High-dimensional Causal Mediation Analysis Based on Partial Linear Structural Equation Models *

Xizhen Cai[a], Yeying Zhu[b], Yuan Huang[c,*], Debashis Ghosh[d]

[a] *Department of Mathematics and Statistics, Williams College, Williamstown, MA 01267, United States*
[b] *Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada*
[c] *Department of Biostatistics, Yale School of Public Health, New Haven, CT 06511, United States*
[d] *Department of Biostatistics & Informatics, Colorado School of Public Health, Aurora, CO 80045, United States*

Abstract

Causal mediation analysis has become popular in recent years, in which researchers not only aim to estimate the causal effect of a treatment, but also try to understand how the treatment affects the outcome through intermediate variables, namely mediators. In this paper, a set of generalized structural equations to estimate the direct and indirect effects for mediation analysis is proposed when the number of mediators is of high-dimensionality. Specifically, a two-step procedure is considered where the penalization framework can be adopted to perform variable selection. A partial linear model is used to account for a nonlinear relationship among pre-treatment confounder (confounders) and the response variable in each model, given that the interest is in estimating the coefficients for the treatment and the mediators in the structural models. The obtained estimators can be interpreted as causal effects without imposing the linear assumption on the model structure. The performance of Sobel's method in obtaining the standard error and confidence interval for the estimated joint indirect effect is also evaluated in simulation studies. Simulation results show a superior performance of our proposed method. The proposed method is applied to investigate how DNA methylation plays a role in the regulation of human stress reactivity impacted by childhood trauma.

Keywords: Adaptive LASSO; Causal inference; Confounding; High-dimensional mediators.

*Corresponding author: Yuan Huang. E-mail address: yuan.huang@yale.edu. Postal address: Ste 815, 60 College Street, New Haven, CT 06520, United States.

*The online version of this article contains a supplementary file.

1. Introduction

Mediation analysis is often used to study how a treatment variable relates to the outcome variable through an intermediate variable, namely, a mediator (see Figure 1 (a) for a visual illustration). The most commonly used approach for mediation analysis is the Baron and Kenny's four-step linear structural equation modeling (LSEM) approach (Baron and Kenny, 1986; Judd and Kenny, 1981). In LSEM, the total effect of the treatment on the outcome is decomposed into two kinds of effects: the direct effect and the indirect effect, where the latter refers to the effect of the treatment on the outcome that goes through the mediator. Based on the framework of counterfactuals, modern mediation approaches interpret the mediation effect as natural effects, controlled effects, and principal stratification effects, all of which can be interpreted causally because they are based on the contrast among the potential outcomes within the same subject. Such approaches include Angrist et al. (1996), who apply two-stage least squares to estimate principal stratification effects among compliers; Ten Have et al. (2007), who propose rank preserving models (RPM) for controlled effects, Gallop et al. (2009), who focus on Bayesian approaches for principal stratification effects, and Imai et al. (2010a,b), who propose nonparametric identification of natural direct and indirect effects. A comprehensive review of these causal approaches can be found in Coffman et al. (2016).

In particular, Imai et al. (2010a) have shown that the estimated direct and indirect effects by LSEM are causal effects under certain assumptions. A strong assumption about the traditional LSEM approach is that the assumed linear models are correct. However, the true relationship among the three set of variables (treatment, outcome, mediators), as well as the pre-treatment confounders are unknown (we will defer the formal definition of pre-treatment confounders to Section 2.1). With the existence of a large number of pre-treatment confounders, this assumption is unlikely to be true (Keele and Keele, 2008; Imai et al., 2010a). In this article, we relax the linear assumption by imposing a set of more flexible models, i.e., partial linear models, where the pre-treatment confounders are regarded as nuisance in the sense that nonparametric smoothing methods can be fitted to capture the relationship between the pre-treatment confounders and the outcome variables in each model. A similar type of model framework has been considered in Hines et al. (2021).

Very often, the causal effect of a treatment on the outcome can be carried out by multiple indirect pathways and recent development in mediation analysis has focused on estimating the direct and indirect effects with the existence of multiple mediators, e.g., Imai and Yamamoto (2013); VanderWeele and Vansteelandt (2014); Huang and Pan (2016); Aung et al. (2020); Jirolon et al. (2020). In many applications, the set of mediators could be high-dimensional or even ultra high-dimensional. For example, researchers might be interested in investigating how an exposure variable could affect the disease outcome through DNA methylation markers. In this case, the number of markers could even be larger than the sample size. Zhang et al. (2016) propose a penalized estimating and inference procedure based on linear models when the mediators are high-dimensional methylation markers, in which the authors employ sure independence screening (SIS) and minimax concave penalty (MCP) for variable selection. As a data application, the authors study how DNA methylations mediate the association between smoking and lung cancer. Chén et al. (2018) consider transforming the mediators into a few of orthogonal components by linear combinations. Gao et al. (2019) propose a sparse mediation model and a high-dimensional testing procedure based on SIS and LASSO for correlated multiple mediators. The authors apply the proposed method to study how DNA methylations mediate the association between alcohol consumption and epithelial ovarian cancer status. Guo et al. (2021) study a statistical inference procedure in the high-dimensional linear mediation models. They propose a new F-type test for the direct and indirect effects and also develop its theoretical properties. Luo et al. (2020) extend the methodologies to a survival outcome for settings with high-dimensional mediators.

However, none of the above-mentioned methods explicitly consider the possible confounders in the study. In observational studies, if the purpose is to draw causal conclusions, not accounting for the pre-treatment confounders will lead to biased estimators of causal direct and indirect effects. Under high-dimensional settings, performance of variable selection can also be affected. In this paper, we propose a two-step procedure to select the important mediators and to estimate the direct and indirect effects simultaneously with the confounders taken into account in the models.

The paper will proceed as follows. In Section 2, we review the traditional linear structural equation modeling approach and the assumptions needed for estimating causal direct and indirect effects. In Section 3, we propose a set of partial linear models that allow flexible modeling for the pre-treatment confounders. We then propose variable selection procedures when the set of mediators is high-dimensional. Simulation studies are conducted to compare the proposed method with alternative approaches in Section 4 and an application to an epigenetic study is provided in Section 5.

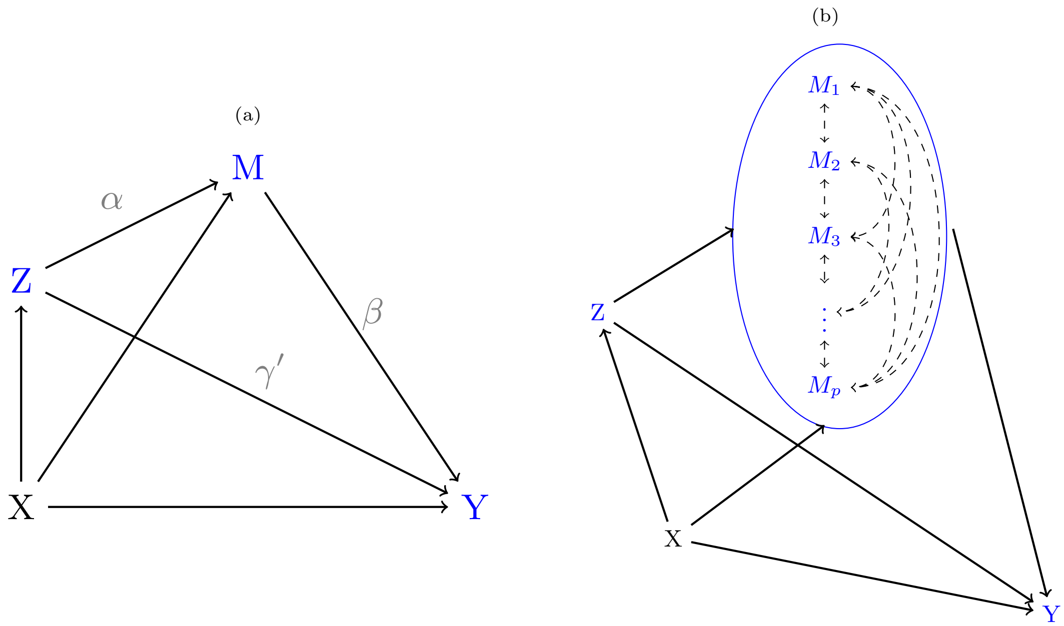

Figure 1: (a) The treatment Z, outcome Y, pre-treatment confounders X, and a single mediator M; (b) The treatment Z, outcome Y, pre-treatment confounders X, and multiple mediators $M_1, \ldots, M_p$. The dashed lines with double-headed arrows on plot (b) represent the correlations between pairs of mediators.

2. Traditional Linear Structural Equation Modeling Approach

In this section, we review the traditional mediation analysis method as well as the causal mediation analysis based on the potential outcomes framework in the context of a single mediator, followed by a description of connections between the two streams of methods.

2.1. Linear Structural Equation Modeling

The traditional approach for mediation analysis uses a linear structural equation modeling (LSEM) approach (Judd and Kenny, 1981; Baron and Kenny, 1986). Based on Baron and Kenny's four-step procedure, the total effect of the treatment can be decomposed into the direct and indirect effects, where the latter implies the amount of mediation. Denote the treatment variable as Z, the outcome variable as Y, the mediator variable as M, and the pre-treatment confounders as $\boldsymbol{X} = (X_1, \ldots, X_k)^\top$. A pre-treatment confounder is a variable that (1) jointly affects Z & Y; or (2) jointly affects Z & M; or (3) jointly affects M &Y; and (4) is not affected by Z (VanderWeele, 2016; Valente et al., 2017).

The mediation effect can be calculated through the following set of regression models:

$$Y = \gamma_0 + \gamma Z + \gamma_X^\top \mathbf{X} + \epsilon_1, \quad (1)$$

$$Y = \beta_0 + \beta M + \gamma' Z + \beta_X^\top \mathbf{X} + \epsilon_2, \quad (2)$$

$$M = \alpha_0 + \alpha Z + \alpha_X^\top \mathbf{X} + \epsilon_3, \quad (3)$$

where $\epsilon_i \sim N(0, \sigma_i^2), i = 1, 2, 3$. In the above models, the direct effect is defined as γ' (i.e., the effect of the treatment on the outcome when M is fixed) while the indirect effect is $\alpha\beta$ (i.e., the effect of the treatment on the outcome that goes through the mediator). As a result, examining whether a mediation effect exists is equivalent to testing $H_0 : \alpha\beta = 0$. Although Model (1) is not used to estimate α and β, if we substitute Model (3) into Model (2) and compare it with Model (1), we get $\gamma - \gamma' = \alpha\beta$. If all of the parameters are estimated by least squares, we also have $\hat{\gamma} - \hat{\gamma}' = \hat{\alpha}\hat{\beta}$ (MacKinnon et al., 1995).

If Z is randomized, the total effect of Z on Y (i.e., γ) and the effect of Z on M (i.e., α) may be interpreted causally. On the other hand, γ' and β do not readily admit a causal interpretation due to the fact that M is a post-treatment variable, which is a variable that can be affected by the treatment. In Section 2.2, we list the assumptions under which the direct and indirect effects based on LSEM have causal interpretations.

2.2. Assumptions

To make causal inference, we employ the potential outcome framework to define causal quantities (Rubin, 1974, 1978). We denote $Y_i(z, m)$ as the potential outcome if subject i was assigned to the treatment level z and the mediator level m. We further define $M_i(z')$ as the potential mediator value if subject i was assigned to the treatment level z'. Assuming the treatment Z is binary, the natural direct effect is defined as $NDE_z = E(Y(1, M(z)) - E(Y(0, M(z))$ and the natural indirect effect is defined as $NIE_z = E(Y(z, M(1)) - E(Y(z, M(0))$ for $z = 0, 1$ (Pearl, 2001). The total effect is $TE = E(Y(1, M(1)) - E(Y(0, M(0))$ and can be written as $NDE_1 + NIE_0$ or $NDE_0 + NIE_1$. Imai et al. (2010a) prove that the mediation effect $\alpha\beta$ can be interpreted as a causal effect and $\alpha\beta = NIE_1 = NIE_0$ if the following assumptions are satisfied:

- *Sequential Ignorability Assumption:*

$$\{Y(z,m), M(z')\} \perp Z | \mathbf{X},$$

$$Y(z,m) \perp M(z') | Z, \mathbf{X},$$

  where $Pr(Z = z|\mathbf{X} = \mathbf{x}) > 0$ and $Pr(M(z') = m|Z = z, \mathbf{X} = \mathbf{x}) > 0$ for all possible values of $\mathbf{x}$, m, and z.

- *Linearity Assumption:* the linear relationship between the predictors and the response variable is satisfied in Models (1) - (3).

- *No Interaction Assumption:* there is no interaction between Z & M.

The first part of the ignorability assumption says that among those who share the same values of baseline covariates, the treatment can be regarded as randomized. This is automatically true if the treatment is randomly assigned. However, in observational studies, where treatment is self-selected, this assumption may not be true and can hardly be tested based on the observed data. As pointed out by Imai et al. (2010b), a common strategy is to collect as many baseline covariates as possible so there is no unmeasured confounders. This argument also applies to the second part of the ignorability assumption; that is among those who have the same values of treatment and baseline covariates, the mediator can be regarded as randomized if there is no unmeasured confounders. Note that in Model (2), there is no treatment-mediator interaction, which is the focus of this study. However, in the literature, this no interaction assumption has been relaxed by Kraemer et al. (2002, 2008); Imai et al. (2010a,b). Since the sequential ignorability assumption is an untestable assumption, we will focus on relaxing the linearity assumption for estimating the natural direct and indirect effects in the following work.

3. Proposed Methodology

3.1. Relaxing the Linearity Assumption

The linearity assumption mentioned above is a strong assumption imposed on the form of the models. To relax this assumption, we propose a set of generalized linear structural equation models:

$$\begin{aligned} Y &= \gamma_0 + \gamma Z + g_1(\mathbf{X}) + \epsilon_1, & (4) \\ Y &= \beta_0 + \beta M + \gamma' Z + g_2(\mathbf{X}) + \epsilon_2, & (5) \\ M &= \alpha_0 + \alpha Z + g_3(\mathbf{X}) + \epsilon_3, & (6) \end{aligned}$$

where $g_j(\boldsymbol{X}), j = 1, 2, 3$ are assumed to be some unknown smoothed functions, and thus are nonparametric. The above models are partial linear models (Härdle et al., 2000) where the nonparametric components act like nuisance parameters. The main interest of partial linear models is to estimate the coefficients for the linear components. Under this model framework, the indirect effect is still $\gamma - \gamma' = \alpha\beta$, but we allow a more flexible relationship between the pre-treatment confounders and the mediator, as well as the outcome variable.

3.2. Multiple Mediators and Mediator Selection

In many applications, especially genetic studies, the causal pathway is often carried out by multiple mediators (see Figure 1 (b) for a visual illustration). Models (5)-(6) can be further extended to accommodate datasets with multiple candidate mediators. Denote the vector of candidate mediators as $\boldsymbol{M} = (M_1, M_2, \ldots, M_p)^\top$ where p is the number of candidate mediators. Model (4) remains unchanged, but Models (5) and (6) can be generalized to the following set of models,

$$\begin{aligned} Y &= \beta_0 + \boldsymbol{\beta}^\top \mathbf{M} + \gamma' Z + g_2(\mathbf{X}) + \epsilon_2, & (7) \\ M_j &= \alpha_{j0} + \alpha_j Z + g_{3j}(\mathbf{X}) + \epsilon_{3j}, \quad j = 1, 2, \ldots, p. & (8) \end{aligned}$$

Note that now each mediator is associated with one equation in (8). It is often not realistic to assume that the mediators are independent. For example, in genetic studies, the DNA methylation markers could be correlated with each other. Therefore, it is assumed that ϵ_{3j}'s are correlated, but they are independent with ϵ_2 in (7). Note here we do not consider the case where the mediators are causally related. The dashed lines with double-headed arrows between any pair of two mediators

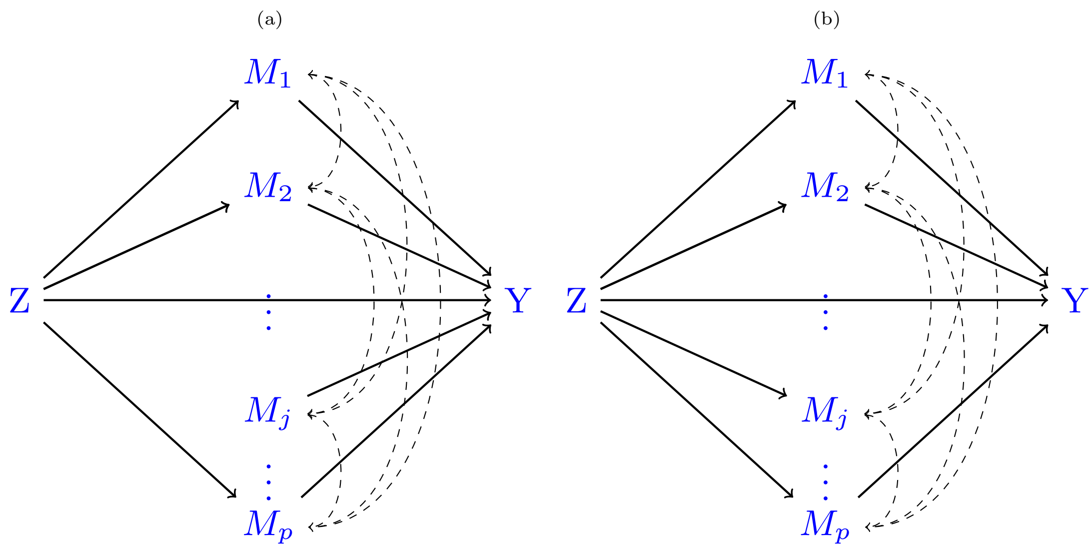

Figure 2: An illustration of an invalid mediator M_j (ignoring the pre-treatment confounders).In (a), $\alpha_j = 0$ but $\beta_j \neq 0$; in (b), $\beta_j = 0$ but $\alpha_j \neq 0$. In both scenarios, M_j is only associated with one of Z and Y, and thus M_j is not a valid mediator.

in Figure 1 (b) indicate that they are only correlated. In the case when one mediator causes the other, the direct and indirect effect have to be redefined in a more complex way. See Imai and Yamamoto (2013); Daniel et al. (2015) for examples. Without independence, the mediation effect for each individual mediator is generally unidentifiable. Thus the interest of such settings is usually in estimating the *joint mediation/indirect effect* of all the mediators. If we stack the coefficients α_j's in (8) as $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)^\top$, then the joint indirect effect is represented by $\boldsymbol{\alpha}^\top \boldsymbol{\beta}$.

When the dimension of the potential mediators is high or ultra-high, identifying a subset of true mediators is of scientific interest and also increases estimation precision of the estimated direct and indirect effects. For an individual mediator M_j, if $\alpha_j \beta_j = 0$, it adds no contribution to the joint indirect effect, and thus should be removed from the set of mediators. Next, we propose a two-step procedure for selecting the important mediators based on the fact that $\alpha_j \beta_j \neq 0$ if and only if $\alpha_j \neq 0$ and $\beta_j \neq 0$. In other words, if $\alpha_j = 0$ but $|\beta_j|$ is large, we do not consider M_j as a valid mediator. The same applies to the scenario when $\beta_j = 0$ but $|\alpha_j|$ is large. These two scenarios can be illustrated in Figure 2. In both scenarios, M_j is not a valid mediator because it is only associated with one of Z and Y.

To proceed, we first assume that the nonlinear components of Models (7) and (8) follow an additive model (Hastie and Tibshirani, 1990), i.e.,

$$g_2(\mathbf{X}) = \sum_{k=1}^{K} g_2^k(X_k), \qquad g_{3j}(\mathbf{X}) = \sum_{k=1}^{K} g_{3j}^k(X_k). \tag{9}$$

Then, we are going to approximate $g_2(\mathbf{X})$ and $g_{3j}(\mathbf{X})$ by smoothing techniques. There are various smoothing techniques we can employ including but not limited to regression spline, smoothing spline, local regression, etc (Fan and Gijbels, 2018). We include more details of the implementation in the simulation study section. Here we use regression spline for its easiness of implementation in a regression setting. Specifically, denote the normalized B-spline basis functions as $B_b(\cdot), b = 1, \ldots, B$, then the transformed expression of $g_2(\mathbf{X})$ and $g_{3j}(\mathbf{X})$ can be written as

$$g_2(\mathbf{X}) \approx \sum_{k=1}^{K}\sum_{b=1}^{B} \xi_b'^k B_b'(X_k), \qquad g_{3j}(\mathbf{X}) \approx \sum_{k=1}^{K}\sum_{b=1}^{B} \xi_{bj}^k B_{bj}(X_k) \tag{10}$$

where $\xi_b'^k$ and ξ_{bj}^k are the coefficients associated with the b-th basis function $B_b'(\cdot)$ and $B_{bj}(X_k)$, respectively.

Denote the subscript i as the index for the ith observation. We propose the following two-step procedure to get a parsimonious set of mediators and to estimate the direct effect and the joint indirect effect based on the selected mediators. First, we estimate and select the nonzero components of $\boldsymbol{\beta}$ by minimizing

$$\frac{1}{2n}\sum_{i=1}^{n}\left\{Y_i - \beta_0 - \boldsymbol{\beta}^\top \mathbf{M}_i - \gamma' Z_i - \sum_{k=1}^{K}\sum_{b=1}^{B} \xi_b'^k B_b'(X_{ik})\right\}^2 + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|), \tag{11}$$

where $p_{\lambda_j}(|\beta_j|)$ is the penalty function placed on the jth covariate with tuning parameter λ_j. It shrinks the magnitude of the estimated $\boldsymbol{\beta}$, so that some $\hat{\beta}_j$'s might be shrunken to zero. Here, we adopt the adaptive LASSO based on L_1 penalty for variable selection due to its oracle property (Zou, 2006). Other adopted penalty functions can also be implemented in this framework, including SCAD (Fan and Li, 2001), elastic net (Zou and Hastie, 2005), and the plain LASSO (Tibshirani, 1996), etc.

Then, denote $\mathcal{B}$ as the set of non-zero β_j's selected by (11) and $q = \dim(\mathcal{B})$, we now only consider a subset of the equations in (8). That is,

$$M_j \quad = \quad \alpha_{j0} + \alpha_j Z + g_{3j}(\mathbf{X}) + \epsilon_{3j}, \quad \text{for} \quad j \in \mathcal{B}.$$

To complete the second step, we further apply one of the following two methods to select a subset of the mediators in $\mathcal{B}$ by identifying nonzero α_j's, with $j \in \mathcal{B}$.

I. Consider all q models at the same time and apply a penalty function on α_j's, $j \in \mathcal{B}$. Specifically, we minimize the following objective function by stacking the corresponding datasets for the q penalization problems together,

$$\frac{1}{2nq} \sum_{i=1}^{n} \sum_{j \in \mathcal{B}} \left\{ M_{ij} - \alpha_{j0} - \alpha_j Z_i - \sum_{k=1}^{K} \sum_{b=1}^{B} \xi_{bj}^{k} B_{bj}(X_{ik}) \right\}^2 + \sum_{j \in \mathcal{B}} p_{\lambda_j}(|\alpha_j|). \tag{12}$$

II. Estimate q models separately, each as one partial linear regression; and obtain the p-values for α_j's. We then adjust the obtained p-values for testing $\alpha_j = 0, j \in \mathcal{B}$ by the Bonferroni correction and select the ones with an adjusted p-value smaller than a given threshold. This is to control the family-wise type I error in multiple testing.

After the above second-step selection, we denote the set of non-zero α_j's selected as $\mathcal{A}$, which is a subset of $\mathcal{B}$. As a result, $\{M_j, j \in \mathcal{A}\}$ is the final set of important mediators selected by our proposed two-step procedure.

To estimate the direct and joint indirect effect, we refit Models (7) & (8) with the selected mediators in $\mathcal{A}$. Specifically, the Models (8) is fitted as a multivariate regression model with a multivariate response vector of dimension $|\mathcal{A}|$, i.e., the size of $\mathcal{A}$. By conducting a multivariate regression, we account for the correlation in the selected mediators, and are able to obtain an estimated covariance matrix for the estimated $\hat{\alpha}_j$'s, which can further be used to estimate the standard error of the estimated joint indirect effect (see section 3.3). Without ambiguity, we still denote the estimated direct effect as $\hat{\gamma}'$ and the estimated indirect effect as $\hat{\boldsymbol{\alpha}}^\top \hat{\boldsymbol{\beta}}$ after model selection and refit.

3.3. Inference of the Indirect Effect and Standard Error Estimation

Under classic low-dimensional settings, the asymptotic distribution of $\hat{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\beta}}$ is a multivariate normal distribution (Sobel, 1982) obtained by the multivariate delta method with the variance estimator as

$$\widehat{Var}[\hat{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\beta}}] = \hat{\boldsymbol{\beta}}^\top Cov[\hat{\boldsymbol{\alpha}}]\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\alpha}}^\top Cov[\hat{\boldsymbol{\beta}}]\hat{\boldsymbol{\alpha}}, \tag{13}$$

where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ are the estimated coefficients in Models (7) & (8); $Cov(\hat{\boldsymbol{\alpha}})$ is the estimated variance-covariance matrix for $\hat{\boldsymbol{\alpha}}$ using multivariate linear regression for Model (8), and $Cov(\hat{\boldsymbol{\beta}})$ is the estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ in the simple partial linear regression for Model (7). Note, similar variance estimation for the joint indirect effect is also discussed and used in Bollen (1987, 1989); Preacher et al. (2007); Preacher and Hayes (2008). Denote $SE[\hat{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\beta}}]$ as the square root of the variance estimate, then a Wald test statistic for testing $H_0 : \boldsymbol{\alpha}^\top\boldsymbol{\beta} = 0, vs. H_a : \boldsymbol{\alpha}^\top\boldsymbol{\beta} \neq 0.$ can be defined as

$$\frac{\hat{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\beta}}}{SE[\hat{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\beta}}]},$$

which is treated as following the standard normal distribution asymptotically (Sobel, 1982).

In our settings with high-dimensionality, the application of the Sobel's method needs a careful choice of post-selection estimates for the parameters and their covariance matrices in Eqn (13). We examine the performance of Sobel's method using the estimates from the refitted models. That is, in Eqn (13), $\hat{\boldsymbol{\alpha}}$ ($\hat{\boldsymbol{\beta}}$) and $Cov(\hat{\boldsymbol{\alpha}})$ ($Cov(\hat{\boldsymbol{\beta}})$) are the estimated coefficients and covariance fitted using only the selected mediators in Models (7) & (8). We note this is a simple choice and the variability of variable selection is not fully accounted. In the simulation, we will evaluate its performance and assess the impact of variable selection.

4. Simulation Studies

4.1. Aim, Data Generation, and Estimand

To examine the performance of our proposed method in terms of variable selection, estimation and inference performance, we conduct several simulation studies.

We consider the case when Y is a continuous variable and the treatment variable Z is binary with an equal chance for 0 or 1. We generate two independent covariates X_1 and X_2 where $X_i \sim N(0, 1), i = 1, 2$. We further assume the mediator vector $\boldsymbol{M}$ is multivariate and follows a $MVN(\mu, \boldsymbol{\Sigma})$ distribution. Here, we consider an AR(1) correlation structure on $\boldsymbol{\Sigma} = (\sigma_{ij})$, i.e.

$\sigma_{ij}^2 = \rho^{|i-j|}$, where $\rho = 0.5$. In addition, we let

$$M_j = \alpha_{0j} + \alpha_j Z + g_{3j}^1(X_1) + g_{3j}^2(X_2) + \epsilon, \quad j = 1, 2, \ldots, p, \tag{14}$$

$$Y = \beta_0 + \gamma' Z + \boldsymbol{\beta}^\top \boldsymbol{M} + g_2^1(X_1) + g_2^2(X_2) + \epsilon, \tag{15}$$

where $\alpha_{0j} = 2, j = 1, 2, \ldots p, \beta_0 = 2, \gamma' = 1$, and $\epsilon \sim N(0, 1)$; the other parameters $\alpha_j, j = 1, \ldots, p$ and $\boldsymbol{\beta}$ are considered in the following two settings:

A. $\boldsymbol{\alpha}_{1:p} = (\underbrace{1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2}_{\text{10 nonzeros}}, 0, \ldots, 0)^\top$,
$\boldsymbol{\beta} = (\underbrace{1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2}_{\text{20 nonzeros}}, 0, \ldots, 0)^\top$.

B. $\boldsymbol{\alpha}_{1:p} = (\underbrace{1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2}_{\text{20 nonzeros}}, 0, \ldots, 0)^\top$,
$\boldsymbol{\beta} = (\underbrace{1, 0.8, 0.6, 0.4, 0.2, 1, 0.8, 0.6, 0.4, 0.2}_{\text{10 nonzeros}}, 0, \ldots, 0)^\top$.

In setting A, the number of significant mediators in Model (15) is double of that in Model (14); and vice versa in setting B. Since our two-step procedure selects β's first and then select α's, by creating the second setting, we would like to investigate the performance of the two-step procedure when the true signals in $\boldsymbol{\beta}$ is a subset of that in $\boldsymbol{\alpha}$.

Furthermore, we consider different types of the relationship between the covariates X_1, X_2, $\boldsymbol{M}$, and Y, as described in Scenarios (I)-(III) below. In particular, Scenarios (I)&(II) are based on a set of nonlinear models, and Scenario (III) is based on a set of linear models.

- Scenario I: $g_{3j}^1(X) = g_{3j}^2(X) = X + 0.5X^2$ and $g_2^1(X) = 0.5X + 0.25X^2$, $g_1^2(X) = 0.1X^2$.
- Scenario II: $g_{3j}^1(X) = sin(2X)$, $g_{3j}^2(X) = cos(X)$ and $g_2^1(X) = cos(X)$, $g_2^2(X) = sin(2X)$.
- Scenario III: $g_{3j}^1(X) = 0.5X$, $g_{3j}^2(X) = X$ and $g_2^1(X) = 0.5X$, $g_2^2(X) = 0.25X$.

We consider four settings for the sample size n and the number of mediators p: $n = 100, p = 500$; $n = 300, p = 500$; $n = 1000, p = 500$; and $n = 300, p = 3000$. Based on the simulation setup, the true direct effect is 1; the joint indirect effect is 1 and the total effect is 4.4.

4.2. Methods and Performance Measures

To estimate the direct and the indirect effect, we implement two proposed approaches based on partial linear models: the one using Bonferroni correction for variable selection in the second step (PLSEM$_B$), and the one using adaptive LASSO for variable selection (PLSEM$_{AL}$). For comparison, we implement two approaches based on the linear structural equation models: the one using adaptive LASSO for variable selection following our two-step procedure (LSEM$_{AL}$) and the joint significance testing approach proposed by Zhang et al. (2016) (HIMA). We use the `bs` function in the R package `spline` to fit the regression spline with degrees of freedom 5. This is equivalent to using a cubic spline with two knots, where the locations of knots are determined by the (33.3%. 66.6%) percentiles of the data. All penalization methods are implemented with R package `glmnet`. The weights for adaptive LASSO are calculated using the inverse of the absolute values of the ridge regression estimators with the tuning parameter selected from 20 equally spaced grid on [0, 1]. The tuning parameter for adaptive LASSO is selected by the function `cv.glmnet`. The HIMA procedure is implemented by the R package `HIMA` available on the github page[2].

For each simulation setting and each model scenario, we report the mediator selection results and the estimation performance. For the variable selection results, we report the true positive (TP) and false positive (FP) number of mediators selected for each approach. For a fair comparison, we also report the corresponding rates. For the estimation results, we evaluate the bias, the empirical standard error, and the square root of mean squared error (RMSE). Additionally, for statistical inference purpose, we also calculate the standard error of the estimated indirect effect using the method described in Section 3.3. The coverage rate of a 95% confidence interval is also evaluated using this asymptotic standard error.

We repeat the simulation for 500 times for each simulation setting. The variable selection results for setting A are summarized in Table 1, and that for setting B are in Table 2. Figure 3 & 4 give the results for the bias and variability of both the direct and joint indirect effect, respectively, with detailed numerical results of estimation provided in the supplementary file. The inference results are displayed in Table 3.

[2] `https://github.com/YinanZheng/HIMA/`

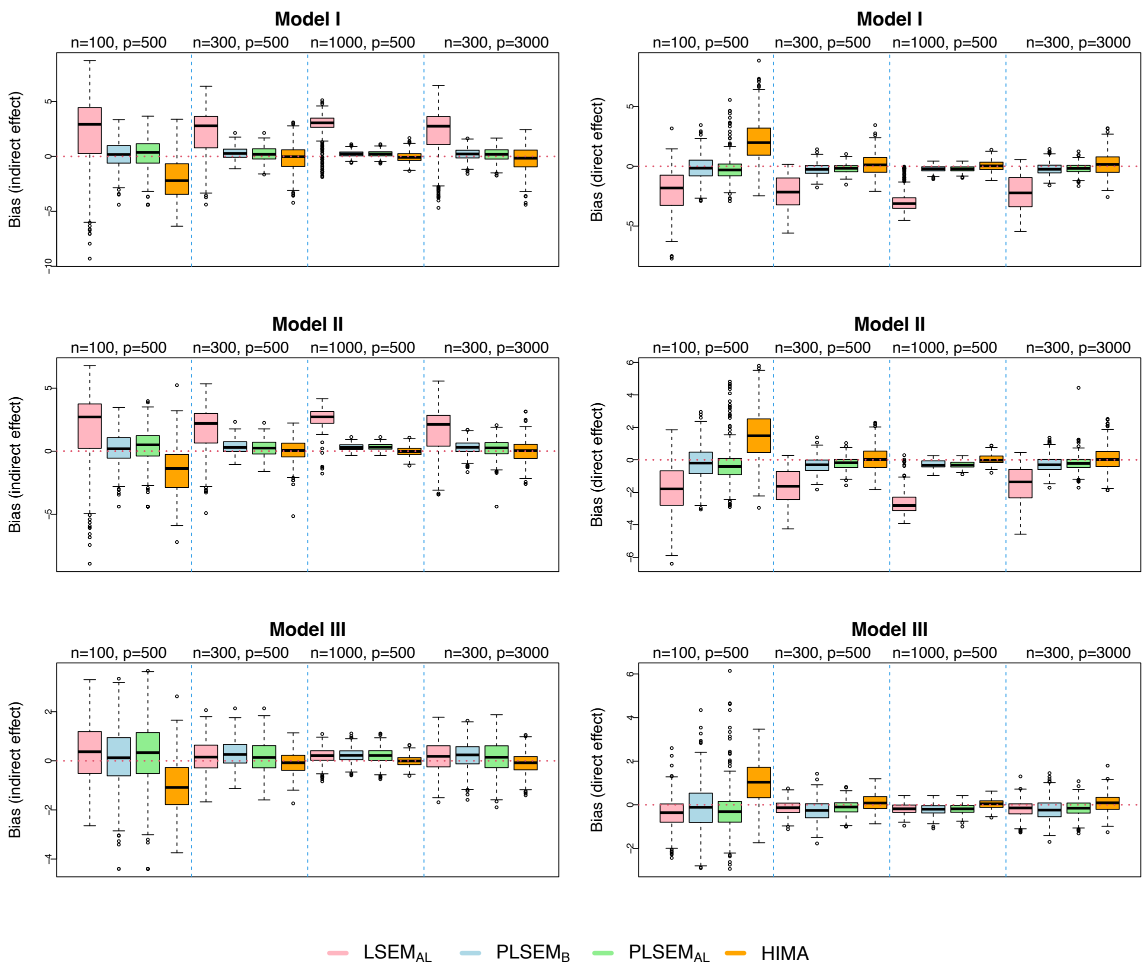

Figure 3: Visualization of estimated bias for indirect effects (left) and direct effects (right) in the setting A of simulated studies. Results are based on 500 replications. Each row is corresponding to a particular model, with models I, II, and III from the top to bottom. Each segmentation in one graph (separated by the vertical blue dashed lines) presents the comparisons among the four methods (colored legend on the bottom), and there are four combinations of sample sizes and dimensions of the mediators in each setting.

4.3. Results

In terms of variable selection, the proposed PLSEM$_{AL}$ approach based on adaptive LASSO has the best performance overall by yielding a high TP rate and a low FP rate. The proposed PLSEM$_B$ approach based on Bonferroni correction is too conservative by yielding relatively low TP rate across all scenarios; the LSEM$_{AL}$ approach has the highest TP rate but also the highest FP

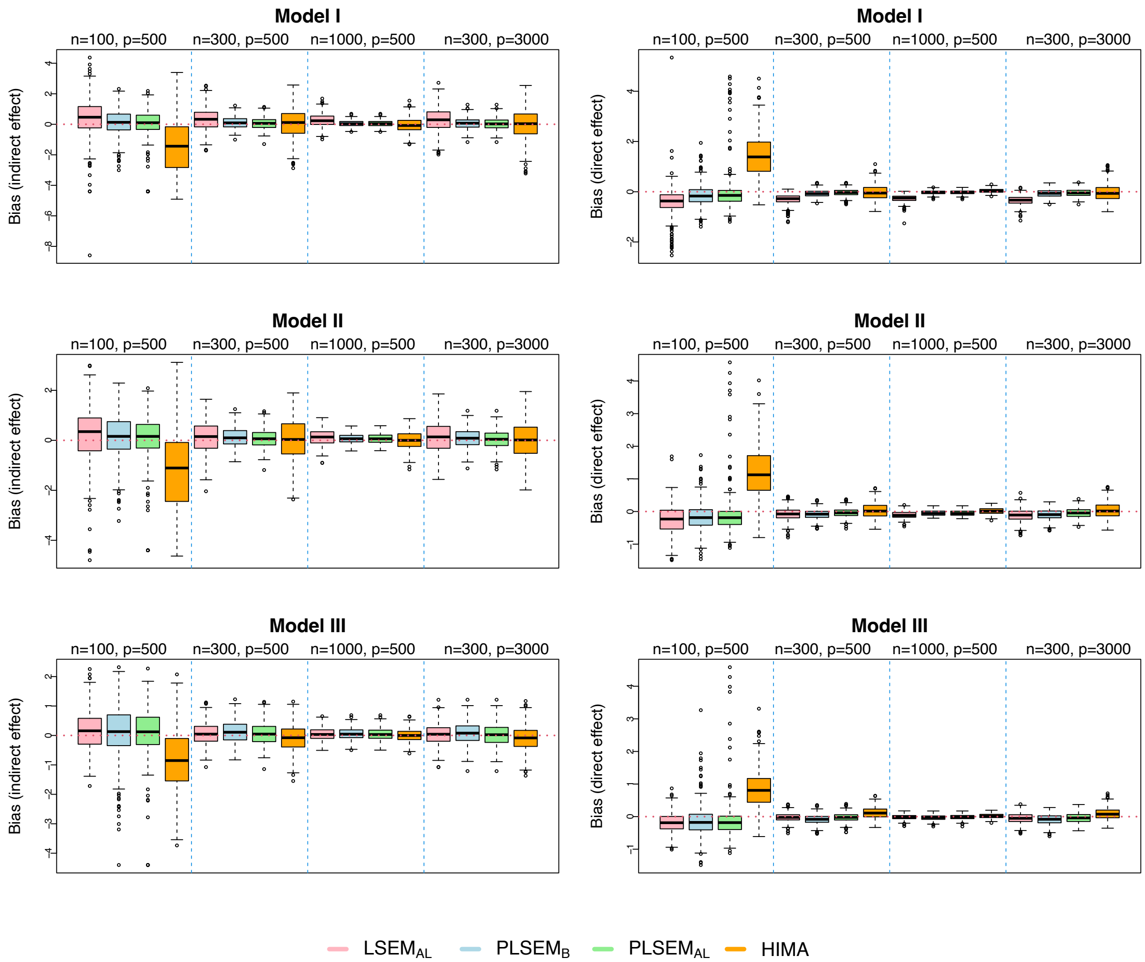

Figure 4: Visualization of estimated bias for indirect effects (left) and direct effects (right) in the setting B of simulated studies. Results are based on 500 replications. Each row is corresponding to a particular model, with models I, II, and III from the top to bottom. Each segmentation in one graph (separated by the vertical blue dashed lines) presents the comparisons among the four methods (colored legend on the bottom), and there are four combinations of sample sizes and dimensions of the mediators in each setting.

rate in almost all the scenarios. On the other hand, the HIMA approach yields the lowest TP rate but also the lowest FP rate.

In terms of bias and variance, in Scenario III where the true underlying models are linear, the bias and RMSE for LSEM$_{AL}$ and HIMA are the smallest as these two approaches are built upon the linear structural equation models. In Scenarios I & II, where the true underlying models are nonlinear, the methods based on partial linear models yield the smaller standard errors and RMSE values across different settings. This indicates that by not assuming a fully parametric model, we eliminate model misspecification for the covariates and can eventually improve the estimation of the direct and indirect effects when the true model is nonlinear. In addition, although the TP rate for HIMA is the lowest, the bias is the smallest in most of the scenarios. We found this is due to the fact that HIMA has difficulty in picking up the weak signals, which may not impact the estimation results in the same magnitude. To verify this, we consider another simulation scenario in supplementary file when all the signals are weak. As can be found in Table S6 & S7 of the supplementary file, the bias of the proposed approaches are much lower than HIMA in this case.

Regarding inference, from Table 3, we find that when $n = 100, p = 500$, the theoretical standard error is much smaller than the empirical standard error due to the small sample size. In this case, we observe that the discrepancy gets much smaller as the sample size increases (see results for $n = 1000, p = 500$). For the proposed $PLSEM_B$, the theoretical standard error is close to the empirical standard error when $n = 300, p = 500$ and $n = 300, p = 3000$ and for $PLSEM_{AL}$, they are close when $n = 1000, p = 500$, indicating that the formula for the theoretical standard error described in Section 3.3 performs relatively well in these specific scenarios. In addition, when the biases are small, such as most of the scenarios in Setting B, the coverage rates of the 95% confidence interval calculated from the standard error formula described in Section 3.3 are close to the nominal level. The above observations suggest that the performance can depend on the settings and a larger sample size can be helpful in improving the performance. In general, the performance of the standard error formula works better in Setting B, compared to that in Setting A. As we indicate in the discussions, further investigation of alternative methods for characterizing confidence intervals and conducting inference is needed.

In addition, we explore additional simulation settings with a slightly more complex confounding structure. Instead of using a completely randomized treatment, we consider a scenario where there is a pre-treatment confounder that jointly affects Z and M. The detailed settings and the corresponding results are presented in Section 3 of the supplementary file. Observations on the comparisons under these new settings are similar to those of the major simulations presented above.

Overall, we recommend $PLSEM_{AL}$ over $PLSEM_B$ as it provides less bias and variance and higher true positive rates in most scenarios.

Table 1: Summary of variable selection results in setting A. "True Positive" (TP) is the number of variables in $\mathcal{A} = (M_1, M_2, \ldots, M_{10})$ that are selected as the important mediators; TP rate is its value divided by 10. "False Positive" (FP) is the number of variables that are selected as the important mediators, but are not in $\mathcal{A}$; FP rate is its value divided by $p - 10$.

n	p	Model	Method	TP	TP rate	FP	FP rate
100	500	I	LSEM_{AL}	6.698	0.670	4.890	0.010
			PLSEM_{B}	4.510	0.451	0.070	0
			PLSEM_{AL}	6.974	0.697	3.856	0.008
			HIMA	1.376	0.138	0.020	0
		II	LSEM_{AL}	7.188	0.719	4.36	0.009
			PLSEM_{B}	4.444	0.444	0.072	0
			PLSEM_{AL}	7.014	0.701	5.006	0.010
			HIMA	1.998	0.200	0.036	0
		III	LSEM_{AL}	7.954	0.795	7.184	0.015
			PLSEM_{B}	4.400	0.440	0.088	0
			PLSEM_{AL}	7.290	0.729	6.254	0.013
			HIMA	3.442	0.344	0.026	0
300	500	I	LSEM_{AL}	8.280	0.828	2.252	0.005
			PLSEM_{B}	7.638	0.764	0.062	0
			PLSEM_{AL}	8.734	0.873	1.332	0.003
			HIMA	3.576	0.358	0.028	0
		II	LSEM_{AL}	8.666	0.867	2.250	0.005
			PLSEM_{B}	7.676	0.768	0.062	0
			PLSEM_{AL}	9.048	0.905	2.136	0.004
			HIMA	4.52	0.452	0.024	0
		III	LSEM_{AL}	9.238	0.924	2.628	0.005
			PLSEM_{B}	7.664	0.766	0.064	0
			PLSEM_{AL}	9.178	0.918	2.872	0.006
			HIMA	7.306	0.731	0.032	0
1000	500	I	LSEM_{AL}	8.954	0.895	0.726	0.001
			PLSEM_{B}	9.170	0.917	0.036	0
			PLSEM_{AL}	9.116	0.912	0.108	0
			HIMA	8.244	0.824	0.018	0
		II	LSEM_{AL}	8.800	0.880	2.330	0.005
			PLSEM_{B}	9.130	0.913	0.060	0
			PLSEM_{AL}	9.130	0.913	0.770	0.002
			HIMA	9.136	0.914	0.518	0.001
		III	LSEM_{AL}	9.506	0.951	0.380	0.001
			PLSEM_{B}	9.220	0.922	0.040	0
			PLSEM_{AL}	9.526	0.953	0.630	0.001
			HIMA	9.086	0.909	0.018	0
300	3000	I	LSEM_{AL}	7.848	0.785	4.372	0.001
			PLSEM_{B}	7.164	0.716	0.050	0
			PLSEM_{AL}	7.886	0.789	1.770	0.001
			HIMA	3.452	0.345	0.022	0
		II	LSEM_{AL}	8.754	0.875	3.896	0.001
			PLSEM_{B}	7.548	0.755	0.066	0
			PLSEM_{AL}	8.876	0.888	2.154	0.001
			HIMA	4.488	0.449	0.038	0
		III	LSEM_{AL}	8.892	0.889	4.078	0.001
			PLSEM_{B}	7.444	0.744	0.068	0
			PLSEM_{AL}	8.854	0.885	4.040	0.001
			HIMA	7.190	0.719	0.030	0

Table 2: Summary of Variable Selection Results for Setting B. “True Positive” (TP) is the number of variables in $A = (M_1, M_2, \ldots, M_{10})$ that are selected as the important mediators; TP rate is its value divided by 10. “False Positive” is the number of variables that are selected as the important mediators, but are not in $\mathcal{A}$; FP rate is its value divided by $p - 10$.

n	p	Model	Method	TP	TP rate	FP	FP rate
100	500	I	LSEM_{AL}	7.222	0.722	2.190	0.004
			PLSEM_B	4.870	0.487	0.424	0.001
			PLSEM_{AL}	7.274	0.727	3.006	0.006
			HIMA	2.506	0.251	0.028	0
		II	LSEM_{AL}	7.346	0.735	2.530	0.005
			PLSEM_B	4.874	0.487	0.532	0.001
			PLSEM_{AL}	7.462	0.746	3.598	0.007
			HIMA	2.984	0.298	0.042	0
		III	LSEM_{AL}	8.198	0.82	4.582	0.009
			PLSEM_B	4.836	0.484	0.494	0.001
			PLSEM_{AL}	7.620	0.762	4.082	0.008
			HIMA	4.298	0.43	0.138	0
300	500	I	LSEM_{AL}	8.460	0.846	0.040	0
			PLSEM_B	7.872	0.787	0.028	0
			PLSEM_{AL}	8.780	0.878	0.142	0
			HIMA	5.356	0.536	0.002	0
		II	LSEM_{AL}	8.714	0.871	0.268	0.001
			PLSEM_B	7.868	0.787	0.108	0
			PLSEM_{AL}	9.044	0.904	0.576	0.001
			HIMA	6.428	0.643	0.006	0
		III	LSEM_{AL}	9.194	0.919	0.552	0.001
			PLSEM_B	7.872	0.787	0.106	0
			PLSEM_{AL}	9.148	0.915	0.552	0.001
			HIMA	7.478	0.748	0.070	0
1000	500	I	LSEM_{AL}	8.954	0.895	0.726	0.001
			PLSEM_B	9.392	0.939	0	0
			PLSEM_{AL}	9.448	0.945	0	0
			HIMA	8.244	0.824	0.018	0
		II	LSEM_{AL}	8.790	0.879	2.890	0.006
			PLSEM_B	9.170	0.917	0.660	0.001
			PLSEM_{AL}	9.130	0.913	0.830	0.002
			HIMA	8.600	0.860	0.010	0
		III	LSEM_{AL}	9.634	0.963	0.178	0
			PLSEM_B	9.356	0.936	0.192	0
			PLSEM_{AL}	9.686	0.969	2.132	0.004
			HIMA	9.262	0.926	0.02	0
300	3000	I	LSEM_{AL}	7.860	0.786	0.448	0
			PLSEM_B	7.400	0.740	0.020	0
			PLSEM_{AL}	7.774	0.777	0.054	0
			HIMA	5.202	0.52	0.004	0
		II	LSEM_{AL}	8.536	0.854	0.414	0
			PLSEM_B	7.812	0.781	0.092	0
			PLSEM_{AL}	8.580	0.858	0.104	0
			HIMA	6.380	0.638	0.002	0
		III	LSEM_{AL}	8.662	0.866	0.486	0
			PLSEM_B	7.774	0.777	0.064	0
			PLSEM_{AL}	8.538	0.854	0.110	0
			HIMA	7.240	0.724	0.068	0

Table 3: Results of the standard error estimation for the joint indirect effect. The SE is the empirical standard error, the SE* is the standard error calculated by Eqn (13) and coverage is the coverage rate of the true indirect effect in the 95% confidence interval using ±2SE*.

					Setting A			Setting B	
n	p	Model	Methods	SE	SE*	Coverage	SE	SE*	Coverage
100	500	I	$LSEM_{AL}$	3.14	2.53	0.86	1.26	1.38	0.97
			$PLSEM_{B}$	1.14	0.90	0.95	0.81	0.68	0.95
			$PLSEM_{AL}$	1.43	0.96	0.91	0.96	0.68	0.94
			HIMA	1.87	0.60	0.57	1.80	0.69	0.71
		II	$LSEM_{AL}$	2.83	2.07	0.82	1.13	1.15	0.98
			$PLSEM_{B}$	1.16	0.91	0.95	0.81	0.68	0.94
			$PLSEM_{AL}$	1.49	0.97	0.88	0.88	0.69	0.93
			HIMA	1.79	0.63	0.69	1.68	0.65	0.75
		III	$LSEM_{AL}$	1.19	0.97	0.91	0.66	0.66	0.96
			$PLSEM_{B}$	1.17	0.9	0.95	0.86	0.67	0.94
			$PLSEM_{AL}$	1.40	0.98	0.91	0.81	0.69	0.95
			HIMA	1.06	0.48	0.85	1.02	0.49	0.92
300	500	I	$LSEM_{AL}$	2.12	1.49	0.53	0.74	0.82	0.95
			$PLSEM_{B}$	0.55	0.50	0.93	0.38	0.38	0.95
			$PLSEM_{AL}$	0.67	0.52	0.91	0.39	0.38	0.96
			HIMA	1.13	0.69	0.94	0.98	0.70	0.95
		II	$LSEM_{AL}$	1.74	1.19	0.55	0.64	0.67	0.96
			$PLSEM_{B}$	0.55	0.51	0.90	0.38	0.38	0.95
			$PLSEM_{AL}$	0.69	0.53	0.89	0.39	0.38	0.95
			HIMA	0.88	0.61	0.93	0.81	0.60	0.94
		III	$LSEM_{AL}$	0.66	0.52	0.91	0.38	0.38	0.96
			$PLSEM_{B}$	0.55	0.50	0.93	0.38	0.38	0.95
			$PLSEM_{AL}$	0.67	0.53	0.92	0.39	0.38	0.95
			HIMA	0.46	0.36	0.97	0.44	0.36	0.97
1000	500	I	$LSEM_{AL}$	1.14	0.81	0.10	1.14	0.81	0.10
			$PLSEM_{B}$	0.27	0.27	0.89	0.27	0.27	0.89
			$PLSEM_{AL}$	0.28	0.27	0.88	0.28	0.27	0.88
			HIMA	0.46	0.41	0.97	0.46	0.41	0.97
		II	$LSEM_{AL}$	1.08	0.64	0.08	0.35	0.36	0.97
			$PLSEM_{B}$	0.26	0.28	0.80	0.20	0.21	0.97
			$PLSEM_{AL}$	0.26	0.28	0.79	0.20	0.21	0.98
			HIMA	0.41	0.35	0.97	0.39	0.35	0.98
		III	$LSEM_{AL}$	0.31	0.27	0.89	0.20	0.21	0.98
			$PLSEM_{B}$	0.27	0.27	0.89	0.19	0.21	0.98
			$PLSEM_{AL}$	0.31	0.27	0.89	0.20	0.21	0.98
			HIMA	0.21	0.2	0.98	0.21	0.20	0.98
300	3000	I	$LSEM_{AL}$	2.16	1.49	0.54	0.80	0.82	0.94
			$PLSEM_{B}$	0.50	0.50	0.95	0.36	0.38	0.97
			$PLSEM_{AL}$	0.63	0.51	0.93	0.37	0.38	0.97
			HIMA	1.08	0.69	0.94	1	0.70	0.93
		II	$LSEM_{AL}$	1.72	1.20	0.61	0.62	0.66	0.97
			$PLSEM_{B}$	0.52	0.50	0.94	0.37	0.38	0.96
			$PLSEM_{AL}$	0.71	0.52	0.90	0.37	0.38	0.97
			HIMA	0.85	0.60	0.96	0.76	0.59	0.96
		III	$LSEM_{AL}$	0.65	0.52	0.91	0.36	0.37	0.97
			$PLSEM_{B}$	0.51	0.50	0.96	0.36	0.38	0.97
			$PLSEM_{AL}$	0.68	0.52	0.91	0.37	0.38	0.97
			HIMA	0.44	0.35	0.97	0.44	0.35	0.97

5. Data Application

It has been shown that DNA methylations play an important role in many human activities. In a genome-wide analysis of blood DNA methylation, the DNA methylation information are recorded on 85 subjects and their stress-related activities are examined (Houtepen et al., 2016). The researchers are interested in knowing whether DNA methylation plays a role in the regulation of human stress reactivity impacted by childhood trauma. The treatment variable is childhood trauma exposure, which was assessed using a version of the Childhood Trauma Questionnaire (Bernstein et al., 2003). The dataset is publicly available[3]. The distribution of the original score has a right skewed distribution ranging from 24 to 63, with a mean of 32 and standard deviation of of 8.22. After standardization (centered and scaled to have mean zero and standard deviation one), the range becomes -1 to 4. The outcome variable is cortisol stress reactivity, whose distribution is close to a bell-shaped curve, with a minimum of -1029.85 and max of 1876.28. The mean is 243.46 with a standard deviation of 420.6. The cortisol stress reactivity is measured from saliva samples collected from a stress induction task, which consists of a public speaking test and subsequent arithmetic task (Vinkers et al., 2013). The span of the experiment is 90 minutes and the participants in this study have an average age of 33 (Houtepen et al., 2016). The mediators are human DNA methylation markers. There are two important covariates that we need to adjust for: "Age" (X_1) and "Sex" (X_2). Figure 5 plots the relationship between the outcome variable and the covariate "Age", as well as the relationship between "Age" and some of the methylation markers identified by the proposed method. The plot shows that the relationships are nonlinear and the partial linear models can be helpful to accommodate the nonlinearity.

In the original dataset, there are a total of 385882 methylation variables. In order to scale down the computational burden, we first apply sure independence screening (Fan and Lv, 2008) to reduce the dimension to a reasonable scale. We follow the suggestion given by Gao et al. (2019) to select the top candidate mediators by fitting the following model:

$$Y = \beta_0 + \beta_j M_j + \gamma' Z + g(X_1) + X_2 + \epsilon_j, j = 1, \ldots, 385882.$$

The "Sex" covariate (i.e., X_2) is a binary variable, so we do not apply the smoothing function

[3] https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-77445

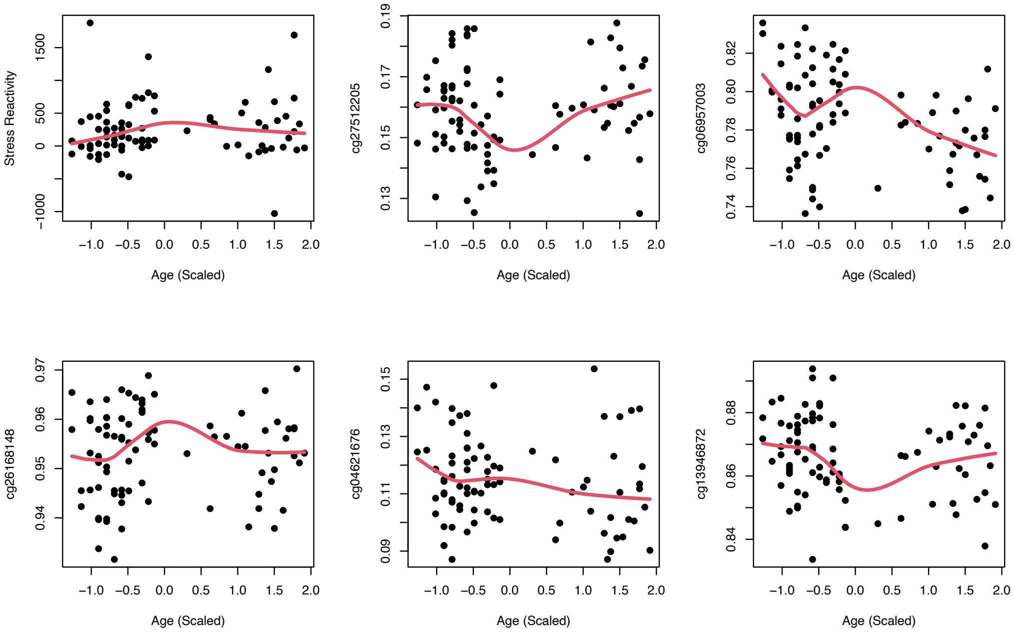

Figure 5: Top left: the relationship between the scaled age (continuous covariate) and the cortisol stress reactivity. Others: the relationship between the scaled age (continuous covariate) and the identified mediators. The red line corresponds to the smoothed curve that describes the relationship between the two variables.

Table 4: Results for real data analysis when $p = 500, 1000$, and 2000. The standard error of the indirect effect is estimated by Eqn (13). The p-value for the indirect effect is calculated by Wald test, and the p-value for the direct effect is reported by t-test. All significant effects at $\alpha = 0.05$ are marked with *. The PLSEM approach using Bonferroni correction is not listed as no mediator is selected.

		#		Indirect			Direct	
p	Method	Selected	Estimate	SE	$p-$value	Estimate	SE	$p-$value
500	LSEM$_{AL}$	7	-12.65	3.93	0.0006*	-2.10	4.49	0.64
	PLSEM$_{AL}$	36	-9.39	5.43	0.04*	-4.30	2.32	0.07
1000	LSEM$_{AL}$	18	-13.53	4.95	0.003*	-1.22	3.21	0.71
	PLSEM$_{AL}$	40	-10.13	5.47	0.03*	-3.56	2.71	0.20
2000	LSEM$_{AL}$	18	-18.47	5.04	0.0001*	3.72	3.73	0.32
	PLSEM$_{AL}$	33	-10.64	5.43	0.02*	-3.04	2.34	0.20

on it. We then pick the first p mediators whose effects are the largest. We consider three different p values: $p = 500, 1000, 2000$. The number of selected methylation markers and the estimated direct and indirect effects are provided in Table 4. The PLSEM approach using Bonferroni correction and the HIMA approach do not select any important mediators so we do not include it in the

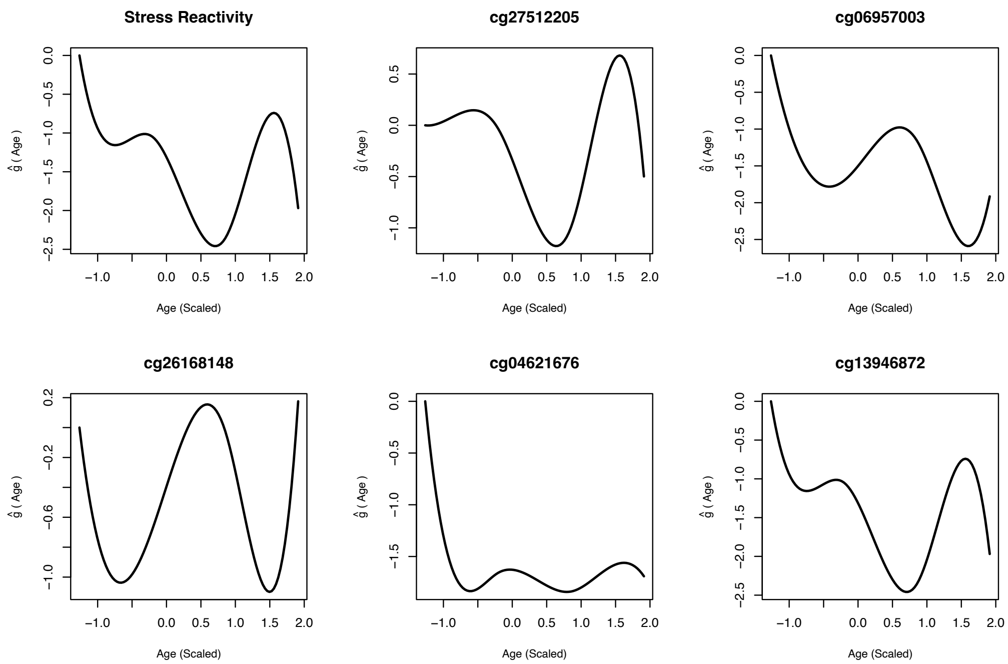

Figure 6: Top left: the estimated function of scaled age (continuous covariate) in the model to predict the cortisol stress reactivity. Others: the estimated function of the scaled age (continuous covariate) in the model to predict the identified mediators.

table. As we can see, the estimation results are consistent across different p values and different methods. Generally speaking, the indirect effect is significant at $\alpha = 0.05$ while the direct effect is not, which indicates some important methylation markers regulate the impact of childhood trauma on adulthood stress level. The selected methylation markers are displayed in Table 5. There is much overlap between LSEM$_{AL}$ and PLSEM$_{AL}$ approaches. In particular, the methylation marker "cg27512205" is selected by both approaches when $p = 500$. This marker is found to be an important locus on the *KITLG* gene that mediates the childhood trauma and cortisol stress reactivity based on three independent samples (Houtepen et al., 2016). The total effect of the treatment on the outcome is -13.69 (s.e.: 5.37 and p-value: 0.01). In mediation analysis, an important estimation quantity is how much of the total effect that can be explained by the given mediator(s).

Table 5: The selected mediators in real data analysis. The highlighted mediators are the ones selected by both methods for a given p. The number under each mediator in parenthesis is $\hat{\alpha}_j\hat{\beta}_j$/Total Effect

Method	Selected Mediators						
$p = 500$							
LSEM_{AL}	**cg27512205** (0.212)	**cg06957003** (0.040)	cg16746576 (0.138)	**cg26168148** (0.124)	cg11876022 (0.013)	**cg04621676** (0.194)	**cg13946872** (0.137)
PLSEM_{AL}	cg09573795 (0.026)	cg14414944 (-0.036)	cg03633948 (0.013)	cg00344209 (0.021)	cg18087143 (0.004)	**cg27512205** (0.019)	**cg25458175** (0.014)
	cg19695521 (0.007)	cg22073766 (0.018)	cg16999495 (0.025)	cg00083399 (0.009)	cg16512390 (0.014)	cg02381064 (-0.034)	cg27155653 (0.037)
	cg13341380 (0.020)	cg22120488 (0.009)	cg04262938 (0.037)	cg06621358 (0.025)	cg06957003 (0.068)	cg13539205 (0.030)	cg22396632 (0.043)
	cg00229532 (-0.001)	cg06144990 (0.054)	cg17880320 (-0.013)	cg16830861 (0.011)	cg23350558 (0.025)	**cg26168148** (0.017)	cg21926402 (0.019)
	cg22815785 (0.013)	cg19386484 (0.006)	cg21063480 (0.118)	cg15604507 (-0.013)	**cg04621676** (0.011)	cg25652781 (0.004)	**cg13946872** (0.051)
	cg23402444 (0.017)						
$p = 1000$							
LSEM_{AL}	cg16376036 (-0.023)	**cg14414944** (-0.058)	cg27512205 (0.115)	**cg22073766** (0.035)	cg05333968 (0.083)	**cg06957003** (0.046)	**cg22396632** (0.123)
	cg26168148 (0.082)	**cg21063480** (0.096)	**cg13946872** (0.135)	**cg21815667** (-0.036)	cg22203081 (0.057)	**cg20247596** (0.081)	cg23686508 (0.015)
	cg06019865 (0.067)	cg00719211 (0.047)	**cg02506717** (-0.013)	**cg09725013** (0.065)			
PLSEM_{AL}	cg09573795 (0.038)	**cg14414944** (-0.063)	cg03633948 (0.003)	cg00344209 (0.018)	cg25458175 (0.053)	cg23092635 (-0.010)	**cg22073766** (0.074)
	cg16999495 (-0.009)	cg00083399 (-0.007)	cg16512390 (0.006)	cg27155653 (0.010)	cg04262938 (0.046)	**cg06957003** (0.048)	cg13539205 (-0.023)
	cg22396632 (0.023)	cg17880320 (-0.021)	cg16830861 (0.028)	cg23350558 (0.018)	**cg26168148** (0.012)	cg02860705 (0.016)	cg12947510 (0.026)
	cg19386484 (0.006)	**cg21063480** (0.102)	cg15604507 (0.001)	cg04621676 (0.131)	**cg13946872** (0.055)	cg23402444 (0.022)	**cg21815667** (0.028)
	cg08504448 (0.040)	**cg20247596** (0.038)	cg16150053 (-0.011)	cg05971891 (0.077)	cg10147507 (-0.038)	cg17840166 (-0.063)	**cg06019865** (0.009)
	cg17189568 (0.038)	**cg02506717** (-0.061)	**cg09725013** (-0.034)	cg24165747 (-0.004)	cg00499707 (0.113)		
$p = 2000$							
LSEM_{AL}	cg27512205 (0.038)	**cg22073766** (0.156)	cg27155653 (0.062)	**cg22396632** (0.040)	cg19230917 (0.168)	cg13946872 (0.134)	**cg21815667** (0.062)
	cg22203081 (0.198)	**cg10147507** (-0.017)	**cg12446629** (0.042)	**cg03341991** (0.148)	**cg19975931** (0.031)	cg22713958 (-0.058)	cg20707780 (0.050)
	cg11753311 (-0.026)	**cg16908740** (0.126)	cg14843651 (-0.009)	**cg25626453** (0.109)			
PLSEM_{AL}	cg14414944 (-0.065)	cg00344209 (-0.002)	cg18087143 (0.006)	cg25458175 (0.072)	cg23092635 (0.019)	cg19695521 (0.017)	**cg22073766** (-0.022)
	cg16512390 (0.014)	cg04262938 (0.040)	cg05333968 (0.076)	**cg22396632** (0.054)	cg00229532 (0.006)	cg17880320 (-0.013)	cg23350558 (0.014)
	cg22815785 (0.028)	cg19386484 (0.011)	cg21063480 (0.080)	**cg21815667** (0.038)	**cg22203081** (0.038)	cg20247596 (0.057)	**cg10147507** (0.018)
	cg12446629 (-0.016)	**cg03341991** (0.077)	cg09725013 (-0.018)	cg00499707 (0.026)	cg08118034 (0.027)	**cg19975931** (-0.007)	cg07314988 (-0.006)
	cg05846894 (-0.015)	**cg16908740** (0.032)	cg04405414 (0.013)	cg05942970 (0.084)	**cg25626453** (0.094)		

6. Discussion

In this study we have examined the performance of a two-step procedure for mediation analysis when the number of mediators is large. With the increasing capability to measure various kinds of -omics data and the growing scientific interest to integrate information in a biologically meaningful way (Richardson et al., 2016), the analysis that can accommodate high-dimensional data will be much needed. We emphasize the importance of incorporating the confounders for drawing causality and allowing flexible models to account for nonlinearity. Although the individual indirect effect for each mediator is generally unidentifiable unless we impose some restrictions on the joint distribution of the mediators or on the correlation structure among the mediators (Wang et al., 2013), the parsimonious set of mediators obtained through the penalized framework provides a data-driven selected mediators for possible downstream scientific investigation. We look into estimation of mediation effects by examining the bias and coverage with the proposed standard error formula. Despite the possible underestimation caused by ignorance of variability in the variable selection step, the performance provides some evidence of the significance with conservativeness as expected.

In the two-step procedure, we choose to estimate Model (11) first and pass the selected set of mediators to Model (12) for further selection. We note that an order-free approach is to use the full set of mediators to estimate both Models (11) and (12), and then choose the subset of mediators that are selected in both models. Our choice is driven by the concern of high computation cost in estimating Model (12) where mediators are stacked to account for their correlation. Although theoretically the selection should be the same regardless of the order of estimation, it is not the case with real data given the high-dimensionality and finite sample size. If in practise, researchers choose to use the order-free approach, we recommend to use SIS to bring down the dimension to make the computation affordable and estimation more reliable. After all, the PLSEM approach using the Bonferroni correction may return a very conservative set. Under our current framework, researchers have flexibility if other smoothing methods or penalties are preferred. One limitation of our two-step methods is that by separately selecting α's and β's, the proposed method may miss the mediator which has a small signal on one but a large signal on the other. Although in examining the estimate and confidence interval of the product term we choose to evaluate the performance of the Sobel's method using the refitted coefficients with those selected in each model, we note that the refitting method is naive. The post-selection inference is an active research field (Gao et al., 2017; Kuchibhotla et al., 2021) and more sophisticated methods such as the debiased method as reviewed in Kuchibhotla et al. (2021) should be investigated in the future work. We also acknowledge that, in terms of the confounding structure, the simulation settings are rather simplistic. In the event where confounders are also high-dimensional, penalties can be applied on the confounders as well. This is an active research area in causal inference (Shortreed and Ertefaie, 2017; Ye et al., 2021). In particular, we note that the confounder selection problem is not the same as a typical variable selection problem in regression and researchers should follow the adjustment criteria (e.g., VanderWeele (2019)) to pick the confounders in mediation analysis.

Acknowledgements

We thank the editor the associate editor and the reviewers for careful reviews and insightful comments, which have led to a significant improvement of this article.

Funding

Zhu's research is supported by the National Sciences and Engineering Research Council of Canada (Grant No. RGPIN-2017-04064).

References

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444–455.

Aung, M.T., Song, Y., Ferguson, K.K., Cantonwine, D.E., Zeng, L., McElrath, T.F., Pennathur, S., Meeker, J.D., Mukherjee, B., 2020. Application of an analytical framework for multivariate mediation analysis of environmental data. Nature Communications 11, 1–13.

Baron, R.M., Kenny, D.A., 1986. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 51, 1173–1182.

Bernstein, D.P., Stein, J.A., Newcomb, M.D., Walker, E., Pogge, D., Ahluvalia, T., Stokes, J., Handelsman, L., Medrano, M., Desmond, D., et al., 2003. Development and validation of a brief screening version of the childhood trauma questionnaire. Child Abuse & Neglect 27, 169–190.

Bollen, K.A., 1987. Total, direct, and indirect effects in structural equation models. Sociological Methodology , 37–69.

Bollen, K.A., 1989. Structural equations with latent variables. volume 210. John Wiley & Sons.

Chén, O.Y., Crainiceanu, C., Ogburn, E.L., Caffo, B.S., Wager, T.D., Lindquist, M.A., 2018. High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics 19, 121–136.

Coffman, D.L., MacKinnon, D.P., Zhu, Y., Ghosh, D., 2016. A comparison of potential outcome approaches for assessing causal mediation, in: Statistical causal inferences and their applications in public health research. Springer, pp. 263–293.

Daniel, R.M., De Stavola, B.L., Cousens, S., Vansteelandt, S., 2015. Causal mediation analysis with multiple mediators. Biometrics 71, 1–14.

Fan, J., Gijbels, I., 2018. Local polynomial modelling and its applications: monographs on statistics and applied probability 66. Routledge.

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

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70, 849–911.

Gallop, R., Small, D.S., Lin, J.Y., 2009. Mediation analysis with principal stratification. Statistics in Medicine 28, 1108–1130.

Gao, X., Ahmed, S., Feng, Y., 2017. Post selection shrinkage estimation for high-dimensional data analysis. Applied Stochastic Models in Business and Industry 33, 97–120.

Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E.L., Cui, Y., 2019. Testing mediation effects in high-dimensional epigenetic studies. Frontiers in Genetics 10, 1195.

Guo, X., Li, R., Liu, J., Zeng, M., 2021. Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to covid-19 pandemic. Manuscript .

Härdle, W., Liang, H., Gao, J., 2000. Partially linear models. Physica Verlag.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models. Chapman & Hall/CRC.

Hines, O., Vansteelandt, S., Diaz-Ordaz, K., 2021. Robust inference for mediated effects in partially linear models. Psychometrika , 1–24.

Houtepen, L.C., Vinkers, C.H., Carrillo-Roa, T., Hiemstra, M., Van Lier, P.A., Meeus, W., Branje, S., Heim, C.M., Nemeroff, C.B., Mill, J., et al., 2016. Genome-wide dna methylation levels and altered cortisol stress reactivity following childhood trauma in humans. Nature Communications 7, 1–10.

Huang, Y.T., Pan, W.C., 2016. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics 72, 402–413.

Imai, K., Keele, L., Tingley, D., 2010a. A general approach to causal mediation analysis. Psychological Methods 15, 309–334.

Imai, K., Keele, L., Yamamoto, T., 2010b. Identification, inference and sensitivity analysis for causal mediation effects. Statistical Science 25, 51–71.

Imai, K., Yamamoto, T., 2013. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. Political Analysis , 141–171.

Jirolon, A., Baglietto, L., Birmeli, E., Alarcon, F., Perduca, V., 2020. Causal mediation analysis in presence of multiple mediators uncausally related. The International Journal of Biostatistics 1.

Judd, C., Kenny, D., 1981. Process analysis. Evaluation Review 5, 602–619.

Keele, L., Keele, L., 2008. Semiparametric regression for the social sciences. volume 230. Wiley Online Library.

Kraemer, H.C., Kiernan, M., Essex, M., Kupfer, D., 2008. How and why criteria defining moderators and mediators differ between the baron & kenny and macarthur approaches. Health Psychology 27, S101–108.

Kraemer, H.C., Wilson, G.T., Fairburn, C.G., Agras, W.S., 2002. Mediators and moderators of treatment effects in randomized clinical trials. Archives of General Psychiatry 59, 877.

Kuchibhotla, A.K., Kolassa, J.E., Kuffner, T.A., 2021. Post-selection inference. Annual Review of Statistics and Its Application 9.

Luo, C., Fa, B., Yan, Y., Wang, Y., Zhou, Y., Zhang, Y., Yu, Z., 2020. High-dimensional mediation analysis in survival models. PLoS Computational Biology 16, e1007768.

MacKinnon, D.P., Warsi, G., Dwyer, J.H., 1995. A simulation study of mediated effect measures. Multivariate Behavioral Research 30, 41–62.

Pearl, J., 2001. Direct and indirect effects, in: Proceedings of the seventeenth conference on uncertainty in artificial intelligence, pp. 411–420.

Preacher, K.J., Hayes, A.F., 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods 40, 879–891.

Preacher, K.J., Rucker, D.D., Hayes, A.F., 2007. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. Multivariate Behavioral Research 42, 185–227.

Richardson, S., Tseng, G.C., Sun, W., 2016. Statistical methods in integrative genomics. Annual Review of Statistics and Its Application 3, 181–209.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D.B., 1978. Bayesian inference for causal effects: The role of randomization. The Annals of Statistics , 34–58.

Shortreed, S.M., Ertefaie, A., 2017. Outcome-adaptive lasso: variable selection for causal inference. Biometrics 73, 1111–1122.

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

Ten Have, T.R., Joffe, M.M., Lynch, K.G., Brown, G.K., Maisto, S.A., Beck, A.T., 2007. Causal mediation analyses with rank preserving models. Biometrics 63, 926–934.

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

Valente, M.J., Pelham III, W.E., Smyth, H., MacKinnon, D.P., 2017. Confounding in statistical mediation analysis: What it is and how to address it. Journal of counseling psychology 64, 659–671.

VanderWeele, T., Vansteelandt, S., 2014. Mediation analysis with multiple mediators. Epidemiologic Methods 2, 95–115.

VanderWeele, T.J., 2016. Mediation analysis: a practitioner's guide. Annual review of public health 37, 17–32.

VanderWeele, T.J., 2019. Principles of confounder selection. European journal of epidemiology 34, 211–219.

Vinkers, C.H., Zorn, J.V., Cornelisse, S., Koot, S., Houtepen, L.C., Olivier, B., Verster, J.C., Kahn, R.S., Boks, M.P., Kalenscher, T., et al., 2013. Time-dependent changes in altruistic punishment following stress. Psychoneuroendocrinology 38, 1467–1475.

Wang, W., Nelson, S., Albert, J.M., 2013. Estimation of causal mediation effects for a dichotomous outcome in multiple-mediator models using the mediation formula. Statistics in Medicine 32, 4211–4228.

Ye, Z., Zhu, Y., Coffman, D.L., 2021. Variable selection for causal mediation analysis using lasso-based methods. Statistical Methods in Medical Research 30, 1413–1427.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., et al., 2016. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics 32, 3150–3154.

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

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