Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

High-Dimensional Mediation Analysis for Selecting DNA Methylation Loci Mediating Childhood Trauma and Cortisol Stress Reactivity

Xu Guo, Runze Li, Jingyuan Liu & Mudong Zeng

To cite this article: Xu Guo, Runze Li, Jingyuan Liu & Mudong Zeng (2022) High-Dimensional Mediation Analysis for Selecting DNA Methylation Loci Mediating Childhood Trauma and Cortisol Stress Reactivity, Journal of the American Statistical Association, 117:539, 1110-1121, DOI: 10.1080/01621459.2022.2053136

To link to this article: https://doi.org/10.1080/01621459.2022.2053136

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 539, 1110–1121: Applications and Case Studies
https://doi.org/10.1080/01621459.2022.2053136

High-Dimensional Mediation Analysis for Selecting DNA Methylation Loci Mediating Childhood Trauma and Cortisol Stress Reactivity

Xu Guo[a], Runze Li[b], Jingyuan Liu[c], and Mudong Zeng[b]

[a]School of Statistics, Beijing Normal University, Beijing, P.R. China; [b]Department of Statistics, The Pennsylvania State University at University Park, University Park, PA; [c]MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen, P.R. China

ABSTRACT

Childhood trauma tends to influence cortisol stress reactivity through the mediating effects of DNA methylation. Houtepen et al. conducted a study to investigate the role of DNA methylation in cortisol stress reactivity and its relationship with childhood trauma. The study collected a dataset consisting of 385,882 DNA methylation loci, cortisol stress reactivity, one-dimensional score on a childhood trauma questionnaire and several covariates for 85 healthy individuals. Of great scientific interest is to identify the active mediating loci out of the 385,882 ones. Houtepen et al. conducted 385,882 linear mediation analyses, in each of which one locus was considered, and identified three active mediating loci. More recently, van Kesteren and Oberski proposed a coordinate-wise mediation filter (CMF) and applied it to the same dataset. They identified five active mediating loci. Unfortunately, the three loci identified by Houtepen et al. are completely different from the five loci identified by van Kesteren and Oberski, probably because both Houtepen et al. and van Kesteren and Oberski did not consider all loci jointly in their analyses. The high dimensional DNA methylation loci indeed necessitate new techniques for identifying active mediating loci and testing the direct and indirect effects of the early life traumatic stress on later cortisol alteration. Motivated by the contradictory results in the aforementioned two scientific works, we develop a new estimating and testing procedure, and apply it to the same dataset as that analyzed by the two works. We identify three new loci: cg19230917, cg06422529 and cg03199124, and their effect sizes and p-values are 321.196 (p-value = 0.035965), 418.173 (p-value = 0.000234) and 471.865 (p-value = 0.001691), respectively. These three loci possess both reasonably neurobiological interpretations and statistically significant effects via our proposed tests. Based on our new procedure, we further confirm that the childhood trauma does not have significant direct effects on cortisol change—it only indirectly affects cortisol through DNA methylation, and the indirect effect is negative. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2021
Accepted March 2022

KEYWORDS
Cortisol stress reactivity; Direct effects; DNA; Early life trauma; High dimensional mediators; Indirect effects; Mediation analysis; Methylation

1. Introduction

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

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

These works mainly concentrated on single-layer linear models, where effects of early life trauma on DNA methylation and effects of DNA methylation on HPAA or cortisol alteration

CONTACT Jingyuan Liu *jingyuan@xmu.edu.cn* MOE Key Laboratory of Econometrics, Department of Statistics, School of Economics, Wang Yanan Institute for Studies in Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen, P.R. China.
*All authors made equal contributions to this work, and are listed in alphabetic order.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

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

From a statistical point of view, this is a high dimensional mediation problem, with DNA methylation loci being potential mediators, the vast majority of which though are supposed to be inactive. Nothwithstanding no shortage of strategies dealing with high dimensional mediators, including those in Houtepen et al. (2016) and van Kesteren and Oberski (2019), most existing literature rely on the marginal screening or penalized regression for sparse estimation. See for instance Preacher and Hayes (2008), Zhang et al. (2016), Serang et al. (2017), and so forth. A pitfall of using these dimension reduction techniques in each or either layer of mediation models lies in the pertinent difference between penalizing paths and finding actual mediators. That is, they choose paths instead of mediators. As a potential insight to break through this obstacle, Zhou, Wang, and Zhao (2020) proposed a debiased Lasso method that can integrate the two layers of high dimensional mediation models, and they also developed significance tests for both direct and indirect effects. However, the method proposed in Zhou, Wang, and Zhao (2020) involves high dimensional matrix estimation and operation, which might bring about a huge computational burden. In addition, the procedure penalizes all parameters, and the debiased step relies on the entire covariance matrix. This leads to inevitable efficiency loss of the estimators. More recently, Guo et al. (2021) observed that despite of high dimensional mediators, the direct and indirect effects are both low dimensional, with sum being the total effect. They thereby proposed a partial penalized approach for estimating the direct effects, which avoids high dimensional matrix estimation and the debiased step, and thus, enhances efficiency of proposed estimators. In spite of the plausible theory and efficient algorithms, Guo et al. (2021) have not yet explicitly elucidated the method with potential confounders, which typically should be considered when studying traumatic effects on cortisol alteration via DNA methylation, as in the literature (Houtepen et al. 2016; van Kesteren and Oberski 2019). Therefore, we in this article extend the work of Guo et al. (2021) to the models with confounders. Then we use our new procedure to study the mediating role of DNA methylation relating childhood trauma and cortisol stress reactivity, with several clinical variables involved as confounders. We further develop relevant tests for the direct and indirect effects of the early life trauma on cortisol stress reactivity.

Aside from the eight DNA methylation loci detected by Houtepen et al. (2016) and van Kesteren and Oberski (2019), we identify three additional loci on the RAB5IF gene (cg19230917), the CPQ gene (cg06422529) and the AGPAT1 gene (cg03199124) as mediators. We look through existing literature, and find reasonably neurobiological interpretations toward these three genes, with details referred to Section 3. Thus, our findings point out a potential direction for deeper neurobiological and epigenetic investigation of the connection between traumatic stress and cortisol alteration. From statistical point of view, we perform several statistical tests, and the results are also in support of the selected genes. According to the tests for the direct and indirect effects proposed in this article, the childhood trauma influences cortisol reactivity only through DNA methylation, since the indirect effect is negatively significant, yet the direct effect is not significant. In the full model with all detected loci, those from the newly identified genes are all significant, while the KITLG gene (cg27512205) selected by Houtepen et al. (2016), the HNRNPF gene (cg12500973) and the ZSCAN30 gene (cg16657538) selected by van Kesteren and Oberski (2019) are no longer significant. However, models with only the genes in Houtepen et al. (2016) yield a contradictory conclusion that KITLG is significant.

In Section 2, we introduce the statistical formulation of the high dimensional mediation problem, including the mediation models with confounders involved, the estimation for direct and indirect effects, and the tests of significance of indirect and direct effects. The detailed analysis is presented and discussed in Section 3. We also conduct a thorough simulation study to validate the finite sample performance of the proposed procedure in Section 4. A brief summary and conclusion are provided in Section 5.

2. Statistical Formulation: High Dimensional Linear Mediation Models with Confounders

In this section, we introduce the high dimensional mediation models with confounders involved, as the statistical formulation associating childhood trauma with cortisol stress reactivity via altering DNA methylation. Then we extend the partial penalization technique in Guo et al. (2021) to these models, for estimating and testing the direct and indirect traumatic effects.

Let y be the response variable, $\boldsymbol{m}$ consist of p-dimensional mediators, $\boldsymbol{x}$ consist of q-dimensional exposure variables, and $\boldsymbol{z}$

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

$$y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_1^T \boldsymbol{x} + \boldsymbol{\alpha}_2^T \boldsymbol{z} + \varepsilon_1, \tag{2.1}$$

$$\boldsymbol{m} = \Gamma_1^T \boldsymbol{x} + \Gamma_2^T \boldsymbol{z} + \boldsymbol{\varepsilon}_2, \tag{2.2}$$

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

Motivated by the real data analysis in Section 3, it is assumed throughout this article that q and d have fixed and finite dimensions, while p is high dimensional. Plugging (2.2) into (2.1), we obtain

$$\begin{aligned} y &= (\boldsymbol{\alpha}_1 + \boldsymbol{\beta})^T \boldsymbol{x} + (\Gamma_2 \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_2)^T \boldsymbol{z} + (\boldsymbol{\alpha}_0^T \boldsymbol{\varepsilon}_2 + \varepsilon_1), \\ &\equiv \boldsymbol{\gamma}_x^T \boldsymbol{x} + \boldsymbol{\gamma}_z^T \boldsymbol{z} + \varepsilon_3, \end{aligned} \tag{2.3}$$

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

2.1. Natural Direct and Natural Indirect Effects

We link the parameters $\boldsymbol{\alpha}_1$ and $\boldsymbol{\beta}$ with natural direct and natural indirect effects on a causal diagram. Let $y(x, m)$ denote the potential outcome that would have been observed had $\boldsymbol{x}$ and $\boldsymbol{m}$ been set to x and m, respectively, and $\boldsymbol{m}(x)$ denote the potential mediator that would have been observed had $\boldsymbol{x}$ been set to x. Following Imai, Keele, and Tingley (2010), Vanderweele and Vansteelandt (2014), and others, for $\boldsymbol{x} = x_1$ versus x_0, the natural direct effect is defined as

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

while the indirect effect is defined as

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

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

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

Furthermore, the independence assumptions of random errors in the mediation models (2.1) and (2.2) ensure the following sequential ignorability conditions (Imai, Keele, and Tingley 2010; Vanderweele and Vansteelandt 2014; Huang 2019; Zhou, Wang, and Zhao 2020).

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

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

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

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

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

$$\begin{aligned} E[y(x_1, \boldsymbol{m}(x_0)) - y(x_0, \boldsymbol{m}(x_0))] &= \boldsymbol{\alpha}_1^T (x_1 - x_0); \\ E[y(x_1, \boldsymbol{m}(x_1)) - y(x_1, \boldsymbol{m}(x_0))] &= \boldsymbol{\beta}^T (x_1 - x_0). \end{aligned}$$

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

2.2. Identifying Active Mediators

In this section, we introduce the procedure of identifying active mediators in the mediation models (2.1) and (2.2), and estimating the direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta}$ that can get around high dimensional matrix estimation. Suppose that $\{\boldsymbol{m}_i, \boldsymbol{x}_i, \boldsymbol{z}_i, y_i\}$, $i = 1, \ldots, n$ is a random sample from (2.1) and (2.2). Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n)^T$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^T$, and $\boldsymbol{W} = (\boldsymbol{X}, \boldsymbol{Z})$.

Despite high dimensionality of $\boldsymbol{m}$, model (2.3) is indeed a fixed-dimensional model. Therefore, the coefficient of $\boldsymbol{x}$, or say the total effect $\boldsymbol{\gamma}_x = \boldsymbol{\alpha}_1 + \boldsymbol{\beta}$, could be naturally estimated via the ordinary least squared estimator in model (2.3), that is,

$$\widehat{\boldsymbol{\gamma}}_x = (\boldsymbol{I}_q, \boldsymbol{O}_{q \times d})(\boldsymbol{W}^T \boldsymbol{W})^{-1} \boldsymbol{W}^T \boldsymbol{y}, \tag{2.4}$$

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

Another key observation is that $\boldsymbol{x}$ and $\boldsymbol{z}$ in model (2.1) are both fixed dimensional, thus, we opt not to impose sparsity on $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$. On the other hand, sparsity on $\boldsymbol{\alpha}_0$, the coefficient associated with the high dimensional mediator $\boldsymbol{m}$, could be naturally and reasonably assumed, as so in most existing high-dimensional literature. Therefore, following Guo et al. (2021), we apply the partial penalized least squared method to fit model (2.1) by only penalizing $\boldsymbol{\alpha}_0$. That is, the objective function subjected to minimization is

$$\frac{1}{2n} \| \boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha}_1 - \boldsymbol{Z}\boldsymbol{\alpha}_2 \|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|), \tag{2.5}$$

where α_{0j} is the jth element in $\boldsymbol{\alpha}_0$, and $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. Throughout this article, we will use the SCAD penalty, whose first-order derivative is

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

and set $a = 3.7$ as suggested by Fan and Li (2001). The tuning parameter λ is selected by HBIC (Wang, Kim, and Li 2013). A numerical algorithm to solve this penalized least squares problem is given in Section S.4 in the supplementary material of this article. Denote the corresponding estimates to be $\widehat{\boldsymbol{\alpha}}_0, \widehat{\boldsymbol{\alpha}}_1$

and $\widehat{\boldsymbol{\alpha}}_2$. Further note that $\boldsymbol{\beta} = \boldsymbol{\gamma}_x - \boldsymbol{\alpha}_1$. Then we can estimate the indirect effect $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\gamma}}_x - \widehat{\boldsymbol{\alpha}}_1$. Let $\mathcal{A} = \{j : \alpha_{0j} \neq 0\}$ and $\widehat{\mathcal{A}} = \{j : \widehat{\alpha}_{0j} \neq 0\}$. Under suitable conditions, we could obtain oracle property for our proposed estimates. That is, with probability tending to 1, $\widehat{\mathcal{A}} = \mathcal{A}$. Note that all truly active mediators are included in $\mathcal{A}$. This then implies that we can identify all truly potential mediators. We will empirically evaluate the performance of the partial penalized least squared method in (2.5) in the simulation section.

2.3. Test of Direct Effect and Indirect Effect

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

To develop tests for the direct effect $\boldsymbol{\alpha}_1$ and indirect effect $\boldsymbol{\beta}$, we first derive the asymptotic distributions of their estimators $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$. Let $\boldsymbol{w} = (\boldsymbol{x}^T, \boldsymbol{z}^T)^T$, $\boldsymbol{\alpha}_w = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T)^T$, $\Gamma_w = (\Gamma_1^T, \Gamma_2^T)^T$ and $\boldsymbol{\beta}_w = \Gamma_w \boldsymbol{\alpha}_0$. Thus, models (2.1) and (2.2) can be rewritten as

$$y = \boldsymbol{\alpha}_0^T \boldsymbol{m} + \boldsymbol{\alpha}_w^T \boldsymbol{w} + \varepsilon_1, \text{ and } \boldsymbol{m} = \Gamma_w^T \boldsymbol{w} + \boldsymbol{\varepsilon}_2, \tag{2.6}$$

which coincide with the causal mediation models considered in Guo et al. (2021). Thus, incorporating the results in Corollary 1 of Guo et al. (2021), the asymptotic distribution of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ can be obtained in a similar fashion. Specifically, define $\boldsymbol{m}_{\mathcal{A}}$ to be the subvector of $\boldsymbol{m}$ corresponding to $\mathcal{A} = \{j : \alpha_{0j} \neq 0\}$. And $\boldsymbol{\Sigma}_{MM} = E(\boldsymbol{m}_{\mathcal{A}} \boldsymbol{m}_{\mathcal{A}}^T)$, $\boldsymbol{\Sigma}_{MW} = \boldsymbol{\Sigma}_{WM}^T = E(\boldsymbol{m}_{\mathcal{A}} \boldsymbol{w}^T)$, $\boldsymbol{\Sigma}_{WW} = E(\boldsymbol{w}\boldsymbol{w}^T)$. Then

$$\sqrt{n}(\widehat{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_1) \to N(0, \sigma_1^2 (\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})(\boldsymbol{\Sigma}_{WW}^{-1} + B_w)(\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})^T), \tag{2.7}$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(0, (\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})(\sigma_2^2 \boldsymbol{\Sigma}_{WW}^{-1} + \sigma_1^2 B_w)(\boldsymbol{I}_q, \boldsymbol{O}_{q\times d})^T), \tag{2.8}$$

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

The asymptotic covariance matrices in (2.7) and (2.8) could be estimated in the same routine as Guo et al. (2021), by replacing quantities related to $\boldsymbol{x}$ in their work with those related to $\boldsymbol{w}$ in this article. These estimates lay the foundation of subsequential tests.

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

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

we modify the F-type lack-of-fit test proposed by Guo et al. (2021) by incorporating confounding effects. In model (2.1), denote RSS_f to be the residual sum of squares (RSS) in the full model fitted by the partial penalized least squares method (2.5), and RSS_r to be the RSS in the reduced model with $\boldsymbol{x}$ deleted from (2.1), obtained by the same partial penalized regression yet with objective function

$$\frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{M}\boldsymbol{\alpha}_0 - \boldsymbol{Z}\boldsymbol{\alpha}_2\|^2 + \sum_{j=1}^{p} p_\lambda(|\alpha_{0j}|). \tag{2.9}$$

The test statistic thereby is defined as

$$T = \frac{\text{RSS}_r - \text{RSS}_f}{\text{RSS}_f/(n-q-d)},$$

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

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

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

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

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

where $\widehat{B}_w = \widehat{\boldsymbol{\Sigma}}_{WW}^{-1} \widehat{\boldsymbol{\Sigma}}_{WM} (\widehat{\boldsymbol{\Sigma}}_{MM} - \widehat{\boldsymbol{\Sigma}}_{MW} \widehat{\boldsymbol{\Sigma}}_{WW}^{-1} \widehat{\boldsymbol{\Sigma}}_{WM})^{-1} \widehat{\boldsymbol{\Sigma}}_{MW} \widehat{\boldsymbol{\Sigma}}_{WW}^{-1}$. The hat versions are the sample counterparts of the covariance matrices. The limiting null distribution of S is $\chi^2(q)$, and the statistic can also detect the local effects with root-n convergence rate, as discussed in Guo et al. (2021). In addition, the Wald test for $H_{0\beta}$ is based on the asymptotical normality of $\widehat{\boldsymbol{\beta}}$. One may construct a Wald test for individual mediation effect β_j or a subvector of $\boldsymbol{\beta}$ based on their marginal asymptotical normality.

3. A Case Study: Exploration of Mediating Effects of DNA Methylation between Childhood Trauma and Cortisol Stress Reactivity

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

Houtepen et al. (2016) performed a genome-wide DNA methylation analysis for cortisol stress reactivity in healthy individuals. Since the number of DNA methylation loci is much greater than the sample size, Houtepen et al. (2016) first ran 385,882 linear regression models—response being cortisol stress reactivity, predictors being each out of the 385,882 DNA methylation loci, respectively, and confounders being several clinical variables. They reported 22,425 loci with p-values less than 0.05, while no statistically significant loci at level 0.05 after

adjustment for multiple testing. The authors then selected three loci that stood out in the p-value distribution of the genome-wide cortisol stress reactivity analysis. The three loci are cg27512205 (denoted by m_1), cg05608730 (m_2) and cg26179948 (m_3), based on which the authors further conducted a mediation analysis, and identified a locus on the KITLG gene (cg27512205) that is not only associated to cortisol stress reactivity, but also partly mediates the relationship between childhood trauma and cortisol stress reactivity. More importantly, they replicated the analysis using two independent samples from the whole blood and buccal (cross-tissue) DNA, respectively, and concluded that the KITLG locus is indeed a mediator.

More recently, van Kesteren and Oberski (2019) proposed a coordinate-wise mediation filter (CMF), which aims to improve the marginal screening method for linear mediation models with high-dimensional mediators. They further applied CMF for an empirical analysis of the same dataset as Houtepen et al. (2016), and identified five loci as the mediators. The five loci are cg16657538 (m_4), cg25626453 (m_5), cg02309301 (m_6), cg13136721(m_7), and cg12500973(m_8), which are completely different from the three loci identified by Houtepen et al. (2016). This contradiction motivates us to conduct a further analysis using the new procedure for studying the mediating role of DNA methylation that relates childhood trauma and cortisol alteration.

3.1. Mediation Analysis via the Proposed Procedures

In our analysis, the exposure variable ($\boldsymbol{x}$) is a one-dimensional score on a childhood trauma questionnaire, and the outcome y is the increased area under the curve (iAUC) in cortisol after a stress test. We consider 385,882 DNA methylation loci in the blood as potential mediators in $\boldsymbol{m}$. Following van Kesteren and Oberski (2019), we first carry out a screening step to retain the top 1000 potential mediators by ranking the absolute value of the product of two correlations—the correlation between $\boldsymbol{x}$ and each element of $\boldsymbol{m}$, and between y and each element of $\boldsymbol{m}$. This indeed is a marginal screening procedure based on Pearson correlation proposed by Fan and Lv (2008). They showed that for linear models, under some regularity conditions, the screening procedure possesses a sure screening property. We also include the eight loci identified by Houtepen et al. (2016) and van Kesteren and Oberski (2019) as domain knowledge and for comparison purpose. Furthermore, eight clinical variables are involved, including age (Z_1), sex (Z_2), B cell proportion (Z_3), CD4 T cell proportion (Z_4), CD8 T cell proportion (Z_5), Monocytes cell proportion (Z_6), Granulocytes cell proportion (Z_7) and Natural Killer cell proportion (Z_8), as confounding variables. This leads to the linear mediation models (2.1) and (2.2), where $\boldsymbol{x}$ (with dimension $q = 1$) and y are defined above; the confounder vector $\boldsymbol{z}$ is $\boldsymbol{z} = (Z_0, Z_1, \ldots, Z_8)^T$, with $Z_0 \equiv 1$ to include an intercept in the model.

We apply the proposed procedure to analyze the data. In the partial penalized least squares approach, we first select the tuning parameter λ by HBIC, and $\hat{\lambda} = 60.8163$. The eight loci $m_1, \ldots, m_8$ are treated as domain knowledge and are not penalized. Aside from them, our proposed method selects three additional loci cg19230917(m_9), cg06422529(m_{10}), and cg03199124(m_{11}). The estimated coefficients $\widehat{\boldsymbol{\alpha}}_0$, $\widehat{\boldsymbol{\alpha}}_1$, and $\widehat{\boldsymbol{\alpha}}_2$, along with their element-wise standard errors, t-values and p-values are listed in Table 1. The estimated coefficients of Γ_1 and Γ_2, with their p-values, are listed in Table 2. It can be seen from Table 1 that the p-values of the newly identified loci (i.e., m_9, m_{10}, and m_{11}) are all less than 0.05, while the p-values of m_1, m_5, and m_6 are greater than 0.10. Some significant effects are positive, while others are negative. From Table 2, childhood trauma $\boldsymbol{x}$ has significant effects on all the eleven mediators $m_1, \ldots, m_{11}$ except for m_8 at level 0.05. To sum up Tables 1 and 2, the newly identified loci m_9, m_{10}, and m_{11} should be considered as mediators since their coefficients are significant, and the coefficients of exposure variable on these loci are also significant at level 0.05.

Table 1. Estimated coefficients, SE, t-values and p-values.

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

Table 3 presents the results for testing the direct and indirect effects. The indirect effect is significant with p-value 0.0016, while the direct effect is insignificant since the p-value is 0.7643. Further note that the estimate of the indirect effect equals $-17.3726 < 0$. This implies that childhood trauma influences the cortisol stress reactivity only through the mediation mechanism of the DNA methylation, and the indirect effect is significantly negative.

Table 4 lists the 11 DNA methylation loci together with the genes to which they belong. It also provides some field knowledge of these genes dug out from existing research, according to which, the newly identified genes m_9, m_{10}, and m_{11} have particularly interesting biological and genetical interpretations. The locus m_9 corresponds to the RAB5IF gene (cg19230917). This gene modulates cell endocytosis process by which cells engulf substances, such as hormones, from outside into the cell (Ravikumar et al. 2008). Cortisol is a steroid hormone produced by the adrenal glands, and it may signal the cells through receptor for endocytosis. Thus, the RAB5IF gene likely plays a mediator rule that transmits the epigenetic alterations evoked by the traumatic stress. m_{10} belongs to the CPQ gene (cg06422529), which is shown by Hauptmann et al. (2017) to function in thyroid and tumor development. Peter (2011) testified this gene as a pathway from stress to cortisol level change

Table 2. Estimated coefficients of Γ_1 and Γ_2 and their p-values.

	x	Z_0	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
					Estimated Γ_1 and Γ_2					
m_1	0.005	−2.999	0.000	0.016	0.870	0.197	−0.311	0.784	0.406	0.010
m_2	0.007	−1.723	−0.002	0.013	1.479	0.032	0.693	0.602	0.935	1.448
m_3	0.006	−3.566	−0.001	0.004	0.985	0.704	−0.809	0.840	0.567	−0.373
m_4	−0.012	−9.222	0.007	−0.115	5.179	3.716	4.666	5.653	4.371	4.901
m_5	−0.011	−3.974	0.004	−0.009	2.051	−2.022	−0.485	0.437	−0.860	−0.594
m_6	0.023	0.322	0.003	0.403	−6.719	−5.542	0.115	−1.608	−5.701	−4.943
m_7	−0.012	5.701	0.002	0.000	0.562	−0.290	0.552	−0.510	−0.285	−0.516
m_8	0.012	1.093	0.001	0.083	−2.086	3.799	1.748	3.922	2.673	1.312
m_9	−0.006	−2.009	0.001	0.018	1.257	−1.623	−1.448	−0.942	−1.181	−1.276
m_{10}	−0.008	8.459	0.003	0.042	−7.930	−4.387	−2.036	−3.341	−4.408	−4.387
m_{11}	−0.006	−5.736	0.003	−0.080	4.745	2.151	1.013	1.080	3.009	2.688
					p-value of estimated coefficients of Γ_1 and Γ_2					
m_1	0.044	0.225	0.711	0.766	0.726	0.936	0.895	0.765	0.868	0.997
m_2	0.016	0.568	0.145	0.848	0.627	0.992	0.811	0.851	0.755	0.613
m_3	0.020	0.201	0.448	0.946	0.724	0.798	0.761	0.776	0.837	0.887
m_4	0.004	0.025	0.001	0.202	0.205	0.356	0.230	0.191	0.277	0.203
m_5	0.002	0.286	0.054	0.917	0.584	0.584	0.892	0.912	0.816	0.866
m_6	0.004	0.968	0.490	0.026	0.406	0.487	0.988	0.850	0.474	0.515
m_7	0.006	0.205	0.444	0.998	0.901	0.948	0.898	0.915	0.949	0.903
m_8	0.056	0.865	0.861	0.563	0.748	0.554	0.777	0.568	0.676	0.830
m_9	0.041	0.528	0.551	0.795	0.695	0.608	0.635	0.781	0.709	0.672
m_{10}	0.048	0.044	0.223	0.646	0.060	0.288	0.608	0.449	0.286	0.266
m_{11}	0.045	0.068	0.108	0.253	0.133	0.488	0.734	0.744	0.332	0.364

Table 3. The estimated coefficients, standard errors, test statistics values and p-values.

Coefficient	Estimated coefficient	SE	Test statistics	p-value
$\boldsymbol{\alpha}_1$	1.3653	4.5529	0.0899	0.7643
$\boldsymbol{\beta}$	−17.3726	5.4945	9.9971	0.0016

Table 4. Annotation of the included mediators.

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

in fish. A further neurobiological exploration is worthwhile to find out whether it has similar mediating effect in human body. m_{11} is located in AGPAT1 gene (cg03199124), which is involved in signal transduction and lipid biosynthesis for creating and storing body fat (Aguado and Campbell 1998). Some existing literature (Gonzalez-Bono et al. 2002; Kuo et al. 2007; Aschbacher et al. 2013) investigated the associations between physical stress like trauma and fat tissue biosynthesis. Vicennati et al. (2009) conducted a retrospective study and showed that women weight gain caused by trauma stress is accompanied by abnormal hormonal level such as cortisol. Our study which finds gene AGPAT1 as a mediator relating trauma stress and cortisol level therefore, may provide clues for such stress pathophysiological mechanism research. In summary, there is a reasonable conjecture that the identified loci, or their located genes, regulate neurobiological pathways and mediate the cortisol stress reactivity in response to childhood trauma, as also indicated by Table 3.

3.2. Some Comparisons

It is worth to compare our results with those in Houtepen et al. (2016) and van Kesteren and Oberski (2019) from statistical point of view. Define $\boldsymbol{m}_{(1)} = (m_1, m_2, m_3)^T$, $\boldsymbol{m}_{(2)} = (m_4, \ldots, m_8)^T$ and $\boldsymbol{m}_{(3)} = (m_1, \ldots, m_8)^T$. For the purpose of comparison, we consider three linear mediation models by replacing $\boldsymbol{m}$ in (2.1) and (2.2) with $\boldsymbol{m}_{(k)}$, $k = 1, 2$, and 3. The mediation models considered in Houtepen et al. (2016) coincide with (2.1) and (2.2) where $\boldsymbol{m}$ is taken to be $\boldsymbol{m}_{(1)}$, and models in van Kesteren and Oberski (2019) correspond to those with $\boldsymbol{m}_{(2)}$. We further consider the mediation models with $\boldsymbol{m}_{(3)}$, which merges $\boldsymbol{m}_{(1)}$ and $\boldsymbol{m}_{(2)}$. The estimated regression coefficients α_j's in model (2.1) are listed in Table 5. The estimated Γ_1 and Γ_2 and their values coincide with those in Table 2 because regressing the multiple responses $\boldsymbol{m}$ over the exposure variable and confounding variables in linear model (2.2) coincides with regressing individual mediator m_j over the exposure variable and confounding variables.

Table 5. Estimated α_j's and their SE and p-values

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

Table 6. The estimated coefficients, standard errors, test statistics values and p-values.

Model	Coefficient	Estimated coefficient	SE	Test statistics	p-value
$\boldsymbol{m}_{(1)}$	$\boldsymbol{\alpha}_1$	−5.487	4.916	1.246	0.2644
	$\boldsymbol{\beta}$	−10.52	3.612	8.447	0.0037
$\boldsymbol{m}_{(2)}$	$\boldsymbol{\alpha}_1$	−10.71	6.310	2.882	0.0896
	$\boldsymbol{\beta}$	−5.2946	4.913	1.161	0.2811
$\boldsymbol{m}_{(3)}$	$\boldsymbol{\alpha}_1$	−2.933	5.791	0.2565	0.6125
	$\boldsymbol{\beta}$	−13.07	5.356	5.959	0.0146

Tables 1 and 5 both suggest that the direct effect of childhood trauma, or say the coefficient of the exposure variable $\boldsymbol{x}$, is not significant in model (2.1). All confounding variables except for Z_2 (i.e., sex) are not significant. Mediators m_5 and m_6 are not significant based on all belonging models under investigation. Furthermore, comparing Tables 1 and 5, we observe that the effect of mediator m_1 may change from significance to insignificance at level 0.05, after inclusion of other mediators into the model. The reversal of test results for mediator m_1, as well as the insignificance of m_5 and m_6, motivates us to explore the relationship among all the identified mediators. Their pairwise correlations, partial correlations given $\boldsymbol{x}$ and $\boldsymbol{z}$, and several multiple regression models all reflect certain degree of association among mediators, which further explain their insignificance given other mediators included in the model. We put the detailed discussion in the supplementary material to save space.

The estimated direct and indirect effects for these three models are presented in Table 6, together with corresponding significance tests. This table indicates that the direct effect is not significant and indirect effect is significant for models with mediators $\boldsymbol{m}_{(1)}$ and $\boldsymbol{m}_{(3)}$, while both direct and indirect effects are not significant for model with mediator $\boldsymbol{m}_{(2)}$.

4. Simulation Studies

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

We set up the experiments to mimic the real data analyzed in Section 3 to the utmost. The sample size is taken to be the same, the dimension of potential mediators is 1000, corresponding to the 1000 candidate DNA methylation loci, and the exposure variable $\boldsymbol{x}$ and confounder $\boldsymbol{z}$ are directly extracted from the dataset. Meanwhile, $\boldsymbol{m}$ is generated via model (2.2), since it needs to be considered as random according to the mechanism of mediation models. Then y is accordingly generated from model (2.1). To accomplish this, we first draw Gaussian noise $\varepsilon_1 \sim N(0, \widehat{\sigma}_1^2)$ in model (2.1), where $\widehat{\sigma}_1^2$ is the estimated σ_1^2 in Section 3. The multivariate noise in model (2.2) is generated from $\boldsymbol{\varepsilon}_2 \sim N(0, \Sigma^*)$, where Σ^* is taken to be autoregressive covariance matrix. That is, the (i, j)-element of Σ^* equals $\rho^{|i-j|}$, and $\rho = 0.5$. The true mediators in $\mathcal{A}$ are designed in accordance with the 11 detected loci, $m_1, \ldots, m_{11}$, from the real data. Their associated coefficients $\boldsymbol{\alpha}_0$ in model (2.1) is taken to be $(1.0, 0.9, 0.8, -0.9, -0.8, -0.7, 0.6, 0.5, 0.4, 0.3, 0.2)$, with signs of elements consistent with those in $\widehat{\boldsymbol{\alpha}}_0$ estimated in Section 3. Moreover, the direct effect $\boldsymbol{\alpha}_1 = c_2$, where $c_2 = 0, 0.1, \ldots, 1.0$, to capture the size and power curve of the test for direct effect. For generating the indirect effect, the true value of Γ_1 is set to be $\Gamma_1 = c_1\widehat{\Gamma}_1$, where $c_1 = 0, \pm 0.1, \ldots, \pm 1.0$ and $\widehat{\Gamma}_1$ is the respective estimate from Section 3, thus, the indirect effect $\boldsymbol{\beta} = \Gamma_1\boldsymbol{\alpha}_0 = c_1\widehat{\Gamma}_1\boldsymbol{\alpha}_0 = -1.5977c_1$. As to the coefficients of confounding variables, we design the following two scenarios of models, without and with confounders, respectively.

4.1. Simulation Studies Without Confounding Variables

We first consider models without confounding variable $\boldsymbol{z}$. That is, $\boldsymbol{\alpha}_2$ and Γ_2 are both taken zero in model (2.1) and (2.2). We evaluate the indirect effect tests, by fixing the direct effect $\boldsymbol{\alpha}_1 = c_2 = 0.5$. The left panel of Figure 1 depicts the size ($c_1 = 0$) and power ($c_1 = \pm 0.1, \ldots, \pm 1.0$) for the three tests with statistics S,

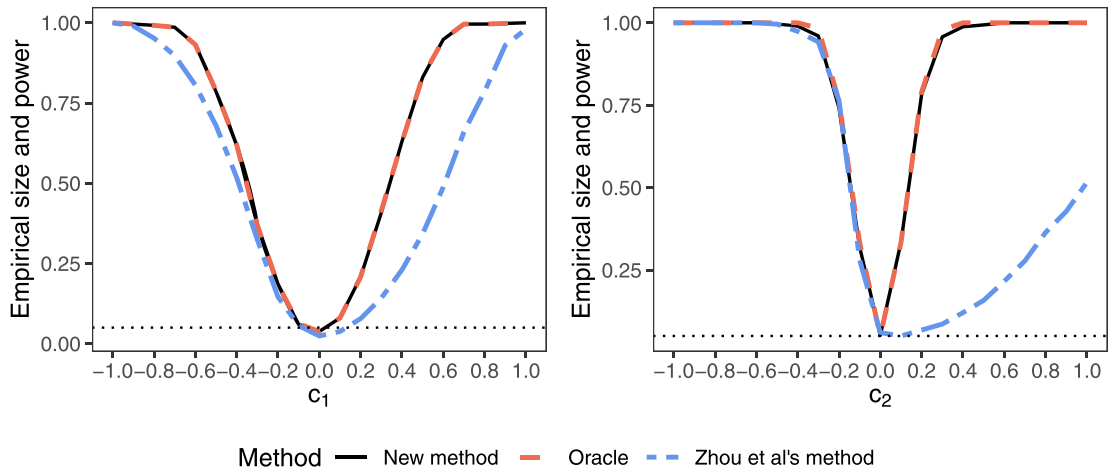

Figure 1. Left panel is the empirical sizes and powers of tests S, S^O, and S^Z at significance level 0.05 over 500 replications for testing indirect effect of mediation model without confounding variables. Solid line, dash line and two-dash line represent the sizes and powers of S, S^O, and S^Z, respectively. Right panel is empirical sizes and powers of tests T, T^O, and T^Z at significance level 0.05 over 500 replications for testing direct effect of mediation model without confounding variables. The solid, dash line and two-dash line represent the sizes and powers of T, T^O, and T^Z, respectively.

Table 7. Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2 when there is absent of confounding variables.

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

NOTE: Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

S^O and S^Z. From this figure, powers of all three tests increase as $|c_1|$ increases, and sizes are well controlled. Our proposed test S performs as well as the oracle test S^O, and is more powerful than S^Z. For instance, when $c_1 = 0.4$, the empirical powers of S and S^O are 0.63, while that of S^Z is 0.23.

We also consider testing direct effect $\boldsymbol{\alpha}_1$ by holding $c_1 = 0.5$, corresponding to true value of indirect effect $\beta = -0.7989$. Similarly, the right panel of Figure 1 shows the empirical size ($c_2 = 0$) and power ($c_2 = \pm 0.1, \ldots, \pm 1.0$) for the proposed test T, the oracle one T^O, and T^Z proposed by Zhou, Wang, and Zhao (2020). The powers of all three tests increase as the value of $|c_2|$ increases. T performs closely with T^O, and is more powerful than T^Z when c_2 is positive. For instance, when $c_2 = 0.2$, the empirical power of T and T^O can reach 0.78, while the empirical power of T^Z test is 0.06.

Moreover, we investigate the performances of the estimators of direct effect $\widehat{\boldsymbol{\alpha}}_1$ and indirect effect $\widehat{\boldsymbol{\beta}}$ in terms of bias and standard deviation. The results are reported in Table 7. From this table, the biases of our proposed estimators $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}$ and oracle ones $\widehat{\boldsymbol{\alpha}}_1^O, \widehat{\boldsymbol{\beta}}^O$ are very small, while the biases of $\widehat{\boldsymbol{\alpha}}_1^Z$ are very large. This in turn results in low power of S^Z and T^Z.

Table 8 depicts the sample standard deviations of the estimates $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ over 500 replications in the column with label "std," which can be regarded the true value of standard error of the estimates. These sample standard deviations are also shown in parentheses of Table 7. In the column with label "se(std)" in Table 8, we report the sample average and sample standard deviation of the 500 estimates of standard errors of $\widehat{\boldsymbol{\alpha}}_1$ and $\widehat{\boldsymbol{\beta}}$ based on the asymptotic covariance matrix formulas in (2.7) and (2.8). Note that the R package "freebird" in Zhou, Wang, and Zhao (2020) does not provide the estimated standard error of $\widehat{\boldsymbol{\alpha}}_1$. The difference between the column "std" and "se(std)" can be used to gauge the performance of the standard error formula based on the asymptotical covariance matrices. From Table 8, both the new method and the oracle have smaller difference than the method proposed by Zhou, Wang, and Zhao (2020).

4.2. Simulation Studies with Confounding Variables

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

Figure 2 shows the empirical sizes and powers of the tests S, S^O, and S^Z for indirect effect, and the tests T, T^O, and T^Z for direct effect. The left panel assesses the performance of tests for

Table 8. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications with different c_1 and c_2 when there is absent of confounding variables.

		Direct effect ($\widehat{\alpha_1}$)				Indirect Effect ($\widehat{\beta}$)					
		New method		Oracle		New method		Oracle		Zhou et al.'s method	
c_1	c_2	std	se(std)	std	se(std)	std	se(std)	std	se(std)	std	se(std)
−0.8	0.5	9.14	$8.10_{(0.94)}$	8.95	$7.91_{(0.94)}$	27.25	$28.70_{(2.04)}$	27.11	$28.70_{(2.04)}$	30.17	$29.73_{(2.32)}$
−0.4	0.5	6.95	$6.92_{(0.67)}$	6.86	$6.19_{(0.68)}$	26.48	$28.19_{(2.09)}$	26.42	$28.18_{(2.09)}$	26.31	$28.36_{(2.73)}$
0	0.5	5.98	$6.18_{(0.54)}$	5.94	$6.17_{(0.55)}$	26.10	$28.01_{(2.11)}$	26.02	$27.99_{(2.11)}$	28.83	$28.69_{(3.12)}$
0.4	0.5	6.87	$6.93_{(0.79)}$	6.69	$6.92_{(0.71)}$	26.16	$28.19_{(2.09)}$	26.23	$28.18_{(2.1)}$	28.11	$29.44_{(2.74)}$
0.8	0.5	9.01	$8.79_{(0.98)}$	8.68	$8.75_{(1.03)}$	26.68	$28.70_{(2.06)}$	26.23	$28.72_{(2.06)}$	27.27	$29.25_{(2.42)}$
0.5	−0.8	7.42	$7.36_{(0.77)}$	7.20	$7.35_{(0.88)}$	26.72	$28.28_{(2.19)}$	26.34	$28.09_{(2.08)}$	26.77	$29.17_{(2.6)}$
0.5	−0.4	7.36	$7.31_{(0.76)}$	7.10	$7.12_{(0.78)}$	26.69	$27.58_{(2.12)}$	26.11	$27.58_{(2.03)}$	27.02	$28.23_{(2.6)}$
0.5	0	7.29	$7.30_{(0.72)}$	6.93	$7.38_{(0.75)}$	24.25	$25.98_{(2.09)}$	23.96	$24.87_{(2.00)}$	26.07	$28.21_{(2.70)}$
0.5	0.4	7.36	$7.32_{(0.74)}$	7.04	$7.21_{(0.77)}$	25.47	$27.08_{(2.13)}$	25.32	$26.11_{(2.09)}$	28.03	$29.01_{(2.64)}$
0.5	0.8	7.47	$7.34_{(0.76)}$	7.11	$7.29_{(0.79)}$	26.75	$27.84_{(2.17)}$	26.44	$27.50_{(2.19)}$	29.13	$28.19_{(2.73)}$

NOTE: Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

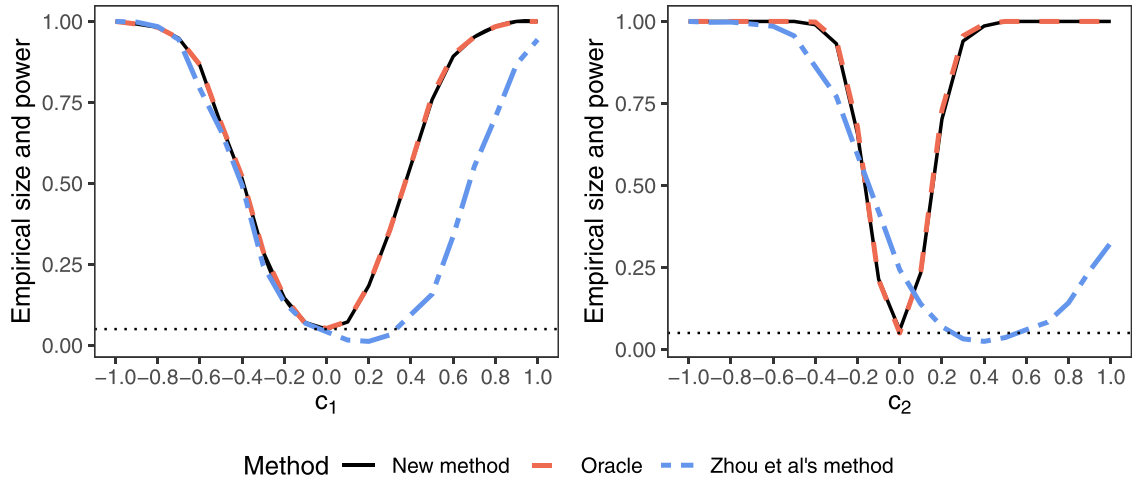

Figure 2. Left panel is the empirical sizes and powers of S, S^O, and S^Z at significance level 0.05 over 500 replications for testing indirect effect of mediation model with confounding variables. Right panel is empirical sizes and powers of T, T^O, and T^Z at significance level 0.05 over 500 replications for testing direct effect of mediation model with confounding variables. Caption is the same as that in Figure 1.

the indirect effect, holding $c_2 = 0.5$ as constant. From Figure 2, S performs as well as S^O, while S^Z exhibits a shifting power curve to the right and the minimum of the curve is not attained at the null hypothesis ($c_1 = 0$). For testing the direct effect $\boldsymbol{\alpha}_1$, we hold $c_1 = 0.5$ and hence, $\beta = -0.7989$. The values of c_2 vary from -1 to 1. The results are shown in the right panel of Figure 2. Not surprisingly, T performs as well as T^O, while T^Z suffers from an even more severe shifting power curve than S^Z for the indirect effects. For instance, when $c_2 = 0.4$, the empirical power of Zhou, Wang, and Zhao (2020) is only 0.024, while the empirical powers of our test and the oracle are 0.986 and 0.992, respectively.

To understand in depth the abnormal behavior of the power curves of Zhou, Wang, and Zhao (2020)'s tests, we investigate the performance of estimated direct effect $\widehat{\boldsymbol{\alpha}}_1$ and indirect effect $\widehat{\boldsymbol{\beta}}$ in terms of bias and standard deviation, as reported in Table 9. The biases of Zhou, Wang, and Zhao (2020)'s estimates are fairly large compared to the proposed estimators $\widehat{\boldsymbol{\alpha}}_1, \widehat{\boldsymbol{\beta}}$ and oracle ones $\widehat{\boldsymbol{\alpha}}_1^O, \widehat{\boldsymbol{\beta}}^O$. When holding $c_2 = 0.5$, the bias of $\widehat{\boldsymbol{\alpha}}_1^Z$ increases dramatically as c_1 increases. Similar phenomenon occurs when holding $c_1 = 0.5$, where the bias of $\widehat{\boldsymbol{\beta}}^Z$ changes substantially. The bias of estimated $\boldsymbol{\alpha}_1^Z$ and $\boldsymbol{\beta}^Z$ results in the shifted curves shown in Figure 2. The large bias of Zhou, Wang, and Zhao (2020)'s estimates and the low power of their tests are possibly due to the penalization on direct effect in the scaled lasso, as also pointed out in Zhou, Wang, and Zhao (2020) (see the discussion in their sec. 7). The penalization on direct effects would make sense when the direct effects are expected to small. This is another main merit of applying partial penalized method toward this problem setting.

We explore Zhou, Wang, and Zhao (2020)'s method more to understand its inferior performance. Note that the proposed method does not penalize the direct effect $\boldsymbol{\alpha}_1$, while Zhou, Wang, and Zhao (2020)'s method does penalize the direct effect in fitting scaled lasso (Sun and Zhang 2012). This leads to a larger estimated error variance $\widehat{\sigma}_1^2$ than the proposed method and the oracle estimator. Figure 3 compares the estimated $\widehat{\sigma}_1$ of the proposed estimate, oracle estimate and Zhou, Wang, and Zhao (2020) when confounding variables are involved in the mediation model. From Figure 3, we can observe that when c_1 or c_2 changes from negative to positive, Zhou, Wang, and Zhao (2020)'s estimated $\widehat{\sigma}_1$ dramatically increases, while the proposed method and oracle estimate do not. The increasing trend of variance estimation results in large biases of initial estimates used in Zhou, Wang, and Zhao (2020), making the debiased step

Table 9. Estimated biases and standard deviations (in parentheses) of different methods with different c_1 and c_2 when confounding variables involved.

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

NOTE: Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

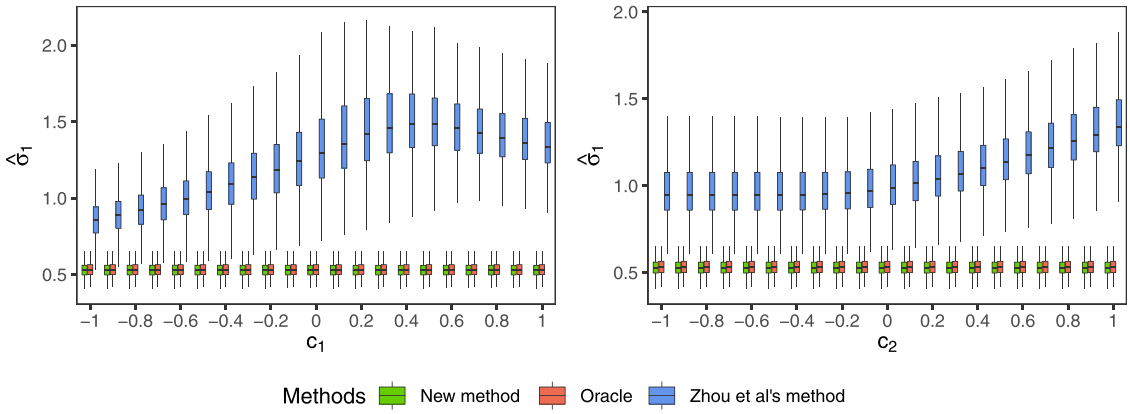

Figure 3. Left panel is the estimated $\widehat{\sigma}_1$ of our proposed new method using (2.5), oracle and Zhou, Wang, and Zhao (2020) (i.e., the scaled Lasso proposed by Sun and Zhang (2012)) over 500 replications by fixing $c_2 = 0.5$ when the mediation model contains confounding variables. Green, red and blue boxes represent the estimate of new method, oracle and Zhou, Wang, and Zhao (2020), respectively. Right panel is the estimated $\widehat{\sigma}_1$ of our method using (2.5), oracle and Zhou, Wang, and Zhao (2020) over 500 replications by fixing $c_1 = 0.5$ when the mediation model contains confounding variables. Green, red and blue boxes represent the estimate of new method, oracle and Zhou, Wang, and Zhao (2020), respectively.

Table 10. Estimated standard deviations and average estimated standard errors with their standard deviations (in parentheses) over 500 replications with different c_1 and c_2 when confounding variables involved.

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

NOTE: Except for c_1 and c_2, the values in this table equal 100 times of the actual ones.

more challenging. In addition, as c_1 or c_2 increases, estimating Ω through $\|\widehat{D} - \widehat{\Omega}\widehat{\Sigma}_{UU}\| \leq \tau$, where $\widehat{\Sigma}_{UU} = \boldsymbol{u}\boldsymbol{u}^T$, $\boldsymbol{u} = (\boldsymbol{m}^T, \boldsymbol{w}^T)^T$, and D and Ω are defined following Zhou, Wang, and Zhao (2020), requires larger tuning parameter τ, corresponding to more penalization on parameters and hence, further biases as well. Moreover, the power loss in the debiased step is attributed in part to multicollinearity, which also increases with c_1 and c_2, and when more confounders are involved.

As in Table 8, we report the empirical standard deviation of the 500 estimates and the average of 500 estimated standard errors over the 500 replications in Table 10 to examine the accuracy of variance estimation. For the new method and oracle, the standard errors estimated by Monte Carlo simulations are close to those calculated from formulas; while the empirical standard deviation and the average standard error of Zhou, Wang, and Zhao (2020) have a large gap.

5. Conclusion

Early life trauma plays a critical role in developing psychiatric disorders, typically via altering certain neuroendocrine substances like cortisol. Various researches thus, have been

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

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

We use this partial penalized least squares method and testing procedures to investigate the high dimensional mediating effects of DNA methylation loci to relate childhood trauma and cortisol stress reactivity, with confounding variables involved. For comparison purpose, we analyze the same dataset as Houtepen et al. (2016) and van Kesteren and Oberski (2019). We choose to include the eight DNA methylation loci discovered by these two papers in the model as domain knowledge. Using the proposed approach, we identified three new loci, on the RAB5IF gene (cg19230917), the CPQ gene (cg06422529) and the AGPAT1 gene (cg03199124), respectively, that actively play the mediator role. We compare our findings with Houtepen et al. (2016) and van Kesteren and Oberski (2019) from statistical perspectives, where tests and related analyses are all in favor of the three newly identified loci. Furthermore, we estimate and test the direct and indirect effects for childhood trauma on cortisol change, and conclude that the early life trauma affects cortisol only indirectly through DNA methylation and the indirect effect is negative, while the direct effect is insignificant.

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

Supplementary Materials

The supplementary materials consist of detailed explanations of confounders and relationship among the mediators for the empirical analysis in Section 3, and additional numerical comparison with the global test (Djordjilović et al. 2019).

Acknowledgments

The authors are grateful to the editor, the associate editor, and two anonymous referees for the constructive comments and suggestions that led to significant improvement of this work.

Funding

Guo's research was supported by National Natural Science Foundation of China grants (NSFC 12071038) and Beijing Natural Science Foundation (1212004). Liu's research was supported by grants from National Natural Science Foundation of China grants 11701034, 11771361 and 71988101. Li and Zeng's research was supported by National Science Foundation, DMS 1820702, 1953196, and 2015539.

References

Aguado, B., and Campbell, R. D. (1998), "Characterization of a Human Lysophosphatidic Acid Acyltransferase that is Encoded by a Gene Located in the Class III Region of the Human major Histocompatibility Complex," *The Journal of Biological Chemistry*, 273, 4096–4105. [1115]

Aschbacher, K., O'Donovan, A., Wolkowitz, O. M., Dhabhar, F. S., Su, Y., and Epel, E. (2013), "Good Stress, Bad Stress and Oxidative Stress: Insights from Anticipatory Cortisol Reactivity," *Psychoneuroendocrinology*, 38, 1698–1708. [1115]

Bremner, J. D., Vythilingam, M., Vermetten, E., Adil, J., Khan, S., Nazeer, A., Afzal, N., McGlashan, T., Elzinga, B., Anderson, G. M., and Heninger, G. (2003), "Cortisol Response to a Cognitive Stress Challenge in Posttraumatic Stress Disorder (PTSD) Related to Childhood Abuse," *Psychoneuroendocrinology*, 28, 733–750. [1110]

Burke, H. M., Davis, M. C., Otte, C., and Mohr, D. C. (2005), "Depression and Cortisol Responses to Psychological Stress: A Meta-analysis," *Psychoneuroendocrinology*, 30, 846–856. [1110]

Carpenter, L. L., Carvalho, J. P., Tyrka, A. R., Wier, L. M., Mello, A. F., Mello, M. F., Anderson, G. M., Wilkinson, C. W., and Price, L. H. (2007), "Decreased Adrenocorticotropic Hormone and Cortisol Responses to Stress in Healthy Adults Reporting Significant Childhood Maltreatment," *Biological Psychiatry*, 62, 1080–1087. [1110]

Carpenter, L. L., Shattuck, T. T., Tyrka, A. R., Geracioti, T. D., and Price, L. H. (2011), "Effect of Childhood Physical Abuse on Cortisol Stress Response," *Psychopharmacology (Berl)*, 214, 367–375. [1110]

Djordjilović, V., Page, C. M., Gran, J. M., Nøst, T. H., Sandanger, T. M., Veierød, M. B., and Thoresen, M. (2019), "Global Test for High-Dimensional Mediation: Testing Groups of Potential Mediators," *Statistics in Medicine*, 38, 3346–3360. [1120]

Edelman, S., Shalev, I., Uzefovsky, F., Israel, S., Knafo, A., Kremer, I., Mankuta, D., Kaitz, M., and Ebstein, R. P. (2012), "Epigenetic and Genetic Factors Predict Women's Salivary Cortisol Following a Threat to the Social Self," *PLoS ONE*, 7, e48597. [1110]

Elzinga, B. M., Spinhoven, P., Berretty, E., de, J. P., and Roelofs, K. (2010), "The Role of Childhood Abuse in HPA-Axis Reactivity in Social Anxiety Disorder: A Pilot Study," *Biological Psychiatry*, 83, 1–6. [1110]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of American Statistical Association*, 96, 1348–1360. [1112]

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. [1114]

Gonzalez-Bono, E., Rohleder, N., Hellhammer, D. H., Salvador, A., and Kirschbaum, C. (2002), "Glucose but not Protein or Fat Load Amplifies the Cortisol Response to Psychosocial Stress," *Hormones and Behavior*, 41, 328–333. [1115]

Guo, X., Li, R., Liu, J., and Zeng, M. (2021), "Statistical Inference for Linear Mediation Models with High-Dimensional Mediators and Application to Studying Stock Reaction to COVID-19 Pandemic," *https://arxiv.org/abs/2108.12329* [1111,1112,1113]

Hauptmann, S., Lange, D., Jarzab, M., Paschke, R., and Jarzab, B. (2017), "Gene Expression (mRNA) Markers for Differentiating between Malignant and Benign Follicular Thyroid Tumours," *International Journal of Molecular Sciences*, 18, 1184. [1114,1115]

Heim, C., Newport, D. J., Heit, S., Graham, Y. P., Wilcox, M., Bonsall, R., Miller, A. H., and Nemeroff, C. B. (2000), "Pituitary-Adrenal and Autonomic Responses to Stress in Women After Sexual and Physical Abuse in Childhood," *Journal of Amecian Medical Association*, 284, 592–597. [1110]

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., and Schalkwyk, L. C. (2016), "Genome-Wide DNA Methylation Levels and Altered Cortisol Stress Reactivity Following Childhood Trauma in Humans," *Nature Communications*, 7, 10967. [1111,1113,1114,1115,1120]

Huang, Y. T.(2019), "Genome-Wide Analyses of Sparse Mediation Effects Under Composite Null Hypotheses," *Annals of Applied Statistics*, 13, 60–84. [1112]

Imai, K., Keele, L., and Tingley, D. (2010), "A General Approach to Causal Mediation Analysis," *Psychological Methods*, 15, 309–334. [1112]

Kraft, A. J., and Luecken, L. J. (2009), "Childhood Parental Divorce and Cortisol in Young Adulthood: Evidence for Mediation by Family Income," *Psychoneuroendocrinology*, 34, 1363–1369. [1110]

Kuo, L. E., Kitlinska, J. B., Tilan, J. U., Li, L., Baker, S. B., Johnson, M. D., Lee, E. W., Burnett, M. S., Fricke, S. T., Kvetnansky, R., and Herzog, H. (2007), "Neuropeptide Y acts Directly in the Periphery on Fat Tissue and Mediates Stress-Induced Obesity and Metabolic Syndrome," *Nature Medicine*, 13, 803–811. [1115]

Lei, X., and Wong, G. W. (2019), "C1q/TNF-related Protein 2 (CTRP2) Deletion Promotes Adipose Tissue Lipolysis and Hepatic Triglyceride Secretion," *The Journal of Biological Chemistry*, 294, 15638–15649. [1115]

Liao, Z. Z., Wang, Y. D., Qi, X. Y., and Xiao, X. H. (2019), "JAZF1, a Relevant Metabolic Regulator in Type 2 Diabetes," *Diabetes/Metabolism Research and Reviews*, 35, e3148. [1115]

Luecken, L. J. (1998), "Childhood Attachment and Loss Experiences Affect Adult Cardiovascular and Cortisol Function," *Psychosomatic Medicine*, 60, 765–772. [1110]

McGowan, P. O., Sasaki, A., D'alessio, A. C., Dymov, S., Labonté, B., Szyf, M., Turecki, G., and Meaney, M. J. (2009), "Epigenetic Regulation of the Glucocorticoid Receptor in Human Brain Associates with Childhood Abuse," *Nature Neuroscience*, 12, 342–348. [1110]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [1113]

Perroud, N., Paoloni-Giacobino, A., Prada, P., Olié, E., Salzmann, A., Nicastro, R., Guillaume, S., Mouthon, D., Stouder, C., Dieben, K., and Huguelet, P. (2011), "Increased Methylation of Glucocorticoid Receptor Gene (NR3C1) in Adults with a History of Childhood Maltreatment: A Link with the Severity and Type of Trauma," *Translational Psychiatry*, 1, e59. [1110]

Pesonen, A. K., Räikkönen, K., Feldt, K., Heinonen, K., Osmond, C., Phillips, D. I., Barker, D. J., Eriksson, J. G., and Kajantie, E. (2010), "Childhood Separation Experience Predicts HPA Axis Hormonal Responses in Late Adulthood: A Natural Experiment of World War II," *Psychoneuroendocrinology*, 35, 758–767. [1110]

Peter, M. C. (2011), "The Role of Thyroid Hormones in Stress Response of Fish," *General and Comparative Endocrinology*, 172, 198–210. [1114]

Petrowski, K., Wintermann, G. B., Schaarschmidt, M., Bornstein, S. R., and Kirschbaum, C. (2013), "Blunted Salivary and Plasma Cortisol Response in Patients with Panic Disorder Under Psychosocial Stress," *International Journal of Psychophysiology*, 88, 35–39. [1110]

Preacher, K. J., and Hayes, A. F. (2008), "Asymptotic and Resampling Strategies for Assessing and Comparing Indirect Effects in Multiple Mediator Models," *Behavior Research Methods*, 40, 879–891. [1111]

Ravikumar, B., Imarisio, S., Sarkar, S., O'Kane, C. J., and Rubinsztein, D. C. (2008), "Rab5 Modulates Aggregation and Toxicity of Mutant Huntingtin through Macroautophagy in Cell and Fly Models of Huntington Disease," *Journal of Cell Science*, 121, 1649–1660. [1114,1115]

Serang, S., Jacobucci, R., Brimhall, K. C., and Grimm, K. J. (2017), "Exploratory Mediation Analysis via Regularization," *Structural Equation Modeling*, 24, 733–744. [1111]

Sun, T., and Zhang, C-H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [1118,1119]

Van de Geer, S., Buhlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [1113]

van Kesteren, E. J., and Oberski, D. L. (2019), "Exploratory Mediation Analysis with Many Potential Mediators," *Structural Equation Modeling: A Multidisciplinary Journal*, 26, 710–723. [1111,1113,1114,1115,1120]

Vanderweele, T. J., and Vansteelandt, S. (2014), "Mediation Analysis with Multiple Mediators," *Epidemiologic Methods*, 2, 95–115. [1112]

Vicennati, V., Pasqui, F., Cavazza, C., Pagotto, U., and Pasquali, R. (2009), "Stress-Related Development of Obesity and Cortisol in Women," *Obesity (Silver Spring)*, 17, 1678–1683. [1115]

Vinkers, C. H., Kalafateli, A. L., Rutten, B. P., Kas, M. J., Kaminsky, Z., Turner, J. D., and Boks, M. P. (2015), "Traumatic Stress and Human DNA Methylation: A Critical Review," *Epigenomics*, 7, 593–608. [1110]

Wang, L., Kim, Y., and Li, R.(2013), "Calibrating Nonconvex Penalized Regression in Ultra-High Dimension," *Annals of Statistics* 41, 2505–2536. [1112]

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. [1111]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1113]

Zhou, R. X., Wang, L. W., and Zhao, S. H. (2020), "Estimation and Inference for the Indirect Effect in High-Dimensional Linear Mediation Models," *Biometrika*, 107, 573–589. [1111,1112,1113,1116,1117,1118,1119,1120]