An Efficient Testing Procedure for High-Dimensional Mediators with FDR Control

Xueyan Bai[1], Yinan Zheng[2], Lifang Hou[2], Cheng Zheng[3], Lei Liu[4] and Haixiang Zhang[1*]

[1]*Center for Applied Mathematics, Tianjin University, Tianjin 300072, China*

[2]*Department of Preventive Medicine, Northwestern University, Chicago IL 60611, USA*

[3]*Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198, USA*

[4]*Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, USA*

Abstract

The field of mediation analysis commonly explores the pathways that connect environmental exposures with health outcomes. With the development of data collection techniques, greater efforts have been dedicated to addressing high-dimensional mediators. In this paper, we present an efficient approach to identify significant mediators while controlling the false discovery rate (FDR). We propose a three-step procedure that incorporates independent screening, variable selection together with refitted partial regression, and divide-aggregate composite-null test (DACT). The simulation includes a comparative analysis of our proposed method in comparison to eight competing approaches, demonstrating that our procedure has significant advantages over other methods. The proposed procedure is applied to investigate the mediation mechanisms of DNA methylation in the relationship between smoking and lung function. Three specific methylation sites (cg26331243, cg19862839, and cg12616487) are identified as potential epigenetic markers involved in mediating this relationship. Our proposed method is available with the R package HIMA at `https://cran.r-project.org/web/packages/HIMA/`.

Keywords: High-dimentional mediation analysis, FDR control, Multiple testing, Variable selection.

*Corresponding author: haixiang.zhang@tju.edu.cn (Haixiang Zhang)

1 Introduction

The method of mediation analysis is a modern statistical approach that investigates the underlying mechanism by which an exposure influences an outcome. It has been widely applied in many fields, including genome-wide association studies (Tam *et al.*, 2019), epidemiological investigations (Stringhini *et al.*, 2017), and socioeconomic research (Baron and Kenny, 1986). In the past few years, substantial research efforts have been devoted for mediation analysis. For example, Coffman (2011) proposed to adjust confounders between the mediator and the outcomes by calculating propensity score; Tchetgen and Shpitser (2012) established a general semiparametric framework for the natural direct and indirect causal effects; Lindquist (2012) extended structural equation models to the functional data analysis; Zhang and Wang (2013) compared four mediation analysis approaches when dealing with the missing data; Boca *et al.* (2014) developed a permutation approach to test multiple mediators; Shen *et al.* (2014) studied the topic on quantile mediation effects; Frölich and Huber (2014) investigated the non-parametric identification of causal direct and indirect effects of a binary treatment based on instrumental variables; Zheng and Zhou (2015) proposed a new model to handle both multilevel intervention and multicomponent mediators; VanderWeele and Tchetgen (2017) provided a weighting approach to direct and indirect effects based on combining the results of two marginal structural models, etc.

The aforementioned studies primarily focused on single or multiple, yet low-dimensional mediators. However, with the advancement of large-scale data collection techniques, considerable attention has been devoted to high-dimensional mediation analysis. For instance, Huang and Pan (2016) presented a hypothesis test for mediation effects of high-dimensional continuous mediators; the three-step procedure named HIMA, proposed by Zhang (2016), is designed to address the challenges posed by high-dimensional mediators; Wang *et al.* (2019) designed a rigorous Sparse Microbial Causal Mediation Model (SparseMCMM) for the high dimensional and compositional microbiome data in a typical three-factor causal study; Zhou *et al.* (2019) developed methods for both incomplete mediation, where a direct effect may exist, and complete mediation, where the direct effect

is known to be absent; Fasanelli *et al.* (2019) considered latent variables in the model for high-dimensional mediation analysis; Zhang *et al.* (2021a) and Zhang *et al.* (2021b) studied mediator selection procedures for compositional microbiome data. Yu *et al.* (2021) proposed a high-dimensional mediation analysis procedure which accommodating the potential confounders by using the propensity score. The HDMT method proposed by Dai et al. (2020) utilized a mixture distribution of three sub-null hypotheses and employed the joint significance (JS) approach to address the issue of conservativeness in JS; Zhang *et al.* (2021c) introduced a method for high-dimensional survival mediation analysis; the DACT method proposed by Liu et al. (2022) involves the computation of a weighted summation of p-values under three sub-null hypotheses. Perera *et al.* (2022) introduced HIMA2 by replacing minimax concave penalty (MCP) with de-biased Lasso. Zhang *et al.* (2024) studied the topic on high-dimensional quantile mediation analysis. Two review papers on high-dimensional mediation analysis were presented by Zeng *et al.* (2021) and Zhang *et al.* (2022).

Mediation analysis has emerged as a prevalent approach for elucidating the causal pathways linking an independent variable to a dependent variable through intermediate variables. Motivated by the ongoing Coronary Artery Risk Development in Young Adults (CARDIA) Study, which will be described in Section 4, our research aims to investigate the mechanistic role of DNA methylations in mediating the pathways between smoking and lung function. The calculation of p-values for high-dimensional mediators is widely acknowledged as challenging. To be specific, the p-values and estimates for the excluded mediators remain unknown due to the common utilization of variable screening or selection methods to reduce mediator dimensionality. To address this gap, we propose an efficient approach for computing p-values across all mediators with refitted partial regression. Then, we employ the DACT method to perform multiple testing on all mediators under FDR control.

The remainder of this paper is organized as follows. In Section 2, we introduce a three-step testing procedure for assessing high-dimensional mediation effects in linear regression models. In Section 3, the performance of our method is compared with eight different methods through numerical simulations. In Section 4, we apply our new pro-

cedure to investigate the mediating role of DNA methylation relating smoking and lung function. In Section 5, a brief conclusion is provided.

2 Statistical Methods

Mediation analysis investigates the intermediate mechanism through which an exposure exerts its influence the outcome. The diagram in Figure 1 depicts a classical mediation model that illustrates the interrelationships among an independent variable (X), multiple mediators ($M_1, \cdots, M_p$), and an outcome variable (Y).

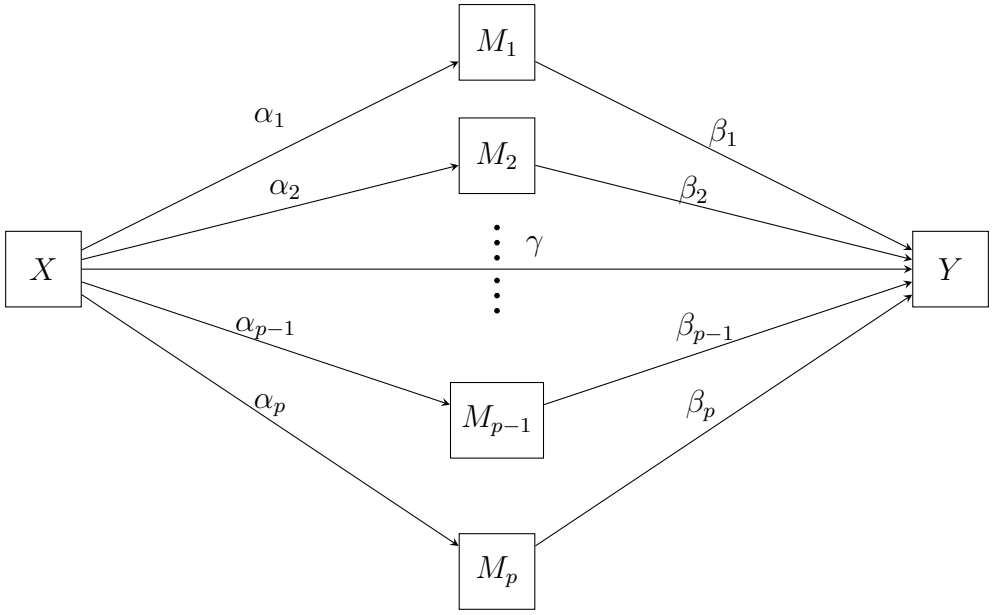

Figure 1. The scenario of a high-dimensional mediation model with omitted covariates.

We adopted the counterfactual framework for the vector of potential mediators $\boldsymbol{M}(x) = (M_1(x), M_2(x), \cdots, M_p(x))'$ under exposure level x, and $Y(x, \boldsymbol{m})$ the potential outcome under exposure level x and mediators level $\boldsymbol{m}$, to express the mediation analysis:

$$Y(x, \boldsymbol{m}) = c + \gamma X + \boldsymbol{m}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \epsilon,$$
$$M_i = c_i + \alpha_i X + \mathbf{Z}'\boldsymbol{\zeta}_i + e_i, \;\; i = 1, \cdots, p,$$

where X is the exposure, γ is the direct effect of exposure on the outcome; $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is the regression parameter vector relating the mediators to the outcome, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)'$

is the parameter vector relating the exposure to mediator, $\boldsymbol{\eta}$ and $\boldsymbol{\zeta}_i$ are corresponding regression coefficients for covariates $\mathbf{Z} = (Z_1, \cdots, Z_q)'$; c and c_i's are intercepts, ϵ and e_i's are mean-zero error terms.

A few causal assumptions that are needed for the identification of natural direct effect (NDE) and natural indirect effects (NIE) are listed below (Mathew *et al.*, 2019; Tsai *et al.*, 2018):

$A1$. Stable unit treatment value assumption (SUTVA) for both the mediators and the outcome. This assumption indicates that there is neither multiple versions of exposures nor interference between individuals, which implies that $\boldsymbol{M}(x)$ and $Y(x, \boldsymbol{m})$ are well defined.

$A2$. Consistency for the mediators and the outcome. That is, there are no measurement errors in the mediators and thus the observed variables satisfy $\boldsymbol{M} = \boldsymbol{M}(X)$ and $Y = Y(X, \boldsymbol{M})$.

$A3$. Sequential ignorability: This assumption contains 4 parts:

$(A3.1)$ $X \perp Y(x, \boldsymbol{m})|\boldsymbol{Z}$, $i.e.$, no unmeasured confounding between exposure and the potential outcome;

$(A3.2)$ $\boldsymbol{M} \perp Y(x, \boldsymbol{m})|X, \boldsymbol{Z}$, $i.e.$, no unmeasured confounding between mediators and the potential outcome;

$(A3.3)$ $X \perp \boldsymbol{M}(x)|\boldsymbol{Z}$, $i.e.$, no unmeasured confounding between exposure and the potential mediators;

$(A3.4)$ $\boldsymbol{M}(x^*) \perp Y(x, \boldsymbol{m})|\boldsymbol{Z}$, $i.e.$, no exposure-induced confounding between mediators and the potential outcome. In other words, the potential mediators under any intervention level $\boldsymbol{m}$ are independent of potential outcomes under any intervention x and mediator level x^* given covariate $\boldsymbol{Z}$.

$A4$. No direct causal relationship between mediators. We do not allow one mediator to be the cause of another, but we do allow them to have shared common causes.

The direct effect is $NDE = E[Y(1, M(0)) - Y(0, M(0))] = \gamma$ under assumption $A1 - A3$, and the indirect effect is $NIE = E[Y(1, M(1)) - Y(1, M(0))] = \sum_{i=1}^{p} \alpha_i \beta_i$. What's more, NIE can be decomposed into the summary of indirect effects through each mediator M_i under the additional assumption $A4$, $i.e.$ $NIE_i = \alpha_i \beta_i$.

We adopted the structural equation model (Zhang *et al.*, 2016) to assess the mediation effects of high-dimensional mediators:

$$Y = c + \gamma X + \mathbf{M}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\eta} + \epsilon, \tag{2.1}$$

$$M_i = c_i + \alpha_i X + \mathbf{Z}'\boldsymbol{\zeta}_i + e_i, \;\; i = 1, \cdots, p, \tag{2.2}$$

where the dimension of potential mediators (p) significantly exceeds the sample size (n). The index set of significant mediators is denoted as $\Omega = \{i : \alpha_i\beta_i \neq 0, i = 1, \cdots, p\}$.

In what follows, we are interested in the multiple testing problem:

$$H_{0i} : \alpha_i\beta_i = 0 \leftrightarrow H_{1i} : \alpha_i\beta_i \neq 0, \;\; i = 1, \cdots, p. \tag{2.3}$$

The proposed method can be described as a three-step approach: First, we utilize the sure independence screening (SIS; Fan and Lv, 2008) method in conjunction with the minimax concave penalty (MCP; Zhang, 2010) to effectively reduce the dimensionality of mediators. Second, the p-values for all mediators are obtained using the refitted partial regression method (Hao and Zhang, 2017). Finally, we perform multiple tests with FDR control for mediation effects using the DACT test. In Figure 2, we present a diagram for our three-step method. Details of the proposed procedure are given as follows.

Step 1: The mediators are initially screened using the following p marginal models,

$$Y = c + \gamma X + \beta_i M_i + \mathbf{Z}^T\boldsymbol{\eta} + \epsilon, \text{ for } i = 1, \cdots, p. \tag{2.4}$$

Let $\Omega_1 = \{i : M_i \text{ is among the top } d = [n/3] \text{ largest effects } |\tilde{\beta}_i| \text{ , } i = 1, \cdots, p\}$, where $\tilde{\beta}_i$ is the ordinary least square (OLS) estimator of β_i based on model (2.4). The MCP-estimators $\hat{\beta}_i^{MCP}$ are obtained by minimizing the criterion

$$Q(\boldsymbol{\beta}_{\Omega_1}; \tau, h) = \sum_{j=1}^{n} \left(Y_j - c - \gamma X_j - \sum_{i \in \Omega_1} \beta_i M_{ji} - \boldsymbol{Z}_j^T \boldsymbol{\eta} \right)^2 + \sum_{i \in \Omega_1} p_{\tau,h}(\beta_i),$$

where the sub-vector $\boldsymbol{\beta}_{\Omega_1}$ represents the subset of elements from vector $\boldsymbol{\beta}$ that correspond to indices in set Ω_1; the MCP function $p_{\tau,h}(\cdot)$ is

$$p_{\tau,h}(\beta_i) = \tau \left[|\beta_i| - \frac{|\beta_i|^2}{2\tau h} \right] I\{0 \leq |\beta_i| < \tau h\} + \frac{\tau^2 h}{2} I\{|\beta_i| \geq \tau h\}.$$

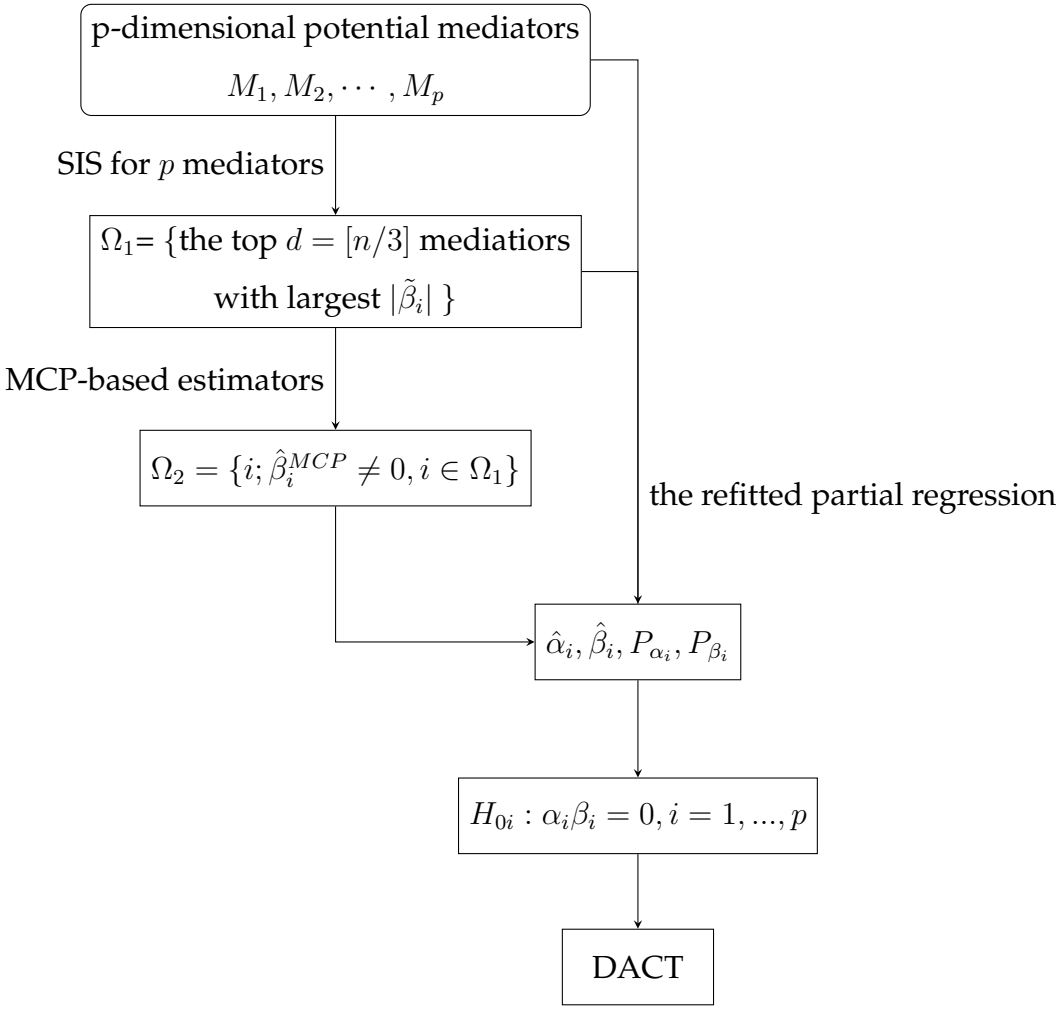

Figure 2. A diagram of multiple testing for high-dimensional mediators.

Here $\tau > 0$ is the regularization parameter, and the concavity of MCP is determined by $h > 0$. From the view of practial application, $\hat{\beta}_i^{MCP}$ can be obtained using the R package `ncvreg`. Denote the selected index as $\Omega_2 = \{i : \hat{\beta}_i^{MCP} \neq 0, i \in \Omega_1\}$.

Step 2: By the refitted partial regression method of Hao and Zhang (2017), we can derive the OLS estimators for β_i's:

• For $i \in \Omega_2$, the OLS estimator $\hat{\beta}_i$ and its standard error (SE) $\hat{\sigma}_{\beta_i}$ can be obtained by the sub-model $Y = c + \gamma X + \mathbf{M}_{\Omega_2}^T \boldsymbol{\beta}_{\Omega_2} + \mathbf{Z}^T \boldsymbol{\eta} + \epsilon$.

• For $i \notin \Omega_2$, the $\hat{\beta}_i$ and $\hat{\sigma}_{\beta_i}$ are obtained by the sub-model $Y = c + \gamma X + \mathbf{M}_{\Omega_2 \cup \{i\}}^T \boldsymbol{\beta}_{\Omega_2 \cup \{i\}} + \mathbf{Z}^T \boldsymbol{\eta} + \epsilon$.

The corresponding p-values for all mediators can be derived as follows:

$$P_{\beta_i} = 2\{1 - \Phi(|\hat{\beta}_i|/\hat{\sigma}_{\beta_i})\}, \text{ for i} = 1, \cdots, \text{p}, \tag{2.5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$. Similarly, we have

$$P_{\alpha_i} = 2\{1 - \Phi(|\hat{\alpha}_i|/\hat{\sigma}_{\alpha_i})\}, \text{ for i} = 1, \cdots, \text{p}. \tag{2.6}$$

Here $\hat{\alpha}_i$ is the OLS estimator of α_i and its SE is $\hat{\sigma}_{\alpha_i}$.

Step 3: We address the issue of multiple testing in (2.3) by the DACT method (Liu *et al.*, 2022). The test statistics are given as

$$P_{DACT,i} = \hat{\omega}_{01} P_{\alpha_i} + \hat{\omega}_{10} P_{\beta_i} + \hat{\omega}_{00} P_{JS,i}^2, \text{ for } i = 1, \cdots, p,$$

where $P_{JS,i} = \max(P_{\alpha_i}, P_{\beta_i})$, P_{β_i} and P_{α_i} are defined in (2.5) and (2.6), respectively; $\hat{\omega}_{01}, \hat{\omega}_{10}$ and $\hat{\omega}_{00}$ are three weights, which can be obtained based on the empirical characteristic function and Fourier analysis (Liu *et al.*, 2022). Let $\rho_i = \Phi^{-1}(1 - P_{DACT,i})$, where the function $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function (CDF) for a standard normal distribution. Denote

$$\widehat{FDR}(\rho_i) = \hat{\pi}_0 \hat{F}_0(\rho_i)/\hat{F}(\rho_i), \text{ for } i = 1, \cdots, p,$$

where $\hat{\pi}_0$ denotes the proportion of null mediation effects (e.g. H_{0i}), $\hat{F}_0(\cdot)$ is the empirical CDF of ρ_i under null mediation hypothesis, and $\hat{F}(\cdot)$ is the empirical CDF of ρ_i under nonnull hypothesis. To control FDR under level δ, the index set of significant mediators is given by $\hat{\Omega}_{DACT} = \{i : \widehat{FDR}(\rho_i) \leq \delta, i = 1, \cdots, p\}$.

3 Simulation Study

The performance of our method (denoted as "eHIMA") is evaluated through simulations in this section. We consider four scenarios for generating the random samples based on models (2.1) and (2.2):

• *Case 1*: We generate the exposure X from $N(0, 2)$. The covariate $Z = (Z_1, Z_2)'$, where Z_1 and Z_2 are independently generated from $N(2, 1)$. The intercept term c_i in model (2.2) is set as 0.5 and ϵ in model (2.1) is generated from $N(0, 1)$. We set $\zeta = (0.5, 0.5)'$, $\gamma = 0.5$ and $\eta = (0.3, 0.3)'$. The true regression coefficients are $\beta_1 = 0.2, \beta_2 = 0.3, \beta_3 = 0.25, \beta_4 = 0.2, \beta_5 = 0.35, \beta_6 = 0.1, \beta_7 = 0.25, \beta_8 = 0.1, \beta_9 = 0.2, \beta_{10} = 0.2$ and $\beta_i = 0$ for others; $\alpha_1 = 0.2, \alpha_3 = 0.15, \alpha_5 = 0.3, \alpha_7 = 0.15, \alpha_9 = 0.2, \alpha_{12} = 0.25, \alpha_{14} = 0.3$ and $\alpha_i = 0$ for others. i.e., $\Omega = \{1, 3, 5, 7, 9\}$. The error terms e_i's follow $N(0, \Sigma_e)$, where Σ_e is a matrix with $(\Sigma_e)_{i,j} = 0.85$ for $i, j = 1, 3, 5, 7, 9; (\Sigma_e)_{i,j} = 0.65$ for $i, j = 2, 4, 6, 8, 10$ and $(\Sigma_e)_{i,j} = 0.5^{|i-j|}$ for others.

• *Case 2*: We set $\Sigma_e = \left(0.8^{|i-j|}\right)_{i,j}$, and other settings are the same as Case 1.

• *Case 3*: The exposure X follows a mixture normal distribution $0.5N(-1, 1)+0.5N(1, 0.5)$. The error terms e_i's follow $N(0, \Sigma_e)$, where Σ_e is a matrix with $(\Sigma_e)_{i,j} = 0.65$ for $i, j = 1, 2, \cdots, 20$ and $(\Sigma_e)_{i,j} = 0.3^{|i-j|}$ for others, the rest settings are the same as Case 1.

• *Case 4*: The error terms e_i's follow $N(0, \Sigma_e)$, where Σ_e is a matrix $(\Sigma_e)_{i,j} = 0.65$ for $i, j = 1, 2, \cdots, 20$ and $(\Sigma_e)_{i,j} = 0.5^{|i-j|}$ for others. The rest settings are the same as Case 3.

To further investigate the impact of correlations among mediators, we manipulate the covariance matrix of error terms e_i's from Case 1 to Case 2. Additionally, the exposures in Cases 3 and 4 exhibit greater complexity compared to those in Cases 1 and 2. The comparison is conducted with two alternative methods, namely Sobel's test (referred to as "Sobel") and joint significant test (referred to as "JS"), in place of the DACT test in Step 3. The Sobel's test statistic is defined as

$$T_{Sobel,i} = \frac{\hat{\alpha}_i \hat{\beta}_i}{\sqrt{\hat{\beta}_i^2 \hat{\sigma}_{\alpha_i}^2 + \hat{\alpha}_i^2 \hat{\sigma}_{\beta_i}^2}}, \text{ for } i = 1, \cdots, p.$$

To control the FDR, we adopt the B-H method (Benjamini and Hochberg, 1995). The asymptotic distribution of $T_{Sobel,i}$ is assumed to follow a standard normal distribution.

Table 1: The simulation results with Case 1 and Case 2 ($p = 5000$).

		Case 1			Case 2		
	Methods	Model size	FDR	Power	Model size	FDR	Power
$n = 500$	**eHIMA**	**4.061**	**0.0350**	**0.7744**	**4.342**	**0.0342**	**0.8442**
	Sobel	1.844	$< 10^{-4}$	0.3689	2.442	0.0100	0.4876
	JS	2.411	$< 10^{-4}$	0.4822	3.282	0.0067	0.6524
	HIMA	2.417	0.0390	0.4567	1.810	0.1150	0.3080
	HIMA2	3.467	0.0106	0.6833	4.052	0.0095	0.8016
	MedFix	3.510	0.0352	0.6680	2.862	0.0421	0.5404
	BSLMM	3.270	$< 10^{-4}$	0.6540	3.130	0.0141	0.6144
	HDMA	4.190	0.0619	0.7720	4.670	0.0938	0.8320
	MT_Comp	6.428	0.2178	0.9822	5.942	0.1695	0.9660
$n = 600$	**eHIMA**	**4.494**	**0.0379**	**0.8556**	**4.810**	**0.0397**	**0.9072**
	Sobel	2.206	0.0019	0.4400	4.020	0.0013	0.8030
	JS	2.767	0.0019	0.5522	4.290	0.0013	0.8570
	HIMA	2.600	0.0434	0.4889	2.495	0.0568	0.4630
	HIMA2	3.783	0.0178	0.7400	4.460	0.0082	0.8830
	MedFix	3.880	0.0234	0.7302	4.030	0.0092	0.7980
	BSLMM	3.242	0.0045	0.6440	3.570	0.0020	0.7120
	HDMA	4.440	0.0694	0.8140	5.050	0.0642	0.9340
	MT_Comp	6.656	0.2308	0.9933	6.495	0.2230	0.9860

†"eHIMA" denotes our proposed method; others are eight competing approaches.

Consequently the p-values can be calculated as $P_{Sobel,i} = 2\{1 - \Phi(|T_{Sobel,i}|)\}$ for each $i = 1, \cdots, p$. We sort those p-values as $P^{(1)}_{Sobel} < P^{(2)}_{Sobel} < \cdots < P^{(p)}_{Sobel}$. Under the FDR level δ, we define the threshold q as

$$q = \max\left\{ i : P^{(i)}_{Sobel} \leq i\frac{\delta}{p},\ i = 1, \cdots, p \right\}.$$

The index set of significant mediators is denoted as $\hat{\Omega}_{sobel} = \{i : P_{Sobel,i} \leq P^{(q)}_{Sobel}, i = 1, \cdots, p\}$. Moreover, The p-value for the JS test is defined as

$$P_{JS,i} = \max(P_{\alpha_i}, P_{\beta_i}), \text{ for } i = 1, \cdots, p, \tag{3.7}$$

where P_{β_i} and P_{α_i} are defined in (2.5) and (2.6), respectively. Following the B-H method, the p-values are ordered as $P^{(1)}_{JS} < P^{(2)}_{JS} < \cdots < P^{(p)}_{JS}$. Under the FDR level δ, the threshold $\tilde{q}$ is set as

$$\tilde{q} = \max\left\{ i : P^{(i)}_{JS} \leq i\frac{\delta}{p} \right\}.$$

The index set of significant mediators is denoted as $\hat{\Omega}_{JS} = \left\{ i : P_{JS,i} \leq P^{(\tilde{q})}_{JS} \right\}$. In addition to the Sobel and JS methods, we also compare eHIMA with HIMA (Zhang *et al.*, 2016), HIMA2 (Perera *et al.*, 2022), MedFix (Zhang, 2022), BSLMM (Song *et al.*, 2020), HDMA (Gao *et al.*, 2019) and MT_Comp (Huang, 2019).

In Tables 1-4, we present the model size (MS) of the selected model, which represents the averaged cardinality of the estimated index of significant mediators; the FDR and Power of tests with a significance level of 0.05 are also provided. The results are based on 500 repetitions, with sample sizes of n=500 and 600 respectively. The dimensions of the mediators are set as p=5000 and 8000 respectively. The results presented in Tables 1-4 indicate that Sobel, JS, and BSLMM exhibit a significantly conservative behavior with low statistical power compared to the other six methods. The tendency of all methods is to select a model smaller than the true model with $|\Omega| = 5$, except for HDMA and MT_Comp. The FDRs of HDMA and MT_Comp exceed the predetermined significance level, resulting in a substantial number of false positive mediators. The proposed eHIMA method demonstrates superior performance in terms of FDR and Power in the simulated settings as a whole.

Table 2: The simulation results with Case 1 and Case 2 ($p = 8000$).

		Case 1			Case 2		
	Methods	Model size	FDR	Power	Model size	FDR	Power
$n = 500$	**eHIMA**	**4.081**	**0.0309**	**0.7838**	**4.465**	**0.0256**	**0.8640**
	Sobel	1.769	$< 10^{-4}$	0.3538	3.215	0.0017	0.6420
	JS	2.344	0.0021	0.4675	3.755	0.0058	0.7470
	HIMA	2.425	0.0422	0.4575	1.965	0.0825	0.3580
	HIMA2	3.531	0.0119	0.6963	4.285	0.0128	0.8430
	MedFix	3.421	0.0380	0.6480	2.802	0.0478	0.5340
	BSLMM	3.634	$< 10^{-4}$	0.7263	3.230	0.0095	0.6380
	HDMA	4.220	0.0784	0.7560	4.590	0.0847	0.8260
	MT_Comp	6.556	0.2283	0.9838	6.280	0.2050	0.9740
$n = 600$	**eHIMA**	**4.536**	**0.0402**	**0.8629**	**4.862**	**0.0265**	**0.9434**
	Sobel	2.186	$< 10^{-4}$	0.4371	4.190	$< 10^{-4}$	0.8382
	JS	2.636	$< 10^{-4}$	0.5271	4.490	$< 10^{-4}$	0.8980
	HIMA	2.636	0.0480	0.4929	2.661	0.1017	0.4780
	HIMA2	3.771	0.0121	0.7429	4.710	0.0140	0.9240
	MedFix	3.850	0.0283	0.7400	4.102	0.0173	0.8064
	BSLMM	3.704	0.0025	0.7375	3.790	0.0020	0.7562
	HDMA	4.788	0.0943	0.8450	5.290	0.0869	0.9480
	MT_Comp	6.629	0.2306	0.9900	6.540	0.2243	0.9964

†"eHIMA" denotes our proposed method; others are eight competing approaches.

Table 3: The simulation results with Case 3 and Case 4 ($p = 5000$)†.

		Case 3			Case 4		
	Methods	Model size	FDR	Power	Model size	FDR	Power
$n = 500$	**eHIMA**	**4.606**	**0.0485**	**0.8660**	**4.612**	**0.0567**	**0.8572**
	Sobel	1.222	0.0020	0.2436	1.258	0.0055	0.2488
	JS	2.652	0.0036	0.5280	2.638	0.0060	0.5232
	HIMA	3.468	0.0374	0.6632	3.476	0.0429	0.6604
	HIMA2	4.364	0.0312	0.8396	4.352	0.0342	0.8336
	MedFix	4.358	0.0328	0.8364	4.472	0.0460	0.8440
	BSLMM	2.696	0.0269	0.5224	2.602	0.0275	0.5024
	HDMA	5.166	0.1267	0.8848	5.142	0.1197	0.8880
	MT_Comp	5.686	0.1570	0.9372	5.684	0.1588	0.9352
$n = 600$	**eHIMA**	**4.922**	**0.0422**	**0.9382**	**4.890**	**0.0471**	**0.9224**
	Sobel	1.607	0.0010	0.3210	1.654	0.0025	0.3296
	JS	3.370	0.0009	0.6732	3.286	0.0026	0.6552
	HIMA	3.999	0.0359	0.7670	4.084	0.0346	0.7828
	HIMA2	4.749	0.0223	0.9244	4.668	0.0218	0.9088
	MedFix	4.808	0.0288	0.9248	4.760	0.0311	0.9144
	BSLMM	3.149	0.0112	0.6218	3.154	0.0128	0.6216
	HDMA	5.426	0.1034	0.9580	5.436	0.1103	0.9460
	MT_Comp	5.969	0.1631	0.9796	5.950	0.1606	0.9788

†"eHIMA" denotes our proposed method; others are eight competing approaches.

Table 4: The simulation results with Case 3 and Case 4 ($p = 8000$)†.

		Case 3			Case 4		
	Methods	Model size	FDR	Power	Model size	FDR	Power
$n = 500$	**eHIMA**	**4.435**	**0.0423**	**0.8580**	**4.525**	**0.0495**	**0.8640**
	Sobel	1.215	$< 10^{-4}$	0.2430	1.185	$< 10^{-4}$	0.2370
	JS	2.615	0.0025	0.522	2.700	0.0025	0.5390
	HIMA	3.540	0.0458	0.6680	3.765	0.0419	0.7160
	HIMA2	4.415	0.0364	0.8410	4.470	0.0277	0.8500
	MedFix	4.495	0.0326	0.8360	4.555	0.0334	0.8450
	BSLMM	2.700	0.0444	0.5100	2.715	0.0793	0.5010
	HDMA	5.170	0.1118	0.8950	5.070	0.1005	0.8980
	MT_Comp	5.6800	0.1681	0.924	5.710	0.1626	0.9380
$n = 600$	**eHIMA**	**4.894**	**0.0427**	**0.9276**	**4.965**	**0.0513**	**0.9328**
	Sobel	1.548	$< 10^{-4}$	0.3096	1.539	0.0012	0.3070
	JS	3.338	0.0009	0.6668	3.242	0.0026	0.6460
	HIMA	3.916	0.0339	0.7540	4.022	0.0283	0.7764
	HIMA2	4.712	0.0276	0.9108	4.728	0.0239	0.9178
	MedFix	4.700	0.0303	0.9052	4.761	0.0300	0.9162
	BSLMM	3.146	0.0235	0.6112	3.265	0.0338	0.628
	HDMA	5.380	0.1057	0.9484	5.393	0.0968	0.9562
	MT_Comp	6.024	0.1767	0.9720	5.943	0.1599	0.9768

†"eHIMA" denotes our proposed method; others are eight competing approaches.

4 Real data example

In this section, we apply the proposed methodology to the CARDIA Study, a longitudinal cohort study that investigates the development and determinants of clinical and subclinical cardiovascular disease as well as their risk factors (Friedman *et al.*, 1988). Similar to Perera *et al.* (2022), we focused on 1042 individuals from the CARDIA participants at Year 15 with 850,000 DNA methylation (DNAm) markers ($p = 850,000$). The FEV1 (forced expiratory volume in 1 s) measured at Year 20 is considered as the lung function outcome. The number of cigarette packs/year in Year 10 is the exposure variable. The analysis adjusts for age, height, weight, study center, gender, and race in (2.1) and (2.2) as confounders. Additionally, the proportions of CD4+T lymphocytes, CD8+T lymphocytes, B lymphocytes, natural killer cells, monocytes, and granulocytes are also adjusted to confounders. The mediation pathway was built in sequence: smoking at Year 10 $\rightarrow$ High dimensional DNAm markers at Year 15 $\rightarrow$ lung function at Year 20. More details of the data can be found in Perera *et al.* (2022). For analysis, the exposure, DNAm markers and continuous confounders are standardized with mean zero and variance one.

Based on the simulations, we mainly use six methods (eHIMA, Sobel, JS, HIMA, HIMA2, MedFix, BSLMM) to select active DNA methylation (DNAm) markers that mediate the relation between smoking and lung function. Based on FDR < 0.05, HIMA identifies cg26331243 only, HIMA2 identifies 2 CpGs: cg26331243 and cg19862839 and our eHIMA method identifies an extra cg12616487 over HIMA2's results. Other methods fail to select any significant mediators. The summary results on selected mediators obtained by the eHIMA method are presented in Table 5. The results of HIMA and HIMA2, which can be found in Perera *et al.* (2022), have been excluded.

The CpG cg26331243 is situated within the genomic region of gene CCDC33, which exhibits differential expression in response to tobacco smoke exposure (Gower *et al.*, 2011). Additionally, CCDC33 has been associated with susceptibility to lung function disorders (Lees *et al.*, 2019). The role of cg26331243 in regulating the expression of CCDC33, which mediates the pathway from smoking to lung function, is plausible (Perera *et al.*, 2022). The CpG cg19862839 is located in the body region of gene TBX4, which has been associ-

Table 5: Summary results of selected CpG with eHIMA method[†].

CpGs	Chromosome	Gene	$\hat{\alpha}_i(SE)$	$\hat{\beta}_i(SE)$
cg26331243	chr15	CCDC33	-0.081(0.016)	0.008 (0.019)
cg19862839	chr17	TBX4	-0.082(0.024)	-0.017 (0.018)
cg12616487	chr11	EML3/AHNAK	-0.128(0.028)	0.024 (0.018)

[†] "SE" is the standard error.

ated with a wide range of lung disorders (Haarman *et al.*, 2020). Additionally, mutations in TBX4 may increase susceptibility to cigarette smoking (Maurac *et al.*, 2019). We hypothesize that cg19862839 may be involved in the regulation of TBX4 expression, thereby serving as a mediator linking smoking and lung function (Perera *et al.*, 2022).

The CpG cg12616487 is located in the body region of the EML3/AHNAK gene, which has been demonstrated to exhibit colocalization with AHNAK expression and lung function (Jamieson *et al.*, 2020). The genes cg12616487, which have been previously implicated in an EWAS of maternal smoking (Burgess and Thompson, 2015), were found to have an effect on lung function. Therefore, we postulate that cg12616487 serves as a mediator in the intricate interplay between smoking and lung function.

5 Concluding Remarks

In this paper, we have proposed a three-step method for conducting high-dimensional mediation analysis. We used the sure independence screening and minmax concave penalty techniques to effectively reduce the dimensionality of potential mediators. To perform multiple testing, we utilized the oracle p-value approach. Furthermore, we applied the DACT to identify significant mediators. The simulation results clearly demonstrate the superiority of our method in terms of FDR, Power, and accuracy in identifying mediators compared to eight other approaches including Sobel, JS, HIMA, HIMA2, MedFix, BSLMM, HDMA and MT_Comp. Subsequently, we applied our proposed procedure to a real dataset investigating the impact of DNA methylation on smoking and lung function.

The future research will focus on two topics. First, we have considered mean regression models for the mediators and outcomes in our analysis. As highlighted by the reviewer, it would be intriguing to explore the impact of low or high DNA methylation levels. One potential approach is to employ the framework of quantile mediation analysis (Shen *et al.*, 2014; Bind *et al.*,2017; Zhang *et al.*, 2024), which falls beyond the scope of this paper and necessitates further investigation. Second, the expansion of our method to encompass other data types, including binary outcomes and longitudinal data, is highly desirable.

Acknowledgements

The authors would like to thank the Editor, the Associate Editor and two reviewers for their constructive and insightful comments that greatly improved the manuscript.

References

Baron, R. M. and Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* **51 6**, 1173–82.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society series b-methodological* **57**, 289–300.

Bind, M.-A., VanderWeele, T. J., Schwartz, J. D., and Coull, B. A. (2017). Quantile causal mediation analysis allowing longitudinal data. *Statistics in Medicine* **36**, 4182–4195.

Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C., and Sampson, J. N. (2014). Testing multiple biological mediators simultaneously. *Bioinformatics* **30 2**, 214–20.

Burgess, S. and Thompson, S. G. (2015). Multivariable mendelian randomization: The use

of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology* **181**, 251 – 260.

Coffman, D. L. (2011). Estimating causal effects in mediation analysis using propensity scores. *Structural Equation Modeling: A Multidisciplinary Journal* **18**, 357 – 369.

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

Fasanelli, F., Giraudo, M. T., Ricceri, F., Valeri, L., and Zugna, D. (2019). Marginal time-dependent causal effects in mediation analysis with survival data. *American journal of epidemiology* **188 5**, 967–974.

Friedman, G. D., Cutter, G., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jacobs, D. R., Liu, K., and Savage, P. J. (1988). Cardia: study design, recruitment, and some characteristics of the examined subjects. *Journal of clinical epidemiology* **41 11**, 1105–16.

Frölich, M. and Huber, M. (2014). Direct and indirect treatment effects–causal chains and mediation analysis with instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**.

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

Gower, A., Steiling, K., Brothers, J., Lenburg, M., and Spira, A. (2011). Transcriptomic studies of the airway field of injury associated with smoking-related lung disease. *Proceedings of the American Thoracic Society* **8**, 173–179.

Haarman, M. G., Kerstjens-Frederikse, W. S., and Berger, R. M. F. (2020). Tbx4 variants and pulmonary diseases: getting out of the 'box'. *Current Opinion in Pulmonary Medicine* **26**, 277 – 284.

Hao, N. and Zhang, H. H. (2017). Oracle p-values and variable screening. *Electronic Journal of Statistics* **11**, 3251–3271.

Huang, Y.-T. (2019). Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics* .

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72**.

Jamieson, E., Korogolou-Linden, R., Wootton, R. E., Guyatt, A. L., Battram, T., Burrows, K., Gaunt, T. R., Tobin, M. D., Munafo, M. R., Smith, G. D., Tilling, K., Relton, C. L., Richardson, T. G., and Richmond, R. C. (2020). Smoking, dna methylation, and lung function: a mendelian randomization analysis to investigate causal pathways. *American Journal of Human Genetics* **106**, 315 – 326.

Lees, J. A., Ferwerda, B., Kremer, P. H., Wheeler, N. E., and et al. (2019). Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nature Communications* **10**.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association* **107**, 1297 – 1309.

Liu, Z., Shen, J., Barfield, R., Schwartz, J., Baccarelli, A. A., and Lin, X. (2022). Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association* **117**, 67–81.

Mathew, A. R., Bhatt, S. P., Colangelo, L. A., Allen, N. B., Jacobs, D. R., Auer, R., Dransfield, M. T., Hitsman, B., Washko, G. R., and Kalhan, R. (2019). Life-course smoking trajectories and risk for emphysema in middle age: The cardia lung study. *American Journal of Respiratory and Critical Care Medicine* **199**, 237–240.

Maurac, A., Émilie Lardenois, Eyries, M., Ghigna, M. R., Petit, I., Montani, D., Guillaumot, A., Caput, B., Chabot, F., and Chaouat, A. (2019). T-box protein 4 mutation causing pulmonary arterial hypertension and lung disease. *European Respiratory Journal* **54**, 1900388.

Perera, C., Zhang, H., Zheng, Y., Hou, L., Qu, A., Zheng, C., Xie, K., and Liu, L. (2022).

Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. *BMC Bioinformatics* **23**.

Shen, E., Chou, C.-P., Pentz, M. A., and Berhane, K. T. (2014). Quantile mediation models: A comparison of methods for assessing mediation across the outcome distribution. *Multivariate Behavioral Research* **49**, 471 – 485.

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S. L. R., Roux, A. V. D., Needham, B. L., Smith, J. A., and Mukherjee, B. (2020). Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics* **76**, 700–710.

Stringhini, S., Zaninotto, P., Kumari, M., Kivimäki, M., Lassale, C., and Batty, G. D. (2017). Socio-economic trajectories and cardiovascular disease mortality in older people: the english longitudinal study of ageing. *International Journal of Epidemiology* **47**, 36 – 46.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics* 1–18.

Tchetgen, E. J. T. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics* **40 3**, 1816–1845.

Tsai, P.-C., Glastonbury, C. A., Eliot, M. N., Bollepalli, S., Yet, I., Castillo-Fernandez, J. E., Carnero-Montoro, E., Hardiman, T., Martin, T. C., Vickers, A., Mangino, M., Ward, K., Pietiläinen, K. H., Deloukas, P., Spector, T. D., Viñuela, A., Loucks, E. B., Ollikainen, M., Kelsey, K. T., Small, K. S., and Bell, J. T. (2018). Smoking induces coordinated dna methylation and gene expression changes in adipose tissue with consequences for metabolic health. *Clinical Epigenetics* **10**.

VanderWeele, T. J. and Tchetgen, E. J. T. (2017). Mediation analysis with time varying exposures and mediators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**.

Wang, C., Hu, J., Blaser, M. J., and Li, H. (2019). Estimating and testing the microbial

causal mediation effect with high-dimensional and compositional microbiome data. *bioRxiv* .

Yu, Z., Cui, Y., Wei, T., Ma, Y., and Luo, C. (2021). High-dimensional mediation analysis with confounders in survival models. *Frontiers in Genetics* **12**.

Zeng, P., Shao, Z., and Zhou, X. (2021). Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. *Computational and Structural Biotechnology Journal* **19**, 3209–3224.

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

Zhang, H., Chen, J., Feng, Y., Wang, C., Li, H., and Liu, L. (2021a). Mediation effect selection in high-dimensional and compositional microbiome data. *Statistics in Medicine* **40**, 885–896.

Zhang, H., Chen, J., Li, Z., and Liu, L. (2021b). Testing for mediation effect with application to human microbiome data. *Statistics in Biosciences* **13**, 313–328.

Zhang, H., Hong, X., Zheng, Y., Hou, L., Zheng, C., Wang, X., and Liu, L. (2024). High-dimensional quantile mediation analysis with application to a birth cohort study of mother–newborn pairs. *Bioinformatics* **40**, DOI:10.1093/bioinformatics/btae055.

Zhang, H., Hou, L., and Liu, L. (2022). A review of high-dimensional mediation analyses in DNA methylation studies. *In Guan, Weihua (Ed.), Epigenome-Wide Association Studies: Methods and Protocols* **2432**.

Zhang, H., Zheng, Y., Hou, L., Zheng, C., and Liu, L. (2021c). Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics* **37**, 3815–3821.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L., and Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32**, 3150–3154.

Zhang, Q. (2022). High dimensional mediation analysis with applications to causal gene identification. *Statistics in Biosciences* **14**, 432–451.

Zhang, Z. J. and Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika* **78**, 154–184.

Zheng, C. and Zhou, X. (2015). Causal mediation analysis in the multilevel intervention and multicomponent mediator case. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**.

Zhou, R. R., Wang, L., and Zhao, S. D. (2019). Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika* **107 3**, 573–589.