A Selective Review of High-dimensional Mediation Analyses in DNA Methylation Studies

Haixiang Zhang[1], Lifang Hou[2], and Lei Liu[3*]

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

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

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

Abstract. DNA methylation alterations have been widely studied as mediators of environmentally-induced disease risks. With new advances in technique, epigenome-wide DNA methylation data (EWAS) have become the new standard for epigenetic studies in human populations. However, to date most epigenetic studies of mediation effects only involve selected (gene-specific) candidate methylation markers. There is an urgent need for appropriate analytical methods for EWAS mediation analysis. In this paper, we provide an overview of recent advances on high-dimensional mediation analysis, with application to two DNA methylation data.

Keywords. Multiple comparison; False discovery rate; Variable selection; Regularization; Joint significance test; Mediation analysis; Epigenetics.

1 Introduction

DNA methylation (DNAm) is a major epigenetic regulator of gene expression (El-Osta and Wolffe 2001; Herman and Baylin 2003;, Esteller 2007). It stands at the intersection of genetic and environmental risk factors for disease, and is critical for improved risk prediction and understanding of the biology of chronic diseases as health care transitions to a new era of precision medicine (Feinberg and Fallin 2015). Unlike genetic variation, which is static throughout the life course, environmental factors and human behaviors can induce changes in DNAm. These epigenetic changes may serve as mediating factors in the causal pathway from exposure or treatment to health outcomes. More importantly, these changes can also be modified or even reversed through preventive and therapeutic interventions (Cortessis et al. 2012).

Mediation analysis plays an important role in the social and behavioral sciences (Baron and Kenny 1986; MacKinnon 2008; Preacher and Hayes 2008; Kenny 2008). The main

*Corresponding author: lei.liu@wustl.edu (L. Liu)

goal of mediation analysis is to investigate whether the effect of an independent variable on a dependent variable is at least partially transmitted through an intermediate variable (*mediator*). For more related literatures, we refer to the monographs (MacKinnon 2008; Hayes 2013; VanderWeele 2015) and the review articles (MacKinnon et al. 2007; Wood, et al. 2008; Ten Have and Joffe 2012; Richiardi et al. 2013; Wang and Sobel 2013; Preacher 2015; VanderWeele 2016; Richmond et al. 2016).

Currently, most mediation studies of DNAm only involve candidate (gene-specific) methylation markers (Bellavia et al. 2013; Tarantini et al. 2013, Bind et al. 2014). Recent advances in measurement techniques, such as Illumina Infinium platforms, have resulted in epigenome-wide DNAm data (EWAS) becoming the standard for studies of epigenetics in human populations. A motivating example is an epigenome-wide DNA methylation study (Zhang et al. 2016), where some of roughly 480K probes on DNA methylation markers could be potential mediators between the exposure (smoking) and the health outcome (lung function). These high-dimensional EWAS data pose great challenges for data analyses (particularly mediation analyses), for which appropriate analytical methods are urgently needed.

Several papers, e.g., Liu et al. (2013), have proposed high-dimensional mediation analysis methods in the framework of adjusting for multiple comparisons. These methods considered each exposure-DNAm mediator relation and each DNAm mediator-outcome relation separately, adjusting for multiple comparisons by Bonferroni's approach or false discovery rate (FDR). However, as shown in Figure 2 below, multiple mediators can lead to the same outcome, meaning that it is necessary to adjust for other mediators when assessing the effect of a given individual mediator. Furthermore, these methods cannot be used for predicting multifactorial disease risk, e.g., by developing a prediction index based on more than one DNAm markers.

To address these gaps in the literature, Zhang et al. (2016) proposed to use the sure independent screening (SIS; Fan and Lv 2008) and minimax concave penalty (MCP; Zhang 2010) based joint significance test approach. There are also other related results on high-dimensional mediation analysis. For example, Huang and Pan (2016) proposed a transformation model using spectral decomposition to test the mediation effects of high-dimensional continuous mediators. Zhao and Luo (2016) proposed a sparse high-dimensional mediation model by introducing a new penalty called Pathway Lasso. Chén et al. (2018) introduced a novel direction of mediation approach by linearly combining potential mediators into a smaller number of orthogonal components in the high-dimensional setting. Wu et al. (2018) studied the mediation effects of DNA methylation between alcohol consumption and epithelial ovarian cancer using high-dimensional logistic regression.

In this paper, we will review the recent advances on high-dimensional mediation analysis,

with application to DNA methylation studies. The remainder of this Chapter is organized as follows. In Section 2, we give the definition of a mediation model with a single mediator, and review some traditional methods to assess the mediation effect. In Section 3, we briefly present a multiple mediators model, together with some recent advances. In Section 4, we pay attention to several new developments on high-dimensional mediation analysis. In Section 5, we showed the application of two selected methods to real data analysis. Some concluding remarks are reported in Section 6.

2 Single mediator model

To characterize the path-specific effect of an exposure on an outcome that is mediated through a mediator (in Figure 1), we consider the three-variable regression equation (MacKinnon, 2008):

$$\begin{aligned} Y &= i_1 + \gamma^* X + e_1, \\ M &= i_2 + \alpha X + e_2, \\ Y &= i_3 + \gamma X + \beta M + e_3, \end{aligned} \tag{2.1}$$

where X is the independent variable (*exposure*), M is the mediator, Y is the dependent variable (*outcome*); i_1, i_2 and i_3 are intercepts, γ^* represents the *total effect* of the independent variable X on the dependent variable Y; γ is the "*direct effect*" of X on Y adjusted for the mediator M; α is the path coefficient relating X and Y; β is the path coefficient relating the mediator M to the dependent variable Y adjusted for X; e_1, e_2 and e_3 are error terms. It is straightforward to derive that the γ^* (*total effect*) is equal to γ (*direct effect*) plus $\alpha\beta$ (*indirect effect*).

To assess whether there exists an indirect effect from X to Y that is mediated by M, a popular technique is the product of coefficients approach, most well known as the Sobel test (Sobel 1982),

$$H_0 : \alpha\beta = 0 \text{ vs. } H_A : \alpha\beta \neq 0. \tag{2.2}$$

The test statistic for (2.2) is given as $\hat{S} = \hat{\alpha}\hat{\beta}/\hat{\sigma}_{\alpha\beta}$, where $\hat{\alpha}$, $\hat{\beta}$, $\hat{\sigma}^2_\alpha$ and $\hat{\sigma}^2_\beta$ are ordinary least squares (OLS) estimates, and $\hat{\sigma}_{\alpha\beta} = \sqrt{\hat{\alpha}^2\hat{\sigma}^2_\beta + \hat{\beta}^2\hat{\sigma}^2_\alpha}$ is derived from the delta method. By Sobel (1982), the asymptotic distribution of $\hat{S}$ is $N(0,1)$. Thus, the p-value is $P_{sobel} = 2\{1 - \Phi(|\hat{S}|)\}$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. Of note, the Sobel test requires the assumption that the sampling distribution of the indirect effect is normal. However, the product of two normal variables tends to be asymmetric with nonzero

skewness and kurtosis, and the performance of Sobel test is usually conservative (Hayes 2009). Another common approach is the joint significance test (Taylor et al. 2008), where the p-value of test (2.2) is given as $P_{joint} = \max(P_a, P_b)$ with $P_a = 2\{1 - \Phi(|\hat{\alpha}|/\hat{\sigma}_\alpha)\}$ and $P_b = 2\{1 - \Phi(|\hat{\beta}|/\hat{\sigma}_\beta)\}$. That is to say, the joint significance test requires that both α and β are significant simultaneously. Moreover, there also exist some alternative mediation testing methods, e.g. difference in coefficients (MacKinnon et al., 2002), distribution of the product (Williams and MacKinnon, 2008), resampling methods (Preacher and Hayes, 2008), and permutation methods (Taylor and MacKinnon, 2012).

3 Multiple mediators model

In practice, there may exist multiple mediators on the causal pathway between an exposure and an outcome (in Figure 2). To describe this causal relationship, we consider the following multiple mediators regression model (MacKinnon, 2008),

$$\begin{aligned} Y &= c^* + \gamma^* X + \epsilon_1, \\ M_k &= c_k + \alpha_k X + e_k, \quad k = 1, \cdots, p, \\ Y &= c + \gamma X + \beta_1 M_1 + \cdots + \beta_p M_p + \epsilon_2, \end{aligned} \tag{3.1}$$

where $\mathbf{M} = (M_1, \cdots, M_p)'$ is the vector of mediators; γ^* represents the relation between the X and Y in the "direct path", γ is the parameter relating X to Y adjusted for the effects of $\mathbf{M}$ in the "indirect path"; α_k is the parameter relating X to the mediating variable M_k, $k = 1, \cdots, p$; $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ is the vector of parameters relating the mediators to Y adjusted for the effects of X; c, c^*, and $\{c_k, k = 1, \cdots, p\}$ are intercept terms; ϵ_1, ϵ_2 and $\{e_k,\ k = 1, \cdots, p\}$ are residuals.

Let X_i, $\mathbf{M}_i = (M_{i1}, \cdots, M_{ip})'$ and Y_i be i.i.d. observations, $i = 1, \cdots, n$. Consider the multiple testing problem,

$$H_{0k} : \alpha_k \beta_k = 0 \text{ vs. } H_{Ak} : \alpha_k \beta_k \neq 0, \quad k = 1, \cdots, p. \tag{3.2}$$

For testing of (3.2), it can be performed with a univariate or multivariate approach. Here, the univariate approach analyzes each mediator separately using a marginal model $Y = c + \gamma X + \beta_k M_k + \epsilon_2$ (Barfield et al., 2017; Sampson et al., 2018). A major drawback of this naive univariate method is the neglect of other possible correlated mediators, which may result in biased estimates and efficiency loss. To solve this issue, the multivariate approach can improve power and accuracy (Boca et al. 2014), since it can adjust for confounding variables (other DNAm mediators) by including them in the model.

Boca et al. (2014) used the max correction (Westfall and Young 1993) and permutation to address the family wise error rate, which can be briefly described as follows: *Step 1*: Calculate the maximum-type test statistics $\hat{S} = \max_{1\leq k\leq p}\{|\hat{\alpha}_k\hat{\beta}_k|\}$, where $\hat{\alpha}_k$ and $\hat{\beta}_k$ are the OLS estimates in (3.1). *Step 2*: Permute X to obtain $\hat{\alpha}_k^*$, and get $\hat{\beta}_k^*$ by permuting the residual of regressing Y on E. Calculate the permutation statistics $\hat{S}^* = \max_{1\leq k\leq p}\{|\hat{\alpha}_k^*\hat{\beta}_k^*|\}$. *Step 3*: Repeat Step 2 to obtain a distribution of $\hat{S}^*$, and the 95th percentile of this distribution is denoted as $\mathcal{Q}_{0.95}$. We declare M_k to be significant if $|\hat{\alpha}_k\hat{\beta}_k| \geq \mathcal{Q}_{0.95}$, $k = 1, \cdots, p$. Of note, this permutation approach that focuses on the maximum of the test statistics can significantly improve the power to detect mediators over the Bonferroni-based multiple adjustment (Boca et al. 2014).

4 High-dimensional mediators model

As the number of mediators increasing, p may be larger than n, the multiple mediators model (3.1) can be generalized to the framework of high-dimensional mediation analysis. Below, we will review some recent advances on high-dimensional mediation effects for continuous outcome and binary outcome, respectively.

4.1 Continuous outcome

For high-dimensional linear regression, the ordinary least squares (OLS) estimate is not available since the number of mediators p is larger than the sample size n (Tibshirani et al., 2015). There are two approaches for these high-dimensional correlated mediators (in Figure 2): orthogonal transformation approach and the variable selection approach.

Huang and Pan (2016) and Chén et al. (2018) proposed to transform the original p mediators to be uncorrelated given the exposure such that we can evaluate the mediation effects using a series of single mediator models. More specifically, let $\tilde{\mathbf{M}} = F(\mathbf{M}) = (\tilde{M}_1, \cdots, \tilde{M}_p)'$ be the vector of new transformed variables, where $\mathbf{M} = (M_1, \cdots, M_p)$ is the vector of original mediators, $F(\cdot) : \mathbb{R}^p \mapsto \mathbb{R}^p$ is an orthogonal transformation. As suggested by Huang and Pan (2016) and Chén et al. (2018), we can assume the following three-variable regression model,

$$\begin{aligned} \tilde{M}_k &= \tilde{c}_k + \tilde{\alpha}_k X + \zeta_k,\ k = 1, \cdots, p, \\ Y &= \tilde{c} + \tilde{\gamma} X + \tilde{\beta}_1\tilde{M}_1 + \cdots + \tilde{\beta}_p\tilde{M}_p + \epsilon, \end{aligned} \tag{4.1}$$

where ϵ and $\{\zeta_k,\ k = 1, \cdots, p\}$ are random error terms. The orthogonal transformation $F(\cdot)$ plays a key role in this method, we can use the spectral decomposition (Huang and Pan,

2016) or the directions of mediation (Chén et al., 2018) as the transformation for the original mediators. Because the new transformed variables $\tilde{M}_k$'s are orthogonal, we can estimate the parameters in (4.1) separately for each $\tilde{M}_k$ using marginal models, $k = 1, \cdots, p$.

However, the orthogonal transformation approach cannot evaluate the contribution from each individual mediator since the transformed variable $\tilde{M}_k$ is a linear combination of the original p mediators. To tackle this issue, Zhang et al. (2016) used the sure independent screening (SIS; Fan and Lv 2008) and minimax concave penalty (MCP; Zhang 2010) to reduce the dimension of mediators, and adopt the joint significance test procedure. The proposed method in Zhang et al. (2016) for mediation analyses has been implemented with the R package `HIMA`. We summarize the details as follows:

Step 1.(*Screening*). Use the SIS (Fan and Lv 2008) to identify a subset $\mathcal{I} = \{1 \leq k \leq p : M_k$ is among the top $d = [2n/\log(n)]$ largest effects for the response $Y\}$.

Step 2. (*MCP-penalized estimate*). Compute $\{\hat{\beta}_k, k \in \mathcal{I}\}$ by minimizing the MCP penalized criterion,

$$Q^{mcp} = \sum_{i=1}^{n} \left(Y_i - c - \gamma X_i - \sum_{k \in \mathcal{I}} \beta_k M_{ik} \right)^2 + \sum_{k \in \mathcal{I}} p_{\lambda,\delta}(\beta_k), \tag{4.2}$$

where $p_{\lambda,\delta}(\cdot)$ is the minimax concave penalty:

$$p_{\lambda,\delta}(\beta_k) = \lambda \left[|\beta_k| - \frac{|\beta_k|^2}{2\delta\lambda} \right] I\{0 \leq |\beta_k| < \delta\lambda\} + \frac{\lambda^2\delta}{2} I\{|\beta_k| \geq \delta\lambda\}. \tag{4.3}$$

Here $\lambda > 0$ is the regularization parameter, and $\delta > 0$ determines the concavity of MCP.

Step 3. (*Joint significance test*). Let $\mathcal{S} = \{k : \hat{\beta}_k \neq 0\}$, which is based on the MCP-penalized estimate in Step 2. The p-value for the joint significance test is given as

$$P_{joint,k} = \max(P_{1k}, P_{2k}),$$

with

$$P_{1k} = \min(|\mathcal{S}| \cdot 2\{1 - \Phi(|\hat{\beta}_k|/\hat{\sigma}_{\beta_k})\}, 1)$$

and

$$P_{2k} = \min(|\mathcal{S}| \cdot 2\{1 - \Phi(|\hat{\alpha}_k|/\hat{\sigma}_{\alpha_k})\}, 1),$$

where $|\mathcal{S}|$ is the number of variables in $\mathcal{S}$, $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. Here $\hat{\beta}_k$ is the MCP estimate in (4.2), whose standard error $\hat{\sigma}_{\beta_k}$ can be obtained from the oracle property of MCP (Zhang 2010); $\hat{\alpha}_k$ is the ordinary least square estimator for α_k, and $\hat{\sigma}_{\alpha_k}$ is the corresponding estimated standard error.

4.2 Binary outcome

To explore the mediation mechanism on binary outcome, Wu et al. (2018) adopted the causal inference test (CIT; Millstein et al. 2009) together with counterfactual mediation procedure in VanderWeele and Vansteelandt (2013). Here we summarize their method in details as follows:

Step 1. (X *is associated with* Y). A logistic regression model is fitted to examine the association between the exposure X and the binary outcome Y with $\text{logit}\{P(Y=1)\} = c^* + \gamma^* X$. In addition, we consider the hypothesis testing $H_0 : \gamma^* = 0$ vs. $H_A : \gamma^* \neq 0$.

Step 2. (M_k *is associated with* Y *conditional on* X). We fit all the mediators M_k into one single multiple logistic regression model conditional on X as $\text{logit}\{P(Y=1)\} = c + \gamma X + \beta_1 M_1 + \cdots + \beta_p M_p$. Since the number of mediators p is much larger than the sample size n, the traditional maximum likelihood does not work for this testing task $H_{0k} : \beta_k = 0$, $k = 1, \cdots, p$. To solve this problem, Wu et al. (2018) proposed to use the de-sparsified Lasso estimator $\hat{\beta}_k$, where van de Geer et al. (2014) have proved the asymptotic normality for $\hat{\beta}_k$. The corresponding p-value $P_k^{(b)}$ is adjusted for multiple testing by the Bonferroni correction. Denote $\mathcal{S}_1 = \{k; P_k^{(b)} < 0.05\}$ as the significant variables in the mediator-outcome causal pathways.

Step 3. (X *is associated with* M_k *conditional on* Y, $k \in \mathcal{S}_1$). The identified significant variables M_k in Step 2 are subsequently regressed on X given Y as $M_k = c_k + \alpha_k X + \eta_k Y + e_k$, $k \in \mathcal{S}_1$. Consider the testing $H_0 : \alpha_k = 0$, and index of significant variables is denoted as $\mathcal{S}_2$.

Step 4. (Y *is independent of* X *conditional on* $\{M_k, k \in \mathcal{S}_2\}$). To check if the outcome Y is independent of X conditional on those significant mediators identified in Step 3, we fit the following logistic regression model

$$\text{logit}\{P(Y=1)\} = c + \gamma X + \sum_{k \in \mathcal{S}_2} \beta_k M_k.$$

Consider the testing $H_0 : \gamma = 0$ vs. $H_A : \gamma \neq 0$. To get the p-value, we can use a bootstrap type approach in Millstein et al. (2009).

Step 5. (*Validation of the CIT results*). To further validate the identified potentially significant mediators by the CIT approach in Steps 1-4, Wu et al. (2018) used the causal multiple mediators framework of VanderWeele and Vansteelandt (2013).

Of note, *Steps 1-4* in the framework of CIT ensures that the effects of X on Y are wholly transmitted through the mediators. However, some effect is likely to impose on Y directly from X, rather than be transmitted indirectly by mediators. In other words, the CIT-based method can only tackle whole-mediation effects. Moreover, as pointed out by Wu et al. (2018), the procedure in *Step 5* regards multiple mediators as joint mediators, hence

it is impossible to weigh the relative importance of individual mediators.

5 Applications

5.1 Normative Aging Study

The first application is the US Department of Veterans Affairs Normative Aging Study, which is an ongoing longitudinal cohort of elderly, predominantly white American veterans (NAS, Spiro and Vokonas 2007). In 1963, 2280 men aged 21 to 80 years and free of hypertension or other chronic conditions were enrolled. Between January 1, 1999 and December 31, 2013, 686 were randomly selected and had blood samples profiled using the Illumina Infinium 450K BeadChip DNA methylation array. Zhang et al. (2016) studied the mechanism of how these methylation markers mediate the relationship between smoking (measured in pack-years) and lung function, which is measured by 4 outcomes: FEV1 (forced expiratory volume in 1 second), FVC (forced expiratory vital capacity), FEV1/FVC, and MMEF (maximum mid expiratory flow). After excluding subjects with lung-related diseases, e.g., asthma, emphysema, and COPD, a sample size of 290 was used in the analysis. The proper temporal relationship (exposure $\rightarrow$ methylation $\rightarrow$ outcome) was ensured by taking the appropriate temporal order of measurement for smoking, DNAm, and lung function. They also adjusts for age, height, and weight in each equation of Model (3.1).

From 486K CpGs, they used Model (3.1) and identified two CpGs as mediators associated with at least one lung function outcome. Specifically, cg05575921 (in AHRR) was associated with FEV1, FVC, and FEV1/FVC. Methylation at this site was previously shown to be a sensitive marker of smoking history (Harlid et al. 2014; Gao et al. 2016). Another CpG, cg24859433 in the intergenic region 6p21.33, was associated with MMEF and also previously associated with smoking (Zeilinger et al. 2013; Ambatipudi et al. 2016). Thus, the overlap between our EWAS results and the current literature demonstrates the validity of this approach. On the other hand, the naive test (Liu et al. 2013) with Bonferroni's adjustment failed to identify any significant mediators.

Zhang et al. (2016) also calculated the extent to which the total effect is mediated through methylation markers, defined as $\alpha_k\beta_k/\gamma^*$ for each CpG site (in the last column of Table 4 of Zhang et al. 2016). CpG cg05575921 mediates about 50% of the total effect of smoking on both FEV1 and FVC, and 40% on FEV1/FVC; while cg24859433 mediates 16% of the total effect of smoking on MMEF.

5.2 Epithelial ovarian cancer

The second application is from the Mayo Clinic Ovarian Cancer Case-Control Study, with 196 cases and 202 age-matched controls ($n = 398$). Data include alcohol consumption (X), DNAm markers (M; the total number p= 25926), epithelial ovarian cancer status (Y). Epithelial ovarian cancer (EOC) is the leading cause of gynecologic cancer death in the United States (Morgan et al., 2011). Bagnardi et al. (2001) showed that a higher daily alcohol intake (100 g/day) is a risk factor for EOC. Philibert et al. (2012) found that alcohol consumption is associated with changes in DNA methylation, and Shen et al. (2013) showed that DNA methylation alterations could represent a mechanism of epithelial ovarian cancer risk. A natural question arises on whether the effect of alcohol consumption on epithelial ovarian cancer is mediated by DNA methylation.

To identify those potential mediators, Wu et al. (2018) adopted the *five-step* procedure in Section 4.2. During the testing process, several covariates were included in the model, including the effects of estimated differential leukocyte cell counts, age, current smoking status, study enrollment year, location of residence, parity and age at first birth, and the first principal component representing within-European population sub-structure. They identified two CpG sites (cg09358725, cg11016563) that represent potential mediators of the relationship between alcohol consumption and EOC case-control status. However, it is impossible to assess the individual effects of cg09358725 and cg11016563, since the mediation testing method in Section 4.2 treats multiple mediators jointly.

6 Concluding remarks

Mediation analysis is often used to investigate the role of intermediate variables that lie on the causal path between an exposure and outcome. Until recently most of the mediation analysis methods have been restricted to a single mediator or multiple (yet low-dimensional) mediators. In this paper we briefly described some basic concepts and methods for single and multiple mediation models. Then we focused on the new developments for high-dimensional mediation analysis, with application in DNA methylation studies.

The research on mediation analysis can be roughly divided into two categories: structural equation modeling (SEM) and counterfactual frameworks. The SEM framework is mainly based on regression to describe the causal relation with the model coefficients interpreted as causal effects. Various topics under the SEM framework have been explored, e.g., Cheung (2007), Jo et al. (2011), Lindquist (2012), Enders et al. (2013), Zhang and Wang (2013), Fritz et al. (2016). The counterfactual approach devotes to decomposing the total effect into

direct and indirect effects in the framework of causal inference (Rubin, 1974). Examples include VanderWeele (2009), Imai (2010), Imai et al. (2010), Albert and Nelson (2011), Valeri and VanderWeele (2013), Albert and Wang (2015), Daniel et al. (2015), Wang and Albert (2017), among others. Tingley et al. (2014) developed an R package `mediation` for conducting counterfactual mediation analysis. The two methods in Section 4 (and correspondingly the two applications in Section 5) represent the exploration in each framework.

Of note, the SIS + MCP based procedure in Zhang et al. (2016) relies on variable screening and cleaning stage, and the screened-out mediators are excluded from the testing process. Therefore, this method may miss some potential mediators. A possible solution is to use the de-biased Lasso method (Zhang and Zhang 2014), and we will report this result in a forthcoming article.

Furthermore, we can consider mediation analysis for high-dimensional survival models, e.g., Rein (2017). The more sophisticated situation where exposures, mediators, and outcomes could be longitudinally measured is another topic of future interest.

Finally, although we reviewed the high dimensional variable selection methods in DNA methylation studies, these methods can be applied to other subject areas, e.g., microbiome studies (Tsilimigras and Fodor 2016; Sohn and Li, 2017; Xia and Sun, 2017; Zhang et al. 2018).

Acknowledgments

The work of Haixiang Zhang is partially supported by Science Foundation of Tianjin University (No. 2018XRG-0038). The work of Lei Liu is partially supported by the Washington University Institute of Clinical and Translational Sciences grant UL1TR000448 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH).

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44, 139-177.

Albert, J. and Nelson, S. (2011). Generalized causal mediation analysis. *Biometrics*, 67, 1028-1038.

Albert, J. and Wang, W. (2015). Sensitivity analyses for parametric causal mediation effect estimation. *Biostatistics*, 16, 339-351.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10, 1100-1120.

Bagnardi, V., Blangiardo, M., La Vecchia, C. and Corrao, G. (2001). A meta-analysis of alcohol drinking and cancer risk. *British Journal of Cancer*, 85, 1700-1705.

Barfield, R., Shen, J., Just, A., Vokonas, P., Schwartz, J., Baccarelli, A., VanderWeele, T. and Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic Epidemiology*, 41, 824-833.

Baron, R. and Kenny, D. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.

Bellavia, A., Urch, B., Speck, M., Brook, R., Scott, J., Albetti, B. et al. (2013). DNA hypomethylation, ambient particulate matter, and increased blood pressure: findings from controlled human exposure experiments. *Journal of the American Heart Association*, 2:e000212.

Bind, M., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A., Coull, B. et al. (2014). Air pollution and gene-specific methylation in the Normative Aging Study: association, effect modification, and mediation analysis. *Epigenetics*, 9,448-458.

Boca, S., Sinha, R., Cross, A., Moore, S. and Sampson, J. (2014). Testing multiple biological mediators simultaneously. *Bioinformatics*, 30, 214 - 220.

Chén, O., Crainiceanu, C., Ogburn, E., Caffo, B., Wager, T. and Lindquist, M. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19, 121-136.

Cheung, M. (2007). Comparison of approaches to constructing confidence intervals for mediating effects using structural equation models. *Structural Equation Modeling*, 14, 227-246.

Cortessis, V., Thomas, D., Levine, A., Breton, C., Mack, T., Siegmund, K., Haile, R. and Laird, P. (2012). Environmental epigenetics: prospects for studying epigenetic mediation of exposure-response relationships. *Human Genetics*, 131, 1565-1589.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 74, 187-220.

Daniel, R., Stavola, B., Cousens, S. and Vansteelandt, S. (2015). Causal mediation analysis with multiple mediators. *Biometrics*, 71, 1-14.

Dezeure, R., Bhlmann, P., Meier, L., Meinshausen, N. (2015). High-dimensional inference: confidence intervals, p-values and R software hdi. *Stat. Sci.* 30:53358.

El-Osta, A. and Wolffe, A. (2001). DNA methylation and histone deacetylation in the control of gene expression: basic biochemistry to human development and disease. *Gene expression*, 9, 63-75.

Enders, C., Fairchild, A. and MacKinnon, D. (2013). A Bayesian approach for estimating mediation effects with missing data. *Multivariate Behavioral Research*, 48, 340-369.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8, 286-298.

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.

Feinberg, A. and Fallin, M. (2015). Epigenetics at the Crossroads of Genes and the Environment. *JAMA*, 314, 1129-1130.

Fritz, M., Kenny, D. and MacKinnon, D. (2016). The combined effects of measurement error and omitting confounders in the single-mediator model. *Multivariate Behavioral Research*, 51, 681-697.

Fulcher, I., Tchetgen Tchetgen, E. and Williams, P. (2017). Mediation snalysis for censored survival data under an accelerated failure time model. *Epidemiology*, 28, 660-666.

Gao, X., Jia, M., Zhang, Y., Breitling, L. and Brenner, H. (2015). DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clinical Epigenetics*. 7:113.

Guo, S. and Zeng, D. (2014). An overview of semiparametric models in survival analysis. *Journal of Statistical Planning and Inference*,151-152, 1-16.

Hayes, A. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication monographs*, 76, 408-420.

Hayes, A. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: The Guilford Press.

Herman, J. and Baylin, S. (2003). Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine*, 349,2042-2054.

Huang, Y. and Cai, T. (2016). Mediation analysis for survival data using semiparametric probit models. *Biometrics*, 72, 563-574.

Huang, Y. and Pan, W. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72, 402-413.

Huang, Y. and Yang, H. (2017). Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology*, 28, 370-378.

Imai, K. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15, 309-334.

Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25, 51-71.

Jo, B., Stuart, E., MacKinnon, D. and Vinokur, A. (2011). The use of propensity scores in mediation analysis. *Multivariate Behavioral Research*, 46, 425-452.

Kalbfleisch, J. and Prentice, R.(2002). *The Statistical Analysis for Failure Time Data*. John Wiley and Sons, New York.

Kenny, D. (2008). Reflections on mediation. *Organizational Research Methods*, 11, 353-358.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature methods*, 7, 813-819.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81, 61-71.

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

Liu, Y., Aryee, M., Padyukov, L., Fallin, M., Hesselberg, E., Runarsson, A. et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31, 142-147.

MacKinnon, D., Lockwood, C., Hoffman, J., West, S. and Sheets,V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, **7**, 83 -104.

MacKinnon, D., Fairchild, A. and Fritz, M. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593-614.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*. New York: Erlbaum and Taylor Francis Group.

Millstein, J., Zhang, B., Zhu, J. and Schadt, E. (2009). Disentangling molecular relationships with a causal inference test. *BMC Genetics*, 10:23.

Morgan Jr, R., Alvarez, R., Armstrong, D., et al. (2011). Epithelial ovarian cancer. *Journal of the National Comprehensive Cancer Network*, 9, 82-113.

Philibert, R., Plume, J., Gibbons, F., Brody, G., Beach, S. (2012). The impact of recent alcohol use on genome wide DNA methylation signatures. *Frontiers in Genetics*, 3:54.

Preacher, K. and Hayes, A. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40, 879-891.

Preacher, K. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology*, 66, 825-852.

Rein, C. (2017). Identification of mediators in high dimensional survival data in the presence of confounding. Avaliable at `http://nbn-resolving.de/urn:nbn:de:bvb:19-epub-41010-0`

Richiardi, L., Bellocco, R. and Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *International Journal of Epidemiology*, 42, 1511-1519.

Richmond, R., Hemani, G., Tilling, K., Davey Smith, G. and Relton, C. (2016). Challenges and novel approaches for investigating molecular mediation. *Human Molecular Genetics*, 25, R149-R156.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688-701.

Sampson, J., Boca, S., Moore, S. and Heller, R. (2018). FWER and FDR control when testing multiple mediators. *Bioinformatics*. In press.

Shen, H., Fridley, B., Song, H., et al. (2013). Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. *Nature Communications*, 4, 1628.

Sohn, M. and Li, H. (2017). Compositional mediation analysis for microbiome studies. *bioRxiv 149419*, doi: https://doi.org/10.1101/149419

Sonnenburg, J. and Bäckhed, F. (2016). Diet-microbiota interactions as moderators of human metabolism. *Nature*, 535, 56-64.

Spiro, A. and Vokonas, P. (2007). Normative aging study. In K. Markides (Ed.), Encyclopedia of Health & Aging. (pp. 422-423). Thousand Oaks, CA: SAGE Publications, Inc.

Swenson, N. G. (2011). Phylogenetic beta diversity metrics, trait evolution and inferring the functional beta diversity of communities. *PLoS ONE*, 6(6), e21264.

Tarantini, L., Bonzini, M., Tripodi, A., Angelici, L., Nordio, F., Cantone, L. et al. (2013). Blood hypomethylation of inflammatory genes mediates the effects of metal-rich airborne pollutants on blood coagulation. *Occupational and Environmental Medicine*, 70,418-425.

Taylor, A., MacKinnon, D. and Tein, J. (2008). Tests of the three-path mediated effect. *Organizational Research Methods*, 11, 241-269.

Taylor, A. and MacKinnon, D. (2012). Four applications of permutation methods to testing a single-mediator model. *Behavior research methods*, 44, 806-844.

Ten Have, T. and Joffe, M. (2012). A review of causal estimation of effects in mediation analyses. *Statistical Methods in Medical Research*, 21, 77-107.

Tibshirani, R., Wainwright, M., Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. New York: Chapman and Hall/CRC.

Tingley, D., Yamamoto, T., Hirose, H., Keele, L. and Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, Vol. 59, Issue 5.

Tsilimigras, M. and Fodor, A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, 26, 330-335.

Valeri, L. and VanderWeele, T. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychological Methods*, 18, 137-150.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for highdimensional models. *Annals of Statistics*, 42, 1166-1202.

VanderWeele, T. (2009). Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*, 20, 18-26.

VanderWeele, T. and Vansteelandt, S. (2013). Mediation analysis with multiple mediators. *Epidemiologic Methods*, 2, 95-115.

VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

VanderWeele, T. (2016). Mediation analysis: A practitioner's guide. *Annual Review of Public Health*, 37, 17-32.

Wang, W. and Albert, J. (2017). Causal mediation analysis for the Cox proportional hazards model with a smooth baseline hazard estimator. *Journal of the Royal Statistical Society: Series C*, 66, 741-757.

Wang, X. and Sobel, M. (2013). New perspectives on causal mediation analysis. *In Handbook of Causal Analysis for Social Research (S. Morgan eds)*. Springer, 215-242.

Westfall, P. and Young, S. (1993) *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*. New York: Wiley-Interscience.

Williams, J. and MacKinnon, D. (2008). Resampling and distribution of the product methods for testing indirect effects in complex models. *Structural Equation Modeling*, 15, 23-51.

Wood, R., Goodman, J., Beckmann, N. and Cook, A. (2008). Mediation testing in management research: A review and proposals. *Organizational Research Methods*, 11, 270-295.

Wu, D., Yang, H., Winham, S., Natanzon, Y., Koestler, D., Luo, T., Fridley, B., Goode, E., Zhang, Y. and Cui, Y. (2018). Mediation analysis of alcohol consumption, DNA methylation, and epithelial ovarian cancer. *Journal of Human Genetics*, 63, 339-348.

Xia, Y. and Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases*, 4, 138-148.

Yuan, Y. and MacKinnon, D. (2014). Robust mediation analysis based on median regression. *Psychological Methods*, 19, 1-20.

Zeilinger,S., Kühnel, B., Klopp, N., et al. (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One*, 8, e63812.

Zeng, D. and Lin, D. Y. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102, 1387-1396.

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

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B*, 76, 217-242.

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

Zhang, J., Wei, Z. and Chen, J. (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*. In press.

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

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: methods and software. *Behavior Research Methods*, 46, 1184-1198.

Zhao, S. and Prentice, R. (2014). Covariate measurement error correction methods in mediation analysis with failure time data. *Biometrics*, 70, 835-844.

Zhao, Y. and Luo, X. (2016). Pathway Lasso: estimate and select sparse mediation pathways with high-dimensional mediators. *arXiv:1603.07749v1*, Preprint.

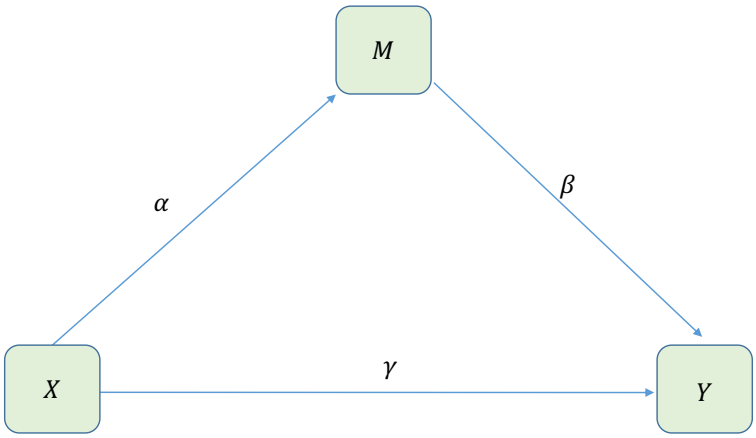

Figure 1. A scenario with a single mediator between exposure and outcome.

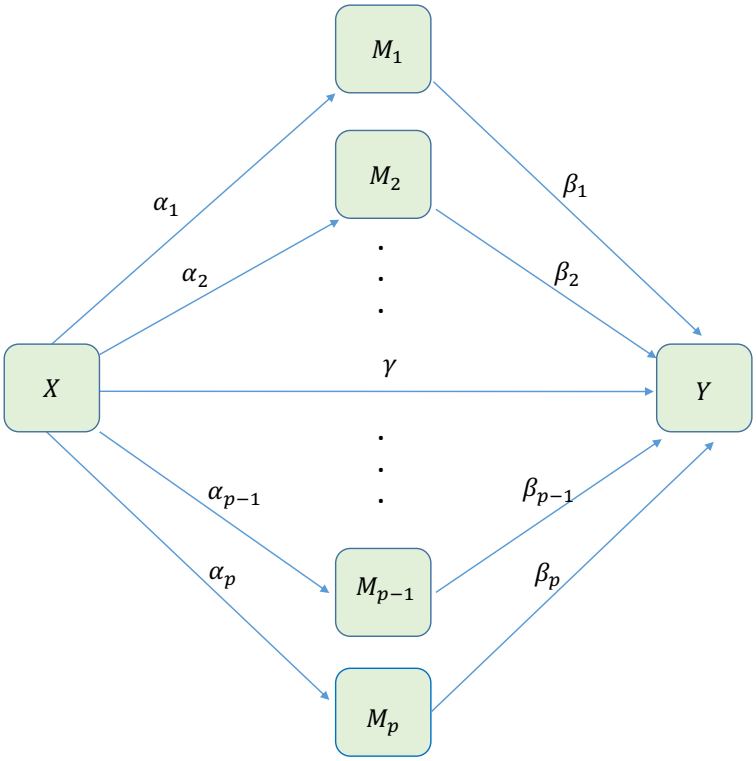

Figure 2. A scenario with multiple/high-dimensional mediators between exposure and outcome.