Heterogeneous Mediation Analysis for Causal Inference

Fei Xue[1], Xiwei Tang[2], Grace Kim[3], Allison Aiello[4], Karestan Koenen[5], Sandro Galea[6], Derek Wildman[7], Monica Uddin[7], and Annie Qu[8]

[1]University of Pennsylvania

[2]University of Virginia

[3]University of Illinois College of Medicine

[4]The University of North Carolina at Chapel Hill

[5]Harvard T.H. Chan School of Public Health

[6]Boston University

[7]University of South Florida

[8]University of California Irvine

Abstract

Mediation analysis is widely used to understand mediating mechanisms of variables in causal inference. However, existing approaches do not consider heterogeneity in mediation effects. Mediators in different sub-populations could have opposite effects on the outcome, and could be difficult to identify under the homogeneous model framework. In this paper, we propose a new mediator selection method, which can identify sub-populations and select mediators in each sub-population for heterogeneous data simultaneously. We perform a multi-directional clustering analysis to determine sub-group mediators and the corresponding subjects. Specifically, to select mediators, we propose a new joint penalty which penalizes

the effect of independent variable on a mediator and the effect of a mediator on the response jointly. The proposed algorithm is implemented through the convex-smooth gradient descent. Our numerical studies show that the proposed method outperforms the existing methods for heterogeneous data. We also apply the proposed mediation method to estimate mediation effects of DNA methylation variations in glucocorticoid receptor regulatory network genes for post-traumatic stress disorder (PTSD) among African-Americans. Based on data from the Detroit Neighborhood Health Study, we have found heterogeneous mediators, which are indeed associated with PTSD or traumatic experiences according to literature but were not selected by existing homogeneous mediation selection methods.

Key words: Post-traumatic stress disorder, DNA methylation, clustering, variable selection, linear structural equation modeling, difference of convex

1 Introduction

Causal mediation analysis aims to identify and explain the underlying relationship between an independent variable, a outcome, and mediator or intermediary variables. The independent variable could have a direct causal effect on the outcome, while it could also influence the mediator variable which in turn influences the outcome. Mediation analysis is widely used in social sciences to understand mediating mechanisms of variables with scientific meanings. Moreover, after identifying true mediators, we can prevent or facilitate the occurrence of outcome through controlling the mediator variables, even when we cannot change the independent variable.

Causal mediation analysis is typically formulated under the framework of linear structural equation modeling (LSEM) (Gunzler et al., 2013), including mediator models and a outcome model. The mediator models treat mediators as response variables and only have one predictor, that is, the independent variable. The outcome model use the outcome variable as the response and treat all mediators and the independent variable as covariates. (Serang et al., 2017) propose an exploratory mediation analysis approach via regularization, which first includes all potential predictors with the regularized structural equation modeling (Jacobucci et al., 2016), selects mediators

with nonzero specific indirect effect, and then refit a structural equation modeling with only the selected mediators. (Boca et al., 2013) propose to test multiple mediators simultaneously while maintain a desired family wise error rate. However, they assume that all the potential mediators are parallel and do not affect one another. (Imai et al., 2010b) propose identifiability assumptions and sensitivity analysis for causally related multiple mediators, while (Jerolon et al., 2018) extend the assumptions and corresponding theoretical results to uncausally correlated multiple mediators. (VanderWeele and Vansteelandt, 2014) propose to incorporate treatment-mediator interactions and mediator-mediator interactions in the outcome model.

To account for high-dimensional data, (Zhao and Luo, 2016) consider mediation pathway selection for a large number of causally dependent mediators, and propose a sparse mediation model using a regularized SEM approach, which introduces a pathway Lasso penalty. The pathway Lasso penalty shrinks estimate for pathway (mediation) effect of each mediator which is a product of coefficients, as well as estimates for individual coefficients. In addition, (Van Kesteren and Oberski, 2018) propose an exploratory coordinate-wise mediation filter approach, which considers correlated mediator and applies a decision function to each mediator conditional on the set of currently selected mediators. They utilize the Sobel test (Sobel, 1986), which tests significance of mediation effect (product of coefficients) under linear regression models, as a decision function. Moreover, (Zhang et al., 2016) consider high-dimensional mediation effects in a DNA methylation (DNAm) study. Specifically, they first adopt the sure independent screening (SIS) (Fan and Lv, 2008) to screen potential mediators, then utilize the minimax concave penalty (MCP) (Zhang et al., 2010) to estimate coefficients of potential mediators which are screened out, and finally conduct a joint significant test based on the MCP estimate to identify significant mediators.

However, these approaches do not consider heterogeneity in mediation effects. Mediators in different types of sub-populations could be different or have different effects on the outcome. For example, as illustrated in Figure 1, we treat trauma as an independent variable, Post-Traumatic Stress Disorder CheckList (PCL) score as a outcome, and DNAm probes as mediator variables. The trauma variable has positive effect on DNAm mediators, and the mediators also have positive

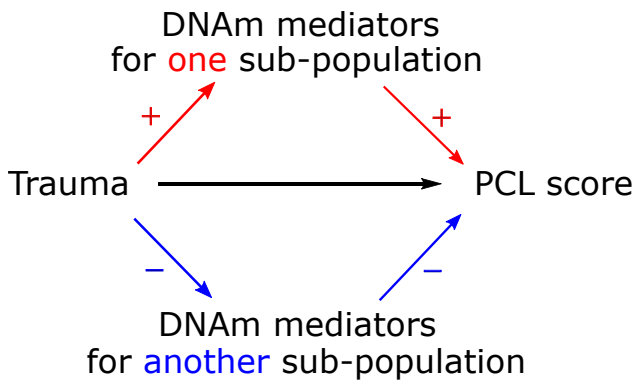

Figure 1: Heterogeneous mediation

effects on PCL score in one sub-population. In contrast, effects in another sub-population are all negative. If we analyze these mediators homogeneously, the mediators are hard to be identified since the effects could be canceled out. (Qin and Hong, 2017) develop a weighting method to identify and estimate site-specific heterogeneous direct and indirect effects, which adopts inverse-probability-of-treatment weight (Rosenbaum, 1987) and ratio-of-mediator-probability weighting (Hong et al., 2015) in each site to relax treatment and mediator assumptions. Nevertheless, the method is not applicable in general especially when the sub-populations are hidden. (Dyachenko and Allenby, 2015) propose a Bayesian mixture model, which combines likelihood functions based on two different outcome models to incorporate heterogeneity, and computes the probability that a specific respondent follows a mediated path. However, they only consider one single mediator variable.

In this paper, we propose a new mediator selection method, which can identify sub-populations and select mediators in each sub-population from high-dimensional data simultaneously. Specifically, we utilize the sum of squared residuals of a subject in mediator models and outcome model of a sub-population as a distance between the subject and the sub-population. For each subject, we find the smallest distance between this subject and all the sub-populations to determine which sub-population the subject should belong to. We estimate parameters for each sub-population based on subjects which are identified in the sub-population. We iteratively identify subjects and estimate parameters in each sub-population. On the other hand, we allow to detect homogeneous predictors

through fused Lasso penalty between parameters of each predictor in different sub-populations.

To select mediators instead of just variables, we propose a new joint penalty which penalizes effect from the independent variable to a mediator (independent-mediator effect) and effect from the mediator to the outcome (mediator-outcome effect) together. The penalization decreases when at least one of independent-mediator effect and mediator-outcome effect is large, which is inspired by the fact that the mediation effect increases when at least one of the two effects increases.

To implement the proposed method, we propose a difference of convex-smooth gradient descent (DC-SmGD) algorithm. Specifically, we approximate the proposed non-convex objective function through a series of convex functions, replace the non-smooth fused Lasso penalty by a smooth approximation function, and minimize the approximated objective function via gradient descent. Numerical studies show that the proposed method performs better than existing methods for heterogeneous data.

The paper is organized as follows. In Section 2, we introduce the counterfactual framework for the mediation problem. In Section 3, we propose the heterogeneous mediation method. In Section 4, the implementation and corresponding algorithm are illustrated. Sections 5 provides numerical studies through simulations.In Sections 6, we apply the proposed method to a real PTSD data.

2 Causal mediation under the counterfactual framework

Let T_i be an independent variable and $\boldsymbol{X}_i$ be a $r \times 1$ vector of pre-treatment confounders (e.g., race or gender) for the i-th subject and $1 \leq i \leq N$. Denote potential mediators by $\boldsymbol{M}_i = (M_{i1}, \ldots, M_{ip})^T$. The mediator could be affected by the independent variable T_i, that is $\boldsymbol{M}_i = \boldsymbol{M}_i(T_i)$. Let $Y_i(T_i, \boldsymbol{M}_i(T_i))$ be the outcome. Throughout the paper, we make the stable unit treatment value assumption (SUTVA) (Rubin, 1980), that is, the potential outcomes of one subject is unaffected by the assignment of treatments of other subjects. Without loss of generality, we also assume that the outcome and all the variables are centered.

If T_i is a binary treatment, the causal joint mediation effect or indirect effect for the i-th subject

is

$$\delta_i(t) = Y_i(t, \boldsymbol{M}_i(1)) - Y_i(t, \boldsymbol{M}_i(0)).$$

The direct effect of the treatment is $\zeta_i(t) = Y_i(1, \boldsymbol{M}_i(t)) - Y_i(0, \boldsymbol{M}_i(t))$, while the total effect of the treatment is $\tau_i = Y_i(1, \boldsymbol{M}_i(1)) - y_i(0, \boldsymbol{M}_i(0))$. We introduce the no-interaction assumption (Imai et al., 2010a), which says that there is no interaction term between treatment and mediators in model for the outcome i.e., $\delta_i = \delta_i(0) = \delta_i(1)$ and $\zeta_i = \zeta_i(0) = \zeta_i(1)$. Then, under this assumption, we have $\tau_i = \delta_i + \zeta_i$.

3 Heterogeneous causal mediation

Traditional methods average over all subjects and estimate the average causal mediation effects (Imai et al., 2010a). However, there might be heterogeneous mediation effects possibly due to racial, age or trauma variations in subjects. In this paper, we propose a new approach to identify heterogeneous mediation effects.

Suppose that the whole population can be partitioned into H subgroups, where mediators and average mediation effects within each subgroup are homogeneous. Denote the index set for subjects in the h-th subgroup by $\mathcal{S}(h)$. Then, the average joint mediation effect and direct effect in the h-th subgroup are $\bar{\delta}^{(h)}(t) = E\{\delta_i(t)\}$ and $\bar{\zeta}^{(h)}(t) = E\{\zeta_i(t)\}$, respectively, where $i \in \mathcal{S}(h)$.

3.1 Identification

To identify valid mediators and causal joint mediation effect, following (Jerolon et al., 2018), we propose a heterogeneous sequential ignorability for multiple mediators assumption (HSIMMA):

$$\{Y_i(t', m), \boldsymbol{M}_i^{(h)}(t)\} \perp\!\!\!\perp T_i | \boldsymbol{X}_i = \boldsymbol{x},$$

and

$$Y_i(t', m) \perp\!\!\!\perp \boldsymbol{M}_i^{(h)}(t) | T_i, \boldsymbol{X}_i = \boldsymbol{x},$$

for each h, $i \in \mathcal{S}(h), t', t$ and m, where $\boldsymbol{M}_i^{(h)}(t)$ represents true mediators in the h-th subgroup. Under HSIMMA, the Group h average mediation effect and direct effect can be identified non-parametrically as

$$\begin{aligned}\bar{\delta}^{(h)}(t) &= \int \Big\{ \int E(Y_i|\boldsymbol{M}_i^{(h)} = \boldsymbol{m}, T_i = t, \boldsymbol{X}_i = \boldsymbol{x}) dF_{\boldsymbol{M}_i^{(h)}|T_i=1,\boldsymbol{X}_i=\boldsymbol{x}} \\ &\quad - \int E(Y_i|\boldsymbol{M}_i^{(h)} = \boldsymbol{m}, T_i = t, \boldsymbol{X}_i = \boldsymbol{x}) dF_{\boldsymbol{M}_i^{(h)}|T_i=0,\boldsymbol{X}_i=\boldsymbol{x}} \Big\} dF_{\boldsymbol{X}_i=\boldsymbol{x}},\end{aligned}$$

$$\begin{aligned}\bar{\zeta}^{(h)}(t) &= \int \int \Big\{ E(Y_i|\boldsymbol{M}_i^{(h)} = \boldsymbol{m}, T_i = 1, \boldsymbol{X}_i = \boldsymbol{x}) \\ &\quad - E(Y_i|\boldsymbol{M}_i^{(h)} = \boldsymbol{m}, T_i = 0, \boldsymbol{X}_i = \boldsymbol{x}) \Big\} dF_{\boldsymbol{M}_i^{(h)}|T_i=t,\boldsymbol{X}_i=\boldsymbol{x}} dF_{\boldsymbol{X}_i=\boldsymbol{x}},\end{aligned}$$

for $i \in \mathcal{S}(h)$, where $F_{Z|W}$ and F_Z are the distribution function of a random variable Z and the conditional distribution function of Z given W, respectively.

3.2 Linear structural equation modeling under heterogeneity

We consider the problem with following linear structural equations:

$$\boldsymbol{M}_i = \boldsymbol{b}_h T_i + \boldsymbol{\Gamma} \boldsymbol{X}_i + \boldsymbol{\varepsilon}_{iM}, \tag{1}$$

$$Y_i = \beta_h T_i + \boldsymbol{\theta}_h^T \boldsymbol{M}_i + \boldsymbol{\gamma}^T \boldsymbol{X}_i + \varepsilon_i, \tag{2}$$

for $i \in \mathcal{S}(h)$, where $\boldsymbol{b}_h, \boldsymbol{\varepsilon}_{iM}, \boldsymbol{\theta}_h \in \mathbb{R}^p$, $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}$, and $\boldsymbol{\gamma} \in \mathbb{R}^r$. Here, β_h represents the average direct effect of the treatment on the outcome in the h-th subgroup, $\boldsymbol{\theta}_h^T \boldsymbol{b}_h$ represents the average joint mediation (indirect) effect of the treatment, $\boldsymbol{b}_h$ is the parameter relating the treatment to potential mediators, and $\boldsymbol{\varepsilon}_{iM} \sim N(\boldsymbol{0}, \Sigma_h)$ and $\varepsilon_i \sim N(0, \sigma_h^2)$ are error terms corresponding to the i-th sample, which are uncorrelated with T_i and $\boldsymbol{X}_i$. We assume that the model (2) is sparse, that is, most true values of $\boldsymbol{\theta}_h$ are zero for each h. There might be common mediators in multiple subgroups. The coefficients in $\boldsymbol{\theta}_h$ corresponding to $\boldsymbol{M}_i^{(h)}$ are non-zero. We assume that the ef-

fect of pre-treatment confounders $\boldsymbol{X}_i$ on each mediator or the outcome is homogeneous, that is, coefficients for $\boldsymbol{X}_i$ do not vary across all the subgroups.

Our goal is to find the subgroups, identify true mediators in each subgroup from high-dimensional potential mediators, and estimate all the parameters.

3.3 Residual-based loss function

Let $\boldsymbol{\Theta}_{2h} = (\beta_h, \boldsymbol{b}_h^T, \boldsymbol{\theta}_h^T)^T$ and $\boldsymbol{\Theta}_3 = (\boldsymbol{\gamma}, \boldsymbol{\Gamma}^T)$ representing parameters for mediators and confounders, respectively, for the h-th subgroup. The residual-based unpenalized loss function for the i-th sample with parameters of the h-th subgroup is

$$\begin{aligned} l_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3) &= \text{tr}\{\boldsymbol{W}_M(\boldsymbol{M}_i - \boldsymbol{b}_h T_i - \boldsymbol{\Gamma}\boldsymbol{X}_i)(\boldsymbol{M}_i - \boldsymbol{b}_h T_i - \boldsymbol{\Gamma}\boldsymbol{X}_i)^T\} \\ &\quad + w_y(Y_i - \beta_h T_i - \boldsymbol{\theta}_h^T \boldsymbol{M}_i - \boldsymbol{\gamma}^T \boldsymbol{X}_i)^2, \end{aligned} \tag{3}$$

where $\boldsymbol{W}_M$ is a positive-definite weighting matrix for the mediator models, and w_y is the weight for the outcome model. Here, $\boldsymbol{W}_M$ is set the importance of mediator models and their relationship, while w_y can adjust the proportion between mediator models and the outcome model. Without loss of generality, we can assume that the sum of diagonal elements equals p. In implementation, we set the diagonal elements of $\boldsymbol{W}_M$ to be 1 since we do not have prior information about the mediators. If the values of diagonal elements are not all the same, we should re-weight the corresponding penalty accordingly, e.g. $\|\boldsymbol{b}_h\|$.

Then, the loss function for all the subjects is

$$L_0(\boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3) = \sum_{1 \le h \le H} \sum_{i \in \mathcal{S}(h)} L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3) = \sum_{1 \le i \le N} \sum_{1 \le h \le H} I\{i \in \mathcal{S}(h)\} L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3), \tag{4}$$

where $I\{\cdot\}$ is an indicator function, $\boldsymbol{\Theta}_2 = (\boldsymbol{\Theta}_{2h}, \ldots, \boldsymbol{\Theta}_{2H})$, and $L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3) = l_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3)$. However, we do not know which subjects are in the h-th subgroup and which are not. To estimate the subgroup label for the i-th subject, we find the subgroup with smallest $L_{i,h}$ among all

subgroups. Then, the loss function becomes

$$L_1(\mathbf{\Theta}_2, \mathbf{\Theta}_3) = \sum_{1 \le i \le N} \min_{1 \le h \le H} \{L_{i,h}(\mathbf{\Theta}_{2h}, \mathbf{\Theta}_3)\}. \tag{5}$$

3.4 Penalties

We propose the following within-group penalty:

$$p_w(\mathbf{\Theta}_{2h}) = p_{SCAD,\lambda_1,a}(|\beta_h|) + \lambda_1 p_m(\boldsymbol{b}_h, \boldsymbol{\theta}_h), \tag{6}$$

where

$$p_m(\boldsymbol{b}_h, \boldsymbol{\theta}_h) = \sum_{j=1}^{p} \left(\frac{1}{c_0} - \frac{1}{c_0(1 + c_0|b_{h,j}|)(1 + c_0|\theta_{h,j}|)} \right), \tag{7}$$

$p_{SCAD,\lambda_1}(\cdot)$ is the SCAD penalty with tuning parameters λ_1 and a, c_0 is a constant to adjust the shrinkage, $\theta_{h,j}$ and $b_{h,j}$ are the j-th elements of $\boldsymbol{\theta}_h$ and $\boldsymbol{b}_h$, respectively. We propose the new two-dimensional joint mediation penalty $p_m(\boldsymbol{b}_h, \boldsymbol{\theta}_h)$ in (7) to identify mediators with large mediation effects. Figure 2 is a graph of the mediation penalty function. We can see that the penalty tends to shrink values around axes to zero, and that the shrinkage gradually levels off as $\theta_{h,1}$ or $b_{h,1}$ increases.

To avoid high-dimensional confounders, we propose confounders penalty function as follows:

$$p_c(\mathbf{\Theta}_3) \quad = \quad \lambda_3 \left\{ w_y \|\boldsymbol{\gamma}\|_1 + \sum_{j=1}^{r} \|\mathbf{\Gamma}_j\|_1 \right\}, \tag{8}$$

where λ_3 is a tuning parameter, and $\mathbf{\Gamma}_j$ denotes the j-th column of $\mathbf{\Gamma}$. To identify homogeneous covariates, we consider the following between-group fused lasso penalties

$$p_b(\mathbf{\Theta}_2) \quad = \quad \lambda_4 \sum_{1 \le h_1, h_2 \le H} \left[w_y |\beta_{h_1} - \beta_{h_2}| + \sum_{j=1}^{p} \{ w_y |\theta_{h_1,j} - \theta_{h_2,j}| + |b_{h_1,j} - b_{h_2,j}| \} \right], \tag{9}$$

where λ_4 is a tuning parameter.

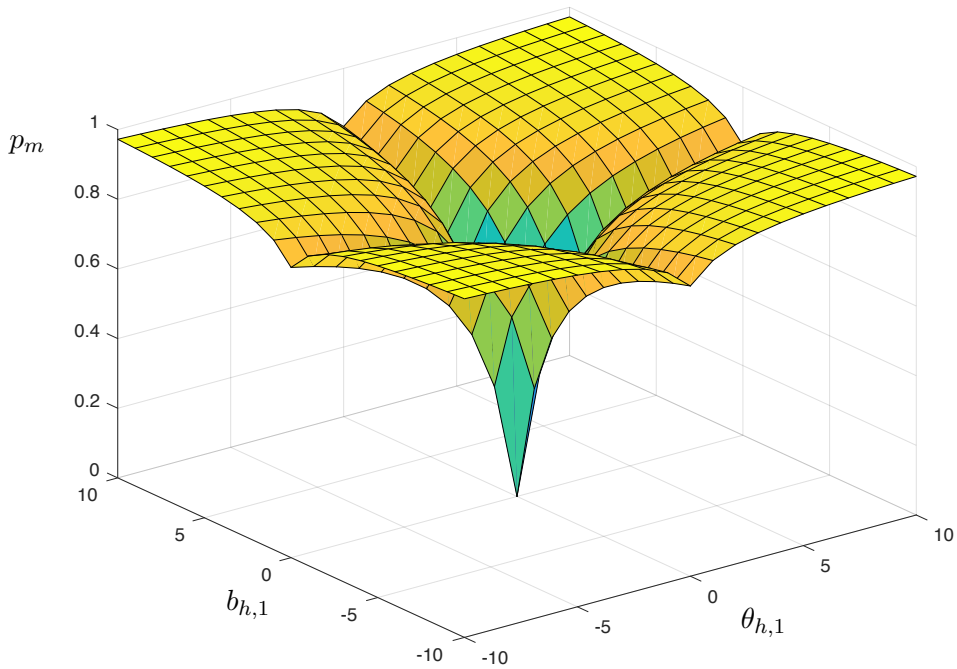

Figure 2: Mediation penalty with $p = 1$, $\lambda_1 = 1$, and $c_0 = 0.5$.

Thus, the final objective function of the proposed method is

$$f(\boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3) = L_1(\boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3) + N \left\{ \sum_{1 \le h \le H} p_w(\boldsymbol{\Theta}_{2h}) + p_b(\boldsymbol{\Theta}_2) + p_c(\boldsymbol{\Theta}_3) \right\}. \tag{10}$$

4 Algorithm

4.1 Difference of convex for within-group penalty

Note that the within-group penalty in (6) is not convex. Let

$$p_d(\boldsymbol{\Theta}_{2h}) = \frac{\beta_h^2}{a-1} + 2c_0\lambda_1(\|\boldsymbol{\theta}_h\|_2^2 + \|\boldsymbol{b}_h\|_2^2). \tag{11}$$

Then, $p_d(\boldsymbol{\Theta}_{2h}) + p_w(\boldsymbol{\Theta}_{2h})$ is convex for each $1 \le h \le H$. The objective function f can be decomposed as

$$f(\boldsymbol{\nu}) = f_1(\boldsymbol{\nu}) - f_2(\boldsymbol{\nu}), \tag{12}$$

where $\boldsymbol{\nu}$ is a vector of all parameters,

$$f_1(\boldsymbol{\nu}) = \sum_{i=1}^{N}\sum_{h=1}^{H} L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3) + N\left\{\sum_{h=1}^{H}\{p_d(\boldsymbol{\Theta}_{2h}) + p_w(\boldsymbol{\Theta}_{2h})\} + p_b(\boldsymbol{\Theta}_2) + p_c(\boldsymbol{\Theta}_3)\right\},$$

and

$$f_2(\boldsymbol{\nu}) = \sum_{i=1}^{N}\max_{1\le k\le H}\left\{\sum_{1\le h\le H, h\ne k} L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3)\right\} + N\sum_{h=1}^{H} p_d(\boldsymbol{\Theta}_{2h}).$$

Since f_1 and f_2 are both convex, f is a difference of convex (DC) function.

We adopt a DC-SNadam type algorithm. Given $\boldsymbol{\nu}^{(m-1)}$, at the m-th iteration, let $\boldsymbol{\mu}^{(m-1)} \in \partial f_2(\boldsymbol{\nu}^{(m-1)})$. We calculate

$$\begin{aligned}\partial f_2(\boldsymbol{\nu}) &= \sum_{i=1}^{N}\partial\left\{\max_{1\le k\le H} F_{ik}(\boldsymbol{\nu})\right\} + N\sum_{h=1}^{H}\partial p_d(\boldsymbol{\Theta}_{2h}) \\ &= \sum_{i=1}^{N}\mathrm{co}\left\{\cup_{k\in J(\boldsymbol{\nu})}\partial F_{ik}(\boldsymbol{\nu})\right\} + N\sum_{h=1}^{H}\partial p_d(\boldsymbol{\Theta}_{2h}) \qquad (13)\end{aligned}$$

where $F_{ik}(\boldsymbol{\nu}) = \sum_{1\le h\le H, h\ne k} L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3)$, $J(\boldsymbol{\nu}) = \{1 \le k \le H : F_{ik}(\boldsymbol{\nu}) = \max_{1\le k\le H} F_{ik}(\boldsymbol{\nu})\}$, and co stands for the convex hull. Based on the idea of DC algorithm, at this iteration, f_2 can be replaced by its affine minorization

$$f_2^{(m)}(\boldsymbol{\nu}) = f_2(\boldsymbol{\nu}^{(m-1)}) + \left\langle \boldsymbol{\nu} - \boldsymbol{\nu}^{(m-1)}, \boldsymbol{\mu}^{(m-1)}\right\rangle. \qquad (14)$$

4.2 Smooth approximation for between-group penalty

Note that $p_b(\boldsymbol{\Theta}_2)$ is non-smooth and non-separable, and that

$$p_b(\boldsymbol{\Theta}_2) = p_b(\boldsymbol{\nu}) = \lambda_4\|\boldsymbol{D}\boldsymbol{\nu}\|_1 = \lambda_4 \max_{\|\boldsymbol{\eta}\|_\infty\le 1} \boldsymbol{\eta}^T\boldsymbol{D}\boldsymbol{\nu}$$

can be reformulated as a maximization problem, where $\boldsymbol{D}$ is a difference operator. We consider to replace it by a smooth approximation

$$\tilde{p}_b(\boldsymbol{\nu};\rho) = \lambda_4 \max_{\|\boldsymbol{\eta}\|_\infty \le 1} \left(\boldsymbol{\eta}^T \boldsymbol{D}\boldsymbol{\nu} - \frac{\rho}{2}\|\boldsymbol{\eta}\|_2^2 \right), \tag{15}$$

where ρ is a positive smoothing parameter. Following (Chen et al., 2012), we choose $\rho = 10^{-4}$ in implementation. The function $\tilde{p}_b(\boldsymbol{\nu};\rho)$ approximates $p_b(\boldsymbol{\nu})$ as $\rho \to 0$ (Nesterov, 2005). Let $\boldsymbol{\eta}^* = S(\boldsymbol{D}\boldsymbol{\nu}/\rho)$, where

$$S(x) = \begin{cases} x, & -1 \le x \le 1, \\ 1, & x > 1, \\ -1, & x < -1. \end{cases}$$

Then, $\tilde{p}_b(\boldsymbol{\nu};\rho) = \lambda_4 \left\{ (\boldsymbol{\eta}^*)^T \boldsymbol{D}\boldsymbol{\nu} - \frac{\rho}{2}\|\boldsymbol{\eta}^*\|_2^2 \right\}$ is convex and differentiable with respect to $\boldsymbol{\nu}$ (Chen et al., 2012). We have

$$\frac{\partial \tilde{p}_b}{\partial \boldsymbol{\nu}}(\boldsymbol{\nu};\rho) = \lambda_4 \boldsymbol{D}^T \boldsymbol{\eta}^*. \tag{16}$$

Therefore, at the m-th iteration, we replace the f_2 and p_b in f by $f_2^{(m)}$ and $\tilde{p}_b$, respectively, and

$$\boldsymbol{\nu}^{(m)} = \underset{\boldsymbol{\nu}}{\operatorname{argmin}} \tilde{f}^{(m)}(\boldsymbol{\nu}), \tag{17}$$

where

$$\begin{aligned} \tilde{f}^{(m)}(\boldsymbol{\nu}) \;=\; & \sum_{i=1}^{N}\sum_{h=1}^{H} L_{i,h}(\boldsymbol{\Theta}_{2h}, \boldsymbol{\Theta}_3) + N \sum_{h=1}^{H} \{p_d(\boldsymbol{\Theta}_{2h}) + p_w(\boldsymbol{\Theta}_{2h})\} \\ & + N \{\tilde{p}_b(\boldsymbol{\nu};\rho) + p_c(\boldsymbol{\Theta}_3)\} - f_2^{(m)}(\boldsymbol{\nu}). \end{aligned} \tag{18}$$

We solve (17) through the gradient descent (GD) algorithm. At the k-th iteration in GD,

$$\boldsymbol{g}^{(k)} = \nabla \tilde{f}^{(m)}(\boldsymbol{\nu}^{(k-1)}), \tag{19}$$

$$\boldsymbol{\nu}^{(k)} = \boldsymbol{\nu}^{(k-1)} - t\boldsymbol{g}^{(k)}, \tag{20}$$

where t is the step size. We adopt the backtracking line search to determine the step size. Specifically, we start from $t = 1$, and repeat $t = \tau_1 t$ until

$$\tilde{f}^{(m)}(\boldsymbol{\nu}^{(k-1)} - t\boldsymbol{g}^{(k)}) < \tilde{f}^{(m)}(\boldsymbol{\nu}^{(k-1)}) - \tau_2 t\{\boldsymbol{g}^{(k)}\}^T \boldsymbol{g}^{(k)},$$

where $\tau_1 \in (0, 1)$ and $\tau_2 \in (1, 1/2)$. Most items in (18) are differentiable respect to $\boldsymbol{\nu}$ except for $p_w(\boldsymbol{\Theta}_{2h})$ and $p_c(\boldsymbol{\Theta}_3)$.

$$\frac{\partial p_w}{\partial \beta_h}(\boldsymbol{\Theta}_{2h}) = \lambda_1\{I(\beta_h \leq \lambda_1) + \frac{(a\lambda_1 - \beta_h)_+}{(a-1)\lambda_1} I(\beta_h > \lambda_1)\} \operatorname{sign}(\beta_h), \tag{21}$$

$$\frac{\partial p_w}{\partial \theta_{h,j}}(\boldsymbol{\Theta}_{2h}) = \lambda_1 \frac{\operatorname{sign}(\theta_{h,j})}{(1 + c_0|b_{h,j}|)(1 + c_0|\theta_{h,j}|)^2} \tag{22}$$

$$\frac{\partial p_w}{\partial b_{h,j}}(\boldsymbol{\Theta}_{2h}) = \lambda_1 \frac{\operatorname{sign}(b_{h,j})}{(1 + c_0|b_{h,j}|)^2(1 + c_0|\theta_{h,j}|)} \tag{23}$$

Based on the first derivatives, we can see that the penalty decreases as the absolute value of $|b_{h,j}|$ or $|\theta_{h,j}|$ increases. The complete algorithm is in Algorithm 1.

Algorithm 1: DC-SmGD

1. Set number of subgroups H, tolerance ϵ, and the tuning parameter λ.
2. Obtain initial subgroup labels $c_i^{(0)}$ for $i = 1, \ldots, N$, based on k-means clustering method.
3. Calculate $\boldsymbol{\nu}^{(0)}$ based on $c_i^{(0)}$ for each subgroup separately.
4. At the m-th iteration, given $\boldsymbol{\nu}^{(m-1)}$ from the $(m-1)$-th iteration:
   (a) Calculate $\boldsymbol{\mu}^{(m-1)} \in \partial f_2(\boldsymbol{\nu}^{(m-1)})$.
   (b) Solve $\boldsymbol{\nu}^{(m)}$ in (17) through the gradient descent: Update $\boldsymbol{g}^{(k)}$ and $\boldsymbol{\nu}^{(k)}$ in (19)-(20) iteratively until converge.
   (c) Re-calculate $\boldsymbol{\nu}^{(m)}$ through (13).
5. Iterate Step 4 until the convergence criterion is satisfied, e.g., $\|\boldsymbol{\nu}^{(m)} - \boldsymbol{\nu}^{(m-1)}\|_1 < \epsilon$.

4.3 Tuning parameter selection

We propose a Bayesian information criterion (BIC) type criterion to tune $\boldsymbol{\lambda}$ and the the number of clusters:

$$\begin{aligned} BIC(\boldsymbol{\lambda}, H) &= N \cdot \log\left\{RSS_M(\boldsymbol{\lambda}, H)/N\right\} + df_M(\boldsymbol{\lambda}, H) \cdot \log(N) \\ &\quad + w_y\left[N \cdot \log\left\{RSS_Y(\boldsymbol{\lambda}, H)/N\right\} + df_Y(\boldsymbol{\lambda}, H) \cdot \log(N)\right], \end{aligned} \tag{24}$$

where $\boldsymbol{\lambda}$ consists all the tuning parameters, $df_M(\boldsymbol{\lambda}, H)$ and $df_Y(\boldsymbol{\lambda}, H)$ are numbers of non-zero estimated coefficients in mediator models and the outcome model with tuning parameters $\boldsymbol{\lambda}$, respectively,

$$RSS_M\left(\boldsymbol{\lambda}, H\right) = \sum_{1 \le h \le H} \sum_{\hat{c}_{i,\boldsymbol{\lambda}} = h} (\boldsymbol{M}_i - \hat{\boldsymbol{b}}_{h,\boldsymbol{\lambda}} T_i - \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\lambda}} \boldsymbol{X}_i)^T \boldsymbol{W}_M (\boldsymbol{M}_i - \hat{\boldsymbol{b}}_{h,\boldsymbol{\lambda}} T_i - \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{\lambda}} \boldsymbol{X}_i),$$

and

$$RSS_Y(\boldsymbol{\lambda}, H) = \sum_{1 \le h \le H} \sum_{\hat{c}_{i,\boldsymbol{\lambda}} = h} (Y_i - \hat{\beta}_{h,\boldsymbol{\lambda}} T_i - \hat{\boldsymbol{\theta}}_{h,\boldsymbol{\lambda}}^T \boldsymbol{M}_i - \hat{\boldsymbol{\gamma}}_{\boldsymbol{\lambda}}^T \boldsymbol{X}_i)^2. \tag{25}$$

Here, we only count once for estimated homogeneous variables in df_M and df_Y, and w_y is utilized to adjust the proportion of information from mediator models and the outcome model.

5 Simulation

In this section, we provide simulation studies to compare the proposed method with existing homogeneous mediation methods. The simulation results show that the proposed method achieves higher mediation effect estimation accuracy and mediator selection accuracy than the homogeneous method. We simulate data following models in (1) and (2) with $r = 0$ using 100 replications.

The proposed method is implemented based on Algorithm 1 with $w_y = 1$, $c_0 = 10$ and $a = 3.7$. We apply the "HIMA" package (https://cran.r-project.org/web/packages/HIMA/index.html) in R to

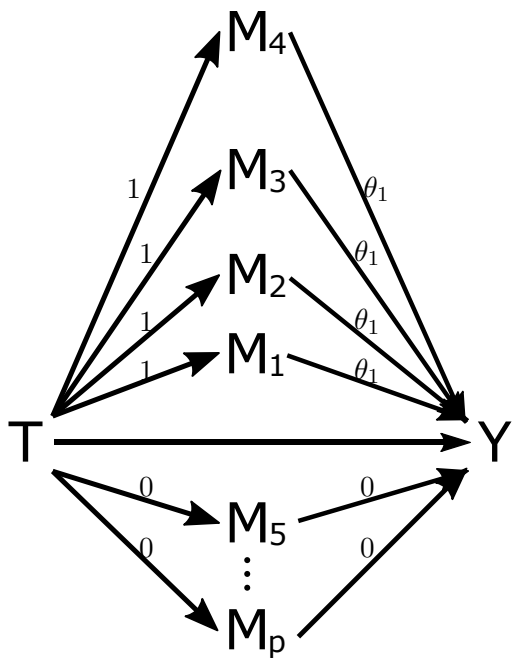

Figure 3: True coefficients for the homogeneous setting

implement the high-dimensional mediation analysis (HIMA) method for comparison, which is a homogeneous mediation approach. For each method, we calculate the average of all individuals' mediator false negative rates (FN) and the average of all individuals' mediator false positive rates (FP) for mediator selection. We also evaluate each method via the average of all individuals' mean-squared-errors (MSE) of mediation effects.

We consider the following settings.

Setting 1**:** $N = 200$, $H = 1$, $n_1 = 200$, $p = 30$. True coefficients in the model are shown in Figure 3, where $\beta_1 = 0.5$, $b_1 = 1$, and $\theta_1 = 0.2, 0.3,$ or 0.8. The results are in Table 1.

Setting 2**:** $N = 200$, $H = 2$, $n_1 = 50$, $n_2 = 150$, $p = 30$. True coefficients in the model are shown in Figure 4, where $\beta_1 = 0.5, b_1 = 1$, $\beta_2 = -0.5, b_2 = -1$, $\theta_1 = 0.5, 1,$ or 4 and $\theta_2 = -0.5, -1,$ or -4. The results are in Table 2.

Setting 3**:** $N = 100$, $H = 2$, $n_1 = 30$, $n_2 = 70$, $p = 150$. True coefficients in the model are shown in Figure 4, where $\beta_1 = b_1 = 1$, $\beta_2 = b_2 = -1$, $\theta_1 = 0.5, 0.8, 1,$ or 4 and $\theta_2 = -0.5, -0.8, -1,$ or -4. The results are in Table 3.

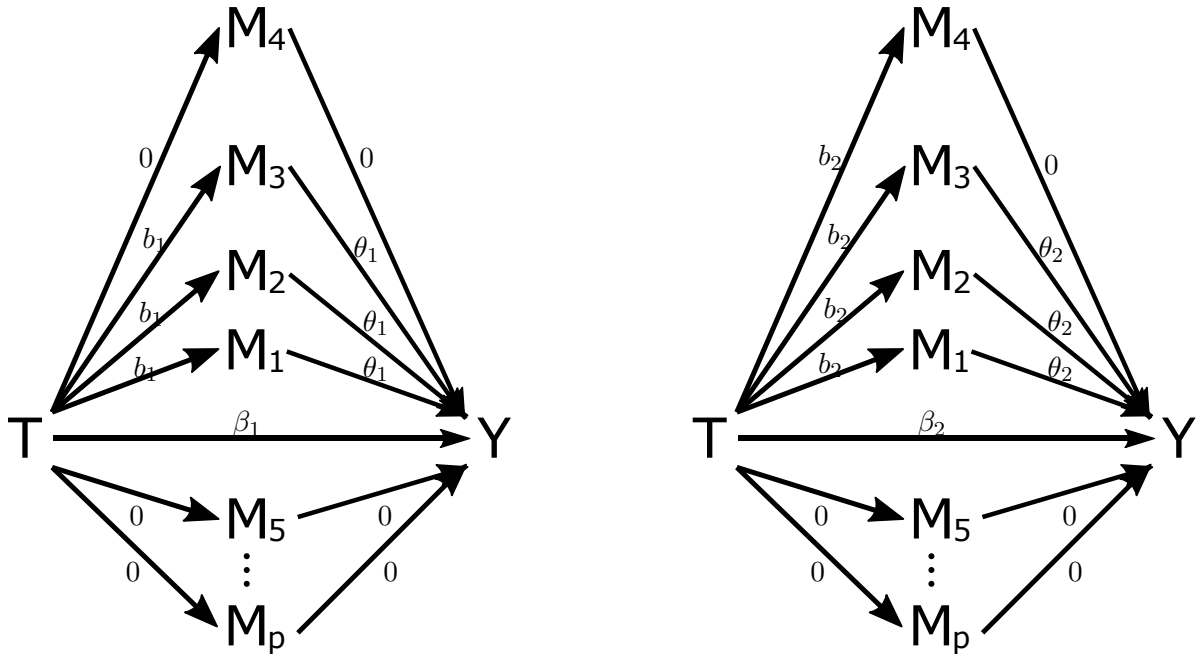

Figure 4: True coefficients for the two sub-populations

Table 1: Results of Setting 1.

θ_1	Method	FN	FP	FN+FP	MSE
0.2	**Proposed**	0.046	0.002	0.048	0.002
	HIMA	0.717	0.023	0.740	0.005
0.3	**Proposed**	0.000	0.000	0.000	0.002
	HIMA	0.188	0.117	0.305	0.005
0.8	**Proposed**	0.000	0.000	0.000	0.002
	HIMA	0.000	0.135	0.135	0.005

Table 2: Results of Setting 2.

(θ_1, θ_2)	Method	FN	FP	FN+FP	MSE
(0.5, -0.5)	**Proposed**	0.031	0.007	0.038	0.004
	HIMA	0.979	0.005	0.984	0.025
(1, -1)	**Proposed**	0.002	0.000	0.002	0.006
	HIMA	0.973	0.006	0.979	0.100
(4, -4)	**Proposed**	0.001	0.000	0.001	0.058
	HIMA	0.805	0.009	0.814	1.569

Table 3: Results of Setting 3.

(θ_1, θ_2)	Method	FN	FP	FN+FP	MSE
(0.5, -0.5)	**Proposed**	0.530	0.001	0.531	0.004
	HIMA	0.907	0.002	0.909	0.005
(0.8, -0.8)	**Proposed**	0.220	0.000	0.220	0.005
	HIMA	0.983	0.003	0.986	0.013
(1, -1)	**Proposed**	0.143	0.001	0.144	0.007
	HIMA	0.990	0.001	0.991	0.020
(4, -4)	**Proposed**	0.092	0.001	0.093	0.068
	HIMA	0.986	0.001	0.987	0.320

Simulation results in Tables 1-3 show that the proposed method produces smaller overall false negative rate plus false positive rate in mediator selection, indicating that the proposed method selects mediator more accurately. Moreover, the proposed method produces smaller MSE of mediation effects, indicating that the proposed method is also more effective in estimation of mediation effects.

6 Real data application

Socially adverse or traumatic experiences of discrimination and isolation appear to be high among African Americans, which could have a pronounced negative impact on mental and physical health. One of the common mental health problems caused by traumatic events is post traumatic stress disorder (PTSD). Studies show that its pathophysiology is associated with DNA methylation (DNAm) in glucocorticoid receptor regulatory network (GRRN) genes, since the process is intrinsically linked to gene regulation (Rusiecki et al., 2013). However, it is unknown whether DNAm mediates the relationship between experiences of social adversity and development of PTSD.

In this section, we apply the proposed method to study the mediation effects of DNAm variations of GRRN genes on PTSD symptom severity using the Detroit Neighborhood Health Study (DNHS) data (`https://dnhs.unc.edu/`). The DNHS is a representative longitudinal cohort study with five waves and a majority of African Americans in Detroit. One main goal of this study

is to investigate the interrelationship between ecologic stressors and the risk of PTSD.

In this paper, we utilize the baseline waves (wave 1 or 2) for our mediation analysis. The DNHS study measures the PTSD symptom severity of participants through a widely-used self-report PTSD Checklist, Civilian Version (PCL-C) Blanchard et al. (1996); Grubaugh et al. (2007). The PCL-C set contains 17 items corresponding to key symptoms of PTSD. Participants indicate how much they have been bothered by each symptom using a 5-point (1–5) scale in reference to their worst traumatic experience. For total severity, we calculate the average of the 17 items and treat the average PCL-C score as a representative of the PTSD symptom severity for each participant. Moreover, the DNHS study records types of traumatic or stressful events that each participant has ever experienced. We calculate the total number of all the experienced event types as a trauma feature representing overall trauma severity for each subject. The DNHS study also collects demographic information and blood specimens for genetic testing of DNA.

In total, there are 125 subjects in the DNHS study who have available PTSD measurements, trauma type numbers, and thousands of DNAm CpG probes ($n = 125$). We screen DNAm CpG probes via an expression quantitative trait methylation (eQTM) analysis on GRRN-annotated DNAm CpG probes and 53 expressed GRRN genes. We find 144 CpG probes significantly correlated with the GRRN genes at significance level 0.05. Thus, we treat the PCL-C score as a dependent variable and the the total number of trauma exposures as an independent variable, and select key mediators between the independent and dependent variables from the 144 DNAm CpG probes.

To compare the performance of the proposed method with existing homogeneous methods, we randomly split the data into a test set (10% of all samples) and a training set (90% of all samples) 100 times. For each method, we train the model on the training data, and calculate the prediction mean-squared error (PMSE) of the dependent variable in the test set. For each subject in the test set, to identify the sub-population label in the proposed method, we calculate the average PMSE of mediators based on estimated parameters in each sub-population. The smallest one corresponds the sub-population that the subject should belong to.

Table 4 shows that the mean PMSE of the proposed method is much smaller than that of the existing HIMA method, which indicates that the proposed method is more accurate in mediation effect estimation. The standard deviation of PMSE of the proposed method is also smaller, indicating that the proposed method is more stable.

In addition, the HIMA does not detect any DNAm probe, while the proposed method selects 10 probes on average. We provide the most frequently selected 10 DNAm probes and corresponding gene names in Table 5. These genes are indeed associated with PTSD or traumatic experiences according to existing literature (Breen et al., 2019; Criado-Marrero et al., 2018; Kim et al., 2019; Raabe and Spengler, 2013). For example, the FKPB5 gene is significantly associated with stress reactivity and PTSD risk. In addition, the FKPB5 is a key modulator of the stress response acting as a co-chaperone modulating glucocorticoid receptor activity in response to stressors (Criado-Marrero et al., 2018; Zannas et al., 2016). Moreover, the NR3C1 gene expression has been found up-regulated in patients with PTSD (Ramírez et al., 2020). In addition, maternal PTSD severity and parenting stress are negatively correlated with the mean percentage of methylation of NR3C1 (Schechter et al., 2015).

Table 4: Real data results. The "NS" represents the mean number of selected mediators. The "Mean" and "SD" represent mean and standard deviation of PMSE based on 100 replications, respectively.

Method	NS	Mean	SD
Proposed	10	1.908	1.091
HIMA	0	7.708	1.741

In summary, the proposed method produces smaller MSE for prediction in test sets than homogeneous mediation methods, indicating that our method achieves better performance in estimation of mediation effects. Moreover, the proposed method identifies important mediators, which are not detected by existing methods.

Table 5: Most frequently selected CpG probes and corresponding gene names

CpG probes (frequency)	Gene names
cg00773696 (72)	*FOS*
cg16012111 (67)	*FKBP5*
cg02675550 (58)	*NFATC1*
cg00431336 (45)	*NFATC1*
cg03995361 (35)	*VIPR1*
cg13025668 (29)	*POMC*
cg02432888 (28)	*SMARCA4*
cg13409449 (26)	*GATA3*
cg13456321 (25)	*NFATC1*
cg20387815 (25)	*POMC*

References

Blanchard, E. B., Jones-Alexander, J., Buckley, T. C., and Forneris, C. A. (1996). Psychometric properties of the ptsd checklist (pcl). *Behaviour research and therapy*, 34(8):669–673.

Boca, S. M., Sinha, R., Cross, A. J., Moore, S. C., and Sampson, J. N. (2013). Testing multiple biological mediators simultaneously. *Bioinformatics*, 30(2):214–220.

Breen, M. S., Bierer, L. M., Daskalakis, N. P., Bader, H. N., Makotkine, I., Chattopadhyay, M., Xu, C., Grice, A. B., Tocheva, A. S., Flory, J. D., et al. (2019). Differential transcriptional response following glucocorticoid activation in cultured blood immune cells: a novel approach to ptsd biomarker development. *Translational psychiatry*, 9(1):1–13.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., Xing, E. P., et al. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752.

Criado-Marrero, M., Rein, T., Binder, E. B., Porter, J. T., Koren III, J., and Blair, L. J. (2018). Hsp90 and fkbp51: complex regulators of psychiatric diseases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1738):20160532.

Dyachenko, T. and Allenby, G. M. (2015). Bayesian analysis of heterogeneous mediation.

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

Grubaugh, A. L., Elhai, J. D., Cusack, K. J., Wells, C., and Frueh, B. C. (2007). Screening for ptsd in public-sector mental health settings: the diagnostic utility of the ptsd checklist. *Depression and Anxiety*, 24(2):124–129.

Gunzler, D., Chen, T., Wu, P., and Zhang, H. (2013). Introduction to mediation analysis with structural equation modeling. *Shanghai archives of psychiatry*, 25(6):390.

Hong, G., Deutsch, J., and Hill, H. D. (2015). Ratio-of-mediator-probability weighting for causal mediation analysis in the presence of treatment-by-mediator interaction. *Journal of Educational and Behavioral Statistics*, 40(3):307–340.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science* 51–71.

Jacobucci, R., Grimm, K. J., and McArdle, J. J. (2016). Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal*, 23(4):555–566.

Jerolon, A., Baglietto, L., Birmele, E., Perduca, V., and Alarcon, F. (2018). Causal mediation analysis in presence of multiple mediators uncausally related. *arXiv preprint arXiv:1809.08018*.

Kim, G., Aiello, A., Koenen, K., Galea, S., Wildman, D., and Uddin, M. (2019). 27. differential methylation at glucocorticoid-relevant regulatory regions associated with ptsd in african americans. *Biological Psychiatry*, 85(10):S11–S12.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152.

Qin, X. and Hong, G. (2017). A weighting method for assessing between-site heterogeneity in causal mediation mechanism. *Journal of Educational and Behavioral Statistics*, 42(3):308–340.

Raabe, F. J. and Spengler, D. (2013). Epigenetic risk factors in ptsd and depression. *Frontiers in psychiatry*, 4:80.

Ramírez, C. G., Queijeiro, A. V., Morales, S. J., López, D. B., Miranda, A. H., Chow, A. R., Cárdenas, L. T., and Estrada, M. G. (2020). The nr3c1 gene expression is a potential surrogate biomarker for risk and diagnosis of posttraumatic stress disorder. *Psychiatry Research*, 284:112797.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rusiecki, J. A., Byrne, C., Galdzicki, Z., Srikantan, V., Chen, L., Poulin, M., Yan, L., and Baccarelli, A. (2013). PTSD and DNA methylation in select immune function gene promoter regions: a repeated measures case-control study of us military service members. *Frontiers in Psychiatry*, 4:56.

Schechter, D. S., Moser, D. A., Paoloni-Giacobino, A., Stenz, L., Gex-Fabry, M., Aue, T., Adouan, W., Cordero, M. I., Suardi, F., Manini, A., et al. (2015). Methylation of nr3c1 is related to maternal ptsd, parenting stress and maternal medial prefrontal cortical activity in response to child separation among mothers with histories of violence exposure. *Frontiers in psychology*, 6:690.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Serang, S., Jacobucci, R., Brimhall, K. C., and Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural equation modeling: a multidisciplinary journal*, 24(5):733–744.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. *Sociological methodology*, 16:159–186.

Van Kesteren, E.-J. and Oberski, D. L. (2018). Exploratory mediation analysis with many potential mediators. *arXiv preprint arXiv:1810.06334*.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

Zannas, A. S., Wiechmann, T., Gassen, N. C., and Binder, E. B. (2016). Gene–stress–epigenetic regulation of fkbp5: clinical and translational implications. *Neuropsychopharmacology*, 41(1):261–274.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154.

Zhao, Y. and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749*.