Winter 12-15-2022

Dealing with Dimensionality: Problems and Techniques in High-Dimensional Statistics

Cezareo Rodriguez
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics and Statistics

Dissertation Examination Committee:

Nan Lin, Chair

Likai Chen

Renato Feres

Jose Figueroa-Lopez

Lei Liu

Dealing with Dimensionality: Problems and Techniques in High-Dimensional Statistics

by

Cezareo Rodriguez

A dissertation presented to
the Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2022

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

First I would first like to thank my advisor Professor Nan Lin for his patience, encouragement, and mentorship throughout my time at WashU. His deep expertise, critical thinking, and attention to detail are traits that I aspire to attain as I move forward in my career as a researcher and statistician.

I would like to thank Professor Lei Liu for collaborating on my research and lending his biostatistical expertise. I also like to thank my other committee members Jose Figueroa-Lopez, Likai Chen, and Renato Feres for all of their help and insights in finalizing my thesis. Thanks also to Todd Kuffner for organizing the WHOA-PSI workshops and for showing me that it is possible to be both a talented musician and a brilliant academic.

I would like thank the rest of the Mathematics and Statistics department for all of the support and opportunities provided to me that I would have never dreamed of. Special thanks to my professors who deepened my understanding and appreciation of mathematics. To my fellow graduate student friends in both statistics and math, you made my graduate experience an amazing one that I will never forget. Thanks to my academic siblings Liqun Yu, Wei Wang, Guanshengrui Hao, and Xiaoyu Dai for their advice and encouragement. Special thanks to Chang Liu, Jiayi Fu, Chuyi Yu, Luis Garcia, Jiaqi Li, Bei Wu, Yuchen Han, Cody Stockdale, Meredith Sargent, Jeff Norton, Jeet Sampat, Atzimba Martinez, Claire Huang, Bowei Zhao, Manasa Vempati, Tyler Williams, Josh Covey, Ben Castor, Chris Felder, Alberto Dayan, Nathan Wagner, and Mary Barker for all the great discussions and fun times.

Finally I would like to thank my parents Cesareo and Mieko and my siblings Alex, Victoria, and Arthur for their constant love and encouragement. Very special thanks goes to Savannah Myers for going above and beyond in supporting me and for being the absolute best partner I could ever ask for.

Cezareo Rodriguez

Washington University in St. Louis

August 2022

Dedicated to my family and friends.

Preface

Near the completion of this dissertation, we became aware of the recent upload of [1]. At first glance, this paper is very similar to our work presented in Chapter 3 in that it uses the composite null distribution of Sobel's test statistic to perform mediation analysis in the high dimensional setting. There are, however, some key differences that make both works valuable contributions to the literature of mediation analysis. [1] uses their Sobel-comp method as an approach to do multiple mediation hypothesis testing, then applies this multiple testing procedure in a high dimensional setting. Our approach, on the other hand, is to use a marginal screening approach based on the Sobel test (our Marginal Sobel Screening method) to reduce the dimensionality of the original problem while retaining the true set of mediators. Since the Sobel test plays a different role in the two approaches, our theoretical foundations and theorems proved also differ considerably from theirs. We cite a result of [1] in Section 3.5.2 for the sake of theoretical completeness.

ABSTRACT OF THE DISSERTATION

Dealing with Dimensionality: Problems and Techniques in High-Dimensional Statistics

by

Cezareo Rodriguez

Ph.D. in Statistics,

Washington University in St. Louis, 2022.

Professor Nan Lin, Chair

In modern data analysis, problems involving high dimensional data with more variables than subjects is increasingly common. Two such cases are mediation analysis and distributed optimization. In Chapter 2 we start with an overview of high dimensional statistics and mediation analysis. In Chapter 3 we motivate and prove properties for a new marginal screening procedure for performing high dimensional mediation analysis. This screening procedure is shown via simulation to perform better than benchmark approaches and is applied to a DNA methylation study. In Chapter 4 we construct a cryptosystem that accurately performs distributed penalized quantile regression in the high-dimensional setting using a divide-and-conquer approach while preserving the privacy of subject data.

1. Introduction

1.1 Motivation

In the modern age of high throughput computing and massive data collection, high dimensional data are increasingly common whose number of variables p is larger than its sample size n. In such a case, traditional approaches to model building and statistical inference fail. This problem has since created an explosion of literature for handling high dimensional data. The LASSO [2], elastic net [3], and other penalized regression methods seeked to perform variable selection and parameter estimation simultaneously while forcing a sparse solution to the regression problem. While this sounds convenient and appealing, optimization of these penalized functions can become computationally intensive and inconsistencies in these estimates are common in ultrahigh dimensions. Another set of approaches uses a multistage strategy in which the practitioner first performs variable screening to significantly reduce the dimensionality, then further refines the set of candidate variables with a penalized regression method. In this dissertation, we examine both classes of solutions via two applications. Using these applications as motivation, we then develop a new method for variable screening in the context of mediation and another new method for penalized quantile regression in a distributed setting.

1.2 Summary of Objectives

Under an overarching theme of high-dimensional statistics, the objective of this dissertation is motivated by:

1. An effective screening approach for conducting high-dimensional mediation analysis,

2. Methods for performing distributed penalized quantile regression in a manner that preserves privacy of individual subject information.

Background information on the models and techniques can be found in Chapter 2. Deeper details on the literature review, theory, and methodologies are further developed in Chapters 3 and 4.

1.3 Data Acknowledgement

The Coronary Artery Risk Development in Young Adults (CARDIA) study is a longitudinal study examining the development and determinants of clinical and subclinical cardiovascular disease [4]. The study began in 1985 with a group of 5115 black and white men and women aged 18-30 years. The participants were selected such that there would be roughly the same number of people in subgroups of race, gender, education, and age in each of 4 centers located in Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. These same subjects were asked to participate in follow-up examinations during 1987-1988 (Year 2), 1990-1991 (Year 5), 1992-1993 (Year 7), 1995-1996 (Year 10), 2000-2001 (Year 15), 2005-2006 (Year 20), 2010-2011 (Year 25), and 2015-2016 (Year 30). A majority of the group attended each of these follow-up examinations where various cardiovascular measurements

were repeatedly taken. Data have also been collected on physical measurements such as weight and body composition in addition to lifestyle factors such as dietary habits, exercise patterns, drug and alcohol use, behavioral and psychological variables, medical and family history, and other chemistries (e.g., insulin). Our studies focus specifically on year 15 from 2000-2001 where we treat tobacco exposure as a treatment variable, DNA methylation measurements as mediators, and lung function as a response variable.

The MIMIC III database [5] is a massive, freely-available database comprising of de-identified medical data associated with over 46,520 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. MIMIC-III includes features such as demographics, bedside vital sign measurements (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). Detailed descriptions for each of these variables are available at [6]. This database is provided largely through the work of researchers at the MIT Laboratory for Computational Physiology and their collaborator research groups.

2. Background

2.1 Inference in High Dimensions

The methods discussed in this dissertation take place in the high dimensional setting where the number of predictor variables p exceeds the sample size n. While each problem considered is specialized in its own way, they are rooted in ordinary least squares (OLS) regression in high dimensions. As such, let us consider an overview of high dimensional regression analysis. Let

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2.1}$$

where $\boldsymbol{X_i} = [X_{i,1}, ..., X_{i,p}]^T$ denotes the covariate vector for the i^{th} subject, Y_i denotes the response value for subject i, and $p > n$.

2.1.1 Variable Screening

In the presense of high dimensional data, one common technique to quickly reduce the dimension of the regression problem is variable screening. Prior to variable selection, a moderate size $d < n < p$ of candidate variables are filtered out from the total set of variables for further refinement to produce a final model. A key assumption here is the sparsity as-

sumption that a small proportion of these candidate variables are actually members of the underlying true model. Some examples of variable screening methods include sure independence screening (SIS) [7, 8], high-dimensional ordinary least-square projection (HOLP) [9], forward regression (FR) [10], and tilting [11]. These variable screening approaches aim to create a computationally fast and easy way to reduce the dimension of the candidate set of variables while retaining as many variables from the true model as possible. This is expressed in the sure screening property:

$$\mathbb{P}(\mathcal{M}^\star \subset \hat{\mathcal{M}}) \overset{n\to\infty}{\to} 1 \tag{2.2}$$

where $\mathcal{M}^\star$ denotes the set of variables in the true model and $\hat{\mathcal{M}}$ denotes the set of variables retained by the screening procedure.

The following regularity conditions are assumed when performing variable screening to ensure the sure screening property when performing variable screening:

1. $p > n$ and $\log(p) = O(n^\xi)$ for some $\xi \in (0, 1 - 2\kappa)$, where κ is given by Condition 3.

2. $\boldsymbol{Z} = X\sigma^{-1/2}$ has a spherically symmetric distribution and property C: that there exist some $c, c_1 > 1$ and $C_1 > 0$ such that the deviation inequality

$$P\left\{\lambda_{\max}\left(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}\right) > c_1 \text{ or } \lambda_{\min}\left(\tilde{p}^{-1}\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}^{\mathrm{T}}\right) < 1/c_1\right\} \leq \exp\left(-C_1 n\right)$$

holds for any $n \times \tilde{p}$ submatrix $\tilde{\boldsymbol{Z}}$ of $\boldsymbol{Z}$ with $cn < \tilde{p} \leq p$. Also, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$.

3. $\text{Var}(Y) = O(1)$ and, for some $\kappa \geq 0$ and $c_2, c_3 > 0$,

$$\min_{i \in \mathcal{M}_*} |\beta_i| \geq \frac{c_2}{n^\kappa} \quad \text{and} \quad \min_{i \in \mathcal{M}_*} \left|\text{cov}\left(\beta_i^{-1} Y, X_i\right)\right| \geq c_3$$

4. There are some $\tau \geq 0$ and $c_4 > 0$ such that $\lambda_{max}(\mathbf{\Sigma}) \leq c_4 n^\tau$

Condition 1 acknowledges the high-dimensional setting where p may grow with n, but not too quickly. Condition 2 is a regularity condition on all subsets of the matrix $\boldsymbol{Z} = X\sigma^{-1/2}$. Condition 3 requires a minimum signal strength for variables in $\mathcal{M}^\star$. Condition 4 requires a bound on the variability of the set of candidate variables. All four assumptions are easy to satisfy in most datasets after proper preprocessing.

2.1.2 Penalized Regression

Given a likelihood function of $\ell(\theta)$ that we would like to maximize, a penalized likelihood function is a functon of the form $\ell(\theta)+\lambda P(\theta)$, where λ is a scalar penalization hyperparameter and $P(\cdot)$ is a penalty function. Common choices for P include the LASSO [2], ridge penalty [12], elastic net [3], smoothly clipped absolute deviations (SCAD) [13], and minimax concave penalty (MCP) [14].

2.1.3 False Discovery Rate Control

Historically, statisticians aimed to control Familywise Error Rate (FWER) while performing simultaneous inference and multiple hypothesis testing [15]. A rich literature of methods emerged for dealing with this problem in low dimensions [16–18], but these methods proved

to be too stringent when the dimensionality of the regression problem gets too large. For this reason, statisticians aim to control false discovery rate (FDR) in high dimensions, which allows for more rejections of hypotheses while still controlling the number of false rejections. Commonly used FDR control methods include the Benjamini-Hochberg procedure [19] and Benjamini-Yekutieli procedure [20].

2.2 Mediation Analysis

Mediation analysis aims to explain the causal mechanism between a treatment A and a response Y by examining the extent to which a third variable or set of variables M contributes to the relationship between A and Y.

In the classical setting there are 4 steps in establishing mediation [21–23]:

1. Step 1: Show that the treatment variable is correlated with the response. Use Y as the dependent variable in a regression equation and A as a predictor. This step estimates the total effect (TE) of A on Y.

2. Step 2: Show that the treatment variable A is correlated with the mediator M. Use M as the dependent variable in the regression equation and A as a predictor.

3. Step 3: Show that the mediator affects the outcome variable. Use Y as the dependent variable in a regression equation and A and M as predictors. Note that it is not sufficient just to correlate the mediator with the outcome. Without including A in this model, the correlation between mediator and response can be inflated because Y and M are both caused by A. The effect of A on Y in the presence of M is known as the direct effect (DE) of A on Y given M, denoted β_A in Figure 2.2.

4. Step 4: To establish that M completely mediates the $A \to Y$ relationship, the effect of A on Y controlling for M calculated in Step 3 should be zero. If it is nonzero but smaller than the effect of A on Y calculated in Step 1, we say that M partially mediates the $A \to Y$ relationship.

If all four of these steps are met, then the data are consistent with the hypothesis that variable M completely or partially mediates the $A \to Y$ relationship. Step 1 is often considered to be optional, since cases can arise where mediation effects are significant but the treatment variable is marginally uncorrelated with the response. In particular, this situation arises when mediation effects have opposite signs and cancel each other out. Mediation effects are quantified by the indirect effect (IE). The quantification of the IE depends on the assumed structure of the mediation model, but in this single mediator case we can express IE as $IE = \gamma - \beta_A$, where γ and β_A are shown in Figures 2.1 and 2.2.

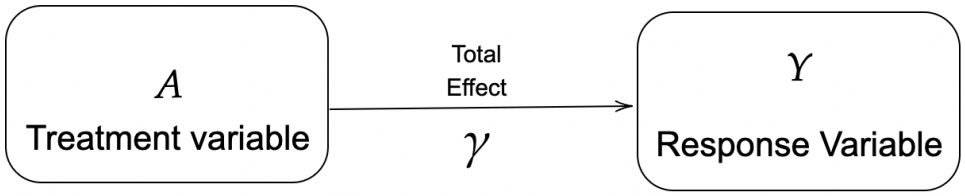

Figure 2.1.: Illustration of the first quantity considered in mediation analysis: the total effect (TE) of A on Y with no mediators present.

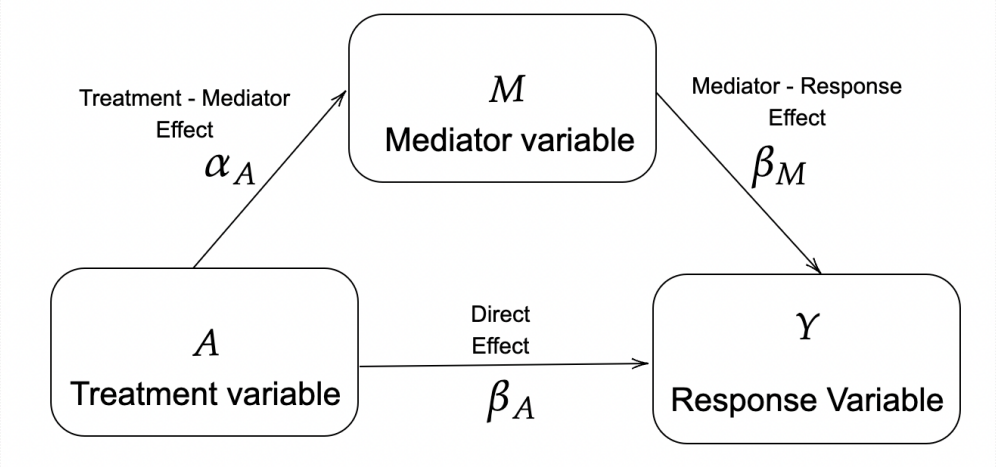

Figure 2.2.: Illustration of the breakdown of TE into direct effect (DE) of A on Y and the indirect effect (IE) of A on Y through M. The treatment-mediator effect α and the mediator-response effect β are estimated in Steps 2 and 3, respectively.

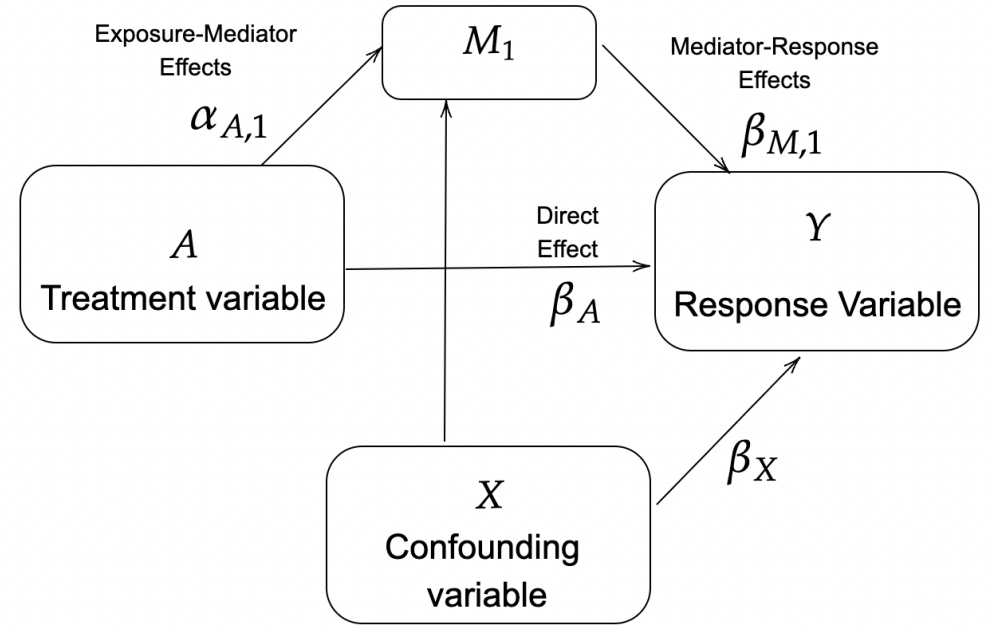

Figure 2.3.: Confounding variables are often taken into account in mediation analyses by including them as covariates at each stage of the mediation model. One important assumption of mediation analysis is no unmeasured confounders, so practitioners must be careful to consider all possible important covariates for a mediation study. In general a link can also exist from X to A to create a sequential mediation diagram, but for now we consider X to be independent of A.

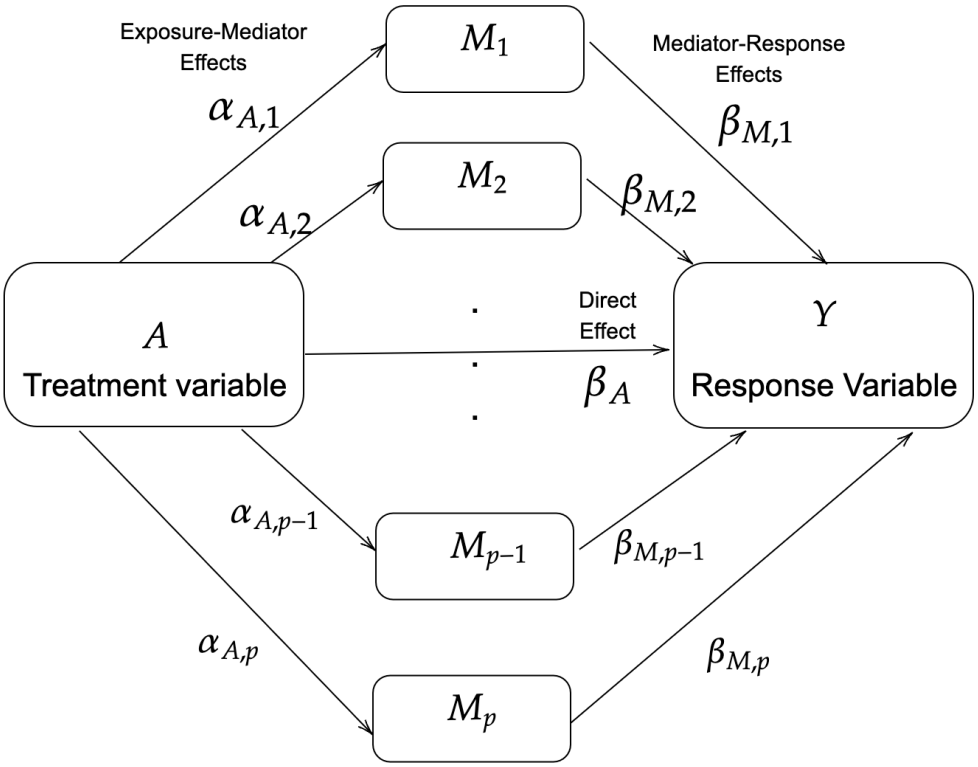

Figure 2.4.: When M becomes multi-dimensional we can consider the IE of A on Y through each mediator M_j, $j \in \{1, \ldots, p\}$ or the IE of A on Y through the entire set M.

Influenced by the works of [24, 25] and formalized for mediation analysis use by [21, 23], linear structural equation models (LSEMs) have been used to describe the relationship between A, M, and Y using linear regression models. Specifically, the system of equations for subject i, $i \in \{1, \ldots, n\}$, is:

$$M_{1,i} = \mu + \alpha_{A,1} A_i + e_i, \text{ where } e_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \tag{2.3}$$

$$Y_i = \beta_0 + \beta_A A_i + \beta_{M,1} M_{1,i} + \epsilon_i, \text{ where } \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \tag{2.4}$$

This pair of equations offers a model representation of Figure 2.2 and is often compared to the equation:

$$Y_i = \gamma A_i + \varepsilon_i \text{ where } \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \eta^2), \tag{2.5}$$

which represents Figure 2.1. The indirect effect of A on Y through M can then be expressed as either the product $\alpha\beta$ or the difference $\gamma - \beta_A$. In the case of LSEM with continuous mediator and response, these two characterizations coincide. It is worth noting that unlike in standard linear regression where there is a clear distinction between dependent and independent variables, LSEMs come with the structure that M is the dependent variable in one equation but an independent variable in the other equation. This role of the variable M, however, allows us to infer causal relationships from mediation analysis [26]. There is also an implied graphical structure as seen in Figures 2.1 and 2.2 that is not as commonly seen or used in standard linear regression.

The LSEM structure allows for several ways to test the IE, with the primary method in literature being Sobel's test of the product of coefficients [27]. After fitting the equation models from Equations (2.3) and (2.4), the product method tests the null hypothesis $H_0 : \alpha\beta = 0$ by approximating the distribution of the product via the central limit theorem. This test is performed by constructing the test statistic $t = (\hat{\alpha}\hat{\beta})/SE$, where SE denotes the standard error of $(\hat{\alpha}\hat{\beta})$: $SE = \sqrt{\hat{\alpha}^2\sigma_\beta^2 + \hat{\beta}^2\sigma_\alpha^2}$. Using the standard normal approximation for the distribution of t, the Sobel test rejects H_0 if $p_{Sobel} = \mathbb{P}(|Z| > |t|) < \alpha_{sig}$, where α_{sig} denotes the desired significance level of the test. Recent improvements to Sobel's test have been made to improve the approximation of the reference distribution by considering the composite nature of the null hypothesis [28]. Note that $H_0 : \alpha\beta = 0$ is also made up of three

subcases: $H_{10} : \alpha \neq 0$ and $\beta = 0$, $H_{01} : \alpha = 0$ and $\beta \neq 0$, and $H_{00} : \alpha = 0$ and $\beta = 0$. Under the nulls H_{10} and H_{01}, t asymptotically follows $\mathcal{N}(0, 1)$, but under null H_{00}, t asymptotically follows $\mathcal{N}(0, 1/4)$. This improvement boosted the performance of the Sobel test, particularly because multiple mediators can be used to estimate the proportion of mediators from each null distribution. MacKinnon [29] proposed the joint significance test, also known as the MaxP test, which tests the hypotheses $H_{0,a} : \alpha = 0$ and $H_{0,b} : \beta = 0$, then rejects the composite null hypothesis if both $H_{0,a}$ and $H_{0,b}$ are rejected [29]. Concretely, two test statistics are constructed $Z_a = \hat{\alpha}/\hat{\sigma}_\alpha$ and $Z_b = \hat{\beta}/\hat{\sigma}_\beta$, and p-values computed $p_a = \mathbb{P}(|Z| > |Z_a|)$ and $p_b = \mathbb{P}(|Z| > |Z_b|)$ where Z comes from a standard normal distribution. Finally, H_0 is rejected if $p_{\max} = \max(p_a, p_b) < \alpha_{sig}$.

Other variations of this approach can be taken when the linearity or normality assumptions in Equations (2.3) and (2.4) are not appropriate. For example, generalized linear models may be employed when the mediators or response are discrete or binary [30]. Pearl's interventionalist approach to causal mediation employs techniques from policy learning literature to forgo functional or distributional form assumptions on the data [31].

3. Sure Independence Screening for Mediation Analysis

3.1 Introduction

Mediation models are an increasingly popular family of statistical models that aim to explain the relationship between an independent (treatment) variable A and a dependent (response) variable Y via a third intermediate "mediator" variable M. Here, A and Y are assumed to be univariate while M is a p-dimensional set of mediator variables (i.e. random vector). In such a model, the total effect (TE) is a measure of the effect of A on Y with no mediator variables present. This TE can be decomposed into two terms: the direct effect (DE) of A on Y, denoted $DE(A, M, Y)$, describes the effect of A on Y in the presence of the mediator, and the indirect effect (IE) of A on Y, denoted $IE(A, M, Y)$, describes the effect of A on Y through the mediator M. We say that M is a mediator if the IE of A on Y through M is nonzero. If $TE(A, Y) = IE(A, M, Y)$, then $DE(A, M, Y) = 0$ and we refer to this case as complete mediation. In the complete mediation case, A no longer has any effect on Y after controlling for M. If $IE(A, M, Y) < TE(A, Y)$ and $DE(A, M, Y) \neq 0$, then we refer to this case as partial mediation. In this case controlling for M reduces the effect of A on Y, but the DE of A on Y is still present.

Two main approaches exist in the literature for approaching a mediation analysis problem. The first approach is rooted in the structural equation model (SEM) literature, dating back to Wright's works on causal paths [24, 25]. In the following decades, this literature

grew with contributions across a variety of disciplines [23, 32–35]. In many of these applied fields, mediation analysis with linear structural equation models (LSEMs) was attractive for their interpretability and ease of implementation, though this approach does require strong assumptions about the underlying data generating process.

The second approach was more recently developed by Pearl [36, 37]. Using the language of counterfactuals and directed acyclic graphs (DAGs), this approach requires fewer assumptions but is more computationally expensive due to its reliance on graph structures and their associated computationally complex algorithms. Pearl's original framework considered a single mediator, but in the past decade these results were extended to the multiple mediator case where $1 < p < n$ [38–41]. This counterfactual and DAG approach is often used in the context of structure learning [42].

We will ultimately be approaching a problem from the view of a practitioner who would like to create an LSEM model from a high-dimensional set of mediators, which is a problem of great interest in recent years. Thus, all mediation models for the rest of the paper will follow the first approach mentioned above.

In recent years, mediation analysis has become a hot topic in the field of causal inference with applications across a wide variety of disciplines from genomewide association studies (GWAS) in biology to psychological studies [43–45]. In modern mediation analysis it is common to consider the mediator M not to be a single variable, but an ultra-high dimensional set of candidate mediators. In such a context, traditional methods for performing mediation analysis fail, and new techniques are needed to handle this massive set of potential mediators.

For $i = 1, ..., n$, let A_i denote the binary treatment variable of the i^{th} subject, M_{ij} the corresponding observation of mediator j ($j = 1, ..., p$), $X_i \in \mathbb{R}^q$ a vector of covariates to ac-

count for confounding, and Y_i the corresponding response variable. Let $\boldsymbol{M}_i = [M_{i1}, ..., M_{ip}]^T$ for $i = 1, ..., N$ be the vector of observed mediator variables for the i^{th} subject, and $\boldsymbol{M} = [\boldsymbol{M}_1, ..., \boldsymbol{M}_N]^T$ the random mediator matrix. We will further denote the j^{th} column of $\mathbf{M}$, i.e. j^{th} mediator in $\mathcal{M}$, as $M^{(j)}$ and the j^{th} mediator observation of subject i with M_{ij}. Assume without loss of generality that the mediators $\boldsymbol{M}^{(j)}$ have been standardized such that $\mathbb{E}[M^{(j)}] = 0$ and $\mathbb{E}[M^{(j)2}] = 1$ for all j. Traditionally, mediation analysis considers the linear structural equation model (LSEM):

$$M_{ij} = \mu_j + X_i^T \boldsymbol{\alpha}_{X,j} + A_i \alpha_{A,j} + e_{ij} \tag{3.1}$$

$$Y_i = \beta_0 + X_i^T \beta_X + \beta_A A_i + \sum_{j=1}^{p} \beta_{M,j} M_{ij} + \epsilon_i \tag{3.2}$$

where $\boldsymbol{e}_i = [e_{i1}, ..., e_{ip}]^T \sim \mathcal{N}(\mathbf{0}_p, \Sigma_M)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Plugging the definition for M_{ij} into the definition of Y_i, it can easily be seen that

$$\mathbb{E}[Y_i | A_i, X_i] = \left(\beta_0 + \sum_{j=1}^{p} \beta_j \mu_j\right) + X_i^T \left(\beta_X + \sum_{j=1}^{p} \beta_j \boldsymbol{\alpha}_{X,j}\right) + \left(\beta_0 + \sum_{j=1}^{p} \alpha_j \beta_j\right) A_i \tag{3.3}$$

The coefficient of A_i, $\left(\beta_0 + \sum_{j=1}^{p} \alpha_j \beta_j\right)$, is the decomposition of $TE(A, Y)$ into $DE(A, M, Y) = \beta_X$ and $IE(A, M, Y) = \left(\beta_0 + \sum_{j=1}^{p} \alpha_j \beta_j\right)$. We refer to each product $\alpha_j \beta_j$ as the indirect effect of A on Y through mediator $M^{(j)}$. Defining $U_i = [1, X_i^T, A_i]^T$,

$V_i = [1, X_i^T, A_i, M_i^T]^T$, $\boldsymbol{\alpha}_j = [\mu_j, \boldsymbol{\alpha}_{X,j}, \alpha_{A,j}]^T$, $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_p]$, and $\boldsymbol{\beta} = [\beta_0, \boldsymbol{\beta}_X^T, \beta_{A,j}^T, \boldsymbol{\beta}_M^T]^T$, equation (3.1) can be written succinctly as

$$M_{ij} = U_i^T \boldsymbol{\alpha}_j + e_{ij} \tag{3.4}$$

$$\boldsymbol{M}_i = (U_i^T \boldsymbol{\alpha})^T + \boldsymbol{e}_i \tag{3.5}$$

$$Y_i = V_i^T \boldsymbol{\beta} + \epsilon_i \tag{3.6}$$

Further defining $U = [U_1, \ldots, U_n]^T$, $V = [V_1, \ldots, V_n]^T = [U, \boldsymbol{M}]$, $E = [\boldsymbol{e}_1, ..., \boldsymbol{e}_n]^T$, and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$, all observations can be combined into matrix form:

$$\boldsymbol{M} = U\boldsymbol{\alpha} + E \tag{3.7}$$

$$\boldsymbol{Y} = V\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.8}$$

To describe the underlying true model of the data generating process, let $\boldsymbol{\alpha}^\star$ and $\boldsymbol{\beta}^\star$ denote the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ underlying the data generating process. $\alpha_A^\star = [\alpha_{A,1}^\star, ..., \alpha_{A,p}^\star]^T$ and $\boldsymbol{\beta}_M^\star = [\beta_{M,1}^\star, \ldots, \beta_{M,p}^\star]^T$ denote the true parameters of interest of the model, while $\boldsymbol{\mu}^\star = [\mu_1^\star, \ldots \mu_p^\star]^T$, $\boldsymbol{\alpha}_X^\star = [\boldsymbol{\alpha}_{X,1}^\star, \ldots, \boldsymbol{\alpha}_{X,p}^\star]$, $\beta_0^\star$, $\beta_X^\star$, and $\beta_A^\star$ denote the true values for the nuisance parameters. Let $\mathcal{M} = \{1, \ldots, p\}$ denote the indices of the candidate mediators and $\mathcal{M}^\star = \{j = 1, ..., p : \alpha_j^\star \neq 0 \text{ and } \beta_j^\star \neq 0\}$ denote the set of true mediators.

In the low dimensional context, this problem has been studied extensively. The original tests for mediation effects considered the case where M is univariate, i.e. $p = 1$. Sobel [27] developed the product method, also known as Sobel's Test, for testing the significance of mediation effect. MacKinnon et al. [29] recommends the joint significance test, also known as

the MaxP test, as a less conservative alternative for testing mediation effects. Both methods rely on estimating $\boldsymbol{\alpha}$ with $\hat{\boldsymbol{\alpha}} = (U^T U)^{-1} U^T M$ and $\hat{\boldsymbol{\beta}} = (V^T V)^{-1} V^T Y$ and applying their respective decision rules. This approach was easily extended to a multiple mediator case in which multiple mediators are added to the mediator-response regression and multiple treatment-mediator regressions are considered [46]. Naturally, this resulted in increased interest in mediation analysis under the high-dimensional setting.

In the high dimensional context ($p > n$), estimation of $\boldsymbol{\beta}$ becomes undetermined due to the non-identifiability of $f_2(Y|M, A, X)$. To remedy this issue, one common technique used in mediation is to apply sure independence screening (SIS) to the mediator variables and only consider mediators with a high marginal correlation with the response variable [8]. Using this screening approach, the set of mediators $\mathcal{M}$ of size p can be reduced to a smaller candidate set $\hat{\mathcal{M}}$ of smaller size $d < n$ to which we can apply traditional mediation techniques. Under mild regularity conditions, SIS possesses the sure screening property:

$$\mathbb{P}(\mathcal{M}^\star \subset \hat{\mathcal{M}}) \to 1 \text{ as } n \to \infty \tag{3.9}$$

where $\hat{\mathcal{M}}$ is the set of mediators chosen by SIS. Other modern approaches include simply conducting multiple testing in the high dimensional context [1, 28] or employing techniques from the post-selection inference literature [47].

To develop our novel screening method, we follow the approach of [8], [48], [49], and others by establishing properties needed at the population level and sample level to achieve the sure screening property. Our screening method is developed in a similar fashion to [1], though to our knowledge our approach is novel in its use as a screening method to reduce the dimension

of a mediation analysis problem before employing low-dimensional techniques. This follows the mediation analysis epigenetic study framework laid out by [50] with an improvement in the screening technique employed.

3.2 Mediation Analysis

3.2.1 Model and Low-Dimensional Case

We first consider linear structural equation models to simplify calculations and proofs, though our results can be readily extended to the case structural equation models of generalized linear models.

Let us consider the linear structural equation model defined by (3.1). Consider the mediators $\boldsymbol{M}$ to be partitioned into sets $\mathcal{M}_1 = \{j : \alpha_{A,j} \neq 0 \text{ and } \beta_{M,j} \neq 0\}$, $\mathcal{M}_2 = \{j : \alpha_{A,j} = 0 \text{ and } \beta_{M,j} \neq 0\}$, $\mathcal{M}_3 = \{j : \alpha_{A,j} \neq 0 \text{ and } \beta_{M,j} = 0\}$, and $\mathcal{M}_4 = \{j : \alpha_{A,j} = 0 \text{ and } \beta_{M,j} = 0\}$. $\mathcal{M}_1$ can be considered to be the true set of true mediators that we wish to identify (i.e. $\mathcal{M}_1 = \mathcal{M}^\star$), while $\mathcal{M}_2, \mathcal{M}_3$, and $\mathcal{M}_4$ are sets of false mediators.

In the modern causal inference framework, there are five standard assumptions for estimating the IE [51, 52]:

1. There are no unmeasured treatment-response confounders given $\boldsymbol{X}$,
2. There are no unmeasured mediator-response confounders given $(\boldsymbol{X}, A)$,
3. There are no unmeasured treatment-mediator confounders given $\boldsymbol{X}$,
4. There is no effect of treatment that confounds the mediator-response relationship,
5. There is no treatment and mediator interaction on the response.

Under these assumptions, the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be identified, as can the IE which depends on them. In our continuous case, the IE of A on Y through mediator j is equal to $\alpha_{A,j}\beta_{M,j}$. Note that in practice, unmeasured confounders may be part of the true underlying model.

These assumptions also imply that the causal effect of the treatment A on the mediator M is independent of the causal effect of the mediator M on the response Y [28]. Indeed, under assumptions (1)-(5), the joint probability density function of (Y, M, A, X) can be factored as $f(Y, M, A, X) = f_2(Y|M, A, X)f_1(M|A, X)f_0(A, X)$. $f_0(A, X)$ is ancillary and does not depend on the model parameters of model (3.1), so it can be discarded. Thus, only $f_2(Y|M, A, X)f_1(M|A, X)$ is needed for screening based on these parameters. $f_2(Y|M, A, X)$ contains the $M \Rightarrow Y$ associations, while $f_1(M|A, X)$ contains the $A \Rightarrow M$ associations. Denoting the log-likelihood function as $\ell(\cdot)$, the previous factorization implies that $\frac{\partial^2 \ell}{\partial\alpha\partial\beta} = 0$, which tells us that the Fisher information matrix is block diagonal, i.e.

$$\mathcal{I}(\alpha, \beta) = \begin{bmatrix} \mathcal{I}_\alpha(\alpha) & 0_{q+1} \\ 0_{q+1} & \mathcal{I}_\beta(\beta) \end{bmatrix}$$

where

$$\mathcal{I}_\alpha(\alpha) = \mathbb{E}\left[\left(\frac{\partial}{\partial\alpha}\log f_2(M; \alpha, X)\right)^2 \mid \alpha\right] = \int_{\mathbb{R}}\left(\frac{\partial}{\partial\alpha}\log f_2(m; \alpha, x)\right)^2 f(m; \alpha, x)dm, \text{ and}$$

$$\mathcal{I}_\beta(\beta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\beta}\log f_1(Y; \beta, M, X)\right)^2 \mid \beta\right] = \int_{\mathbb{R}}\left(\frac{\partial}{\partial\beta}\log f_1(y; \beta, m, x)\right)^2 f(y; \beta, m, x)dy$$

This tells us that the covariance matrix of the maximum likelihood estimator

$$\Sigma_{\alpha,\beta} = \mathcal{I}(\alpha,\beta)^{-1} = \begin{bmatrix} \mathcal{I}_\alpha(\alpha)^{-1} & 0_{q+1} \\ 0_{q+1} & \mathcal{I}_\beta(\beta)^{-1} \end{bmatrix}$$

is also block diagonal, which implies that $\hat{\alpha}$ and $\hat{\beta}$ are independent in their limiting distribution. Using this property, it becomes easy to derive the joint distribution of $(\hat{\alpha}, \hat{\beta})$ and construct our test for mediation.

3.2.2 Mediation Analysis in High Dimensions

In the high-dimensional setting ($p > n$), the maximum likelihood estimator $\hat{\theta}_{MLE}$ is noisy. In a similar spirit to SIS, we aim to utilize marginal information from individual to screen out mediators that are not likely to be important in the mediation model while retaining as many important mediators as possible. In terms of our model, we would like to reduce our candidate set of mediators $\boldsymbol{M}$ to a low dimensional size $d < n$ by filtering out as many mediators in $\mathcal{M}_2, \mathcal{M}_3$, and $\mathcal{M}_4$ as possible while also retaining as many mediators in $\mathcal{M}_1$ as possible.

To make use of marginal information of the mediators, we first define the marginal log-likelihood function. For $\theta_j = (\alpha_{A,j}, \beta_{M,j})$, define

$$\begin{aligned} \ell_j(\theta_j) &= \log(f_{1,j}(M_j; \alpha_{A,j}, A, X)) + \log(f_{2,j}(Y; \beta_{M,j}, M_j, A, X)) \\ &= -\sum_{i=1}^{N} \left\{ (M_{ij} - \mu_j - A_i\alpha_{A,j} - X_i^T\alpha_{X,j})^2 + (Y_i - \beta_0 - \beta_A A_i - X_i^T\beta_X - \beta_{M,j}M_{ij})^2 \right\} \end{aligned} \tag{3.10}$$

This marginal likelihood function ℓ_j can be viewed as a misspecified or underspecified likelihood function in which the only mediator considered is M_j. Maximizing this likelihood function with respect to θ_j then yields $\hat{\alpha}_{A,j} = \text{argmax}_{\alpha_{A,j}} \log(f_1(M_j; \alpha_{A,j}, A, X))$ and $\hat{\beta}^M_{M,j} = \text{argmax}_{\beta_{M,j}} \log(f_2(Y; \beta_{M,j}, M_j, A, X))$. Define $\theta^*_j = \text{argmax}_{\theta_j} \ell_j(\theta_j)$. θ^*_j is the value that minimizes the KL divergence between the marginal model and the true model defined by $\theta^\star$, and we assume this value is unique [53]. Indeed, considering the distributions Q denoting the cumulative distribution function corresponding to $\mathcal{N}(Z_j\boldsymbol{\beta}^*, \sigma_1^2)$ and P the cumulative distribution function corresponding to $\mathcal{N}(Z\boldsymbol{\beta}^\star, \sigma_w^2)$, expanding the form of the KL divergence yields

$$KL(\boldsymbol{\beta}^\star || \boldsymbol{\beta}^*) = \frac{1}{2}\log\left(\frac{\sigma_1^2}{\sigma_2^2}\right) - \frac{1}{2} + \frac{\sigma_1^2 + (Z\boldsymbol{\beta}^\star - Z_j\boldsymbol{\beta}^*)^2}{2\sigma_2^2}, \tag{3.11}$$

which has a unique maximum with respect to $\boldsymbol{\beta}^*$ provided that $(Z_j^T Z_j)$ is invertible. As a misspecified maximum likelihood estimator, the asymptotic distribution of $\hat{\theta}_j$ is given by $\hat{\theta}_j = [\hat{\alpha}_{A,j}, \hat{\beta}^M_{M,j}]^T \sim \mathcal{N}\left([\alpha_{A,j}, \beta^M_{M,j}]^T, \Sigma_\theta\right)$, where $\beta^M_{M,j} = \text{argmax}_{\beta_{M,j}} \mathbb{E}\left[\log(f_2(Y; \beta_{M,j}, M_j, A, X))\right]$ where expectation is taken under the true model of the data [53]. In other words, $\beta^M_{M,j}$ is the population version of $\hat{\beta}^M_{M,j}$.

Since $\hat{\theta}_j$ satisfies the score equations $\nabla_{\theta_j}\ell_j(\hat{\theta}_j) = 0$, performing a Taylor series expansion of the score equations around $\hat{\theta}_j = \theta^*_j$ yields

$$0 \approx \nabla_{\theta_j}\ell_j(\theta^*_j) + (\hat{\theta}_j - \theta^*_j)\nabla^2_{\theta_j}\ell_j(\theta^*_j). \tag{3.12}$$

Moving terms around then yields

$$\sqrt{n}(\hat{\theta}_j - \theta_j^*) \approx \left\{ \frac{1}{n} \nabla_\theta^2 \ell_j(\theta_j^*) \right\}^{-1} \left\{ \frac{1}{\sqrt{n}} \nabla_\theta \ell_j(\theta^*) \right\}. \tag{3.13}$$

Applying the Central Limit Theorem to the right hand term yields

$$\frac{1}{\sqrt{n}} \nabla_{\theta_j} \ell_j(\theta_j^*) = \frac{-1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_{\theta_j} \log(f_j(Y, M_j; \theta_j)) \to \mathcal{N}\left(\mathbf{0}, \mathrm{Var}_{\theta_j^*} \left[\nabla_\theta \log(f_j(Y, M_j; \theta_j)) \right] \right). \tag{3.14}$$

Similarly,

$$\frac{1}{n} \nabla_{\theta_j}^2 \ell(\theta^*) \to \mathbb{E}_{\theta^*} \left[\nabla_{\theta_j}^2 \ell_j(\theta^*), \right] \tag{3.15}$$

where the expectation is taken under the true model $\theta^\star$. By Slutsky's Theorem, $\sqrt{n}(\hat{\theta} - \theta^*) \to \mathcal{N}(0, W)$, where

$$\begin{aligned} W &= \left\{ \mathbb{E}_{\theta^\star} \left[\nabla_{\theta_j}^2 \ell_j(\theta^*) \right] \right\}^{-1} \mathrm{Var}_{\theta_j^*} \left(\nabla_{\theta_j} \log(f_j(Y, M; \theta_j)) \right) \left\{ \mathbb{E}_{\theta^\star} \left[\nabla_{\theta_j}^2 \ell_j(\theta^*) \right] \right\}^{-1} \\ &= B^{-1} C B^{-1}, \end{aligned} \tag{3.16}$$

where $B = \mathbb{E}_{\theta_j^\star} \left[\nabla_{\theta_j}^2 \ell_j(\theta_j^*) \right]$ and $C = \mathrm{Var}_{\theta_j^*} \left(\nabla_{\theta_j} \log(f_j(Y, M; \theta_j)) \right)$. Expanding these terms by applying the derivatives to ℓ_j then gives us

$$B = \mathbb{E}_{\theta^\star} \left[\nabla_{\theta_j}^2 \ell_j(\theta_j^*) \right] = \mathbb{E}_{\theta^\star} \begin{bmatrix} \sum_{i=1}^N A_i^2 & 0 \\ 0 & \sum_{i=1}^N M_{ij}^2 \end{bmatrix} = \begin{bmatrix} A^T A & 0 \\ 0 & \sum_{i=1}^N \left\{ (\mu_j + X_i^T \boldsymbol{\alpha}_{X,j} + \alpha_{A,j} A_i)^2 + \sigma_M^2 \right\} \end{bmatrix}$$

and

$$
\begin{aligned}
C &= \mathrm{Var}_{\theta_j^*}\left(\nabla_{\theta_j}\log(f_j(Y,M;\theta_j))\right) \\
&= \mathrm{Var}_{\theta^\star}\left(\begin{bmatrix} -2\sum_{i=1}^N(M_{ij}-\mu_j-\alpha_{A,j}A_i-X_i^T\boldsymbol{\alpha}_{X,j})A_i \\ -2\sum_{i=1}^N(Y_i-\beta_0-X_i^T\boldsymbol{\beta}_X-\beta_A A_i-\beta_{M,j}M_{ij})M_{ij}\end{bmatrix}\right) \\
&= \mathrm{Var}_{\theta^\star}\left(\begin{bmatrix} -2S_{1,j} \\ -2S_{2,j}\end{bmatrix}\right) = 4\,\mathrm{Var}_{\theta^\star}\left(\begin{bmatrix} S_{1,j} \\ S_{2,j}\end{bmatrix}\right)
\end{aligned}
$$

where $S_{1,j} = \sum_{i=1}^N(M_{ij}-\mu_j-\alpha_{A,j}A_i-X_i^T\boldsymbol{\alpha}_{X,j})A_i$ and

$S_{2,j} = \sum_{i=1}^N(Y_i-\beta_0-X_i^T\boldsymbol{\beta}_X-\beta_A A_i-\beta_{M,j}M_{ij})M_{ij}$. Computing the first and second moments of $S_{1,j}$ and $S_{2,j}$ then yields

$$
\mathbb{E}_{\theta^\star}[S_{1,j}] = \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^N(M_{ij}-\mu_j-\alpha_{A,j}A_i-X_i^T\boldsymbol{\alpha}_{X,j})A_i\right] = 0,
$$

$$
\begin{aligned}
\mathbb{E}_{\theta^\star}[S_{2,j}] &= \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^N(Y_i-\beta_0-X_i^T\boldsymbol{\beta}_X-\beta_A A_i-\beta_{M,j}M_{ij})M_{ij}\right] = \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^N\sum_{k\neq j}\beta_{M,k}M_{ik}M_{ij}\right] \\
&= \sum_{i=1}^N\sum_{k\neq j}\beta_{M,k}\left\{(\mu_j+X_i^T\boldsymbol{\alpha}_{X,j}+\alpha_{A,j}A_i)(\mu_k+X_i^T\boldsymbol{\alpha}_{X,k}+\alpha_{A,k}A_i)+\rho_{jk}\right\},
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\left[S_{1,j}^2\right] &= \mathbb{E}_{\theta^\star}\left[\left(\sum_{i=1}^{N}(M_{ij}-\mu_j-\alpha_{A,j}A_i-X_i^T\boldsymbol{\alpha}_{X,j})A_i\right)^2\right] \\
&= \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^{N}\sum_{i'=1}^{N}(M_{ij}-\mu_j-\alpha_{A,j}A_i-X_i^T\boldsymbol{\alpha}_{X,j})A_i(M_{i'j}-\mu_j-\alpha_{A,j}A_{i'}-X_{i'}^T\boldsymbol{\alpha}_{X,j})A_{i'}\right] \\
&= \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^{N}(M_{ij}-\mu_j-\alpha_{A,j}A_i-X_i^T\boldsymbol{\alpha}_{X,j})^2A_i^2\right] \text{ (cross terms independent)} \\
&= \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^{N}e_{ij}^2A_i^2\right] = \sigma_j^2A^TA,
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[S_{1,j}S_{2,j}] = \sum_{i=1}^{N}A_i\,\mathbb{E}_{\theta^\star}[\,&\sum_{k\neq j}\beta_{M,k}(\mu_j+X_i^T\boldsymbol{\alpha}_{X,j}+\alpha_{A,j}A_i)e_{ij}e_{ik} \\
&+e_{ij}^2\beta_{M,k}(\mu_k+X_i^T\boldsymbol{\alpha}_{X,k}+\alpha_{A,k}A_i)+e_{ij}^2e_{ik}\beta_{M,k}] \\
= \sum_{i=1}^{N}A_i\sum_{k\neq j}\beta_{M,k}\{&(\mu_j+X_i^T\boldsymbol{\alpha}_{X,j}+\alpha_{A,j}A_i)\,\mathbb{E}_{\theta^\star}\left[e_{ij}e_{ik}\right] \\
&+\mathbb{E}_{\theta^\star}\left[e_{ij}^2\right](\mu_k+X_i^T\boldsymbol{\alpha}_{X,k}+\alpha_{A,k}A_i)\beta_{M,k}\} \\
= \sum_{i=1}^{N}A_i\sum_{k\neq j}\beta_{M,k}\{&(\mu_j+X_i^T\boldsymbol{\alpha}_{X,j}+\alpha_{A,j}A_i)\rho_{jk} \\
&+\rho_{jj}(\mu_k+X_i^T\boldsymbol{\alpha}_{X,k}+\alpha_{A,k}A_i)\beta_{M,k}\}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}_{\theta^\star}\left[S_{2,j}^2\right] &= \mathbb{E}_{\theta^\star}\left[\sum_{i=1}^{N}(Y_i - \beta_0 - X_i^T\boldsymbol{\beta}_X - \beta_A A_i - \beta_{M,j}M_{ij})^2 M_{ij}^2\right] \\
&= \sum_{i=1}^{N}\mathbb{E}_{\theta^\star}\left[(Y_i - \beta_0 - X_i^T\boldsymbol{\beta}_X - \beta_A A_i - \beta_{M,j}M_{ij})^2 M_{ij}^2\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[\sum_{k\neq j}(\beta_{M,k}M_{ik} + \epsilon_i)^2 M_{ij}^2\right] = \sum_{i=1}^{N}\mathbb{E}\left[\sum_{k\neq j}(\beta_{M,k}M_{ik})^2 M_{ij}^2 + \epsilon_i^2 M_{ij}^2\right] \\
&= \sum_{i=1}^{N}\sum_{k\neq j}\left\{\beta_{M,k}^2\,\mathbb{E}\left[M_{ik}^2 M_{ij}^2\right] + \mathbb{E}\left[\epsilon_i^2 M_{ij}^2\right]\right\} \\
&= \sum_{i=1}^{N}\sum_{k\neq j}\{\beta_{M,k}^2(\mu_j + X_i^T\boldsymbol{\alpha}_{X,j} + \boldsymbol{A}\alpha_{A,j})^2(\mu_k + X_i^T\boldsymbol{\alpha}_{X,k} + \boldsymbol{A}\alpha_{A,k})^2 \\
&+ \sigma_{kk}(\mu_j + X_i^T\boldsymbol{\alpha}_{X,j} + \boldsymbol{A}\alpha_{A,j})^2 + 4(\mu_j + X_i^T\boldsymbol{\alpha}_{X,j} + \boldsymbol{A}\alpha_{A,j})(\mu_k + X_i^T\boldsymbol{\alpha}_{X,k} + \boldsymbol{A}\alpha_{A,k}) \\
&+ \sigma_{jj}(\mu_k + X_i^T\boldsymbol{\alpha}_{X,k} + \boldsymbol{A}\alpha_{A,k})^2 + \sigma_{jj} + \sigma_{kk} + 4\sigma_{jk} \\
&+ \sigma^2\left((\mu_j + X_i^T\boldsymbol{\alpha}_{X,j} + A\alpha_{A,j})^2 + \sigma_{jj}\right)\}
\end{aligned}
$$

Finally putting these together, we obtain

$$
\begin{aligned}
\mathrm{Var}_{\theta^\star}\left(\begin{bmatrix} S_{1,j} \\ S_{2,j} \end{bmatrix}\right) &= \begin{bmatrix} \mathbb{E}[S_{1,j}^2] & \mathbb{E}[S_{1,j}S_{2,j}] \\ \mathbb{E}[S_{1,j}S_{2,j}] & \mathbb{E}[S_{2,j}^2] \end{bmatrix} - \begin{bmatrix} \mathbb{E}[S_{1,j}] \\ \mathbb{E}[S_{2,j}] \end{bmatrix}\begin{bmatrix} \mathbb{E}[S_{1,j}] & \mathbb{E}[S_{2,j}] \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{E}[S_{1,j}^2] & \mathbb{E}[S_{1,j}S_{2,j}] \\ \mathbb{E}[S_{1,j}S_{2,j}] & \mathbb{E}[S_{2,j}^2] - \mathbb{E}[S_{2,j}]^2 \end{bmatrix}
\end{aligned} \tag{3.17}
$$

Since $W = B^{-1}CB^{-1}$, plugging everything in gives us the asymptotic covariance matrix of $(\hat{\alpha}_{A,j}, \hat{\beta}_{M,j}^M)$. In particular,

$$
\mathrm{Cov}(\hat{\alpha}_{A,j}, \hat{\beta}_{M,j}^M) \to \frac{4\,\mathbb{E}[S_{12}]}{B_{11}B_{22}} \tag{3.18}
$$

$$= \frac{4\sum_{i=1}^{N} A_i \sum_{k\neq j} \beta_{M,k} \left\{(\mu_j + X_i^T \boldsymbol{\alpha}_{X,j} + \alpha_{A,j} A_i)\rho_{jk} + \rho_{jj}(\mu_k + X_i^T \boldsymbol{\alpha}_{X,k} + \alpha_{A,k} A_i)\beta_{M,k}\right\}}{(A^T A)\left(\sum_{i=1}^{N} \left\{(\mu_j + X_i^T \boldsymbol{\alpha}_{X,j} + \alpha_{A,j} A_i)^2 + \sigma_M^2\right\}\right)}$$

$$= \frac{4\left\{\boldsymbol{A}^T(\mu_j \mathbb{1}_n + \boldsymbol{A}\alpha_{A,j} + \boldsymbol{X}\boldsymbol{\alpha}_{X,j}) I_{-j}^T \Sigma_M \tilde{\boldsymbol{\beta}}_{-j} + \sigma_{jj}\boldsymbol{A}^T(\mathbb{1}_n \boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\alpha}_X + \boldsymbol{A}\boldsymbol{\alpha}_A)\tilde{\boldsymbol{\beta}}_{-j}\right\}}{\boldsymbol{A}^T\boldsymbol{A}\left\{(\mu_j \mathbb{1}_n + \boldsymbol{X}\boldsymbol{\alpha}_{X,j} + \boldsymbol{A}\alpha_{A,j})^T(\mu_j \mathbb{1}_n + \boldsymbol{X}\boldsymbol{\alpha}_{X,j} + \boldsymbol{A}\alpha_{A,j}) + n\sigma_{jj}\right\}}$$

where I_{-j} denotes the j^{th} column of the $p \times p$ identity matrix I_p, and $\tilde{\boldsymbol{\beta}}_{-j}$ is a copy of $\boldsymbol{\beta}^\star$ with the j^{th} entry replaced by 0. This quantity is easily nonzero for many parameter settings, especially in the high dimensional case. Thus, hypothesis tests for mediation that assume independence between $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are easily invalidated in high dimensions.

3.2.3 Screening

In the high dimensional context, the likelihood function f does not have a unique maximum, causing $\hat{\theta}_{MLE} = (\hat{\boldsymbol{\alpha}}_{MLE}, \hat{\boldsymbol{\beta}}_{MLE})$ to be an ill-conditioned estimator. Our screening approach aims to lower the dimensionality of the high-dimensional mediation problem by making use of marginal information of each mediator in a similar spirit to SIS [8]. The goal when developing a screening method is twofold. The first goal is to create a method with low computational cost, since screening is predominantly used as a crude and quick way to reduce the dimensionality of the regression problem considered. The second goal is to create a method that satisfies the sure screening property under reasonable assumptions.

Our screening approach for mediation modifies the conditional sure independence screening (CSIS) approach of Barut, Fan, & Verhasselt which uses a conditioning set of variables to further screen for more variables associated with a response [49]. We will consider $U = [\mathbb{1}_n, X, A]$ to be the conditional set of variables for the $A \to M$ relationships,

$U_{\mathcal{C}} = [\mathbb{1}_n, X]$ to be the conditional set of variables with no treatment, and $Z_j = [U, M_j]$ to be the conditional set of variables for the $M \to Y$ relationships. While taking advantage of the structure of mediation models, we hope to use this idea to detect more significant mediators than traditional approaches that assume independence between the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}^M$.

Given a set of covariates X, our goal is to screen the high-dimensional set of mediators $\boldsymbol{M}$ for variables that mediate the relationship between A and Y. That is, we would like to identify the set $\{j \in \{1, ..., p\} : \alpha_{A,j} \neq 0 \text{ and } \beta_{M,j} \neq 0\}$. In the language of causal paths, we would like to identify which M_j lies on a path from A to Y in the graph corresponding to the true underlying LSEM.

Given a random $\{(X_i, A_i, \boldsymbol{M}_i, Y_i)\}_{i=1}^n$ from the LSEM model, the conditional maximum marginal likelihood estimator $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$ for $j \in \mathcal{M}$ is defined as the minimizer of the negative marginal log-likelihood:

$$\hat{\beta}_{\mathcal{C}j}^M = \text{argmin}_{\beta_{\mathcal{C}}, \beta_{M,j}} \sum_{i=1}^{n} \{\log f_2(Y_i | Z_{j,(i)}, \beta_{\mathcal{C}j})\}, \tag{3.19}$$

where $Z_{j,(i)}^T$ denotes the i^{th} row of Z_j, $\log f_2(Y_i | Z_{j,(i)}$ is the conditional marginal log likelihood function for $\boldsymbol{\beta}_{\mathcal{C}j}$, and $\beta_{\mathcal{C}j}^M = [\hat{\beta}_{\mathcal{C}}^T, \beta_{M,j}^M]^T$. This can also be compactly written as

$$\ell_{2,j}(\boldsymbol{\beta}_{\mathcal{C}j}) = \boldsymbol{\beta}_{\mathcal{C}j}^T \boldsymbol{Z}_j^T Y - \frac{1}{2} \boldsymbol{\beta}_{\mathcal{C}j}^T \boldsymbol{Z}_j^T \boldsymbol{Z}_j \boldsymbol{\beta}_{\mathcal{C}j} \tag{3.20}$$

Similarly, for the $A \to M$ relationship, consider the conditional maximum marginal likelihood estimator $\hat{\alpha}_j$ for $j \in \mathcal{M}$ which minimizes the negative marginal log-likelihood:

$$\hat{\alpha}_j = \text{argmin}_{\alpha_j} \sum_{i=1}^{n} \{\log f_1(M_{ij}|U_i, \alpha_j)\} = \text{argmin}_{\alpha_j} \ell_1(\alpha_j), \tag{3.21}$$

where $\log f_1(M_{ij}|U_i, \alpha_j)$ is the marginal likelihood for M_j given U. This similarly defines the marginal log-likelihood function for $\boldsymbol{\alpha}_j$:

$$\ell_{1,j}(\boldsymbol{\alpha}_j) = \boldsymbol{\alpha}_j^T U^T M^{(j)} - \frac{1}{2}\boldsymbol{\alpha}_j^T U^T U \boldsymbol{\alpha}_j. \tag{3.22}$$

We will similarly consider $\boldsymbol{\alpha}_j$ to be partitioned $\boldsymbol{\alpha}_j = [\hat{\alpha}_{\mathcal{C}}^T, \hat{\alpha}_{A,j}]^T$.

Our mediation screening procedure based on estimated marginal magnitudes keeps the mediators belong to the set:

$$\hat{\mathcal{M}}_\lambda = \{j \in \mathcal{M} : |\hat{\alpha}_{A,j}^M \hat{\beta}_{M,j}^M| > \lambda\}, \tag{3.23}$$

given a thresholding parameter $\lambda > 0$. This thresholding rule depends on the scale of the parameters $\mathbb{E}[M_j|U]$ and $\mathbb{E}[Y|Z_j]$, so a scale-free thresholding may instead be preferred. Rather than thresholding based on the absolute scale of $|\hat{\alpha}_{A,j}^M \hat{\beta}_{M,j}^M|$, a scale-free thresholding rule is based on the reduction in the negative log-likelihood function when adding mediator M_j conditional on U. For this rule we compute

$$T_{Sobel,j} = \frac{\hat{\alpha}_{A,j}\hat{\beta}_{M,j}}{\sqrt{\hat{\alpha}_{A,j}^2 \hat{\sigma}_{\beta,j}^2 + \hat{\beta}_{M,j}^2 \hat{\sigma}_{\alpha,j}^2}}, \tag{3.24}$$

which scales the product of marginal coefficients $|\hat{\alpha}^M_{A,j}\hat{\beta}^M_{M,j}|$ by its standard error $\sqrt{\hat{\alpha}^2_{A,j}\hat{\sigma}^2_{\beta,j} + \hat{\beta_{M,j}}^2\hat{\sigma}^2_{\alpha}}$, where $\hat{\sigma}_{\alpha,j}$ and $\hat{\sigma}_{\beta,j}$ denote the standard errors of $\hat{\alpha}_{A,j}$ and $\hat{\beta}^M_{M,j}$, respectively [54]. This can especially be seen as a scaled version of the previous statistic if we let $Z_{\alpha,j} = \hat{\alpha}_{A,j}/\hat{\sigma}_{\alpha}$ and $Z_{\beta,j} = \hat{\beta}^M_{M,j}/\hat{\sigma}_{\beta,j}$. Then $T_{Sobel,j}$ can be written

$$T_{Sobel,j} = \frac{Z_{\alpha,j}Z_{\beta,j}}{\sqrt{Z^2_{\alpha,j} + Z^2_{\beta,j}}}, \tag{3.25}$$

where the standard error of $Z_{\alpha,j}Z_{\beta,j}$ is $\sqrt{Z^2_{\alpha,j} + Z^2_{\beta,j}}$. The scale-free thresholding rule then keeps variables according to

$$\hat{\mathcal{M}}_{\tilde{\lambda}} = \{j \in \mathcal{M} : |T_{Sobel,j}| > \tilde{\lambda}\}, \tag{3.26}$$

where $\tilde{\lambda}$ is a thresholding parameter associated with $T_{Sobel,j}$. In particular, the asymptotic distribution of $T_{Sobel,j}$ can be used to calculate thresholds $\tilde{\lambda}$ based on the inverse CDF of $T_{Sobel,j}$. Due to this screening procedure's influence from Sobel's test, we refer to is as the marginal Sobel screening (MSS) procedure.

Intuitively, we would like to keep the mediators that have large indirect effects on the response, i.e. they lie on a more significant path from A to Y. Note that this approach differs from traditional methods that only screen based on correlations between the mediators and the response. By incorporating information about the relationships between the treatment and the mediators, this approach helps to prevent removal of mediators with small $\beta_{M,j}$ but large $\alpha_{A,j}$. It also does not use this screening as a simultaneous hypothesis test, but rather a simple dimension reduction technique before further model selection takes place.

3.3 Sure Screening Properties

In order to prove the sure screening property for our new approach, properties of the data at the population and the sample levels must be addressed. We follow the approach of [49] in developing these properties and related conditions in order to establish a sure screening property.

3.3.1 Population-Level Properties

Each marginal fitting of the mediator - response relationships can be seen as a misspecified or underspecified model. Due to this misspecification, the marginal regression estimator $\beta_{M,j}^{M}$ can differ from the true model parameters $\beta_{M,j}^{\star}$. We investigate conditions under which $\beta_{M,j}^{M}$ is still informative about $\beta_{M,j}^{\star}$ in the sense that if $|\beta_{M,j}^{M}|$ is larger than some threshold value, then $|\beta_{M,j}^{\star}|$ is larger than another threshold value.

The marginal regression coefficients $\boldsymbol{\alpha}_j$ satisfy the score equations:

$$\mathbb{E}[U^T U \boldsymbol{\alpha}_j] = \mathbb{E}[U^T M_j] = \mathbb{E}[U^T U \boldsymbol{\alpha}_j^{\star}], \tag{3.27}$$

which implies

$$\mathbb{E}[U^T U \boldsymbol{\alpha}] = \mathbb{E}[U^T M] = \mathbb{E}[U^T U \boldsymbol{\alpha}^{\star}]. \tag{3.28}$$

Similarly, the marginal regression coefficients $\boldsymbol{\beta}_{\mathcal{C}j}^{M}$ satisfy the score equations:

$$\mathbb{E}[Z_j^T Z_j \beta_{\mathcal{C}j}^{M}] = \mathbb{E}[Z_j^T Y] = \mathbb{E}[Z_j^T V \boldsymbol{\beta}^{\star}]. \tag{3.29}$$

Without additional mediator variable M_j, the baseline parameter is

$$\boldsymbol{\beta}_{\mathcal{C}}^{M} = \text{argmin}_{\boldsymbol{\beta}_{\mathcal{C}}} \mathbb{E}\, l(U\boldsymbol{\beta}_{\mathcal{C}}, Y), \tag{3.30}$$

which satisfies the equations

$$\mathbb{E}[U^T U \beta_{\mathcal{C}}^{M}] = \mathbb{E}[U^T Y] = \mathbb{E}[U^T V \boldsymbol{\beta}^{\star}]. \tag{3.31}$$

In the following theorem, the following notation is used for linear conditional covariances:

$$\text{Cov}_L(A, M_j|X) = \mathbb{E}\left[(M_j - \mathbb{E}[M_j|X])(A - \mathbb{E}[A|X])\right], \text{ and}$$

$$\text{Cov}_L(Y, M_j|A, X) = \mathbb{E}\left[(Y - \mathbb{E}[Y|A, X])(M_j - \mathbb{E}[M_j|A, X])\right].$$

Theorem 3.3.1 *For $j \in \mathcal{M}$,*

(i) $\alpha_{A,j} = 0$ if and only if $\text{Cov}_L(A, M_j|X) = 0$, and

(ii) the marginal regression parameters $\beta_{M,j}^{M} = 0$ if and only if $\text{Cov}_L(Y, M_j|A, X) = 0$.

Proof (i) Sufficiency:

Suppose $\alpha_{A,j} = 0$. The marginal regression coefficients $\boldsymbol{\alpha}_j$ satisfy the score equations

$$\mathbb{E}[U^T U \boldsymbol{\alpha}_j] = \mathbb{E}[U^T M_j] = \mathbb{E}[U^T U \boldsymbol{\alpha}_j^{\star}]. \tag{3.32}$$

The baseline parameter $\boldsymbol{\alpha}_{\mathcal{C}j} = \text{argmin}_{\alpha_{\mathcal{C}j}} \mathbb{E}\left[\ell(U_{\mathcal{C}}\boldsymbol{\alpha}_{\mathcal{C}j}, \boldsymbol{M}^{(j)})\right]$, where $U_{\mathcal{C}} = [\mathbb{1}_n, X]$, satisfies

$$\mathbb{E}\left[U_{\mathcal{C}}^T U_{\mathcal{C}}\boldsymbol{\alpha}_{\mathcal{C}j}\right] = \mathbb{E}\left[U_{\mathcal{C}}^T \boldsymbol{M}^{(j)}\right] = \mathbb{E}\left[U_{\mathcal{C}}^T U_{\mathcal{C}}\boldsymbol{\alpha}_j^\star\right], \tag{3.33}$$

which can be rewritten

$$\mathbb{E}\left[U_{\mathcal{C}}^T(\boldsymbol{M}^{(j)} - \mathbb{E}_L[\boldsymbol{M}^{(j)}|U_{\mathcal{C}})]\right] = 0. \tag{3.34}$$

When $\alpha_{A,j} = 0$, the first q components of $\boldsymbol{\alpha}_j$ are equal to $\boldsymbol{\alpha}_{\mathcal{C}j}$ by uniqueness of equation (3.33). Performing a similar manipulation to equation (3.32) for the treatment A results in

$$\mathbb{E}\left[A^T(U_{\mathcal{C}}\boldsymbol{\alpha}_{\mathcal{C}j})\right] = \mathbb{E}\left[A^T\boldsymbol{M}^{(j)}\right],$$

or

$$\mathbb{E}\left[A^T(\boldsymbol{M}^{(j)} - \mathbb{E}_L\left[\boldsymbol{M}^{(j)}|U_{\mathcal{C}}\right])\right] = 0.$$

Combining this expression with equation (3.34) then yields

$$\text{Cov}_L(\boldsymbol{M}^{(j)}, A|U_{\mathcal{C}}]) = \mathbb{E}\left[(A - \mathbb{E}_L[A|U_{\mathcal{C}}])^T(\boldsymbol{M}^{(j)} - \mathbb{E}_L[\boldsymbol{M}^{(j)}|U_{\mathcal{C}}]\right] = 0,$$

as needed.

Necessity: Suppose $\text{Cov}_L(\boldsymbol{M}^{(j)}, A|U_{\mathcal{C}}) = 0$, which is equivalent to

$$\mathbb{E}[A^T U_{\mathcal{C}}\boldsymbol{\alpha}_{\mathcal{C}j}] = \mathbb{E}\, A^T\boldsymbol{M}^{(j)}.$$

This equivalency, along with equation (3.32), imply that $((\boldsymbol{\alpha}_{\mathcal{C}})^T, 0)^T$ is a solution to the score equations (Equation 3.33). Uniqueness of this solution implies that

$$\boldsymbol{\alpha}_j = ((\boldsymbol{\alpha}_{\mathcal{C}j})^T, 0)^T \Rightarrow \alpha_{A,j} = 0.$$

(ii) Sufficiency: Suppose $\beta^M_{M,j} = 0$. Since the marginal regression coefficients $\beta^M_{\mathcal{C}j}$ satisfy the score equations

$$\mathbb{E}[Z_j^T Z_j \beta^M_{\mathcal{C}j}] = \mathbb{E}[Z_j^T Y] = \mathbb{E}[Z_j^T V \boldsymbol{\beta}^\star],$$

the first $q+2$ components of $\boldsymbol{\beta}^M_{\mathcal{C}j}$ should be equal to $\boldsymbol{\beta}^M_{\mathcal{C}}$ by uniqueness of equation (3.31). Then the above expression becomes

$$\mathbb{E}[M_j^T U \beta^M_{\mathcal{C}}] = \mathbb{E}[M^{(j)T} Y], \text{ or } \mathbb{E}_L[M^{(j)T}(Y - \mathbb{E}[Y|U])] = 0,$$

which can then be expressed as

$$\mathrm{Cov}_L\left(Y, M_j \mid U\right) \equiv \mathbb{E}\left(M_j - \mathbb{E}_L\left(M_j \mid U\right)\right)\left(Y - \mathbb{E}_L\left(Y \mid U\right)\right) = 0.$$

Necessity: Suppose now $\mathrm{Cov}_L(Y, M_j|U) = 0$, which is equivalent to

$$\mathbb{E}[M^{(j)T} U \boldsymbol{\beta}^M_{\mathcal{C}}] = \mathbb{E}[M^{(j)T} Y].$$

This equivalency, along with Equation (3.31), imply that $((\boldsymbol{\beta}_{\mathcal{C}}^{M})^T, 0)^T$ is a solution to the score equations (3.29). Uniqueness of this solution implies that

$$\boldsymbol{\beta}_{\mathcal{C}j}^{M} = ((\boldsymbol{\beta}_{\mathcal{C}}^{M})^T, 0)^T \Rightarrow \beta_{M,j}^{M} = 0.$$

■

Indeed at the population level in order for a mediator to be identified, it should be both linearly correlated with A given X and with Y given (X, A). Further, if the mediator is conditionally correlated with A given X and with Y given (X, A), then the corresponding marginal regression coefficients will be nonzero. The following condition is required to guarantee minimum signal strength of both treatment-mediator effects and mediator-response effects for important mediators, which will be proven in the following theorem.

Condition 1 *1. There exist positive constants $c_1, c_2 > 0$ and $\kappa_1, \kappa_2 < \frac{1}{2}$ such that for all $j \in \mathcal{M}^\star$, $|\operatorname{Cov}_L(M_j, A|X)| \geq c_1 n^{-\kappa_1}$ and $|\operatorname{Cov}_L(Y, M_j|U)| \geq c_2 n^{-\kappa_2}$*

2. $\mathbb{E}\, X_{ik}^2 \leq c_3$ uniformly in $k = 1, ..., q$

3. $\mathbb{E}\, M_{ij}^2 \leq c_4$ uniformly in $j = 1, ..., p$

Theorem 3.3.2 *If Condition 1 holds, then there exist $c_3, c_4 > 0$ such that*

(i) $\min_{j \in \mathcal{M}^\star} |\alpha_{A,j}| \geq c_3 n^{-\kappa_1}$, *and*

(ii) $\min_{j \in \mathcal{M}^\star} |\beta_{M,j}^{M}| \geq c_4 n^{-\kappa_2}$.

Proof (i) By the score equations

$$\mathbb{E}[U_{\mathcal{C}}^T U \boldsymbol{\alpha}_j] = \mathbb{E}[U_{\mathcal{C}}^T U_{\mathcal{C}} \boldsymbol{\alpha}_{\mathcal{C}j}],$$

which can be rewritten as

$$\mathbb{E}\left[U_{\mathcal{C}}^T (U\boldsymbol{\alpha}_j - U_{\mathcal{C}}\boldsymbol{\alpha}_{\mathcal{C}j})\right] = 0.$$

Let $\boldsymbol{\alpha}_{\Delta,j} = \boldsymbol{\alpha}_{\mathcal{C}j1} = \boldsymbol{\alpha}_{\mathcal{C}j}$ denote the subvector created from the first $q+1$ entries of $\boldsymbol{\alpha}_j$. Then the previous expression can be rewritten as

$\mathbb{E}[U_{\mathcal{C}}^T (U_{\mathcal{C}}\boldsymbol{\alpha}_{\Delta,j} + A\boldsymbol{\alpha}_{A,j})] = 0$. Solving for $\boldsymbol{\alpha}_{\Delta,j}$ results in

$$\boldsymbol{\alpha}_{\Delta,j} = -(U_{\mathcal{C}}^T U_{\mathcal{C}})^{-1} U_{\mathcal{C}}^T A \alpha_{A,j}. \tag{3.35}$$

Observe next the linear covariance

$$\mathrm{Cov}_L(\boldsymbol{M}^{(j)}, A|U_{\mathcal{C}}) = \mathbb{E}[A^T(U_{\mathcal{C}}\boldsymbol{\alpha}_{\Delta,j} + A\alpha_j)] = U_j^T U_{\mathcal{C}}\boldsymbol{\alpha}_{\Delta,j} + U_j^T A\alpha_{A,j}.$$

Plugging (3.35) into the previous expression then yields

$$\mathrm{Cov}_L(\boldsymbol{M}^{(j)}, A|U_{\mathcal{C}}) = (U_j^T A - U_j^T U_{\mathcal{C}}(U_{\mathcal{C}}^T U_{\mathcal{C}})^{-1} U_{\mathcal{C}}^T A)\alpha_{A,j}. \tag{3.36}$$

Under Condition 1, taking the modulus of both sides of (3.36) then results in

$$|\alpha_{A,j}| \geq c^{-1} |\operatorname{Cov}_L(\boldsymbol{M}^{(j)}, A|U_{\mathcal{C}}| \geq c_3 n^{-\kappa_1}/c,$$

where $c = (U_j^T A - U_j^T U_{\mathcal{C}} (U_{\mathcal{C}}^T U_{\mathcal{C}})^{-1} U_{\mathcal{C}}^T A)$. Taking minimum over all $j \in \mathcal{M}^\star$ results in the conclusion.

(ii) Let $\Omega_{2,j} = \begin{bmatrix} \mathbb{E}\, U^T U & \mathbb{E}\, U^T M_j \\ \mathbb{E}\, M_j^T U & \mathbb{E}\, M_j^T M_j \end{bmatrix} = \begin{bmatrix} \Omega_{\mathcal{C}\mathcal{C}} & \Omega_{\mathcal{C}j} \\ \Omega_{j\mathcal{C}} & \Omega_{jj} \end{bmatrix}$.

By score equations (3.29) and (3.31)

$$\mathbb{E}[U^T U \beta_{\mathcal{C}}^M] = \mathbb{E}[U^T Z_j \beta_{\mathcal{C}j}^M)],$$

$$\Rightarrow \mathbb{E}[U^T U \beta_{\mathcal{C}}^M)] - \mathbb{E}[U^T Z_j \beta_{\mathcal{C}j}^M] = \mathbb{E}[U^T (U\beta_{\mathcal{C}}^M - Z_j \beta_{\mathcal{C}j}^M)] = 0.$$

Let $\boldsymbol{\beta}_{\Delta,j} = \boldsymbol{\beta}_{\mathcal{C}j1}^M - \boldsymbol{\beta}_{\mathcal{C}}^M$. Then the previous expression can be rewritten

$$\mathbb{E}\left[U^T (U\boldsymbol{\beta}_{\Delta,j}^M + M_j \boldsymbol{\beta}_{M,j}^M)\right] = 0,$$

$$\Rightarrow \boldsymbol{\beta}_{\Delta,j} = -\Omega_{\mathcal{C}\mathcal{C}}^{-1} \Omega_{\mathcal{C}j} \boldsymbol{\beta}_{M,j}^M.$$

and $\operatorname{Cov}_L(Y, M_j|U) = \mathbb{E}[M_j(Y - \mathbb{E}_l[Y|U)]]$. Using the score equations, this expression can then be rewritten as

$$\operatorname{Cov}_L(Y, M_j|U) = \mathbb{E}[M_j(Z_j \boldsymbol{\beta}_{\mathcal{C}j}^M - U\boldsymbol{\beta}_{\mathcal{C}}))]. \tag{3.37}$$

Again using the definition of $\boldsymbol{\beta}_{\Delta,j}$, this becomes

$$\mathrm{Cov}_L(Y, M_j|U) = \mathbb{E}[M_j(Z_j^T \boldsymbol{\beta}_{\mathcal{C}j}^M - U^T \boldsymbol{\beta}_{\mathcal{C}}^M)],$$

$$= \mathbb{E}[M_j(U^T \boldsymbol{\beta}_{\Delta,j}^M + M_j^T \boldsymbol{\beta}_{M,j}^M)] = \Omega_{\mathcal{C}j}^T \boldsymbol{\beta}_{\Delta,j} + \Omega_{jj}\beta_{M,j}^M,$$

$$\Rightarrow \mathrm{Cov}_L(Y, M_j|U) = (\Omega_{jj} - \Omega_{\mathcal{C}j}^T \Omega_{\mathcal{C}\mathcal{C}}^{-1} \Omega_{\mathcal{C}j})\beta_{M,j}^M.$$

By Condition 1, $|\beta_{M,j}^M| \geq c^{-1}|\,\mathrm{Cov}_L(Y, M_j|U)| \geq c^{-1}c_4 n^{-\kappa_2}$. Finally, taking the minimum over all $j \in \mathcal{M}^\star$ yields the conclusion.

■

Corollary 1 *If Condition 1 holds, then there exists $c_5 > 0$ such that*

$$\min_{j \in \mathcal{M}^\star} |\alpha_{A,j}\beta_{M,j}^M| \geq c_5 n^{-2\kappa}.$$

Proof Under Condition 1, by Theorem 3.3.2

$$\min_{j \in \mathcal{M}^\star} |\alpha_{A,j}\beta_{M,j}^M| = (\min_{j \in \mathcal{M}^\star} |\alpha_{A,j}||\beta_{M,j}^M|) \geq (\min_{j \in \mathcal{M}^\star} |\alpha_{A,j}|)(\min_{j \in \mathcal{M}^\star} |\beta_{M,j}^M|) \geq c_3 c_4 n^{-2\kappa} := c_5 n^{-2\kappa}.$$

■

3.3.2 Sample-Level Properties

Next, we prove the uniform convergence of the mediation marginal maximum likelihood estimator and the sure screening property of the mediation sure independence screening

method. Further, we establish an upper bound for the size of the set $\hat{\mathcal{M}}$ selected by our method.

The log-likelihood of the linear model is concave and thus has a unique minimizer over $\boldsymbol{\beta}_{\mathcal{C}j} \in \mathcal{B}$ at an interior point $\beta^M_{\mathcal{C}j}$ where $\mathcal{B}(B) = \{\boldsymbol{\beta} \in \mathbb{R}^{q+3} : |\beta_1| \leq B, ..., |\beta_{q+2}| \leq B, |\beta_{q+3}| \leq B\}$ for sufficiently large B is the parameter space over which we maximize the marginal likelihood. At the sample level, a few more conditions are required to obtain uniform convergence.

Condition 2 *1. The operator norm $\|I_j(\boldsymbol{\beta}_{\mathcal{C}j})\|_{\mathcal{B}}$ of the Fisher Information $I_j(\boldsymbol{\beta}_{\mathcal{C}j}) = \mathbb{E}\left[Z_j Z_j^T\right]$ is bounded, where $\|I_j(\boldsymbol{\beta}_{\mathcal{C}j})\|_{\mathcal{B}} = \sup_{\beta_{\mathcal{C}j} \in \mathcal{B}, \|Z_j\|=1} \|I_j(\boldsymbol{\beta}_{\mathcal{C}j})^{1/2} Z_j\|$.*

2. There exist some positive constants r_0, r_1, s_0, s_1, and a such that

$$\mathbb{P}(|M_j| > t) \leq r_1 \exp(-r_0 t^a) \text{ for } j = 1, ..., p,$$

for sufficiently large t and that

$$\mathbb{E}\left[\exp(b(V\boldsymbol{\beta}^\star + s_0) - b(V\boldsymbol{\beta}^\star))\right] + \mathbb{E}\left[\exp(b(V^T\boldsymbol{\beta}^\star - s_0) - b(V\boldsymbol{\beta}^\star))\right] \leq s_1,$$

where $b(\eta)$ denotes the log-partition function for the canonical form of the likelihood of Y given (X, A, M).

3. There exists an $\epsilon_1 > 0$ such that for all $j \in \mathcal{M}$,

$$\sup_{\boldsymbol{\beta}_{\mathcal{C}_j} \in \mathcal{B}, \left\|\boldsymbol{\beta}_{\mathcal{C}j} - \boldsymbol{\beta}^M_{\mathcal{C}_j}\right\| \leq \varepsilon_1} \left| \mathbb{E}\left[\left(\frac{1}{2} U \boldsymbol{\beta}_{\mathcal{C}_j}\right)^2 I\left(|\mathbf{M}_j| > K_n\right) \right] \right| \leq o\left(n^{-1}\right)$$

where K_n is an arbitrarily large constant such that for a given $\boldsymbol{\beta} \in \mathcal{B}$, the function $\ell(Z_j\beta, y)$ is Lipschitz for all (z_j, y) in $\Lambda_n = \{z_j, y : \|z_j\|_\infty \leq K_n, |y| \leq K_n^\star\}$ with $K_n^\star = r_0 K_n^\alpha / s_0$.

4. *For all $\boldsymbol{\beta}_{\mathcal{C}j} \in \mathcal{B}$, we have*

$$\mathbb{E}(l(Z_j\boldsymbol{\beta}_{\mathcal{C}j}, Y) - l(Z_j\boldsymbol{\beta}_{\mathcal{C}j}^M, Y)) \geq S_n \|\boldsymbol{\beta}_{\mathcal{C}j} - \boldsymbol{\beta}_{\mathcal{C}j}^M\|^2$$

for some positive constant S_n, uniformly over $j = 1, ..., p$.

The following theorem uses these conditions to establish the uniform convergence of our conditional likelihood estimators and the sure screening property for our approach.

Theorem 3.3.3 *Under Condition 2, it holds that for any $t > 0$,*

$$\mathbb{P}(\sqrt{n}\|\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M - \boldsymbol{\beta}_{\mathcal{C}j}^\star\| \geq 16k_n(1+t)/S_n) \leq \exp\{-2t^2/K_n^2\}_n \, \mathbb{P}(\Lambda_n^c).$$

where $\boldsymbol{\beta}_{\mathcal{C}j}^\star$ is the subvector of $\boldsymbol{\beta}^\star$ whose indices correspond to those of $\hat{\boldsymbol{\beta}}_{\mathcal{C}j}^M$.

Proof To prove this statement we follow the approach of [48] in proving their Lemma 1. This approach uses several lemmas including the symmetrization theorem in [55], contraction theorem in [56], and concentration theorem of [57]. We return to this proof after going through these theorems, as well as Lemmas 1 and 5 from [48]. ■

Lemma 1 (Symmetrization; Lemma 2.3.1 [55]) *Let $Z_1, ..., Z_n$ be independent random variables with values in $\mathcal{Z}$ and $\mathcal{F}$ is a class of real valued functions on $\mathbb{Z}$. Then*

$$\mathbb{E}\left\{\sup_{f\in\mathcal{F}}|(\mathbb{P}_n - P)f(Z)|\right\} \leq 2\,\mathbb{E}\left\{\sup_{f\in\mathcal{F}}|\,\mathbb{P}_n\,\epsilon f(Z)|\right\},$$

where $\epsilon_1, ..., \epsilon_n$ is a Rademacher sequence (i.e. an i.i.d. sequence taking values ± 1 with probability .5) independent of $Z_1, ..., Z_n$ and $Pf(Z) = \mathbb{E}\,f(Z)$.

Lemma 2 (Contraction Theorem; [56]) *Let $z_1, ..., z_n$ be nonrandom elements of some space $\mathcal{Z}$, and let $\mathcal{F}$ be a class of real valued functions on $\mathcal{Z}$. Let $\epsilon_1, ..., \epsilon_n$ be a Rademacher sequence. Consider Lipschitz functions $g_i : \mathbb{R} \mapsto \mathbb{R}$ (i.e. satisfying $|g_i(s) - g_i(\tilde{s})| \leq |s - \tilde{s}|\ \ \forall s, \tilde{s} \in \mathbb{R}$ Then for any function $\tilde{f} : \mathcal{Z} \mapsto \mathbb{R}$,*

$$\mathbb{E}\left\{\sup_{f\in\mathcal{F}}|\,\mathbb{P}_n\,\epsilon(g(f) - g(\tilde{f}))|\right\} \leq 2\,\mathbb{E}\left\{\sup_{f\in\mathcal{F}}|\,\mathbb{P}_n\,\epsilon(f - \tilde{f})|\right\}.$$

Lemma 3 (Concentration Theorem [57]) *Let $Z_1, ..., Z_n$ be independent random variables with values in some space $\mathcal{Z}$ and let $\gamma \in \Gamma$, a class of real valued functions on $\mathcal{Z}$. Assume that for some positive constants $l_{i,\gamma}$ and $u_{i,\gamma}, l_{i,\gamma} \leq \gamma(Z_i) \leq u_{i,\gamma}\ \forall \gamma \in \Gamma$. Define*

$$L^2 = \sup_{\gamma\in\Gamma}\sum_{i=1}^{n}(u_{i,\gamma} - l_{i,\gamma})^2/n$$

and

$$\mathbf{Z} = \sup_{\gamma\in\Gamma}|(\mathbb{P}_n - P)\gamma(Z)|.$$

Then for any $t > 0$,

$$P(\mathbf{Z} \geq \mathbb{E}\,\mathbf{Z} + t) \leq \exp\left(-\frac{nt^2}{2L^2}\right).$$

For $N > 0$, let $\mathcal{B}(N) = \{\beta \in \mathcal{B} : \|\beta - \beta^{\star}_{\mathcal{C}j}\| \leq N\}$. Let

$$\mathbb{G}_1(N) = \sup_{\boldsymbol{\beta} \in \mathcal{B}(N)} |(\mathbb{P}_n - \mathbb{P}) \left\{ l(Z_j \boldsymbol{\beta}^{\star}_{\mathcal{C}j}, Y \right\} I_n(Z_j, Y)|,$$

where $I_n(Z_j, Y) = I((Z_j, Y) \in \Lambda_n)$ with Λ_n is defined in Condition 2. The following lemma is about the upper bound of the tail probability for $\mathbb{G}_1(N)$ in the neighborhood of $\mathcal{B}(N)$.

Lemma 4 (Lemma 5 from [48]) *For all $t > 0$, it holds that*

$$\mathbb{P}(\mathbb{G}_1(N) \geq 4Nk_n(q/n)^{1/2}(1 + t)) \leq \exp(-2t^2/K_n^2).$$

Lemma 5 (Lemma 1 from [48]) *If Condition 2.2 holds, then for any $t > 0$,*

$$\mathbb{P}(|M_j| > r_0 t^a / s_0) \leq s_1 \exp(-r_0 t^a).$$

Given these lemmas, we return to the proof of Theorem 3.3.3:

Proof [Proof of Theorem 3.3.3] Step 1: Bound $\|\hat{\boldsymbol{\beta}}^M_{\mathcal{C}j} - \boldsymbol{\beta}^{\star}_{\mathcal{C}j}\|$ with $\mathbb{G}(N)$ for some small N chosen such that Conditions 2.3 and 2.4 hold. [48] took inspiration from [58] by defining the convex combination $\boldsymbol{\beta}_s = s\hat{\boldsymbol{\beta}}^M_{\mathcal{C}j} + (1 - s)\boldsymbol{\beta}^{\star}_{\mathcal{C}j}$, with

$$s = (1 + \|\hat{\boldsymbol{\beta}}^M_{\mathcal{C}j} - \boldsymbol{\beta}^{\star}_{\mathcal{C}j}\|/N)^{-1}.$$

Then by definition

$$\|\boldsymbol{\beta}_s - \boldsymbol{\beta}^\star_{\mathcal{C}j}\| = s\|\hat{\boldsymbol{\beta}}^M_{\mathcal{C}j} - \boldsymbol{\beta}^\star_{\mathcal{C}j}\| \leq N,$$

so $\boldsymbol{\beta}_s \in \mathcal{B}(N)$. By convexity

$$\mathbb{P}_n\, l(Z_j\boldsymbol{\beta}_s, Y) \leq s\,\mathbb{P}_n\, l(Z_j\hat{\boldsymbol{\beta}}^M_{\mathcal{C}j}, Y) + (1-s)\,\mathbb{P}_n\, l(Z_j\boldsymbol{\beta}^\star_{\mathcal{C}j}, Y)$$

$$\leq \mathbb{P}_n\, l(Z_j\boldsymbol{\beta}^\star_{\mathcal{C}j}, Y).$$

Since $\boldsymbol{\beta}^\star_{\mathcal{C}j}$ is the minimizer

$$\mathbb{E}\left[l(Z_j\boldsymbol{\beta}_s, Y) - l(Z_j\boldsymbol{\beta}^\star_{\mathcal{C}j}, Y)\right] \geq 0.$$

Putting these together yields

$$\begin{aligned}
&\mathbb{E}\left[l\left(Z_j\boldsymbol{\beta}_s, Y\right) - l\left(Z_j\boldsymbol{\beta}^\star_{\mathcal{C}j}, Y\right)\right] \\
&\quad \leq (\mathbb{E} - \mathbb{P}_n)\left[l\left(Z_j\boldsymbol{\beta}_s, Y\right) - l\left(Z_j\boldsymbol{\beta}^\star_{\mathcal{C}j}, Y\right)\right] \\
&\quad \leq \mathbb{G}(N),
\end{aligned}$$

where $\mathbb{G}(N) = \sup_{\beta\in\mathcal{B}(N)}\left|(\mathbb{P}_n - P)\left\{l\left(Z_j\beta, Y\right) - l\left(Z_j\beta^\star_{\mathcal{C}j}, Y\right)\right\}\right|$. By Condition 2.4,

$$\|\boldsymbol{\beta}_s - \boldsymbol{\beta}^\star_{\mathcal{C}j}\| \leq [\mathbb{G}(N)/S_n]^{1/2}. \tag{3.38}$$

Note that for any x,

$$\mathbb{P}(\|\boldsymbol{\beta}_s - \boldsymbol{\beta}^\star_{\mathcal{C}j}\| \geq x) \leq \mathbb{P}(\mathbb{G}(N) \geq S_n x^2).$$

Setting $x = N/2$, we have

$$\mathbb{P}(\|\boldsymbol{\beta}_s - \boldsymbol{\beta}^\star_{\mathcal{C}j}\| \geq N/2) \leq \mathbb{P}(\mathbb{G}(N) \geq S_n N^2/4).$$

By definition of $\boldsymbol{\beta}_s$, the left hand side of the previous expression is equal to $\mathbb{P}(\|\hat{\boldsymbol{\beta}}^M_{\mathcal{C}j} - \boldsymbol{\beta}^\star_{\mathcal{C}j}\| \geq N)$.

Taking $N = 4a_n(1+t)/S_n$ with $a_n = 4k_n\sqrt{q/n}$, we have

$$\begin{aligned} P\left\{\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star_{\mathcal{C}j}\right\| \geq N\right\} &\leq P\left\{\mathbb{G}(N) \geq V_n N^2/4\right\} \\ &= P\left\{\mathbb{G}(N) \geq N a_n(1+t)\right\}, \end{aligned}$$

where the last probability is bounded by

$$P\left\{\mathbb{G}(N) \geq N a_n(1+t), \Lambda_n\right\} + P\left\{\Lambda_n^c\right\}, \tag{3.39}$$

where $\Lambda_n = \{\|Z_{j,(i)}\| \leq K_n, |Y_i| \leq K_n^\star\}$ as in Condition 2.

On the set Λ_n, since

$$\sup_{\beta \in \mathcal{B}(N)} \mathbb{P}_n \left| l\left(Z_j\boldsymbol{\beta}, Y\right) - l\left(Z_j\boldsymbol{\beta}^\star_{\mathcal{C}j}, Y\right)\right| \left(1 - I_n(Z_j, Y)\right) = 0,$$

by the triangle inequality

$$\mathbb{G}(N) \leq \mathbb{G}_1(N) + \sup_{\beta \in \mathcal{B}(N)} \left| E\left[l\left(Z_j\beta, Y\right) - l\left(Z_j\beta_0, Y\right)\right]\left(1 - I_n(Z_j, Y)\right)\right|.$$

Then by Condition 2.3, the upper bound in Equation (3.39) is bounded by

$$P\left\{\mathbb{G}_1(N) \geq Na_n(1+t) + o(q/n)\right\} + nP\left\{(Z_j, Y) \in \Omega_n^c\right\}.$$

Finally, the conclusion follows from Lemma 4. ■

Theorem 3.3.4 *Suppose that Condition 2 holds. Let* $k_n = b'(K_nB(q+1)) + r_0K_n^a/s_0$ *with* K_n *given in Condition 2, and assume that* $n^{1-2\kappa}k_n^{-2}K_n^{-2} \to \infty$.

1. *For any* $c_3 > 0$, *there exists a positive constant* c_4 *such that*

$$\mathbb{P}\left(\max_{q+1\leq j\leq p} |\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq c_3 n^{-\kappa}\right) \leq 2d\exp\{-c_4 n^{1-2\kappa}\}.$$

2. *For any* $c_5 > 0$, *there exists a positive constant* c_6 *such that*

$$\mathbb{P}\left(\max_{q+1\leq j\leq p} |\hat{\beta}_j^M - \beta_{M,j}^M| \geq c_5 n^{-\kappa}\right) \leq d\exp(-c_6 n^{1-2\kappa}(k_nK_n)^{-2}) + dnr_2\exp(-r_0K_n^a),$$

where $r_2 = qr_1 + s_1$.

3. *For any* $c_7 > 0$, *there exists a positive constant* c_8 *such that*

$$\begin{aligned}\mathbb{P}\left(\max_{q+1\leq j\leq p} |\hat{\alpha}_j\hat{\beta}_j^M - \alpha_{A,j}\beta_{M,j}^M| \geq c_7 n^{-\kappa}\right) \leq & 2d\exp\{-cn^{1-2\kappa}\} + d\exp\{-cn^{1-2\kappa}(k_nK_n)^{-2}\} \\ & + dnr_2\exp\{-r_0K_n^a\} + 2d\exp\{-cn^{1-\kappa}\},\end{aligned}$$

where $r_2 = qr_1 + s_1$.

4. If Condition 1 also holds, then by taking $\lambda = c_5 n^{-\kappa}$ with $c_5 \leq c_3/2$, we have

$$\mathbb{P}(\mathcal{M}^\star \subset \hat{\mathcal{M}}_\lambda) \geq 1 - s\exp(-c_4 n^{1-2\kappa})(k_n K_n)^{-2}) - nr_2 s\exp(-r_0 K_n^a),$$

for some constant c_5, where $s = |\mathcal{M}^\star|$ is the size of the set of true mediators

Proof 1. By our model assumptions, $\hat{\alpha}_j \sim \mathcal{N}\left(\alpha_{A,j}, (U^T U)^{-1}_{(q+2),(q+2)}\sigma_M^2\right)$ where $(U^T U)^{-1}_{(q+2),(q+2)}$ denotes the bottom-right corner entry of $(U^T U)^{-1}$. This can be expressed in closed form as

$$(U^T U)^{-1}_{(q+2),(q+2)} = \left\{A^T(I - U_C(U_C^T U_C)^{-1}U_C^T)A\right\}^{-1}.$$

Thus,

$$\mathbb{P}(|\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq c_3 n^{-\kappa}) =$$

$$= \mathbb{P}\left(\frac{|\hat{\alpha}_{A,j} - \alpha_{A,j}|}{\sqrt{\left\{A^T(I - U_C(U_C^T U_C)^{-1}U_C^T)A\right\}^{-1}\sigma_M^2}} \geq \frac{c_3 n^{-\kappa}}{\sqrt{\left\{A^T(I - U_C(U_C^T U_C)^{-1}U_C^T)A\right\}^{-1}\sigma_M^2}}\right)$$

$$= 2\left(1 - \Phi\left(\frac{c_3 n^{-\kappa}}{\sqrt{\left\{A^T(I - U_C(U_C^T U_C)^{-1}U_C^T)A\right\}^{-1}\sigma_M^2}}\right)\right).$$

Conditional on (X, A), $A^T(I - U_C(U_C^T U_C)^{-1}U_C^T)A = n\,\mathrm{Var}[A|U_{\mathcal{C}}]$, so the above is

$$= 2\left(1 - \Phi\left(\frac{c_3 n^{-\kappa}}{\sqrt{(n\,\mathrm{Var}[A|U_{\mathcal{C}}])^{-1}\sigma_M^2}}\right)\right)$$

$$= 2\left(1 - \Phi\left(\frac{c_3 n^{1/2-\kappa}}{\sqrt{(\mathrm{Var}[A|U_{\mathcal{C}}])^{-1}\sigma_M^2}}\right)\right) \xrightarrow{n\to\infty} 0.$$

In particular, using the Chernoff bound for Gaussian random variables yields

$$\mathbb{P}(\hat{\alpha}_{A,j} \geq \alpha_{A,j} + t) \leq \exp\{-t^2/(2\,\mathrm{Var}(\hat{\alpha}_{A,j})\}$$

$$\Rightarrow \mathbb{P}(|\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq t) \leq 2\exp\{-t^2/(2\left\{A^T(I - U_C(U_C^TU_C)^{-1}U_C^T)A\right\}^{-1}\sigma_M^2)\}$$

$$= 2\exp\{-t^2 n/(2\,\mathrm{Var}[A|U_C]\sigma_M^2\}$$

$$\Rightarrow \mathbb{P}(|\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq c_3 n^{-\kappa}) \leq 2\exp\{-(c_3 n^{-\kappa})^2 n/(2\,\mathrm{Var}[A|U_C]\sigma_M^2\} = 2\exp\{-c_4 n^{1-2\kappa}\}.$$

Finally, the conclusion follows from applying Bonferonni's inequality to obtain

$$\mathbb{P}\left(\max_{q+1\leq j\leq p} |\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq c_3 n^{-\kappa}\right) \leq 2d\exp\{-c_4 n^{1-2\kappa}\}.$$

2. Let the event $\Lambda_n = \{\|M\|_\infty \leq K_n; |Y| \leq K_n^\star\}$. By Lemma 5, Condition 2.2 implies the bound

$$\mathbb{P}(|Y| \geq u) \leq s_1\exp\{-s_0 u\}.$$

Looking at the complementary event to Λ_n then yields

$$\mathbb{P}(\Lambda_n^c) \leq \mathbb{P}(\|M\|_\infty > K_n) + \mathbb{P}(|Y| \geq K_n^\star) \leq r_2\exp\{-r_0 K_n^a\}.$$

Combining the previous statement with Theorem 3.3.3, let $1+t = cS_n n^{-1/2-\kappa}/(16k_n)$, we have

$$\begin{aligned}\mathbb{P}\left(|\hat{\beta}^M_{M,j} - \beta^M_{M,j}| \geq c_3 n^{-\kappa}\right) &\leq \mathbb{P}\left(\|\hat{\beta}^M_{\mathcal{C}j} - \beta^M_{\mathcal{C}j}\| \geq c_3 n^{-\kappa}\right) \\ &\leq \exp\left(-c_4 n^{1-2\kappa}/(k_n K_n)^2\right) + nr_2 \exp(-r_0 K_n^a)\end{aligned}$$

for some constant $c_4 > 0$ for all $j \in \mathcal{M}^\star$. Applying Bonferroni's inequality then yields

$$\mathbb{P}\left(\max_{q+1\leq j\leq p} |\hat{\beta}^M_{M,j} - \beta^M_{M,j}| \geq c_3 n^{-\kappa}\right) \leq d\left(\exp\left(-c_4 n^{1-2\kappa}/(k_n K_n)^2\right) + nr_2 \exp(-r_0 K_n^a)\right),$$

which proves the statement.

3. Since

$$\begin{aligned}|\hat{\alpha}_{A,j}\hat{\beta}^M_{M,j} - \alpha_{A,j}\beta^M_{M,j}| &= |\hat{\alpha}_{A,j}\hat{\beta}^M_{M,j} - \hat{\alpha}_{A,j}\beta^M_{M,j} + \hat{\alpha}_{A,j}\beta^M_{M,j} - \alpha_{A,j}\beta^M_{M,j}| \\ &\leq |\hat{\alpha}_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| + |\beta^M_{M,j}||\hat{\alpha}_{A,j} - \alpha_{A,j}|,\end{aligned}$$

by the triangle inequality,

$$\begin{aligned}&= |\hat{\alpha}_{A,j} - \alpha_{A,j} + \alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| + |\beta^M_{M,j}||\hat{\alpha}_{A,j} - \alpha_{A,j}| \\ &\leq |\hat{\alpha}_{A,j} - \alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| + |\alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| \\ &+ |\beta^M_{M,j}||\hat{\alpha}_{A,j} - \alpha_{A,j}|,\end{aligned}$$

again by application of the triangle inequality. Then $\mathbb{P}\left(|\hat{\alpha}_{A,j}\hat{\beta}^M_{M,j} - \alpha_{A,j}\beta^M_{M,j}| \geq cn^{-\kappa}\right)$ is bounded above by

$$\mathbb{P}\left(|\hat{\alpha}_{A,j} - \alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| + |\alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| + |\beta^M_{M,j}||\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq cn^{-\kappa}\right),$$

which is further bounded above by

$$\mathbb{P}\left(|\hat{\alpha}_{A,j} - \alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| \geq cn^{-\kappa}\right) + \mathbb{P}\left(|\alpha_{A,j}||\hat{\beta}^M_{M,j} - \beta^M_{M,j}| \geq cn^{-\kappa}\right)$$
$$+ \mathbb{P}\left(|\beta^M_{M,j}||\hat{\alpha}_{A,j} - \alpha_{A,j}| \geq cn^{-\kappa}\right).$$

Parts (1) and (2) of this theorem give us bounds for the last two terms, so we must still bound $\mathbb{P}\left(|\hat{\alpha}_{A,j} - \alpha||\hat{\beta}^M_j - \beta^M_{M,j}| \geq cn^{-\kappa}\right)$. Since

$$\mathbb{P}\left(|\hat{\alpha}_{A,j} - \alpha||\hat{\beta}^M_j - \beta^M_{M,j}| \geq cn^{-\kappa}\right) = \mathbb{P}\left(\left|\left(\frac{(\hat{\alpha}_{A,j} - \alpha)}{\sigma_\alpha}\right)\left(\frac{\hat{\beta}^M_j - \beta^M_{M,j}}{\sigma_\beta}\right)\right| \geq \frac{cn^{-\kappa}}{\sigma_{\alpha,j}\sigma_{\beta,j}}\right),$$

where $\sigma^2_{\alpha,j}$ and $\sigma^2_{\beta,j}$ denote the variances of $\hat{\alpha}_{A,j}$ and $\hat{\beta}_j$, respectively. This bound can be found using the distribution of the product of two Gaussian variables. Letting $W = \left(\frac{(\hat{\alpha}_{A,j} - \alpha)}{\sigma_\alpha}\right)\left(\frac{\hat{\beta}^M_j - \beta^M_{M,j}}{\sigma_\beta}\right)$, W has characteristic function [59]

$$\mathbb{E}[e^{-itW}] = \varphi_W(t) = (1 - i(1+\rho)t)^{-1/2}(1 + i(1-\rho)t)^{-1/2},$$

where $\rho = \text{Corr}(\hat{\alpha}_{A,j}, \hat{\beta}^M_{M,j})$. This allows us to apply the Chernoff bound

$$\mathbb{P}(W \geq a) \leq \frac{\mathbb{E}[e^{tW}]}{e^{ta}} = \frac{\varphi(-it)}{e^{ta}} = \exp\{-ta\} \left\{(1-(1+\rho)t)(1+(1-\rho)t)\right\}^{-1/2},$$

$$\Rightarrow \mathbb{P}(|W| \geq a) \leq 2\exp\{-ta\} \left\{(1-(1+\rho)t)(1+(1-\rho)t)\right\}^{-1/2}.$$

Minimizing the right hand side (or equivalently, its log) yields a minimum at $t^* = \frac{\rho^2 - 2a\rho - 1 + \sqrt{4a^2}}{2a(1-\rho^2}$

Plugging this in, along with $a = cn^{-\kappa}/(\sigma_{\alpha,j}\sigma_{\beta,j}) = c'n^{1-\kappa}$, then yields that $\mathbb{P}\left(\left|\left(\frac{(\hat{\alpha}_{A,j} - \alpha)}{\sigma_\alpha}\right)\left(\frac{\hat{\beta}^M_j -}{\sigma_\beta}\right.\right.\right.$ is bounded above by

$$\exp\left\{\frac{2c'n^{1-\kappa}\rho + 1 - \rho^2 - \sqrt{4c'^2n^{2-2\kappa} + (1-\rho^2)^2}}{2(1-\rho^2)}\right\} \times \left\{1 - 2\rho t^* - (1-\rho^2)t^{*2}\right\}^{-1/2}.$$

Assuming that $|\rho| \neq 1$, the first term of the product is bounded by $\mathcal{O}(\exp\{-cn^{1-\kappa}\})$, while the second term is $\mathcal{O}(1)$. Both terms are positive functions, so the total term is $\mathcal{O}(\exp\{-cn^{1-\kappa}\})$.

Putting these together, we have

$$\begin{aligned}\mathbb{P}(|\hat{\alpha}_{A,j}\hat{\beta}^M_j - \alpha_{A,j}\beta^M_{M,j}| \geq cn^{-\kappa}) \leq & 2\exp\{-cn^{1-2\kappa}\} + \exp\{-cn^{1-2\kappa}(k_nK_n)^{-2}\} \\ & + nr_2\exp\{-r_0K_n^a\} + 2\exp\{-cn^{1-\kappa}\}.\end{aligned}$$

The conclusion then follows from applying Bonferonni's inequality.

4. Consider the event $\mathcal{A}_n = \left\{\max_{j\in\mathcal{M}^\star}\left|\hat{\alpha}_{A,j}\hat{\beta}^M_{M,j} - \alpha_j\beta^M_{M,j}\right| \leq c_3 n^{-2\kappa}/2\right\}$. On A_n, Corollary 1 implies that for all $j \in \mathcal{M}^\star$,

$$|\hat{\alpha}_{A,j}\hat{\beta}^M_{M,j}| \geq c_3 n^{-2\kappa}.$$

Let $\lambda = c_5 n^{-2\kappa} \leq c_3 n^{-2\kappa}/2$. On the event $\mathcal{A}_n$ this leads to $\mathcal{M}^\star \subset \hat{\mathcal{M}}_\lambda$, i.e. we have the sure screening property on $\mathcal{A}_n$. Using the first result and the Bonferonni inequality over all considered j, we obtain the concluding probability statement.

■

After proving the sure screening property, we consider other desirable properties for our method. In particular the following condition allows us to bound the size of the set selected by MSS, which helps us limit the number of false positives selected by the screener.

Condition 3 *1. The variances* $\mathrm{Var}(U\boldsymbol{\alpha}^\star) = \boldsymbol{\alpha}^{\star T}\Sigma_U\boldsymbol{\alpha}^\star$, $\mathrm{Var}(V\boldsymbol{\beta}^\star) = \boldsymbol{\beta}^{\star T}\Sigma_M\boldsymbol{\beta}^\star$, *and* $b''(\cdot)$ *are bounded.*

2. The minimum eigenvalues of the matrices $\mathbb{E}[Z_j^T Z_j]$ *is larger than a positive constant, uniformly over* j.

3. Letting

$$\boldsymbol{Z} = \mathbb{E}\left\{\mathbb{E}_L\left[\mathbf{M} \mid U\right]^T\left[V\beta^\star - U\beta^M_{\mathcal{C}}\right]\right\},$$

it holds that $\|\boldsymbol{Z}\|^2 = o\{\lambda_{max}(\Sigma_{M|U})\}$, *with* $\lambda_{max}(\Sigma_{M|U})$ *the largest eigenvalue of*

$$\Sigma_{M|U} = \mathbb{E}\left[\mathbf{M} - \mathbb{E}_L\left(\mathbf{M} \mid U\right)\right]\left[\mathbf{M} - \mathbb{E}_L(\mathbf{M} \mid U)\right]^T.G$$

Condition 3.1 is a regularity condition on the conditional means of M given U and Y given V. For Gaussian LSEMS, $b(\theta) = \theta^2/2$, so all of Condition 3.1 applies in our scenario. Condition 3.2 is a mild stability condition for our marginal regression estimators - in particular it guarantees that $(Z_j^T Z_j)$ is invertible and that the variance of our marginal maximum likelihood estimators is not too high. For Condition 3.3, in our Gaussian LSEM setting $\mathbb{E}_L(M|U) = U\alpha^\star$, so $\boldsymbol{Z} = \mathbb{E}\left[\boldsymbol{\alpha}^{\star T}(U^T V\beta^\star - U^T U\beta^M_{\mathcal{C}})\right] = 0$ by the score equations, so this condition is also satisfied.

Theorem 3.3.5 *Under Conditions 2 and 3, for $\lambda = c_6 n^{-2\kappa}$, there exists a $c_4 > 0$ such that*

$$\mathbb{P}\left(\left|\hat{\mathcal{M}}_\lambda\right| \leq O\left(n^{2\kappa}\lambda_{\max}\left(\boldsymbol{\Sigma}_{M|U}\right)\right)\right) \geq 1 - p\left(\exp\left(-c_4 n^{1-2\kappa}\left(k_n K_n\right)^{-2}\right) + nr_2\exp\left(-r_0 K_n^\alpha\right)\right).$$

Proof Claim 1: $\|\boldsymbol{\beta}_M\|^2 = \mathcal{O}(\lambda_{\max}(\Sigma_{M|U}))$. This claim implies that the size of the set $\{j = q+1, ..., p : |\beta^M_{M,j}| > \epsilon n^{-\kappa}\}$ is bounded by $\mathcal{O}(n^{2\kappa}\lambda_{\max}(\Sigma_{M|U}))$ for every $\epsilon > 0$. We will first accept this claim then prove it later.

Let $\epsilon > 0$, and consider the events $\mathcal{B}_n = \left\{\max_{q_1 \leq j \leq p} |\hat{\beta}^M_{M,j} - \beta^M_{M,j}| \leq \epsilon n^{-\kappa}\right\}$. On the event $\mathcal{B}_n$, the set $\{j = q+1, ..., p : |\hat{\beta}^M_{M,j}| > 2\epsilon n^{-\kappa}\}$ is a subset of $\{j = q+1, ..., p : |\beta^M_{M,j}| > \epsilon n^{-\kappa}\}$, which has size bounded by $\mathcal{O}(n^{2\kappa}\lambda_{\max}(\Sigma_{M|U}))$ Taking $\epsilon = c_5/2$ yields

$$\mathbb{P}(|\hat{\mathcal{M}}_\lambda| \leq \mathcal{O}(n^{2\kappa}\lambda_{\max}(\Sigma_{M|U}))) \geq \mathbb{P}(\mathcal{B}_n).$$

By Theorem (3.3.4),

$$\mathbb{P}(\mathcal{B}_n) \geq 1 - d\left(\exp\{-c_4 n^{1-2\kappa}(k_n K_n)^{-2}\} + nr_2\exp\{-r_0 K_n^a\}\right).$$

Proof of Claim 1: By Condition 3.2, $\Omega_{jj} - \Omega_{\mathcal{C}j}^T \Omega_{\mathcal{C}\mathcal{C}}^{-1} \Omega_{\mathcal{C}j}$ is uniformly bounded below. Since

$$\mathrm{Cov}_L(Y, M_j | M_{\mathcal{C}}) = (\Omega_{jj} - \Omega_{\mathcal{C}j}^T \Omega_{\mathcal{C}\mathcal{C}}^{-1} \Omega_{\mathcal{C}j}) \beta_{M,j}^M,$$

then

$$|\beta_{M,j}^M| \leq D_1 |\,\mathrm{Cov}_L(Y, M_j | U)$$

for some constant $D_1 > 0$ Using equations (3.29), (3.37), and Lipschitz continuity of b', we then obtain

$$\begin{aligned} |\,\mathrm{Cov}_L(Y, M_j|U)| &= \mathbb{E}\,|M_j\{b'(V\boldsymbol{\beta}^\star) - b'(U\boldsymbol{\beta}_{\mathcal{C}}^M)\}| \\ &\leq D_2\,\mathbb{E}\,|M_j(V\boldsymbol{\beta}^\star - U\boldsymbol{\beta}_{\mathcal{C}}^M)| \\ &= D_2\,\mathbb{E}\,|M_j(U\boldsymbol{\beta}_{\mathcal{C}}^\Delta + M\boldsymbol{\beta}_M^\star)|, \end{aligned}$$

where $\boldsymbol{\beta}_{\mathcal{C}}^\Delta = \boldsymbol{\beta}_{\mathcal{C}}^\star - \boldsymbol{\beta}_{\mathcal{C}}^M$. Bounding the right hand side of this inequality requires us to bound

$$\|\,\mathbb{E}\,M_j^T \boldsymbol{M}\boldsymbol{\beta}^\star + M_j^T U \boldsymbol{\beta}_{\mathcal{C}}^\Delta\|.$$

Combining the property of least-squares $\mathbb{E}\left[\mathbb{E}_L(\boldsymbol{M}|U)^T U\right] = \mathbb{E}\left[M^T U\right]$ with the definition of $\Sigma_{M|U}$ in Condition 3.3, we obtain

$$\left\|\Sigma_{M|U}\boldsymbol{\beta}_M^\star + \mathbb{E}\,\mathbb{E}_L(\boldsymbol{M}|U)\left[U\boldsymbol{\beta}_{\mathcal{C}}^\Delta + \mathbb{E}_L(\boldsymbol{M}^T|U)\boldsymbol{\beta}_M^\star\right]\right\|^2 = \|\Sigma_{M|U}\boldsymbol{\beta}_M^\star + \boldsymbol{Z}\|^2,$$

where $\boldsymbol{Z} = \mathbb{E}\,\mathbb{E}_L(\boldsymbol{M}|U)(\boldsymbol{M}\boldsymbol{\beta}_M^\star - U\boldsymbol{\beta}_{\mathcal{C}}^M)$ as in Condition 3. By the Law of Total Variance, this term is bounded:

$$\|\Sigma_{M|U}\boldsymbol{\beta}_M^\star + \boldsymbol{Z}\|^2 = \boldsymbol{\beta}^{\star T}(\Sigma_{M|U})^2\boldsymbol{\beta}_M^\star + 2\boldsymbol{Z}^T\Sigma_{M|U} + \boldsymbol{Z}^T\boldsymbol{Z}$$

$$\leq \lambda_{\max}(\Sigma_{M|U})(\boldsymbol{\beta}^{\star T}\Sigma_{M|U}\boldsymbol{\beta}_M^\star) + 2\boldsymbol{Z}^T\Sigma_{M|U} + \boldsymbol{Z}^T\boldsymbol{Z}$$

$$\leq \lambda_{\max}(\Sigma_{M|U})\,\mathrm{Var}(\boldsymbol{M}^T\boldsymbol{\beta}^\star) + 2\boldsymbol{Z}^T\Sigma_{M|U} + \boldsymbol{Z}^T\boldsymbol{Z}.$$

By Condition 3, the last two terms are $o(\lambda_{\max}(\Sigma_{M|U}))$, so

$$\|\boldsymbol{\beta}_M\|^2 = \mathcal{O}(\lambda_{\max}(\Sigma_{M|U})),$$

which proves Claim 1.

Claim 2: $\|\boldsymbol{\alpha}_A \circ \boldsymbol{\beta}_M^M\|^2 = \mathcal{O}(\mathrm{Var}(X)^{-1}\lambda_{\max}(\Sigma_{M|U}))$, where $\boldsymbol{\alpha}_A \circ \boldsymbol{\beta}_M^M$ here denotes the Hadamard (entrywise) product between $\boldsymbol{\alpha}_A$ and $\boldsymbol{\beta}_M^M$. (to be proven)

Accepting Claim 2 for now, we have $|\{j = q{+}1, ..., p : |\alpha_{A,j}\beta_{M,j}^M| > \epsilon n^{-\kappa}\}| \leq \mathcal{O}(n^{2\kappa}\lambda_{\max}(\Sigma_M)\lambda_{\max}(\Sigma_{M|U}))$ $\forall\epsilon > 0$. Let $\epsilon > 0$, and consider events $\mathcal{B}_n = \{\max_{q+1\leq j\leq p} |\hat{\alpha}_{A,j}\hat{\beta}_{M,j}^M - \alpha_{A,j}\beta_{M,j}^M| \leq \epsilon n^{-\kappa}\}$. On event $\mathcal{B}_n$, we have

$$\{j = q+1, ..., p : |\hat{\alpha}_{A,j}\hat{\beta}_{M,j}^M| > 2\epsilon n^{-\kappa}\} \subset \{j = q+1, ..., p : |\hat{\alpha}_{A,j}\hat{\beta}_{M,j}^M| > \epsilon n^{-\kappa}\},$$

with the right hand set having size bounded by $\mathcal{O}(n^{2\kappa}\lambda_{\max}(\Sigma_M)\lambda_{\max}(\Sigma_{M|U}))$. Letting $\epsilon = c/2$ yields

$$\mathbb{P}\left(|\hat{\mathcal{M}}_\lambda| \leq \mathcal{O}(n^{2\kappa}\lambda_{\max}(\Sigma_M)\lambda_{\max}(\Sigma_{M|U}))\right) \geq \mathbb{P}(\mathcal{B}_n).$$

By Theorem (3.3.4) $\mathbb{P}(\mathcal{B}_n) \geq 1 - \mathcal{O}\left(n^{(1-\kappa)/2} \exp\left\{-cn^{1-\kappa}\right\}\right)$, as needed.

Proof of Claim 2: Since $\|\boldsymbol{\alpha}_A \circ \boldsymbol{\beta}_M^M\| \leq \mathrm{Tr}(\boldsymbol{\alpha}_A(\boldsymbol{\beta}_M^M)^T) \leq \|\boldsymbol{\alpha}_A\|\|\boldsymbol{\beta}_M^M\|$, and $\|\boldsymbol{\beta}_M^M\|$ is bounded by Claim 1, it remains to bound $\|\boldsymbol{\alpha}_A\|$.

Indeed, by Condition (3.1) we have that $\mathrm{Var}(U\boldsymbol{\alpha}^\star)$ is bounded, i.e. $\mathrm{Var}(U\boldsymbol{\alpha}^\star) \leq \mathcal{O}(1)$, which in particular implies that $\mathrm{Var}(A\boldsymbol{\alpha}_A^\star) \leq \mathcal{O}(1)$. Expanding this yields

$$\mathrm{Var}(A\boldsymbol{\alpha}_A^\star) = \boldsymbol{\alpha}_A^{\star T} \mathrm{Var}(A)\boldsymbol{\alpha}_A^\star = \mathrm{Var}(A)\boldsymbol{\alpha}_A^{\star T}\boldsymbol{\alpha}_A^\star =\leq \mathcal{O}(1)$$

which implies

$$\|\boldsymbol{\alpha}_A^\star\|_2^2 \leq \mathcal{O}(\mathrm{Var}(A)^{-1}).$$

■

3.4 Selection of Thresholding Parameter

Clearly, the exact choice of λ affects which mediators are screened out from $\hat{\mathcal{M}}_\lambda$. In this section, we derive some choices for this threshold. The following will be used:

$\hat{\boldsymbol{\alpha}}_j = (U^TU)^{-1}U^TM_j$, $\mathbb{E}[\hat{\alpha}_{A,j}] = \alpha_{A,j}$, $\hat{\sigma}_\alpha^2 := \mathrm{Var}(\hat{\alpha}_{A,j}) = \sigma_M^2(U^TU)^{-1}_{q+2,q+2}$,

$\hat{\beta}_{M,j}^M = (Z_j^TZ_j)^{-1}Z_j^TY$, $\mathbb{E}[\hat{\beta}_{M,j}^M] = \beta_{M,j}^M$,

$\hat{\sigma}_\beta^2 := \mathrm{Var}(\hat{\beta}_{M,j}^M) = \mathbb{E}[\mathrm{Var}(\hat{\beta}_{M,j}^M|X,U,M_j)] + \mathrm{Var}(\mathbb{E}[\hat{\beta}_{M,j}^M|X,U,M_j)] = \sigma^2\,\mathbb{E}[(Z_j^TZ_j)^{-1}]_{q+3,q+3}$.

Consider the potential mediators to be in four groups:

$\mathcal{M}_1 = \{j \in \mathcal{M} : \alpha_{A,j}^\star \neq 0, \beta_{M,j}^\star \neq 0\}$,

$\mathcal{M}_2 = \{j \in \mathcal{M} : \alpha_{A,j}^\star \neq 0, \beta_{M,j}^\star = 0\}$,

$\mathcal{M}_3 = \{j \in \mathcal{M} : \alpha_{A,j}^\star = 0, \beta_{M,j}^\star \neq 0\}$, and

$\mathcal{M}_4 = \{j \in \mathcal{M} : \alpha^{\star}_{A,j} = 0, \beta^{\star}_{M,j} = 0\}$.

Define $Z_{\alpha,j} = \dfrac{\hat{\alpha}_{A,j}}{\hat{\sigma}_{\alpha}}$ and $Z_{\beta,j} = \dfrac{\hat{\beta}_j}{\hat{\sigma}_{\beta,j}}$, and $T_{Sobel,j} = \dfrac{\hat{\alpha}_{A,j}\hat{\beta}^M_{M,j}}{\sqrt{(\hat{\beta}^M_{M,j})^2\hat{\sigma}^2_{\alpha} + \hat{\alpha}^2_{A,j}\hat{\sigma}^2_{\beta}}} = \dfrac{Z_{\alpha,j}Z_{\beta,j}}{\sqrt{Z^2_{\alpha} + Z^2_{\beta}}}$.

Liu et al. [28] showed that the distribution of $Z_{\alpha,j}Z_{\beta,j}$ depends on which group the potential mediator comes from. In particular, for non-mediators in $\mathcal{M}_2 \cup \mathcal{M}_3$, $Z_{\alpha,j}Z_{\beta,j} \sim \mathcal{N}(0,1)$, while for non-mediators in $\mathcal{M}_4$, $Z_{\alpha,j}Z_{\beta,j} \sim \mathcal{N}(0, \frac{1}{4})$. This results in a mixture distribution for the asymptotic null distribution of $T_{Sobel,j}$. Denoting this asymptotic null cumulative distribution distribution $F(x) = (\pi_2 + \pi_3)\Phi(x) + \pi_4\Phi(2x)$, where Φ denotes the standard normal cdf. We use a similar idea to [28] where these distributions in combination with the relative proportion of potential mediators in $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ select the most appropriate choice of λ.

We consider two types of thresholds when selecting the set $\hat{\mathcal{M}}$ according to the conventions of [60]. The first type, soft thresholding, uses the null distribution of the computed statistic for screening to compute a threshold. The second type, hard thresholding, prespecifies a size d of variables to screen out of $\mathcal{M}$. After the test statistics T_j's are computed, the mediators with the largest values of $|T_j|$ are recruited into $\hat{\mathcal{M}}$.

3.4.1 Soft Thresholding: FDR Control Approach

In the high-dimensional setting, Bonferonni's approach of controlling Type I error is too conservative and often misses important signals. In this setting, it often more useful to control the false discovery rate (FDR) of a screening approach. The following condition allows FDR control when using $T_{Sobel,j}$ to recruit variables into $\hat{\mathcal{M}}_{\lambda}$.

Condition 4 *(i) For any j, let $e_{ij} = Y_i - \boldsymbol{z}_{i,\mathcal{C}j}^T \beta_{\mathcal{C}j}$ for $i = 1, ..., n$. For a given j, $\mathrm{Var}(e_{ij}) \geq c$ for some positive $c > 0$ and $i = 1, ..., n$, and $\sup_{i \geq 1} \mathbb{E}\,|e_i|^{2+\chi} < \infty$ for some $\chi > 0$.*

(ii) For $j \in (\mathcal{M}^\star)^c$, $\mathrm{Cov}_L(A, M_j | U_{\mathcal{C}}) = 0$ or $\mathrm{Cov}_L(Y, M_j | U) = 0$.

Theorem 3.4.1 *Under Conditions 2, 3, and 4, if we choose*

$$\hat{\mathcal{M}}_\lambda = \{j : |T_{Sobel,j}| \geq \lambda\},$$

where $\lambda := F^{-1}(1 - f/(2p))$ and f is the maximum tolerated number of false positives, then for some constant $c > 0$,

$$\mathbb{E}\left(\frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|(\mathcal{M}^\star)^c|}\right) \leq \frac{f}{p} + cn^{-1/2}.$$

Proof Note that $\mathbb{E}\left(\dfrac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|(\mathcal{M}^\star)^c|}\right) = \dfrac{1}{p - |\mathcal{M}^\star|} \displaystyle\sum_{j \in \mathcal{M}_1^c} \mathbb{P}\{|T_{Sobel,j}| \geq \lambda\}$.

Under Conditions 2, 3, and 4, Theorem 3.3.1 says we have $\alpha_{A,j} = 0$ or $\beta_{M,j}^M = 0$ for $j \in \mathcal{M}_\star^c$. Since U includes an intercept term, $\mathbb{E}\, e_i = 0$. Since $T_{Sobel,j}$ asymptotically follows a mixture distribution with probability $(\pi_2 + \pi_3)$ of coming from $\mathcal{N}(0, 1)$ and probability π_4 of coming from $\mathcal{N}(0, \frac{1}{4})$, the right hand side of the previous inequality is

$$\frac{1}{p - |\mathcal{M}^\star|} \sum_{j \in \mathcal{M}_\star^c} \sum_{k=2}^{4} \mathbb{P}\left\{|T_{Sobel,j}| \geq \lambda \middle| j \in \mathcal{M}_k\right\} \pi_k$$

$$= \frac{1}{p - |\mathcal{M}^\star|} \sum_{j \in \mathcal{M}_\star^c} 2(1 - \Phi(\lambda))(\pi_2 + \pi_3) + 2(1 - \Phi(2\lambda)\pi_4$$

$$= 2(1 - \Phi(\lambda))(\pi_2 + \pi_3) + 2(1 - \Phi(2\lambda))\pi_4.$$

The asymptotic null cumulative distribution function of $T_{Sobel,j}$ is $F(x) = (\pi_2 + \pi_3)\Phi(x) + \pi_4\Phi(2x)$. Thus, there exists $c > 0$ such that

$$\sup_z |\mathbb{P}(T_{Sobel,j} \leq z) - F(z)| \leq cn^{-1/2}.$$

Putting these statements together yields

$$\mathbb{E}\left(\frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|(\mathcal{M}^\star)^c|}\right) \leq \frac{1}{p - |\mathcal{M}^\star|}\sum_{j \in \mathcal{M}_\star^c} 2(1 - F(\lambda)) + cn^{-1/2}$$

$$= 2(1 - F(\lambda)) + cn^{-1/2}.$$

Plugging in $\lambda = F^{-1}(1 - f/(2p))$ then yields

$$2(1 - F(\lambda)) + cn^{-1/2} = f/p + cn^{-1/2},$$

as needed. ■

Note that the left hand side of the previous inequality, $\mathbb{E}\left(\frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|(\mathcal{M}^\star)^c|}\right)$, is the false positive rate rather than the FDR typically defined as $\mathbb{E}\left(\frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|\hat{\mathcal{M}}_\lambda|}\right)$. [61] pointed out that since

$$\frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|\hat{\mathcal{M}}_\lambda|} = \frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|\mathcal{M}_\star^c|}\frac{|(\mathcal{M}^\star)^c|}{|\hat{\mathcal{M}}_\lambda|} < \frac{|\hat{\mathcal{M}}_\lambda \cap (\mathcal{M}^\star)^c|}{|\mathcal{M}_\star^c|}\frac{p}{|\hat{\mathcal{M}}_\lambda|},$$

controlling the false positive rate at level f/p is equivalent to controlling the FDR at level $f/|\hat{\mathcal{M}}_\lambda|$, conditional on $|\hat{\mathcal{M}}_\lambda|$.

In practice, the true sizes of the sets $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ are unknown, so the proportions π_2, π_3, and π_4 are also unknown. Jin and Cai [62] developed the JC-Method, which uses the empirical characteristic function and Fourier analysis, for estimating the proportion of non-mediators in each group, say with $\hat{\pi}_2$, $\hat{\pi}_3$, and $\hat{\pi}_4$. These estimates can then be used to construct weights $\hat{w}_2 = \hat{\pi}_2/(\hat{\pi}_2 + \hat{\pi}_3 + \hat{\pi}_4)$, $\hat{w}_3 = \hat{\pi}_3/(\hat{\pi}_2 + \hat{\pi}_3 + \hat{\pi}_4)$, and $\hat{w}_4 = \hat{\pi}_4/(\hat{\pi}_2 + \hat{\pi}_3 + \hat{\pi}_4)$, as in [28]. Finally, the asymptotic null distribution of T_{Sobel} can be estimated with $\hat{F}(x) = (\hat{w}_2 + \hat{w}_3)\Phi(x) = \hat{w}_4\Phi(2x)$ to create the cutoff value $\lambda = \hat{F}^{-1}(1 - f/(2p))$.

3.4.2 Hard Threshold

Fan and Lv [8] recommend to fix a number, say d, of predictors to keep after screening. Commonly recommended values of d are $d = \lfloor N/\log(N) \rfloor$ or $d = N - 1$. This equates to sorting the statistics $|\hat{T}_{Sobel}|$ to obtain the order statistics $T_{(1)}, ..., T_{(p)}$ then setting $\lambda := T_{(p-d+1)}$. This thresholding rule is the standard choice in the literature [63] and works well with the MSS method as well as shown in our simulation studies. This approach to thresholding is called hard thresholding by [60] as opposed to soft thresholding which uses FDR control to generate a probabilistic threshold. A researcher can choose a large value of d (such as $d = n - 1$) to conservatively ensure that all true mediators are included in the selected set with high probability, but too large of d will also cause the screening method to select too many false mediators. Too small of d, on the other hand, causes the sure screening property to fail and it becomes impossible for the screening method to select all of the true mediators. While the standard $d = \lfloor N/\log(N) \rfloor$ is common in practice, [63] also developed the REflection via Data Splitting (REDS) procedure for selecting d in a general screening

setting. In practice, hard thresholding performs very well and is very quick to implement as it allows the user to skip estimating the composite null proportions while still maintaining the sure screening property. The standard $d = \lfloor N/\log(N) \rfloor$ or any $d < n$ also guarantees that the screened set of variables is smaller than the sample size, which is not always the case with the soft thresholded methods.

3.5 Comparison with Other Methods

3.5.1 Comparison with SIS

To fully understand the behavior of MSS, we investigate the theory of how it compares to SIS in terms of power and FDR. Specifically we would like to show that for $j \in \mathcal{M}_1$,

$$\mathbb{P}(j \in \hat{\mathcal{M}}_{MSS} | j \in \mathcal{M}_1) \geq \mathbb{P}(j \in \hat{\mathcal{M}}_{SIS} | j \in \mathcal{M}_1),$$

and that for $j \in \mathcal{M} \setminus \mathcal{M}_1$,

$$\mathbb{P}(j \in \hat{\mathcal{M}}_{MSS} | j \in \mathcal{M} \setminus \mathcal{M}_1) \leq \mathbb{P}(j \in \hat{\mathcal{M}}_{SIS} | j \in \mathcal{M} \setminus \mathcal{M}_1).$$

SIS targets the partial correlations between M_j and Y, say $\rho_j = \frac{\text{Cov}(M_j, Y)}{\sigma_{Mj}\sigma_Y}$. Assuming without loss of generality that the columns of $\boldsymbol{M}$ have been scaled such that $\text{Var}(M_j) = 1$ for all j,

this correlation learning is equivalent to choosing the mediators M_j with the largest absolute covariance with Y. According to our model,

$$\begin{aligned}
\operatorname{Cov}(M_j, Y|X) &= \operatorname{Cov}(M_j, \beta_0 + X^T\boldsymbol{\beta}_X + A\beta_A + \sum_{k=1}^{p} \beta_{M,k} M_k + \epsilon|X) \\
&= \beta_A \operatorname{Cov}(M_j, A|X) + \sum_{k=1}^{p} \beta_{M,k} \operatorname{Cov}(M_j, M_k|X) \\
&= \operatorname{Var}(A)\alpha_{A,j} \left(\beta_A + \sum_{k=1}^{p} \alpha_{A,k}\beta_{M,k} \right) + \sum_{k=1}^{p} \beta_{M,k} \operatorname{Cov}(e_j, e_k) \\
&= \alpha_{A,j} \operatorname{Var}(A) TE + \sum_{k=1}^{p} \beta_{M,k}\sigma_{Mjk}, \qquad (3.40)
\end{aligned}$$

where $TE = \beta_A + \sum_{k=1}^{p} \alpha_{A,k}\beta_{M,k}$ denotes the total effect of A on Y. By application of the triangle inequality, we also have

$$|\operatorname{Cov}(M_j, Y)| \leq |\alpha_{A,j}| \operatorname{Var}(X)|TE| + \left| \sum_{k=1}^{p} \beta_{M,k}\sigma_{Mjk} \right|. \qquad (3.41)$$

This decomposition suggests several components of the mediation model that may affect performance of SIS:

- For $j \in \mathcal{M}_1$ and $\mathcal{M}_3$, i.e. when $\alpha_{A,j} \neq 0$, $|\operatorname{Cov}(M_j, Y)|$ increases as $|TE|$ increases, and decreases as $|TE|$ decreases.
- In particular, large $|TE|$ may cause SIS to recruit too many variables from $\mathcal{M}_3$, while small $|TE|$ may cause SIS to recruit too few variables from $\mathcal{M}_1$.
- Note that $|TE|$ is maximized when all $\alpha_{A,j}\beta_{M,j}$ have the same sign for all $j \in \mathcal{M}_1$.

- Correlation of M_j with mediators in $\mathcal{M}_1$ and $\mathcal{M}_2$, i.e. mediators for which $\beta_{M,k} \neq 0$, can further increase false positives and false negatives.

- In particular, if $|TE|$ is small, SIS may recruit mediators from $\mathcal{M}_2$ with large β values.

3.5.2 Comparison with HDMT

[1] derived a similar statistical procedure to ours for the context of multiple hypothesis testing. We include their Proposition 1 here for completeness, as it outlines a condition for Z_α and Z_β under which Sobel-Comp has greater power than HDMT. Details of proof can be found [1].

Proposition 1 (Proposition 1 from [1]) *Suppose* $|Z_{\beta,j}| > |Z_{\alpha,j}| \geq 0$. *The case-specific p-value under* $H_{00,j}$ *from Sobel-comp is smaller than that from HDMT if*

$$|Z_{\beta,j}| > \max\left(|Z_{\alpha,j}|, \left\{4\left(\Phi^{-1}\left(2\Phi\left(|Z_{\alpha,j}|\right)^2\right)\right)^{-2} - Z_{\alpha,j}^{-2}\right\}^{-1/2}\right),$$

where $\Phi(\cdot)$ *is the cumulative distribution function of a standard normal random variable.*

The authors further explain that the above result requires π_4 to be near one to hold. Further, this result is also true when $Z_{\beta,j}$ and $Z_{\alpha,j}$ are interchanged.

3.6 Numerical Studies

To evaluate the performance of model selection based on our marginal Sobel screener and compare its performance with other model selection approaches, we present a variety of simulations and a real data example. The other model selection procedures considered

are SIS, CSIS, MaxP, and DACT. MaxP screens using $p_{MaxP,j} = \max(p_{\alpha,j}, p_{\beta,j})$, where $p_{\alpha,j} = 2*(1-\Phi(|\hat{\alpha}_{A,j}|/\hat{\sigma}_{\alpha,j}))$ and $p_{\beta,j} = 2*(1-\Phi(|\hat{\beta}^{M}_{M,j}|/\hat{\sigma}_{\beta,j}))$ and is found to be conservative in practice. DACT [28] corrects this formulation by accounting for the composite nature of the null hypothesis and screening using $p_{DACT,j} = \dfrac{\pi_2 p_{\alpha,j} + \pi_3 p_{\beta,j} + \pi_4 p^2_{MaxP,j}}{\pi_2 + \pi_3 + \pi_4}$. We consider both soft-thresholded and hard-thresholded approaches to screening to reflect both the case when the researcher may have no knowledge of the true sparsity of the mediators (soft thresholding) and the case when the researcher has some prior knowledge of how many mediators should be kept (hard thresholding). Note that the hard and soft threshold of variable screening [60] differs from hard and soft thresholding in penalized regression and variable selection literature.

3.6.1 Simulation Studies

For our first set of simulation study, we consider the LSEM model:

$$M = X\boldsymbol{\alpha} + e, e \overset{iid}{\sim} \mathcal{N}_p(0, I_p),$$

$$Y = [X, \boldsymbol{M}]\boldsymbol{\beta} + \epsilon, \epsilon \overset{iid}{\sim} \mathcal{N}(0, 1).$$

For $j \in \mathcal{M}_1$, $\alpha_{A,j} \neq 0$ and $\beta_{M,j} \neq 0$. For $j \in \mathcal{M}_2$, $\alpha_{A,j} = 0$ while $\beta_{M,j} \neq 0$. For $j \in \mathcal{M}_3$, $\alpha_{A,j} \neq 0$ while $\beta_{M,j} = 0$. Finally, for $j \in \mathcal{M}_4$, both $\alpha_{A,j} = 0$ and $\beta_{M,j} = 0$. This structure is chosen so that non-mediators, i.e. M_j, $j \notin \mathcal{M}_1$, come from the composite null distribution described throughout the paper. For this study the following procedure is used:

1. Generate sets $(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4)$ and parameters $\boldsymbol{\alpha}^\star$, $\boldsymbol{\beta}^\star$ according to settings.

2. For $i = 1, 2, ..., n_{sim}$:

   (a) Generate data $(\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{M}, \boldsymbol{Y})$: For our numerical studies $X = \emptyset$ and $A \sim BIN(p = .5)$ are first generated. Then, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are generated from a normal or uniform distribution with difference variace values for each case. Once all parameters and preconditioned variables are generated, $(\boldsymbol{M}, \boldsymbol{Y})$ are then generated according to its LSEM model.

   (b) Apply MSS to obtain $\hat{\mathcal{M}}_{MSS}$. If a soft thresholding rule is used, save the random size of the selected set $|\hat{\mathcal{M}}_{MSS}| = \hat{m}_{MSS}$. For a hard thresholding rule, $\hat{m}_{MSS}$ is preset to $\log(N)/N$.

   (c) Apply SIS using $\hat{m}_{MSS}$ as the prespecified number of variables to keep, yielding $\hat{\mathcal{M}}_{SIS}$ also of size $\hat{m}_{MSS}$.

   (d) Apply MaxP screener and DACT using the same soft thresholding or hard thresholding rule as in (b) and (c) to obtain $\hat{\mathcal{M}}_{MaxP}$ and $\hat{\mathcal{M}}_{DACT}$.

   (e) For each procedure $PROC \in \{SIS, CSIS, MSS, MaxP, DACT\}$:

      i. Compute sample TPR: $\text{TPR}_{PROC} = \dfrac{|\hat{\mathcal{M}}_{PROC} \cap \mathcal{M}_1|}{m_1}$.

      ii. Compute sample FDR: $\text{FDR}_{PROC} = \dfrac{|\hat{\mathcal{M}}_{PROC} \cap \mathcal{M}_1^c|}{\hat{m}_{PROC}}$.

3. Estimate Power and FDR by averaging over simulated results.

For each setting, we generate $n_{sim} = 200$ replications to estimate TPR and FDR of each method. Further details are given in the captions of each figure. Inspired by the decomposition of the marginal covariance shown in (3.41), the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are drawn from distributions symmetric around zero as is common in both real life situations and in

non-informative priors in Bayesian statistics. In this common scenario, $|TE(A, M, Y)| = |\beta_X + \sum_{j=1}^{p} \alpha_j \beta_j|$ was small, resulting in small marginal correlation between the mediators and response even for mediators M_j for whom $\alpha_j \beta_j \neq 0$. As shown in our results, this drives up the FDR of the SIS and related methods while MSS outperforms them in terms of both higher TPR (i.e. better power) and lower FDR.

Comparision 1a: Fixed p, varying N, hard thresholds

For our first set of simulations, we consider the case where the number of candidate mediators p is fixed. Each selection procedure computes some test statistic T_j for each mediator, then selects $d = \lfloor \log N/N \rfloor$ mediators with the most significant values of T_j. For each case $p = 100,000$ while $N \in \{200, 500, 1000\}$ varies. Of the p mediators, let $\pi = (.01, .02, .02, .95)$ denote the proportions of mediators that are in $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$, respectively. We also consider different distributions for the generation of the nonzero parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, indexed by a variance parameter $\sigma^2 \in \{.1, .3, .7, 1\}$. One point of note is that under this hard thresholding rule and these parameter settings, MaxP and DACT yielded the same selected sets, causing their lines to be indistinguishable. This is because when π_4 is close to 1, $p_{DACT,j} \propto \pi_2 p_{\alpha,j} + \pi_3 p_{\beta,j} + \pi_4 p_{MaxP}^2 \approx p_{MaxP,j}^2$, which keeps the same ordering as $p_{MaxP,j}$.

Comparision 1b: Fixed p, varying N, soft thresholds

Our second set of simulations mirrors the first, but applies a soft thresholding rule for each selection procedure. For each case $p = 100,000$ while $N \in \{200, 500, 1000\}$ varies. Of the p mediators, let $\pi = (.01, .02, .02, .95)$ denote the proportions of mediators that are in $\mathcal{M}_1$,

$\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$, respectively. We also consider different distributions for the generation of the nonzero parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, indexed by a variance parameter $\sigma^2 \in \{.1, .3, .7, 1\}$.

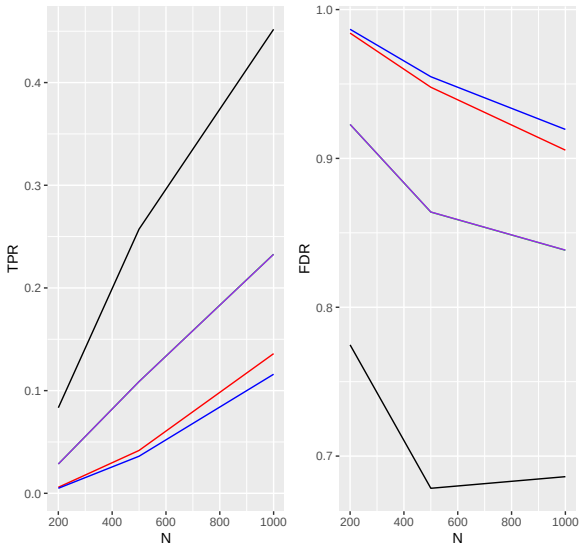

Figure 3.1.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .1)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

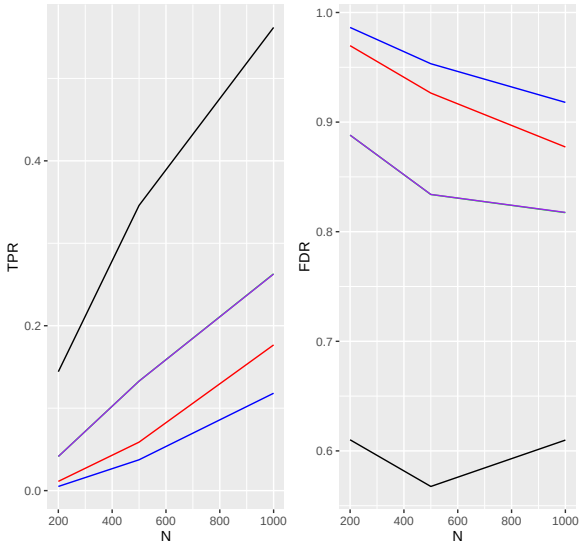

Figure 3.2.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .3)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

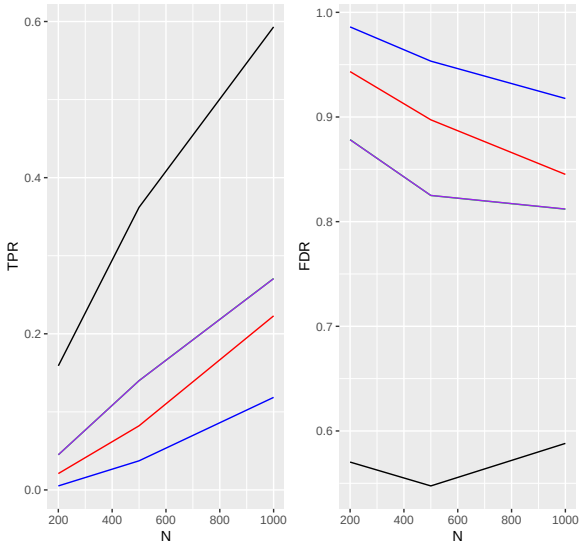

Figure 3.3.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .5)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

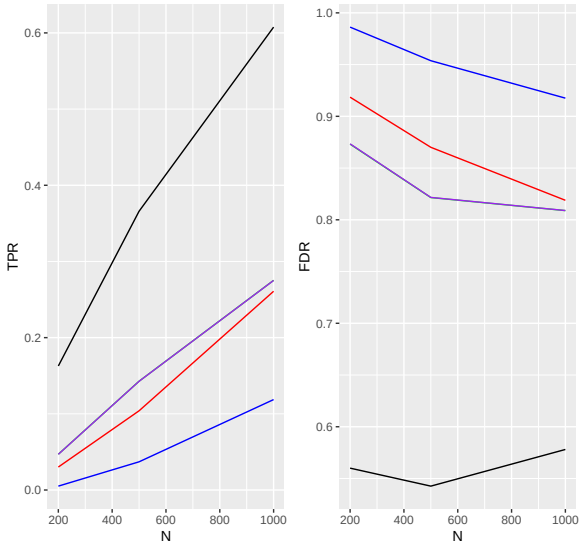

Figure 3.4.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .7)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

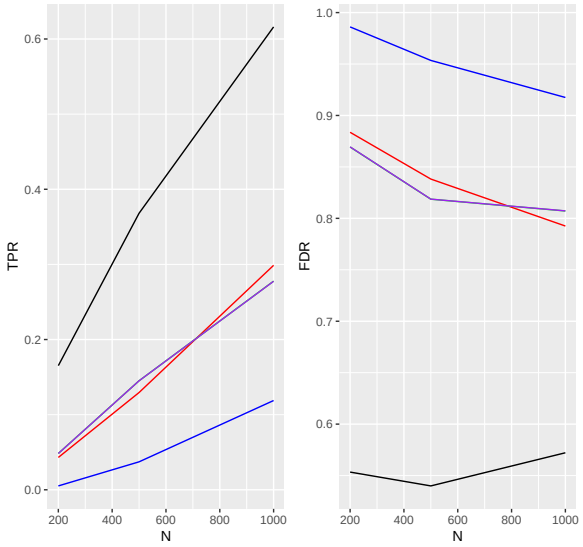

Figure 3.5.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0,1)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

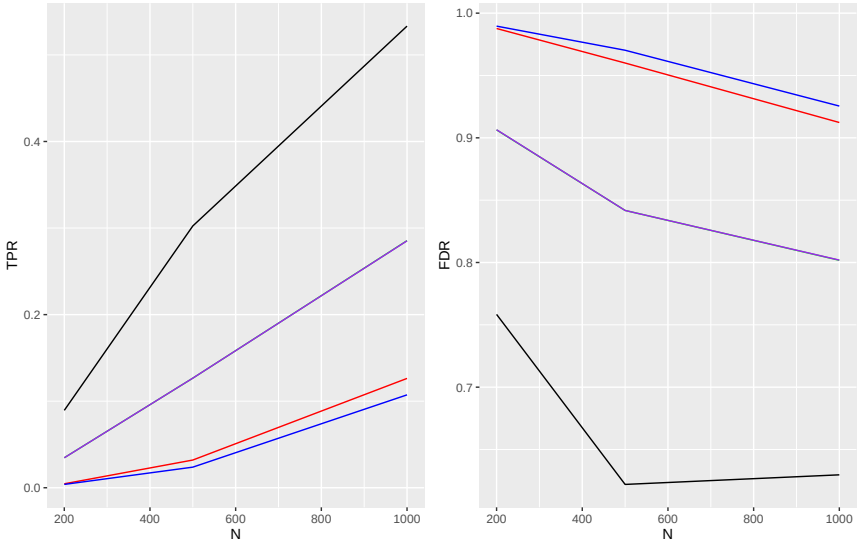

Figure 3.6.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .1$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .1$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

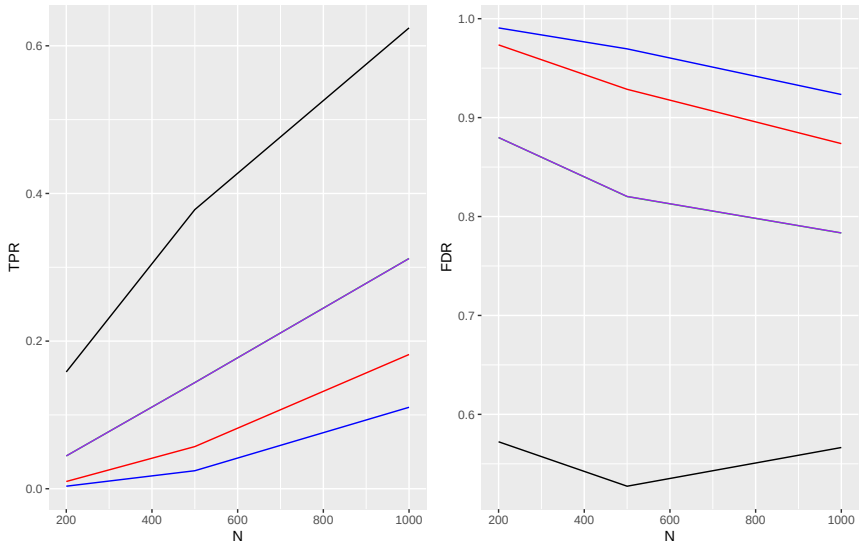

Figure 3.7.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .3$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .3$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

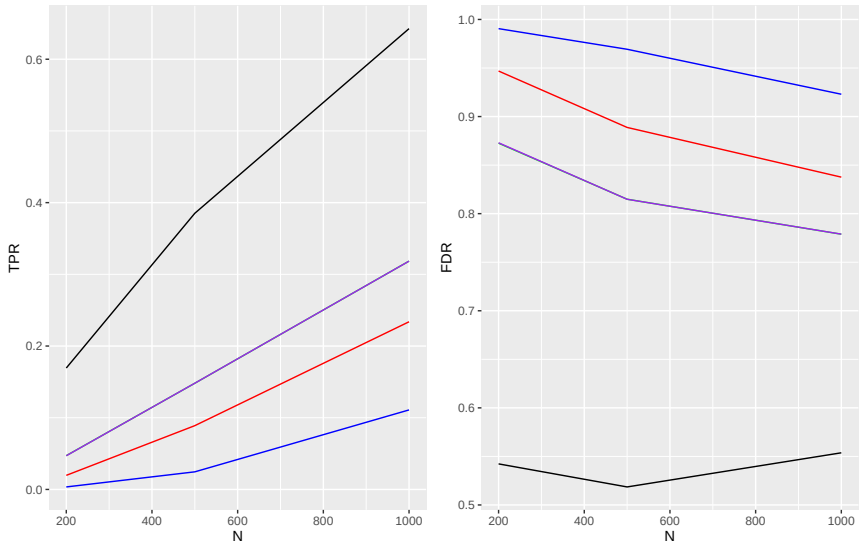

Figure 3.8.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .5$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .5$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

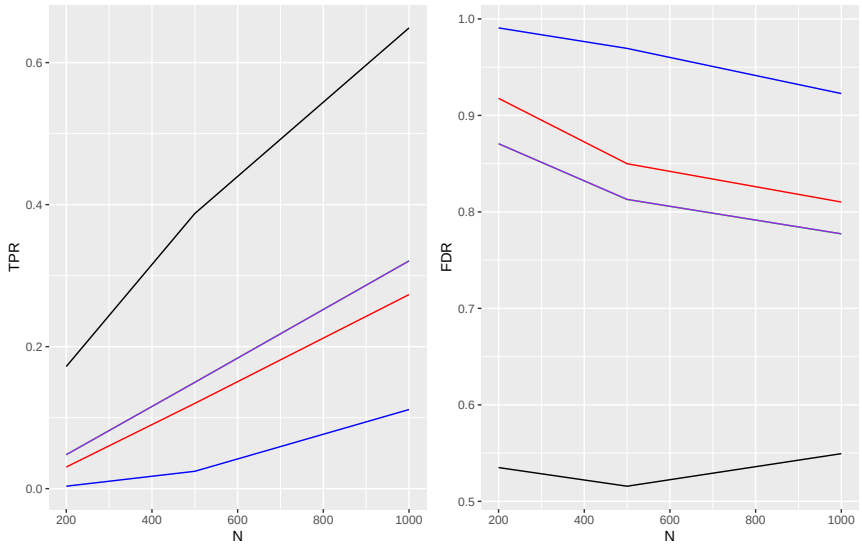

Figure 3.9.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .7$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .7$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

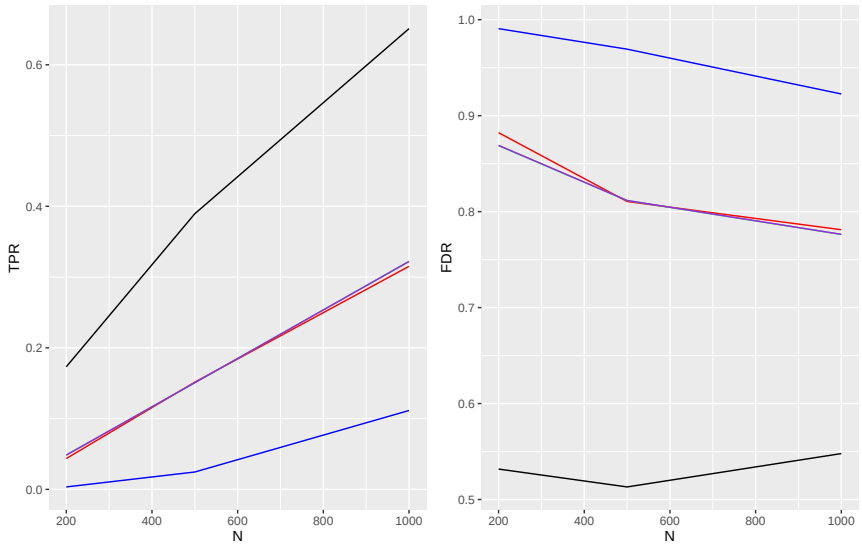

Figure 3.10.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = 1$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = 1$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

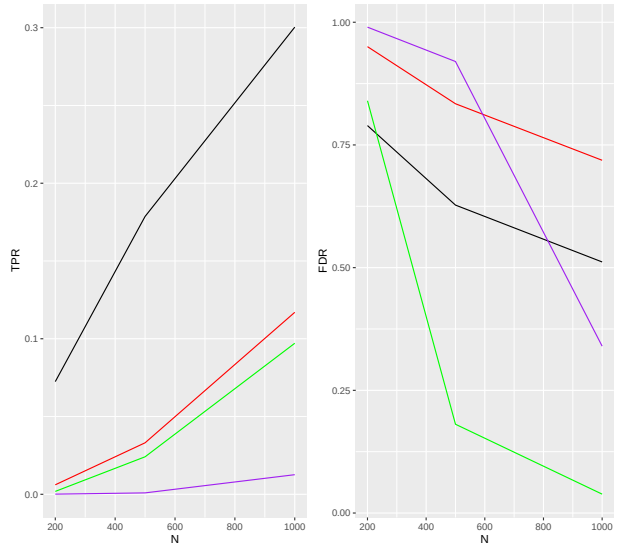

Figure 3.12.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .3)$. Here the line representing MSS is in black, SIS in red, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

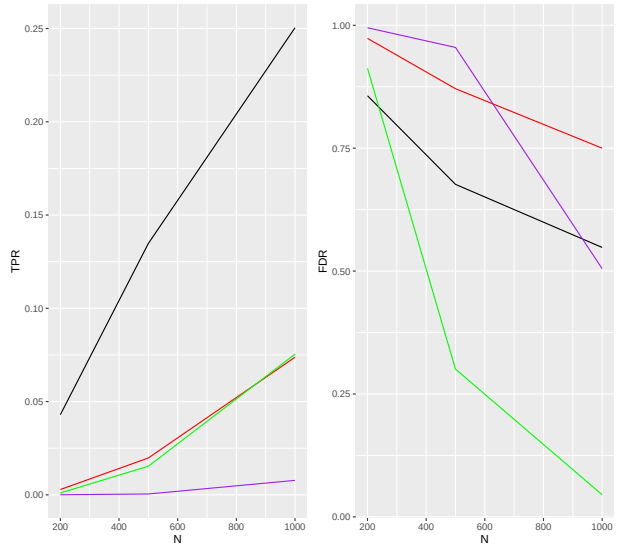

Figure 3.11.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .1)$. Here the line representing MSS is in black, SIS in red, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

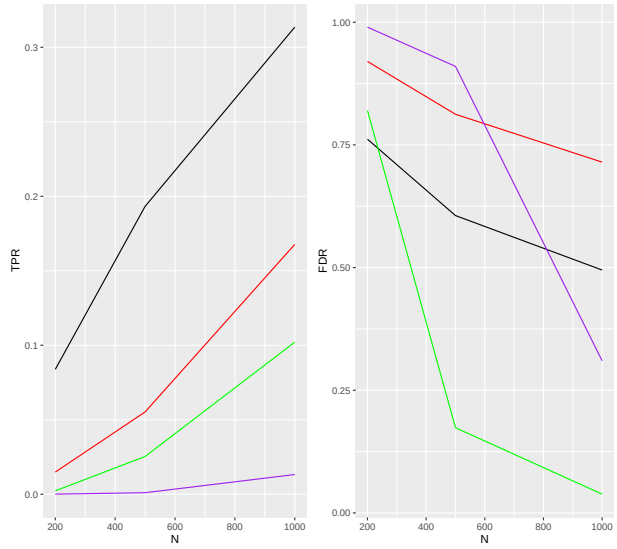

Figure 3.13.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .5)$. Here the line representing MSS is in black, SIS in red, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

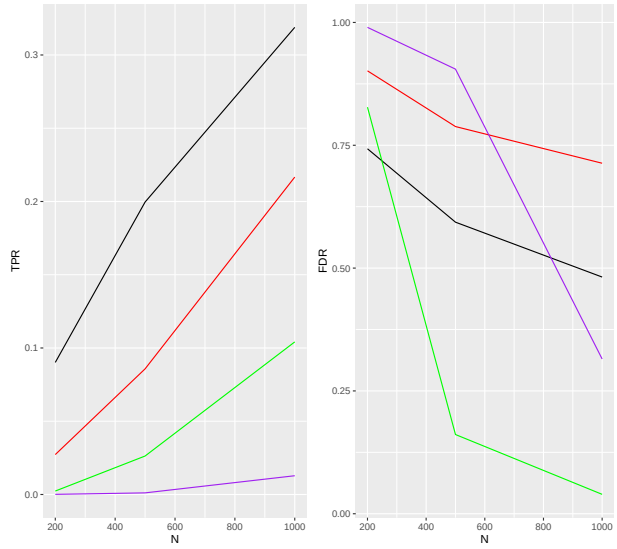

Figure 3.14.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .7)$. Here the line representing MSS is in black, SIS in red, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

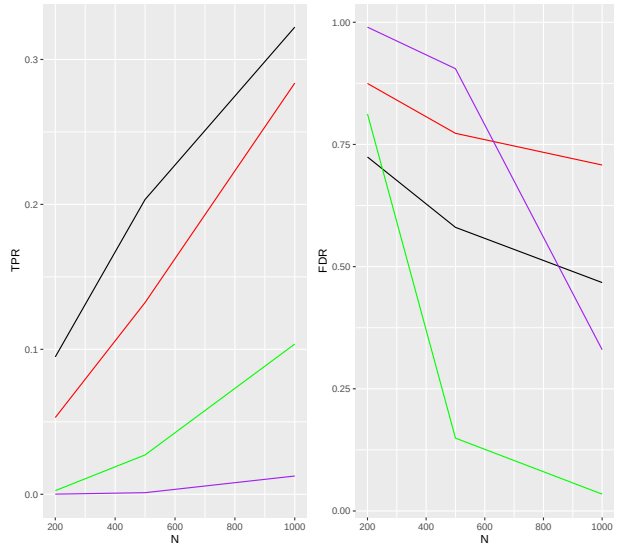

Figure 3.15.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero α and β parameters are generated from $\mathcal{N}(0,1)$. Here the line representing MSS is in black, SIS in red, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

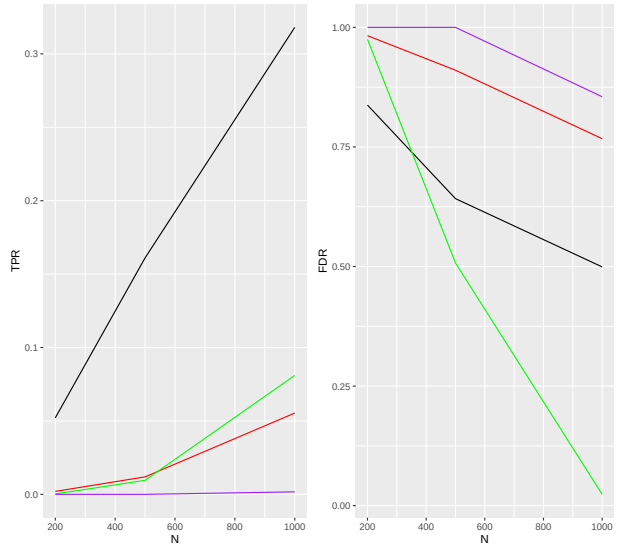

Figure 3.16.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .1$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .1$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

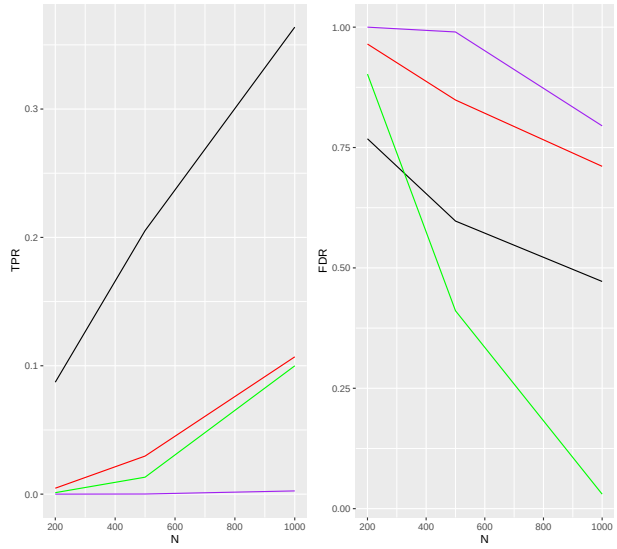

Figure 3.17.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .3$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .3$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

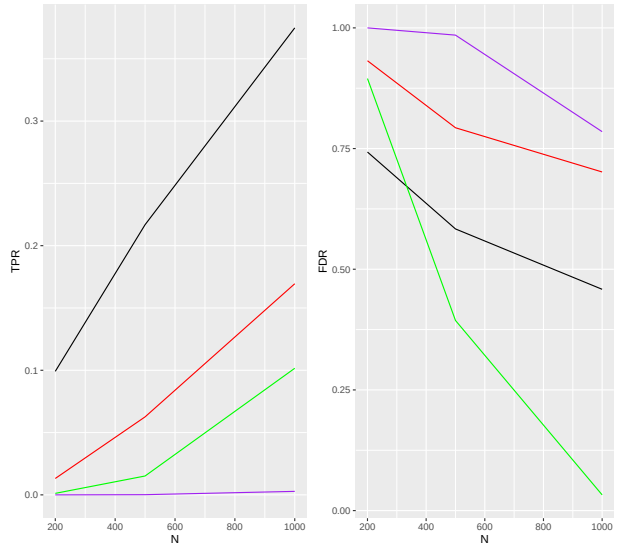

Figure 3.18.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .5$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .5$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

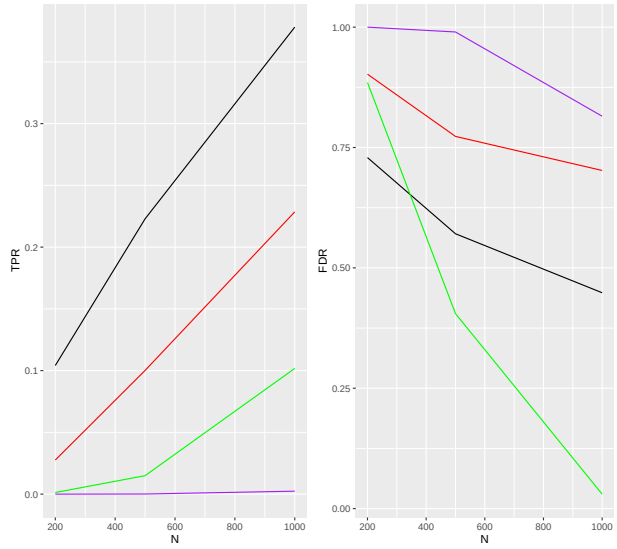

Figure 3.19.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .7$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .7$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Soft thresholding rule is applied with $f = 250$.

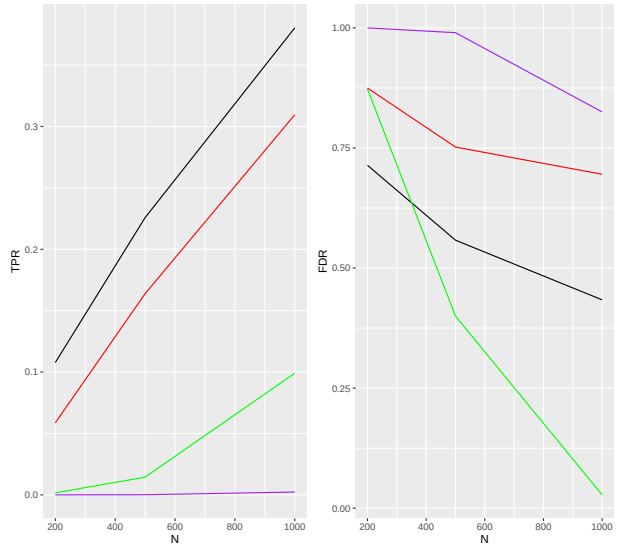

Figure 3.20.: Comparison of TPR and FDR when $p = 100,000$ is fixed and N varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = 1$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = 1$ for nonzero $\beta_{M,j}$. Soft thresholding rule is applied with $f = 250$.

Comparison 2: Fixed N, varying p, hard thresholds

For this set of simulations, we test how the hard-thresholded versions of MSS and SIS perform as N is fixed and the total number of mediators $p \in \{1000, 10000, 100000\}$ varies. The sizes of the mediator sets $(m1, m2, m3, m4)$ vary as fixed proportions of p determined by $(m1, m2, m3, m4) = p * (.01, .02, .02, .95)$. As with Comparison 1, under this hard thresholding rule and these parameter settings MaxP and DACT yielded the same selected sets, causing their lines to be indistinguishable.

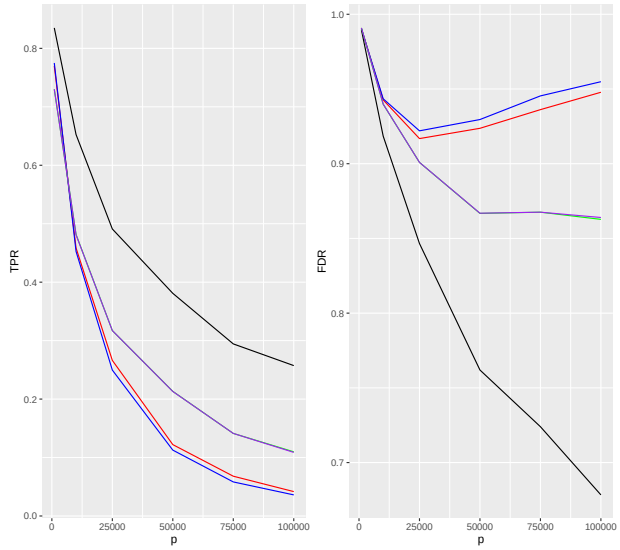

Figure 3.21.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .1)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

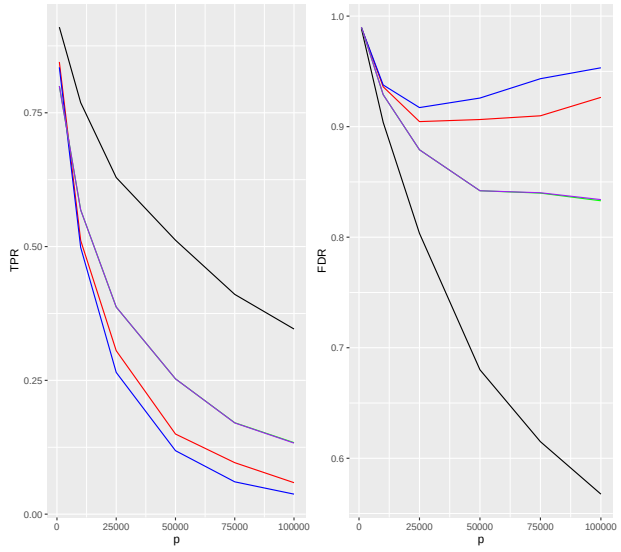

Figure 3.22.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .3)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

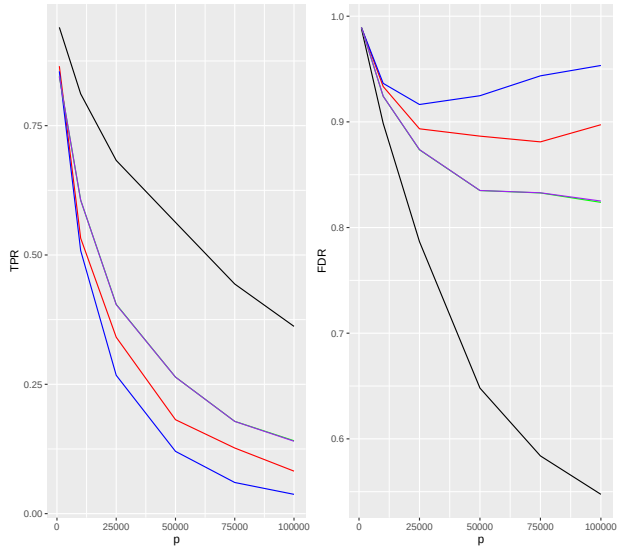

Figure 3.23.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .5)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

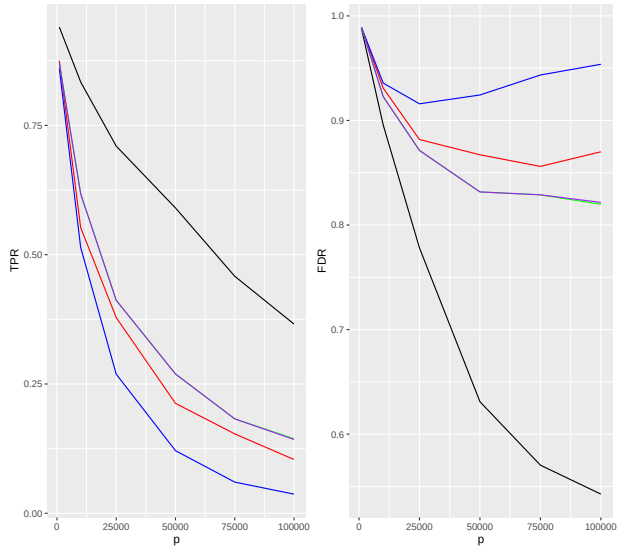

Figure 3.24.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, .7)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

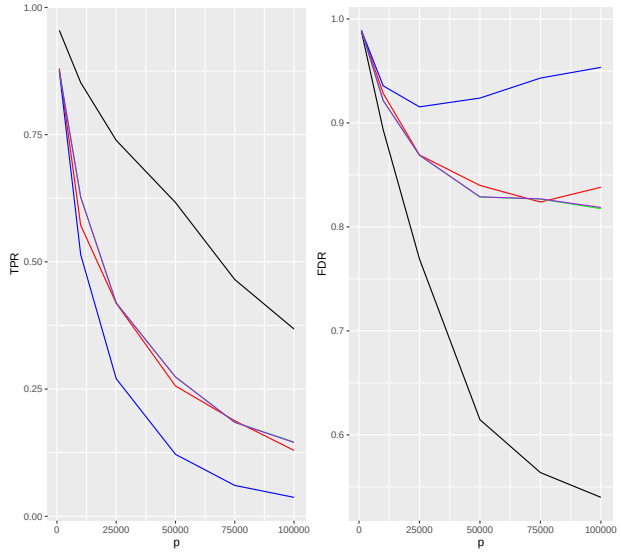

Figure 3.25.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero α and β parameters are generated from $\mathcal{N}(0, 1)$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

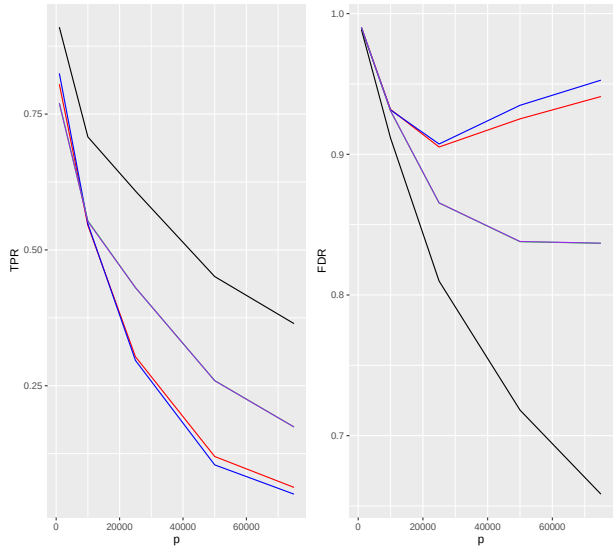

Figure 3.26.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .1$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .1$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

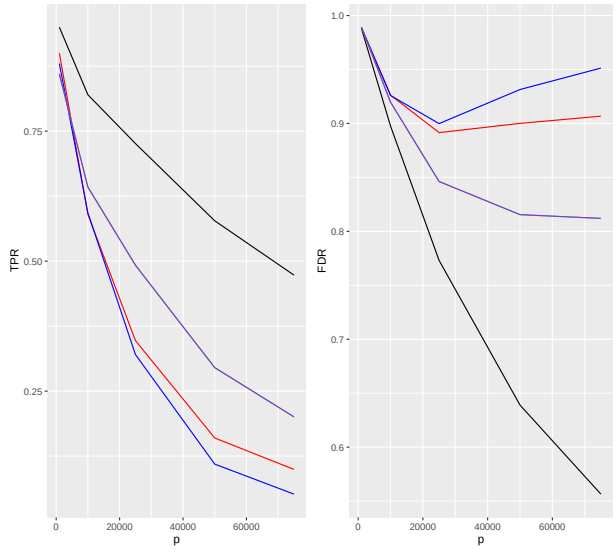

Figure 3.27.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = .3$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = .3$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

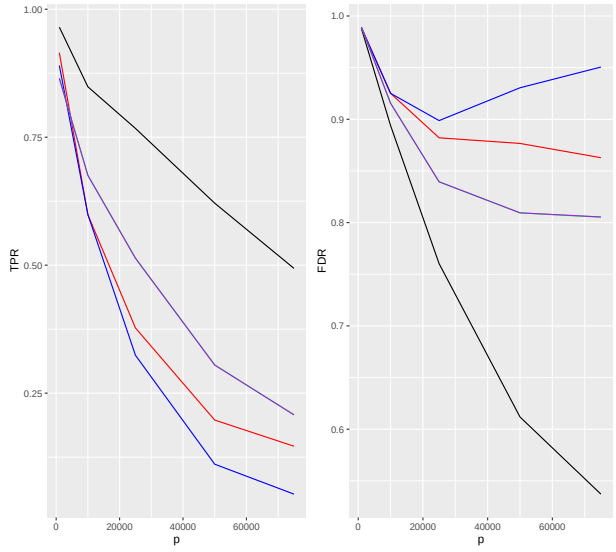

Figure 3.28.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = 0.5$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = 0.5$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

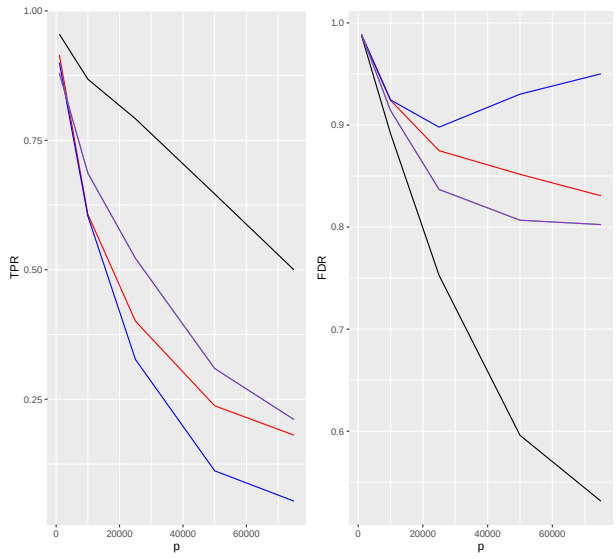

Figure 3.29.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = 0.7$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = 0.7$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

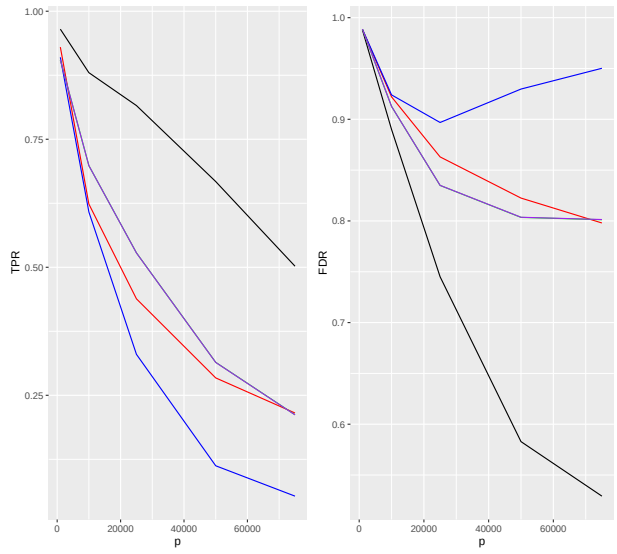

Figure 3.30.: Comparison of TPR and FDR when $N = 500$ is fixed and $p \in \{1000, 10000, 100000\}$ varies. Nonzero $\alpha_{A,j}$ and $\beta_{M,j}$ parameters are generated from UNIF$(-c, c)$, where c is chosen such that $\text{Var}(\alpha_{A,j}) = 1$ for nonzero $\alpha_{A,j}$ and $\text{Var}(\beta_{M,j}) = 1$ for nonzero $\beta_{M,j}$. Here the line representing MSS is in black, SIS in red, CSIS (conditional on X) in blue, DACT in green, and MaxP in purple. Hard thresholding rule is applied with $d = \lfloor N/\log N \rfloor$.

Comparison 3: Simulation of Settings of a Real Data Set

To better understand how our screening approach performs on real data, we create a simulation under the same (N, p) settings as the CARDIA dataset considered in our real data analysis section. This comparison allows us to see how well the screening methods perform when the underlying LSEM structure is known. For this study $p = 860,627$ and $N = 892$. As a "prior" for nonzero parameters, nonzero α and β values are drawn from $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 \in \{.1, .3, .5, .7\}$. In this case, we only consider the new MSS method vs the benchmark SIS method.

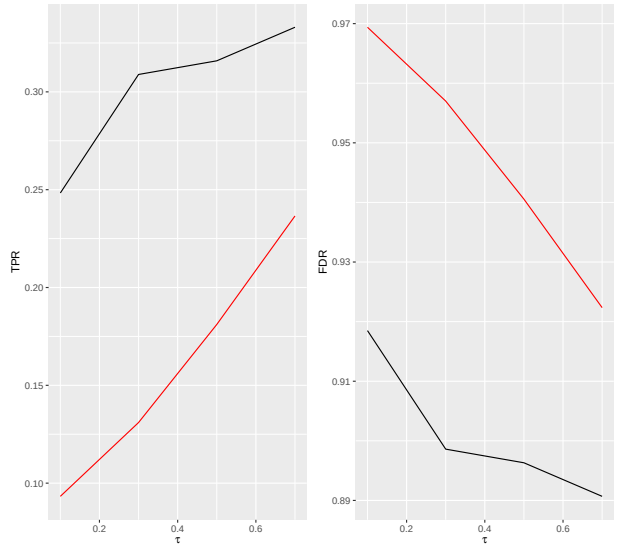

Figure 3.31.: Comparison of TPR and FDR when $N = 892$ and $p = 860,627$. Our MSS screening method performance is shown in black while the benchmark SIS method performance is shown in red. Hard thresholding rule is applied with
$d = 2\lfloor N/\log N \rfloor = 262$

For all levels of σ^2, we see that MSS has a higher true positive rate (TPR) and lower FDR compared to those of SIS. As expected, both methods perform better as the signal strength of nonzero α and β increase.

3.6.2 Real Data Analysis: CARDIA Data Set

As a real world application of our screening approach, we consider mediation analysis in the context of a DNA methylation study. Methyl groups are added to DNA at binding locations known as cytosine-phosphate-guanine (CpG) sites, resulting in changes to gene expression [64]. These methylation markers are potential mediators between an exposure variable and a health response variable. In past studies, scientists usually considered a small set of these potential mediators using ad-hoc selection based on prior information [50].

Using the entire set of methylation markers for a mediation analysis allows a more objective approach that is less dependent on subjective knowledge.

The methylation data come from the Coronary Artery Risk Development in Young Adults (CARDIA) Study, which is a longitudinal study designed to investigate lifestyle and other factors that influence the evolution of coronary heart disease risk factors during young adulthood [65]. The exposure is smoking status at 2 levels: never smokers and current smokers. The response variable in our study is the change in lung function between year 10 and year 15 of the study. The candidate set of mediators $\boldsymbol{M}$ consists of 860,627 DNA methylation measurements. To account for confounding, other subject information such as age, sex, race, testing center, height, and measurement center are included as covariates. Our goal is to use our screening method as part of the multi-step procedure developed by [50] for estimating and identifying which DNA methylation sites mediate the relationship between smoking and lung function.

In our analysis, we use the HIMA and HIMA2 procedures, which are both three-stage procedures for performing mediation analysis [50, 66]. The two procedures have the same basic steps, but they differ in the way each step is performed. The first step is to use SIS to screen the set of candidate mediators and reduce its the dimensionality of the problem from high dimensional p to lower dimensional $d < n$. HIMA applies SIS to the marginal coefficients $\beta_{M,j}^M$, while HIMA2 applies SIS using the product of coefficients $\alpha_{A,j}\beta_{M,j}^M$. Note that the screening step for HIMA2 is equivalent to the version of MSS when the product $\alpha_{A,j}\beta_{M,j}^M$ is used rather than the scaled value $T_{Sobel,j} = \frac{\alpha_{A,j}\beta_{M,j}^M}{\sqrt{(\hat{\beta}_{M,j}^M)^2\hat{\sigma}_\alpha^2 + \hat{\alpha}_{A,j}^2}}$. Our work supports the use of either screener in HIMA2, although the scaled version that uses T_{Sobel} detects more mediators that are declared significant by the overall joint significance test. The second step

of each is to use penalized regression to perform model selection to estimate $\hat{\beta}$. HIMA uses the MCP penalty to perform this step, while HIMA2 produces more appropriate p-values by using the debiased LASSO. During this step, $\hat{\alpha}_{A,j}$ is also estimated for $j \in \mathcal{S}$ where $\mathcal{S} = \{j : \hat{\beta}_{M,j} \neq 0\}$. Finally in the third step, the MaxP test is used to perform a final joint significance test for mediation effect with FDR control. HIMA uses the uniform distribution as the reference distribution of the MaxP statistic, which results in a statistically valid but overly conservative test [67]. HIMA2, on the other hand, uses a null mixture distribution developed by [68] to achieve FDR control while allowing for greater power. For both procedures we compare the performance of MSS as the screening procedure used in Step 1 vs the previously mentioned SIS procedures. To keep explanations shorter and easier to follow, we will refer to our settings as four cases:

- Case 1: HIMA2 procedure with MSS screener
- Case 2: HIMA2 procedure with SIS screener
- Case 3: HIMA procedure with MSS screener
- Case 4: HIMA procedure with SIS screener

The statistically significant mediators identified by each procedure is reported below, along with their corresponding estimates $\hat{\beta}_j$, $\hat{\sigma}_{\beta,j}$, $\hat{\alpha}_{A,j}$, $\hat{\sigma}_{\alpha,j}$, the raw p-value p_{raw}, and indirect effect IE.

In the screening step for HIMA2, $d = 2\lfloor n/\log n \rfloor = 262$ mediators with the highest respective statistic were recruited into Step 2. That is, for SIS the d mediators with highest value of $|\hat{\alpha}_{A,j}\hat{\beta}^{M}_{M,j}|$ are recruited, while for MSS the d mediators with the highest value of

CpG	$\hat{\beta}_{M,j}$	$\hat{\sigma}_{\beta,j}$	$\hat{\alpha}_{A,j}$	$\hat{\sigma}_{\alpha,j}$	p_{raw}	IE
cg01114441	-0.07463	0.01835	-0.09708	0.02134	0.00004779	0.007245
cg01873760	-0.05523	0.01539	-0.08258	0.02610	0.001557	0.004561
cg01907945	-0.04499	0.01572	0.07829	0.02584	0.004210	-0.003523
cg03820608	0.06107	0.01696	-0.08947	0.02353	0.0003175	-0.005464
cg04747388	0.05355	0.01517	0.07926	0.02523	0.001678	0.004244
cg04946953	0.03921	0.01285	-0.1103	0.03033	0.002271	-0.004324
cg05890855	-0.08780	0.02280	-0.08593	0.01759	0.0001172	0.007545
cg05974483	-0.06654	0.01701	-0.1148	0.02418	0.00009171	0.007636
cg06273376	-0.04947	0.01651	0.08539	0.02458	0.002724	-0.004224
cg08544271	-0.04398	0.01280	-0.1120	0.03192	0.0005933	0.004925
cg09277086	-0.03610	0.01272	-0.1013	0.03206	0.004554	0.003655
cg11346960	-0.07749	0.01637	-0.09529	0.02429	0.00008771	0.007384
cg14685642	-0.06156	0.01585	-0.1046	0.02560	0.0001032	0.006439
cg18375153	-0.04561	0.01534	-0.1193	0.02655	0.002941	0.005440
cg21596426	-0.05107	0.01479	-0.09996	0.02635	0.0005523	0.005105
cg23058194	0.03823	0.01273	0.1493	0.03170	0.002663	0.005706
cg24919394	-0.06404	0.01398	-0.1017	0.02880	0.0004167	0.006510
cg25512179	0.06995	0.01499	0.07930	0.02558	0.001939	0.005547
cg26120924	-0.04423	0.01186	-0.1121	0.03247	0.0005578	0.004957
cg26331243	0.06428	0.02183	-0.08424	0.01721	0.003241	-0.005415
cg26946015	-0.08017	0.02095	-0.09033	0.02003	0.0001302	0.007242
cg27363280	-0.06639	0.01386	0.08935	0.02890	0.001991	-0.005932
cg27626216	-0.06860	0.02151	-0.07389	0.01724	0.001428	0.005069

Table 3.1: Results for HIMA2 using MSS as screening procedure

$|T_{Sobel}|$ are recruited. It is worth noting that two screening procedures shared 148 mediators in common. After performing penalized debiased LASSO regression on the screened mediators to obtain $(\hat{\alpha}, \hat{\beta})$ and performing the joint significance test the HIMA2 procedure that used MSS in the screening step identified 23 mediators while the procedure that used SIS identified 15 mediators. After the final step, CpG sites cg05890855, cg05974483, cg14685642, cg18375153, cg25512179, and cg26331243 are common to both sets of significant mediators.

For the HIMA procedure, we see that the choice of screening method makes a huge difference in the number of mediators that are declared significant after the final step of the

CpG	$\hat{\beta}_j$	$\hat{\sigma}_{\beta,j}$	$\hat{\alpha}_{A,j}$	$\hat{\sigma}_{\alpha,j}$	p_{raw}	IE
cg02044044	-0.08810	0.02792	-0.1336	0.01733	0.001604	0.01177
cg02671671	-0.05786	0.01659	-0.09744	0.02750	0.0004859	0.005638
cg04136921	-0.06791	0.01814	-0.09697	0.02509	0.0001820	0.006585
cg05753553	0.05970	0.01667	0.2086	0.02925	0.0003430	0.01246
cg05890855	-0.1014	0.02643	-0.08593	0.01759	0.0001250	0.008713
cg05974483	-0.06845	0.01950	-0.1148	0.02418	0.0004487	0.007855
cg06060595	0.05601	0.01481	0.1663	0.03063	0.0001552	0.009315
cg08644506	-0.09480	0.02302	-0.1332	0.02122	0.00003829	0.01263
cg11152412	-0.04671	0.01533	-0.1736	0.03096	0.002314	0.008108
cg14685642	-0.07954	0.01790	-0.1046	0.02560	0.00004403	0.008320
cg18375153	-0.06614	0.01735	-0.1193	0.02655	0.0001379	0.007889
cg19862839	0.05787	0.01910	-0.08429	0.02469	0.002443	-0.004877
cg22335872	-0.08007	0.02490	-0.1061	0.01780	0.001303	0.008494
cg25512179	0.07508	0.01717	0.07930	0.02558	0.001939	0.005954
cg26331243	0.08013	0.02553	-0.08424	0.01721	0.001695	-0.006750

Table 3.2: Results for HIMA2 using SIS as screening procedure

CpG	$\hat{\beta}_j$	$\hat{\sigma}_{\beta,j}$	$\hat{\alpha}_{A,j}$	$\hat{\sigma}_{\alpha,j}$	p_{raw}	IE
cg03820608	0.07109	0.01744	-0.08947	-0.08947	0.0001432	-0.006360
cg05974483	-0.06818	0.01793	-0.1148	-0.1148	0.0001427	0.007825
cg06273376	-0.05996	0.01636	0.08539	0.08540	0.0005114	-0.005120
cg08544271	-0.05248	0.01261	-0.1120	-0.1120	0.0004498	0.005878
cg08751854	-0.05351	0.01581	-0.1639	-0.1639	0.0007154	0.008771
cg18100580	0.06393	0.01887	-0.1047	-0.1047	0.0007059	-0.006691
cg21596426	-0.05397	0.01507	-0.1000	-0.1000	0.0003435	0.005395
cg24919394	-0.05867	0.01396	-0.1017	-0.1017	0.0004167	0.005963
cg26120924	-0.04527	0.01202	-0.1121	-0.1121	0.0005578	0.005073
cg26331243	0.09671	0.02517	-0.08424	-0.08424	0.0001219	-0.008147

Table 3.3: Results for HIMA when MSS is used in screening step

CpG	$\hat{\beta}_j$	$\hat{\sigma}_{\beta,j}$	$\hat{\alpha}_{A,j}$	$\hat{\sigma}_{\alpha,j}$	p_{raw}	IE
cg01320698	-0.1129	0.03440	-0.03815	-0.03815	0.001993	0.004308
cg05127574	-0.08570	0.02825	0.07125	0.07125	0.002419	-0.006106
cg07487014	-0.08333	0.02839	0.06634	0.06634	0.003338	-0.005528

Table 3.4: Results for HIMA when SIS is used in screening step

procedure. When using MSS in Step 1, the HIMA procedure declares 10 mediators to be significant in the mediation model, while only 3 are significant when SIS applied to the $\hat{\beta}^M_{M,j}$ is used.

It is also worth noting that case 4 shares no CpGs with any of the other methods. Case 3 shares 2 mediators with case 2: cg05974483 and cg26331243. Case 3 also shares 8 mediators with case 1: cg03820608, cg05974483, cg06273376, cg08544271, cg21596426, cg24919394, cg26120924, and cg26331243. CpG sites cg05974483 and cg26331243 are common to the results of all three procedures and may be of special interest. CpG cg26331243 is located in the body region of gene CCDC33, which is has been shown in previous studies to be differentially expressed under the experimental condition of tobacco smoke exposure [66, 69, 70]. CCDC33 is also linked to susceptibility to lung disease such as pneumococcal meningitis and SARS-CoV-2 infection [71, 72]. It is plausible that cg26331243 plays a role in regulating the expression of CCDC33, which in turn mediates the pathway from smoking to lung function. CpG cg05974483 is located on gene NXPH3, which enables signaling receptor binding activity via a protein called Neurexophilin-3 [73, 74]. This newly identified gene is functionally important for sensorimotor gating and motor coordination [75] and may be worth investigating for further evidence of mediation effect.

3.7 Discussion

After reviewing mediation analysis in the traditional low-dimensional cases, we have developed and compared methods for screening mediators in a high dimensional setting. In particular, we have proved the Sure Screening Property for the Marginal Sobel Screen-

ing method and demonstrated cases in which it performs better than traditional screening methods such as Sure Independence Screening in the mediation setting. After reducing a high-dimensional mediation problem to a low dimensional problem via screening, traditional methods can then be applied to perform inference on the reduced set of candidate mediators. This screening method is easy to implement and interpret for practitioners. We then applied the rest of the multi-stage mediation analysis developed by [50, 66] by fitting a penalized regression model after screening, then performing final mediation testing on the reduced set of mediator variables. In our real data analysis we showed the improvement to HIMA and HIMA2 when using our new screener for screening.

One point of note when performing the multi-stage mediation analysis procedure is that inference conducted at later stages is performed conditional on earlier screening stages. Valid post-selection p-values and confidence intervals can be obtained by considering the asymptotic distribution of T_{Sobel} or $\hat{\alpha}\hat{\beta}^M$ and the behavior of the MSS screener.

Clearly, this work can be expanded into many directions. The framework described in this paper can be readily extended from LSEMs to generalized structural equation models or sequential mediation models. We also would like to consider various correlation structure between the mediators $\boldsymbol{M}$ to improve our screening approaches in the future.

4. Privacy-Preserving Penalized Quantile Regression ADMM

4.1 Introduction

Quantile regression is a popular and versatile regression technique that offers a systematic strategy for examining how covariates influence the entire response distribution [76]. Instead of estimating the conditional mean as in least squares regression, quantile regression models the relationship between the covariates and the conditional quantiles of the response. Compared with its least squares counterpart, quantile regression provides a more flexible approach to learn about responses that come from a distribution that does not follow the standard OLS assumptions. For example, quantile regression estimates are robust to outliers and heteroscedastic data which is common in medical, educational, and social science settings, among many others [76].

In our modern era of big data, it is increasingly common to build high-dimensional models from huge numbers of observations. As the number of parameters in these models increases, the required number of observations can quickly grow beyond the capacity or capability of a single computational entity. Such an analysis would depend heavily on collaboration between institutions, i.e. data sharing. This raises concerns for user privacy, especially after high-profile attacks on widely-used online platforms [77, 78]. Summary statistics and

local parameter estimates can also compromise sensitive user data or institutional data, and should therefore be shared with caution [79]. Privacy laws such as HIPAA and FERPA make medical data and academic data examples of cases where a privacy-preserving collaborative scheme may be especially useful [80, 81].

When discussing data sharing, it is important to also discuss network structure. One such structure we consider is a centralized network in which all agents are in communication with a secure and powerful central computer that takes inputs from the agents and returns an output to all agents. Centralized systems are attractive because they reduce data redundancy (repeated copies of data) and are financially cost effective [82]. Some famous institutions that use centralized structures include IBM and the National Informatics Center in India [83, 84]. The strong central computer can also unfortunately be a weakness for this architecture because the system fails if the agents lose connectivity to the central computer.

We also consider a decentralized network in which all agents have the computational resources to carry all steps of the algorithm without dependence on a central entity. Decentralized networks have become increasingly popular in recent years due to the increased power of modern computers and the rise of bitcoin and blockchain technology, which use decentralized networks to preserve privacy of personal data [85]. Decentralized networks have the advantage of robustness to network bottlenecks or failures of single links [86]. That is, if direct communication between agent A and agent B fails, there is still a path in the network through which the two can exchange information. Although the initial cost of such a network is high, decentralized algorithms also lend themselves well to scalabilty. One possible disadvantage of these structures is that decentralized networks require more information to be exchanged between agents which can cause failure for low-bandwidth networks. As

personal computers become more powerful, the line between centralized and decentralized systems is blurring [83]. For example, cloud computing architectures may involve agents directly communicating with one another as well as with a cloud server.

To make the problem concrete, suppose that K agents wish to collaborate on a quantile regression analysis to estimate the τth quantile of a response variable given their combined data. Each agent j collects n_j observations independently such that $\sum_{j=1}^{K} n_j = N$. In our schemes, we make the assumption that all agents honestly compute the proper calculations when told but also passively listen to other agents for information. This is a security model commonly known as "honest but curious" in the cryptography literature [87]. In this case, we do not consider cases in which agents are corrupted and provide false information. Our goal is to devise a scheme that allows the agents to perform quantile regression efficiently while maintaining security of the shared information. Here, "security" means that no agent inside or outside the scheme learns anything of the true values data transmitted by other agents.

Let $\mathbf{y} \in \mathbb{R}^N$ denote the observed response vector, $\mathbf{X} \in \mathbb{R}^{N\times(p+1)}$ the design matrix, and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ a parameter vector. For $\tau \in (0,1)$ suppose that the τth conditional quantile of $y|X$ is linear in $\boldsymbol{\beta}$. That is, $Q_\tau(y|X) = X\boldsymbol{\beta}$, where Q_τ denotes the τth quantile. Then the quantile regression estimate $\hat{\boldsymbol{\beta}}(\tau)$ is given by

$$\hat{\beta}(\tau) = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\, \rho_\tau(y - X\boldsymbol{\beta}), \tag{4.1}$$

where for $u \in \mathbb{R}^n$, $\rho_\tau(\cdot)$ denotes the check loss function $\rho_\tau(u) = \sum_{i=1}^n u_i[\tau - \mathbb{I}(u_i < 0)]$, and $\mathbb{I}$ denotes the indicator function. This estimation is typically solved using linear programming methods such as interior point search or the simplex method [76].

In the high dimensional setting, i.e. large p, quantile regression is often coupled with regularization for variance reduction and feature selection. The penalized quantile regression (PQR) problem is typically formulated as

$$\text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \rho_\tau(\mathbf{y} - X\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}_{-0}), \tag{4.2}$$

where $\boldsymbol{\beta}_{-0}$ denotes the parameter vector with the intercept removed and P_λ is a penalty function. Popular choices for penalty functions include the lasso [2], elastic net [3], MCP [14], and SCAD [13]. The aforementioned linear programming methods for solving (4.2) often prove too computationally intensive to scale effectively for big data [76]. To remedy this, recent works use variants of the alternating direction method of multipliers (ADMM) algorithm which has been shown to work well for distributed convex optimization problems [88]. Gu et al. (2017) proposed a proximal ADMM and sparse coordinate descent ADMM to solve the penalized quantile regression problem with the lasso, adaptive lasso, and folded concave penalties. Yi and Huang (2016) proposed a coordinate descent algorithm for solving the elastic-net penalized Huber regression and used that to approximate the penalized quantile regression estimator. QPADM [89] offers a distributed algorithm for solving (4.2) based on ADMM with guaranteed convergence for convex penalty functions and empirical success for some commonly used nonconvex penalties, such as MCP and SCAD [89]. Our proposed schemes use the QPADM formulation to transform (4.2) into a distributed opti-

mization problem for the agents to solve. Based on this framework, we develop a privacy preservation step to ensure that no agents can obtain the shared data from any other agent.

Section 2 of this paper reviews the QPADM algorithm and its computation. Section 3 will detail some of the privacy-preservation methods considered for our schemes. We then outline the overall centralized algorithm in Section 4, and the decentralized algorithm in Section 5. An algorithm based on differential privacy is outlined in Section 6. Finally, we compare the computation speed of our privacy-preserving QPADM algorithms to the plain non-secure algorithm in Section 7 with discussions in Section 8.

4.2 QPADM

The key to QPADM lies in converting equation (4.2) to its equivalent parallelized ADMM form. First, assume there are K agents with combined data

$$\mathbf{y} = [\mathbf{y}_1^T, \mathbf{y}_2^T, ..., \mathbf{y}_K^T]^T \text{ and } X = [X_1^T, X_2^T, ..., X_K^T]^T,$$

where $\mathbf{y}_j$ is a n_j-dimensional response vector for the jth agent and X_j is a $n_j \times (p+1)$ dimensional design matrix for the jth agent. Each agent j's data, $(X_j, \mathbf{y}_j)$, can be seen as a block or partition of the combined data. Problem (4.2) can then be rewritten as a constrained optimization problem by introducing auxillary variables $\mathbf{r}_j$

$$\min_{\mathbf{r}_j, \boldsymbol{\beta}} \left\{ \sum_{j=1}^{K} \rho_\tau(\mathbf{r}_j) + P_\lambda(\boldsymbol{\beta}_{-0}) \right\} \text{ s.t. } \mathbf{y}_j - X_j\boldsymbol{\beta}_j = \mathbf{r}_j, \boldsymbol{\beta}_j = \boldsymbol{\beta}, j = 1, 2, ..., K$$

with the Lagrangian

$$L_\gamma(\mathbf{r}, \mathbf{u}, \boldsymbol{\beta}) = \sum_{j=1}^{K} \left[\rho_\tau(\mathbf{r}_j) + \mathbf{u}_j^T(\mathbf{y}_j - X_j\boldsymbol{\beta}_j - \mathbf{r}_j) + \frac{\gamma}{2}\|\mathbf{y}_j - X_j\boldsymbol{\beta}_j - \mathbf{r}_j\|_2^2 \right] + P_\lambda(\boldsymbol{\beta}_{-0}),$$

This function can then be minimized using ADMM alternatively over $r, u,$ and $\boldsymbol{\beta}$. At iteration $k+1$, the update rule is

$$\begin{aligned}
\boldsymbol{\beta}^{(k+1)} &:= \text{argmin}_{\boldsymbol{\beta}} \frac{K\gamma}{2}\|\boldsymbol{\beta} - \bar{\beta}^{(k)} - \bar{\eta}^{(k)}\|_2^2 + P_\lambda(\boldsymbol{\beta}_{-0}), \\
\mathbf{r}_j^{(k+1)} &:= \text{argmin}_{\mathbf{r}_j} \rho_\tau(\mathbf{r}_j) + \frac{\gamma}{2}\|\mathbf{y}_j - X_j\beta_j^{(k+1)} + \mathbf{u}_j^{(k)} - \mathbf{r}_j\|_2^2, \\
\boldsymbol{\beta}_j^{(k+1)} &:= (X_j^T X_j + I)^{-1}\left(X_j^T(\mathbf{y}_j - \mathbf{r}_j^{(k+1)} + \mathbf{u}_j^{(k)}) - \eta_j^{(k)} + \boldsymbol{\beta}^{(k+1)}\right), \\
\mathbf{u}_j^{(k+1)} &:= \mathbf{u}_j^{(k)} + \mathbf{y}_j - X_j\boldsymbol{\beta}_j^{(k+1)} - \mathbf{r}_j^{(k+1)}, \\
\boldsymbol{\eta}_j^{(k+1)} &:= \eta_j^{(k)} + \boldsymbol{\beta}_j^{(k+1)} - \boldsymbol{\beta}^{(k+1)}.
\end{aligned} \tag{4.3}$$

where $\bar{\beta}^{(k)} = K^{-1}\sum_{j=1}^{K}\boldsymbol{\beta}_j^{(k)}$ and $\bar{\eta}^{(k)} = K^{-1}\sum_{j=1}^{K}\eta_j^{(k)}$ [89].

Notice that the last four updates with subscript j only depend on the jth block of data, so these updates can be performed locally and independently by each agent in the study. The bulk of the computation difficulty lies in computing the β-update, especially when the penalty function P_λ is nonconvex. This step has a closed form solution for many common

penalties, including the lasso, elastic net, MCP, and SCAD penalties [89]. The $\mathbf{r}$-update also has a closed form solution

$$\mathbf{r}_j^{(k+1)} := \left[\gamma^{-1}\mathbf{u}_j^{(k)} + \mathbf{y}_j - X_j\boldsymbol{\beta}_j^{(k+1)} - \tau\gamma^{-1}\mathbb{1}_{n_j}\right]_+ - \left[-\gamma^{-1}\mathbf{u}_j^{(k)} - \mathbf{y}_j + X_j\boldsymbol{\beta}_j^{(k+1)} + (\tau-1)\gamma^{-1}\mathbb{1}_{n_j}\right]_+. \tag{4.4}$$

Both of our schemes take advantage of this parallelizability to assign agents to computing local parameter estimates. The key difference in these two algorithms is how secure aggregation of the local estimates is performed.

4.3 Privacy Preservation

For agents involved in the study the most important part of our algorithms is individual data privacy and protection of summary statistics, i.e. the goal for agent j is to protect each individual entry of $(X_j, \mathbf{y}_j)$ (referred to henceforth as individual subject data) as well as any statistic computed from $(X_j, \mathbf{y}_j)$. Our schemes make the assumption that agents protect the privacy of individual data by storing individual data in a secure location controlled only by that agent. In other words, other agents in the scheme are unable to view the individual subject data belonging to other agents. Under the "honest but curious" assumption, any data transmitted between agents are public to all agents inside or outside the scheme. Under this condition, we define privacy to mean that if an agent A has his message overheard, any listeners do not obtain any information about any of A's individual subject data or summary statistics.

Our schemes also preserve privacy of summary statistics, which can again pose a danger to individual data within the institution who shares its summary statistics. The summary statistics may also contain information that an institution wishes to keep private, especially in situations where the collaborating institutions are competitors. In the case of QPADM, these summary statistics are the local regression parameter estimates $\beta_j^{(k+1)}$ and $\eta_j^{(k+1)}$ which are exchanged to compute the global parameter estimates $\bar{\beta}^{(k)}$ and $\bar{\eta}^{(k)}$ (Equation 4.3). For our centralized scheme, this summary statistic privacy preservation takes place at two levels using the help of J computation centers ($J \leq K$) and a global center (Figure 4.1). At the first level, agents encrypt their local summary statistics before sending them to the J computation centers, which are independent and secure locations that aggregate the encrypted local summary statistics. Note that since the computation centers only perform operations on encrypted data, no one at the computation centers can see the local statistics of any agent. The computation centers then send their encrypted aggregates to a global center, which decrypts the aggregates and sends the update back to the agents. For our decentralized scheme there is only one step that involves exchanging local parameter estimates, but this step is carried out between every pair of agents in a way that preserves the parameter estimates of each agent (Figure 4.2).

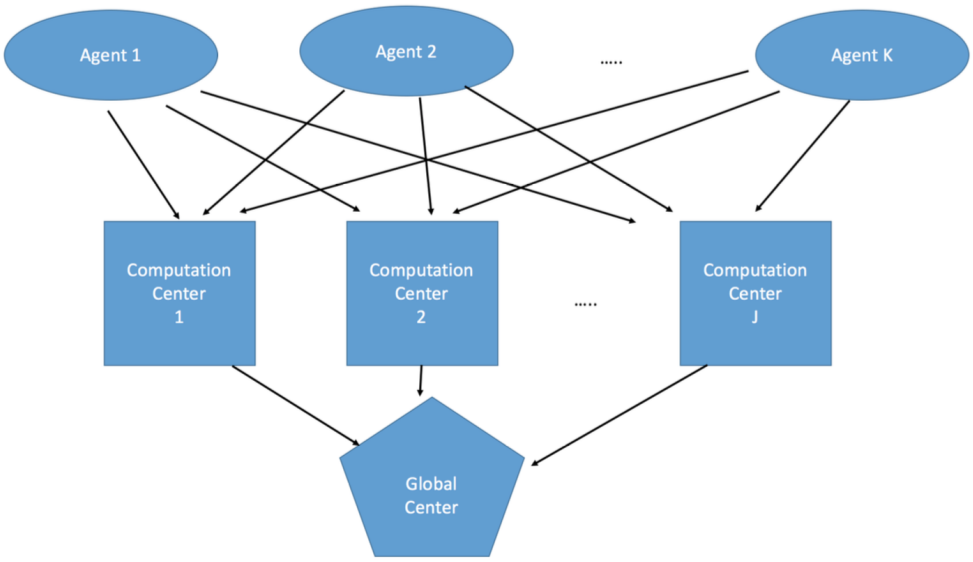

Figure 4.1.: Illustration of Centralized Scheme. Each arrow denotes the transmission of encoded data that can only be decoded by the global center after receiving J encoded aggregates from the J computation centers.

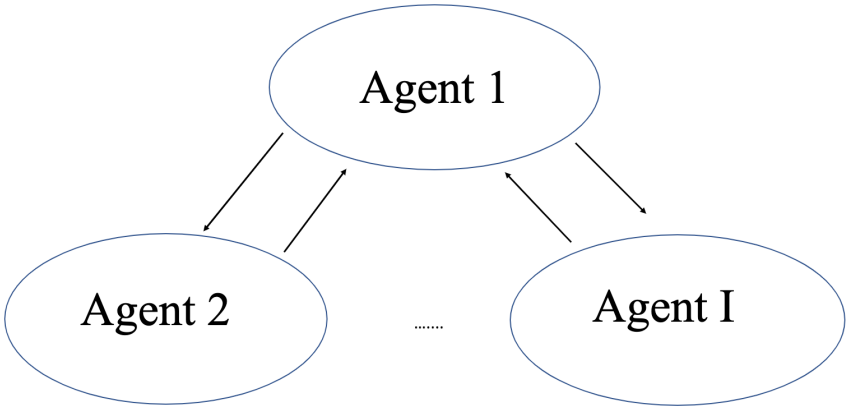

Figure 4.2.: Illustration of Decentralized Scheme: Specifically Algorithm 5 for agent 1. The outer arrows denote the sending of public keys to other agents. Upon receiving public key(s), each agent encodes their local statistics using the public key and sends these encoded statistics back to agent 1. Agent 1 then decodes these quantities using their private key.

The key to the secure aggregation used in these schemes is additively homomorphic encryption. An encryption scheme is said to be additively homomorphic if certain mathematical

operations can be applied directly to the ciphertext in such a way that decrypting the output results in the sum of original unencrypted data [90]. Formally, the plaintext and ciphertext are related by a group homomorphism. Specifically, if G is the domain of the plaintext and H an Abelian group with group operation "$\star$", then the group homomorphism from $(G, +)$ to $(H, \star)$ is a mapping $f : G \to H$ that satisfies the condition $f(x_1) \star f(x_2) = f(x_1 + x_2)$ for all $x_1, x_2 \in G$. This mapping f denotes an encryption function. Given a plaintext number x, $f(x)$ denotes the encoded ciphertext and $f^{-1}(f(x)) = x$ denotes decrypting the cyphertext to obtain the original plaintext x. The homomorphic property of such an f allows us to perform computation on encrypted data without the need for decryption during intermediate steps of our algorithm. After combining this encrypted data, the result can be decrypted to obtain the desired plaintext result. In creating our schemes we consider two classical additively homomorphic cryptosystems, but our idea can be readily extended to use other additively homomorphic schemes. More complicated cryptosystems may provide further security but also result in higher computational burden (i.e. they may take more time or memory). Encrypting large amounts of data is known to be computationally expensive [91], so for our study we only consider two foundational homomorphic cryptosystems: Shamir's secret-sharing algorithm (centralized) and the Paillier cryptosystem (decentralized).

The method of privacy preservation used in our centralized algorithm study is Shamir's secret-sharing algorithm [92]. This algorithm is commonly researched in the privacy preservation literature and is a valuable tool for multiparty computation schemes [93]. In Shamir secret-sharing, a plaintext number S is divided into pieces $S_1, S_2, ..., S_n$ in such a way that knowledge of k or more pieces allows easy computation of S while knowledge of $k - 1$ or fewer pieces leaves all possible values of S equally likely [92]. These desired characteristics

are attained by using a key fact: to uniquely determine a polynomial of degree $k-1$, k coordinate points are needed. To protect a plaintext S, we generate a random polynomial $p(x)$ of degree $k-1$ with the plaintext S as its intercept. That is, $p(x) = S + \sum_{i=1}^{k-1} a_i x^i$, where S is the secret to be protected and the a_i are randomly generated coefficients. Using this polynomial, S is split into pieces $S_i = p(i)$, $i = 1, 2, ..., n$. Given k of these pieces $p(x)$ is simple to compute using Legendre interpolation, and S can be recovered by evaluating $p(0)$. Note that we can only encrypt scalars in this manner, so each entry of the parameter vectors must be encoded separately. Thus, each agent must generate p polynomials of degree $k-1$ to encode their parameter estimates. Since Shamir's secret-sharing algorithm is additively homomorphic, the encoded secrets can then be aggregated securely by the computation centers to create an encoded version of the aggregated statistic desired.

Our decentralized algorithm study depends on another homomorphic encryption scheme known as the Pallier cryptosystem [94]. This scheme is also additively homomorphic like Shamir's algorithm, but does not break the secret into pieces. Instead it utilizes a public-private key system that allows for agents to securely send information directly to one another. Each agent in the decentralized scheme possesses a public key that can be used to encode data by anyone and a private key that allows the agent to decrypt information encoded by the public key. Just as with Shamir's secret sharing, the Paillier cryptosystem can only encode scalars so p public and private keys must be used to encode parameter vectors. The public keys allow agents to exchange information safely without giving this information away to outside listeners, while the private key allows an agent to decrypt summary statistics once they have received enough data. See Algorithms 4 and 5 for further details.

Another notion of privacy preservation is statistical privacy in the sense that if an attacker has access to all of a database D except for one entry, a statistic computed from the full database D offers no new information about the final entry. Differential privacy is a popular technique that offers this statistical privacy [95].

To make this clear, define two databases D and D' as neighboring or adjacent if their Hammig distance is 1, i.e. they differ by only one element. A randomized mechanism M provides ϵ-differential privacy if for all neighboring databases D and D' and for any $S \subseteq \text{Range}(M)$, $P[M(D) \in S] \leq e^{\epsilon} P[M(D') \in S]$ [95]. The term ϵ is often referred to as the leakage, which represents how much privacy is leaked by the mechanism. When ϵ is small, $e^{\epsilon} \approx 1 + \epsilon$, so this can be thought of as the requirement that for all $S \in \text{Range}(M)$, $\frac{P[M(D) \in S]}{P[M(D') \in S]} \in (1 - \epsilon, 1 + \epsilon)$.

Differentially private QPADM protects the privacy of agents' local data through statistical privacy rather than a cryptosystem [95]. That is, we perturb the individual user data in such a way that an attacker cannot determine whether or not a particular user is a part of any differentially private database. Since this scheme does not need to encrypt and decrypt data for each step of the algorithm, this privacy scheme is expected to perform much faster than the previously mentioned schemes. As a tradeoff, the perturbation of the data introduces error that must be accounted for.

4.4 Centralized Algorithm

Our centralized algorithm takes place in three stages. In the first stage, each agent uses the previously computed global parameter estimate to compute local estimates of the

QPADM parameters. That is, each agent A_i computes local $\mathbf{r}_i, \boldsymbol{\beta}_i, \mathbf{u}_i$, and $\boldsymbol{\eta}_i$ $(i = 1, 2, ..., K)$ (Equation 4.3). Agent i then encodes $\boldsymbol{\beta}_i$ and $\boldsymbol{\eta}_i$ using Shamir's secret sharing algorithm and send an encoded estimate to each of the J computation centers, say $\beta_{is}^{k+1}(j)$ and $\eta_{is}^{k+1}(j)$ denote the encoded parameter estimates sent from agent i to computation center j. Via a secure aggregation technique, each computation center is able to aggregate the agents' parameter estimates into a new global estimate without revealing the local information or the global information. Each computation center j computes secret-protected aggregates $\bar{\beta}_s^{(k+1)}(j)$ and $\bar{\eta}_s^{(k+1)}(j)$. Finally, the J computation centers submit their secure global estimates to the global center, which can then decode the global estimate and update the global estimates for the agents.

Algorithm 1 Secure Local Computation

procedure SECURELOC(β^k) ▷ Returns Shamir secret shares of $\boldsymbol{\beta}_j^{k+1}, \boldsymbol{\eta}_j^{k+1}$
  for Agent $i = 1, 2, ..., K$ **do**
    Compute r_i^{k+1} and $\boldsymbol{\beta}_i^{k+1}$.
    Compute u_i^{k+1} and $\boldsymbol{\eta}_i^{k+1}$.
    Make encoding polynomials β_{is}^{k+1} and η_{is}^{k+1} such that $\beta_{is}^{k+1}(0) = \beta_i^{k+1}$ and $\eta_{is}^{k+1}(0) = \eta_i^{k+1}$
    for Computation Center $j = 1, 2, ..., J$ **do**
      Securely submit the encoded $\beta_{is}^{k+1}(j)$ and $\eta_{is}^{k+1}(j)$ to center j.

Algorithm 2 Secure Aggregation at jth Computation Center

procedure SECUREAGG($\beta_{1s}^{k+1}(j), ..., \beta_{Is}^{k+1}(j), \eta_{1s}^{k+1}(j), ..., \eta_{Is}^{k+1}(j), \lambda$)
  Aggregate $\bar{\beta}_s(j)^{(k+1)} = \sum_{i=1}^{I} \beta_{is}(j)^{(k+1)}/I$.
  Aggregate $\bar{\eta}_s(j)^{(k+1)} = \sum_{i=1}^{I} \eta_{is}^{(k+1)}(j)/I$.
  Securely submit secret-protected aggregates $\bar{\beta}_s^{(k+1)}(j)$ and $\bar{\eta}_s^{(k+1)}(j)$ to global aggregator.

Algorithm 3 Global Update

procedure GLOBUPDATE($\beta_{1s}^{k+1}, ..., \beta_{Js}^{k+1}, \eta_{1s}^{k+1}, ..., \eta_{Js}^{k+1}, \lambda$)
  Use computation center aggregates to decrypt $\bar{\beta}^{(k+1)}$ and $\bar{\eta}^{(k+1)}$.
  Compute $\beta^{(k+1)}$, where

$$\beta^{\text{new}} = \min_{\boldsymbol{\beta}} \frac{J \cdot \gamma}{2} \|\boldsymbol{\beta} - \bar{\beta}^{(k+1)} - \bar{\eta}^{(k+1)}\|_2^2 + P_\lambda(\boldsymbol{\beta}_{-0}^{(k+1)})$$

  Transmit $\boldsymbol{\beta}^{(k+1)}$ to all agents.

4.5 Decentralized Algorithm

Compared to the centralized scheme, the decentralized scheme offers some advantages and disadvantages. The clear advantage is that no single entity is entrusted to aggregate the local statistics. As a tradeoff, more data transmission is required: each agent sends a private key to all other agents, resulting in up to $2(K-1)K$ data transmissions total. Our decentralized scheme may also be useful in situations where a centralized computation center may pose a security risk. For example, rival companies or institutions may wish to collaborate in a study to mutually improve performance without entrusting a third party to aggregate their data. Decentralized schemes are also attractive for their robustness to network traffic bottlenecks and ease of scalability [86].

As mentioned, this decentralized case makes use of the Paillier cryptosystem (Pailier, 1999). This is a public-private key system, so each agent i has a public key E_i and a private key D_i such that $D_i(E_i(x)) = x$ for all integers x. The public key can be shared with other agents for them to encode messages while the private key is known only to agent i. This cryptosystem is also additively homomorphic in that $E_i(x_1)E_i(x_2) = E_i(x_1 + x_2)$ for any integers x_1 and x_2. This allows agent i to securely aggregate encrypted data via multiplication without the need to decrypt the data he receives.

There is one remaining trivial security flaw that must be accounted for. If agent i receives a local statistic $E_i(\beta_j^{k+1})$ or $E_i(\eta_j^{k+1})$ from agent j, he can simply decode this statistic and compromise agent j's security. To prevent this, agent j instead sends $E_i(\beta_j^{k+1} + b_j)$ and $E_i(\eta_j^{k+1} + c_j)$, where b_j and c_j are secret constants known only to agent j, and $\sum_{j=1}^{K} b_j = B$ and $\sum_{j=1}^{K} c_j = C$ are constants known to all agents in the scheme.

Just as in the centralized case, the algorithm starts with each agent performing secure computation of local parameter estimates. Once these are computed, each agent i sends its public key E_i to all other agents. Each agent j, after receiving a public key E_i, replies with $\mathcal{E}_i(\beta_j + b_j)$ and $\mathcal{E}_i(\eta_j + c_j)$. After each agent collects the encrypted data, aggregates $\bar{\beta}$ and $\bar{\eta}$ are computed by each agent, and the β-update is computed as before.

Algorithm 4 Secure Local Computation for Agent i

procedure SECURELOC($\boldsymbol{\beta}^k$) ▷ Returns $\boldsymbol{\beta}_i^{k+1}, \boldsymbol{\eta}_i^{k+1}$
    Compute r_i^{k+1} and $\boldsymbol{\beta}_i^{k+1}$.
    Compute u_i^{k+1} and $\boldsymbol{\eta}_i^{k+1}$.

Algorithm 5 Secure Local Aggregation for Agent i

procedure PAILLIERAGG($\boldsymbol{\beta}_i^{k+1}$, $\boldsymbol{\eta}_i^{k+1}$) ▷ Compute update $\boldsymbol{\beta}^{k+1}$
    for agent $j = 1, 2, ..., i-1, i+1, ..., K$ **do**
        Send public key(s) E_i to agent j.
        Agent j returns $E_i(\boldsymbol{\beta}_j^{k+1} + b_j)$ and $E_j(\boldsymbol{\eta}_i^{k+1} + c_j)$ to agent i
    Compute $\prod_{j=1}^K E_i(\boldsymbol{\beta}_j^{k+1} + b_j) = \mathcal{E}_i(\sum_{j=1}^K \beta_j^{k+1} + B)$
    Compute $\prod_{j=1}^K E_i(\boldsymbol{\eta}_j^{k+1} + c_j) = \mathcal{E}_i(\sum_{j=1}^K \eta_j^{k+1} + C)$
    Use private key(s) D_i to decode the above quantities. Subtract the known constants and divide by K to obtain $\bar{\beta}^{(k+1)}$ and $\bar{\eta}^{(k+1)}$.
    $\beta^{(k+1)} = \min_{\boldsymbol{\beta}} \frac{J \cdot \gamma}{2} \|\boldsymbol{\beta} - \bar{\beta}^{(k+1)} - \bar{\eta}^{(k+1)}\|_2^2 + P_\lambda(\boldsymbol{\beta}_{-0}^{(k+1)})$

4.6 Simulation Study

In our simulation study, we compare the runtime and accuracy of our two algorithms. We also include preliminary results for an algorithm based on differential privacy, with further comments in the discussion section. For each algorithm, we test for accuracy in case information is lost during the encryption-decryption process. Naturally, the decentralized scheme is expected to take significantly longer in total runtime, due to each agent needing to

calculate the global summary statistics. For both simulations we used $p = 1000$ parameters and varied the sample size n. For each experiment, the data is generated from the true model :

$$y_i = 2x_{2i} - 3x_{5i} + x_{6i} - x_{9i} + x_{1i}\epsilon_{1i}, \text{ where } \epsilon_{1i} \sim N(0, 1)$$

(i.e. a heteroskedastic normal model). We then applied our three methods to each dataset using the SCAD penalty with hyperparameter $a = 3.7$ (determined empirically) to obtain the quantile regression estimator for the quantiles $\tau = .3, .5$, and $.7$. For each sample size n we considered, we repeated this experiment 100 times and recorded the Monte Carlo error, runtime, and final estimate for β. We then averaged over these 100 trials for each n and recorded our results. All experiments were run on a linux-based 8-core server with an Intel Xeon 3.3GHz processor and 16GB RDIMM.

From this, we can also see how the runtime of the three algorithms scales with the sample size. We also run a controlled case without any privacy preservation so we can get a better idea of how the privacy constraint affects our distributed optimization. In each study, we have 3 agents jointly performing quantile regression. To control for network latency, all agents were simulated on one computer with the total runtime between all agents recorded. Due to time and memory constraints, smaller sample sizes were used for the decentralized algorithm experiment compared to the centralized algorithm experiments. These experiments were repeated 100 times per case, with means and standard errors plotted in Figures 4.3 and 4.4.

As expected, adding encryption steps to the QPADM increases the runtime in both cases. The centralized method takes roughly 4 times longer to perform than the non-private method, while the decentralized method takes nearly 10 times longer in total runtime. For

each generated dataset, the exact same quantile regression estimate is obtained from our secure algorithm as standard unsecured QPADM. As shown in Figure 4.5, no information is lost during the encryption, decryption, or secure aggregation.

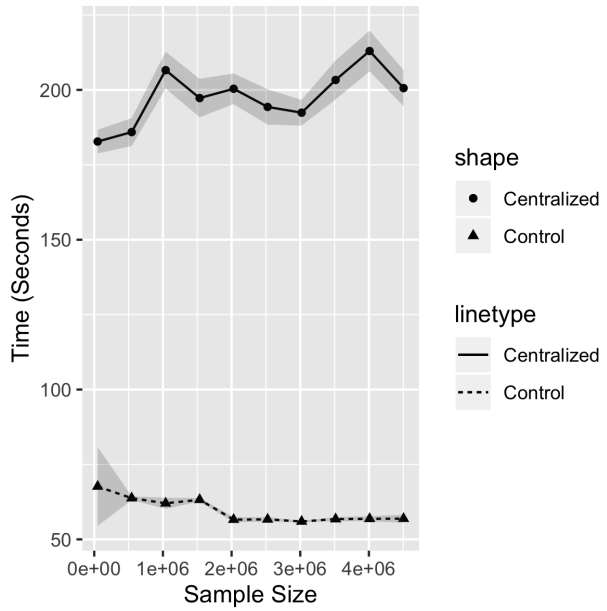

Figure 4.3.: Comparison of Runtimes for Control Case and Centralized Scheme. Shaded regions denote 95% bootstrap confidence bands.

Table 4.1: Performance Analysis of Centralized and Decentralized Algorithms on Synthetic Data, $\tau = 0.5$

Algorithm	Control	Centralized	DP	Control	Decentralized	DP
Number of Samples	4,505,000	4,505,000	4,505,000	10,000	10,000	10,000
Number of Parameters	1,000	1,000	1,000	100	100	100
Total Runtime (s)	56.90476	200.5945	60.811	20.01143	1818.368	26.371

Table 4.2: Performance Analysis of Centralized and Decentralized Algorithms on Synthetic Data, $\tau = 0.3$

Algorithm	Control	Centralized	DP	Control	Decentralized	DP
Number of Samples	4,505,000	4,505,000	4,505,000	10,000	10,000	10,000
Number of Parameters	1,000	1,000	1,000	100	100	100
Total Runtime (s)	54.818	204.298	55.014	21.007	1824.693	24.816

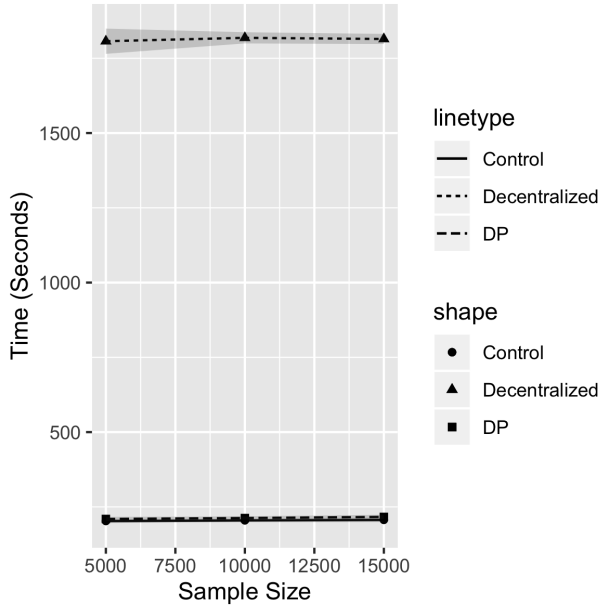

Figure 4.4.: Comparison of Runtimes for Control Case, Decentralized Scheme, and Differential Privacy. Shaded regions denote pointwise 95% confidence intervals.

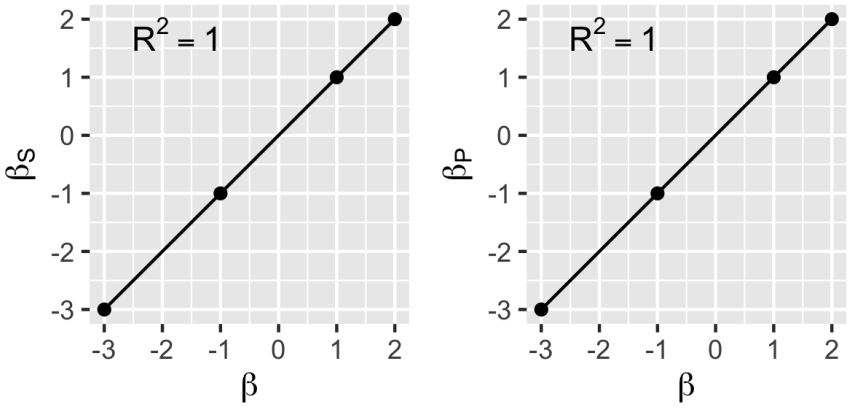

Figure 4.5.: Comparison of Beta Estimates without Encryption vs Estimate Obtained from the Centralized Scheme β_S and Estimate Obtained from the Decentralized Scheme β_P. We see that no accuracy is sacrificed in the encryption/decryption process.

4.7 Application to a Real Data Set

The MIMIC III database [96] is a massive, freely-available database comprising of de-identified medical data associated with over 46,520 patients who stayed in critical care units

Table 4.3: Performance Analysis of Centralized and Decentralized Algorithms on Synthetic Data, $\tau = 0.7$

Algorithm	Control	Centralized	DP	Control	Decentralized	DP
Number of Samples	4,505,000	4,505,000	4,505,000	10,000	10,000	10,000
Number of Parameters	1,000	1,000	1,000	100	100	100
Total Runtime (s)	55.230	202.075	55.419	19.172	1843.944	25.673

of the Beth Israel Deaconess Medical Center between 2001 and 2012. Access to this dataset can be requested at https://mimic.physionet.org/. MIMIC-III includes features such as demographics, bedside vital sign measurements (1 data point per hour), laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality (both in and out of hospital). To illustrate our procedure, we perform distributed quantile regression using the following variables from the MIMIC III tables described below:

- Admissions: Ethnicity, insurance, religion, marital_status, admission_type, admission_location
- CPT Events: cost_center
- ICU Stays: first_careunit, last_careunit, los
- Patients: Gender, dob

Detailed descriptions for each of these variables are available at [6]. We designate the total admission time at the medical centers (los) as the response while the remaining variables are used as predictor variables. We also consider interaction effects between demographic related categorical variables (ethnicity and gender) and age to demonstrate the flexibility of quantile regression. That is,

$$Y_\tau = \beta_{0,\tau} + X_{age}\beta_{age,\tau} + X_{eth}\beta_{eth,\tau} + X_{gen}\beta_{gen,\tau} + X_{ins}\beta_{ins,\tau} + X_{admtype}\beta_{admtype,\tau}$$

$$+X_{admtime}\beta_{admtime,\tau} + X_{mar}\beta_{mar,\tau} + X_{cos}\beta_{cos,\tau} + X_{fcu}\beta_{fcu,\tau} + X_{lcu}\beta_{lcu,\tau} + \epsilon,$$

where $\mathbb{E}[\epsilon] = \mathbf{0}$ (no other distributional assumptions on ϵ), Y_τ denotes the τth quantile of the response Y, and the τ subscripts on the regression parameters indicate parameters for regression on the τth quantile. These models are fitted using our centralized scheme. To simulate 3 institutions collaborating to create these quantile regression models, we randomly partitioned the patients horizontally into 3 blocks by sampling without replacement. To avoid correlations between individual hospital visits we only consider unique patients, which reduced our sample size from 100,000 to 46,520. Categorical variables were one-hot encoded to allow use in regression, i.e. they were transformed from multicategorical variables . To reduce the number of one-hot encoded variables, ethnicities were combined into 5 natural categories: white, black, Hispanic/Latino, Asian, and other. The resulting quantile regression coefficient curve plotted is plotted in Figure 4.6 and quantile regression error recorded in Table 4.4. Figure 4.7 shows that no information is lost during the encryption, decryption, or secure aggregation just as in the simulation study. We compare our results with a previous neural network regression study [97], which aimed to classify MIMIC III cases into serious cases lasting longer than 5 days and less serious cases lasting less than 5 days. The classification accuracy of our median regression model here is 79.852%, which is higher than that of the neural network approach. Our method has the advantage of being much faster to train and a much more interpretable model that can be explained to researchers in other disciplines.

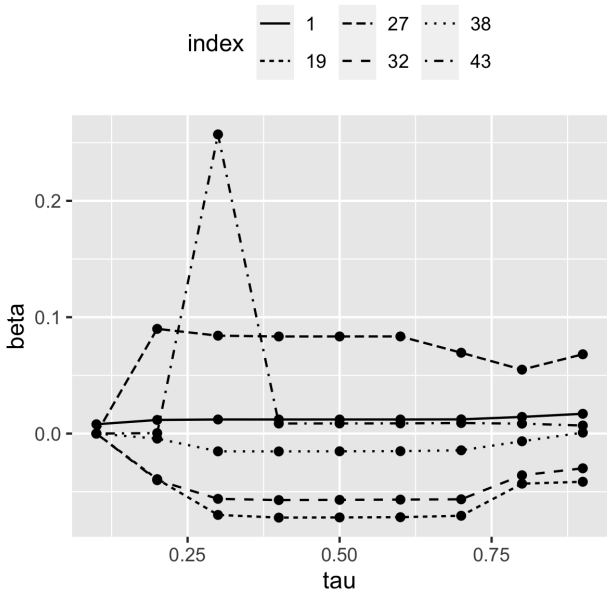

Figure 4.6.: Observation of variation in parameter estimates as quantile τ increases from 0.1 to 0.9

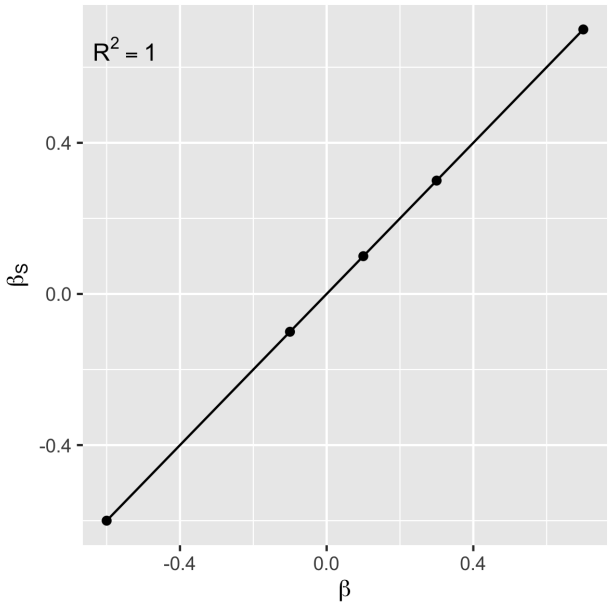

Figure 4.7.: Comparison of β Estimates without Encryption vs Estimate Obtained from Centralized Scheme (β_S) for MIMIC-III Data

4.8 Discussion

In this study, we combined two cryptographic methods with distributed ADMM to develop methods for performing distributed penalized quantile regression. We found that both

Table 4.4: Performance Analysis of Centralized Algorithm on Real Data

Quantile	.1	.2	.3	.4	.5	.6	.7	.8	.9
Error	0.46073	0.90944	1.33026	1.71759	2.07093	2.39027	2.67689	2.96761	3.25719

methods were able to maintain security of local data and estimates while preserving accuracy of the quantile regression estimator obtained by the QPADM algorithm with no privacy preservation. We also find that the same properties hold when applying the scheme to a real data set, making the scheme useful in a variety of real life applications. We hope that these methods may allow for expanded joint research between institutions while preserving the safety of all customers and institutions involved.

The aforementioned two schemes can be useful in a variety of scenarios where heteroskedastic data is common and individual data must be protected. An example of such a scenario is a genetic study in which the hospitals act as agents who collect data from patients. When the response variable of a genetic study is continuous, quantile regression is used to obtain a comprehensive examination of covariate effects on multiple quantile levels of the response [98]. Examples of such cases include time-varying coefficient models [99], GWAS [98], reference growth charts [100], etc. Many such studies depend on high-volume data sharing or sharing of local statistics, which can give away sensitive information about the subjects at an institution. Our scheme allows such analyses to be performed while still respecting this privacy constraint.

Our schemes also work well in the context of educational studies. In this case, schools would act as agents who collect data from students, whose data must again be protected. For these studies, the response variable is often a continuous quantity such as school income, average student grades, graduation rate, etc., and the results of such studies drive school

policies [101]. Through our schemes, universities can collaborate on studies to obtain inference for the general student population while preserving the privacy of student information and distributional information about the university's sensitive data.

Another major scheme that looks promising, particularly in the big data setting, is differential privacy. The algorithms presented in our work do not add extra noise to the data which means it sacrifices no optimality in the ADMM optimization. Our empirical results show that even though differential privacy trades data accuracy and utility for privacy, it is much faster to perform especially in the big data setting. Differential privacy can be advantageous in that there is no decoding step required, since all estimates are obtained by the perturbed data. Our preliminary empirical studies show that differential privacy adds a large amount of bias to the penalized quantile regression estimator, but it is currently unclear if bias correction is possible to produce a more accurate estimate without sacrificing privacy.

5. Conclusion

High dimensional statistics is a rapidly evolving field in which each new answer gives rise to even more questions. In this dissertation we considered two problems under the overarching theme of high dimensional statistics and examined the challenges faced when taking them from the traditional setting to the high-dimensional setting. Our first objective was to develop an effective screening approach for conducting high-dimensional mediation analysis. We described the linear structural equation model framework and key assumptions underlying causal mediation analysis, then proposed a novel screening procedure for high dimensional mediation analysis. This new procedure screens mediators based on the product of coefficients obtained from the treatment-mediator model and the mediator - response model in a similar way to how Sobel's test statistic is computed in traditional mediation analysis. After showing that this procedure has the sure screening property under our causal mediation analysis assumptions, we demonstrate its use as a screening step within a multi-stage mediation analysis procedure in a variety of synthetic datasets and a real data example.

Our second objective was to develop methods for performing distributed penalized quantile regression in a manner that preserves privacy of individual subject information. This problem was motivated by an example in which multiple agents want to perform a collaborative high dimensional quantile regression analysis with their shared data in such a way that no agent inside or outside the scheme learns anything of the true values of data transmitted by other agents. We first went over the QPADM algorithm, which is a version of the

alternating direction method of multipliers algorithm applied to penalized quantile regression. Using the parallelizability of QPADM, we then developed a centralized cryptosystem and a decentralized cryptosystem so that the agents in the study can perform the algorithm collaboratively in a secure manner. Our cryptosystems were shown in simulations and real data to be fast and lossless in the sense that models calculated by our scheme were the same as those calculated without any encryption.

The work presented in this dissertation has great potential to be taken in many research directions in the future. In the mediation analysis setting, our marginal screening approach can be modified in similar ways to the other marginal screening methods that inspired it. For example it can be extended from the LSEM framework to a generalized LSEM, where the treatment - mediator model and mediator - response model are generalized linear models rather than linear models. Just as sure independence screening has an iterated version [8] that repeatedly screens to further reduce dimension and a conditional version [49] that screens after conditioning on prior information, these same ideas can be applied to marginal Sobel screening to improve its versatility. The distributional properties of our marginal screener can also be used to create marginally valid inference as is sought after in the post-selection inference literature. Possible correlation structures between mediators can also be considered to improve screening.

In the distributed optimization and cryptography setting, our cryptosystem provides a blueprint for training other penalized models under similar privacy constraints. Using the parallelizability of ADMM or other optimization techniques allows researchers to distribute the work of an optimization problem, while homomorphic encryption algorithms allow the agents in the scheme to trade information in a secure manner while still allowing mathemat-

ical operations to be performed on encrypted data. Differential privacy [95] is another more recent form of encryption that may work in a similar cryptosystem with the advantage of improved speed, although more research needs to be done on how to maintain accuracy after injecting noise into the data.

The proposed methods in this dissertation provide some improvements for methods in high dimensional statistics. We hope that these methods not only improve the ability to make better scientific discoveries and policy decisions, but also inspire others who run into problems when working with high dimensional problems. We further hope that this dissertation provides valuable techniques to the field of high dimensional statistics and science as a whole.

References

[1] Jiacong Du, Xiang Zhou, Wei Hao, Yongmei Liu, Jennifer A Smith, and Bhramar Mukherjee. Methods for large-scale single mediator hypothesis testing: Possible choices and comparisons. *arXiv preprint arXiv:2203.13293*, 2022.

[2] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

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

[4] *CARDIA: Coronary Artery Risk Development in Young Adults.* `https://www.cardia.dopm.uab.edu/`. Accessed: 2022-07-13.

[5] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

[6] *Overview of the MIMIC-III data.* `https://mimic.mit.edu/docs/iii/tables/`. Accessed: 2022-07-14.

[7] Jianqing Fan. Variable screening in high-dimensional feature space. In *Proceedings of the 4th International Congress of Chinese Mathematicians*, volume 2, pages 735–747, 2007.

[8] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[9] Xiangyu Wang and Chenlei Leng. High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):589–611, 2016.

[10] Hansheng Wang. Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association*, 104(488):1512–1524, 2009.

[11] Haeran Cho and Piotr Fryzlewicz. High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):593–622, 2012.

[12] Donald E Hilt and Donald W Seegrist. *Ridge, a Computer Program for Calculating Ridge Regression Estimates*. U.S. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station, Upper Darby, PA, 1977.

[13] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[14] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

[15] John Wilder Tukey. The problem of multiple comparisons. Unpublished manuscript. In *The Collected Works of John W. Tukey: Multiple Comparisons 1948-1983*, volume 8, pages 1–300. Chapman and Hall, New York, NY, 1953.

[16] Carlo Bonferroni. Teoria statistica delle classi e calcolo delle probabilita [statistical class theory and calculation of probability]. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[17] Henry Scheffé. *The Analysis of Variance*, volume 72. John Wiley & Sons, Hoboken, NJ, 1999. (Original work published in 1959).

[18] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[19] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[20] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.

[21] Charles M Judd and David A Kenny. Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, 5(5):602–619, 1981.

[22] Lawrence R James and Jeanne M Brett. Mediators, moderators, and tests for mediation. *Journal of Applied Psychology*, 69(2):307–321, 1984.

[23] Reuben M Baron and David A Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.

[24] Sewall Wright. The theory of path coefficients: a reply to Niles's criticism. *Genetics*, 8(3):239–255, 1923.

[25] Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

[26] Douglas Gunzler, Tian Chen, Pan Wu, and Hui Zhang. Introduction to mediation analysis with structural equation modeling. *Shanghai Archives of Psychiatry*, 25(6):390–394, 2013.

[27] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13:290–312, 1982.

[28] Zhonghua Liu, Jincheng Shen, Richard Barfield, Joel Schwartz, Andrea A Baccarelli, and Xihong Lin. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, 117(537):67–81, 2021.

[29] David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104, 2002.

[30] Jeffrey M Albert and Suchitra Nelson. Generalized causal mediation analysis. *Biometrics*, 67(3):1028–1038, 2011.

[31] Judea Pearl. Interpretation and identification of causal mediation. *Psychological Methods*, 19(4):459–504, 2014.

[32] Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 11(1):1–12, 1943.

[33] Otis Dudley Duncan. *Introduction to Structural Equation Models*. Elsevier, Amsterdam, NL, 1973.

[34] John Fox. Effect analysis in structural equation models: Extensions and simplified methods of computation. *Sociological Methods & Research*, 9(1):3–28, 1980.

[35] Kenneth A Bollen. *Structural Equations with Latent Variables*. John Wiley & Sons, Hoboken, NJ, 1989.

[36] Judea Pearl. Direct and indirect effects. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 411–420, 2001.

[37] Judea Pearl. The causal mediation formula — a guide to the assessment of pathways and mechanisms. *Prevention Science*, 13(4):426–436, 2012.

[38] Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2(1):95–115, 2014.

[39] Rhian M Daniel, Bianca L De Stavola, SN Cousens, and Stijn Vansteelandt. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14, 2015.

[40] Stijn Vansteelandt and Rhian M Daniel. Interventional effects for mediation analysis with multiple mediators. *Epidemiology*, 28(2):258–265, 2017.

[41] Johan Steen, Tom Loeys, Beatrijs Moerkerke, and Stijn Vansteelandt. Flexible mediation analysis with multiple mediators. *American Journal of Epidemiology*, 186(2):184–193, 2017.

[42] Kun Zhang, Mingming Gong, Joseph D Ramsey, Kayhan Batmanghelich, Peter Spirtes, and Clark Glymour. Causal discovery with linear non-Gaussian models under measurement error: Structural identifiability results. In *Association for Uncertainty in Artificial Intelligence*, pages 1063–1072, 2018.

[43] David P MacKinnon. *Introduction to Statistical Mediation Analysis*. Routledge, Oxfordshire, UK, 2012.

[44] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference, and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

[45] Zhongheng Zhang, Cheng Zheng, Chanmin Kim, Sven Van Poucke, Su Lin, and Peng Lan. Causal mediation analysis in the context of clinical research. *Annals of Translational Medicine*, 4(21):425–435, 2016.

[46] David P MacKinnon. Contrasts in multiple mediator models. In Jennifer S. Rose, Laurie Chassin, Clark C. Presson, and Steven J. Sherman, editors, *Multivariate Applications in Substance Use Research: New Methods for New Questions*, chapter 5, pages 141–160. Psychology Press, London, UK, 2000.

[47] Jeremiah Jones, Ashkan Ertefaie, and Robert L Strawderman. Causal mediation analysis: Selection with asymptotically valid inference. *arXiv preprint arXiv:2110.06127*, 2021.

[48] Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

[49] Emre Barut, Jianqing Fan, and Anneleen Verhasselt. Conditional sure independence screening. *Journal of the American Statistical Association*, 111(515):1266–1277, 2016.

[50] Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[51] Tyler J VanderWeele and Stijn Vansteelandt. Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4):457–468, 2009.

[52] Linda Valeri and Tyler J VanderWeele. 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, 2013.

[53] Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50(1):1–25, 1982.

[54] Michael E Sobel. Some new results on indirect effects and their standard errors in covariance structure models. *Sociological Methodology*, 16:159–186, 1986.

[55] Aad W Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

[56] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, New York, NY, 1991.

[57] Pascal Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

[58] Sara van de Geer. M-estimation using penalties or sieves. *Journal of Statistical Planning and Inference*, 108(1-2):55–69, 2002.

[59] Saralees Nadarajah and Tibor K Pogány. On the distribution of the product of correlated normal random variables. *Comptes Rendus Mathematique*, 354(2):201–204, 2016.

[60] Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011.

[61] Sihai Dave Zhao and Yi Li. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411, 2012.

[62] Jiashun Jin and T Tony Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007.

[63] Xu Guo, Haojie Ren, Changliang Zou, and Runze Li. Threshold selection in feature screening for error rate control. *Journal of the American Statistical Association*, Advance online publication:1–13, 2022.

[64] Peri H Tate and Adrian P Bird. Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development*, 3(2):226–231, 1993.

[65] Gary D Friedman, Gary R Cutter, Richard P Donahue, Glenn H Hughes, Stephen B Hulley, David R Jacobs Jr, Kiang Liu, and Peter J Savage. CARDIA: Study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology*, 41(11):1105–1116, 1988.

[66] Chamila Perera, Haixiang Zhang, Yinan Zheng, Lifang Hou, Annie Qu, Cheng Zheng, Ke Xie, and Lei Liu. HIMA2.0: High-dimensional mediation analysis and its application in genome-wide DNA methylation data. *BMC Bioinformatics*, 23(296), 2022.

[67] Yen-Tsung Huang. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. *The Annals of Applied Statistics*, 12(3):1535–1557, 2018.

[68] James Y Dai, Janet L Stanford, and Michael LeBlanc. A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association*, 117(537):198–213, 2020.

[69] Jennifer Beane, Paola Sebastiani, Gang Liu, Jerome S Brody, Marc E Lenburg, and Avrum Spira. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biology*, 8(9):1–17, 2007.

[70] Adam C Gower, Katrina Steiling, John F Brothers, Marc E Lenburg, and Avrum Spira. Transcriptomic studies of the airway field of injury associated with smoking-related lung disease. *Proceedings of the American Thoracic Society*, 8(2):173–179, 2011.

[71] John A Lees, Bart Ferwerda, Philip HC Kremer, Nicole E Wheeler, Mercedes Valls Serón, Nicholas J Croucher, Rebecca A Gladstone, Hester J Bootsma, Nynke Y Rots, Alienke J Wijmega-Monsuur, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nature Communications*, 10(1):1–14, 2019.

[72] Basavaraj Vastrad, Chanabasayya Vastrad, and Anandkumar Tengli. Bioinformatics analyses of significant genes, related pathways, and candidate diagnostic biomarkers and molecular targets in SARS-CoV-2/COVID-19. *Gene Reports*, 21:100956, 2020.

[73] Alexander G Petrenko, Beate Ullrich, Markus Missler, Valery Krasnoperov, Thomas W Rosahl, and Thomas C Südhof. Structure and evolution of neurexophilin. *Journal of Neuroscience*, 16(14):4360–4369, 1996.

[74] Markus Missler and Thomas C Südhof. Neurexophilins form a conserved family of neuropeptide-like glycoproteins. *Journal of Neuroscience*, 18(10):3630–3638, 1998.

[75] Vassilios Beglopoulos, Monique Montag-Sallaz, Astrid Rohlmann, Kerstin Piechotta, Mohiuddin Ahmad, Dirk Montag, and Markus Missler. Neurexophilin 3 is highly localized in cortical and cerebellar regions and is functionally important for sensorimotor gating and motor coordination. *Molecular and Cellular Biology*, 25(16):7278–7288, 2005.

[76] Roger Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, UK, 2005.

[77] Emma Graham-Harrison and Carole Cadwalladr. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*, March 2018.

[78] Lee Mathews. Equifax data breach impacts 143 million Americans. *Forbes*, September 2017.

[79] Homer et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLOS Genetics*, 2008.

[80] U.S. Department of Health & Human Services. *Health Information Privacy*. `https://www.hhs.gov/hipaa/index.html`, 2019.

[81] US Department of Education. *Family Educational Rights and Privacy Act (FERPA)*. `https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html`, 2018.

[82] John Leslie King. Centralized versus decentralized computing: Organizational considerations and management options. *ACM Computing Surveys (CSUR)*, 15(4):319–349, 1983.

[83] International Business Machines Corporation. *z/OS Basic Skills Information Center: Mainframe Concepts*. Number 0-521-83084-2. International Business Machines Corporation, Armonk, New York, 2008.

[84] Neeta Varma and Savita Dawar. Digital transformation in the Indian government. *Communications of the ACM*, 62(1):50–53, 2019.

[85] Guy Zyskind, Oz Nathan, et al. Decentralizing privacy: Using blockchain to protect personal data. In *2015 IEEE Security and Privacy Workshops*, pages 180–184. IEEE, 2015.

[86] Chunlei Zhang, Muaz Ahmad, and Yongqiang Wang. ADMM based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, 14(3):565–580, 2018.

[87] Oded Goldreich. *Foundations of Cryptography Volume II: Basic Applications*. Cambridge University Press, Cambridge, UK, 2004.

[88] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2011.

[89] Liqun Yu, Nan Lin, and Lan Wang. A parallel algorithm for large-scale nonconvex penalized quantile regression. *Journal of Computational and Graphical Statistics*, 26(4):935–939, 2017.

[90] Douglas Stinson and Maura Paterson. *Cryptography: Theory and Practice*. Cambridge University Press, Cambridge, UK, 2018.

[91] P Ram Mohan Rao, S Murali Krishna, and AP Siva Kumar. Privacy preservation techniques in big data analytics: a survey. *Journal of Big Data*, 5(33):1–12, 2018.

[92] Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

[93] Ed Dawson and Diane Donovan. The breadth of Shamir's secret-sharing scheme. *Computers & Security*, 13(1):69–78, 1994.

[94] Pascal Paillier. Public-key cryptosystems based on composite eegree residuosity classes. In *International Conference on the Theory and Applications of Cryptographic Techniques*, pages 223–238. Springer, 1999.

[95] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pages 1–12. Springer-Verlag, 2006.

[96] T.J. Pollard and A.E.W. Johnson. The MIMIC code repository: Enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2017.

[97] Thanos Gentimis, Alnaser Ala'J, Alex Durante, Kyle Cook, and Robert Steele. Predicting hospital length of stay using neural networks on MIMIC III data. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 1194–1201. IEEE, 2017.

[98] Ying Wei, Xiaoyu Song, Mengling Liu, Iuliana Ionita-Laza, and Joan Reibman. Quantile regression in the secondary analysis of case-control data. *Journal of the American Statistical Association*, 111(513):344–354, 2016.

[99] Ying Wei, Xinran Ma, Xinhua Liu, and Mary Beth Terry. Using time-varying quantile regression approaches to model the influence of prenatal and infant exposures on childhood growth. *Biostatistics & Epidemiology*, 1(1):133–147.

[100] Ying Wei, Anneli Pere, Roger Koenker, and Xuming He. Quantile regression methods for reference growth charts. *Statistics in Medicine*, 25:1369–1382, 2006.

[101] Beatrice Schindler Rangvid. School composition effects in Denmark: quantile regression evidence from PISA 2000. In Christian Dustmann, Bernd Fitzenberger, and Stephen Manchin, editors, *The Economics of Education and Training*, pages 179–208. Springer, 2008.