsymbol{h}_n^T \Phi^{-1} \boldsymbol{h}_n/\sigma_1^2) \leq x)| \to 0. \tag{3.1.16}$$

Here $\Phi = (I_q, 0_{q\times s})\Sigma^{-1}(I_q, 0_{q\times s})^T$ and $\chi_q^2(n\boldsymbol{h}_n^T \Phi^{-1} \boldsymbol{h}_n/\sigma_1^2)$ is a chi square random variable with q degrees of freedom and noncentrality parameter $n\boldsymbol{h}_n^T \Phi^{-1} \boldsymbol{h}_n/\sigma_1^2$.

Theorem 2 implies that under H_{02}, T_n asymptotically follows χ_q^2 distribution, which does not depend on any parameter in the model. This is similar to the Wilks phenomenon for likelihood ratio test in classical statistical setting. In other words, the Wilks phenomenon still holds in this high dimensional mediation model. Theorem 2 also implies that T_n can detect local alternatives that are distinct from the null hypothesis at the rate of $1/\sqrt{n}$.

3.1.5 Algorithm and tuning parameter selection

To compute the partial penalized estimators $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$, we apply the local linear approximation algorithm (LLA) in Zou and Li (2008) with the SCAD penalty in Fan and Li (2001),

$$p'_\lambda(t) = \lambda\{I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\},$$

and set $a = 3.7$. The tuning parameter λ for our method is chosen based on the high-dimensional BIC (HBIC) method in Wang et al. (2013). For a fixed regularization parameter λ, define

$$(\widehat{\boldsymbol{\alpha}}_0^\lambda, \widehat{\boldsymbol{\alpha}}_1^\lambda) = \min_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1} \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0,j}|).$$

The minimization of the partial penalized least squares method can be carried out as follows.

1. Get initial values for $\boldsymbol{\alpha}_0^{(0)}, \boldsymbol{\alpha}_1^{(0)}$ by minimizing a partial L_1-penalized least squares: $(\widehat{\boldsymbol{\alpha}}_0^{(0)}, \widehat{\boldsymbol{\alpha}}_1^{(0)}) = \min_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2 + \lambda \sum_{j=1}^{p} |\alpha_{0,j}|$.
2. Solve $(\widehat{\boldsymbol{\alpha}}_0^{(k+1)}, \widehat{\boldsymbol{\alpha}}_1^{(k+1)}) = \min_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2 + \sum_{j=1}^{p} p'_\lambda(|\alpha_{0,j}^{(k)}|)|\alpha_{0,j}|$ for $k = 1, 2, \cdots$, until $\{(\widehat{\boldsymbol{\alpha}}_0^{(k)}, \widehat{\boldsymbol{\alpha}}_1^{(k)})\}$ converges.

In practice, we use a data-driven method to choose the tuning parameter λ. Following Wang et al. (2013), we use the HBIC criterion to choose λ. The HBIC score is defined as $\text{HBIC}(\lambda) = \log(\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2) + \text{df}\log(\log(n))\log(p+q)/n$, where df is the number of variables with nonzero coefficients in $(\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T)^T$. Minimizing $\text{HIBIC}(\lambda)$ yields a selection of λ.

3.2 Numerical studies

In this section, we examine the finite sample performance of the proposed procedures via Monte Carlo simulation studies and illustrate the proposed procedure by a real data example.

3.2.1 Simulation studies

We first examine finite sample performances of the proposed partial-penalization based test statistics, along with comparisons with the oracle test statistics which know the true set $\mathcal{A} = \{j : \alpha_{0j}^\star \neq 0\}$, denoted as S_n^O and T_n^O as a benchmark, and the debiased test statistics S_n^Z and T_n^Z in Zhou et al. (2020), denoted by Zhou et al. (2020)'s method in the tables and figures in this section. Note that Zhou et al. (2020) focuses on the test of indirect effects. One can derive a valid Wald test for direct effects based on the asymptotical normality established in their paper.

Example 1. In this example, we set $n = 300$, $q = 1$, and $p = 500$. $\boldsymbol{x} \sim N(0, 1)$ and $\boldsymbol{m} = \Gamma^T\boldsymbol{x} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(0, \Sigma^*)$ with Σ^* being an AR correlation structure. That is, the (i, j)-element of Σ^* equals $\rho^{|i-j|}$ and ρ is set to be 0.5. Take $\Gamma = c_1(\tau_1, \cdots, \tau_p)^T$, where $\tau_k = 0.2k$ for $k = 1, \cdots, 5$, and when $k > 5$, τ_k's are

independently generated from $N(0, 0.1^2)$. Set $c_1 = 0$ to examine Type I error rate and $c_1 = \pm 0.1, \pm 0.2, \cdots, \pm 1$ for power when testing the indirect effects.

We generate the response y from model $y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 0.5^2)$, $\boldsymbol{\alpha}_0 = [1, 0.8, 0.6, 0.4, 0.2, 0, \cdots 0]^T$ and $\boldsymbol{\alpha}_1 = c_2$ is set in the same fashion as c_1. The simulation results are based on 500 replications. The significance level is set to be 0.05.

We first compare the performances of S_n, S_n^O and S_n^Z for testing the indirect effect $\boldsymbol{\beta}$. We set $c_2 = 0.5$ and $\boldsymbol{\beta} = \Gamma \boldsymbol{\alpha}_0 = 1.4 c_1$. The left panel of Figure 3.1 depicts power functions of the three tests versus the values of c_1 over $[-0.3, 0.3]$. All the three tests gain larger powers as $|c_1|$ increases. S_n performs as well as the oracle S_n^O, and is generally more powerful than S_n^Z. For instance, when $c_1 = -0.2$, the empirical power of S_n^Z is 0.516, while the empirical powers of S_n and S_n^O are 0.596. These observations are in consistent with the theoretical results in Section 2.

Next, we turn to test the direct effect. Set $c_1 = 0.5$. And c_2 is taken from $0, \pm 0.1, \pm 0.2, \cdots, \pm 1$, where $c_2 = 0$ corresponds to the null hypothesis. The right panel of Figure 3.1 depicts the power function of the three tests versus the values of c_2 over $[-0.3, 0.3]$. The proposed test T_n performs almost the same as the oracle one, and is obviously more powerful than the test T_n^Z proposed in Zhou et al. (2020), whose power curve is asymmetric. In fact, when $c_2 = -0.2$, the empirical powers of our test statistic T_n and the oracle test T_n^O are about 1, while that of T_n^Z is only about 0.780.

Furthermore, T_n^Z performs unstably according to our simulation studies. To gain insight of this, we explore more on $\widehat{\boldsymbol{\alpha}}_1^Z, \widehat{\boldsymbol{\beta}}^Z$. The estimates $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}_1^O, \widehat{\boldsymbol{\beta}}^O$ are reported in Table 3.1 from which it can be seen that the biases of $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}_1^O, \widehat{\boldsymbol{\beta}}^O$ are very small, while $\widehat{\boldsymbol{\alpha}}_1^Z$ has a large bias. This may be due to that the direct effect $\boldsymbol{\alpha}_1$ is also penalized in Zhou et al. (2020) 's estimation procedure based on scaled lasso. This makes sense only if the direct effect is expected to be zero. As seen in Table 3.1, the bias of $\widehat{\boldsymbol{\alpha}}_1^Z$ is very small when $c_2 = 0$, yet inversely when $c_2 \neq 0$. Table 3.1 also reports standard errors of corresponding estimates. Both the proposed method and oracle outperform Zhou et al. (2020), especially when estimating $\boldsymbol{\alpha}_1$.

To assess the accuracy of variance estimation of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$, Table 3.2 reports

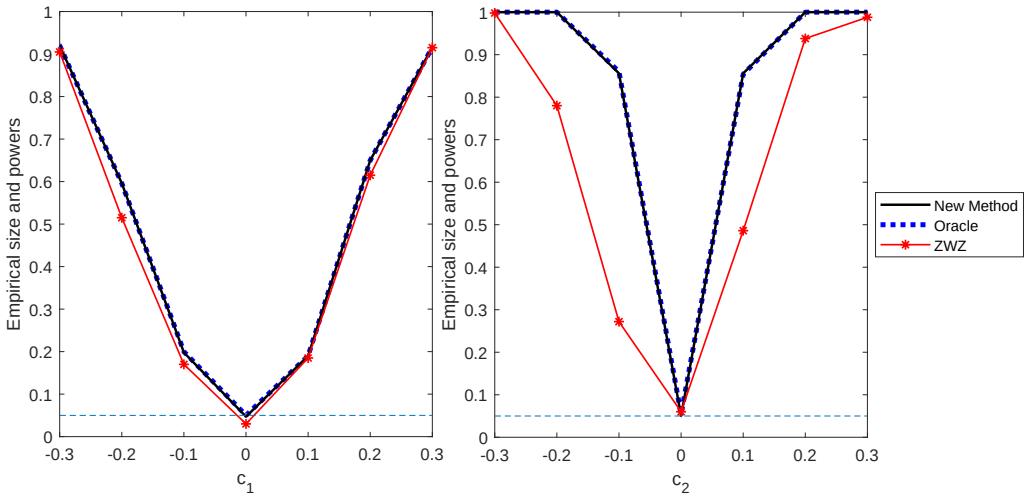

Figure 3.1. Left panel is the empirical sizes and powers of S_n, S_n^Z and S_n^O at level $\alpha = 0.05$ over 500 replications for testing indirect effect when $\alpha_1 = 0.5$. Solid line, dotted line and solid line marked by '*' represent the sizes and powers of S_n, S_n^O, and S_n^Z, respectively. Right panel is empirical sizes and powers of T_n, T_n^Z, and T_n^O at level $\alpha = 0.05$ over 500 replications for testing direct effect when $\beta = 0.7$. The solid line, dotted line, and solid line marked by '*' represent the sizes and powers of T_n, T_n^O, and T_n^Z, respectively.

their estimated standard errors in two ways. As to each method - new, oracle and Zhou et al. (2020)'s method, the first column lists the empirical standard deviations of point estimates $\widehat{\boldsymbol{\alpha}}_1$ or $\widehat{\boldsymbol{\beta}}$ over 500 replications (they are also recorded in parentheses of Table 3.1); for the second column, we estimate standard errors of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ using formula (3.1.12) in each simulation run, and reports the average together with standard deviations (in parentheses) over the 500 runs. Note that the R package "freebird" (Zhou et al., 2020) does not provide the estimated standard error of $\widehat{\boldsymbol{\alpha}}_1$. From Table 3.2, for the new method and oracle, the standard errors estimated by Monte Carlo simulations are close to those calculated from formulas; while the two versions of Zhou et al. (2020) depart more.

Furthermore, Figure 3.2 visually compares the standard deviations of $\widehat{\boldsymbol{\beta}}$ over 500 point estimates using the new method (x-axis) with those using oracle or Zhou et al. (2020)'s method (y-axis), respectively. Each blue diamond or red dot in the figure corresponds to each of the 21 different simulation settings - when holding $c_2 = 0.5$, vary $c_1 = 0, \pm 0.1, \cdots, \pm 1$ in (a) and holding $c_1 = 0.5$, vary $c_2 = 0, \pm 0.1, \cdots, \pm 1$ in (b). The figures imply that the estimated standard errors

Table 3.1. Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		New method		Oracle		Zhou et al. (2020)'s method	
c_1	c_2	$\widehat{\boldsymbol{\alpha}}_1$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\alpha}}_1^O$	$\widehat{\boldsymbol{\beta}}^O$	$\widehat{\boldsymbol{\alpha}}_1^Z$	$\widehat{\boldsymbol{\beta}}^Z$
-0.8	0.5	$-0.23_{(4.15)}$	$-0.22_{(13.73)}$	$-0.11_{(4.11)}$	$-0.35_{(13.70)}$	$-11.77_{(6.56)}$	$11.31_{(14.05)}$
-0.4	0.5	$0.18_{(3.13)}$	$-0.33_{(11.98)}$	$0.25_{(3.08)}$	$-0.40_{(11.95)}$	$-3.49_{(5.10)}$	$3.37_{(12.20)}$
0	0.5	$-0.02_{(2.99)}$	$0.39_{(12.61)}$	$-0.00_{(2.99)}$	$0.37_{(12.63)}$	$-0.13_{(8.65)}$	$0.47_{(15.00)}$
0.4	0.5	$0.02_{(3.15)}$	$0.08_{(11.83)}$	$-0.02_{(3.11)}$	$0.12_{(11.81)}$	$-0.60_{(5.31)}$	$0.77_{(12.66)}$
0.8	0.5	$0.31_{(3.79)}$	$0.26_{(12.69)}$	$0.16_{(3.72)}$	$0.42_{(12.63)}$	$-1.57_{(8.57)}$	$2.19_{(15.05)}$
0.5	-0.8	$0.16_{(3.38)}$	$0.79_{(11.62)}$	$0.11_{(3.37)}$	$0.85_{(11.64)}$	$16.37_{(5.61)}$	$-7.63_{(13.13)}$
0.5	-0.4	$-0.01_{(3.43)}$	$0.16_{(12.58)}$	$-0.09_{(3.36)}$	$0.26_{(12.57)}$	$16.05_{(4.00)}$	$-8.08_{(13.64)}$
0.5	0	$0.10_{(3.35)}$	$-0.15_{(12.52)}$	$0.01_{(3.33)}$	$-0.06_{(12.52)}$	$0.66_{(6.56)}$	$-0.71_{(13.82)}$
0.5	0.4	$0.35_{(3.39)}$	$0.01_{(12.26)}$	$0.32_{(3.37)}$	$0.04_{(12.26)}$	$-0.96_{(5.69)}$	$1.30_{(13.10)}$
0.5	0.8	$0.13_{(3.29)}$	$0.24_{(12.10)}$	$0.05_{(3.26)}$	$0.32_{(12.17)}$	$-0.53_{(5.58)}$	$0.84_{(12.86)}$

Table 3.2. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		Direct effect ($\widehat{\alpha}_1$)				Indirect Effect ($\widehat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al. (2020)'s method	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
-0.8	0.5	4.15	$3.88_{(0.23)}$	4.11	$3.89_{(0.23)}$	13.73	$12.56_{(0.72)}$	13.70	$12.56_{(0.72)}$	14.05	$13.43_{(1.03)}$
-0.4	0.5	3.13	$3.16_{(0.18)}$	3.08	$3.17_{(0.18)}$	11.98	$12.38_{(0.73)}$	11.95	$12.38_{(0.73)}$	12.20	$13.14_{(0.85)}$
0	0.5	2.99	$2.90_{(0.17)}$	2.99	$2.91_{(0.17)}$	12.61	$12.26_{(0.66)}$	12.63	$12.26_{(0.66)}$	15.00	$13.12_{(2.62)}$
0.4	0.5	3.15	$3.18_{(0.18)}$	3.11	$3.19_{(0.18)}$	11.83	$12.35_{(0.71)}$	11.81	$12.35_{(0.71)}$	12.66	$13.09_{(0.82)}$
0.8	0.5	3.79	$3.88_{(0.24)}$	3.72	$3.88_{(0.23)}$	12.69	$12.47_{(0.73)}$	12.63	$12.47_{(0.73)}$	15.05	$13.37_{(1.79)}$
0.5	-0.8	3.38	$3.31_{(0.19)}$	3.37	$3.32_{(0.19)}$	11.62	$12.43_{(0.71)}$	11.64	$12.42_{(0.71)}$	13.13	$14.30_{(0.76)}$
0.5	-0.4	3.43	$3.30_{(0.19)}$	3.36	$3.31_{(0.20)}$	12.58	$12.30_{(0.70)}$	12.57	$12.30_{(0.70)}$	13.64	$13.19_{(0.71)}$
0.5	0	3.35	$3.32_{(0.18)}$	3.33	$3.33_{(0.18)}$	12.52	$12.35_{(0.75)}$	12.52	$12.34_{(0.75)}$	13.82	$13.78_{(3.73)}$
0.5	0.4	3.39	$3.32_{(0.19)}$	3.37	$3.33_{(0.19)}$	12.26	$12.39_{(0.71)}$	12.26	$12.39_{(0.71)}$	13.10	$13.14_{(0.75)}$
0.5	0.8	3.29	$3.33_{(0.20)}$	3.26	$3.34_{(0.20)}$	12.10	$12.37_{(0.74)}$	12.17	$12.37_{(0.74)}$	12.86	$13.27_{(1.31)}$

of the new method are close to oracle, and are generally smaller than those of Zhou et al. (2020)'s method. This in turn intuitively illustrates the precision of proposed estimators.

Lastly, Table 3.3 reports the computing times, where the new method is nearly 1000 times faster than Zhou et al. (2020)'s method. The proposed method is very fast and stable because initialized by LASSO estimator, LLA algorithm converges in one step.

Example 2. In this example, we examine the finite sample performances of proposed method when heavy-tail errors are encountered. Specifically, assume now $\varepsilon_1 \sim t_6/\sqrt{6}$. The multiplier $\sqrt{6}$ ensures the equality of variance of ε_1 to that when

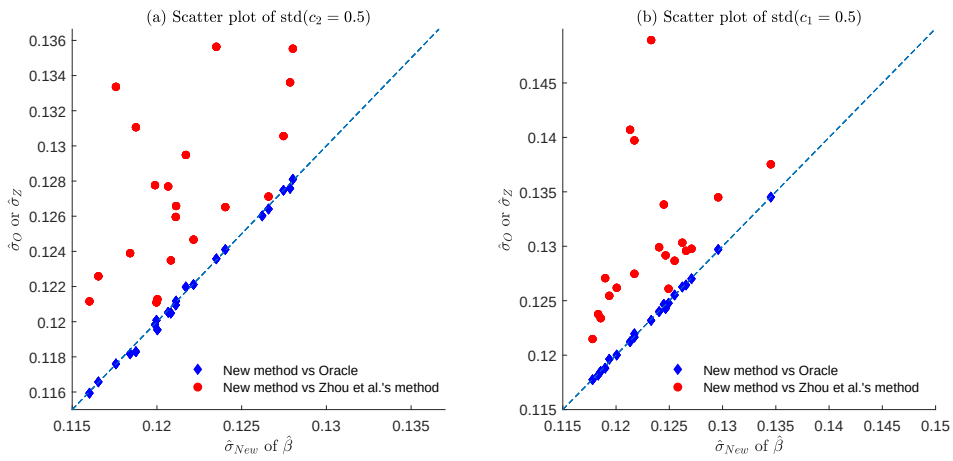

Figure 3.2. Scatter plot of standard deviations of $\widehat{\boldsymbol{\beta}}$ over 500 point estimates by the new method (x-axis) and by oracle or Zhou et al. (2020)'s method (y-axis). Each dot (blue and red) corresponds each of the 21 different simulation settings - when holding $c_2 = 0.5$, vary $c_1 = 0, \pm 0.1, \cdots, \pm 1$ in (a) and holding $c_1 = 0.5$, vary $c_2 = 0, \pm 0.1, \cdots, \pm 1$ in (b).

Table 3.3. Comparison results of the average computing time (in seconds) over 500 replications.

c_1	c_2	New method	Zhou et al. (2020)'s method
-0.8	0.5	1.38	1,207.88
-0.4	0.5	1.47	1,327.82
0	0.5	1.31	1,197.66
0.4	0.5	1.52	1,614.84
0.8	0.5	1.22	1,332.24
0.5	-0.8	1.35	1,192.32
0.5	-0.4	1.33	1,329.48
0.5	0	1.48	1,544.23
0.5	0.4	1.50	1,790.34

$\varepsilon_1 \sim N(0, 0.5^2)$. All other settings are identical to those in Example 1. We first investigate the performances of S_n, S_n^O and S_n^Z for testing indirect effect $\boldsymbol{\beta}$ via the left panel of Figure 3.3. The proposed test S_n performs as well as the oracle one S_n^O in terms of controlling Type-I error rate ($c_1 = 0$) and possessing much larger power than S_n^Z (when $c_1 \neq 0$), especially when $c_1 < 0$. Similar phenomenons are observed in the right penal of Figure 3.3 when examining T_n, T_n^O and T_n^Z. The proposed test T_n performs as well as the oracle one, and is more powerful than the test T_n^Z. In fact, when $c_2 = -0.2$, the empirical powers of our test statistic T_n and

the oracle test T_n^O are about 1, while that of T_n^Z is only about 0.756. In addition, we also evaluate the accuracy and precision of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ through Tables 3.4 and 3.5. The overall pattern in these two tables with $\varepsilon_1 \sim t_6/\sqrt{6}$ is very similar to that for $\varepsilon_1 \sim N(0, 0.5^2)$. In sum, the proposed method retains its validity for heavy-tailed error distributions.

Table 3.4. Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2 when $\varepsilon_1 \sim t_6/\sqrt{6}$. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		New method		Oracle		Zhou et al. (2020)'s method	
c_1	c_2	$\widehat{\boldsymbol{\alpha}}_1$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\alpha}}_1^O$	$\widehat{\boldsymbol{\beta}}^O$	$\widehat{\boldsymbol{\alpha}}_1^Z$	$\widehat{\boldsymbol{\beta}}^Z$
-0.8	0.5	$0.14_{(4.06)}$	$-0.30_{(12.46)}$	$0.22_{(3.93)}$	$-0.38_{(12.43)}$	$-13.93_{(6.09)}$	$13.50_{(12.84)}$
-0.4	0.5	$0.01_{(1.93)}$	$-0.14_{(6.24)}$	$0.06_{(1.89)}$	$-0.19_{(6.23)}$	$-3.34_{(2.81)}$	$3.23_{(6.43)}$
0	0.5	$0.16_{(3.03)}$	$-0.36_{(12.21)}$	$0.14_{(3.01)}$	$-0.34_{(12.21)}$	$-1.13_{(4.68)}$	$0.86_{(12.74)}$
0.4	0.5	$0.16_{(3.29)}$	$-0.36_{(12.30)}$	$0.09_{(3.26)}$	$-0.28_{(12.29)}$	$-0.77_{(5.19)}$	$0.52_{(13.01)}$
0.8	0.5	$0.28_{(3.07)}$	$-0.26_{(6.67)}$	$0.21_{(3.02)}$	$-0.18_{(6.63)}$	$0.75_{(4.06)}$	$-0.70_{(7.15)}$
0.5	-0.8	$0.19_{(3.44)}$	$-0.37_{(12.34)}$	$0.10_{(3.40)}$	$-0.28_{(12.33)}$	$6.50_{(5.61)}$	$-6.73_{(12.89)}$
0.5	-0.4	$0.16_{(3.45)}$	$-0.32_{(12.32)}$	$0.09_{(3.41)}$	$-0.25_{(12.30)}$	$5.92_{(12.67)}$	$-6.16_{(16.26)}$
0.5	0	$0.19_{(3.42)}$	$-0.34_{(12.34)}$	$0.09_{(3.41)}$	$-0.25_{(12.30)}$	$0.70_{(4.56)}$	$-0.95_{(12.95)}$
0.5	0.4	$0.20_{(3.44)}$	$-0.39_{(12.39)}$	$0.09_{(3.41)}$	$-0.28_{(12.33)}$	$-1.20_{(5.30)}$	$0.93_{(12.98)}$
0.5	0.8	$0.18_{(3.44)}$	$-0.34_{(12.32)}$	$0.09_{(3.41)}$	$-0.25_{(12.30)}$	$-1.17_{(5.29)}$	$0.96_{(13.07)}$

Table 3.5. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) of different methods with different c_1 and c_2 when $\varepsilon_1 \sim t_6/\sqrt{6}$. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		Direct effect ($\widehat{\alpha}_1$)				Indirect Effect ($\widehat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al. (2020)'s method	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
-0.8	0.5	4.06	$3.87_{(0.28)}$	3.93	$3.88_{(0.27)}$	12.46	$12.55_{(0.70)}$	12.43	$12.55_{(0.70)}$	12.84	$13.12_{(0.90)}$
-0.4	0.5	1.93	$1.94_{(0.14)}$	1.89	$1.95_{(0.14)}$	6.24	$6.29_{(0.33)}$	6.23	$6.29_{(0.33)}$	6.43	$6.47_{(0.36)}$
0	0.5	3.03	$2.91_{(0.21)}$	3.01	$2.92_{(0.21)}$	12.21	$12.29_{(0.72)}$	12.21	$12.29_{(0.72)}$	12.74	$12.80_{(0.77)}$
0.4	0.5	3.29	$3.17_{(0.23)}$	3.26	$3.18_{(0.23)}$	12.30	$12.35_{(0.72)}$	12.29	$12.35_{(0.72)}$	13.01	$12.95_{(0.82)}$
0.8	0.5	3.07	$2.92_{(0.22)}$	3.02	$2.93_{(0.22)}$	6.67	$6.66_{(0.30)}$	6.63	$6.66_{(0.30)}$	7.15	$6.57_{(0.64)}$
0.5	-0.8	3.44	$3.31_{(0.24)}$	3.40	$3.32_{(0.24)}$	12.34	$12.39_{(0.71)}$	12.33	$12.39_{(0.71)}$	12.89	$12.98_{(0.74)}$
0.5	-0.4	3.45	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.32	$12.39_{(0.71)}$	12.30	$12.39_{(0.71)}$	16.26	$13.01_{(0.87)}$
0.5	0	3.42	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.34	$12.39_{(0.71)}$	12.30	$12.39_{(0.71)}$	12.95	$12.96_{(0.74)}$
0.5	0.4	3.44	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.39	$12.39_{(0.71)}$	12.33	$12.39_{(0.71)}$	12.98	$12.98_{(0.81)}$
0.5	0.8	3.44	$3.31_{(0.24)}$	3.41	$3.32_{(0.24)}$	12.32	$12.39_{(0.71)}$	12.30	$12.39_{(0.71)}$	13.07	$12.99_{(0.86)}$

3.2.2 Real data analysis

We apply the proposed method to an empirical analysis to examine whether financial statements items and metrics mediate the relationship between company sec-

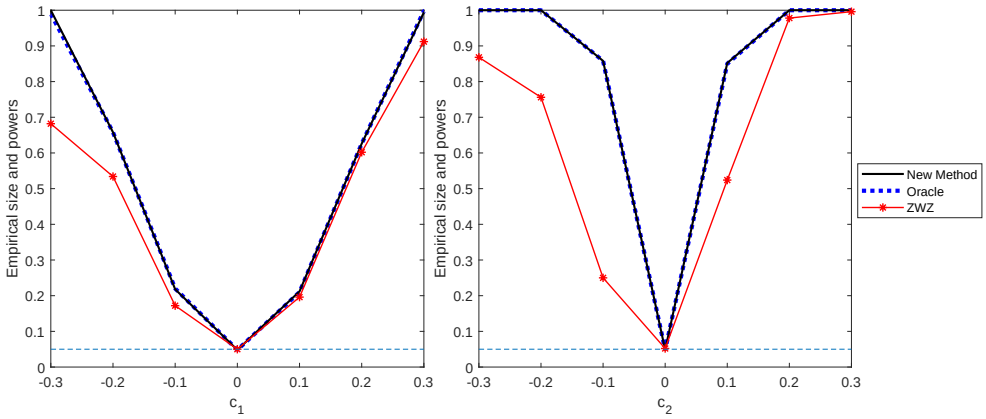

Figure 3.3. Left panel is empirical sizes and powers of S_n, S_n^Z and S_n^O when $\varepsilon_1 \sim t_6/\sqrt{6}$ at level $\alpha = 0.05$ over 500 replications for testing indirect effect when $\alpha_1 = 0.5$. Dotted line, solid line, and solid line marked by '*' represent the sizes and powers of S_n, S_n^O and S_n^Z, respectively. Right panel is empirical sizes and powers of T_n, T_n^Z and T_n^O for testing direct effect when $\beta = 0.7$. The dotted line, solid line, and solid line marked by '*' represent the sizes and powers of T_n, T_n^O and T_n^Z, respectively.

tors and stock price recovery after COVID-19 pandemic outbreak. While investors and researchers have reached a consensus ages ago that stock returns highly rely on companies' belonging sectors, recent studies more focus on using financial statements or market conditions to predict stock returns. Fama and French (1993)'s pioneering proposal of the three-factor model started this era, which captures patterns of return using market return, firm size and book-to-market ratio factors. Callen and Segal (2004) showed that accruals, cash flow, growth in operating income significantly influence stocks return. Edirisinghe and Zhang (2007) developed a relative financial strength metric based on data envelopment analysis (Farrell, 1957; Charnes et al., 1978), and found that return on assets and solvency ratio has high correlation with stock price return. To enhance prediction accuracy, deep neural network and data mining techniques were developed, with model inputs as historical financial statements and output as stock price return (Enke and Thawornwong, 2005; Huang et al., 2019; Lee et al., 2019). Meanwhile, it is reasonable to hypothesize that companies' sectors affect stock performances via influencing the associated financial metrics. Few existing works, however, study the mediating effects of such financial metrics. Hence our analysis aims to fill in this gap, and use the proposed mediation analysis to select important financial metrics, as well

as to test the direct and indirect effects of companies' sectors on returns.

In addition, we in this analysis are specifically interested in the stock performance of S&P 500 component companies during the COVID-19 pandemic period. As is known, the outbreak of the COVID-19 dealt a shock to the U.S. economy with unprecedented speed, and the government had to take a lockdown to stop spread of virus. The lockdown took a toll in the U.S. economy: business were closed, millions of people lost jobs and the price of an oil futures contract fell below zero. The crisis spread to the U.S. stock market, dragging down the major index S&P 500 by 33.92%. To help businesses, households and the economy, the Federal Reserve and the White House launched various rescue programs and take measures to stabilize energy prices from the end of March, 2020. Therefore, all these events and measures led the U.S. stock market to a V-shape pattern, thanks to which, the general financial rules from classical literature may not directly apply any more.

Admittedly, a number of recent literature studied the economic reaction to COVID-19 pandemic from sector or company level data (Ramelli and Wagner, 2020; Zhang et al., 2020; Baker et al., 2020; Gormsen and Koijen, 2020; De Vito and Gomez, 2020). Thorbecke (2020) analyzed sector-specific and macroeconomic variables as contributing factors to stock return in COVID-19 downturn and found that idiosyncratic factors negatively affected energy and consumer cyclical sectors. Hassan et al. (2020) investigated companies' transcripts of quarterly earnings call from January to September 2020 to investigate senior management's and major market participants' opinions about future prospects. They discovered several important factors related to accounting and business fundamentals, including supply chain, production and operations and financing, that are highly associated with stock market recovery from COVID-19. However, these methods mainly rely on prior financial knowledge to select low dimensional data for modeling, while ignore important company level factors. Besides, these methods only consider the relation of stock return to either sector level or company level while failing to recognize that the company's financial plays a role in mediating stock sector effects to stock price return. Therefore, we use the proposed method to study the financial statement items or metrics that mediate the relationship between firm sectors and stock performance in this special period. This work may then shed light on how

to select valuable stocks during a pandemic or any adverse event likewise.

In the mediation models, the response is taken to be the stock return from its highest price before the pandemic in February, 2020 to April 30th, 2020. The closed price is adjusted for both dividends and splits. The potential mediators in $\boldsymbol{m}$ are 550 accounting metrics from financial statements of associated companies, scratched from Yahoo Finance on April 30, 2020. We obtain firms' annual reports from fiscal year 2015 to 2019 and the first three quarterly reports in 2019. We use the firms' latest annual report to compute financial metrics and use previous annual reports to compute average growth rate of each financial metrics. The exposure variables in $\boldsymbol{x}$, are companies' sectors according to Global Industry Classification Standard (GICS) that are coded as dummy variables. GICS classifies companies into eleven sectors: basic materials, communication services, consumer cyclical, consumer defensive, energy, financial services, healthcare, industrials, real estate, technology and utilities. We set energy sector as baseline level.

Table 3.6 presents the estimated direct and indirect effects of companies' sectors, together with their standard errors. We also calculate Wald's test for the indirect effect and generalized likelihood test for direct effect, with p-values smaller than 10^{-9} and 10^{-15}, respectively, indicating both the direct and indirect effect are significant. As for direct effect, stocks in sectors such as healthcare and technology are more likely to outperform benchmark than ones from utilities sector. Furthermore, sectors influence the stocks performance partly through business operation reflected by selected financial metrics, and the indirect effects are significantly positive.

The selected mediating metrics, their associated estimated coefficients in model (3.1.1), as well as their brief descriptions, are presented in Table 3.7. These selected metrics are of their own significance. For instance, the first three chosen metrics in Table 3.7, namely return on assets, gross margin and annual growth rate of operating income, reflect firms' revenue. Return on assets is an indicator of how well a firm utilizes its assets, by determining how profitable a firm is relative to its total assets. A firm with a higher return-on-assets value is preferred, as the firm squeezes more out of limited resources to make a profit. Gross margin is the portion of sales revenue a firm retains after subtracting costs of producing the goods it sells and the services it provides. It measures the gross profit of a

Table 3.6. The estimated coefficients, standard errors, test statistics values and p-values for real data.

Sectors	Direct effect	std	Indirect effect	std
Intercept	-0.4634	0.1558	-0.5216	0.1016
Basic materials	0.5725	0.2226	0.4450	0.1407
Communication services	0.9231	0.2698	0.4227	0.1691
Consumer cyclical	0.0793	0.1805	0.4154	0.1165
Consumer defensive	0.9808	0.2087	0.6265	0.1386
Financial services	0.1363	0.1844	0.3452	0.1206
Healthcare	1.0176	0.1887	0.7601	0.1232
Industrials	0.3658	0.1816	0.5899	0.1181
Real.Estate	0.0736	0.2185	0.5010	0.1365
Technology	0.6537	0.1823	0.7655	0.1203
Utilities	0.6798	0.2121	0.3717	0.1343
p-value	$< 1 \times 10^{-9}$		$< 1 \times 10^{-15}$	

firm. A firm that has higher gross margin is more likely to retain more profit for every dollar of good sold. Annual growth rate of operating income shows the firm's growth of generating operating income compared with previous year. Operating income measures the amount of profit realized from a business's operation, after deducting operating expenses such as wages, depreciation, and cost of goods sold. A firm with high growth of operating income can avoids unnecessary production costs, and improve core business efficiency. In a word, a firm with higher return on assets, gross margin and growing operating income is considered profitable, and hence, is likely to attract investors.

On the other hand, both the average growth rate of quick ratio and debt to assets are indicators of financial leverage of a firm. Quick ratio of a firm is defined as the dollar amount of liquid assets dividing that of current liabilities, where liquid assets are the portion of assets that can be quickly converted into cash with minimal impact on the price received in open market, while current liabilities are a firm's debts or obligations to be paid to creditors within one year. Thus a large quick ratio indicates that the firm is fully equipped with enough assets to be instantly liquidated to pay off its current liabilities. Debt to assets is the total amount of debt relative to assets owned by a firm. It reflects a firm's financial stability. Therefore, a firm with a higher quick ratio or a lower debt to assets might be more likely to survive when it is difficult to finance through borrowing and cover

Table 3.7. Selected importance mediators and their coefficients

Selected mediator	Estimated coefficient (std)	Description
Return on assets	0.4246 (0.0379)	Net income divided by the total assets
Gross margin	0.0841 (0.0393)	The difference between the revenue and cost of goods sold divided by revenue
AGR* Operating Income	0.1063 (0.0347)	Revenue subtract cost of goods sold and operating expenses
AGR* Quick ratio	0.1194 (0.0345)	Total current assets minus inventory divided by total current liabilities
Debt to assets	-0.1209 (0.0369)	Total debts divided by total assets
Receivables turnover (days)	-0.0947 (0.0346)	Average receivables divided by net credit sales times 360 days

* AGR: average growth rate, calculated as the average of growth rates for the metrics from 2015 to 2019.

its debts, thus are more favorable to investors during the economy lockdown.

Lastly, receivables turnover quantifies a firm's effectiveness in collecting its receivables or money owed by clients. It shows how well a firm uses and manages the credit it extends to customers and how quickly that short-term debt is paid. Receivables turnover can be negative when net credit sale is negative because the client pre-pay for the product or service. A negative receivables turnover means that the firm are less susceptible to counter-party credit risk because it already receives the cash from its client before delivering the service or shipping out the product. This is especially important during liquidity dry periods when the clients may default or delay payment due to lack of cash. Therefore, a firm that has a negative receivables turnover is preferred.

On all accounts, one might incorporate the analysis results as reference when seeking for a stock portfolio during the financial crisis caused by pandemic. First, the sectors in 'Healthcare', 'Consumer defensive', 'Communication service', 'Utility' and 'Technology' have the top five positive direct effects on stock return. In terms of the financial metrics, we may focus on those reported in Table 3.7 to filter stocks. For example, we shall select firms that have higher values in AGR operating income, gross margin, quick ratio, and return on assets but lower values in debt to assets and receivable turnover.

Moreover, we compare our findings with those selected in established models. For instance, our method picks profitability factors like return on assets, which is also selected in Fama and French (2015), as profitability is the core of a firm's stock performance. But we do not include metrics representing size of firm, valuation of stock price or investment that were covered by Fama and French (2015). For firm size factor, there is no evidence that small-size firms recovered faster or slower than larger-size ones. For valuation of stock price factor, previous price valuation ratio changed significantly due to stock price change and is no longer reliable to predict future stock return. For investment factor, it is less important for a short-term stock price movement. Compared with Edirisinghe and Zhang (2007), our method also picks profitability (return on assets), liquidity (quick ratio) and solvency (debt to assets) metrics, as in Edirisinghe and Zhang (2007). During the crisis, a firm facing liquidity crunch could not access to credit. Therefore, a firm with sufficient cash and less debt is more easily to survive and less likely to be forced to liquidate valuable assets at unfavorable prices. And its stock would be safer and more attractive to investors. But we did not select metrics of earnings per share or about capital intensity as in Edirisinghe and Zhang (2007). The lockdown dramatically changes a firm's revenue structure and capital allocation, and hence reduces predictive capability of these metrics to short-term recovery.

3.3 Conclusion

In this paper, we propose statistical inference procedures for the indirect effects in high dimensional mediation model. We introduce a partial penalized least squares method and study its statistical properties under random design. We show that the proposed estimators are more efficient than existing ones. We further propose a partial penalized Wald test to detect the indirect effect, with a χ^2 limiting null distribution. In this paper, we also propose an F-type test for the direct effect and reveal Wilks phenomenon in the high-dimensional mediation model. We further utilize the proposed inference procedures to analyze the mediation effects of various financial metrics on the relationship between company's sector and the stock return.

3.4 Proofs of Theorem 1 and 2

Define

$$Q_n(\boldsymbol{\theta}) \;=\; \frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{M}\boldsymbol{\alpha}_0-\boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2+\sum_{j=1}^{p}p_\lambda(|\alpha_{0,j}|).$$

Proof of Theorem 1: To enhance the readability, we divide the proof of Theorem 1 into three steps. In the first step, we show that there exists a local minimizer $\bar{\boldsymbol{\theta}}$ of $Q_n(\boldsymbol{\theta})$ with the constraints $\bar{\boldsymbol{\alpha}}_{0,\mathcal{A}^c}=0$, such that $\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2=O_P(\sqrt{s/n})$. In the second step, we prove that $\bar{\boldsymbol{\theta}}$ is indeed a local minimizer of $Q_n(\boldsymbol{\theta})$. This implies $\widehat{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}$. In the final step, we derive the asymptotic expansion of $\widehat{\boldsymbol{\theta}}$.

Step 1: Consistency in the $(s+q)$-dimensional subspace: We first constrain $Q_n(\boldsymbol{\theta})$ on the $(s+q)$-dimensional subspace of $\{\boldsymbol{\theta}\in R^{p+q}:\boldsymbol{\alpha}_{0,\mathcal{A}^c}=0\}$. This constrained partial penalized least squares function is given by

$$\bar{Q}_n(\boldsymbol{\vartheta})=\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\delta}-\boldsymbol{X}\boldsymbol{\alpha}_1\|_2^2+\sum_{j=1}^{s}p_\lambda(|\delta_j|).$$

Here $\boldsymbol{\vartheta}=(\boldsymbol{\alpha}_1^T,\boldsymbol{\delta}^T)^T$ and $\boldsymbol{\delta}=(\delta_1,\cdots,\delta_s)^T$. We now show that there exists a strict local minimizer $\bar{\boldsymbol{\vartheta}}$ of $\bar{Q}_n(\boldsymbol{\vartheta})$ such that $\|\bar{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0\|_2=O_P(\sqrt{s/n})$. To this end, we consider an event

$$H_n=\{\min_{\boldsymbol{\vartheta}\in\partial\mathcal{N}_\tau}\bar{Q}_n(\boldsymbol{\vartheta})>\bar{Q}_n(\boldsymbol{\vartheta}_0)\}.$$

where $\mathcal{N}_\tau=\{\boldsymbol{\vartheta}\in R^{s+q}:\|\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0\|_2\le\tau\sqrt{s/n}\}$ with $\tau\in(0,\infty)$, and $\partial\mathcal{N}_\tau$ denotes the boundary of the closed set $\mathcal{N}_\tau$. Clearly, on the event H_n, there exists a local minimizer of $\bar{Q}_n(\boldsymbol{\vartheta})$ in $\mathcal{N}_\tau$. Thus, we only need to show that $P(H_n)\to 1$ as $n\to\infty$ when τ is large. To this aim, we next analyze the function $\bar{Q}_n$ on the boundary $\partial\mathcal{N}_\tau$.

For any $\boldsymbol{\vartheta}$, it follows from a second order Taylor's expansion that

$$\bar{Q}_n(\boldsymbol{\vartheta})-\bar{Q}_n(\boldsymbol{\vartheta}_0)=-(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0)^T\boldsymbol{\nu}+\frac{1}{2}(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0)^TD(\boldsymbol{\vartheta}-\boldsymbol{\vartheta}_0). \quad (3.4.1)$$

Here

$$\boldsymbol{\nu} = \begin{pmatrix} \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^{\star}_{0,\mathcal{A}} - \boldsymbol{X}\boldsymbol{\alpha}^{\star}_1) \\ \frac{1}{n}\boldsymbol{M}^T_{\mathcal{A}}(\boldsymbol{y} - \boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^{\star}_{0,\mathcal{A}} - \boldsymbol{X}\boldsymbol{\alpha}^{\star}_1) - \lambda_n\bar{\rho}(\boldsymbol{\alpha}^{\star}_{0,\mathcal{A}}) \end{pmatrix},$$

and

$$\begin{aligned} D &= \frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} \\ \boldsymbol{M}^T_{\mathcal{A}}\boldsymbol{X} & \boldsymbol{M}^T_{\mathcal{A}}\boldsymbol{M}_{\mathcal{A}} \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & \Lambda(\boldsymbol{\alpha}^*_{0,\mathcal{A}}) \end{pmatrix} \\ &=: D_1 + D_2. \end{aligned}$$

where $\boldsymbol{\alpha}^*_{0,\mathcal{A}}$ lies in the line segment jointing $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}^{\star}_{0,\mathcal{A}}$, and $\Lambda(\boldsymbol{\alpha}^*_{0,\mathcal{A}})$ is a diagonal matrix with nonnegative diagonal elements. Clearly $\boldsymbol{\alpha}^*_{0,\mathcal{A}} \in \mathcal{N}_0$. By condition (A2), the maximum eigenvalue of $\Lambda(\boldsymbol{\alpha}^*_{0,\mathcal{A}})$ is upper bounded by $\lambda_n\kappa_0$. Recall that

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XM} \\ \Sigma_{MX} & \Sigma_{MM} \end{pmatrix}.$$

Further note that

$$P(\|D_1 - \Sigma\|_2 \geq \eta) \leq \frac{1}{\eta^2}E[\|D_1 - \Sigma\|_2^2] \leq \frac{cn}{\eta^2 n^2}E[\sum_{i,j}^{s}[m_{1i}m_{1j} - E(m_{1i}m_{1j})]^2$$
$$+ \sum_{i=1}^{s}\sum_{j=1}^{q}[m_{1i}x_{1j} - E(m_{1i}x_{1j})]^2 + \sum_{i,j}^{q}[x_{1i}x_{1j} - E(x_{1i}x_{1j})]^2] = \frac{cs^2}{\eta^2 n}.$$

Thus $\|D_1 - \Sigma\|_2 = O_P(s/\sqrt{n}) = o_P(1)$, when $s = o(n^{1/2})$.
Since $\lambda_{\min}(\Sigma) \geq c$ and $\lambda_n\kappa_0 = o(1)$,

$$\lambda_{\min}(D) \geq \bar{c} > 0. \tag{3.4.2}$$

Consequently, we obtain

$$\begin{aligned} &\min_{\boldsymbol{\vartheta}\in\partial\mathcal{N}_\tau} \bar{Q}_n(\boldsymbol{\vartheta}) - \bar{Q}_n(\boldsymbol{\vartheta}_0) \geq \min_{\boldsymbol{\vartheta}\in\partial\mathcal{N}_\tau}\left(-\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2\|\boldsymbol{\nu}\|_2 + \frac{1}{2}\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2^2\bar{c}\right) \\ = \; & -\sqrt{\frac{s}{n}}\tau\|\boldsymbol{\nu}\|_2 + \frac{1}{2}\frac{s}{n}\tau^2\bar{c}. \end{aligned}$$

By the Markov inequality, it entails that

$$P(H_n) \geq P(\|\boldsymbol{\nu}\|_2 \leq \frac{1}{2}\sqrt{\frac{s}{n}}\tau\bar{c}) \geq 1 - \frac{4nE\|\boldsymbol{\nu}\|_2^2}{s\tau^2\bar{c}^2}. \tag{3.4.3}$$

In the following, we aim to show that $E\|\boldsymbol{\nu}\|_2^2 = O(s/n)$.

Note that

$$\boldsymbol{\nu} = \begin{pmatrix} \frac{1}{n}\boldsymbol{X}^T\epsilon_1 \\ \frac{1}{n}\boldsymbol{M}_{\mathcal{A}}^T\epsilon_1 \end{pmatrix} - \begin{pmatrix} 0 \\ \lambda_n\bar{\rho}(\boldsymbol{\alpha}_{0,\mathcal{A}}^\star) \end{pmatrix} = \boldsymbol{\nu}_1 - \boldsymbol{\nu}_2,$$

Then by condition (A1),

$$\begin{aligned} E\|\boldsymbol{\nu}_1\|_2^2 &= \frac{1}{n^2}\mathrm{tr}\left[E\begin{pmatrix} \boldsymbol{X}^T\epsilon_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\epsilon_1 \end{pmatrix}\begin{pmatrix} \boldsymbol{X}^T\epsilon_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\epsilon_1 \end{pmatrix}^T\right]. \\ &= \frac{\sigma_1^2}{n}\mathrm{tr}(\Sigma) \leq \sigma_1^2\frac{s+q}{n}\lambda_{\max}(\Sigma) = O(\frac{s}{n}). \end{aligned}$$

It follows from the concavity of $\rho(\cdot)$, $d_n < |\alpha_{0j,\mathcal{A}}|$, and condition (A2) that:

$$\|\boldsymbol{\nu}_2\|_2^2 \leq (s^{1/2}p_\lambda'(d_n))^2 = o(\frac{1}{n}).$$

Consequently, step 1 is completed.

Step 2: Sparsity: According to Theorem 1 in Fan and Lv (2011), it suffices to show that with probability tending to 1,

$$\frac{1}{n}\|\boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{y} - \boldsymbol{M}\bar{\boldsymbol{\alpha}}_0 - \boldsymbol{X}\bar{\boldsymbol{\alpha}}_1)\|_\infty \ll \lambda_n. \tag{3.4.4}$$

Here $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\alpha}}_1^T, \bar{\boldsymbol{\alpha}}_0^T)^T$ satisfies that $\bar{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ and $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_P(\sqrt{s/n})$. Note that

$$\boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{y} - \boldsymbol{M}\bar{\boldsymbol{\alpha}}_0 - \boldsymbol{X}\bar{\boldsymbol{\alpha}}_1) = \boldsymbol{M}_{\mathcal{A}^c}^T\epsilon_1 - \boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0). \tag{3.4.5}$$

For the second term,

$$\|\boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)\|_\infty \leq \|\boldsymbol{M}_{\mathcal{A}^c}^T(\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})\|_{2,\infty}\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 = O_P(\sqrt{ns}).$$

Next we come to determine the rate of the first term $\|\boldsymbol{M}_{\mathcal{A}^c}^T \epsilon_1\|_\infty$.
Let $a_n = n^{1/\varpi+\varsigma} K_n, b = \sqrt{Cn \log p}$ with C being large enough and note that

$$\begin{aligned} m_{ij}\varepsilon_{i1} &= m_{ij}\varepsilon_{i1} I(|m_{ij}\varepsilon_{i1}| \le a_n) - E[m_j\varepsilon_1 I(|m_j\varepsilon_1| \le a_n)] \\ &\quad + m_{ij}\varepsilon_{i1} I(|m_{ij}\varepsilon_{i1}| > a_n) - E[m_j\varepsilon_1 I(|m_j\varepsilon_1| > a_n)] \\ &=: \epsilon_{ij,1} + \epsilon_{ij,2}. \end{aligned}$$

$$\begin{aligned} & P\left(|\sum_{i=1}^n m_{ij}\varepsilon_{i1}| > b, \text{ for some } j \in \mathcal{A}^c \right) \\ & \le P\left(|\sum_{i=1}^n \epsilon_{ij,1}| + |\sum_{i=1}^n \epsilon_{ij,2}| > b, \text{ for some } j \in \mathcal{A}^c \right) \\ & \le P\left(|\sum_{i=1}^n \epsilon_{ij,1}| > b/2, \text{ for some } j \in \mathcal{A}^c \right) \\ & + P\left(|\sum_{i=1}^n \epsilon_{ij,2}| > b/2, \text{ for some } j \in \mathcal{A}^c \right) \\ & =: P_1 + P_2. \end{aligned}$$

Firstly consider the term P_1. Note that $\epsilon_{1j,1}, \ldots, \epsilon_{nj,1}$ are independent centered random variables a.s. bounded by $2a_n$ in absolute value. Then the Bernstein inequality yields that

$$\begin{aligned} P_1 &\le 2(p-s) \max_j \exp\left\{ -\frac{b^2/4}{2nE(\epsilon_{j,1}^2) + 2 \cdot 2a_n \cdot b/(2 \cdot 3)} \right\} \\ &\le 2p \max_j \exp\left\{ -\frac{C \log p/4}{2E(\epsilon_{j,1}^2) + 2a_n\sqrt{C \log p/n}/3} \right\} \to 0. \end{aligned}$$

Next we turn to consider P_2. First note that

$$\begin{aligned} P_2 \le P[&\sum_{i=1}^n \max_j |m_{ij}\varepsilon_{i1}| I(|m_{ij}\varepsilon_{i1}| > a_n) \\ &+ \max_j nE[|m_j\varepsilon_1| I(|m_j\varepsilon_1| > a_n)] > b/2] \end{aligned}$$

Further note that

$$E^2[|m_j\varepsilon_1|I(|m_j\varepsilon_1|>a_n)] \le E[m_j^2\varepsilon_1^2]P(|m_j\varepsilon_1|>a_n) \le E[m_j^2\varepsilon_1^2]\frac{E[|m_j\varepsilon_1|^{\varpi}]}{a_n^{\varpi}}.$$

We then conclude that

$$\max_j nE[|m_j\varepsilon_1|I(|m_j\varepsilon_1|>a_n)] \le \max_j n\sqrt{\frac{E[m_j^2\varepsilon_1^2]E[|m_j\varepsilon_1|^{\varpi}]}{a_n^{\varpi}}} = o(\sqrt{n}).$$

From this, we then have

$$\begin{aligned} P_2 &\le P\left(\sum_{i=1}^{n}\max_j|m_{ij}\varepsilon_{i1}|I(|m_{ij}\varepsilon_{i1}|>a_n)>b/4\right) \\ &\le P\left(\max_j|m_{ij}\varepsilon_{i1}|>a_n \ \text{ for some } i\right) \\ &\le n\frac{E[\max_j|m_j\varepsilon_1|^{\varpi}]}{a_n^{\varpi}} = o(1). \end{aligned}$$

Thus $\|\boldsymbol{M}_{\mathcal{A}^c}^T\epsilon_1\|_\infty = O_P(\sqrt{n\log p})$.

Consequently, given condition A2, step 2 is finished.

Step 3: Asymptotic expansions: Steps 1 and 2 show that $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ with probability tending to 1, and further $\|\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}^\star_{0,\mathcal{A}}\|_2 = O_P(\sqrt{s/n})$.

First denote

$$\dot{L}(\boldsymbol{\vartheta}_0) \;=\; \begin{pmatrix} \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}}-\boldsymbol{X}\boldsymbol{\alpha}^\star_1) \\ \boldsymbol{M}_{\mathcal{A}}^T(\boldsymbol{y}-\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}}-\boldsymbol{X}\boldsymbol{\alpha}^\star_1) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\epsilon_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\epsilon_1 \end{pmatrix}. \tag{3.4.6}$$

For $\widehat{\boldsymbol{\vartheta}}$, denote

$$\dot{L}(\widehat{\boldsymbol{\vartheta}}) \;=\; \begin{pmatrix} \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}-\boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1) \\ \boldsymbol{M}_{\mathcal{A}}^T(\boldsymbol{y}-\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}-\boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1) \end{pmatrix} = \begin{pmatrix} 0 \\ n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}) \end{pmatrix}. \tag{3.4.7}$$

Notice that

$$\dot{L}(\boldsymbol{\vartheta}_0) = \dot{L}(\widehat{\boldsymbol{\vartheta}}) + nD_1(\widehat{\boldsymbol{\vartheta}}-\boldsymbol{\vartheta}_0).$$

Or equivalently,

$$\frac{1}{\sqrt{n}}(\dot{L}(\boldsymbol{\vartheta}_0) - \dot{L}(\widehat{\boldsymbol{\vartheta}})) = \Sigma\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) + (D_1 - \Sigma)\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0).$$

Recall that $\|D_1 - \Sigma\|_2 = O_P(s/\sqrt{n})$, and $\|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\| = O_P(\sqrt{s/n})$. Then,

$$(D_1 - \Sigma)\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) = o_P(1),$$

when $s = o(n^{1/3})$. Thus,

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \;\; = \;\; \Sigma^{-1}\frac{1}{\sqrt{n}}(\dot{L}(\boldsymbol{\vartheta}_0) - \dot{L}(\widehat{\boldsymbol{\vartheta}})) + o_P(1).$$

Under condition (A2), we have $\|\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}^\star_{0,\mathcal{A}}\|_\infty = O_P(\sqrt{s/n}) \ll d_n$. This implies that

$$\min_{j\in\mathcal{A}} |\widehat{\alpha}_{0j,\mathcal{A}}| > \min_{j\in\mathcal{A}} |\alpha^\star_{0j,\mathcal{A}}| - d_n = d_n.$$

By the concavity of $p(\cdot)$ and condition (A2), we obtain that

$$\|n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}})\|_2 \le ns^{1/2}p'_{\lambda_n}(d_n) = o(n^{1/2}).$$

Since $\lambda_{\max}(\Sigma^{-1}) = O(1)$, it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \;\; = \;\; \Sigma^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + o_P(1). \tag{3.4.8}$$

Proof of Corollary 1: Recall that

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{XX}^{-1} + \Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1} & -\Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1} \\ -\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1} & \Sigma_{MM.X}^{-1} \end{pmatrix}.$$

Here $\Sigma_{MM.X} = \Sigma_{MM} - \Sigma_{MX}\Sigma_{XX}^{-1}\Sigma_{XM}$.

As a result, it follows that

$$\begin{aligned} &\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) = (I_{q\times q}, 0_{q\times s})\Sigma^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + o_P(1) \\ = \;\; &\frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\boldsymbol{X}^T\epsilon_1 + \frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}(\Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{X}^T - \boldsymbol{M}_{\mathcal{A}}^T)\epsilon_1 + o_P(1). \end{aligned} \tag{3.4.9}$$

The asymptotic variance matrix of $\widehat{\boldsymbol{\alpha}}_1$ is

$$\sigma_1^2(I_{q\times q}, 0_{q\times s})\Sigma^{-1}(I_{q\times q}, 0_{q\times s})^T = \sigma_1^2\left(\Sigma_{XX}^{-1} + \Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1}\right).$$

Recall that

$$\sqrt{n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^\star) = \frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\boldsymbol{X}^T(\epsilon_1 + \epsilon_2) + o_P(1). \tag{3.4.10}$$

Consequently we obtain that

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) &= \frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\boldsymbol{X}^T\epsilon_2 \\ &\quad + \frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\Sigma_{MX}\Sigma_{MM.X}^{-1}(\boldsymbol{M}_{\mathcal{A}}^T - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{X}^T)\epsilon_1 + o_P(1) \\ &= \frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\sum_{i=1}^n W_{1i} \\ &\quad + \frac{1}{\sqrt{n}}\Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\sum_{i=1}^n W_{2i} + o_P(1).\end{aligned} \tag{3.4.11}$$

Here $W_{1i} = \boldsymbol{x}_i\varepsilon_{2i}$ and $W_{2i} = (\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{x}_i)\varepsilon_{1i}$.

It is easy to show that $E[W_{1i}] = E[\boldsymbol{x}_i E(\varepsilon_{2i}|\boldsymbol{x}_i)] = 0$. Similarly, $E[W_{2i}] = E[(\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{x}_i)E(\varepsilon_{1i}|\boldsymbol{x}_i, \boldsymbol{m}_{i,\mathcal{A}})] = 0$.

Further we obtain that $\mathrm{var}(W_{1i}) = \sigma_2^2\Sigma_{XX}$, $\mathrm{var}(W_{2i}) = \sigma_1^2\Sigma_{MM.X}$, and

$$\begin{aligned}\mathrm{cov}(W_{1i}, W_{2i}) &= E[\boldsymbol{x}_i\varepsilon_{2i}(\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{x}_i)\varepsilon_{1i}] \\ &= E[\boldsymbol{x}_i\varepsilon_{2i}(\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{x}_i)E(\varepsilon_{1i}|\boldsymbol{x}_i, \boldsymbol{m}_{i,\mathcal{A}}, \varepsilon_{2i})] = 0.\end{aligned}$$

As a result, it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \to N(0, \sigma_2^2\Sigma_{XX}^{-1} + \sigma_1^2\Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1}). \tag{3.4.12}$$

Proof of Theorem 2: Similar to the arguments in the proof of Theorem 1, we can also show that $\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ with probability 1, and further $\|\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}^\star\|_2 = O_P(\sqrt{s/n})$.

Denote $\Delta\widehat{\boldsymbol{\vartheta}} = \widehat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}} = (\Delta\widehat{\boldsymbol{\vartheta}}_1, \Delta\widehat{\boldsymbol{\vartheta}}_2)$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \ \Sigma^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}.$$

It is noted that

$$\begin{aligned} \begin{pmatrix} 0 \\ n\lambda_n \bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}) \end{pmatrix} &= \dot{L}(\widehat{\boldsymbol{\vartheta}}) = \dot{L}(\tilde{\boldsymbol{\vartheta}}) - nD_1\Delta\widehat{\boldsymbol{\vartheta}} \\ &= \begin{pmatrix} L_1(\tilde{\boldsymbol{\vartheta}}) \\ n\lambda_n \bar{\rho}(\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}}) \end{pmatrix} - \Sigma n\Delta\widehat{\boldsymbol{\vartheta}} - (D_1 - \Sigma)n\Delta\widehat{\boldsymbol{\vartheta}}. \end{aligned} \quad (3.4.13)$$

Here $D_1 = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} \\ \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{X} & \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{M}_{\mathcal{A}} \end{pmatrix}/n.$

From the proof of Theorem 1, it is known that $\|n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}})\|_2 = o_P(n^{1/2})$ and similarly $\|n\lambda_n\bar{\rho}(\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}})\|_2 = o_P(n^{1/2})$. Further recall that $\|D_1 - \Sigma\|_2 = O_P(s/\sqrt{n})$ and $\|\Delta\widehat{\boldsymbol{\vartheta}}\|_2 = \|(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) - (\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)\|_2 = O_P(\sqrt{s/n})$. Thus under condition that $s = o(n^{1/3})$,

$$o_P(1) = \begin{pmatrix} \frac{1}{\sqrt{n}}L_1(\tilde{\boldsymbol{\vartheta}}) \\ 0 \end{pmatrix} - \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_1 \\ \sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_2 \end{pmatrix}, \quad (3.4.14)$$

from which

$$\sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_2 = -\Sigma_{22}^{-1}\Sigma_{21}\sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_1 + o_P(1), \text{ and} \sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_1 = \Sigma^{11}\frac{1}{\sqrt{n}}L_1(\tilde{\boldsymbol{\vartheta}}) + o_P(1). \quad (3.4.15)$$

Note that $\sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_1 = \sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) + \sqrt{n}\boldsymbol{h}_n = O_P(1)$ from Corollary 1. Thus, $\sqrt{n}\Delta\widehat{\boldsymbol{\vartheta}}_2 = O_P(1)$, which further implies that $\Delta\widehat{\boldsymbol{\vartheta}}_2^T n\lambda_n\bar{\rho}(\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}}) = o_P(1)$.

Now we are ready to investigate the asymptotic distribution of T_n. Under the event $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = \tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ and recalling equation (3.4.15), we can show that

$$\begin{aligned} \mathrm{RSS}_1 - \mathrm{RSS}_0 &= -2\Delta\widehat{\boldsymbol{\vartheta}}^T\dot{L}(\tilde{\boldsymbol{\vartheta}}) + \Delta\widehat{\boldsymbol{\vartheta}}^T nD_1\Delta\widehat{\boldsymbol{\vartheta}} \\ &= -n\Delta\widehat{\boldsymbol{\vartheta}}_1^T(\Sigma^{11})^{-1}\Delta\widehat{\boldsymbol{\vartheta}}_1 + o_P(1). \end{aligned} \quad (3.4.16)$$

Now denote $\Phi = (I_q, 0_{q\times s})\Sigma^{-1}(I_q, 0_{q\times s})^T$. It is easy to know that $\Phi = \Sigma^{11}$.

From the proof of Corollary 1, it is known that $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) \to N(0, \sigma_1^2\Phi)$. Thus we obtain that

$$\mathrm{RSS}_0 - \mathrm{RSS}_1 \;=\; \|\Phi^{-1/2}[\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star)] + \sqrt{n}\Phi^{-1/2}\boldsymbol{h_n}\|_2^2 + o_P(1). \quad (3.4.17)$$

On the other hand,

$$\begin{aligned}\frac{\mathrm{RSS}_1}{n-q} &= \frac{1}{n-q}\|\boldsymbol{y} - \boldsymbol{M}\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1\|_2^2 = \frac{1}{n-q}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0^\star - \boldsymbol{X}\boldsymbol{\alpha}_1^\star\|_2^2 \\ &\quad -2\frac{1}{n-q}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^T\dot{L}(\boldsymbol{\vartheta}_0) + \frac{1}{n-q}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)^T n D_1(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \\ &= I_1 - 2I_2 + I_3.\end{aligned}$$

It is easy to know that $I_1 \to \sigma_1^2$, while $I_3 \le \|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2^2\|D_1\|_2 = O_P(s/n) = o_P(1)$. Further note that

$$I_2 \le \|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2\|\dot{L}(\boldsymbol{\vartheta}_0)\|_2/(n-q) = O_P((1/n)\sqrt{s/n}\sqrt{ns}) = O_P(\frac{s}{n}) = o_P(1).$$

In sum, it follows that

$$\frac{\mathrm{RSS}_1}{n-q} = \sigma_1^2 + o_P(1). \quad (3.4.18)$$

As a result,

$$T_n = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1/(n-q)} \to \chi_q^2(n\boldsymbol{h}_{\boldsymbol{n}}^T\Phi^{-1}\boldsymbol{h_n}/\sigma_1^2).$$

Chapter 4

An Application and Case Study of High-dimensional Linear Mediation Models

Childhood trauma tends to influence cortisol stress reactivity through the mediating effects of DNA methylation (Carpenter et al., 2007; Vinkers et al., 2015; Heim et al., 2000). Early works mainly adopted single-layer linear models to study that childhood trauma effects on DNA methylation, and that DNA methylation effects on cortisol level separately (McGowan et al., 2009; Perroud et al., 2011; Edelman et al., 2012). These works neglected the fact that DNA methylation should act as mediators in the relation between childhood trauma and cortisol stress reactivity. In addition, all of their discoveries are based on epigenetic modifications of a single gene instead of variable selections across all possible genes, which are likely to incur estimation bias. To overcome such flaws, Houtepen et al. (2016) conducted a study to investigate the role of DNA methylation in cortisol stress reactivity and its relationship with childhood trauma. The study collected a data set consisting of 385,882 DNA methylation loci, cortisol stress reactivity, one-dimensional score on a childhood trauma questionnaire and several covariates for 85 healthy individuals. Of great scientific interest is to identify the active mediating loci out of the 385,882 ones. Houtepen et al. (2016) conducted 385,882 linear mediation analyses, in each of which one locus was considered, and identified three active mediating loci. More recently, van Kesteren and Oberski (2019) proposed a coordinate-wise

mediation filter (CMF) and applied it to the same data set. They identified five active mediating loci. Unfortunately, the three loci identified by Houtepen et al. (2016) are completely different from the five loci identified by van Kesteren and Oberski (2019), probably because both Houtepen et al. (2016) and van Kesteren and Oberski (2019) did not consider all loci jointly in their analyses. The high dimensional DNA methylation loci indeed necessitate new techniques for identifying active mediating loci and testing the direct and indirect effects of the early life traumatic stress on later cortisol alteration.

Motivated by the contradictory results in the aforementioned two scientific works, we, in this chapter, develop a new estimating and testing procedure, and apply it to the same data set as that analyzed by the two works. We identify three new loci that possess both reasonably neurobiological interpretations and statistically significant effects via our proposed tests. Based on our new procedure, we further confirm that the childhood trauma does not have significant direct effects on cortisol change - it only indirectly affects cortisol through DNA methylation, and the indirect effect is negative. We also conduct simulation studies to validate the proposed method, and to compare its performance with both oracle and a debiased procedure proposed by Zhou et al. (2020).

In section 4.1, we introduce the statistical formulation of the high dimensional mediation problem, including the mediation models with confounders involved, the estimation for direct and indirect effects, and the tests of significance of indirect and direct effects. The detailed analysis is presented and discussed in section 4.2. We also conduct a thorough simulation study to validate the finite sample performance of the proposed procedure in section 4.3. A brief summary and conclusion are provided in section 4.4.

4.1 Statistical formulation: high dimensional linear mediation models with confounders

In this section, we introduce the high dimensional mediation models with confounders involved, as the statistical formulation associating childhood trauma with cortisol stress reactivity via altering DNA methylation. Then we extend the partial

penalization technique in Chapter 3 to these models, for estimating and testing the direct and indirect traumatic effects.

Let y be the response variable, $\boldsymbol{m}$ consist of p-dimensional mediators, $\boldsymbol{x}$ consist of q-dimensional exposure variables, and $\boldsymbol{z}$ consist of d-dimensional confounding variables. In our study, y is designated as the cortisol stress reactivity, $\boldsymbol{x}$ is childhood trauma, and elements in $\boldsymbol{m}$ correspond to DNA methylation loci that potentially mediate relations between trauma and cortisol. Moreover, we take several clinical variables as confounders in $\boldsymbol{z}$, with detailed descriptions in Section 3. Consider linear mediation models

$$y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \boldsymbol{\alpha}_2^T \boldsymbol{z} + \varepsilon_1, \tag{4.1.1}$$

$$\boldsymbol{m} = \Gamma_1^T \boldsymbol{x} + \Gamma_2^T \boldsymbol{z} + \boldsymbol{\varepsilon}_2, \tag{4.1.2}$$

where ε_1 is a random error with $E\varepsilon_1 = 0$ and $\text{var}(\varepsilon_1) = \sigma_1^2$ and $\boldsymbol{\varepsilon}_2$ is a random error vector with $E(\boldsymbol{\varepsilon}_2) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\varepsilon}_2) = \boldsymbol{\Sigma}^*$. Assume that ε_1 is independent of $\boldsymbol{m}, \boldsymbol{x}$ and $\boldsymbol{z}$, and $\boldsymbol{\varepsilon}_2$ is independent of $\boldsymbol{x}$ and $\boldsymbol{z}$. Furthermore, assume that ε_1 and $\boldsymbol{\varepsilon}_2$ are independent.

Motivated by the real data analysis in Section 3, it is assumed throughout this paper that q and d have fixed and finite dimensions, while p is high dimensional. Plugging (4.1.2) into (4.1.1), we obtain

$$\begin{aligned} y &= (\boldsymbol{\alpha}_1 + \boldsymbol{\beta})^T \boldsymbol{x} + (\Gamma_2 \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_2)^T \boldsymbol{z} + (\boldsymbol{\alpha}_0^T \boldsymbol{\varepsilon}_2 + \varepsilon_1), \\ &\equiv \boldsymbol{\gamma}_x^T \boldsymbol{x} + \boldsymbol{\gamma}_z^T \boldsymbol{z} + \varepsilon_3, \end{aligned} \tag{4.1.3}$$

where $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta} = \Gamma_1 \boldsymbol{\alpha}_0$ are called the direct and indirect effect of exposure $\boldsymbol{x}$ in mediation literature, respectively, and $\boldsymbol{\gamma}_x = \boldsymbol{\alpha}_1 + \boldsymbol{\beta}$ is called the total effect of $\boldsymbol{x}$. Often of primary interest from mediation point of view is to estimate and test $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}$. And these two parameters possess their own interpretations as natural indirect effect and natural direct effect in causal inference.

4.1.1 Natural direct and natural indirect effects

We link the parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}$ with natural direct and natural indirect effects on a causal diagram. Let $y(x, m)$ denote the potential outcome that would have been

observed had $\boldsymbol{x}$ and $\boldsymbol{m}$ been set to x and m, respectively, and $\boldsymbol{m}(x)$ denote the potential mediator that would have been observed had $\boldsymbol{x}$ been set to x. Following Imai et al. (2010), VanderWeele and Vansteelandt (2014), and others, for $\boldsymbol{x} = x_1$ versus x_0, the natural direct effect is defined as

$$E[y(x_1, \boldsymbol{m}(x_0)) - y(x_0, \boldsymbol{m}(x_0))],$$

while the indirect effect is defined as

$$E[y(x_1, \boldsymbol{m}(x_1)) - y(x_1, \boldsymbol{m}(x_0))].$$

The total effect is then naturally defined as the sum of natural direct and indirect effects

$$E[y(x_1, \boldsymbol{m}(x_1)) - y(x_0, \boldsymbol{m}(x_0))].$$

Furthermore, the independence assumptions of random errors in the mediation models (4.1.1) and (4.1.2) ensure the following sequential ignorability conditions (Imai et al., 2010; VanderWeele and Vansteelandt, 2014; Huang, 2019; Zhou et al., 2020).

(A1) $\boldsymbol{x} \perp\!\!\!\perp y(x, m)|\boldsymbol{z}$: that is, no unmeasured confounders between the exposure and outcome.

(A2) $\boldsymbol{m} \perp\!\!\!\perp y(x, m)|(\boldsymbol{x}, \boldsymbol{z})$: no unmeasured confounders between the mediators and outcome.

(A3) $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{m}(x)|\boldsymbol{z}$: no unmeasured confounders between the exposure and mediator.

(A4) $\boldsymbol{m}(\tilde{x}) \perp\!\!\!\perp y(x, m)|\boldsymbol{z}$: no exposure-dependent confounders between the mediators and outcome, where $\tilde{x}$ is the realization of exposure at a different value from x.

Under these sequential ignorability conditions, VanderWeele and Vansteelandt (2014) showed that

$$E[y(x_1, \boldsymbol{m}(x_0)) - y(x_0, \boldsymbol{m}(x_0))] = \boldsymbol{\alpha}_1^T(x_1 - x_0);$$

$$E[y(x_1, \boldsymbol{m}(x_1)) - y(x_1, \boldsymbol{m}(x_0))] = \boldsymbol{\beta}^T(x_1 - x_0).$$

Thus $\boldsymbol{\alpha}_1$ can be interpreted as the average natural direct effect, and $\boldsymbol{\beta} = \Gamma_1\boldsymbol{\alpha}_0$ can be interpreted as the average natural indirect effect.

4.1.2 Partial penalized least squares estimate

In this section, we introduce the estimation procedure of the direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta}$ that can get around high dimensional matrix estimation. Suppose that $\{\boldsymbol{m}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}$, $i = 1, \cdots, n$ is a random sample from (4.1.1) and (4.1.2). Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$, $\boldsymbol{M} = (\boldsymbol{m}_1, \cdots, \boldsymbol{m}_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)^T$, and $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$.

Despite high dimensionality of $\boldsymbol{m}$, model (4.1.3) is indeed a fixed-dimensional model. Therefore, the coefficient of $\boldsymbol{x}$, or say the total effect $\boldsymbol{\gamma}_x = \boldsymbol{\alpha} + \boldsymbol{\beta}$, could be naturally estimated via the ordinary least squared estimator in model (4.1.3), i.e.,

$$\widehat{\boldsymbol{\gamma}}_x = (\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{y}, \tag{4.1.4}$$

where $\boldsymbol{I}_q$ is $q \times q$ dimensional identity matrix, and $\boldsymbol{O}_{q\times d}$ is a $q \times d$ zero matrix.

Another key observation is that $\boldsymbol{x}$ and $\boldsymbol{z}$ in model (4.1.1) are both fixed dimensional, thus we opt not to impose sparsity on $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. On the other hand, sparsity on $\boldsymbol{\alpha}_0$, the coefficient associated with the high dimensional mediator $\boldsymbol{m}$, could be naturally and reasonably assumed, as so in most existing high-dimensional literature. Therefore, following Chapter 3, we apply the partial penalized least squared method to fit model (4.1.1) by only penalizing $\boldsymbol{\alpha}_0$. That is, the objective function subjected to minimization is

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1 - \boldsymbol{Z}\boldsymbol{\alpha}_2\|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|), \tag{4.1.5}$$

where α_{0j} is the jth element in $\boldsymbol{\alpha}_0$, and $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. Throughout this paper, we will use the SCAD penalty (Fan and Li, 2001) with λ being selected by HBIC (Chapter 3). Denote the corresponding estimates to be $\widehat{\boldsymbol{\alpha}}_0, \widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\alpha}}_2$. Further note that $\boldsymbol{\beta} = \boldsymbol{\gamma}_x - \boldsymbol{\alpha}_1$. Then we can estimate the indirect effect $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\gamma}}_x - \widehat{\boldsymbol{\alpha}}_1$.

Penalizing $\boldsymbol{\alpha}_0$ gains efficiency when estimating the coefficients, and hence enhances power towards the subsequential tests. Meanwhile, not penalizing $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ renders their respective estimators unbiasedness, thus there is no need for conducting any of the debiased, desparsified or decorrelated procedures (Zhang and Zhang, 2014; Van de Geer et al., 2014; Ning and Liu, 2017; Zhou et al., 2020), which admittedly correct estimation biases brought by ordinary regularization methods yet sacrifice efficiency.

4.1.3 Test of direct effect and indirect effect

To develop tests for the direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta}$, we first derive the asymptotic distributions of their estimators $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$. Let $\boldsymbol{w} = (\boldsymbol{x}^T, \boldsymbol{z}^T)^T$, $\boldsymbol{\alpha}_w = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T)^T$, $\Gamma_w = (\Gamma_1^T, \Gamma_2^T)^T$ and $\boldsymbol{\beta}_w = \Gamma_w \boldsymbol{\alpha}_0$. Thus models (4.1.1) and (4.1.2) can be rewritten as

$$y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_w^T \boldsymbol{w} + \varepsilon_1, \text{ and } \boldsymbol{m} = \Gamma_w^T \boldsymbol{w} + \boldsymbol{\varepsilon}_2, \tag{4.1.6}$$

which coincide with the causal mediation models considered in Chapter 3. Thus, incorporating the results in Corollary 1 of Chapter 3, the asymptotic distribution of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ can be obtained in a similar fashion. Specifically, define $\mathcal{A} = \{j : \alpha_{0j} \neq 0\}$ to be the truly active mediator index set, $\boldsymbol{m}_{\mathcal{A}}$ to be the subvector of $\boldsymbol{m}$ corresponding to the true mediators. And $\boldsymbol{\Sigma}_{MM} = E(\boldsymbol{m}_{\mathcal{A}} \boldsymbol{m}_{\mathcal{A}}^T)$, $\boldsymbol{\Sigma}_{MW} = \boldsymbol{\Sigma}_{WM}^T = E(\boldsymbol{m}_{\mathcal{A}} \boldsymbol{w}^T)$, $\boldsymbol{\Sigma}_{WW} = E(\boldsymbol{w}\boldsymbol{w}^T)$. Then

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) \to N(0, \sigma_1^2 (\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})(\boldsymbol{\Sigma}_{WW}^{-1} + B_w)(\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})^T), \tag{4.1.7}$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, (\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})(\sigma_2^2 \boldsymbol{\Sigma}_{WW}^{-1} + \sigma_1^2 B_w)(\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})^T), \tag{4.1.8}$$

where $B_w = \boldsymbol{\Sigma}_{WW}^{-1} \boldsymbol{\Sigma}_{WM} (\boldsymbol{\Sigma}_{MM} - \boldsymbol{\Sigma}_{MW} \boldsymbol{\Sigma}_{WW}^{-1} \boldsymbol{\Sigma}_{WM})^{-1} \boldsymbol{\Sigma}_{MW} \boldsymbol{\Sigma}_{WW}^{-1}$, $\sigma_2^2 = \boldsymbol{\alpha}_0^T \boldsymbol{\Sigma}^* \boldsymbol{\alpha}_0$.

The asymptotic covariance matrices in (4.1.7) and (4.1.8) could be estimated in the same routine as Chapter 3, by replacing quantities related to $\boldsymbol{x}$ in their work by those related to $\boldsymbol{w}$ in this paper. These estimates lay the foundation of subsequential tests.

For testing the direct effect $\boldsymbol{\alpha}_1$ with hypotheses

$$H_{0\alpha}: \ \boldsymbol{\alpha}_1 = \mathbf{0}, \text{ versus } H_{1\alpha}: \ \boldsymbol{\alpha}_1 \neq \mathbf{0},$$

we modify the F-type lack-of-fit test proposed by Chapter 3 by incorporating confounding effects. In model (4.1.1), denote RSS_f to be the residual sum of squares (RSS) in the full model fitted by the partial penalized least squares method (4.1.5), and RSS_r to be the RSS in the reduced model with $\boldsymbol{x}$ deleted from (4.1.1), obtained by the same partial penalized regression yet with objective function

$$\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{Z}\boldsymbol{\alpha}_2\|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|). \tag{4.1.9}$$

The test statistic thereby is defined as

$$T = \frac{RSS_r - RSS_f}{RSS_f/(n-q-d)},$$

which follows $\chi^2(q)$, the chi-squared distribution with degrees of freedom q, under the null hypothesis. And it possesses local power for local alternatives which converge to the null at the rate of $n^{-1/2}$.

For testing the indirect effect $\boldsymbol{\beta}$ with hypotheses

$$H_{0\beta}: \ \boldsymbol{\beta} = \mathbf{0}, \text{ versus } H_{1\beta}: \ \boldsymbol{\beta} \neq \mathbf{0},$$

we construct the Wald test statistic with the estimated covariance matrices,

$$S = n\widehat{\boldsymbol{\beta}}^T\{(\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})(\widehat{\sigma}_2^2\widehat{\boldsymbol{\Sigma}}_{WW}^{-1} + \widehat{\sigma}_1^2\widehat{B}_w)(\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})^T\}^{-1}\widehat{\boldsymbol{\beta}},$$

where $\widehat{B}_w = \widehat{\boldsymbol{\Sigma}}_{WW}^{-1}\widehat{\boldsymbol{\Sigma}}_{WM}(\widehat{\boldsymbol{\Sigma}}_{MM} - \widehat{\boldsymbol{\Sigma}}_{MW}\widehat{\boldsymbol{\Sigma}}_{WW}^{-1}\widehat{\boldsymbol{\Sigma}}_{WM})^{-1}\widehat{\boldsymbol{\Sigma}}_{MW}\widehat{\boldsymbol{\Sigma}}_{WW}^{-1}$. The hat versions are the sample counterparts of the covariance matrices. The limiting null distribution of S is $\chi^2(q)$, and the statistic can also detect the local effects with root-n convergence rate, as discussed in Chapter 3.

4.2 A case study: exploration of mediating effects of DNA methylation between childhood trauma and cortisol stress reactivity

This section is devoted to an empirical analysis of the same data set as that in Houtepen et al. (2016) and van Kesteren and Oberski (2019), for studying how DNA methylation plays a role in the regulation of human stress reactivity. More specifically, Houtepen et al. (2016) aimed to provide an unbiased investigation of the role of DNA methylation in cortisol stress reactivity and its relationship with childhood trauma. The data can be downloaded from the following website: https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-77445, and the data set consists of 385,882 DNA methylation loci and various variables for 85 people.

Houtepen et al. (2016) performed a genome-wide DNA methylation analysis for cortisol stress reactivity in healthy individuals. Since the number of DNA methylation loci is much greater than the sample size, Houtepen et al. (2016) first ran 385,882 linear regression models - response being cortisol stress reactivity, predictors being each out of the 385,882 DNA methylation loci, respectively, and confounders being several clinical variables. They reported 22,425 loci with p-values less than 0.05, while no statistically significant loci at level 0.05 after adjustment for multiple testing. The authors then selected three loci that stood out in the p-value distribution of the genome-wide cortisol stress reactivity analysis. The three loci are cg27512205 (denoted by m_1), cg05608730 (m_2) and cg26179948 (m_3), based on which the authors further conducted a mediation analysis, and identified a locus on the KITLG gene (cg27512205) that is not only associated to cortisol stress reactivity, but also partly mediates the relationship between childhood trauma and cortisol stress reactivity. More importantly, they replicated the analysis using two independent samples from the whole blood and buccal (cross-tissue) DNA, respectively, and concluded that the KITLG locus is indeed a mediator.

More recently, van Kesteren and Oberski (2019) proposed a coordinate-wise mediation filter (CMF), which aims to improve the marginal screening method for linear mediation models with high-dimensional mediators. They further ap-

plied CMF for an empirical analysis of the same data set as Houtepen et al. (2016), and identified five loci as the mediators. The five loci are cg16657538 (m_4), cg25626453 (m_5), cg02309301 (m_6), cg13136721(m_7) and cg12500973(m_8), which are completely different from the three loci identified by Houtepen et al. (2016). This contradiction motivates us to conduct a further analysis using the new procedure for studying the mediating role of DNA methylation that relates childhood trauma and cortisol alteration.

4.2.1 Mediation analysis via the proposed procedures

In our analysis, the exposure variable ($\boldsymbol{x}$) is a one-dimensional score on a childhood trauma questionnaire, and the outcome y is the increased area under the curve (iAUC) in cortisol after a stress test. We consider 385,882 DNA methylation loci in the blood as potential mediators in $\boldsymbol{m}$. Following van Kesteren and Oberski (2019), we first carry out a screening step to retain the top 1000 potential mediators by ranking the absolute value of the product of two correlations - the correlation between $\boldsymbol{x}$ and each element of $\boldsymbol{m}$, and between y and each element of $\boldsymbol{m}$. This indeed is a marginal screening procedure based on Pearson correlation proposed by Fan and Lv (2008). They showed that for linear models, under some regularity conditions, the screening procedure possesses a sure screening property. We also include the eight loci identified by Houtepen et al. (2016) and van Kesteren and Oberski (2019) as domain knowledge and for comparison purpose. Furthermore, eight clinical variables are involved, including age (Z_1), sex (Z_2), B cell proportion (Z_3), CD4 T cell proportion (Z_4), CD8 T cell proportion (Z_5), Monocytes cell proportion (Z_6), Granulocytes cell proportion (Z_7) and Natural Killer cell proportion (Z_8), as confounding variables. Refer to Table 4.1 for detailed explanations of these confounders. This leads to the linear mediation models (4.1.1) and (4.1.2), where $\boldsymbol{x}$ (with dimension $q = 1$) and y are defined above; the confounder vector $\boldsymbol{z}$ is $\boldsymbol{z} = (Z_0, Z_1, \cdots, Z_8)^T$, with $Z_0 \equiv 1$ to include an intercept in the model.

We apply the developed procedure to analyze the data. In the partial penalized least squares approach, we first select the tuning parameter λ by HBIC, and $\widehat{\lambda} = 60.8163$. In addition to the eight loci $m_1, \ldots, m_8$, our proposed method selects three additional loci cg19230917(m_9), cg06422529(m_{10}) and cg03199124(m_{11}). The

Table 4.1. Details of Confounding Variables

Variables	Name	Details
Z_1	Age	age of the participant
Z_2	Sex	gender of the participant
Z_3	B cell proportion	centre of the adaptive humoral immune system
Z_4	CD4 T cell proportion	instigate and shape adaptive immune responses
Z_5	CD8 T cell proportion	immune defence against intracellular pathogens
Z_6	Monocytes cell proportion	a type of white blood cell that is amoeboid
Z_7	Granulocytes cell proportion	a type of white blood cell that has small granules
Z_8	Natural Killer cell proportion	a type of lymphocyte that kills virally infected cells

estimated coefficients α_j's along with their standard errors, t-values and p-values are listed in Table 4.2. The estimated coefficients γ_{ij}'s in Γ_1 and Γ_2, with their p-values, are listed in Tables 4.3. It can be seen from Table 4.2 that the p-values of the newly identified loci (i.e., m_9, m_{10} and m_{11}) are all less than 0.05, while the p-values of m_1, m_5 and m_6 are greater than 0.10. Some significant effects are positive, while others are negative. From Table 4.3, childhood trauma $\boldsymbol{x}$ has significant effects on all the eleven mediators $m_1, \cdots, m_{11}$ except for m_8 at level 0.05. To sum up Tables 4.2 and 4.3, and the newly identified loci m_9, m_{10} and m_{11} should be considered as mediators since their coefficients are significant, and the coefficients of exposure variable on these loci are also significant at level 0.05.

Table 4.4 presents the results for testing the direct and indirect effects and compares the results with the Zhou et al. (2020)'s method (using R package 'freebird'). The test statistics in Table 4.4 are t-statistic. Under significance level 0.05, our new method's indirect effect is significant with p-value 0.0016, while the direct effect is insignificant since the p-value is 0.7643. Zhou et al. (2020)'s method also agrees with the significance of indirect effect and insignificance of direct effect under significance level 0.05, but their tests' p-value are larger than our new method's. If we choose the significance level 0.01, our indirect effect will still be significant but Zhou et al. (2020)'s indirect effect will be insignificant. The 'freebird' package only

Table 4.2. Estimated Coefficients, SE, t-values and p-values

Locus or Variables	Coefficients	SE	t-value	p-value
cg27512205(m_1)	-237.547	199.506	-1.191	0.238178
cg05608730(m_2)	-301.168	151.038	-1.994	0.050418
cg26179948(m_3)	-474.486	160.042	-2.965	0.004252
cg02309301(m_4)	259.730	108.633	2.391	0.019759
cg12500973(m_5)	30.029	116.354	0.258	0.797173
cg16657538(m_6)	84.330	53.236	1.584	0.118104
cg25626453(m_7)	369.183	97.988	3.768	0.000361
cg13136721(m_8)	260.990	65.585	3.979	0.000179
cg19230917(m_9)	321.196	149.918	2.142	0.035965
cg06422529(m_{10})	418.173	107.252	3.899	0.000234
cg03199124(m_{11})	471.865	143.943	3.278	0.001691
$\boldsymbol{x}$	1.365	4.553	0.300	0.765240
Z_0	-3110.834	3805.517	-0.817	0.416702
Z_1	-1.864	2.056	-0.906	0.368135
Z_2	349.037	82.811	4.215	0.000080
Z_3	1843.451	3702.100	0.498	0.620228
Z_4	406.642	3533.801	0.115	0.908748
Z_5	781.938	3368.749	0.232	0.817189
Z_6	967.962	3745.283	0.258	0.796890
Z_7	123.714	3544.597	0.035	0.972266
Z_8	341.974	3401.318	0.101	0.920229

reports estimated coefficients and p-values but we extracted their standard error of estimated coefficients and the t-statistic based on their codes, which are shown in Table 4.4. Besides, the 'freebird' package does not report their selected important mediators, so we cannot make comparisons on the selected important mediators between the two methods. Further note that the estimate of the indirect effect equals $-17.3726 < 0$. This implies that childhood trauma influences the cortisol stress reactivity only through the mediation mechanism of the DNA methylation, and the indirect effect is significantly negative.

Table 4.5 lists the 11 DNA methylation loci together with the genes to which they belong. It also provides some field knowledge of these genes dug out from existing research, according to which, the newly identified genes m_9, m_{10} and m_{11} have particularly interesting biological and genetical interpretations. The locus m_9

Table 4.3. Estimated Coefficients $\widehat{\gamma}_{ij}$ and their p-values

	$\boldsymbol{x}$	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
	Estimated γ_{ij}									
m_1	0.005	-2.999	0.000	0.016	0.870	0.197	-0.311	0.784	0.406	0.010
m_2	0.007	-1.723	-0.002	0.013	1.479	0.032	0.693	0.602	0.935	1.448
m_3	0.006	-3.566	-0.001	0.004	0.985	0.704	-0.809	0.840	0.567	-0.373
m_4	-0.012	-9.222	0.007	-0.115	5.179	3.716	4.666	5.653	4.371	4.901
m_5	-0.011	-3.974	0.004	-0.009	2.051	-2.022	-0.485	0.437	-0.860	-0.594
m_6	0.023	0.322	0.003	0.403	-6.719	-5.542	0.115	-1.608	-5.701	-4.943
m_7	-0.012	5.701	0.002	0.000	0.562	-0.290	0.552	-0.510	-0.285	-0.516
m_8	0.012	1.093	0.001	0.083	-2.086	3.799	1.748	3.922	2.673	1.312
m_9	-0.006	-2.009	0.001	0.018	1.257	-1.623	-1.448	-0.942	-1.181	-1.276
m_{10}	-0.008	8.459	0.003	0.042	-7.930	-4.387	-2.036	-3.341	-4.408	-4.387
m_{11}	-0.006	-5.736	0.003	-0.080	4.745	2.151	1.013	1.080	3.009	2.688
	p-value of Estimated Coefficients $\widehat{\gamma}_{ij}$									
m_1	0.044	0.225	0.711	0.766	0.726	0.936	0.895	0.765	0.868	0.997
m_2	0.016	0.568	0.145	0.848	0.627	0.992	0.811	0.851	0.755	0.613
m_3	0.020	0.201	0.448	0.946	0.724	0.798	0.761	0.776	0.837	0.887
m_4	0.004	0.025	0.001	0.202	0.205	0.356	0.230	0.191	0.277	0.203
m_5	0.002	0.286	0.054	0.917	0.584	0.584	0.892	0.912	0.816	0.866
m_6	0.004	0.968	0.490	0.026	0.406	0.487	0.988	0.850	0.474	0.515
m_7	0.006	0.205	0.444	0.998	0.901	0.948	0.898	0.915	0.949	0.903
m_8	0.056	0.865	0.861	0.563	0.748	0.554	0.777	0.568	0.676	0.830
m_9	0.041	0.528	0.551	0.795	0.695	0.608	0.635	0.781	0.709	0.672
m_{10}	0.048	0.044	0.223	0.646	0.060	0.288	0.608	0.449	0.286	0.266
m_{11}	0.045	0.068	0.108	0.253	0.133	0.488	0.734	0.744	0.332	0.364

Table 4.4. The estimated coefficients, standard errors, t-statistic values and p-values.

Method	Coefficient	Estimated Coefficient	SE	t-statistic	p-value
New method	α_1	1.3653	4.5529	0.2998	0.7643
	β	-17.3726	5.4945	-3.161	0.0016
Zhou et al.(2020)	α_1	-2.5551	3.9293	-0.6475	0.5172
	β	-11.4304	5.5653	-2.0538	0.0399

corresponds to the RAB5IF gene (cg19230917). This gene modulates cell endocytosis process by which cells engulf substances, such as hormones, from outside into the cell (Ravikumar et al., 2008). Cortisol is a steroid hormone produced by the adrenal glands, and it may signal the cells through receptor for endocytosis. Thus the RAB5IF gene likely plays a mediator rule that transmits the epigenetic alter-

Table 4.5. Annotation of the Included Mediators

Locus	Gene	Field knowledge from literature
cg27512205(m_1)	KITLG	Associated with germ cell and neural cell development.
cg05608730(m_2)	C1QTNF2	Involved in regulation of insulin action, sugar and fat metabolisms (Lei and Wong, 2019)
cg26179948(m_3)	JAZF1	Involved in regulation of glucose and lipid homeostasis (Liao et al., 2019)
cg02309301(m_4)	ARGLU1	Associated with sexual development
cg12500973(m_5)	HNRNPF	Involved in regulation of mRNA
cg16657538(m_6)	ZSCAN30	Involved in transcriptional regulation
cg25626453(m_7)	PRRC2A	Associated with the age-at-onset of diabetes
cg13136721(m_8)	RPTOR	Involved in regulation of cell growth and survival
cg19230917(m_9)	RAB5IF	Involved in endocytosis and macroautophagy (Ravikumar et al., 2008)
cg06422529(m_{10})	CPQ	Involved in thyroid development and tumours (Wojtas et al., 2017)
cg03199124(m_{11})	AGPAT1	Involved in signal transduction and lipid biosynthesis (Aguado and Campbell, 1998)

ations evoked by the traumatic stress. m_{10} belongs to the CPQ gene (cg06422529), which is shown by Wojtas et al. (2017) to function in thyroid and tumour development. Peter (2011) testified this gene as a pathway from stress to cortisol level change in fish. A further neurobiological exploration is worthwhile to find out whether it has similar mediating effect in human body. m_{11} is located in AGPAT1 gene (cg03199124), which is involved in signal transduction and lipid biosynthesis for creating and storing body fat (Aguado and Campbell, 1998). Some existing literature (Aschbacher et al., 2013; Gonzalez-Bono et al., 2002; Kuo et al., 2007) investigated the associations between physical stress like trauma and fat tissue biosynthesis. Vicennati et al. (2009) conducted a retrospective study and showed that women weight gain caused by trauma stress is accompanied by abnormal hormonal level such as cortisol. Our study which finds gene AGPAT1 as a mediator relating trauma stress and cortisol level therefore may provide clues for such stress pathophysiological mechanism research. In summary, there is a reasonable

conjecture that the identified loci, or their located genes, regulate neurobiological pathways and mediate the cortisol stress reactivity in response to childhood trauma, as also indicated by Table 4.4.

4.2.2 Some comparisons

It is worth to compare our results with those in Houtepen et al. (2016) and van Kesteren and Oberski (2019) from statistical point of view. Define $\boldsymbol{m}_{(1)} = (m_1, m_2, m_3)^T$, $\boldsymbol{m}_{(2)} = (m_4, \cdots, m_8)^T$ and $\boldsymbol{m}_{(3)} = (m_1, \cdots, m_8)^T$. For the purpose of comparison, we consider three linear mediation models by replacing $\boldsymbol{m}$ in (4.1.1) and (4.1.2) with $\boldsymbol{m}_{(k)}$, $k = 1, 2$ and 3. The mediation models considered in Houtepen et al. (2016) coincide with (4.1.1) and (4.1.2) where $\boldsymbol{m}$ is taken to be $\boldsymbol{m}_{(1)}$, and models in van Kesteren and Oberski (2019) correspond to those with $\boldsymbol{m}_{(2)}$. We further consider the mediation models with $\boldsymbol{m}_{(3)}$, which merges $\boldsymbol{m}_{(1)}$ and $\boldsymbol{m}_{(2)}$. The estimated regression coefficients α_j's in model (4.1.1) are listed in Table 4.6. The estimated γ_{ij}'s and their values coincide with those in Table 4.3 because regressing the multiple responses $\boldsymbol{m}$ over the exposure variable and confounding variables in linear model (4.1.2) coincides with regressing individual mediator m_j over the exposure variable and confounding variables.

Tables 4.2 and 4.6 both suggest that the direct effect of childhood trauma, or say the coefficient of the exposure variable $\boldsymbol{x}$, is not significant in model (4.1.1). All confounding variables except for Z_2 (i.e., sex) are not significant. Mediators m_5 and m_6 are not significant based on all belonging models under investigation. Furthermore, comparing Tables 4.2 and 4.6, we observe that the effect of mediator m_1 may change from significance to insignificance at level 0.05, after inclusion of other mediators into the model.

The estimated direct and indirect effects for these three models are presented in Table 4.7, together with corresponding significance tests. This table indicates that the direct effect is not significant and indirect effect is significant for models with mediators $\boldsymbol{m}_{(1)}$ and $\boldsymbol{m}_{(3)}$, while both direct and indirect effects are not significant for model with mediator $\boldsymbol{m}_{(2)}$.

Table 4.6. Estimated α_j's and their SE and p-values

	$\boldsymbol{m}_{(1)}$		$\boldsymbol{m}_{(2)}$		$\boldsymbol{m}_{(3)}$	
	Estimate(SE)	p-value	Estimate(SE)	p-value	Estimate(SE)	p-value
m_1	-590.5(251.4)	0.022			-296.4(245.9)	0.232
m_2	-576.3(193.8)	0.004			-473.6(187.5)	0.013
m_3	-560.3(217.6)	0.012			-535.6(202.2)	0.010
m_4			283.2(152.2)	0.067	223.4(135.3)	0.103
m_5			248.4(162.9)	0.132	156.0(144.3)	0.283
m_6			98.66(75.04)	0.193	44.43(67.58)	0.513
m_7			395.8(132.2)	0.004	321.5(121.2)	0.010
m_8			280.0(92.06)	0.003	198.8(83.17)	0.020
X	-5.487(4.916)	0.268	-10.71(6.310)	0.094	-2.933(5.791)	0.614
Z_0	-4861(4853)	0.320	905.0(5260)	0.864	-3098(4698)	0.512
Z_1	3.281(2.553)	0.203	0.8145(2.880)	0.778	0.3883(2.592)	0.881
Z_2	370.3(106.6)	0.001	322.7(118.8)	0.008	356.8(104.7)	0.001
Z_3	2471(4826)	0.610	-397.5(5186)	0.939	1096(4563)	0.811
Z_4	526.3(4755)	0.912	-955.4(5090)	0.852	-483.1(4469)	0.914
Z_5	1822(4584)	0.692	139.6(4880)	0.977	365.4(4292)	0.932
Z_6	2473(5091)	0.629	-1254(5436)	0.818	284.8(4785)	0.953
Z_7	991.9(4752)	0.835	-1202(5081)	0.813	-266.5(4465)	0.953
Z_8	726.2(4544)	0.873	-821.0(4860)	0.866	-294.0(4278)	0.945

4.2.3 Relationship among the mediators

The reversal of test results for mediator m_1, as well as the insignificance of m_5 and m_6, motivates us to explore the relationship among all the identified mediators. Table 4.8 exhibits the correlation structure among $m_1, \ldots, m_{11}$. The upper triangle of the table lists their pairwise Pearson correlations, and the lower triangle provides the corresponding p-values for testing the pairwise correlations. The table indeed detects several significantly correlated mediator pairs, such as m_1 and m_3, m_4 and m_5, m_5 and m_{11}, among others. Partial correlations among mediators given $\boldsymbol{x}$ and $\boldsymbol{z}$ are also presented in Table 4.9, together with their p-values. The table reflects similar phenomenon as Table 4.8. Furthermore, we fit several regression models, and extract their associated R^2, F-statistics and p-value of F-statistics in Table 4.10. It can be seen that the variation of m_1 can be partially explained by $m_4, \cdots, m_8$. This may interpret why the effect of m_1 becomes insignificant when

Table 4.7. The estimated coefficients, standard errors, test statistics values and p-values.

Model	Coefficient	Estimated Coefficient	SE	Test statistics	p-value
$\boldsymbol{m}_{(1)}$	α_1	-5.487	4.916	1.246	0.2644
	β	-10.52	3.612	8.447	0.0037
$\boldsymbol{m}_{(2)}$	α_1	-10.71	6.310	2.882	0.0896
	β	-5.2946	4.913	1.161	0.2811
$\boldsymbol{m}_{(3)}$	α_1	-2.933	5.791	0.2565	0.6125
	β	-13.07	5.356	5.959	0.0146

Table 4.8. Sample Pearson Correlation $\widehat{\rho}(m_j, m_k)$ and its p-values for $H_0 : \rho(m_j, m_k) = 0$

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}
						p-values					
m_1	$--$	.0023	.00003	.1236	.1245	.3121	.0067	.5546	.0705	.4381	.9664
m_2	.3251	$--$	.0310	.0477	.0975	.4100	.7073	.0831	.3186	.2841	.0388
m_3	.4327	.2339	$--$	.2500	.1163	.9144	.0662	.9424	.5203	.0332	.9829
m_4	-.1684	-.2152	-.1263	$--$	.0029	.3184	.7900	.8896	.0600	.7308	.3932
m_5	-.1680	-.1810	-.1717	.3171	$--$	.9710	.4336	.6590	.1132	.3724	.0054
m_6	-.1112	-.0907	-.0119	.1097	-.0040	$--$	.6043	.6849	.2166	.7042	.1541
m_7	-.2906	-.0414	-.2001	-.0294	.0862	-.0572	$--$	.6226	.4898	.1591	.6431
m_8	-.0652	-.1891	-.0080	-.0153	.0487	.0448	-.0543	$--$	.3715	.8036	.3788
m_9	-.1971	-.1097	-.0709	.2047	.1732	-.1356	-.0761	-.0984	$--$	.0132	.0045
m_{10}	-.0854	-.1178	-.2310	.0380	.0982	-.0419	.1543	-.0275	.2671	$--$	.9098
m_{11}	-.0046	-.2243	-.0024	.0940	.2978	-.1561	-.0511	-.0969	.3036	-.0125	$--$
	Sample Pearson Correlations										

$m_4, \cdots, m_8$ are included in the model. Similarly, the variation of m_4 and m_5 can be explained by other mediators in $\boldsymbol{m}_{(2)}$. This interprets their insignificant effects in models based on $\boldsymbol{m}_{(2)}$ and $\boldsymbol{m}_{(3)}$.

4.3 Simulation studies

We in this section conduct Monte Carlo simulation studies to investigate the finite sample performances of the statistical procedure described in Section 2, and compare it with the oracle tests that know the true mediator set $\mathcal{A}$, with statistics S^O and T^O, and those in Zhou et al. (2020), with statistics denoted by S^Z and T^Z. The results are based on 500 replications and significance level 0.05.

Table 4.9. Sample Partial Correlation $\widehat{\rho}(m_j, m_k|X, \boldsymbol{z})$ and its p-values for H_0 : $\rho(m_j, m_k|X, \boldsymbol{z}) = 0$

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}
						p-values					
m_1	$--$	.0082	.0007	.1833	.1860	.0687	.0359	.2038	.0942	.8357	.9873
m_2	.3030	$--$	.0670	.1639	.1423	.0994	.7079	.0470	.4010	.6658	.0556
m_3	.3835	.2126	$--$	.9051	.3575	.6418	.2421	.4120	.6685	.1508	.8175
m_4	-.1553	-.1624	-.0140	$--$	.0918	.0710	.1922	.7734	.1266	.5887	.7082
m_5	-.1544	-.1711	-.1077	.1960	$--$	.3815	.8615	.1406	.6583	.6323	.0580
m_6	-.2114	-.1917	-.0542	.2083	.1018	$--$	.7336	.6982	.5894	.8178	.8066
m_7	-.2427	.0437	-.1358	-.1512	-.0204	.0397	$--$	.8690	.1495	.3479	.2335
m_8	-.1484	-.2286	-.0955	.0336	.1706	-.0452	.0192	$--$	.8419	.7777	.6490
m_9	-.1947	-.0977	-.0499	.1768	.0516	-.0629	-.1670	-.0233	$--$	.0042	.0280
m_{10}	-.0243	-.0503	-.1664	-.0630	.0558	-.0269	.1092	-.0329	.3244	$--$	.9810
m_{11}	.0019	-.2219	.0271	-.0439	.2199	-.0288	-.1392	-.0534	.2539	.0028	$--$
	Sample Partial Correlations										

Table 4.10. The R^2, F statistics values and p-values of regression models between mediators to investigate multi-collinearity between mediators.

Dependent variable	Independent variable	R^2	F-stat	p-value
m_1	m_4,m_5,m_6,m_7,m_8	0.142	2.620	0.030
m_1	m_9,m_{10},m_{11}	0.043	1.213	0.310
m_4	m_5,m_6,m_7,m_8	0.117	2.650	0.039
m_5	m_4,m_6,m_7,m_8	0.115	2.587	0.043
m_6	m_4,m_5,m_7,m_8	0.018	0.374	0.827

We set up the experiments to mimic the real data analyzed in Section 3 to the utmost. The sample size is taken to be the same, the dimension of potential mediators is 1000, corresponding to the 1000 candidate DNA methylation loci, and the exposure variable $\boldsymbol{x}$ and confounder $\boldsymbol{z}$ are directly extracted from the data set. Meanwhile, $\boldsymbol{m}$ is generated via model (4.1.2), since it needs to be considered as random according to the mechanism of mediation models. Then y is accordingly generated from model (4.1.1). To accomplish this, we first draw Gaussian noise $\varepsilon_1 \sim N(0, \widehat{\sigma}_1^2)$ in model (4.1.1), where $\widehat{\sigma}_1^2$ is the estimated σ_1^2 in Section 3. The multivariate noise in model (4.1.2) is generated from $\boldsymbol{\varepsilon}_2 \sim N(0, \Sigma^*)$, where Σ^* is taken to be autoregressive covariance matrix. That is, the (i, j)-element of Σ^* equals $\rho^{|i-j|}$, and $\rho = 0.5$. The true mediators in $\mathcal{A}$ are designed in accordance with the 11 detected loci, $m_1, \ldots, m_{11}$, from the real data. Their associated coefficients $\boldsymbol{\alpha}_0$

in model (4.1.1) is taken to be $(1.0, 0.9, 0.8, -0.9, -0.8, -0.7, 0.6, 0.5, 0.4, 0.3, 0.2)$, with signs of elements consistent with those in $\widehat{\boldsymbol{\alpha}}_0$ estimated in Section 3. Moreover, the direct effect $\boldsymbol{\alpha}_1 = c_2$, where $c_2 = 0, 0.1, \cdots, 1.0$, to capture the size and power curve of the test for direct effect. For generating the indirect effect, the true value of Γ_1 is set to be $\Gamma_1 = c_1\widehat{\Gamma}_1$, where $c_1 = 0, \pm 0.1, \cdots, \pm 1.0$ and $\widehat{\Gamma}_1$ is the respective estimate from Section 3, thus the indirect effect $\boldsymbol{\beta} = \Gamma_1\boldsymbol{\alpha}_0 = c_1\widehat{\Gamma}_1\boldsymbol{\alpha}_0 = -1.5977c_1$. As to the coefficients of confounding variables, we design the following two scenarios of models, without and with confounders, respectively.

4.3.1 Simulation studies without confounding variables

We first consider models without confounding variable $\boldsymbol{z}$. That is, $\boldsymbol{\alpha}_2$ and Γ_2 are both taken zero in model (4.1.1) and (4.1.2). We evaluate the indirect effect tests, by fixing the direct effect $\boldsymbol{\alpha}_1 = c_2 = 0.5$. The left panel of Figure 4.1 depicts the size ($c_1 = 0$) and power ($c_1 = \pm 0.1, \cdots, \pm 1.0$) for the three tests with statistics S, S^O and S^Z. From this figure, powers of all three tests increase as $|c_1|$ increases, and sizes are well controlled. Our proposed test S performs as well as the oracle test S^O, and is more powerful than S^Z. For instance, when $c_1 = 0.4$, the empirical powers of S and S^O are 0.63, while that of S^Z is 0.23.

We also consider testing direct effect $\boldsymbol{\alpha}_1$ by holding $c_1 = 0.5$, corresponding to true value of indirect effect $\beta = -0.7989$. Similarly, the right panel of Figure 4.1 shows the empirical size ($c_2 = 0$) and power ($c_2 = \pm 0.1, \cdots, \pm 1.0$) for the proposed test T, the oracle one T^O, and T^Z proposed by Zhou et al. (2020). The powers of all three tests increase as the value of $|c_2|$ increases. T performs closely with T^O, and is more powerful than T^Z when c_2 is positive. For instance, when $c_2 = 0.2$, the empirical power of T and T^O can reach 0.78, while the empirical power of T^Z test is 0.06.

Moreover, we investigate the performances of the estimators of direct effect $\widehat{\boldsymbol{\alpha}}_1$ and indirect effect $\widehat{\boldsymbol{\beta}}$ in terms of bias and standard deviation. The results are reported in Table 4.11. From this table, the biases of our proposed estimators $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}$ and oracle ones $\widehat{\boldsymbol{\alpha}}_1^O, \widehat{\boldsymbol{\beta}}^O$ are very small, while the biases of $\widehat{\boldsymbol{\alpha}}_1^Z$ are very large. This in turn results in low power of S^Z and T^Z.

Table 4.12 depicts the sample standard deviations of the estimates $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$

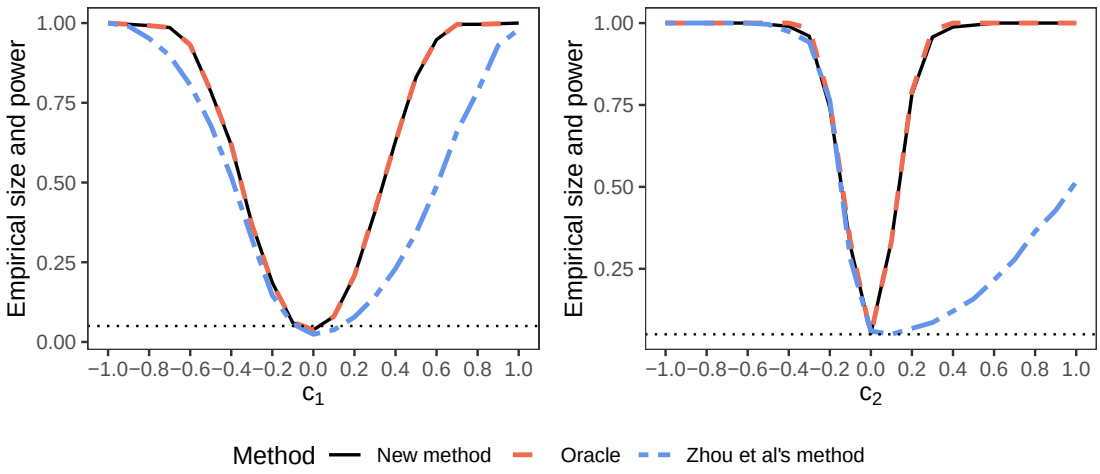

Figure 4.1. Left panel is the empirical sizes and powers of tests S, S^O and S^Z at significance level 0.05 over 500 replications for testing indirect effect of mediation model without confounding variables. Solid line, dash line and two-dash line represent the sizes and powers of S, S^O and S^Z respectively. Right panel is empirical sizes and powers of tests T, T^O and T^Z at significance level 0.05 over 500 replications for testing direct effect of mediation model without confounding variables. The solid, dash line and two-dash line represent the sizes and powers of T, T^O and T^Z, respectively.

Table 4.11. Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2 when there is absent of confounding variables. Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

		New method		Oracle		Zhou et al.'s method	
c_1	c_2	$\widehat{\boldsymbol{\alpha}}_1$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\alpha}}_1^O$	$\widehat{\boldsymbol{\beta}}^O$	$\widehat{\boldsymbol{\alpha}}_1^Z$	$\widehat{\boldsymbol{\beta}}^Z$
-0.8	0.5	$0.28_{(9.14)}$	$-1.69_{(27.25)}$	$0.23_{(8.95)}$	$-1.64_{(27.11)}$	$-8.29_{(15.7)}$	$7.9_{(30.17)}$
-0.4	0.5	$0.17_{(6.95)}$	$-1.58_{(26.48)}$	$0.12_{(6.86)}$	$-1.53_{(26.42)}$	$-4.85_{(14.42)}$	$4.44_{(26.31)}$
0	0.5	$0.01_{(5.98)}$	$-1.00_{(26.10)}$	$0.01_{(5.94)}$	$-0.82_{(26.02)}$	$-22.7_{(12.15)}$	$19.29_{(28.83)}$
0.4	0.5	$-0.22_{(6.87)}$	$-1.21_{(26.16)}$	$-0.09_{(6.69)}$	$-1.01_{(26.23)}$	$-25.04_{(7.65)}$	$44.63_{(28.11)}$
0.8	0.5	$-0.37_{(9.01)}$	$-1.01_{(26.68)}$	$-0.20_{(8.68)}$	$-0.99_{(26.23)}$	$-31.6_{(3.17)}$	$51.17_{(27.27)}$
0.5	-0.8	$-0.24_{(7.42)}$	$-1.9_{(26.72)}$	$-0.15_{(7.20)}$	$-1.29_{(26.34)}$	$2.62_{(15.18)}$	$-3.03_{(26.77)}$
0.5	-0.4	$-0.34_{(7.36)}$	$-1.7_{(26.69)}$	$-0.14_{(7.10)}$	$-1.56_{(26.11)}$	$2.22_{(14.34)}$	$-2.63_{(27.02)}$
0.5	0	$-0.23_{(7.29)}$	$-1.73_{(24.25)}$	$-0.12_{(6.93)}$	$-1.47_{(23.96)}$	$-9.82_{(7.05)}$	$8.40_{(26.07)}$
0.5	0.4	$-0.21_{(7.36)}$	$-2.03_{(25.47)}$	$-0.13_{(7.04)}$	$-1.98_{(25.32)}$	$-30.03_{(7.1)}$	$19.62_{(28.03)}$
0.5	0.8	$-0.18_{(7.47)}$	$-2.18_{(26.75)}$	$-0.16_{(7.11)}$	$-2.17_{(26.44)}$	$-50.92_{(8.79)}$	$50.44_{(29.13)}$

Table 4.12. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications with different c_1 and c_2 when there is absent of confounding variables. Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

		Direct effect ($\widehat{\alpha}_1$)				Indirect Effect ($\widehat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al.'s method	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
-0.8	0.5	9.14	$8.10_{(0.94)}$	8.95	$7.91_{(0.94)}$	27.25	$28.70_{(2.04)}$	27.11	$28.70_{(2.04)}$	30.17	$29.73_{(2.32)}$
-0.4	0.5	6.95	$6.92_{(0.67)}$	6.86	$6.19_{(0.68)}$	26.48	$28.19_{(2.09)}$	26.42	$28.18_{(2.09)}$	26.31	$28.36_{(2.73)}$
0	0.5	5.98	$6.18_{(0.54)}$	5.94	$6.17_{(0.55)}$	26.10	$28.01_{(2.11)}$	26.02	$27.99_{(2.11)}$	28.83	$28.69_{(3.12)}$
0.4	0.5	6.87	$6.93_{(0.79)}$	6.69	$6.92_{(0.71)}$	26.16	$28.19_{(2.09)}$	26.23	$28.18_{(2.1)}$	28.11	$29.44_{(2.74)}$
0.8	0.5	9.01	$8.79_{(0.98)}$	8.68	$8.75_{(1.03)}$	26.68	$28.70_{(2.06)}$	26.23	$28.72_{(2.06)}$	27.27	$29.25_{(2.42)}$
0.5	-0.8	7.42	$7.36_{(0.77)}$	7.20	$7.35_{(0.88)}$	26.72	$28.28_{(2.19)}$	26.34	$28.09_{(2.08)}$	26.77	$29.17_{(2.6)}$
0.5	-0.4	7.36	$7.31_{(0.76)}$	7.10	$7.12_{(0.78)}$	26.69	$27.58_{(2.12)}$	26.11	$27.58_{(2.03)}$	27.02	$28.23_{(2.6)}$
0.5	0	7.29	$7.30_{(0.72)}$	6.93	$7.38_{(0.75)}$	24.25	$25.98_{(2.09)}$	23.96	$24.87_{(2.00)}$	26.07	$28.21_{(2.70)}$
0.5	0.4	7.36	$7.32_{(0.74)}$	7.04	$7.21_{(0.77)}$	25.47	$27.08_{(2.13)}$	25.32	$26.11_{(2.09)}$	28.03	$29.01_{(2.64)}$
0.5	0.8	7.47	$7.34_{(0.76)}$	7.11	$7.29_{(0.79)}$	26.75	$27.84_{(2.17)}$	26.44	$27.50_{(2.19)}$	29.13	$28.19_{(2.73)}$

over 500 replications in the column with label 'std', which can be regarded the true value of standard error of the estimates. These sample standard deviations are also shown in parentheses of Table 4.11. In the column with label 'se(std)' in Tabel 4.12, we report the sample average and sample standard deviation of the 500 estimates of standard errors of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ based on the asymptotic covariance matrix formulas in (4.1.7) and (4.1.8). Note that the R package "freebird" in Zhou et al. (2020) does not provide the estimated standard error of $\widehat{\boldsymbol{\alpha}}_1$. The difference between the column 'std' and 'se(std)' can be used to gauge the performance of the standard error formula based on the asymptotical covariance matrices. From Table 4.12, both the new method and the oracle have smaller difference than the method proposed by Zhou et al. (2020).

4.3.2 Simulation studies with confounding variables

We next examine the performances of the proposed methods for the models with confounding variables. In our simulation, we set the associated coefficients $\boldsymbol{\alpha}_2$ and Γ_2 to be those estimated from the real data.

Figure 4.2 shows the empirical sizes and powers of the tests S, S^O and S^Z for indirect effect, and the tests T, T^O and T^Z for direct effect. The left panel assesses the performance of tests for the indirect effect, holding $c_2 = 0.5$ as constant. From Figure 4.2, S performs as well as S^O, while S^Z exhibits a shifting power curve

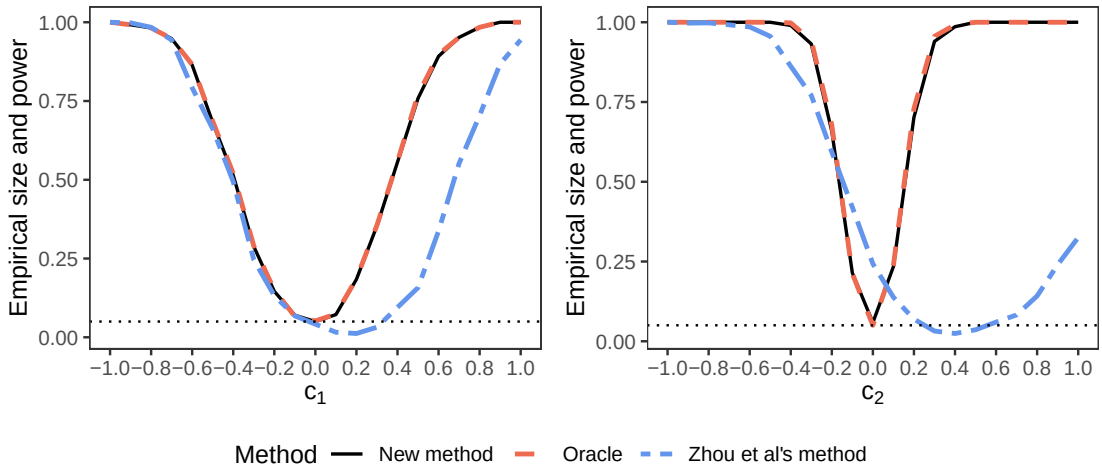

Figure 4.2. Left panel is the empirical sizes and powers of S, S^O and S^Z at significance level 0.05 over 500 replications for testing indirect effect of mediation model contains confounding variables. Right panel is empirical sizes and powers of T, T^O and T^Z at significance level 0.05 over 500 replications for testing direct effect of mediation model contains confounding variables. Caption is the same as that in Figure 4.1.

to the right and the minimum of the curve is not attained at the null hypothesis ($c_1 = 0$). For testing the direct effect $\boldsymbol{\alpha}_1$, we hold $c_1 = 0.5$ and hence $\beta = -0.7989$. The values of c_2 vary from -1 to 1. The results are shown in the right panel of Figure 4.2. Not surprisingly, S performs as well as T^O, while T^Z suffers from an even more severe shifting power curve than S^Z for the indirect effects. For instance, when $c_2 = 0.4$, the empirical power of Zhou et al. (2020) is only 0.024, while the empirical powers of our test and the oracle are 0.986 and 0.992, respectively.

To understand in depth the abnormal behavior of the power curves of Zhou et al. (2020)'s tests, we investigate the performance of estimated direct effect $\widehat{\boldsymbol{\alpha}}_1$ and indirect effect $\widehat{\boldsymbol{\beta}}$ in terms of bias and standard deviation, as reported in Table 4.13. The biases of Zhou et al. (2020)'s estimates are fairly large compared to the proposed estimators $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}$ and oracle ones $\widehat{\boldsymbol{\alpha}}_1^O, \widehat{\boldsymbol{\beta}}^O$. When holding $c_2 = 0.5$, the bias of $\widehat{\boldsymbol{\alpha}}_1^Z$ increases dramatically as c_1 increases. Similar phenomenon occurs when holding $c_1 = 0.5$, where the bias of $\widehat{\boldsymbol{\beta}}^Z$ changes substantially. The bias of estimated $\boldsymbol{\alpha}_1^Z$ and $\boldsymbol{\beta}^Z$ results in the shifted curves shown in Figure 4.2. The large bias of Zhou et al. (2020)'s estimates and the low power of their tests are possibly due to the penalization on direct effect in the scaled lasso, as also pointed out in Zhou et al. (2020) (see the discussion in their Section 7). The penalization on di-

Table 4.13. Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2 when confounding variables involved. Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

		New method		Oracle		Zhou et al.'s method	
c_1	c_2	$\widehat{\boldsymbol{\alpha}}_1$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\alpha}}_1^O$	$\widehat{\boldsymbol{\beta}}^O$	$\widehat{\boldsymbol{\alpha}}_1^Z$	$\widehat{\boldsymbol{\beta}}^Z$
-0.8	0.5	$0.23_{(9.75)}$	$-1.35_{(30.39)}$	$0.08_{(9.48)}$	$-1.2_{(30.25)}$	$-2.64_{(17.97)}$	$1.54_{(25.84)}$
-0.4	0.5	$0.14_{(7.63)}$	$-1.27_{(29.58)}$	$0.02_{(7.44)}$	$-1.14_{(29.56)}$	$-1.87_{(15.81)}$	$0.74_{(22.85)}$
0	0.5	$-0.2_{(6.91)}$	$-0.64_{(29.60)}$	$-0.17_{(6.70)}$	$-0.67_{(29.57)}$	$-21.03_{(14.00)}$	$20.18_{(21.69)}$
0.4	0.5	$-0.45_{(7.88)}$	$-0.39_{(29.68)}$	$-0.34_{(7.57)}$	$-0.5_{(29.74)}$	$-45.44_{(7.84)}$	$44.58_{(27.02)}$
0.8	0.5	$-0.5_{(9.96)}$	$-0.27_{(30.30)}$	$-0.36_{(9.59)}$	$-0.40_{(30.38)}$	$-53.09_{(5.79)}$	$52.31_{(29.18)}$
0.5	-0.8	$-0.19_{(8.25)}$	$-0.93_{(29.36)}$	$-0.12_{(7.91)}$	$-1.00_{(29.42)}$	$7.48_{(18.33)}$	$-8.59_{(21.14)}$
0.5	-0.4	$-0.18_{(8.26)}$	$-0.95_{(29.36)}$	$-0.12_{(7.91)}$	$-1.00_{(29.42)}$	$5.70_{(15.61)}$	$-6.81_{(22.62)}$
0.5	0	$-0.08_{(8.32)}$	$-0.35_{(29.79)}$	$-0.03_{(8.04)}$	$-0.40_{(29.85)}$	$-9.98_{(7.07)}$	$9.55_{(28.13)}$
0.5	0.4	$-0.08_{(8.31)}$	$-0.35_{(29.79)}$	$-0.03_{(8.04)}$	$-0.40_{(29.85)}$	$-40.65_{(6.85)}$	$40.23_{(28.06)}$
0.5	0.8	$-0.18_{(8.27)}$	$-0.95_{(29.36)}$	$-0.12_{(7.91)}$	$-1.01_{(29.42)}$	$-71.1_{(8.51)}$	$69.97_{(26.48)}$

rect effects would make sense when the direct effects are expected to sparse. This is another main merit of applying partial penalized method towards this problem setting.

We explore Zhou et al. (2020)'s method more to understand its inferior performance. Note that the proposed method does not penalize the direct effect α_1, while Zhou et al. (2020)'s method does penalize the direct effect in fitting scaled lasso (Sun and Zhang, 2012). This leads to a larger estimated error variance $\widehat{\sigma}_1^2$ than the proposed method and the oracle estimator. Figure 4.3 compares the estimated $\widehat{\sigma}_1$ of the proposed estimate, oracle estimate and Zhou et al. (2020) when confounding variables are involved in the mediation model. From Figure 4.3, we can observe that when c_1 or c_2 changes from negative to positive, Zhou et al. (2020)'s estimated $\widehat{\sigma}_1$ dramatically increases, while the proposed method and oracle estimate do not. The increasing trend of variance estimation results in large biases of initial estimates used in Zhou et al. (2020), making the debiased step more challenging. In addition, as c_1 or c_2 increases, estimating Ω through $\|\widehat{D} - \widehat{\Omega}\widehat{\Sigma}_{UU}\| \leq \tau$, where $\widehat{\Sigma}_{UU} = \boldsymbol{u}\boldsymbol{u}^T$, $\boldsymbol{u} = (\boldsymbol{m}^T, \boldsymbol{w}^T)^T$, and D and Ω are defined following Zhou et al. (2020), requires larger tuning parameter τ, corresponding to more penalization on parameters and hence further biases as well. Moreover, the power loss in the debiased step is attributed in part to multicollinearity, which also increases with c_1 and c_2, and when more confounders are involved.

As in Table 4.12, we report the empirical standard deviation of the 500 esti-

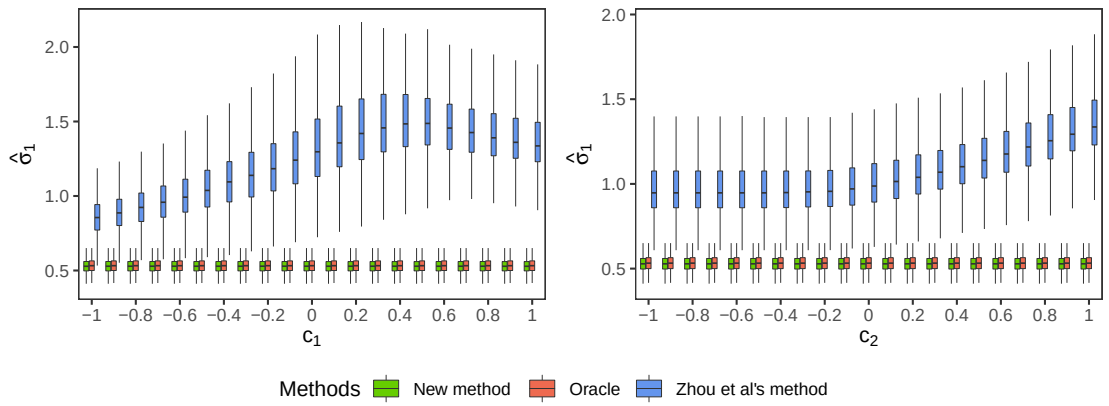

Figure 4.3. Left panel is the estimated $\widehat{\sigma}_1$ of our proposed new method using (4.1.5), oracle and Zhou et al. (2020) (i.e. the scaled Lasso proposed by Sun and Zhang (2012)) over 500 replications by fixing $c_2 = 0.5$ when the mediation model contains confounding variables. Green, red and blue boxes represent the estimate of new method, oracle and Zhou et al. (2020), respectively. Right panel is the estimated $\widehat{\sigma}_1$ of our method using (4.1.5), oracle and Zhou et al. (2020) over 500 replications by fixing $c_1 = 0.5$ when the mediation model contains confounding variables. Green, red and blue boxes represent the estimate of new method, oracle and Zhou et al. (2020), respectively.

Table 4.14. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications with different c_1 and c_2 when confounding variables involved. Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

		Direct effect ($\widehat{\alpha}_1$)				Indirect Effect ($\widehat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al.'s method	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
-0.8	0.5	9.75	$9.58_{(1.12)}$	9.48	$9.71_{(1.11)}$	30.39	$31.94_{(2.38)}$	30.25	$31.93_{(2.38)}$	25.84	$30.88_{(2.60)}$
-0.4	0.5	7.63	$7.67_{(0.83)}$	7.44	$7.77_{(0.82)}$	29.58	$31.41_{(2.43)}$	29.56	$31.4_{(2.43)}$	22.85	$30.17_{(2.87)}$
0	0.5	6.91	$6.93_{(0.7)}$	6.70	$7.03_{(0.70)}$	29.6	$31.21_{(2.45)}$	29.57	$31.19_{(2.46)}$	21.69	$29.22_{(3.22)}$
0.4	0.5	7.88	$7.69_{(0.86)}$	7.57	$7.79_{(0.88)}$	29.68	$31.39_{(2.44)}$	29.74	$31.37_{(2.45)}$	27.02	$29.69_{(3.01)}$
0.8	0.5	9.96	$9.6_{(1.18)}$	9.59	$9.71_{(1.19)}$	30.3	$31.94_{(2.44)}$	30.38	$31.94_{(2.44)}$	29.18	$30.31_{(2.84)}$
0.5	-0.8	8.25	$8.08_{(0.93)}$	7.91	$8.19_{(0.94)}$	29.36	$31.52_{(2.43)}$	29.42	$31.5_{(2.43)}$	21.14	$29.87_{(3.02)}$
0.5	-0.4	8.26	$8.08_{(0.93)}$	7.91	$8.19_{(0.94)}$	29.36	$31.52_{(2.43)}$	29.42	$31.5_{(2.43)}$	22.62	$29.9_{(3.00)}$
0.5	0	8.32	$8.04_{(0.90)}$	8.04	$8.16_{(0.91)}$	29.79	$31.41_{(2.42)}$	29.85	$31.40_{(2.42)}$	28.13	$30.03_{(2.90)}$
0.5	0.4	8.31	$8.04_{(0.90)}$	8.04	$8.16_{(0.91)}$	29.79	$31.41_{(2.42)}$	29.85	$31.4_{(2.42)}$	28.06	$29.86_{(2.91)}$
0.5	0.8	8.27	$8.08_{(0.93)}$	7.91	$8.19_{(0.94)}$	29.36	$31.52_{(2.43)}$	29.42	$31.5_{(2.43)}$	26.48	$29.48_{(3.00)}$

mates and the average of 500 estimated standard errors over the 500 replications in Table 4.14 to examine the accuracy of variance estimation. For the new method and oracle, the standard errors estimated by Monte Carlo simulations are close to those calculated from formulas; while the empirical standard deviation and the average standard error of Zhou et al. (2020) have a large gap.

4.4 Conclusion

Early life trauma plays a critical role in developing psychiatric disorders, typically via altering certain neuroendocrine substances like cortisol. Various researches thus have been conducted to understand the mechanism relating cortisol change to different circumstances of early life trauma. Along with such prolific research output, scientists gradually realized the bridging role of DNA methylation toward the relation between childhood trauma and cortisol stress reactivity. On genome-wide level, Houtepen et al. (2016) conducted a study to investigate how DNA methylation affects cortisol stress reactivity and its relationship with childhood trauma. They identified three top-rated DNA methylation loci by ranking the p-values in an increasing order, one of which, on the KITLG gene (cg27512205), was shown not only to associate with cortisol change, but also partly mediate the relationship between childhood trauma and cortisol stress reactivity. Another study by van Kesteren and Oberski (2019), however, yielded a completely different set of loci, based on the same data set while using their proposed CMF algorithm.

Motivated by such contradictory results in Houtepen et al. (2016) and van Kesteren and Oberski (2019), in which the authors did not consider the potentially active mediating loci jointly, we propose a partial penalized least squared method for linear mediation models with high-dimensional mediators in the presence of confounders. We further develop relevant tests for the direct and indirect effects in such high-dimensional linear mediation models. Simulation studies validate the capability of the proposed approach for efficiently estimating and testing the direct and indirect effects, and the numerical comparisons also imply that the proposed procedure outperforms the debiased method advocated by Zhou et al. (2020).

We utilize this partial penalized least squares method and testing procedures to investigate the high dimensional mediating effects of DNA methylation loci to relate childhood trauma and cortisol stress reactivity, with confounding variables involved. For comparison purpose, we analyze the same data set as Houtepen et al. (2016) and van Kesteren and Oberski (2019). We choose to include the eight DNA methylation loci discovered by these two papers in the model as domain knowledge. Using the proposed approach, we identified three new loci, on the RAB5IF gene (cg19230917), the CPQ gene (cg06422529) and the AGPAT1 gene

(cg03199124), respectively, that actively play the mediator role. We compare our findings with Houtepen et al. (2016) and van Kesteren and Oberski (2019) from statistical perspectives, where tests and related analyses are all in favor of the three newly identified loci. Furthermore, we estimate and test the direct and indirect effects for childhood trauma on cortisol change, and conclude that the early life trauma affects cortisol only indirectly through DNA methylation and the indirect effect is negative, while the direct effect is insignificant.

From domain knowledge in existing literature, we also provide biological and genetical interpretations for the three selected loci and their belonging genes. The RAB5IF gene takes charge of cell endocytosis process, by which cells engulf substances like cortisol, thus reasonably serves as a mediator which transmits the hormone change brought by the traumatic stress. As to the CPQ gene, previous research has verified it as a pathway from stress to cortisol change in fish. Thus incorporating our findings, an neurobiological exploration toward its role in human is worthwhile. The AGPAT1 gene, on the other hand, was shown to control fat tissue biosynthesis; while some retrospective studies demonstrated that fat biosynthesis and storage caused by trauma stress is accompanied with abnormal hormonal level such as cortisol. Therefore, our findings may offer potential clues for such pathophysiological mechanism research. In short, statistical tests and scientific interpretations both show convincing evidence that the newly identified three DNA methylation loci, or their located genes, should be considered as active mediators that relate childhood trauma and cortisol stress reactivity.

Chapter 5

Generalized Mediation Models with High-Dimensional Potential Mediators

The outbreak of COVID-19 came as a shock to the U.S. economy, and the stock market witnessed an unprecedented roller-coaster ride. It is challenging in making investment decisions in the volatile markets for investors because "normal" strategies are unlikely to work (Yousfi et al., 2021; Broadstock et al., 2021). Therefore, it is of great interest to investigate market reaction to exogenous shock and seek for drivers of company values during the crisis. To accomplish this, we propose a generalized mediation model with high-dimensional potential mediators to study the mediation effects of financial metrics that bridge company's sector and stock value.

We, in chapter 3, propose a new procedure for high dimensional mediation inference for continuous outcome but in practice, the response may be binary or counts. For example, in clinical trial, the subjects can be either healthy or have a certain disease; and in financial industry, the stocks can be either classified as valuable stocks, which have growth potential, or not. In the chapter, we propose an estimation procedure via a partial penalized maximum likelihood method and establish its theoretical properties. We develop a Wald test for the indirect effect and show that the proposed test has a χ^2 limiting null distribution. We also develop a partial penalized likelihood ratio test for the direct effect and show that

the proposed test asymptotically follows a χ^2-distribution under null hypothesis. A more efficient estimator of indirect effect under complete mediation model is also developed. Simulation studies are conducted to examine the finite sample performance of the proposed procedures and compare with some existing methods. We further illustrate the proposed methodology with an empirical analysis of stock reaction to COVID-19 pandemic via exploring the underlying mechanism of the relationship between companies' sectors and their stock values.

In section 5.1, we introduce a new statistical inference procedure for the indirect effect and establish its theoretical properties for HDGMM. In section 5.2, we construct a partial penalized likelihood ratio test for the direct effect. section 5.3 presents numerical studies. Conclusion and discussion are given in section 5.4. All proofs about HDGMM are presented in section 5.5.

5.1 Inference for the indirect effect

Let Y be the response, $\boldsymbol{m}$ consist of the p-dimensional mediator variables, and $\boldsymbol{x}$ consist of q-dimensional exposure variables. Consider HDGMM

$$g\{E(Y|\boldsymbol{m},\boldsymbol{x})\} = \boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x}, \tag{5.1.1}$$

$$\boldsymbol{m} = \Gamma^T\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{5.1.2}$$

where $g(\cdot)$ is the link function, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are p- and q-dimensional regression coefficient vectors, Γ is a $q \times p$ coefficient matrix, and $\boldsymbol{\varepsilon}$ is random error with $var(\boldsymbol{\varepsilon}) = \Sigma^*$ and is independent with $\boldsymbol{x}$. In this paper, q is fixed, while p is allowed to diverge to infinity.

It is assumed throughout this paper that conditioning on $\boldsymbol{x}, \boldsymbol{m}$, the conditional distribution of Y belongs to the exponential family with canonical link. Thus, the conditional density function of Y can be written as

$$\exp[\{y(\boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x}) - b(\boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x})\}/\phi_0 + c(y)], \tag{5.1.3}$$

where ϕ_0 is some positive scale parameter. It is further assumed that the link function $b(\cdot)$ is known, thrice continuously differentiable with $b''(\cdot) > 0$. Denote

$\varepsilon_1 = Y - b'(\boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x})$. It is noted that $b'(\boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x}) = \mu(\boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x}) = E(Y|\boldsymbol{m}, \boldsymbol{x})$ and thus $E(\varepsilon_1|\boldsymbol{m}, \boldsymbol{x}) = 0$.

Define $\boldsymbol{\beta} = \Gamma\boldsymbol{\alpha}_0$. It follows by (5.1.2) that

$$\boldsymbol{\alpha}_0^T\boldsymbol{m} = \boldsymbol{\beta}^T\boldsymbol{x} + \varepsilon_2, \tag{5.1.4}$$

where $\varepsilon_2 = \boldsymbol{\alpha}_0^T\boldsymbol{\varepsilon}$ with $var(\varepsilon_2) = \sigma_2^2 = \boldsymbol{\alpha}_0^T\Sigma^*\boldsymbol{\alpha}_0$. In (5.1.4), $\boldsymbol{\beta}$ is the indirect effect of the exposure variable $\boldsymbol{x}$ on the response Y via mediators $\boldsymbol{m}$ under some causal assumptions (Valeri and VanderWeele, 2013; Preacher, 2015; VanderWeele, 2015).

In mediational analysis, of great interest is to test whether the indirect effect exists or not

$$H_{01} : \boldsymbol{\beta} = \mathbf{0} \text{ versus } H_{11} : \boldsymbol{\beta} \neq \mathbf{0}. \tag{5.1.5}$$

Suppose that $\{\boldsymbol{m}_i, \boldsymbol{x}_i, Y_i\}$, $i = 1, \cdots, n$, is a random sample from model with (5.1.1) and (5.1.2). We next propose an estimation procedure for $\boldsymbol{\beta}$. Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$, $\boldsymbol{M} = (\boldsymbol{m}_1, \cdots, \boldsymbol{m}_n)^T$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$. Denote $\widehat{\Gamma} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}$, the least squares estimate of Γ. If $\boldsymbol{\alpha}_0$ was known, a natural estimator of $\boldsymbol{\beta}$ would be

$$\widehat{\Gamma}\boldsymbol{\alpha}_0 = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\boldsymbol{\alpha}_0. \tag{5.1.6}$$

Of course, $\boldsymbol{\alpha}_0$ is a high-dimensional unknown vector. Let $\widehat{\boldsymbol{\alpha}}_0$ be an estimator of $\boldsymbol{\alpha}_0$, and let $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\widehat{\boldsymbol{\alpha}}_0$. Then it follows from (5.1.4) that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon_2} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}(\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0), \tag{5.1.7}$$

where $\boldsymbol{\varepsilon_2} = \{\varepsilon_{21}, \cdots, \varepsilon_{2n}\}^T$. Thus, we need to control the estimation error in the last term $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}(\widehat{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_0)$ in order to establish the asymptotic normality of $\widehat{\boldsymbol{\beta}}$. To achieve good estimation accuracy of $\widehat{\boldsymbol{\alpha}}_0$ to $\boldsymbol{\alpha}_0$, one needs to impose sparsity assumption on $\boldsymbol{\alpha}_0$. That is, although the dimension p of $\boldsymbol{\alpha}_0$ could be very large, only a small proportion of elements in $\boldsymbol{\alpha}_0$ are nonzero, and the corresponding mediators in $\boldsymbol{m}$ are truly relevant to the response Y. This sparsity assumption is mild and commonly imposed for high-dimensional mediation model with continuous response (Zhou et al., 2020; Song et al., 2020). Under this sparsity assumption, we consider the following partial penalized likelihood function to simultaneously

estimate $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1$ and select important mediator variables

$$(\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\alpha}}_0) = \arg\max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0} \frac{1}{n}\sum_{i=1}^{n}[Y_i(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i) - b(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i)] - \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|), \tag{5.1.8}$$

where $p_\lambda(\cdot)$ is a penalty function with a tuning parameter λ. The partial penalized likelihood (5.1.8) penalizes only the high-dimensional $\boldsymbol{\alpha}_0$ and does not penalize the direct effect $\boldsymbol{\alpha}_1$. This enables us to construct statistical inference procedure on $\boldsymbol{\alpha}_1$.

5.1.1 Asymptotic results

In this section, we investigate the statistical properties of the estimators. We assume that the penalty function $p_\lambda(t_0)$ is increasing and concave in $t_0 \in [0, \infty)$, and has a continuous derivative $p'_\lambda(t_0)$ with $p'_\lambda(0+) > 0$. Let $\rho(t_0, \lambda) = p_\lambda(t_0)/\lambda$ for $\lambda > 0$. In addition, assume $\rho'(t_0, \lambda)$ is increasing in $\lambda \in (0, \infty)$ and $\rho'(0+, \lambda)$ is independent of λ. For any vector $\mathbf{v} = (v_1, \cdots, v_l)^T$, define

$$\bar{\rho}(\mathbf{v}, \lambda) = \{\mathrm{sgn}(v_1)\rho'(|v_1|, \lambda), \cdots, \mathrm{sgn}(v_l)\rho'(|v_l|, \lambda)\}^T,$$

where $\mathrm{sgn}(v_1) = I(v_1 > 0) - I(v_1 < 0)$. We further define the local concavity of the penalty function ρ at $\mathbf{v}$ as

$$\kappa(\rho, \mathbf{v}, \lambda) = \lim_{\epsilon \to 0^+} \max_{1 \le j \le l} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\rho'(t_2, \lambda) - \rho'(t_1, \lambda)}{t_2 - t_1}.$$

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_0^T)^T$ and $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_1^{\star T}, \boldsymbol{\alpha}_0^{\star T})^T$, the true value of $\boldsymbol{\theta}$. Further let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\alpha}}_1^T, \widehat{\boldsymbol{\alpha}}_0^T)$ be the estimator of $\boldsymbol{\theta}_0$. Denote $\mathcal{A} = \{j : \boldsymbol{\alpha}_{0j}^\star \neq 0\}$ and $s = |\mathcal{A}|$ is the number of elements in $\mathcal{A}$, which is also called sparsity level. Let $\boldsymbol{M}^j$ denote the jth column of $\boldsymbol{M}$, $\boldsymbol{M}_\mathcal{A}$ be the submatrix of $\boldsymbol{M}$ formed by columns in $\mathcal{A}$ and $\boldsymbol{m}_{i,\mathcal{A}}$ be the ith column of the matrix $\boldsymbol{M}_\mathcal{A}^T$. Similarly, let $\boldsymbol{\alpha}_{0,\mathcal{A}}^\star$ be the subvector of $\boldsymbol{\alpha}_0^\star$ formed by elements in $\mathcal{A}$. Define $\mathcal{A}^c = [1, \cdots, p]/\mathcal{A}$ be the complement set of $\mathcal{A}$. Moreover $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_{0,\mathcal{A}}^T)^T$. And $\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\vartheta}}$ are similarly defined.

Let d_n be the half minimum signal of $\boldsymbol{\alpha}_{0,\mathcal{A}}^\star$, i.e. $d_n = \min_{j \in \mathcal{A}} |\alpha_{0j}^\star|/2$. Define $\mathcal{N}_0 = \{\boldsymbol{\vartheta} \in \mathbb{R}^{s+q} : \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \le d_n\}$. Let $z(\boldsymbol{\vartheta}) = \boldsymbol{\alpha}_1^T\boldsymbol{x} + \boldsymbol{\alpha}_{0,\mathcal{A}}^T\boldsymbol{m}_\mathcal{A}$, $\Sigma_{XX}(\boldsymbol{\vartheta}) = E[\boldsymbol{x}b''(z(\boldsymbol{\vartheta}))\boldsymbol{x}^T]$, $\Sigma_{XM}(\boldsymbol{\vartheta}) = E[\boldsymbol{x}b''(z(\boldsymbol{\vartheta}))\boldsymbol{m}_\mathcal{A}^T]$, $\Sigma_{MM}(\boldsymbol{\vartheta}) = E[\boldsymbol{m}_\mathcal{A}b''(z(\boldsymbol{\vartheta}))\boldsymbol{m}_\mathcal{A}^T]$,

and

$$\Sigma(\boldsymbol{\vartheta}) \;=\; \begin{pmatrix} \Sigma_{XX}(\boldsymbol{\vartheta}) & \Sigma_{XM}(\boldsymbol{\vartheta}) \\ \Sigma_{MX}(\boldsymbol{\vartheta}) & \Sigma_{MM}(\boldsymbol{\vartheta}) \end{pmatrix}.$$

$\Sigma_0 = \Sigma(\boldsymbol{\vartheta}_0)$. Further denote $\tilde{\Sigma}_{XX} = E[\boldsymbol{x}\boldsymbol{x}^T]$ and $\tilde{\Sigma}_{XM} = E[\boldsymbol{x}\boldsymbol{m}_{\mathcal{A}}^T]$. In our paper, for any vector $\boldsymbol{v} = (v_1, \cdots, v_l)^T$, $\|\boldsymbol{v}\|_\infty = \max_i |v_i|$ and $\|\boldsymbol{v}\|_2 = (\boldsymbol{v}^T\boldsymbol{v})^{1/2}$. $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denotes the minimum and maximum eigenvalues of the matrix A. $\|A\|_{2,\infty} = \sup_{\boldsymbol{v}:\|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_\infty$.

We impose the following conditions:

A1. $\inf_{\boldsymbol{\vartheta}\in\mathcal{N}_0} \lambda_{\min}(\Sigma(\boldsymbol{\vartheta})) \geq c$, $\lambda_{\max}(\Sigma_0) = O(1)$ and $\|\boldsymbol{M}_{\mathcal{A}^c}^T b''(\boldsymbol{W}\boldsymbol{\vartheta}_0)\boldsymbol{W}\|_{2,\infty} = O_p(n)$ with $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})$.

A2. $d_n \gg \lambda_n \gg \max\{\sqrt{s/n}, \sqrt{\log p/n}\}$, $p'_{\lambda_n}(d_n) = o((ns)^{-1/2})$, $\lambda_n\kappa_0 = o(1)$ where $\kappa_0 = \max_{\boldsymbol{\delta}\in\mathcal{N}_0} \kappa(\rho, \boldsymbol{\delta}, \lambda_n)$. Further $d_n \gg \lambda_n$ means $\lim_{n\to\infty} d_n/\lambda_n = \infty$.

A3. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i, \boldsymbol{m}_{i,\mathcal{A}}), i = 1, \cdots, n$. With $0 \leq a < 1/2$,

$$\max_{j\in\mathcal{A}^c} \sup_{\boldsymbol{\vartheta}\in\mathcal{N}_0} \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i m_{ij} b'''(\boldsymbol{w}_i^T\boldsymbol{\vartheta})\boldsymbol{w}_i^T) = O_p((n/s)^a).$$

For some $\varpi > 2$, there exists a positive sequence K_n such that $E[\|\boldsymbol{m}_{\mathcal{A}^c}\varepsilon_1\|_\infty^\varpi] \leq K_n^\varpi$ and $K_n^2 \log p/n^{1-2/\varpi-\varsigma} \to 0$ for some arbitrary small $\varsigma > 0$. Further assume that

$$\max_{1\leq j\leq s+q} \sup_{\boldsymbol{\vartheta}\in\mathcal{N}_0} E(w_j^4 b''^2(\boldsymbol{w}^T\boldsymbol{\vartheta})) < C < \infty.$$

To emphasize the dependence on the sample size, in the above conditions and the Appendix, we use λ_n to denote the tuning parameter. These conditions are mild and commonly imposed. See, for instance, Fan and Lv (2011) and Shi et al. (2019). Condition A2 imposes a minimal signal condition on nonzero elements in $\boldsymbol{\alpha}_0$. However, no minimal signal condition is imposed on the direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta} = \boldsymbol{\gamma} - \boldsymbol{\alpha}_1$, which are of our primary interest. Note that in mediation analysis, we are interested in the direct effect $\boldsymbol{\alpha}_1$ and the indirect effect $\boldsymbol{\beta}$ instead of $\boldsymbol{\alpha}_0$. Thus, Condition A2 is reasonable in practice. It is worth to pointing out that both $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}$ are finite dimensional, while $\boldsymbol{\alpha}_0$

is high-dimensional under the model setting studied in this paper. Compared with existing literature, A3 is also mild. In fact, Shi et al. (2019) assumed that $\max_{j\in\mathcal{A}^c}\sup_{\boldsymbol{\vartheta}\in\mathcal{N}_0}\lambda_{\max}(\frac{1}{n}\sum_{i=1}^n \boldsymbol{w}_i m_{ij} b'''(\boldsymbol{w}_i^T\boldsymbol{\vartheta})\boldsymbol{w}_i^T) = O_p(1)$. That is, take a to be 0. Here we allow the maximum eigenvalues to be diverging. Different from some existing works, for instance Wang et al. (2012), in which it is assumed that all covariates are uniformly bounded, in this paper we assume some moment conditions. Under bounded covariates condition, $E[\|\boldsymbol{m}_{\mathcal{A}^c}\varepsilon_1\|_\infty^\varpi] \le K_n^\varpi$ reduces to $E(|\varepsilon_1|^\varpi) \le C$ by taking K_n as a constant. From Conditions A2 and A3, it is clear that the dimension p is allowed to be an exponential order of the sample size n. Under these conditions, we have the following theorem.

Theorem 3. *Suppose that Conditions (A1)-(A3) hold, and $s = o(n^{1/2})$. It follows that (i) with probability tending to 1, $\widehat{\boldsymbol{\alpha}}_0$ must satisfy (i) $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$; (ii) $\|\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}^\star_{0,\mathcal{A}}\|_2 = O_p(\sqrt{s/n})$. If further $s = o(n^{(1-2a)/(3-2a)})$,*

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \;=\; \Sigma_0^{-1}\frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{X}^T\\ \boldsymbol{M}_{\mathcal{A}}^T\end{pmatrix}[\boldsymbol{Y} - b'(\boldsymbol{M}\boldsymbol{\alpha}_0^\star + \boldsymbol{X}\boldsymbol{\alpha}_1^\star)] + o_p(1).$$

Theorem 3 establishes the sparsity, consistency, and asymptotic representation of the estimators $\widehat{\boldsymbol{\alpha}}_0$ and $\widehat{\boldsymbol{\vartheta}}$. In Theorem 3, the number of nonzero elements of $\boldsymbol{\alpha}_0$ is allowed to diverge at rate $o(n^{1/2})$, while for asymptotic expansion, stronger condition on the sparsity level s is required. It depends on the size of a in condition A3. If $a = 0$ as assumed in Shi et al. (2019), s is allowed to be $o(n^{1/3})$. While if a is relatively large, then sparsity condition on s would be stronger.

Let $\Sigma_{XX} = \Sigma_{XX}(\boldsymbol{\vartheta}_0)$, $\Sigma_{XM} = \Sigma_{XM}(\boldsymbol{\vartheta}_0)$ and $\Sigma_{MM} = \Sigma_{MM}(\boldsymbol{\vartheta}_0)$. Further

$$B = \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}(0_{s\times q}, I_{s\times s})\Sigma_0^{-1}(0_{s\times q}, I_{s\times s})^T\tilde{\Sigma}_{XM}\tilde{\Sigma}_{XX}^{-1}.$$

Using Theorem 3, we may obtain the asymptotic normality of the estimates of the direct effect and the indirect effect in the following corollary.

Corollary 2. *Under Conditions (A1)-(A3), and assume $s = o(n^{(1-2a)/(3-2a)})$, it follows that*

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) \to N(0, \phi_0(I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}(I_{q\times q}, 0_{q\times s})^T),$$

and further

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \to N(0, \sigma_2^2 \tilde{\Sigma}_{XX}^{-1} + \phi_0 B).$$

For ease of notation, models (5.1.1) and (5.1.2) does not include potential confounders. If certain confounders are noteworthy, we can accordingly modify the model by

$$g(E(Y|\boldsymbol{m}, \boldsymbol{x}, \boldsymbol{u})) = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \boldsymbol{\alpha}_u^T \boldsymbol{u}, \text{ and } \quad \boldsymbol{m} = \Gamma^T \boldsymbol{x} + \Gamma_u^T \boldsymbol{u} + \boldsymbol{\varepsilon}, \quad (5.1.9)$$

where $\boldsymbol{u}$ consists of r-dimensional potential confounders. Here we assume the dimension r is fixed and finite. In this model, $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta} = \Gamma \boldsymbol{\alpha}_0$ still stand for the direct effect and indirect effect, respectively.

Now let $\boldsymbol{v} = (\boldsymbol{x}^T, \boldsymbol{u}^T)^T, \boldsymbol{\alpha}_v = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_u^T)^T, \Gamma_v = (\Gamma^T, \Gamma_u^T)^T$ and $\boldsymbol{\beta}_v = \Gamma_v \boldsymbol{\alpha}_0$. Thus, the model can be rewritten as:

$$g(E(Y|\boldsymbol{m}, \boldsymbol{v})) = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_v^T \boldsymbol{v}, \text{ and } \quad \boldsymbol{m} = \Gamma_v^T \boldsymbol{v} + \boldsymbol{\varepsilon}. \quad (5.1.10)$$

This implies that the parametric vectors $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ and the random vector $\boldsymbol{v}$ can be viewed as $\boldsymbol{\alpha}_1, \boldsymbol{\beta}$, and $\boldsymbol{x}$ in the original mediation model, respectively. Then we can estimate $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$ in the same way as that (5.1.7) and (5.1.8). The asymptotic distributions of corresponding estimators $\widehat{\boldsymbol{\alpha}}_v$ and $\widehat{\boldsymbol{\beta}}_v$ can be easily derived by Corollary 1. Further note that $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}$ are the first part of $\boldsymbol{\alpha}_v$ and $\boldsymbol{\beta}_v$, respectively. Parallel to Corollary 1, we have the following corollary.

Corollary 3. *Suppose that Conditions (A1)-(A3) with $\boldsymbol{x}$ replaced by $\boldsymbol{v}$ hold, and further $s = o(n^{(1-2a)/(3-2a)})$, it follows that*

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) \to N\left(0, \phi_0 (I_{q\times q}, 0_{q\times r})(\Sigma_{VV}^{-1} + \Sigma_{VV}^{-1}\Sigma_{VM}\Sigma_{MM.V}^{-1}\Sigma_{MV}\Sigma_{VV}^{-1}) \begin{pmatrix} I_{q\times q} \\ 0_{q\times r} \end{pmatrix}\right),$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \to N\left(0, (I_{q\times q}, 0_{q\times r})(\sigma_2^2\tilde{\Sigma}_{VV}^{-1} + \phi_0 \tilde{\Sigma}_{VV}^{-1}\tilde{\Sigma}_{VM}\Sigma_{MM.V}^{-1}\tilde{\Sigma}_{MV}\tilde{\Sigma}_{VV}^{-1}) \begin{pmatrix} I_{q\times q} \\ 0_{q\times r} \end{pmatrix}\right).$$

Here $\Sigma_{MM.V} = \Sigma_{MM} - \Sigma_{MV}\Sigma_{VV}^{-1}\Sigma_{VM}$ and the matrices $\Sigma_{VV}, \Sigma_{VM}, \tilde{\Sigma}_{VV}$ and $\tilde{\Sigma}_{VM}$ are similarly defined as $\Sigma_{XX}, \Sigma_{XM}, \tilde{\Sigma}_{XX}$ and $\tilde{\Sigma}_{XM}$ respectively by replacing $\boldsymbol{x}$ with $\boldsymbol{v}$.

5.1.2 Test for indirect effect

Corollaries 1 and 2 enable us to construct a test statistic for indirect effect using Wald test statistic. This requires estimating unknown quantities in the asymptotic covariance matrix. First, Σ_{XM} can be estimated by it moment estimator

$$\widehat{\Sigma}_{XM} = \frac{1}{n}\boldsymbol{X}^T b''(z(\widehat{\boldsymbol{\vartheta}}))\boldsymbol{M}_{\widehat{\mathcal{A}}}$$

with $\widehat{\mathcal{A}} = \{j : \widehat{\alpha}_{0j} \neq 0\}$. Under Condition (A3) and $s = o(n^{1/2})$, it can be shown that $\widehat{\Sigma}_{XM} = \Sigma_{XM} + o_p(1)$. The other matrices $\Sigma_{XX}, \Sigma_{MM}, \tilde{\Sigma}_{XX}, \tilde{\Sigma}_{XM}, \Sigma_0$ and B can be estimated by their sample counterparts. For logistic or Poisson regression models, $\phi_0 = 1$. For normal linear regression models, ϕ_0 equals the error variance. The consistent estimate of error variance in high-dimensional linear regression model has been studied in Fan et al. (2012). Thus, assume that there exists a consistent estimate $\widehat{\phi}$ of ϕ_0. By definition of σ_2^2, we may construct its estimator as

$$\widehat{\sigma}_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\alpha}}_0^T\boldsymbol{m}_i - \widehat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)^2.$$

It is shown to be a consistent estimator of σ_2^2 in the appendix. Thus the asymptotic variance matrices of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ can be consistently estimated by:

$$\widehat{\phi}_0(I_{q\times q}, 0_{q\times\widehat{s}})\widehat{\Sigma}_0^{-1}(I_{q\times q}, 0_{q\times\widehat{s}})^T;\ \widehat{\sigma}_2^2\widehat{\tilde{\Sigma}}_{XX}^{-1} + \widehat{\phi}_0\widehat{B}. \tag{5.1.11}$$

Here $\widehat{s} = |\widehat{\mathcal{A}}|$. Then the Wald test for the hypotheses (5.1.5) can be derived as

$$S_n = n\widehat{\boldsymbol{\beta}}^T\left(\widehat{\sigma}_2^2\widehat{\tilde{\Sigma}}_{XX}^{-1} + \widehat{\phi}_0\widehat{B}\right)^{-1}\widehat{\boldsymbol{\beta}}.$$

Clearly, under H_{01}, $S_n \to \chi_q^2$, where χ_q^2 stands for a χ^2-distribution with q degrees of freedom.

To investigate the power performance of S_n, we consider the following local alternative hypotheses:

$$H_{1n}:\ \boldsymbol{\beta} = \Delta/\sqrt{n}.$$

It follows from Corollary 1 that under H_{1n}, $S_n \to \chi_q^2(\Delta^T(\sigma_2^2\tilde{\Sigma}_{XX}^{-1} + \phi_0 B)^{-1}\Delta)$.

Here $\chi^2_q(C)$ is the χ^2-distribution with q degrees of freedom and noncentrality parameter C. Thus, the proposed test S_n can detect local alternative hypotheses which converge to the null hypothesis with root-n rate.

5.2 Test for direct effect

The coefficient $\boldsymbol{\alpha}_1$ is the direct effect of $\boldsymbol{x}$ on Y. When $\boldsymbol{\alpha}_1 = 0$ in model (5.1.1), the mediation models are called complete or full mediation models. Otherwise, they are called incomplete or partial mediation models. Thus, it is of interest in testing

$$H_{02} : \boldsymbol{\alpha}_1 = \mathbf{0} \text{ versus } H_{12} : \boldsymbol{\alpha}_1 \neq \mathbf{0}. \tag{5.2.1}$$

There are two possible approaches for addressing this hypothesis testing problem. One approach is the Wald test based on the asymptotical normality of $\widehat{\boldsymbol{\alpha}}_1$. The other one is based on the likelihood ratio test, which is our focus in this section. Under H_{02}, we consider a penalized likelihood

$$\frac{1}{n}\sum_{i=1}^{n}[Y_i(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i) - b(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i)] - \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|). \tag{5.2.2}$$

Denote $\tilde{\boldsymbol{\alpha}}_0$ to be the penalized likelihood estimate of $\boldsymbol{\alpha}_0$. The likelihood ratio test may be defined

$$T_n = 2n\{L_n(\widehat{\boldsymbol{\theta}}) - L_n(\tilde{\boldsymbol{\theta}})\}/\widehat{\phi}_0,$$

where $L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n}[Y_i(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i) - b(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i)]/n$, $\tilde{\boldsymbol{\theta}} = (0, \tilde{\boldsymbol{\alpha}}_0^T)^T$ and $\widehat{\phi}_0$ is a consistent estimator of ϕ_0.

Theorem 4. *Suppose that Conditions (A1)-(A3) hold and* $s = o(n^{(1-2a)/(3-2a)})$. *Under* $H_{2n} : \boldsymbol{\alpha}_1 = \boldsymbol{h}/\sqrt{n}$, *it follows that*

$$\sup_x |P(T_n \leq x) - P(\chi^2_q(\boldsymbol{h}^T\Phi^{-1}\boldsymbol{h}/\phi_0) \leq x)| \to 0, \tag{5.2.3}$$

where $\Phi = (I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}(I_{q\times q}, 0_{q\times s})^T$, *and with slight abuse of notation,* $\chi^2_q(C)$ *stands for the* χ^2 *random variable with* q *degrees of freedom and noncentrality parameter* C.

From Theorem 4, the limiting null distribution of T_n follow a χ^2_q, which does not depend on any unknown parameters. This reveals the Wilks phenomenon in this setting, in which there exists a high dimensional nuisance parameter $\boldsymbol{\alpha}_0$. Furthermore, the proposed test T_n can detect local alternatives that are distinct from the null hypothesis at $n^{-1/2}$ rate.

Under $H_{02}: \boldsymbol{\alpha}_1 = \mathbf{0}$, we can further obtain the following proposition. Denote $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}\tilde{\boldsymbol{\alpha}}_0$ and $\tilde{B} = \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}\Sigma_{MM}^{-1}\tilde{\Sigma}_{XM}\tilde{\Sigma}_{XX}^{-1}$.

Proposition 1. *Suppose that Conditions (A1)-(A3) hold and* $s = o(n^{(1-2a)/(3-2a)})$. *It follows that under* $H_{02}: \boldsymbol{\alpha}_1 = \mathbf{0}$,

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star) \to N(0, \sigma_2^2\tilde{\Sigma}_{XX}^{-1} + \phi_0\tilde{B}).$$

The proof of Proposition 1 is similar to the Corollary 1 and thus is omitted.

Under complete mediation model, the indirect effect $\boldsymbol{\beta}$ can be consistently estimated by $\widehat{\boldsymbol{\beta}}$, which was proposed under incomplete mediation. Note that

$$(0_{s\times q}, I_{s\times s})\Sigma_0^{-1}(0_{s\times q}, I_{s\times s})^T = (\Sigma_{MM} - \Sigma_{MX}\Sigma_{XX}^{-1}\Sigma_{XM})^{-1} > \Sigma_{MM}^{-1}.$$

Here $A > B$ means that $A - B$ is positive semi-definite. Then from Proposition 1 and Corollary 1, $\tilde{\boldsymbol{\beta}}$ is asymptotically more efficient than $\widehat{\boldsymbol{\beta}}$. This result is expected since $\tilde{\boldsymbol{\beta}}$ uses the extra information that $\boldsymbol{\alpha}_1 = \mathbf{0}$ under complete mediation model.

5.3 Numerical studies

In this section, we assess the finite sample performance of the proposed procedures via Monte Carlo simulation and illustrate the proposed methodology via a real data example. We first provide the practical implementation algorithm of the proposed procedures.

5.3.1 Algorithm for practical implementation

In the numerical study, we adopt the SCAD penalty (Fan and Li, 2001) in the partial penalized likelihood, and its first derivative is defined to be

$$p'_\lambda(t) = \lambda\{I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\}.$$

with $p_\lambda(0) = 0$ and $a = 3.7$. The tuning parameter λ for our method is chosen based on the high-dimensional BIC (HBIC) method in Wang et al. (2013).

For a given λ, define

$$(\widehat{\boldsymbol{\alpha}}_1^\lambda, \widehat{\boldsymbol{\alpha}}_0^\lambda) = \arg\max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_0} \frac{1}{n}\sum_{i=1}^n [Y_i(\boldsymbol{\alpha}_0^T \boldsymbol{m}_i + \boldsymbol{\alpha}_1^T \boldsymbol{x}_i) - b(\boldsymbol{\alpha}_0^T \boldsymbol{m}_i + \boldsymbol{\alpha}_1^T \boldsymbol{x}_i)] - \sum_{j=1}^p p_\lambda(|\alpha_{0j}|). \tag{5.3.1}$$

Following Wang et al. (2013), we apply the local linear approximation (LLA) in Zou and Li (2008) to the penalty, and carry out the corresponding maximization problem via the following algorithm.

Step 1 Set initial values $\boldsymbol{\alpha}_0^{(0)}, \boldsymbol{\alpha}_1^{(0)}$ to be

$$(\widehat{\boldsymbol{\alpha}}_0^{(0)}, \widehat{\boldsymbol{\alpha}}_1^{(0)}) = \max_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1} \frac{1}{n}\sum_{i=1}^n [Y_i(\boldsymbol{\alpha}_0^T \boldsymbol{m}_i + \boldsymbol{\alpha}_1^T \boldsymbol{x}_i) - b(\boldsymbol{\alpha}_0^T \boldsymbol{m}_i + \boldsymbol{\alpha}_1^T \boldsymbol{x}_i)] - \lambda \sum_{j=1}^p |\alpha_{0j}|$$

by the above partial penalized likelihood function with the Lasso penalty.

Step 2 Solve

$$\begin{aligned}(\widehat{\boldsymbol{\alpha}}_0^{(1)}, \widehat{\boldsymbol{\alpha}}_1^{(1)}) \quad = \max_{\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1} \quad & \{\frac{1}{n}\sum_{i=1}^n [Y_i(\boldsymbol{\alpha}_0^T \boldsymbol{m}_i + \boldsymbol{\alpha}_1^T \boldsymbol{x}_i) - b(\boldsymbol{\alpha}_0^T \boldsymbol{m}_i + \boldsymbol{\alpha}_1^T \boldsymbol{x}_i)] \\ & - \sum_{j=1}^p p'_\lambda(|\alpha_{0,j}^{(0)}|)|\alpha_{0,j}|\}. \end{aligned} \tag{5.3.2}$$

In practice, it is desirable to use a data-driven method to choose the tuning parameter λ. In our numerical study, λ is selected by HBIC, whose score is defined

as

$$\begin{aligned}\text{HBIC}(\lambda) &= \log(\sum_{i=1}^{n}[Y_i(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i+\boldsymbol{\alpha}_1^T\boldsymbol{x}_i)-b(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i+\boldsymbol{\alpha}_1^T\boldsymbol{x}_i)]/n)\\ &\quad +\text{df}\cdot\log(\log(n))\cdot\log(p+q)/n,\end{aligned}$$

where df is the number of variables with nonzero coefficients in $(\widehat{\boldsymbol{\alpha}}_0^T,\widehat{\boldsymbol{\alpha}}_1^T)^T$. Minimizing HIBIC($\lambda$) yields an estimate of λ.

5.3.2 Simulation studies

We examine the finite sample performance of the proposed tests, and compare them with the oracle test statistics that include only mediators in the true set $\mathcal{A}=\{j:\alpha_{0j}^\star\neq 0\}$, denoted as S_n^O and T_n^O. We also compare the test for indirect effect with the global test S_n^G proposed by Djordjilović et al. (2019), using R package 'globaltest'.

5.3.2.1 Logistic regression

In this example, we set $n=300$, $q=1$, and $p=500$. We first generate the exposure variable $\boldsymbol{x}\sim N(0,1)$ and mediators $\boldsymbol{m}=\Gamma^T\boldsymbol{x}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}\sim N(\mathbf{0},\Sigma^*)$ with Σ^* being an AR correlation structure. That is, the (i,j)-element of Σ^* equals $\rho^{|i-j|}$ with $\rho=0.5$. Take $\Gamma=c_1(\tau_1,\cdots,\tau_p)^T$, where $\tau_k=0.1k$ for $k=1,\cdots,5$, and when $k>5$, τ_k's are independently generated from $N(0,0.1^2)$. Set $c_1=0$ to examine Type I error rate and $c_1=\pm0.1,\pm0.2,\cdots,\pm1$ for power when testing the indirect effects.

We then generate the response Y from model

$$\text{logit}\{\Pr(Y=1|\boldsymbol{m},\boldsymbol{x})\}=\boldsymbol{\alpha}_0^T\boldsymbol{m}+\boldsymbol{\alpha}_1^T\boldsymbol{x}, \tag{5.3.3}$$

where $\text{logit}(p)=\log\{p/(1-p)\}$, $\boldsymbol{\alpha}_0=[3,1.5,0,0,2,0,\cdots 0]^T$ and $\boldsymbol{\alpha}_1=c_2$ is set in the same fashion as c_1. The simulation results are based on 500 replications. The significance level is set to be 0.05.

We first compare the performances of S_n, S_n^O, S_n^G for testing the indirect effect $\boldsymbol{\beta}$. We set $c_2=1$ and $\boldsymbol{\beta}=\Gamma\boldsymbol{\alpha}_0=1.6c_1$. The left panel of Figure 5.1 depicts power

functions of the three tests versus the values of c_1 over $[-0.8, 0.8]$. All the tests gain larger powers as $|c_1|$ increases. S_n performs as well as the oracle S_n^O. For instance, when $c_1 = -0.2$, the empirical power of S_n is 0.216, while the empirical powers of S_n^O is 0.218 and S_n^G is 0.072. Note that, when $|c_1|$ grows larger, S_n^G's powers grow slowly and are lower than those of S_n and S_n^O. These observations are consistent with the theoretical results in Section 2.

Next, we turn to test the direct effect. Set $c_1 = 1$. And c_2 is taken from $0, \pm 0.1, \pm 0.2, \cdots, \pm 1$, where $c_2 = 0$ corresponds to the null hypothesis. The right panel of Figure 5.1 depicts the power function of the two tests versus the values of c_2 over $[-1, 1]$. The proposed test T_n performs almost the same as the oracle one. In fact, when $c_2 = -0.2$, the empirical powers of our test statistic T_n is 0.134 and the oracle test T_n^O is 0.126.

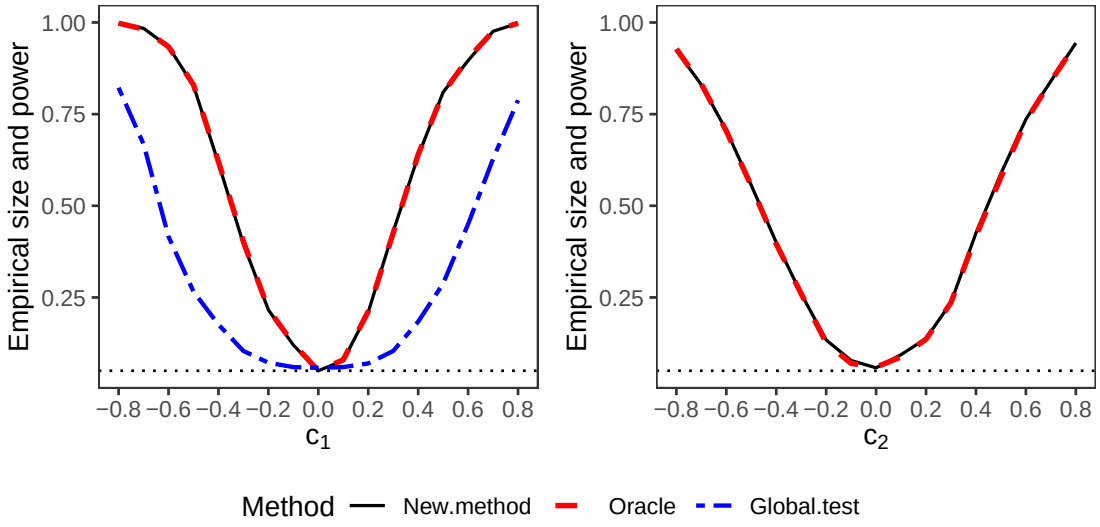

Figure 5.1. Left panel is the empirical sizes and powers of S_n, S_n^O and S_n^G at level $\alpha = 0.05$ over 500 replications for testing indirect effect of logistic regression when $\alpha_1 = 1$. Solid line, dash line and two-dash line represent the sizes and powers of S_n, S_n^O and S_n^G, respectively. Right panel is empirical sizes and powers of T_n and T_n^O at level $\alpha = 0.05$ over 500 replications for testing direct effect of logistic regression when $\beta = 1.6$. The solid and dash line represent the sizes and powers of T_n and T_n^O, respectively.

Table 5.1 depicts biases and standard errors of the estimated indirect and direct effects. It can be seen from Table 5.1 that the proposed estimator can estimate the effects accurately, and its performance is close to the oracle ones.

To assess accuracy of the estimation of the standard error of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$, Table 5.2

Table 5.1. Estimated biases and standard deviations (in parentheses) of logistic regression with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		New Method		Oracle	
c_1	c_2	$\widehat{\boldsymbol{\alpha}}_1$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\alpha}}_1^O$	$\widehat{\boldsymbol{\beta}}^O$
-0.4	1	$3.99_{(26.12)}$	$-2.74_{(30.79)}$	$3.24_{(25.93)}$	$-2.25_{(30.32)}$
-0.2	1	$5.23_{(26.32)}$	$-1.99_{(29.53)}$	$4.39_{(26.02)}$	$-1.65_{(29.31)}$
0	1	$5.51_{(26.96)}$	$-0.75_{(29.28)}$	$4.58_{(26.27)}$	$-0.62_{(29.06)}$
0.2	1	$5.40_{(25.24)}$	$0.75_{(29.66)}$	$4.94_{(24.87)}$	$0.60_{(29.62)}$
0.4	1	$6.28_{(26.36)}$	$2.12_{(31.20)}$	$5.57_{(25.49)}$	$1.71_{(30.88)}$
1	-0.4	$-0.87_{(25.56)}$	$4.60_{(39.28)}$	$-0.73_{(24.88)}$	$4.51_{(38.82)}$
1	-0.2	$0.64_{(26.02)}$	$7.20_{(39.98)}$	$0.65_{(25.18)}$	$5.89_{(39.04)}$
1	0	$2.15_{(25.82)}$	$4.27_{(39.16)}$	$2.31_{(25.21)}$	$4.00_{(38.30)}$
1	0.2	$2.86_{(25.65)}$	$6.67_{(38.98)}$	$2.52_{(25.05)}$	$5.45_{(38.13)}$
1	0.4	$4.30_{(25.71)}$	$4.87_{(39.90)}$	$3.66_{(24.82)}$	$4.03_{(38.80)}$

Table 5.2. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications of logistic regression with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		Direct effect ($\widehat{\alpha}_1$)				Indirect Effect ($\widehat{\beta}$)			
		New Method		Oracle		New Method		Oracle	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)
-0.4	1	26.12	$25.20_{(3.12)}$	25.93	$25.04_{(3.06)}$	30.79	$29.82_{(4.68)}$	30.32	$29.56_{(4.56)}$
-0.2	1	26.32	$25.14_{(3.26)}$	26.02	$24.93_{(3.16)}$	29.53	$28.95_{(4.59)}$	29.31	$28.64_{(4.43)}$
0	1	26.96	$25.21_{(3.50)}$	26.27	$24.98_{(3.32)}$	29.28	$28.62_{(4.66)}$	29.06	$28.33_{(4.49)}$
0.2	1	25.24	$25.38_{(3.22)}$	24.87	$25.23_{(3.13)}$	29.66	$28.99_{(4.68)}$	29.62	$28.78_{(4.53)}$
0.4	1	26.36	$26.03_{(3.42)}$	25.49	$25.79_{(3.18)}$	31.20	$30.21_{(4.97)}$	30.88	$29.87_{(4.66)}$
1	-0.4	25.56	$24.89_{(2.62)}$	24.88	$24.69_{(2.55)}$	39.28	$37.16_{(6.63)}$	38.82	$36.79_{(6.49)}$
1	-0.2	26.02	$24.72_{(2.59)}$	25.18	$24.47_{(2.40)}$	39.98	$37.39_{(6.72)}$	39.04	$36.91_{(6.37)}$
1	0	25.82	$24.69_{(2.48)}$	25.21	$24.49_{(2.33)}$	39.16	$37.21_{(6.60)}$	38.30	$36.80_{(6.25)}$
1	0.2	25.65	$25.02_{(2.57)}$	25.05	$24.83_{(2.49)}$	38.98	$37.39_{(6.58)}$	38.13	$36.96_{(6.29)}$
1	0.4	25.71	$25.69_{(2.77)}$	24.82	$25.51_{(2.70)}$	39.90	$37.59_{(6.79)}$	38.80	$37.26_{(6.50)}$

depicts their estimated standard errors in two ways. As to each method - new and oracle - the first column lists the empirical standard deviations of point estimates $\widehat{\boldsymbol{\alpha}}_1$ or $\widehat{\boldsymbol{\beta}}$ over 500 replications (they are also recorded in parentheses of Table 5.1); for the second column, we estimate standard errors of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ using formula (5.1.11) in each simulation run, and reports the average together with standard deviations (in parentheses) over the 500 replications. Table 5.2 implies that the proposed estimators of the standard errors perform well.

Table 5.3. Estimated biases and standard deviations (in parentheses) of Possion regression with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		New Method		Oracle	
c_1	c_2	$\widehat{\boldsymbol{\alpha}}_1$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\alpha}}_1^O$	$\widehat{\boldsymbol{\beta}}^O$
-0.4	0.1	$-0.02_{(3.06)}$	$-0.29_{(9.81)}$	$-0.02_{(3.05)}$	$-0.31_{(9.8)}$
-0.2	0.1	$-0.16_{(3.09)}$	$-0.28_{(9.75)}$	$-0.15_{(3.08)}$	$-0.30_{(9.74)}$
0	0.1	$0.06_{(3.00)}$	$-0.30_{(9.77)}$	$0.06_{(2.99)}$	$-0.31_{(9.78)}$
0.2	0.1	$0.02_{(3.08)}$	$-0.30_{(9.71)}$	$-0.02_{(3.08)}$	$-0.30_{(9.71)}$
0.4	0.1	$0.19_{(3.05)}$	$-0.23_{(9.77)}$	$0.15_{(3.05)}$	$-0.22_{(9.77)}$
0.1	-0.4	$-0.12_{(3.06)}$	$-0.31_{(9.81)}$	$-0.11_{(3.05)}$	$-0.31_{(9.81)}$
0.1	-0.2	$-0.07_{(3.11)}$	$-0.36_{(9.76)}$	$-0.08_{(3.12)}$	$-0.37_{(9.76)}$
0.1	0	$0.02_{(3.08)}$	$-0.3_{(9.79)}$	$-0.01_{(3.10)}$	$-0.31_{(9.79)}$
0.1	0.2	$0.01_{(3.21)}$	$-0.23_{(9.74)}$	$-0.01_{(3.19)}$	$-0.22_{(9.74)}$
0.1	0.4	$0.10_{(2.87)}$	$-0.20_{(9.73)}$	$0.09_{(2.84)}$	$-0.20_{(9.73)}$

5.3.2.2 Poisson regression

We generate the response Y from model

$$Y|\boldsymbol{m},\boldsymbol{x} \sim \text{Poisson}(\lambda(\boldsymbol{m},\boldsymbol{x})), \quad \text{with} \quad \lambda(\boldsymbol{m},\boldsymbol{x}) = \exp\{\boldsymbol{\alpha}_0^T\boldsymbol{m} + \boldsymbol{\alpha}_1^T\boldsymbol{x}\},$$

where $\boldsymbol{\alpha}_0 = [0.8, 0, 0.6, 1, 0, \cdots 0]^T$ and $\boldsymbol{\alpha}_1 = c_2$. Further $\boldsymbol{m} = \Gamma^T\boldsymbol{x} + \boldsymbol{\varepsilon}$ and take $\Gamma = c_1(\tau_1, \cdots, \tau_p)^T$, where $\tau_k = 0.1k$ for $k = 1, \cdots, 5$, and when $k > 5$, τ_k's are independently generated from $N(0, 0.01^2)$. Other settings are the same as those in Section 4.2.1.

The simulation results of the proposed test statistics and estimates are summarized in Figure 5.2, Tables 5.3 and 5.4. From these figures and tables, it can be seen clearly that the proposed test statistics S_n for indirect effect and T_n for direct effect perform similarly as the oracle test statistics S_n^O and T_n^O in terms of empirical sizes and powers. And S_n is much more powerful than global test S_n^G. The estimators $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ have similar performance as $\widehat{\boldsymbol{\alpha}}_1^O$ and $\widehat{\boldsymbol{\beta}}^O$ in terms of estimated biases and standard deviations. The results in Table 5.4 confirm that the proposed standard error estimators perform well.

5.3.3 An application

Stock market had experienced a dramatic crash and then a rapid recovery during the Covid-19 pandemic. It brings about an opportunity to reassess stock selection strategy. One question worthy to think through is whether financial statements

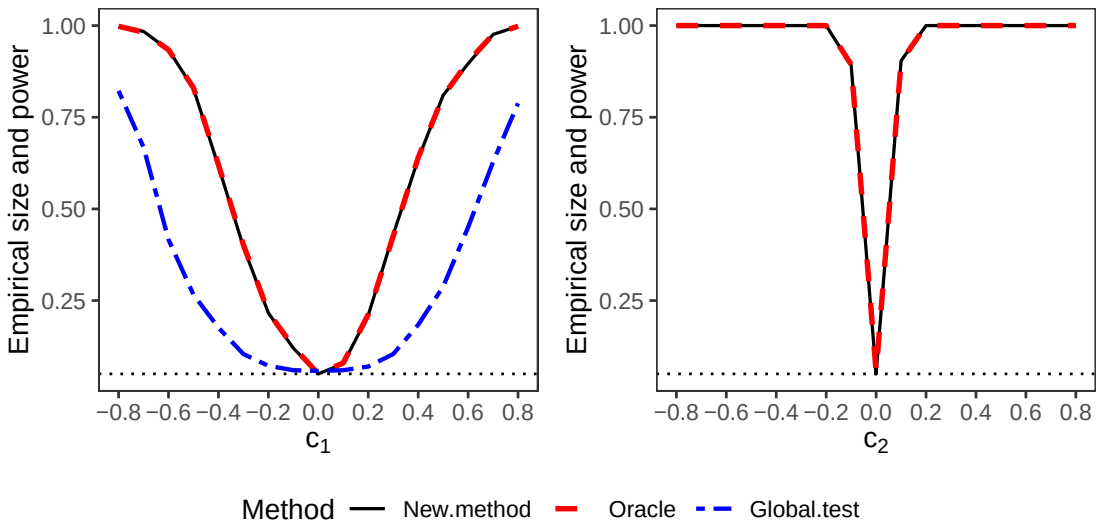

Figure 5.2. Left panel is the empirical sizes and powers of S_n, S_n^O and S_n^G at level $\alpha = 0.05$ over 500 replications for testing indirect effect of poisson regression when $\alpha_1 = 0.1$. Solid line, dash line and two-dash line represent the sizes and powers of S_n, S_n^O and S_n^G, respectively. Right panel is empirical sizes and powers of T_n and T_n^O at level $\alpha = 0.05$ over 500 replications for testing direct effect of poisson regression when $\beta = 0.082$. The solid line and dash line represent the sizes and powers of T_n and T_n^O, respectively.

mediate the company sectors and stock price performance during the pandemic. To address this question, we apply the proposed procedures to a dataset consisting of the S&P 500 companies' financial statements and their stock performance during the first few months of the COVID-19 outbreak. The proposed mediation analysis might help investors to obtain business insights from important financial metrics, and invest their money in the companies that can quickly recover if a similar disaster had occurred in the future.

In the pandemic, the stock market first experienced a large drop, dragging the major S&P 500 index down 33.92% from its prior peak on February 19, 2020, and triggering four trading halts in two weeks. Yet from March 25th, 2020, the market started an unpredictably recovery from the dip, with S&P 500 up 30.17% from its relative bottom by April 30th, 2020. In this market turmoil, a few stocks fell less but bounced higher than the rest of the market. These stocks as valuable stocks, and therefore, it is worthwhile to study the underlying mechanisms that can detect the valuable stocks during the pandemic.

Table 5.4. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications of Possion regression with different c_1 and c_2. Except for c_1 and c_2, the values in this table equals 100 times of the actual ones.

		Direct effect ($\widehat{\alpha}_1$)				Indirect Effect ($\widehat{\beta}$)			
		New Method		Oracle		New Method		Oracle	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)
-0.4	0.1	3.06	$3.16_{(0.4)}$	3.05	$3.16_{(0.4)}$	9.81	$9.75_{(0.55)}$	9.8	$9.75_{(0.55)}$
-0.2	0.1	3.09	$3.15_{(0.4)}$	3.08	$3.14_{(0.4)}$	9.75	$9.73_{(0.56)}$	9.74	$9.74_{(0.56)}$
0	0.1	3	$3.12_{(0.39)}$	2.99	$3.12_{(0.39)}$	9.77	$9.74_{(0.57)}$	9.78	$9.74_{(0.56)}$
0.2	0.1	3.08	$3.11_{(0.39)}$	3.08	$3.11_{(0.39)}$	9.71	$9.73_{(0.58)}$	9.71	$9.73_{(0.57)}$
0.4	0.1	3.05	$3.1_{(0.39)}$	3.05	$3.1_{(0.39)}$	9.77	$9.74_{(0.57)}$	9.77	$9.74_{(0.57)}$
0.1	-0.4	3.06	$3.03_{(0.4)}$	3.05	$3.02_{(0.4)}$	9.81	$9.73_{(0.56)}$	9.81	$9.74_{(0.56)}$
0.1	-0.2	3.11	$3.12_{(0.4)}$	3.12	$3.12_{(0.4)}$	9.76	$9.74_{(0.56)}$	9.76	$9.74_{(0.56)}$
0.1	0	3.08	$3.13_{(0.39)}$	3.1	$3.13_{(0.39)}$	9.79	$9.74_{(0.57)}$	9.79	$9.74_{(0.56)}$
0.1	0.2	3.21	$3.08_{(0.39)}$	3.19	$3.08_{(0.39)}$	9.74	$9.73_{(0.57)}$	9.74	$9.74_{(0.57)}$
0.1	0.4	2.87	$2.95_{(0.39)}$	2.84	$2.95_{(0.39)}$	9.73	$9.73_{(0.58)}$	9.73	$9.74_{(0.58)}$

We use python web scraper to collect the stock price data from February, 2020 to April 30th, 2020 on Yahoo Finance. After removing missing values, there are 490 companies under study. To explicitly capture the value of stocks, we first define two quantities, maximum loss and recovery return. A stock's maximum loss is defined as the ratio between the lowest closing price in March, 2020 and the highest closing price in February, 2020. And a stock's recovery return is defined as the ratio between the closing price on April 30, 2020 and the highest closing price in February, 2020. The closing price is adjusted for both dividends and splits. Thus, a valuable stock is to have lower maximum loss and higher recovery return compared with the S&P 500 index. We code the valuable stocks as '1' and otherwise '0', to serve as the response, Y, in the mediation model.

We also scrapped the financial statements from Yahoo Finance and generated 550 accounting metrics from these financial statements as the potential mediators in $\boldsymbol{m}$. We obtain firms' annual reports from fiscal year 2015 to 2019 and the first three quarterly reports in 2019. The exposure variables, $\boldsymbol{x}$, are companies' belonging sectors according to Global Industry Classification Standard (GICS). GICS classifies companies into eleven sectors: basic materials, communication services, consumer cyclical, consumer defensive, energy, financial services, healthcare, indus-

trials, real estate, technology and utilities. With 'Energy' set to be the reference level, we adopt ten dummy variables for $\boldsymbol{x}$.

We conduct the Wald test for indirect effect and the partial penalized likelihood ratio test for direct effect. For testing indirect effect $H_0 : \boldsymbol{\beta} = 0$, the test statistic $S_n = 26.367$ with P-value 0.003. As for testing direct effect $H_0 : \boldsymbol{\alpha}_1 = 0$, the test statistic $T_n = 85.952$ with P-value$< 10^{-14}$. Both the indirect and direct effect are significant at level 0.05. Table 5.5 presents the estimated indirect effects and direct effects of each stock sectors, together with their standard errors. It can be observed from Table 5.5 that all direct effects are positive. This implies that the stocks in 'Energy' sector got a big hit by the COVID-19 pandemic outbreak and generally performed worse than stocks of other sectors. The 'Communication service' and 'Technology' sectors have high direct effects, suggesting that stocks in these sectors were more likely to perform better than those in 'Energy' sector in terms of direct effect. This can be well explained that people had increasing need for stay-at-home entertainments, such as internet service, during the lockdown. And the 'Healthcare' sector also exhibited outstanding performance because the demands for medicines and vaccines were tremendous. Futhermore, the selected mediators and their mutational effects are presented in Table 5.6.

In a nutshell, an investor may refer to this analysis to select stocks during the pandemic-led stock market crisis. For sector effects, the sectors in 'Healthcare', 'Consumer defensive', 'Communication service', 'Utility' and 'Technology' are likely to outperform. As for the financial statements effects, one may focus on those reported in Table 5.6 to filter stocks. For example, we shall pick companies that have higher values in return on assets, gross margin and AGR ROIC but lower values in debt to assets.

5.4 Conclusion

In this chapter, we studied the statistical inference of the high dimensional generalized mediation model. We proposed a partial penalized maximum likelihood method and investigated its statistical properties. We further proposed a Wald test, whose limiting null distribution follows a χ^2 distribution. We also provided a likelihood ratio test for the direct effect and revealed the Wilks phenomenon un-

Table 5.5. The estimated coefficients, standard errors, test statistics values and p-values for real data.

Sectors	Indirect effect	SE	Direct effect	SE
Basic materials	0.383	0.373	1.860	1.179
Communication services	0.610	0.447	3.781	1.247
Consumer cyclical	0.529	0.319	0.123	1.161
Consumer defensive	0.625	0.394	3.893	1.143
Financial services	0.699	0.348	0.393	1.145
Healthcare	1.429	0.355	2.573	1.086
Industrials	0.912	0.331	1.294	1.100
Real estate	0.849	0.342	1.358	1.160
Technology	1.321	0.342	1.762	1.086
Utilities	0.811	0.369	2.158	1.125

Table 5.6. Selected importance mediators and their coefficients

Selected mediator	Estimated coefficient (std)	Description
Return on assets	0.722(0.144)	Net income divided by the total assets
Gross margin	0.409(0.159)	The difference between the revenue and cost of goods sold divided by revenue
AGR* ROIC	0.531(0.203)	Net operating profit after tax divided by the sum of debt and equity
Inventory growth 2019Q3	0.394(0.200)	The growth of inventory in 2019Q3 comparing to 2018Q3
Debt to assets	-0.327(0.138)	Total debts divided by total assets

* AGR: average growth rate, calculated as the average of growth rates for the metrics from 2015 to 2019.

der this model setting. Simulation studies validated the finite sample performances of the proposed tests and estimates. And a real data analysis was conducted for studying the mediation roles of financial metrics for relating the companies' sectors and its investment values during the Covid-19 pandemic.

5.5 Proof of theorem 3 and 4

Proof of Theorem 3: The proof of this theorem follows from the arguments of Theorem 1 and Corollary 2 in Loh and Wainwright (2015). We only give a brief

proof here.

Firstly, as shown in Loh and Wainwright (2015), the loglikelihood loss function

$$-\frac{1}{n}\sum_{i=1}^{n}[Y_i(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i+\boldsymbol{\alpha}_1^T\boldsymbol{x}_i)-b(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i+\boldsymbol{\alpha}_1^T\boldsymbol{x}_i)]=\mathcal{L}_n(\boldsymbol{\theta})$$

satisfies the so-called restricted strong convexity (RSC) condition. Secondly note that the estimator $\widehat{\boldsymbol{\theta}}$ satisfies

$$\langle\nabla\mathcal{L}_n(\widehat{\boldsymbol{\theta}})+\nabla P_\lambda(\widehat{\boldsymbol{\theta}}),\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\rangle\geq 0,\ \text{ for all feasible } \boldsymbol{\theta}.$$

Here $\nabla P_\lambda(\widehat{\boldsymbol{\theta}})=(\nabla p_\lambda(\widehat{\boldsymbol{\alpha}}_0)^T,\mathbf{0}^T)^T$. Then following similar argument of Loh and Wainwright (2015), we conclude that $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2\leq 1$. This further implies that

$$\begin{aligned}(\psi-\frac{\gamma}{2})\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2 &\leq p_\lambda(\boldsymbol{\alpha}_0^\star)-p_\lambda(\widehat{\boldsymbol{\alpha}}_0)+\frac{\lambda L}{2}\left(\|\widehat{\boldsymbol{\alpha}}_0-\boldsymbol{\alpha}_0^\star\|_1+\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^\star\|_1\right)\\ &\leq \frac{3}{2}p_\lambda(\boldsymbol{\alpha}_0^\star)-\frac{1}{2}p_\lambda(\widehat{\boldsymbol{\alpha}}_0)+\frac{\gamma}{4}\|\widehat{\boldsymbol{\alpha}}_0-\boldsymbol{\alpha}_0^\star\|_2^2+\frac{\lambda L}{2}\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^\star\|_1.\end{aligned}$$

Here ψ is a constant which depends on the population distribution and we assume that $\psi>3\gamma/4$. Since $\lambda\to 0$, then $\lambda\leq\gamma/(2Lq)$ and

$$\frac{\lambda L}{2}\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^\star\|_1 \leq \frac{\lambda L}{2}\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^\star\|_1^2\leq\frac{\lambda Lq}{2}\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^\star\|_2^2\leq\frac{\gamma}{4}\|\widehat{\boldsymbol{\alpha}}_1-\boldsymbol{\alpha}_1^\star\|_2^2.$$

It then follows that:

$$0\leq(\psi-\frac{3\gamma}{4})\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2 \leq \frac{3}{2}p_\lambda(\boldsymbol{\alpha}_0^\star)-\frac{1}{2}p_\lambda(\widehat{\boldsymbol{\alpha}}_0).$$

The other steps are similar to those in Loh and Wainwright (2015) and thus omitted here.

To finish the proof, we still need to show that $\|\nabla\mathcal{L}_n(\boldsymbol{\theta}_0)\|_\infty=O_p(\sqrt{\log p/n})$. Let $\boldsymbol{z}_i=(\boldsymbol{x}_i^T,\boldsymbol{m}_i^T)^T, i=1,\cdots,n$ and $\boldsymbol{\varepsilon}_1=\boldsymbol{Y}-b'(\boldsymbol{M}\boldsymbol{\alpha}_0^\star+\boldsymbol{X}\boldsymbol{\alpha}_1^\star)$. Then

$$\nabla\mathcal{L}_n(\boldsymbol{\theta}_0)=-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{1i}\boldsymbol{z}_i.$$

Let $a_n = n^{1/\varpi+\varsigma}K_n, b = \sqrt{Cn\log p}$ with C being large enough and note that

$$\begin{aligned} z_{ij}\varepsilon_{i1} &= z_{ij}\varepsilon_{i1}I(|z_{ij}\varepsilon_{i1}| \le a_n) - E[z_j\varepsilon_1 I(|z_j\varepsilon_1| \le a_n)] \\ &\quad + z_{ij}\varepsilon_{i1}I(|z_{ij}\varepsilon_{i1}| > a_n) - E[z_j\varepsilon_1 I(|z_j\varepsilon_1| > a_n)] \\ &=: \epsilon_{ij,1} + \epsilon_{ij,2}. \end{aligned}$$

$$\begin{aligned} &P\left(|\sum_{i=1}^n z_{ij}\varepsilon_{i1}| > b, \text{ for some } j = 1, \cdots, p+q\right) \\ &\le P\left(|\sum_{i=1}^n \epsilon_{ij,1}| + |\sum_{i=1}^n \epsilon_{ij,2}| > b, \text{ for some } j = 1, \cdots, p+q\right) \\ &\le P\left(|\sum_{i=1}^n \epsilon_{ij,1}| > b/2, \text{ for some } j\right) + P\left(|\sum_{i=1}^n \epsilon_{ij,2}| > b/2, \text{ for some } j\right) \\ &=: P_1 + P_2. \end{aligned}$$

Firstly consider the term P_1. Note that $\epsilon_{1j,1}, \ldots, \epsilon_{nj,1}$ are independent centered random variables a.s. bounded by $2a_n$ in absolute value. Then the Bernstein inequality yields that

$$\begin{aligned} P_1 &\le 2(p+q)\max_j \exp\left\{-\frac{b^2/4}{2nE(\epsilon_{j,1}^2) + 2\cdot 2a_n\cdot b/(2\cdot 3)}\right\} \\ &\le 2(p+q)\max_j \exp\left\{-\frac{C\log p/4}{2E(\epsilon_{j,1}^2) + 2a_n\sqrt{C\log p/n}/3}\right\} \to 0. \end{aligned}$$

Next we turn to consider P_2. First note that

$$P_2 \le P\left(\sum_{i=1}^n \max_j |z_{ij}\varepsilon_{i1}|I(|z_{ij}\varepsilon_{i1}| > a_n) + \max_j nE[|z_j\varepsilon_1|I(|z_j\varepsilon_1| > a_n)] > b/2\right)$$

Further note that

$$E^2[|z_j\varepsilon_1|I(|z_j\varepsilon_1| > a_n)] \le E[z_j^2\varepsilon_1^2]P(|z_j\varepsilon_1| > a_n) \le E[z_j^2\varepsilon_1^2]\frac{E[|z_j\varepsilon_1|^\varpi]}{a_n^\varpi}.$$

We then conclude that

$$\max_j nE[|z_j\varepsilon_1|I(|z_j\varepsilon_1| > a_n)] \leq \max_j n\sqrt{\frac{E[z_j^2\varepsilon_1^2]E[|z_j\varepsilon_1|^{\varpi}]}{a_n^{\varpi}}} = o(\sqrt{n}).$$

From this, we then have

$$\begin{aligned} P_2 &\leq P\left(\sum_{i=1}^n \max_j |z_{ij}\varepsilon_{i1}|I(|z_{ij}\varepsilon_{i1}| > a_n) > b/4\right) \\ &\leq P\left(\max_j |z_{ij}\varepsilon_{i1}| > a_n \text{ for some } i\right) \\ &\leq n\frac{E[\max_j |z_j\varepsilon_1|^{\varpi}]}{a_n^{\varpi}} = o(1). \end{aligned}$$

Thus $\|\nabla\mathcal{L}_n(\boldsymbol{\theta}_0)\|_\infty = O_p(\sqrt{\log p/n})$. The Theorem is then proven. □

Define

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n [Y_i(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i) - b(\boldsymbol{\alpha}_0^T\boldsymbol{m}_i + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i)] - \sum_{j=1}^p p_\lambda(|\alpha_{0j}|).$$

Proof of Theorem 4: To enhance the readability, we divide the proof of Theorem 2 into three steps. In the first step, we show that there exists a local maximizer $\bar{\boldsymbol{\theta}}$ of $Q_n(\boldsymbol{\theta})$ with the constraints $\bar{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$, such that $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_P(\sqrt{s/n})$. In the second step, we prove that $\bar{\boldsymbol{\theta}}$ is indeed a local maximizer of $Q_n(\boldsymbol{\theta})$. This implies $\widehat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$. In the final step, we derive the asymptotic expansions of $\widehat{\boldsymbol{\theta}}$.

Step 1: Consistency in the $(s+q)$-dimensional subspace: We first constrain $Q_n(\boldsymbol{\theta})$ on the $(s+q)$-dimensional subspace of $\{\boldsymbol{\theta} \in \mathbb{R}^{s+q} : \boldsymbol{\alpha}_{0,\mathcal{A}^c} = 0\}$ of $\mathbb{R}^{p+q}$. This constrained partial penalized likelihood function is given by

$$\bar{Q}_n(\boldsymbol{\vartheta}) = \frac{1}{n}\sum_{i=1}^n [Y_i(\boldsymbol{\delta}^T\boldsymbol{m}_{i,\mathcal{A}} + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i) - b(\boldsymbol{\delta}^T\boldsymbol{m}_{i,\mathcal{A}} + \boldsymbol{\alpha}_1^T\boldsymbol{x}_i)] - \sum_{j=1}^s p_\lambda(|\alpha_{0j}|)$$

Here $\boldsymbol{\vartheta} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\delta}^T)^T$ and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_s)^T$. We now show that there exists a strict local maximizer $\bar{\boldsymbol{\vartheta}}$ of $\bar{Q}_n(\boldsymbol{\vartheta})$ such that $\|\bar{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\|_2 = O_P(\sqrt{s/n})$. To this end, we

consider an event

$$H_n = \{\bar{Q}_n(\boldsymbol{\vartheta}_0) > \max_{\boldsymbol{\vartheta}\in\partial\mathcal{N}_\tau} \bar{Q}_n(\boldsymbol{\vartheta})\}.$$

where $\mathcal{N}_\tau = \{\boldsymbol{\vartheta} \in \mathbb{R}^{s+q} : \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \leq \tau\sqrt{s/n}\}$ with $\tau \in (0, \infty)$, and $\partial\mathcal{N}_\tau$ denotes the boundary of the closed set $\mathcal{N}_\tau$. Clearly, on the event H_n, there exists a local maximizer of $\bar{Q}_n(\boldsymbol{\vartheta})$ in $\mathcal{N}_\tau$. Thus, we only need to show that $P(H_n) \to 1$ as $n \to \infty$ when τ is large. To this aim, we next analyze the function $\bar{Q}_n$ on the boundary $\partial\mathcal{N}_\tau$.

For any $\boldsymbol{\vartheta}$, it follows from a second order Taylor's expansion that

$$\bar{Q}_n(\boldsymbol{\vartheta}) - \bar{Q}_n(\boldsymbol{\vartheta}_0) = (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^T\boldsymbol{\nu} - \frac{1}{2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^T D(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0). \qquad (5.5.1)$$

Here

$$\boldsymbol{\nu} = \begin{pmatrix} \frac{1}{n}\boldsymbol{X}^T(\boldsymbol{Y} - b'(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1)) \\ \frac{1}{n}\boldsymbol{M}^T_{\mathcal{A}}(\boldsymbol{Y} - b'(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1)) - \lambda_n\bar{\rho}(\boldsymbol{\alpha}^\star_{0,\mathcal{A}}) \end{pmatrix},$$

and

$$\begin{aligned} D &= \frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}^T_{\mathcal{A}} \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^*_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^*_1)\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}^T_{\mathcal{A}} \end{pmatrix}^T + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{\Lambda}(\boldsymbol{\vartheta}^*) \end{pmatrix} \\ &=: D_1 + D_2. \end{aligned}$$

Here $\boldsymbol{\vartheta}^* = (\boldsymbol{\alpha}^*_1, \boldsymbol{\alpha}^*_{0,\mathcal{A}})$ and $\boldsymbol{\vartheta}^*$ lies in the line segment jointing $\boldsymbol{\vartheta}$ and $\boldsymbol{\vartheta}_0$. Clearly $\boldsymbol{\vartheta}^* \in \mathcal{N}_0$. $\boldsymbol{\Lambda}(\boldsymbol{\vartheta}^*)$ is a diagonal matrix with nonnegative elements. By condition (A2), the maximum eigenvalue of $\boldsymbol{\Lambda}^*$ is upper bounded by $\lambda_n\kappa_0$.

Let $z(\boldsymbol{\vartheta}) = \boldsymbol{\alpha}_1^T\boldsymbol{x} + \boldsymbol{\alpha}^T_{0,\mathcal{A}}\boldsymbol{m}_{\mathcal{A}}$, $\Sigma_{XX}(\boldsymbol{\vartheta}) = E[\boldsymbol{x}b''(z(\boldsymbol{\vartheta}))\boldsymbol{x}^T]$,
$\Sigma_{XM}(\boldsymbol{\vartheta}) = E[\boldsymbol{x}b''(z(\boldsymbol{\vartheta}))\boldsymbol{m}^T_{\mathcal{A}}]$, $\Sigma_{MM}(\boldsymbol{\vartheta}) = E[\boldsymbol{m}_{\mathcal{A}}b''(z(\boldsymbol{\vartheta}))\boldsymbol{m}^T_{\mathcal{A}}]$, and

$$\Sigma(\boldsymbol{\vartheta}) = \begin{pmatrix} \Sigma_{XX}(\boldsymbol{\vartheta}) & \Sigma_{XM}(\boldsymbol{\vartheta}) \\ \Sigma_{MX}(\boldsymbol{\vartheta}) & \Sigma_{MM}(\boldsymbol{\vartheta}) \end{pmatrix}.$$

Further note that

$$P(\|D_1 - \Sigma(\boldsymbol{\vartheta}^*)\|_2 \geq \epsilon) \leq \frac{1}{\epsilon^2}E[\|D_1 - \Sigma(\boldsymbol{\vartheta}^*)\|_2^2]$$

$$\leq \frac{cn}{\epsilon^2 n^2} E[\sum_{i,j}^{s}[m_{1i}m_{1j}b''(z_1) - E(m_{1i}m_{1j}b''(z_1))]^2$$
$$+ \sum_{i=1}^{s}\sum_{j=1}^{q}[m_{1i}x_{1j}b''(z_1) - E(m_{1i}x_{1j}b''(z_1))]^2$$
$$+ \sum_{i,j}^{q}[x_{1i}x_{1j}b''(z_1) - E(x_{1i}x_{1j}b''(z_1))]^2] = \frac{cs^2}{\epsilon^2 n}.$$

Thus $\|D_1 - \Sigma(\boldsymbol{\vartheta}^*)\|_2 = O_P(s/\sqrt{n}) = o_P(1)$, when $s = o(n^{1/2})$.

Since $\inf_{\boldsymbol{\vartheta}\in\mathcal{N}_0} \lambda_{\min}(\Sigma(\boldsymbol{\vartheta})) \geq c$ and $\lambda_n\kappa_0 = o(1)$,

$$\lambda_{\min}(D) \geq \bar{c} > 0. \tag{5.5.2}$$

Consequently,

$$\max_{\boldsymbol{\vartheta}\in\partial\mathcal{N}_\tau} \bar{Q}_n(\boldsymbol{\vartheta}) - \bar{Q}_n(\boldsymbol{\vartheta}_0) \leq \max_{\boldsymbol{\vartheta}\in\partial\mathcal{N}_\tau} \left(\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 \|\boldsymbol{\nu}\|_2 - \frac{1}{2}\|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2^2 \bar{c} \right)$$
$$= \sqrt{\frac{s}{n}}\tau\|\boldsymbol{\nu}\|_2 - \frac{1}{2}\frac{s}{n}\tau^2\bar{c}.$$

By Markov's inequality, it entails that

$$P(H_n) \geq P(\|\boldsymbol{\nu}\|_2 \leq \frac{1}{2}\sqrt{\frac{s}{n}}\tau\bar{c}) \geq 1 - \frac{4nE\|\boldsymbol{\nu}\|_2^2}{s\tau^2\bar{c}^2}. \tag{5.5.3}$$

In the following, we aim to show that $E\|\boldsymbol{\nu}\|_2^2 = O(s/n)$.

Let $\boldsymbol{\varepsilon}_1 = \boldsymbol{Y} - b'(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1)$. Note that:

$$\boldsymbol{\nu} = \begin{pmatrix} \frac{1}{n}\boldsymbol{X}^T\boldsymbol{\varepsilon}_1 \\ \frac{1}{n}\boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{\varepsilon}_1 \end{pmatrix} - \begin{pmatrix} 0 \\ \lambda_n\bar{\rho}(\boldsymbol{\alpha}^\star_{0,\mathcal{A}}) \end{pmatrix} = \boldsymbol{\nu}_1 - \boldsymbol{\nu}_2,$$

Then by condition (A1),

$$E\|\boldsymbol{\nu}_1\|_2^2 = \frac{1}{n^2} tr\left[E \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{\varepsilon}_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{\varepsilon}_1 \end{pmatrix} \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{\varepsilon}_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{\varepsilon}_1 \end{pmatrix}^T \right]$$

$$
\begin{aligned}
&= \frac{1}{n^2} tr\left[E\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} E[\boldsymbol{\varepsilon}_1 \boldsymbol{\varepsilon}_1^T | \boldsymbol{X}, \boldsymbol{M}] \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T \right] \\
&= \frac{\phi_0}{n^2} tr\left[E\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{\vartheta}_0) \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T \right] \\
&= \frac{\phi_0}{n} tr(\Sigma_0) \leq \phi_0 \frac{s+q}{n} \lambda_{max}(\Sigma_0) = O(\frac{s}{n}).
\end{aligned}
$$

It follows from the concavity of $\rho(\cdot)$, $d_n < |\alpha_{0j,\mathcal{A}}|$, and condition (A2) that:

$$\|\boldsymbol{\nu}_2\|_2^2 \leq (s^{1/2} p_\lambda'(d_n))^2 = o(\frac{1}{n}).$$

Consequently, step 1 is obtained.

Step 2: Sparsity: According to Theorem 1 in Fan and Lv (2011), it suffices to show that with probability tending to 1,

$$\frac{1}{n} \|\boldsymbol{M}_{\mathcal{A}^c}^T (\boldsymbol{Y} - b'(\boldsymbol{M}\bar{\boldsymbol{\alpha}}_0 + \boldsymbol{X}\bar{\boldsymbol{\alpha}}_1))\|_\infty \ll \lambda_n. \tag{5.5.4}$$

Here $\bar{\boldsymbol{\theta}} = (\bar{\boldsymbol{\alpha}}_1^T, \bar{\boldsymbol{\alpha}}_0^T)^T$ satisfies that $\bar{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ and $\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 = O_P(\sqrt{s/n})$. Let $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})$. Note that

$$
\begin{aligned}
& \boldsymbol{M}_{\mathcal{A}^c}^T (\boldsymbol{Y} - b'(\boldsymbol{M}\bar{\boldsymbol{\alpha}}_0 + \boldsymbol{X}\bar{\boldsymbol{\alpha}}_1)) \\
= \; & \boldsymbol{M}_{\mathcal{A}^c}^T \boldsymbol{\varepsilon}_1 - \boldsymbol{M}_{\mathcal{A}^c}^T b''(\boldsymbol{W}\boldsymbol{\vartheta}_0) \boldsymbol{W}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0) + \boldsymbol{R}.
\end{aligned} \tag{5.5.5}
$$

For the second term,

$$\|\boldsymbol{M}_{\mathcal{A}^c}^T b''(\boldsymbol{W}\boldsymbol{\vartheta}_0) \boldsymbol{W}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)\|_\infty \leq \|\boldsymbol{M}_{\mathcal{A}^c}^T b''(\boldsymbol{W}\boldsymbol{\vartheta}_0)\boldsymbol{W}\|_{2,\infty} \|\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0\|_2 = O_P(\sqrt{ns}).$$

While the jth component of $\boldsymbol{R}$, R_j is

$$R_j = (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^T \frac{1}{n} \sum_{i=1}^n \boldsymbol{w}_i m_{ij} b'''(\boldsymbol{w}_i^T \bar{\boldsymbol{\vartheta}}) \boldsymbol{w}_i^T (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0).$$

Here $\bar{\boldsymbol{\vartheta}}$ is between $\boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}$. Under condition A3, we have that $\|\boldsymbol{R}\|_\infty = O_p((s/n)^{1-a}) = o_p(\sqrt{s/n})$.

Similarly to prove that $\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_0)\|_\infty = O_p(\sqrt{\log p/n})$, we can show that $\|\boldsymbol{M}_{\mathcal{A}^c}^T \boldsymbol{\varepsilon}_1\|_\infty = O_p(\sqrt{n \log p})$.

Consequently, given condition A2, step 2 is finished.

Step 3: Asymptotic expansions: Steps 1 and 2 show that $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ with probability 1, and further $\|\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}^\star_{0,\mathcal{A}}\|_2 = O_P(\sqrt{s/n})$.

First we let:

$$\dot{L}(\boldsymbol{\vartheta}_0) \;=\; \begin{pmatrix} \boldsymbol{X}^T(\boldsymbol{Y} - \mu(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1)) \\ \boldsymbol{M}_{\mathcal{A}}^T(\boldsymbol{Y} - \mu(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1)) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{\varepsilon}_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{\varepsilon}_1 \end{pmatrix}. \quad (5.5.6)$$

For $\widehat{\boldsymbol{\vartheta}}$, we let:

$$\dot{L}(\widehat{\boldsymbol{\vartheta}}) \;=\; \begin{pmatrix} \boldsymbol{X}^T(\boldsymbol{Y} - \mu(\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} + \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1)) \\ \boldsymbol{M}_{\mathcal{A}}^T(\boldsymbol{Y} - \mu(\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} + \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1)) \end{pmatrix} = \begin{pmatrix} 0 \\ n\lambda_n \bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}) \end{pmatrix}. \quad (5.5.7)$$

By using second-order Taylor expansion,

$$\dot{L}(\widehat{\boldsymbol{\vartheta}}) = \dot{L}(\boldsymbol{\vartheta}_0) - \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1) \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T (\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) + \boldsymbol{R}.$$

Or equivalently,

$$\begin{aligned} & \frac{1}{\sqrt{n}}(\dot{L}(\boldsymbol{\vartheta}_0) - \dot{L}(\widehat{\boldsymbol{\vartheta}})) \\ = \;& \frac{1}{n} \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^\star_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^\star_1) \begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T \sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) - \frac{\boldsymbol{R}}{\sqrt{n}}. \end{aligned}$$

From the proof in the Step 2, we know that the remainder term $\boldsymbol{R}$ satisfies $\|\boldsymbol{R}\|_\infty = O_p(s(n/s)^a)$. Hence,

$$\|\boldsymbol{R}\|_2 = O_p(s^{3/2}(n/s)^a) = o_p(\sqrt{n}),$$

under the condition that $s = o(n^{(1-2a)/(3-2a)})$.

Note that

$$\|\frac{1}{n}\left(\begin{array}{c}\boldsymbol{X}^T\\ \boldsymbol{M}_{\mathcal{A}}^T\end{array}\right) b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}_{0,\mathcal{A}}^\star + \boldsymbol{X}\boldsymbol{\alpha}_1^\star)\left(\begin{array}{c}\boldsymbol{X}^T\\ \boldsymbol{M}_{\mathcal{A}}^T\end{array}\right)^T - \Sigma_0\| = O_p(\frac{s}{\sqrt{n}}),$$

while $\|\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0\| = O_p(\sqrt{\frac{s}{n}})$. Consequently,

$$\begin{aligned}&\frac{1}{n}\left(\begin{array}{c}\boldsymbol{X}^T\\ \boldsymbol{M}_{\mathcal{A}}^T\end{array}\right) b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}_{0,\mathcal{A}}^\star + \boldsymbol{X}\boldsymbol{\alpha}_1^\star)\left(\begin{array}{c}\boldsymbol{X}^T\\ \boldsymbol{M}_{\mathcal{A}}^T\end{array}\right)^T \sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0)\\ =\ & \Sigma_0\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) + O_p(\sqrt{\frac{s^3}{n}}) = \Sigma_0\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) + o_p(1).\end{aligned}$$

When $s = o(n^{1/3})$.

Thus,

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \ = \ \Sigma_0^{-1}\frac{1}{\sqrt{n}}(\dot{L}(\boldsymbol{\vartheta}_0) - \dot{L}(\widehat{\boldsymbol{\vartheta}})) + o_P(1).$$

Under condition (A2), $\|\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}^\star\|_\infty = O_P(\sqrt{s/n}) \ll d_n$. This implies that

$$\min_{j\in\mathcal{A}}|\widehat{\boldsymbol{\alpha}}_{0j,\mathcal{A}}| > \min_{j\in\mathcal{A}}|\boldsymbol{\alpha}_{0j,\mathcal{A}}^\star| - d_n = d_n.$$

By the concavity of $p(\cdot)$ and condition (A2),

$$\|n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}})\|_2 \le ns^{1/2}p'_{\lambda_n}(d_n) = o(n^{1/2}).$$

Since $\lambda_{\max}(\Sigma_0^{-1}) = O(1)$,

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \ = \ \Sigma_0^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + o_P(1). \tag{5.5.8}$$

Proof of Corollary 2: Recall that

$$\Sigma_0^{-1} = \left(\begin{array}{cc}\Sigma_{XX}^{-1} + \Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1} & -\Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\\ -\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1} & \Sigma_{MM.X}^{-1}\end{array}\right).$$

Here $\Sigma_{MM.X} = \Sigma_{MM} - \Sigma_{MX}\Sigma_{XX}^{-1}\Sigma_{XM}$.

As a result,

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) = (I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + o_P(1). \tag{5.5.9}$$

The asymptotic variance matrix of $\widehat{\boldsymbol{\alpha}}_1$ is

$$\phi_0(I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}(I_{q\times q}, 0_{q\times s})^T = \phi_0\left(\Sigma_{XX}^{-1} + \Sigma_{XX}^{-1}\Sigma_{XM}\Sigma_{MM.X}^{-1}\Sigma_{MX}\Sigma_{XX}^{-1}\right).$$

Recall that

$$\widehat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}. \tag{5.5.10}$$

Consequently,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\varepsilon}_2 + (\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}}\sqrt{n}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}). \tag{5.5.11}$$

We first note that

$$\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \tilde{\Sigma}_{XX}\|_2 = O_p(\frac{1}{\sqrt{n}});\ \lambda_{\max}(\tilde{\Sigma}_{XX}^{-1}) = \lambda_{\min}^{-1}(\tilde{\Sigma}_{XX}) = O(1).$$
$$(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1} - \tilde{\Sigma}_{XX}^{-1} = \tilde{\Sigma}_{XX}^{-1}\left[\tilde{\Sigma}_{XX} - \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right](\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

As a result,

$$\|(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1} - \tilde{\Sigma}_{XX}^{-1}\|_2 = O_p(\frac{1}{\sqrt{n}}).$$

Further note that

$$P(\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XM}\|_2 \geq \epsilon) \leq \frac{1}{\epsilon^2}E[\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XM}\|_2^2]$$
$$\leq \frac{cn}{\epsilon^2 n^2}E\left\{\sum_{i=1}^{s}\sum_{j=1}^{q}[m_{1i}x_{1j} - E(m_{1i}x_{1j})]^2\right\} = \frac{cs}{\epsilon^2 n}.$$

Thus

$$\|\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XM}\|_2 = O_p(\sqrt{\frac{s}{n}}).$$

Consider that

$$\begin{aligned}(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM} &= (\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}[\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XM}] \\ +[(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1} - \tilde{\Sigma}_{XX}^{-1}]\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} &+ [(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1} - \tilde{\Sigma}_{XX}^{-1}][\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XM}].\end{aligned}$$

Since $\|\tilde{\Sigma}_{XX}^{-1}\|_2 = O(1)$, and $\|\tilde{\Sigma}_{XM}\|_2 \leq \|\tilde{\Sigma}_{XM}\|_F = O(\sqrt{s})$,

$$\begin{aligned}\|(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}} - \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}\|_2 &= O_p(1)O_p(\sqrt{\frac{s}{n}}) + O_p(\sqrt{s})O_p(\sqrt{\frac{1}{n}}) \\ &\quad + O_p(\sqrt{\frac{1}{n}})O_p(\sqrt{\frac{s}{n}}).\end{aligned}$$

Further note that $\frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\varepsilon}_2 = O_p(1)$ and $\sqrt{n}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}) = \sqrt{s}$. Then,

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= (\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\varepsilon}_2 + (\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X})^{-1}\frac{1}{n}\boldsymbol{X}^T\boldsymbol{M}_{\mathcal{A}}\sqrt{n}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}) \\ &= \tilde{\Sigma}_{XX}^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\varepsilon}_2 + \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}\sqrt{n}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}_{0,\mathcal{A}}) + o_P(1), \qquad (5.5.12)\end{aligned}$$

under the condition that $s = o(n^{1/2})$. From equation (5.5.8),

$$\begin{aligned}&\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= \tilde{\Sigma}_{XX}^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\varepsilon}_2 + \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}(0_{s\times q}, I_s)\Sigma_0^{-1}\frac{1}{\sqrt{n}}\begin{pmatrix}\boldsymbol{X}^T\boldsymbol{\varepsilon}_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{\varepsilon}_1\end{pmatrix} + o_p(1) \\ &= \tilde{\Sigma}_{XX}^{-1}\frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\varepsilon}_2 + \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}\Sigma_{MM.X}^{-1}\frac{1}{\sqrt{n}}(\boldsymbol{M}_{\mathcal{A}}^T - \Sigma_{MX}\Sigma_{XX}^{-1})\boldsymbol{\varepsilon}_1 + o_p(1) \\ &= \tilde{\Sigma}_{XX}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n W_{1i} + \tilde{\Sigma}_{XX}^{-1}\tilde{\Sigma}_{XM}\Sigma_{MM.X}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n W_{2i} + o_p(1). \qquad (5.5.13)\end{aligned}$$

Here $W_{1i} = \boldsymbol{x}_i\varepsilon_{2i}$ and $W_{2i} = (\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{x}_i)\varepsilon_{1i}$.

It is easy to know that $E[W_{1i}] = E[\boldsymbol{x}_i E(\varepsilon_{2i}|\boldsymbol{x}_i)] = 0$. Similarly, $E[W_{2i}] = E[(\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX}\Sigma_{XX}^{-1}\boldsymbol{x}_i)E(\varepsilon_{1i}|\boldsymbol{x}_i, \boldsymbol{m}_{i,\mathcal{A}})] = 0$.

Further, $var(W_{1i}) = \sigma_2^2 \Sigma_{XX}$, $var(W_{2i}) = \phi_0 \Sigma_{MM.X}$, and

$$\begin{aligned} cov(W_{1i}, W_{2i}) &= E[\boldsymbol{x}_i \varepsilon_{2i} (\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX} \Sigma_{XX}^{-1} \boldsymbol{x}_i) \varepsilon_{1i}] \\ &= E[\boldsymbol{x}_i \varepsilon_{2i} (\boldsymbol{m}_{i,\mathcal{A}} - \Sigma_{MX} \Sigma_{XX}^{-1} \boldsymbol{x}_i) E(\epsilon_{1i} | \boldsymbol{x}_i, \boldsymbol{m}_{i,\mathcal{A}}, \varepsilon_{2i})] = 0. \end{aligned}$$

As a result, we get that:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, \sigma_2^2 \tilde{\Sigma}_{XX}^{-1} + \phi_0 \tilde{\Sigma}_{XX}^{-1} \tilde{\Sigma}_{XM} \Sigma_{MM.X}^{-1} \tilde{\Sigma}_{MX} \tilde{\Sigma}_{XX}^{-1}). \tag{5.5.14}$$

Proof of Theorem 3: We first obtain the asymptotic expansion of $\widehat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}$ and show that $\widehat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}} = O_p(\sqrt{1/n})$.

Similar to the arguments for Theorem 1, we can also show that $\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ with probability 1, and further $\|\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \boldsymbol{\alpha}^\star_{0,\mathcal{A}}\|_2 = O_P(\sqrt{s/n})$. For $\tilde{\boldsymbol{\vartheta}}$,

$$\dot{L}(\tilde{\boldsymbol{\vartheta}}) = \begin{pmatrix} \boldsymbol{X}^T(\boldsymbol{Y} - \mu(\boldsymbol{M}_{\mathcal{A}} \tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}} + \boldsymbol{X} \tilde{\boldsymbol{\alpha}}_1)) \\ \boldsymbol{M}_{\mathcal{A}}^T(\boldsymbol{Y} - \mu(\boldsymbol{M}_{\mathcal{A}} \tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}} + \boldsymbol{X} \tilde{\boldsymbol{\alpha}}_1)) \end{pmatrix} = \begin{pmatrix} \sqrt{n}\boldsymbol{v} \\ n\lambda_n \bar{\rho}(\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}}) \end{pmatrix} \tag{5.5.15}$$

Similar to the argument for $\widehat{\boldsymbol{\vartheta}}$,

$$\begin{aligned} \sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) &= \Sigma_0^{-1} \frac{1}{\sqrt{n}} (\dot{L}(\boldsymbol{\vartheta}_0) - \dot{L}(\tilde{\boldsymbol{\vartheta}})) \\ &= \Sigma_0^{-1} \frac{1}{\sqrt{n}} \dot{L}(\boldsymbol{\vartheta}_0) - \Sigma_0^{-1} \begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix} \boldsymbol{v} + o_P(1). \end{aligned}$$

Recall that $\tilde{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star = 0 - \boldsymbol{h_n} = -\boldsymbol{h_n}$. Then,

$$\begin{aligned} -\sqrt{n}\boldsymbol{h_n} &= \sqrt{n}(\tilde{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1^\star) = (I_{q\times q}, 0_{q\times s}) \sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) \\ &= (I_{q\times q}, 0_{q\times s}) \Sigma_0^{-1} \frac{1}{\sqrt{n}} \dot{L}(\boldsymbol{\vartheta}_0) - (I_{q\times q}, 0_{q\times s}) \Sigma_0^{-1} \begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix} \boldsymbol{v} + o_P(1). \end{aligned}$$

Let

$$\Phi = (I_{q\times q}, 0_{q\times s}) \Sigma_0^{-1} \begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix}.$$

Under condition (A1), $\lambda_{\max}(\Sigma_0) = O(1)$. This implies that $\lambda_{\min}(\Sigma_0^{-1}) > 0$, and then $\lambda_{\min}(\Phi) > 0$. Finally, $\lambda_{\max}(\Phi^{-1}) = O(1)$.

Then we obtain that:

$$\boldsymbol{v} = \Phi^{-1}(I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + \sqrt{n}\Phi^{-1}\boldsymbol{h_n} + o_P(1).$$

Consequently,

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) &= \Sigma_0^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) - \Sigma_0^{-1}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q}\end{pmatrix}\Phi^{-1}(I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) \\ &\quad -\Sigma_0^{-1}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q}\end{pmatrix}\sqrt{n}\Phi^{-1}\boldsymbol{h_n} + o_P(1). \qquad (5.5.16)\end{aligned}$$

Or equivalently

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}_0) &= \Sigma_0^{-1/2}(I - P_n)\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) \\ &\quad -\Sigma_0^{-1}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q}\end{pmatrix}\sqrt{n}\Phi^{-1}\boldsymbol{h_n} + o_p(1). \qquad (5.5.17)\end{aligned}$$

Here

$$P_n = \Sigma_0^{-1/2}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q}\end{pmatrix}\Phi^{-1}(I_{q\times q}, 0_{q\times s})\Sigma_0^{-1/2}.$$

It is easy to see that P_n is an idempotent matrix with rank q.

From the asymptotic expansions of $\widehat{\boldsymbol{\vartheta}}$ and $\tilde{\boldsymbol{\vartheta}}$ in equations (5.5.8) and (5.5.17),

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}) &= \Sigma_0^{-1/2}P_n\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) \\ &\quad +\Sigma_0^{-1}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q}\end{pmatrix}\sqrt{n}\Phi^{-1}\boldsymbol{h_n} + o_P(1). \qquad (5.5.18)\end{aligned}$$

Recall that

$$\dot{L}(\boldsymbol{\vartheta}_0) = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{\varepsilon}_1 \\ \boldsymbol{M}_{\mathcal{A}}^T\boldsymbol{\varepsilon}_1\end{pmatrix}. \qquad (5.5.19)$$

Then we can get:

$$\frac{1}{\phi_0}E\|P_n\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0)\|_2^2 = tr(P_n\Sigma_0^{-1/2}\Sigma_0\Sigma_0^{-1/2}P_n) = tr(P_n) = rank(P_n) = q.$$

It follows that:

$$E\|\Sigma_0^{-1/2}P_n\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0)\|_2^2 \leq \|\Sigma_0^{-1/2}\|_2^2 E\|P_n\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0)\|_2^2 = O(1).$$

Consequently, under assumption (A4),

$$\sqrt{n}(\widehat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}) = O_p(1). \tag{5.5.20}$$

Now we are ready to investigate the asymptotic distribution of the likelihood ratio test. Under the event $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = \tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$ and recalling equation (5.5.7), we obtain that

$$\begin{aligned}
& n(L_n(\tilde{\boldsymbol{\theta}}) - L_n(\widehat{\boldsymbol{\theta}})) \\
= \; & (\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}})^T n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}) \\
& -\frac{1}{2}\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})^T\frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} + \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1)\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}}) \\
& +(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})^T\boldsymbol{R}.
\end{aligned} \tag{5.5.21}$$

From the proof of step 2 in theorem 1, $\|\boldsymbol{R}\|_\infty = O_p(s(n/s)^a)$. As a result,

$$|(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})^T\boldsymbol{R}| \leq \|\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}}\|_2\|\boldsymbol{R}\|_2 = O_p(\frac{1}{\sqrt{n}}s(n/s)^a\sqrt{s}) = o_p(1),$$

when $s = o(n^{(1-2a)/(3-2a)})$. Similarly we can show that

$$\begin{aligned}
& \|(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})^T\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}} + \boldsymbol{X}\widehat{\boldsymbol{\alpha}}_1)\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}}) \\
& -(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})^T\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}_{0,\mathcal{A}}^\star + \boldsymbol{X}\boldsymbol{\alpha}_1^\star)\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})\|_2 = o_p(1).
\end{aligned}$$

From equation (5.5.20) and $n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}) = o_P(n^{1/2})$ based on assumption (A2),

$$(\tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}} - \widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}})^T n\lambda_n\bar{\rho}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}) = o_p(1).$$

Further note that $\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}}) = O_p(1)$ and

$$\|\frac{1}{n}\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix} b''(\boldsymbol{M}_{\mathcal{A}}\boldsymbol{\alpha}^{\star}_{0,\mathcal{A}} + \boldsymbol{X}\boldsymbol{\alpha}^{\star}_1)\begin{pmatrix} \boldsymbol{X}^T \\ \boldsymbol{M}_{\mathcal{A}}^T \end{pmatrix}^T - \Sigma_0\|_2 = o_p(1).$$

Consequently,

$$\begin{aligned} & n(L_n(\tilde{\boldsymbol{\theta}}) - L_n(\widehat{\boldsymbol{\theta}})) = -\frac{1}{2}\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}})^T\Sigma_0\sqrt{n}(\tilde{\boldsymbol{\vartheta}} - \widehat{\boldsymbol{\vartheta}}) + o_p(1) \\ = \; & -\frac{1}{2}\|P_n\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + \Sigma_0^{-1/2}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix}\sqrt{n}\Phi^{-1}\boldsymbol{h_n}\|_2^2 + o_P(1) \quad (5.5.22) \end{aligned}$$

The last equation holds due to equation (5.5.18). Recall that

$$P_n = \Sigma_0^{-1/2}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix}\Phi_0^{-1}(I_{q\times q}, 0_{q\times s})\Sigma_0^{-1/2}.$$

Further denote that

$$\boldsymbol{\omega}_n = (I_{q\times q}, 0_{q\times s})\Sigma_0^{-1}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0).$$

$$\begin{aligned} & \|P_n\Sigma_0^{-1/2}\frac{1}{\sqrt{n}}\dot{L}(\boldsymbol{\vartheta}_0) + \Sigma_0^{-1/2}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix}\sqrt{n}\Phi^{-1}\boldsymbol{h_n}\|_2^2 \\ = \; & \|\Sigma_0^{-1/2}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix}\Phi^{-1}\boldsymbol{\omega}_n + \Sigma_0^{-1/2}\begin{pmatrix} I_{q\times q} \\ 0_{s\times q} \end{pmatrix}\sqrt{n}\Phi^{-1}\boldsymbol{h_n}\|_2^2 \\ = \; & \|\Phi^{-1/2}\boldsymbol{\omega}_n + \sqrt{n}\Phi^{-1/2}\boldsymbol{h_n}\|_2^2. \quad (5.5.23) \end{aligned}$$

It is easy to know that $\Phi^{-1/2}\boldsymbol{\omega}_n \sim N(0, \phi_0 I_{q\times q})$. Since $\widehat{\phi}_0 = \phi_0 + o_p(1)$. Based on Slusky theorem,

$$\sup_x |P(T_n \le x) - P(\chi_q^2(n\boldsymbol{h}_n^T\Phi^{-1}\boldsymbol{h}_n/\phi_0) \le x)| \to 0. \quad (5.5.24)$$

The consistency of $\widehat{\sigma}_2^2$. We give a short proof for the consistency of $\widehat{\sigma}_2^2$. Under the event $\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = \tilde{\boldsymbol{\alpha}}_{0,\mathcal{A}^c} = 0$, we have with probability 1,

$$\widehat{\sigma}_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^T \boldsymbol{m}_{i,\mathcal{A}} - \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i)^2.$$

Note that

$$\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^T \boldsymbol{m}_{i,\mathcal{A}} - \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_i = \boldsymbol{\alpha}_{0,\mathcal{A}}^T \boldsymbol{m}_{i,\mathcal{A}} - \boldsymbol{\beta}^T \boldsymbol{x}_i + (\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^T - \boldsymbol{\alpha}_{0,\mathcal{A}}^T)^T \boldsymbol{m}_{i,\mathcal{A}} + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^T \boldsymbol{x}_i.$$

Thus

$$\begin{aligned}\widehat{\sigma}_2^2 &= \frac{1}{n-1}\sum_{i=1}^{n}[\varepsilon_{2i} + \{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}), (\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^T - \boldsymbol{\alpha}_{0,\mathcal{A}}^T)\}^T (\boldsymbol{x}_i^T, \boldsymbol{m}_{i,\mathcal{A}}^T)^T\}]^2 \\ &= I_1 + 2I_2 + I_3.\end{aligned}$$

It is easy to know that $I_1 \to \sigma_2^2$, while $I_3 \leq \|\{(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}), (\widehat{\boldsymbol{\alpha}}_{0,\mathcal{A}}^T - \boldsymbol{\alpha}_{0,\mathcal{A}}^T)\}\|_2^2 \|D_3\|_2 = O_p(s/n) = o_p(1)$. Here $D_3 = n^{-1}(\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})^T(\boldsymbol{X}, \boldsymbol{M}_{\mathcal{A}})$. By Cauchy-Schwarz inequality, I_2 is also $o_p(1)$. Thus we conclude that $\widehat{\sigma}_2^2 = \sigma_2^2 + o_p(1)$.

Chapter 6

Future Works And Conclusion

In this chapter, we will discuss potential future works and conclude the results of high-dimensional mediation.

6.1 Future works

There are two main tracks for future works, model application and theoretical development. For real world application, mediation analysis are widely used in nature science and social science, where it is critical to understand the mechanism between the exposure variables and response variable. For example, environmental stress such as drought alters gene expression and then has adverse effect on plant growth and development. Scientists can use mediation analysis find out the important genes that regulate the produce of protein and help the plants survive in the drought condition. For developing new statistical procedures, current works can be extended to high dimensional survival model for analyzing the expected duration of time until one event occurs. Survival mediation model can have huge potential in engineering and insurance industry. For example, engineering experts can use mediation analysis to understand the underlying factors resulting in failure in mechanical systems, and actuaries may be able to use mediation model to understand the factors contributing to the risk of insured person death.

6.1.1 Drought stress impact on plant yield study

Environmental stress such as drought has adverse effect on plant growth and development, one common scenario is drought. Drought causes plants defects by losing cellular water, decreasing enzyme activities, and reducing corp yield (Zhu, 2002). Su et al. (2013) studied Arabidopsis thaliana plants development under drought during flowering, and analyzed differential expression of certain genes compared with well-water treatment group. Ma et al. (2014) compared the phenotype and transcriptomes under several drought conditions and found that plants would activate different pathway to acclimate different drought stress. However, the biological pathway from drought stress to crop yield reduction is complex. We may apply HDMM to select important genes regulating the reproduction activities under different drought stress and conduct statistical inference for both indirect and direct effects.

The independent variable of the mediation model is the water content in the soil and the potential mediators are gene expression measured by RNA-seq. The response variables are a group of certain gene expression of interest. To be specific, if the number of genes of interest is small, for example less than five genes, we can build mediation model for each gene. If there are a few number of genes of interest, for example between five to thirty genes, we can let outcome variable be the count number of detected genes expression, and then fit a binomial mediation model. If the number of genes of interest is large, for example greater than 30 genes, we can similarly define the outcome variable but fit a Poisson mediation model. When the number of genes in a binomial distribution is large and the value of probability to detect a gene's expression is very small, the binomial distribution can be approximated by a Poisson distribution. These models will be helpful to find the genes that are activated in the early days of drought treatment, which act as mediators influence subsequent reproductive development.

The analysis will have two steps: 1. Feature screening. This step is to quickly and coarsely screen out insignificant genes that associate with plants' biological function in the drought in order to reduce the dimension from twenty thousands to a few thousands. We will use sure independence screening (SIS) (Fan and Lv 2008) to selects variables based on marginal utilities such as their marginal correlations with the response variable. With probability tending to one, the mediators selected

by SIS would not miss any truly important mediator, and hence the false negative rate can be controlled. 2. Variable selection. This step is to further finely select significant genes based on the set of genes screened in Step 1. In this step, we will use penalized method for example SCAD penalty introduced by Fan and Li (2001) to select important genes and conduct hypothesis testing for direct and indirect effect. Also, the order of significant gene will be provided.

6.1.2 Survival mediation analysis

Survival analysis captures the total effect of a subject exposure on a time-to-event outcome (Kleinbaum and Klein, 2010), while survival mediation analysis can further decompose this effect into its direct and indirect pathways, mediated by variables observed after the exposure prior to the outcome (VanderWeele, 2011). Survival mediation analysis has been used in various scientific research, such as socioeconomics, epidemiology, and so on. For example, in social science, it has been a heated research topic to understand the relationship between socioeconomic status (SES) and mental health. SES is commonly defined as from three dimensions: family income, parents' education level, and parents' occupational prestige. Christensen et al. (2008) study the effect of SES on the risk of experiencing long-term sickness absence, which seems to be partially mediated via physical working environment, but their work is lack of quantitative analysis. To address this problem, Lange and Hansen (2011) use an additive hazard model to analyze direct and indirect effects for time-to-event data but this approach is confined to additive hazard. In sight of this issue, (VanderWeele, 2011) extend the Lange and Hansen (2011)'s work to accommodate proportional hazards with a rare outcome or accelerated failure time models. Wang and Albert (2017) develop procedures for the Cox mediation model with a smooth baseline hazard estimator. There are also other statistical methods for mediation model with low-dimensional mediators (Gelfand et al., 2016; Yu et al., 2019; Cho and Huang, 2019), however, little work has been done for high-dimensional survival mediation model. To our best knowledge, Luo et al. (2020) is the first work toward high-dimensional survival mediator selection. Luo et al. (2020) adapted sure independent screening and minimax concave penalty techniques for variable selection, but lack of theoretical justification

of proposed test statistics. It is promising to adapt our proposed procedures to develop theoretical results for estimation and inference high-dimensional survival mediation models.

6.2 Conclusion

We propose statistical inference procedures for the indirect effects in high dimensional mediation model. We introduce a partial penalized least squares method and study its statistical properties under random design. We show that the proposed estimators are more efficient than existing ones. We further propose a partial penalized Wald test to detect the indirect effect, with a χ^2 limiting null distribution. We also propose an F-type test for the direct effect and reveal Wilks phenomenon in the high-dimensional mediation model. We further utilize the proposed inference procedures to analyze the mediation effects of various financial metrics on the relationship between company's sector and the stock return. In addition, we extend our methods to models with confounding variables and apply the method to analyze the mediator role of DNA methylation bridge childhood trauma and cortisol level. Finally, we extend the method to generalized linear mediation model and use the method to classify valuable stocks when COVID-19 outbreak, which is helpful to understand the key financial metrics for stock selection in the future pandemic.

Bibliography

Aguado, B. and Campbell, R. D. (1998). Characterization of a human lysophosphatidic acid acyltransferase that is encoded by a gene located in the class iii region of the human major histocompatibility complex. *Journal of Biological Chemistry*, 273(7):4096–4105.

Ai, C., Linton, O., and Zhang, Z. (2021). Estimation and inference for the counterfactual distribution and quantile functions in continuous treatment models. *Journal of Econometrics*.

Aschbacher, K., O'Donovan, A., Wolkowitz, O. M., Dhabhar, F. S., Su, Y., and Epel, E. (2013). Good stress, bad stress and oxidative stress: insights from anticipatory cortisol reactivity. *Psychoneuroendocrinology*, 38(9):1698–1708.

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society, Series B*, 80:597–623.

Baker, S. R., Bloom, N., Davis, S. J., Kost, K., Sammon, M., and Viratyosin, T. (2020). The unprecedented stock market reaction to covid-19. *The Review of Asset Pricing Studies*, 10(4):742–758.

Barfield, R., Shen, J., Just, A. C., Vokonas, P. S., Schwartz, J., Baccarelli, A. A., VanderWeele, T. J., and Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic Epidemiology*, 41(8):824–833.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Bollen, K. A. and Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, pages 115–140.

Bremner, J. D., Vythilingam, M., Vermetten, E., Adil, J., Khan, S., Nazeer, A., Afzal, N., McGlashan, T., Elzinga, B., Anderson, G. M., et al. (2003). Cortisol response to a cognitive stress challenge in posttraumatic stress disorder (ptsd) related to childhood abuse. *Psychoneuroendocrinology*, 28(6):733–750.

Broadstock, D. C., Chan, K., Cheng, L. T., and Wang, X. (2021). The role of esg performance during times of financial crisis: Evidence from covid-19 in china. *Finance Research Letters*, 38:101716.

Burke, H. M., Davis, M. C., Otte, C., and Mohr, D. C. (2005). Depression and cortisol responses to psychological stress: a meta-analysis. *Psychoneuroendocrinology*, 30(9):846–856.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ 1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Callen, J. L. and Segal, D. (2004). Do accruals drive firm-level stock returns? a variance decomposition analysis. *Journal of Accounting Research*, 42(3):527–560.

Carpenter, L. L., Carvalho, J. P., Tyrka, A. R., Wier, L. M., Mello, A. F., Mello, M. F., Anderson, G. M., Wilkinson, C. W., and Price, L. H. (2007). Decreased adrenocorticotropic hormone and cortisol responses to stress in healthy adults reporting significant childhood maltreatment. *Biological Psychiatry*, 62(10):1080–1087.

Carpenter, L. L., Shattuck, T. T., Tyrka, A. R., Geracioti, T. D., and Price, L. H. (2011). Effect of childhood physical abuse on cortisol stress response. *Psychopharmacology*, 214(1):367–375.

Chakrabortty, A., Nandy, P., and Li, H. (2018). Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*.

Chandrima, R. M., Kircaburun, K., Kabir, H., Riaz, B. K., Kuss, D. J., Griffiths, M. D., and Mamun, M. A. (2020). Adolescent problematic internet use and parental mediation: A bangladeshi structured interview study. *Addictive Behaviors Reports*, 12:100288.

Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136.

Cho, S.-H. and Huang, Y.-T. (2019). Mediation analysis with causally ordered mediators using cox proportional hazards model. *Statistics in medicine*, 38(9):1566–1581.

Christensen, K. B., Labriola, M., Lund, T., and Kivimäki, M. (2008). Explaining the social gradient in long-term sickness absence: a prospective study of danish employees. *Journal of Epidemiology & Community Health*, 62(2):181–183.

Conti, G., Heckman, J. J., and Pinto, R. (2016). The effects of two influential early childhood interventions on health and healthy behaviour. *The Economic Journal*, 126(596):F28–F65.

De Vito, A. and Gomez, J.-P. (2020). Estimating the covid-19 cash crunch: Global evidence and policy. *Journal of Accounting and Public Policy*, 39(2):106741.

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

Djordjilović, V., Page, C. M., Gran, J. M., Nøst, T. H., Sandanger, T. M., Veierød, M. B., and Thoresen, M. (2019). Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in Medicine*, 38(18):3346–3360.

Donald, S. G. and Hsu, Y.-C. (2014). Estimation and inference for distribution functions and quantile functions in treatment effect models. *Journal of Econometrics*, 178:383–397.

Edelman, S., Shalev, I., Uzefovsky, F., Israel, S., Knafo, A., Kremer, I., Mankuta, D., Kaitz, M., and Ebstein, R. P. (2012). Epigenetic and genetic factors predict women's salivary cortisol following a threat to the social self. *PloS One*, 7(11):e48597.

Edirisinghe, N. C. and Zhang, X. (2007). Generalized dea model of fundamental analysis and its application to portfolio optimization. *Journal of Banking & Finance*, 31(11):3311–3335.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Elzinga, B. M., Spinhoven, P., Berretty, E., de Jong, P., and Roelofs, K. (2010). The role of childhood abuse in hpa-axis reactivity in social anxiety disorder: A pilot study. *Biological Psychology*, 83(1):1–6.

Enke, D. and Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 29(4):927–940.

Esubalew, A. A. and Raghurama, A. (2020). The mediating effect of entrepreneurs' competency on the relationship between bank finance and performance of micro,

small, and medium enterprises (msmes). *European Research on Management and Business Economics*, 26(2):87–95.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):37–65.

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

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020a). *Statistical foundations of data science*. Chapman and Hall/CRC.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fan, J., Xue, L., and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Annals of Statistics*, 42(3):819.

Fan, Y., Demirkaya, E., Li, G., and Lv, J. (2019). Rank: large-scale inference with graphical nonlinear knockoffs. *Journal of the American Statistical Association*.

Fan, Y., Lv, J., Sharifvaghefi, M., and Uematsu, Y. (2020b). Ipad: stable interpretable forecasting with knockoffs inference. *Journal of the American Statistical Association*, 115(532):1822–1834.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society: Series A (General)*, 120(3):253–281.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Gelfand, L. A., MacKinnon, D. P., DeRubeis, R. J., and Baraldi, A. N. (2016). Mediation analysis with survival outcomes: accelerated failure time vs. proportional hazards models. *Frontiers in psychology*, 7:423.

Gonzalez-Bono, E., Rohleder, N., Hellhammer, D. H., Salvador, A., and Kirschbaum, C. (2002). Glucose but not protein or fat load amplifies the cortisol response to psychosocial stress. *Hormones and Behavior*, 41(3):328–333.

Gormsen, N. J. and Koijen, R. S. (2020). Coronavirus: Impact on stock prices and growth expectations. *The Review of Asset Pricing Studies*, 10(4):574–597.

Hassan, T. A., Hollander, S., Van Lent, L., Schwedeler, M., and Tahoun, A. (2020). Firm-level exposure to epidemic diseases: Covid-19, sars, and h1n1. Technical report, National Bureau of Economic Research.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications.

Heim, C., Newport, D. J., Heit, S., Graham, Y. P., Wilcox, M., Bonsall, R., Miller, A. H., and Nemeroff, C. B. (2000). Pituitary-adrenal and autonomic responses to stress in women after sexual and physical abuse in childhood. *Journal of American Medical Association*, 284(5):592–597.

Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., Van Lier, P. A., Meeus, W., Branje, S., Heim, C. M., Nemeroff, C. B., and Mill, J. (2016). Genome-wide dna methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature Communications*, 7(1):1–10.

Huang, Y., Capretz, L. F., and Ho, D. (2019). Neural network models for stock selection based on fundamental analysis. In *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pages 1–4. IEEE.

Huang, Y.-T. (2019). Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics*, 13(1):60–84.

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413.

Huang, Y.-T. and Yang, H.-I. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology (Cambridge, Mass.)*, 28(3):370.

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

Hyman, M. (1955). An experimental study of artificial-larynx and esophageal speech. *Journal of Speech and Hearing Disorders*, 20(3):291–299.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309.

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

Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*, volume 3. Springer.

Kraft, A. J. and Luecken, L. J. (2009). Childhood parental divorce and cortisol in young adulthood: evidence for mediation by family income. *Psychoneuroendocrinology*, 34(9):1363–1369.

Kuo, L. E., Kitlinska, J. B., Tilan, J. U., Li, L., Baker, S. B., Johnson, M. D., Lee, E. W., Burnett, M. S., Fricke, S. T., Kvetnansky, R., et al. (2007). Neuropeptide y acts directly in the periphery on fat tissue and mediates stress-induced obesity and metabolic syndrome. *Nature Medicine*, 13(7):803–811.

Lange, T. and Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology*, pages 575–581.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lee, T. K., Cho, J. H., Kwon, D. S., and Sohn, S. Y. (2019). Global stock market investment strategies based on financial network indicators using machine learning techniques. *Expert Systems with Applications*, 117:228–242.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lei, X. and Wong, G. W. (2019). C1q/tnf-related protein 2 (ctrp2) deletion promotes adipose tissue lipolysis and hepatic triglyceride secretion. *Journal of Biological Chemistry*, 294(43):15638–15649.

Liao, Z.-Z., Wang, Y.-D., Qi, X.-Y., and Xiao, X.-H. (2019). Jazf1, a relevant metabolic regulator in type 2 diabetes. *Diabetes/metabolism Research and Reviews*, 35(5):e3148.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616.

Luecken, L. J. (1998). Childhood attachment and loss experiences affect adult cardiovascular and cortisol function. *Psychosomatic Medicine*, 60(6):765–772.

Luo, C., Fa, B., Yan, Y., Wang, Y., Zhou, Y., Zhang, Y., and Yu, Z. (2020). High-dimensional mediation analysis in survival models. *PLoS computational biology*, 16(4):e1007768.

Ma, X., Sukiran, N. L., Ma, H., and Su, Z. (2014). Moderate drought causes dramatic floral transcriptomic reprogramming to ensure successful reproductive development in arabidopsis. *BioMed Central Plant Biology*, 14(1):164.

MacCorquodale, K. and Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55(2):95.

MacKinnon, D. (2012). *Introduction to statistical mediation analysis*. Routledge.

McCullagh, P. and Nelder, J. (1989). Generalized linear models., 2nd edn.(chapman and hall: London). *Standard Book on Generalized Linear Models*.

McGowan, P. O., Sasaki, A., D'alessio, A. C., Dymov, S., Labonté, B., Szyf, M., Turecki, G., and Meaney, M. J. (2009). Epigenetic regulation of the glucocorticoid receptor in human brain associates with childhood abuse. *Nature Neuroscience*, 12(3):342–348.

Nguyen, Q. C., Osypuk, T. L., Schmidt, N. M., Glymour, M. M., and Tchetgen Tchetgen, E. J. (2015). Practical guidance for conducting mediation analysis with multiple mediators using inverse odds ratio weighting. *American Journal of Epidemiology*, 181(5):349–356.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Ottaviani, C. and Vandone, D. (2018). Financial literacy, debt burden and impulsivity: A mediation analysis. *Economic Notes: Review of Banking, Finance and Monetary Economics*, 47(2-3):439–454.

Peng, W., Li, D., Li, D., Jia, J., Wang, Y., and Sun, W. (2019). School disconnectedness and adolescent internet addiction: Mediation by self-esteem and moderation by emotional intelligence. *Computers in Human Behavior*, 98:111–121.

Perroud, N., Paoloni-Giacobino, A., Prada, P., Olié, E., Salzmann, A., Nicastro, R., Guillaume, S., Mouthon, D., Stouder, C., Dieben, K., et al. (2011). Increased methylation of glucocorticoid receptor gene (nr3c1) in adults with a history of childhood maltreatment: a link with the severity and type of trauma. *Translational Psychiatry*, 1(12):e59–e59.

Pesonen, A.-K., Räikkönen, K., Feldt, K., Heinonen, K., Osmond, C., Phillips, D. I., Barker, D. J., Eriksson, J. G., and Kajantie, E. (2010). Childhood separation experience predicts hpa axis hormonal responses in late adulthood: a natural experiment of world war ii. *Psychoneuroendocrinology*, 35(5):758–767.

Peter, M. S. (2011). The role of thyroid hormones in stress response of fish. *General and Comparative Endocrinology*, 172(2):198–210.

Petrowski, K., Wintermann, G.-B., Schaarschmidt, M., Bornstein, S. R., and Kirschbaum, C. (2013). Blunted salivary and plasma cortisol response in patients with panic disorder under psychosocial stress. *International Journal of Psychophysiology*, 88(1):35–39.

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

Preacher, K. J. and Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3):879–891.

Ramelli, S. and Wagner, A. F. (2020). Feverish stock price reactions to covid-19. *The Review of Corporate Finance Studies*, 9(3):622–655.

Ravikumar, B., Imarisio, S., Sarkar, S., O'Kane, C. J., and Rubinsztein, D. C. (2008). Rab5 modulates aggregation and toxicity of mutant huntingtin through macroautophagy in cell and fly models of huntington disease. *Journal of Cell Science*, 121(10):1649–1660.

Serang, S., Jacobucci, R., Brimhall, K. C., and Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(5):733–744.

Shi, C., Song, R., Chen, Z., and Li, R. (2019). Linear hypothesis testing for high dimensional generalized linear models. *The Annals of Statistics*, 47(5):2671–2703.

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

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L., Roux, A. V. D., Needham, B. L., Smith, J. A., and Mukherjee, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76(3):700–710.

Su, Z., Ma, X., Guo, H., Sukiran, N. L., Guo, B., Assmann, S. M., and Ma, H. (2013). Flower development under drought stress: morphological and transcriptomic analyses reveal acute responses and long-term acclimation in arabidopsis. *The Plant Cell*, 25(10):3785–3807.

Sun, T. and Zhang, C.-H. (2010). Comments on: ℓ_1-penalization for mixture regression models. *Test*, 19(2):270–275.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Thorbecke, W. (2020). The impact of the covid-19 pandemic on the us economy: evidence from the stock market. *Journal of Risk and Financial Management*, 13(10):233.

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

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Valeri, L. and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with sas and spss macros. *Psychological Methods*, 18(2):137–150.

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

van Kesteren, E.-J. and Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(5):710–723.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass.)*, 22(4):582.

Vicennati, V., Pasqui, F., Cavazza, C., Pagotto, U., and Pasquali, R. (2009). Stress-related development of obesity and cortisol in women. *Obesity*, 17(9):1678–1683.

Vinkers, C. H., Kalafateli, A. L., Rutten, B. P., Kas, M. J., Kaminsky, Z., Turner, J. D., and Boks, M. P. (2015). Traumatic stress and human dna methylation: a critical review. *Epigenomics*, 7(4):593–608.

Wang, L., Kim, Y., and Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics*, 41(5):2505.

Wang, L., Wu, Y., and Li, R. (2012). Quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of the American Statistical Association*, 107(497):214–222.

Wang, W. and Albert, J. M. (2017). Causal mediation analysis for the cox proportional hazards model with a smooth baseline hazard estimator. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 66(4):741.

Wojtas, B., Pfeifer, A., Oczko-Wojciechowska, M., Krajewska, J., Czarniecka, A., Kukulska, A., Eszlinger, M., Musholt, T., Stokowy, T., Swierniak, M., et al. (2017). Gene expression (mrna) markers for differentiating between malignant and benign follicular thyroid tumours. *International Journal of Molecular Sciences*, 18(6):1184.

Xu, M., Wang, Y., and Tu, Y. (2021). Uncovering the invisible effect of air pollution on stock returns: A moderation and mediation analysis. *Finance Research Letters*, 39:101646.

Yousfi, M., Dhaoui, A., and Bouzgarrou, H. (2021). Risk spillover during the covid-19 global pandemic and portfolio management. *Journal of Risk and Financial Management*, 14(5):222.

Yu, Q., Wu, X., Li, B., and Scribner, R. A. (2019). Multiple mediation analysis with survival outcomes: with an application to explore racial disparity in breast cancer survival. *Statistics in medicine*, 38(3):398–412.

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

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

Zhang, D., Hu, M., and Ji, Q. (2020). Financial markets under the global pandemic of covid-19. *Finance Research Letters*, 36:101528.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., and Colicino, E. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154.

Zhao, Y., Lindquist, M. A., and Caffo, B. S. (2020). Sparse principal component based high-dimensional mediation analysis. *Computational Statistics & Data Analysis*, 142:106835.

Zhou, R. R., Wang, L., and Zhao, S. D. (2020). Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3):573–589.

Zhu, J.-K. (2002). Salt and drought stress signal transduction in plants. *Annual review of plant biology*, 53(1):247–273.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509.

Vita

Mudong Zeng

EDUCATION

Pennsylvania State University *University Park, PA*
Ph.D. in Statistics, GPA:3.94/4.00 Advisor: Runze Li. Aug 2017-Dec 2021
Fudan University *Shanghai, China*
B.S. in Electrical Engineering, GPA:3.73/4.00 Jun 2017

PUBLICATIONS **(* stands for that all authors are equally contributed to this paper and are listed in alphabetical order)**

1. Guo. X., Li. R., Liu, J. and Zeng, M. (2021)* "Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to COVID-19 pandemic", *Journal of Econometrics*, Submitted
2. Guo. X., Li. R., Liu, J. and Zeng, M. (2021)* "High-dimensional mediation analysis for selecting DNA methylation loci mediatingchildhood trauma and cortisol stress reactivity", *Journal of the American Statistical Association*, Submitted
3. Guo. X., Li. R., Liu, J. and Zeng, M. (2021)* "Estimations and tests for generalized mediation models with high-dimensional potential mediators" *Journal of Business & Economic Statistics*, Submitted
4. Zeng, M., Liao, Y., Li. R. and Agus, S. (2021) "Efficient scoring algorithm for improving neural network design and inference", *AAAI* 2022, Submitted

WORK EXPERIENCES

J.P. Morgan Chase & Co. Jun 2021–Aug 2021
Quant Research Intern *New York, NY*

· Developed MLpricer model to improve pricing accuracy of exotic options.
· Conducted back-test to show MLpricer improved VaR accuracy by 90%.
· Explained the underlying mechanism of the MLpricer using mathematics proves.

Wells Fargo & Co. May 2019–Aug 2019 & Jul 2020–Aug 2020
Quant Research Intern *Charlotte, NC*

· Implement Partial Dependence Plot (PDP) & Accumulated Local Effects (ALE) to interpret machine learning models and quantified estimation uncertainty.
· Expedited PDP and ALE program using sub-sampling and multi-layer parallel.