Biometrika (2020), **107**, 3, *pp.* 573–589
Printed in Great Britain

doi: 10.1093/biomet/asaa016
Advance Access publication 4 May 2020

Estimation and inference for the indirect effect in high-dimensional linear mediation models

By RUIXUAN RACHEL ZHOU

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, Illinois 61820, U.S.A.

rzhou14@illinois.edu

LIEWEI WANG

Division of Clinical Pharmacology, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, 200 First St. SW, Rochester, Minnesota 55905, U.S.A.

Wang.Liewei@mayo.edu

and SIHAI DAVE ZHAO

Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign, Illinois 61820, U.S.A.

sdzhao@illinois.edu

Summary

Mediation analysis is difficult when the number of potential mediators is larger than the sample size. In this paper we propose new inference procedures for the indirect effect in the presence of high-dimensional mediators for linear mediation models. We develop methods for both incomplete mediation, where a direct effect may exist, and complete mediation, where the direct effect is known to be absent. We prove consistency and asymptotic normality of our indirect effect estimators. Under complete mediation, where the indirect effect is equivalent to the total effect, we further prove that our approach gives a more powerful test compared to directly testing for the total effect. We confirm our theoretical results in simulations, as well as in an integrative analysis of gene expression and genotype data from a pharmacogenomic study of drug response. We present a novel analysis of gene sets to understand the molecular mechanisms of drug response, and also identify a genome-wide significant noncoding genetic variant that cannot be detected using standard analysis methods.

Some key words: High-dimensional inference; Integrative genomics; Mediation analysis.

1. Introduction

Mediation analysis is of great interest in many areas of research, such as psychology, epidemiology and genomics (MacKinnon, 2008; Hayes, 2013; Huang et al., 2014). A major goal is to understand the direct and indirect effects of an exposure variable on an outcome variable, potentially mediated through several intervening variables. Statistical methods for estimating and testing direct and indirect effects are well-developed when the number of mediator variables is relatively small (Hayes, 2013; VanderWeele & Vansteelandt, 2014; VanderWeele, 2015), but problems arise when the number of potential mediators exceeds the sample size. This high-dimensional scenario is common in genomics applications. For example, the effects of genetic variants may be mediated through the regulation of gene expression, but it is usually not known a priori which genes are regulated, so the total number of potential mediators can be very large.

General methods for high-dimensional inference are currently the subject of intense research. Techniques based on debiasing penalized regression estimators have been shown to provide asymptotically normal and unbiased estimators for certain parametric sparse regression models (Javanmard & Montanari, 2014, 2018; Van de Geer et al., 2014; Zhang & Zhang, 2014). The sparsity level of the regression parameter is not typically known. Cai & Guo (2017) discussed the construction of confidence intervals that can adapt to this unknown sparsity, and Zhu & Bradic (2018) proposed a test that avoids the sparsity requirement by instead assuming that the precision matrix is known or has certain sparsity properties. While these methods can be used for testing direct effects, they cannot be directly applied to perform inference on indirect effects. One approach is to use them to extend low-dimensional mediation analysis methods such as VanderWeele & Vansteelandt (2014), but it may be difficult to achieve valid inference for reasons that will be explained in § 2.1.

Several semiparametric high-dimensional methods have recently been proposed in the causal inference literature for peforming inference on causal effects in the presence of high-dimensional controls (Belloni et al., 2017; Athey et al., 2018). In particular, the procedure of Athey et al. (2018) is closely related to the method proposed here, and is discussed in detail in § 2.5. However, these approaches do not directly apply to estimating indirect effects in high-dimensional mediation models. Chen et al. (2015) and Huang & Pan (2016) use principal components analysis to reduce the dimensionality of the mediators, and employ the bootstrap for inference. Hanson et al. (2016) and Zhang et al. (2016) first screen the mediators according to their marginal correlations with the response.

In this paper we propose and provide asymptotic guarantees for two new inferential procedures for the indirect effect in high-dimensional linear mediation analysis models. We first consider the incomplete mediation setting, where both direct and indirect effects might exist. This is a common scenario, for example in genome-wide methylation studies that investigate whether environmental exposures exert their effects on phenotype by altering DNA methylation patterns. The exposures may also act through a nonmethylation pathway, giving rise to potential direct effects. We illustrate another application in § 6, where we identify gene sets that may mediate the effect of a gene of interest on a drug response phenotype.

We then consider the complete mediation setting when it is known that a direct effect does not exist. This setting is common when studying genetic variants located in noncoding regions of the genome, which typically can only exert their effects on a phenotype by regulating gene expression. Recent work has shown that in the low-dimensional case, testing for the indirect effect can be much more powerful than directly testing the total effect, even though both are equal under complete mediation (Kenny & Judd, 2014; Zhao et al., 2014b; Loeys et al., 2015). We show theoretically and in simulations that this is also true for our proposed high-dimensional method. Our work can thus be useful in genome-wide association studies where powerful tests are required to detect important variants. In an analysis of the genetics of drug response in § 6, our method was able to identify a genome-wide significant noncoding genetic variant that could not be detected by the standard approach.

2. Proposed methods

2.1. *Mediation model and notation*

For the ith subject, $i = 1, \ldots, n$, let Y_i be the outcome, G_i be a vector of p mediators and S_i be a vector of q exposures. We allow p to be larger than the sample size n, but we assume that S_i is low-dimensional. Finally, assume that the Y_i, G_i and S_i have all been centred to have zero mean. We consider the following linear mediation model:

$$Y_i = G_i^{\mathrm{T}}\alpha_0 + S_i^{\mathrm{T}}\alpha_1 + \epsilon_{1i}, \quad G_i = \gamma S_i + E_i, \tag{1}$$

where ϵ_{1i} are mean-zero random variables and E_i are mean-zero random vectors that are independent of G_i and S_i. Model (1) implies that $G_i^{\mathrm{T}}\alpha_0 = S_i^{\mathrm{T}}\gamma^{\mathrm{T}}\alpha_0 + \epsilon_{2i}$, where $\epsilon_{2i} = E_i^{\mathrm{T}}\alpha_0$. Let σ_1^2 denote the variance of ϵ_{1i}, and σ_2^2 denote the variance of ϵ_{2i}.

We are interested in performing inference on the indirect effect

$$\gamma^{\mathrm{T}}\alpha_0 \equiv \beta_0 \tag{2}$$

of S_i on Y_i when the dimension of G_i exceeds the sample size. We will describe separate methods for the incomplete mediation setting, where S_i may have a direct effect on Y_i through α_1, and the complete mediation setting, where α_1 is assumed to equal zero. We will assume throughout that α_0 is sparse, so that only a small number of variables mediate the effect of S_i on Y_i.

Assuming that model (1) is correctly specified with no unmeasured confounders, β_0 and the direct effect α_1 admit causal interpretations under a counterfactual framework, analogous to low-dimensional mediation models (Huang et al., 2014; VanderWeele & Vansteelandt, 2014). See the Supplementary Material for a detailed discussion. Our method can also accommodate measured confounders or covariates. If Z_i is a low-dimensional vector of potential confounders, we could write

$$Y_i = G_i^{\mathrm{T}}\alpha_0 + S_i^{\mathrm{T}}\alpha_1 + Z_i^{\mathrm{T}}\alpha_z + \epsilon_{1i}, \quad G_i = \gamma S_i + \gamma_z Z_i + E_i.$$

For example, in our data analysis in § 6 we let Z_i be a set of principal components to adjust for population stratification; in the Supplementary Material we describe how our proposed procedures can be modified for this setting. In this paper we do not consider more complicated models, such as interactions between the Z_i and G_i or between S_i and G_i. These may require additional methodological development, which we leave for future work.

The remainder of the paper will use the following notation. Let S be an $n \times q$ matrix of the S_i, G be an $n \times p$ matrix of the G_i, Y be an $n \times 1$ vector of the Y_i, ϵ_1 be an $n \times 1$ vector of the ϵ_{1i}, and E_i be an $n \times p$ matrix of the E_i. Define the vector $X_i = (G_i^{\mathrm{T}}, S_i^{\mathrm{T}})^{\mathrm{T}}$. Also define the sample matrices $\hat{\Sigma}_{SS} = n^{-1}\sum_i(S_iS_i^{\mathrm{T}})$, $\hat{\Sigma}_{SG} = n^{-1}\sum_i(S_iG_i^{\mathrm{T}})$, $\hat{\Sigma}_{GG} = n^{-1}\sum_i(G_iG_i^{\mathrm{T}})$, $\hat{\Sigma}_{GY} = n^{-1}\sum_i(G_iY_i)$, $\hat{\Sigma}_{XY} = n^{-1}\sum_i X_iY_i$, and $\hat{\Sigma}_{XX} = n^{-1}\sum_i X_iX_i^{\mathrm{T}}$, as well as their population-level versions Σ_{SS}, Σ_{SG}, Σ_{GG}, Σ_{GY}, Σ_{XX} and Σ_{XY}. Finally, for any matrix A, let a_{ij} denote the ijth entry and let $\|A\|_{L_1} = \max_j \sum_i |a_{ij}|$, $\|\cdot\|_1$ denote the elementwise ℓ_1 norm, and $\|\cdot\|_\infty$ denote the elementwise ℓ_∞ norm of either a vector or a matrix.

2.2. *Intuitions*

This section provides an intuitive description of the challenges of performing inference on the indirect effect β_0 (2) with high-dimensional mediators. For simplicity, in this subsection we assume that the direct effect $\alpha_1 = 0$. In the low-dimensional problem when $p < n$, γ^{T} and α_0

can be estimated using the ordinary least squares estimates $\tilde{\gamma}^{\mathrm{T}} = \hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}$ and $\tilde{\alpha}_0 = \hat{\Sigma}_{GG}^{-1}\hat{\Sigma}_{GY}$, respectively. Then β_0 can be estimated using $\tilde{\gamma}^{\mathrm{T}}\tilde{\alpha}_0$. Inference is straightforward because this product estimator typically has an asymptotically normal distribution (Sobel, 1982; Zhao et al., 2014b), though see the last paragraph of § 3.1. In high dimensions, when p exceeds n, the challenge is that the ordinary least squares estimator of α_0 does not exist. Since α_0 is sparse, one solution would be to use penalized regression, such as the lasso, to estimate α_0. However, these do not have tractable limiting distributions, so inference on β_0 using this approach is difficult.

An alternative might be to instead use a debiased lasso estimator $\breve{\alpha}_0$ of α_0, whose components do have nice asymptotic distributions (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014). We first briefly introduce $\breve{\alpha}_0$ following Javanmard & Montanari (2014). In high dimensions, the ordinary least squares estimator $\hat{\Sigma}_{GG}^{-1}\hat{\Sigma}_{GY}$ is not feasible because $\hat{\Sigma}_{GG}$ is no longer invertible, but we can still consider estimators of the form $\hat{\Omega}\hat{\Sigma}_{GY}$ for a different data-dependent matrix $\hat{\Omega}$. By model (1),

$$\hat{\Omega}\hat{\Sigma}_{GY} - \alpha_0 = (\hat{\Omega}\hat{\Sigma}_{GG} - I)\alpha_0 + \frac{1}{n}\hat{\Omega}G^{\mathrm{T}}\epsilon_1, \tag{3}$$

where I is the $p \times p$ identity matrix. In general, $\hat{\Omega}\hat{\Sigma}_{GY}$ will therefore be a biased estimator, with bias equal to $(\hat{\Omega}\hat{\Sigma}_{GG} - I)\alpha_0$. When α_0 is sparse, it turns out that this bias can be well-estimated by $(\hat{\Omega}\hat{\Sigma}_{GG} - I)\tilde{\alpha}_0$, if we carefully construct $\hat{\Omega}$ so that $\|\hat{\Omega}\Sigma_{GG} - I\|_\infty$ is small and $\tilde{\alpha}_0$ is a lasso estimate of α_0 so that $\|\tilde{\alpha}_0 - \alpha_0\|_1$ is small; for more details see Javanmard & Montanari (2014). The debiased lasso estimator is then constructed by subtracting the estimated bias from $\hat{\Omega}\hat{\Sigma}_{GY}$:

$$\breve{\alpha}_0 = \hat{\Omega}\hat{\Sigma}_{GY} - (\hat{\Omega}\hat{\Sigma}_{GG} - I)\tilde{\alpha}_0 = \tilde{\alpha}_0 + \frac{1}{n}\hat{\Omega}G^{\mathrm{T}}(Y - G\tilde{\alpha}) = \alpha_0 + \frac{1}{n}\hat{\Omega}G^{\mathrm{T}}\epsilon_1 + \Delta, \tag{4}$$

where $\Delta = (\hat{\Omega}\hat{\Sigma}_{GG} - I)(\alpha_0 - \tilde{\alpha})$. It can be shown for suitably constructed $\hat{\Omega}$ that each component of Δ is $o_P(n^{-1/2})$, so that each component of $n^{1/2}(\breve{\alpha}_0 - \alpha_0)$ is asymptotically normal. Javanmard & Montanari (2014) chose $\hat{\Omega}$ to minimize the variance of $\breve{\alpha}_0$, while Van de Geer et al. (2014) and Zhang & Zhang (2014) chose $\hat{\Omega}$ to estimate the precision matrix Σ_{GG}^{-1}.

Despite these encouraging properties, inference using the corresponding estimator $\tilde{\gamma}^{\mathrm{T}}\breve{\alpha}_0$ for β_0 is still not always possible. Using (4),

$$\tilde{\gamma}^{\mathrm{T}}\breve{\alpha}_0 = \tilde{\gamma}^{\mathrm{T}}\hat{\Omega}\hat{\Sigma}_{GY} - \tilde{\gamma}^{\mathrm{T}}(\hat{\Omega}\hat{\Sigma}_{GG} - I)\tilde{\alpha}_0 = \beta_0 + (\tilde{\gamma}^{\mathrm{T}} - \gamma)\alpha_0 + \frac{1}{n}\tilde{\gamma}^{\mathrm{T}}\hat{\Omega}G^{\mathrm{T}}\epsilon_1 + \tilde{\gamma}^{\mathrm{T}}\Delta,$$

which can be interpreted as a debiased version of $\tilde{\gamma}^{\mathrm{T}}\hat{\Omega}\hat{\Sigma}_{GY}$ for β_0. However, the error $\tilde{\gamma}^{\mathrm{T}}\Delta$ is no longer negligible: even though each component of Δ is $o_P(n^{-1/2})$, the linear combination $\tilde{\gamma}^{\mathrm{T}}\Delta$ may not be, so $n^{-1/2}(\tilde{\gamma}^{\mathrm{T}}\breve{\alpha}_0 - \beta_0)$ may not have an easily characterized asymptotic distribution. We argue in the Supplementary Material that we would need to at least assume either that $p\log(p)/n^{1/2} \to 0$ or that γ is sparse in order for $\tilde{\gamma}^{\mathrm{T}}\Delta = o_P(n^{-1/2})$. However, these conditions are restrictive.

In this paper we propose an estimate of β_0 under the weaker assumption that $\log(p)/n^{1/2} \to 0$, and without assumptions on the sparsity of γ. Our central idea is to develop a debiased estimator not of α_0 or β_0, but of $\Sigma_{SG}\alpha_0$. We will show that the bias of our initial estimator for this quantity can be estimated sufficiently accurately as long as we construct the matrix $\hat{\Omega}$ appropriately. By premultiplying our debiased estimate of $\Sigma_{SG}\alpha_0$ by the low-dimensional quantity $\hat{\Sigma}_{SS}^{-1}$, we will obtain an asymptotically normal estimate of β_0.

2.3. *Inference for the indirect effect under incomplete mediation*

We first estimate the indirect effect β_0 (2) under incomplete mediation, where α_1 is allowed to be nonzero. Let $X = (G, S)$ be the $n \times (p+q)$ design matrix and $\alpha = (\alpha_0^{\mathrm{T}}, \alpha_1^{\mathrm{T}})^{\mathrm{T}}$. As described in § 2.2, our strategy is to first obtain a debiased estimator for $\Sigma_{SG}\alpha_0$, which we will then premultiply by $\hat{\Sigma}_{SS}^{-1}$. First, define

$$D = \begin{pmatrix} \Sigma_{SG} & 0 \\ 0 & \Sigma_{SS} \end{pmatrix}, \quad \hat{D} = \begin{pmatrix} \hat{\Sigma}_{SG} & 0 \\ 0 & \hat{\Sigma}_{SS} \end{pmatrix}. \tag{5}$$

Following (3), we first consider estimators of the form $\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XY}$, for a matrix $\hat{\Omega}_{\mathrm{I}}$ that we will construct later. By model (1),

$$\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XY} - \begin{pmatrix} \Sigma_{SG}\alpha_0 \\ \Sigma_{SS}\alpha_1 \end{pmatrix} = (\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XX} - D)\alpha + \frac{1}{n}\hat{\Omega}_{\mathrm{I}}X^{\mathrm{T}}\epsilon_1.$$

Since α is sparse, we can estimate the bias term using $(\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XX} - \hat{D})\tilde{\alpha}$ as in (3), using a carefully constructed $\hat{\Omega}$ and where $\tilde{\alpha}$ is a lasso estimate of α. We will use the scaled lasso of Sun & Zhang (2012) because it also provides a consistent estimate of the variance of the Y_i, which will be useful later. We may also leave α_1 unpenalized, which is further discussed in § 7.

We can therefore construct a debiased estimate of $\Sigma_{SG}\alpha_0$ by subtracting the estimated bias from $\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XY}$, analogous to (4). We then premultiply the debiased estimator by $I_2 \otimes \hat{\Sigma}_{SS}^{-1}$, where I_2 denotes the 2×2 identity matrix and $\otimes$ denotes the Kronecker product. This gives our proposed estimator $\hat{b}$ for the indirect effect β_0 under incomplete mediation, as well as an estimate $\hat{a}$ of the direct effect α_1:

$$\begin{aligned} \begin{pmatrix} \hat{b} \\ \hat{a} \end{pmatrix} &= (I_2 \otimes \hat{\Sigma}_{SS}^{-1})\{\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XY} - (\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XX} - \hat{D})\tilde{\alpha}\} \\ &= \begin{pmatrix} \hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\tilde{\alpha}_0 \\ \tilde{\alpha}_1 \end{pmatrix} + (I_2 \otimes \hat{\Sigma}_{SS}^{-1})\frac{1}{n}\hat{\Omega}_{\mathrm{I}}X^{\mathrm{T}}(Y - X\tilde{\alpha}), \end{aligned} \tag{6}$$

where $\tilde{\alpha}_1$ is the component of $\tilde{\alpha}$ that estimates α_1.

Analogous to (4), it remains to find a suitable matrix $\hat{\Omega}_{\mathrm{I}}$ so that $(\hat{\Omega}_{\mathrm{I}}\hat{\Sigma}_{XX} - \hat{D})$ is small. We propose to choose $\hat{\Omega}_{\mathrm{I}}$ to estimate the matrix $D\Sigma_{XX}^{-1}$, for D defined in (5) and $\Sigma_{XX} = E(X_i X_i^{\mathrm{T}})$. Our estimator is based on constrained ℓ_1 optimization, similar to the precision matrix estimation procedure of Cai et al. (2011):

$$\hat{\Omega}_{\mathrm{I}} = \arg\min_{\Omega} \|\Omega\|_1 \text{ subject to } \|\Omega\hat{\Sigma}_{XX} - \hat{D}\|_\infty \leqslant \tau_n, \tag{7}$$

where τ_n is a tuning parameter. We will show in § 3 that $\hat{\Omega}_{\mathrm{I}}$ will converge to $D\Sigma_{XX}^{-1}$ under the condition that $D\Sigma_{XX}^{-1}$ is sparse.

We show in § 3 that under certain conditions, $(\hat{b}, \hat{a})$ is asymptotically normal and centred at the true (β_0, α_1). We also provide estimates of the asymptotic variance of $\hat{b}$, which will allow us to construct confidence intervals and conduct Wald tests for the indirect effects. Though this paper focuses on the indirect effect, (6) also gives an estimate $\hat{a}$ for the direct effect. As pointed out by a referee, the direct effect could also be estimated by subtracting $\hat{b}$ from the ordinary least squares estimate of the total effect of S_i on Y_i. We show in § 4.3 in the Supplementary Material that these two approaches are asymptotically equivalent.

2.4. *Inference for the indirect effect under complete mediation*

In some applications, for example in the analysis of noncoding genetic variants, it may be known that exposure does not act directly on the outcome, and only acts through mediators. We can make use of the extra information that $\alpha_1 = 0$ to develop a more efficient procedure for estimating the indirect effect β_0 (2). As above, we first obtain a debiased estimator for $\Sigma_{SG}\alpha_0$ and then premultiply by $\hat{\Sigma}_{SS}^{-1}$. We again first consider estimators of the form $\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GY}$, which by model (1) satisfy

$$\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GY} - \Sigma_{SG}\alpha_0 = (\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GG} - \Sigma_{SG})\alpha_0 + \frac{1}{n}\hat{\Omega}_{\mathrm{C}}G^{\mathrm{T}}\epsilon_1.$$

We construct $\hat{\Omega}_{\mathrm{C}}$ to estimate $\Sigma_{SG}\Sigma_{GG}^{-1}$, analogous to $\hat{\Omega}_{\mathrm{I}}$ (7) above:

$$\hat{\Omega}_{\mathrm{C}} = \arg\min_{\Omega} \|\Omega\|_1 \text{ subject to } \|\Omega\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG}\|_\infty \leqslant \tau_n', \tag{8}$$

where τ_n' is a tuning parameter. We show in § 3 that $\hat{\Omega}_{\mathrm{C}}$ will converge to $\Sigma_{SG}\Sigma_{GG}^{-1}$ if the latter is sparse. If $\tilde{\alpha}_0$ is the scaled lasso estimate of α_0, we can estimate the bias of $\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GY}$ using $(\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG})\tilde{\alpha}_0$. Subtracting this from $\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GY}$ and premultiplying by $\hat{\Sigma}_{SS}^{-1}$ gives our proposed estimate of β_0 under complete mediation:

$$\tilde{b} = \hat{\Sigma}_{SS}^{-1}\{\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GY} - (\hat{\Omega}_{\mathrm{C}}\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG})\tilde{\alpha}_0\} = \hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\tilde{\alpha}_0 + \hat{\Sigma}_{SS}^{-1}\frac{1}{n}\hat{\Omega}_{\mathrm{C}}G^{\mathrm{T}}(Y - G^{\mathrm{T}}\tilde{\alpha}_0). \tag{9}$$

We show in § 3 that $\tilde{b}$ is asymptotically normal and centred at the true β_0, and provide estimates for its asymptotic variance.

This estimator has an interesting efficiency property. Under complete mediation, β_0 can also be estimated by directly regressing Y_i on S_i and ignoring the mediating gene expression information. We will show that the asymptotic variance of the ordinary least squares estimator of β_0 is always greater than or equal to the variance of our $\tilde{b}$. The same phenomenon has been observed in a low-dimensional mediation model (Kenny & Judd, 2014; Zhao et al., 2014b; Loeys et al., 2015). Intuitively, our procedure achieves this efficiency gain by denoising the outcome Y_i, replacing it with an estimate $G_i^{\mathrm{T}}\tilde{\alpha}_0$ of its conditional expectation $G_i^{\mathrm{T}}\alpha_0$ and thus removing much of the variation from the error term ϵ_{1i}.

2.5. *Connections to existing work*

Estimating the indirect effect in high dimensions is challenging because β_0 (2) is a linear combination of the high-dimensional vector α_0. Athey et al. (2018) encountered a similar issue studying inference for a causal effect in the presence of high-dimensional controls, and also took a debiasing approach. Both of our approaches can be viewed as debiasing a pilot estimator by subtracting a weighted sum of the residuals from a fitted penalized regression model for Y_i. Athey et al. (2018) chose the weights in this weighted sum to minimize the estimation error of the desired linear combination, while our weights are equal to $(I_2 \otimes \hat{\Sigma}_{SS}^{-1})\hat{\Omega}_{\mathrm{I}}$ in (6) and $\Sigma_{SS}^{-1}\hat{\Omega}_{\mathrm{C}}$ in (9). The coefficients of the desired linear combination are known in the setting of Athey et al. (2018), while in our approach they are equal to Σ_{SG} and must be estimated, so the method of Athey et al. (2018) is not directly applicable here. It would be interesting to apply their strategy to our mediation framework in the future.

There are alternative approaches to constructing the matrices $\hat{\Omega}_{\mathrm{I}}$ (7) and $\hat{\Omega}_{\mathrm{C}}$ (8). One method might be to choose them to minimize the variances of the resulting estimators $\hat{a}$, $\hat{b}$ and $\tilde{b}$ while controlling their biases. In the standard linear regression setting with high-dimensional covariates, Javanmard & Montanari (2014) showed that this strategy can give asymptotically optimal inference without requiring the precision matrix of the covariates to be sparse. As pointed out by a referee, applying this strategy to the present mediation setting may obviate the need to assume sparsity of $D\Sigma_{XX}^{-1}$ and $\Sigma_{SG}\Sigma_{GG}^{-1}$. This is an important direction for future work, and the Supplementary Material contains a detailed discussion and simulation study exploring the robustness of our procedure to the accuracy of estimating $D\Sigma_{XX}^{-1}$ and $\Sigma_{SG}\Sigma_{GG}^{-1}$. On the other hand, our current strategy of choosing $\hat{\Omega}_{\mathrm{I}}$ and $\hat{\Omega}_{\mathrm{C}}$ to estimate $D\Sigma_{XX}^{-1}$ and $\Sigma_{SG}\Sigma_{GG}^{-1}$ allows us to characterize the asymptotic variances of our proposed estimators in terms of population-level quantities, as well as to construct consistent estimates of those variances. Hirshberg & Wager (2019) studied a similar approach for a more general class of debiased estimators.

3. Theoretical results

3.1. *Incomplete mediation*

This section presents the theoretical properties of our proposed indirect effect inference procedure under incomplete mediation. We first require G_{ij}, S_i and residual error ϵ_{1i} to have exponential-type tails and make several sparsity assumptions.

Assumption 1. For each $j = 1, \ldots, p$, G_{ij} has mean zero and $\mathrm{E}\{\exp(tG_{ij}^2)\} \leqslant K < \infty$ for some constant K and all $|t| \leqslant \eta$, where $\eta \in (0, 1/4)$ and $\{\log(p+q)\}/n \leqslant \eta$. The same tail conditions hold for S_i and ϵ_{i1}.

Assumption 2. For D defined in (5), there exist constants M_X and N_X such that $\|\Sigma_{XX}^{-1}\|_{L_1} \leqslant M_X$ and $\|(D\Sigma_{XX}^{-1})^{\mathrm{T}}\|_{L_1} \leqslant N_X$. Furthermore, if ω_{ij} denotes the ijth entry of $D\Sigma_{XX}^{-1}$, then $\max_i \sum_j |\omega_{ij}|^{\theta} < s_0$ for some s_0 and $\theta \in [0, 1)$.

The quantity s_0 in Assumption 2 measures the degree of sparsity of $D\Sigma_{XX}^{-1}$. The condition on $\|\Sigma_{XX}^{-1}\|_{L_1}$ requires that none of the rows contain too many large entries. This is reasonable, as precision matrices are frequently used to model conditional dependencies between genes in a gene network (Danaher et al., 2014; Zhao et al., 2014a), and gene networks are typically thought to be sparse. The condition on $\|D\Sigma_{XX}^{-1}\|_{L_1}$ is related to the irrepresentable condition of Zhao & Yu (2006), and is similar to requiring that S_i cannot be completely explained by G_i.

Theorem 1. *Let* $\hat{\Omega}_{\mathrm{I}}$ *solve* (7) *with tuning parameter* $\tau_n = (N_X + 1)C_1\{(\log(p+q))/n\}^{1/2}$ *for* $C_1 = 2\eta^{-2}(2 + \tau + \eta^{-1}\mathrm{e}^2K^2)^2$, *where* K *and* η *are from Assumption* 2 *and* $\tau > 0$. *Then, under Assumptions* 1 *and* 2, *with probability greater than* $(1 - 4p^{-\tau})$ *and with* D *defined in* (5),

$$\|\hat{\Omega}_{\mathrm{I}} - D\Sigma_{XX}^{-1}\|_{\infty} \leqslant (4N_X + 2)C_1M_X\{(\log p)/n\}^{1/2}.$$

Theorem 1 shows that our $\hat{\Omega}_{\mathrm{I}}$ (7) is a consistent estimate of the population-level matrix $D\Sigma_{XX}^{-1}$. As discussed in § 2.3, in the standard linear regression setting Javanmard & Montanari (2014) proposed a method for high-dimensional inference that does not require consistent estimation of precision matrices. In the Supplementary Material we discuss whether their approach can be applied here as well, which would avoid the need for the sparsity conditions in Assumption 2.

We can now characterize the asymptotic behaviour of our incomplete mediation estimators $(\hat{b}, \hat{a})$ (6). We require additional assumptions necessary for the good performance of the scaled lasso of Sun & Zhang (2012).

Theorem 2. *Let $\hat{b}$ and $\hat{a}$ be calculated such that both tuning parameters λ_n and τ_n are $O\{(n^{-1}\log p)^{1/2}\}$. Assume the model for Y_i (1) satisfies the conditions of Theorem 2 of Sun & Zhang (2012) and that α_0 has at most $s_0 = o(n^{1/2}/\log p)$ nonzero components. Under Assumptions 1 and 2, if $(\log p)/n^{1/2} \to 0$, and α_0 and Σ_{SG} are not both zero, and if $\Gamma \equiv \Sigma_{SS}^{-1}\Sigma_{SG}(\Sigma_{GG} - \Sigma_{GS}\Sigma_{SS}^{-1}\Sigma_{SG})^{-1}\Sigma_{GS}\Sigma_{SS}^{-1}$ converges to a positive-definite matrix, then*

$$n^{1/2}\begin{pmatrix}\hat{b}-\beta_0\\ \hat{a}-\alpha_1\end{pmatrix} \to N(0,V), \quad \textit{where } V = \begin{pmatrix}\sigma_1^2\Gamma + \sigma_2^2\Sigma_{SS}^{-1} & -\sigma_1^2\Gamma\\ -\sigma_1^2\Gamma & \sigma_1^2(\Gamma+\Sigma_{SS}^{-1})\end{pmatrix}.$$

The ultra-sparsity assumption on s_0 in Theorem 2 is standard in the debiased lasso literature (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014). The choice of τ_n controls the coherence parameter $\|\hat{\Omega}_{\rm I}\hat{\Sigma}_{XX} - \hat{D}\|_\infty$ at rate $(n^{-1}\log p)^{1/2}$, which is necessary for showing that the bias of our proposed estimator goes to 0 when n and p go to infinity. The proof of Theorem 2 shows that the asymptotic variance V can be consistently estimated using

$$\hat{\sigma}_1^2(I_2\otimes\hat{\Sigma}_{SS}^{-1})\hat{\Omega}_{\rm I}\hat{\Sigma}_{XX}\hat{\Omega}_{\rm I}^{\rm T}(I_2\otimes\hat{\Sigma}_{SS}^{-1}) + \begin{pmatrix}\hat{\sigma}_2^2\hat{\Sigma}_{SS}^{-1} & 0\\ 0 & 0\end{pmatrix}.$$

Consistency of $\hat{\Sigma}_{XX}$ and $\hat{\Sigma}_{SS}$ is standard, and consistency of $\hat{\Omega}_{\rm I}$ is given by Theorem 1. Estimation of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ is discussed in § 4.

We caution that Theorem 2 does not cover the setting where both $\Sigma_{SG} = 0$ and $\alpha_0 = 0$. This would cause $n^{1/2}(\hat{b}-\beta_0)$ to asymptotically equal zero, rather than be normally distributed. A related issue arises even for standard low-dimensional Wald-type tests for the indirect effect, such as Sobel's test (Sobel, 1982; Hayes, 2013; Barfield et al., 2017). In practice, these tests can be conservative when the exposure, the mediator and the outcome are only weakly associated. In this case, the true finite-sample distribution of the Wald test statistic has higher kurtosis than a normal distribution, so that critical values calculated assuming a normal distribution lead to a conservative test (Barfield et al., 2017). This setting is different from the weak instrumental variable problem, which we discuss in the Supplemetary Material.

3.2. *Complete mediation*

We now present the theoretical properties of our indirect effect inference procedure under complete mediation. Similar to Assumption 2, we first make several sparsity assumptions, under which we can show that $\hat{\Omega}_{\rm C}$ (8) is a consistent estimate of $\Sigma_{SG}\Sigma_{GG}^{-1}$.

Assumption 3. There exist constants M_G and N_G such that $\|\Sigma_{GG}^{-1}\|_{L_1} \leqslant M_G$ and $\|(\Sigma_{SG}\Sigma_{GG}^{-1})^{\rm T}\|_{L_1} \leqslant N_G$. Furthermore, if ω_{ij} denotes the ijth entry of $\Sigma_{SG}\Sigma_{GG}^{-1}$, then $\max_i\sum_j|\omega_{ij}|^\theta < s_0$ for some s_0 and $\theta \in [0,1)$.

Theorem 3. *Let $\hat{\Omega}_{\rm C}$ solve (8) with tuning parameter $\tau_n = (N_G+1)C_1\{(\log p)/n\}^{1/2}$. Then, under Assumptions 1 and 3, with probability greater than $(1-4p^{-\tau})$, and with C_1 and τ as in Theorem 1,*

$$\|\hat{\Omega}_{\rm C} - \Sigma_{SG}\Sigma_{GG}^{-1}\|_\infty \leqslant (4N_G+2)C_1M_G\{(\log p)/n\}^{1/2}.$$

We can now characterize the asymptotic behaviour of our complete mediation indirect effect estimator $\tilde{b}$ (9). The proof of Theorem 4 indicates that the asymptotic variance of $\tilde{b}$ can be consistently estimated by $\hat{\sigma}_1^2 \hat{\Sigma}_{SS}^{-1} \hat{\Omega}_C \hat{\Sigma}_{GG} \hat{\Omega}_C \hat{\Sigma}_{SS}^{-1} + \hat{\sigma}_2^2 \hat{\Sigma}_{SS}^{-1}$. As with incomplete mediation case, the requirement that α_0 and Σ_{SG} are not both zero arises here as well.

THEOREM 4. *Let $\tilde{b}$ be calculated such that both tuning parameters λ_n and τ_n are of order $O\{(n^{-1}\log p)^{1/2}\}$. Assume the model for Y_i in mediation model* (1) *has $\alpha_1 = 0$, but otherwise satisfies the conditions of Theorem* 2 *of Sun & Zhang (2012), and that α_0 has at most $s_0 = o(n^{1/2}/\log p)$ nonzero components. Under Assumptions* 1 *and* 3, *if $(\log p)/n^{1/2} \to 0$, α_0 and Σ_{SG} are not both zero, and if $\Sigma_{SG}\Sigma_{GG}^{-1}\Sigma_{GS}$ converges to a positive-definite matrix, then*

$$n^{1/2}(\tilde{b} - \beta_0) \to N(0, \sigma_1^2 \Sigma_{SS}^{-1} \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS} \Sigma_{SS}^{-1} + \sigma_2^2 \Sigma_{SS}^{-1}).$$

As mentioned in § 2.4, under complete mediation the indirect effect β_0 can also be consistently estimated by directly regressing Y_i on S_i. The expression for the asymptotic variance of $\tilde{b}$ from Theorem 4 now allows us to analytically compare our estimator with the ordinary least squares estimate of β_0.

PROPOSITION 1. *In model* (1), *assume that $\alpha_1 = 0$, so that $\tilde{b}_{\text{OLS}} = (S^{\mathrm{T}}S)^{-1}S^{\mathrm{T}}Y$ is a consistent estimator of β_0. Then, under the conditions of Theorem* 4, $\text{var}\{n^{1/2}(\tilde{b}_{\text{OLS}} - \beta_0)\} - \text{var}\{n^{1/2}(\tilde{b} - \beta_0)\}$ *converges to a positive semidefinite matrix.*

Proposition 1 shows that our $\tilde{b}$ always has equal or lower asymptotic variance compared to the ordinary least squares estimator, even when the mediators are high-dimensional. This extends similar findings in low dimensions (Kenny & Judd, 2014; Zhao et al., 2014b; Loeys et al., 2015). In fact, we show in the Supplementary Material that for any fixed p, our estimator $\tilde{b}$ achieves the minimum asymptotic variance among all asymptotically unbiased estimators of β_0 with the same convergence rate. Tests based on $\tilde{b}$ will thus have higher power to detect nonzero β_0 than tests based on $\tilde{b}_{\text{OLS}}$, as confirmed by simulations in § 5.3. In practice, the Wald test based on $\tilde{b}$ can still be conservative when α_0 and Σ_{SG} are close to zero, for reasons discussed in § 3.1, but simulations show that our proposed $\tilde{b}$ can still have significant power gains over the majority of the parameter space.

In a closely related setting, Athey et al. (2020) found that when estimating the causal effect of a treatment on a long-term outcome, leveraging intermediate outcomes can increase efficiency. In the Supplementary Material we provide a detailed comparison. Together, these results converge on a common principle, and provide theoretical justification for recent work in genomics showing that data integration using mediation analysis can increase the power to detect important biological signals (Wang et al., 2012; Huang et al., 2015).

The improved efficiency guaranteed by Proposition 1 requires strong scientific or expert knowledge to justify the absence of the direct effect. Furthermore, it also depends on the correct specification of both stages of the linear mediation model (1). In low dimensions, this has been pointed out by Loeys et al. (2015). This is in contrast to the usual ordinary least squares estimator, which requires fewer modelling assumptions. We illustrate the effect of model misspecification on our proposed estimator in the Supplementary Material.

Our estimator $\hat{b}$ (6), proposed in § 2.3 under incomplete mediation, could also be used to estimate the indirect effect under complete mediation. Proposition 2 shows that under complete mediation, $\tilde{b}$ is asymptotically more efficient.

PROPOSITION 2. *In model* (1), *assume that $\alpha_1 = 0$. Under the conditions of Theorems* 2 *and* 4, $\text{var}\{n^{1/2}(\hat{b} - \beta_0)\} - \text{var}\{n^{1/2}(\tilde{b} - \beta_0)\}$ *converges to a positive semidefinite matrix.*

4. Implementation

We first centre the Y_i, G_i and S_i. To apply the scaled lasso, we standardize all covariates to have unit variance and then choose the tuning parameter λ_n using the quantile-based penalty procedure in the `R` package `scalreg` (R Development Core Team, 2020).

To estimate the asymptotic variances of our estimators, given in Theorems 2 and 4, we need estimates of the residual variances σ_1^2 and σ_2^2 from our mediation model (1). Sun & Zhang (2012) showed that the scaled lasso can provide a consistent estimate $\hat{\sigma}_1^2$ for σ_1^2. Since model (1) implies that $Y_i = S_i^{\mathrm{T}}(\beta_0 + \alpha_1) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_1^2 + \sigma_2^2)$, we can estimate σ_2^2 by first regressing Y_i on S_i to obtain the ordinary least squares residual variance estimator $\hat{\sigma}^2$, and using $\hat{\sigma}_2^2 = \hat{\sigma}^2 - \hat{\sigma}_1^2$. In practice, $\hat{\sigma}_1$ may sometimes be larger than $\hat{\sigma}$, in which case we estimate $\hat{\sigma}_2 = 0$. This is sensible because $\hat{\sigma}_1 > \hat{\sigma}$ likely occurs when no mediators are associated with the outcome, i.e., $\alpha_0 = 0$, in which case σ_2 indeed equals zero.

We construct the matrices $\hat{\Omega}_{\mathrm{I}}$ (7) and $\hat{\Omega}_{\mathrm{C}}$ (8) by setting the tuning parameters $\tau_n = \tau_n' = \{(\log p)/n\}^{1/2}/3$. This choice is guided by Theorems 1 and 3. We also tried choosing the tuning parameters by minimizing an ad hoc information criterion-type measure, but this resulted in confidence intervals with poor coverage in some cases. Finding a more data-adaptive tuning procedure is an important direction for future research.

The time-consuming part of our method is the constrained ℓ_1 optimization, in (7) and (8), which we implement using fast algorithms from the `flare` package. For $n = 300$ subjects, $p = 1000$ mediators and $q = 1$ exposure, our procedure with tuning parameter $\tau_n = \{(\log p)/n\}^{1/2}/3$ takes 66 seconds on a single core of an Intel Xeon X5675 processor at 3.07 GHz and with 8 GB of RAM, and larger τ_n results in shorter computation time: our procedure with $\tau_n = \{(\log p)/n\}^{1/2}$ takes 5 seconds. Our procedure is available in the `R` package `freebird`, available at `https://github.com/rzhou14/freebird`.

5. Numerical results

5.1. *Comparison of methods*

We compared our methods to a naive nondebiased remax method, discussed in § 2.2. This estimates β_0 using $\hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\tilde{\alpha}_0$, where $\tilde{\alpha}_0$ is a standard lasso estimate of α_0 implemented using the `R` package `glmnet`. There is no tractable limiting distribution for this estimate of β_0, so we used the bootstrap to obtain percentile confidence intervals, and obtained average power and coverages based on those intervals. Bootstrapping the lasso is not theoretically justified (Dezeure et al., 2015), but this at least allows us to have a comparable baseline method.

We also compared our procedures to the high-dimensional mediation analysis method of Zhang et al. (2016), using their `R` package `HIMA`. The method first uses marginal screening on mediators to reduce dimensionality. It then regresses the outcome on the retained mediators using penalized regression with the minimax concave penalty (Zhang, 2010). Using only the selected mediators, it uses ordinary least squares to compute a pair of p-values for each mediator, for its associations with the outcome and the exposure. These p-values are Bonferroni-corrected for the number of selected mediators, and Zhang et al. (2016) identify a mediator as significant if both of its adjusted p-values are less than the desired significance level. However, this testing approach based on the maximum of two p-values does not provide confidence intervals.

Under complete mediation, the indirect effect is equal to the total effect, which can be tested directly using ordinary least squares. In this setting we therefore also compared our complete mediation method to ordinary least squares.

5.2. *Simulations under incomplete mediation*

We first studied our estimators under incomplete mediation. Following model (1), for samples $i = 1, \ldots, n = 300$ we generated $q = 1$ exposure $S_i \sim N(0, 1)$ and $p = 500$ potential mediators G_i following $G_i = c\gamma S_i + E_i$, where c was a scalar, γ was a $p \times 1$ coefficient vector and $E_i \sim N(0, \Sigma_E)$. We generated Σ_E following procedures in Danaher et al. (2014) such that Σ_E^{-1} was sparse in the sense of Assumption 2 and had diagonal entries equal to 1. Finally, we generated the outcome according to $Y_i = G_i^{\mathrm{T}}\alpha_0 + S_i^{\mathrm{T}}\alpha_1 + \epsilon_{1i}$, where $\epsilon_{1i} \sim N(0, 5)$. In the Supplementary Material we show that our simulation results were similar even when ϵ_{1i} was not normally distributed. We let γ have 15 nonzero components randomly generated between $[-1, 1]$, fixing γ across replications, and let α_0 have 15 nonzero components equal to one. We chose either one or five of these nonzero components to correspond to variables whose entries in γ were also nonzero; these were the true mediators. Here we set the direct effect $\alpha_1 = 0.1$, and in the Supplementary Material we present results when $\alpha_1 = 0.5$. In this simulation scheme the indirect effect $\beta_0 = c\gamma^{\mathrm{T}}\alpha_0$, and we varied β_0 by varying the constant c.

When there was only one true mediator, we used all three competing methods to test $H_0 : \beta_0 = 0$ at the $\alpha = 0.05$ significance level, and used our proposed method and naive method to calculate 95% confidence intervals for different values of β_0. When there were five true mediators, we did not apply the method of Zhang et al. (2016) because it considers each mediator separately and does not provide inference for the overall indirect effect. The existence of multiple significant mediators does not imply that the indirect effect is nonzero because the effects of the different mediators may cancel each other out, a phenomenon known as inconsistent mediation.

Figure 1 reports average coverage probabilities and power curves over 200 replications. The naive method had worse coverage than our approach, and though it had excellent power, it was not theoretically justified, as mentioned in § 5.1. The method of Zhang et al. (2016) had counter-intuitive behaviour when β_0 was large, and surprisingly high power when β_0 was small. Its power was poor for large β_0 because its model selection step performed poorly: larger β_0 corresponded to larger c, and therefore to increased collinearity between G_i and S_i in the regression for Y_i, making consistent model selection difficult. Its power was surprisingly high for small β_0 because it did not appropriately account for the variability of its model selection step. In the Supplementary Material we describe a slightly modified version of their approach that gives confidence intervals and show that it has poor coverage, and also construct a setting where it fails to maintain Type I error because of its improper post-model selection inference.

5.3. *Simulations under complete mediation*

We next studied the performance of our indirect effect estimator under complete mediation. We considered four simulation settings based on the same data generation scheme used above, but with $\alpha_1 = 0$. We generated $p = 500$ potential mediators with either one or five true mediators, and in the Supplementary Material we present results for $p = 1000$.

Figure 2 reports average coverage probabilities and power curves over 200 replications. Our method was always able to maintain the nominal coverage probability and significance level, and in every case had higher power than ordinary least squares and the naive method for sufficiently large β_0, consistent with Proposition 1. The average lengths of 95% confidence intervals were also smaller for our method compared to ordinary least squares; see the Supplementary Material. Similar to the incomplete mediation setting, the method of Zhang et al. (2016) had high power when β_0 was small and counter-intuitive behaviour when β_0 was large. Our test was slightly conservative for β_0 close to zero because the normal approximation to the distribution of our Wald-type test statistic is poor under weak mediation, as discussed in § 3.1.

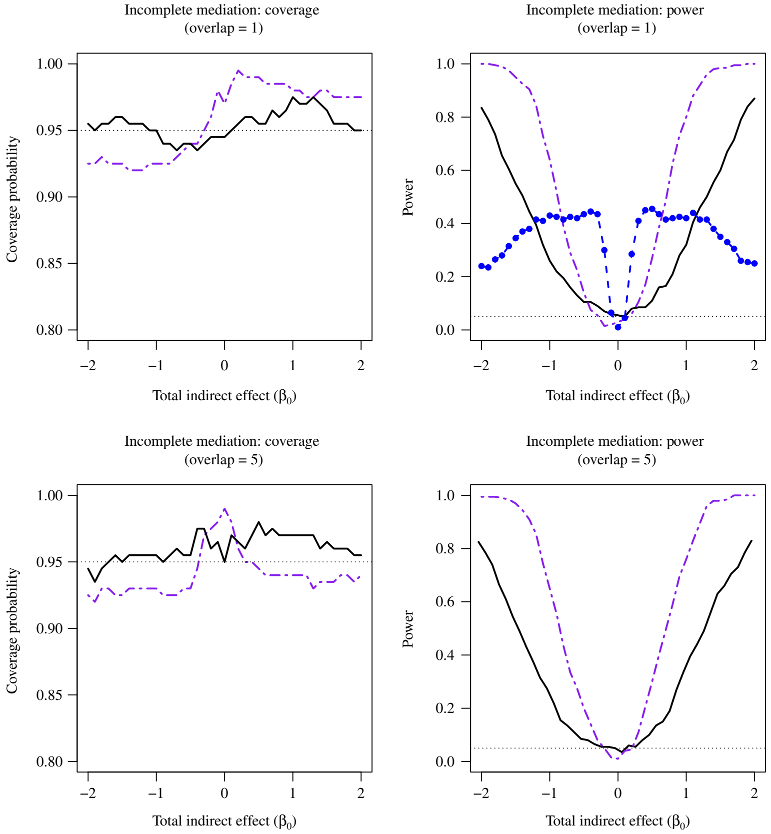

Fig. 1. Average coverage probabilities of 95% confidence intervals (left panels) and average power curves at significance level $\alpha = 0.05$ (right panels) for estimating and testing the indirect effect under incomplete mediation, over 200 replications. The direct effect was $\alpha_1 = 0.1$. The number of true mediators was 1 in the upper panels and 5 in the lower panels. Proposed (solid): $\hat{b}$ from (6); Naive (dot-dash): the naive method discussed in § 5.1; Zhang (large dot-dash): method of Zhang et al. (2016).

6. Data analysis

6.1. *Data description*

Understanding the mechanisms behind individual variation in drug response is an important step in the development of personalized medicine. We applied our proposed methods to pharmacogenetic studies of the response to the cancer drug docetaxel in human lymphoblastoid cell line (Niu et al., 2012; Hanson et al., 2016). The data consists of genotype data on 1 362 849 single nucleotide polymorphisms and expression data on 54 613 probes, after pre-processing, from cell lines from 95 Han-Chinese, 96 Caucasian and 93 African-American individuals. These data are available from the Gene Expression Omnibus under accession number GSE24277. Niu et al. (2012) exposed these cells to docetaxel and quantified their responses using EC_{50}, the concentration at which a drug reduces the population of cells by half (Hanson et al., 2016).

6.2. *Gene set analysis*

It is common in gene expression profiling experiments to identify genes that are significantly associated with the phenotype being studied. A natural next step is to identify gene sets, representing biological pathways, through which these significant genes may act. This is a difficult analysis problem because the intervening pathways may contain a large number of genes,

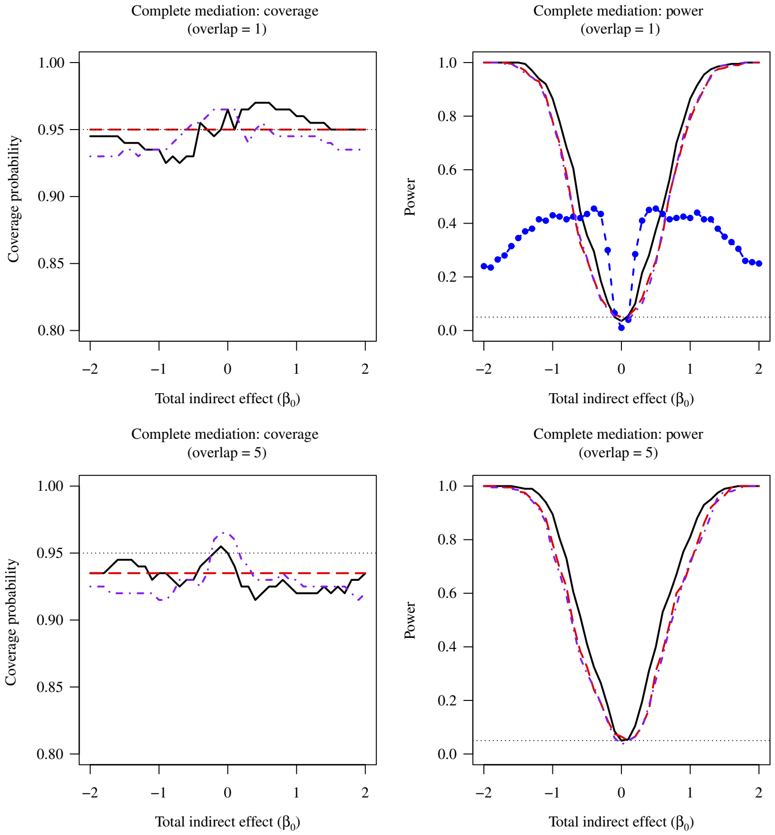

Fig. 2. Average coverage probabilities of 95% confidence intervals (left panels) and average power curves at significance level $\alpha = 0.05$ (right panels) for estimating and testing the indirect effect under complete mediation, over 200 replications. The number of true mediators was 1 in the upper panels and 5 in the lower panels, with 500 potential mediators. Proposed (solid): $\tilde{b}$ from (9); Naive (dot-dash): the naive method discussed in § 5.1; OLS (log dash): ordinary least squares estimate; Zhang (large dot-dash): the method of Zhang et al. (2016).

resulting in a high-dimensional mediation analysis problem. This is different from standard gene set enrichment analysis (Subramanian et al., 2005), as the latter does not allow for direct testing of mediation by the gene set.

We applied our proposed procedures to test whether a candidate gene set mediates the indirect effect of a given gene of interest on the phenotype. We used our incomplete mediation estimator $\hat{b}$ (6), because the gene of interest may have a direct effect on the phenotype that does not proceed through the candidate gene set. As in our simulations, we set the tuning parameter $\tau_n = \{(\log p)/n\}^{1/2}/3$ when estimating $\hat{\Omega}_{\mathrm{I}}$ (7). As an illustration, we studied the indirect effects of TMED10, a transmembrane trafficking protein whose corresponding gene was the most significantly associated with docetaxel response in our data. We retrieved biological process Gene Ontology gene sets with at least 50 genes from the Molecular Signatures Database (Subramanian et al., 2005; Liberzon et al., 2011), then applied our proposed approach to test the indirect effect of TMED10 through each of the 4436 candidates. Of these, 420 gene sets contained more genes than there were samples, making our high-dimensional approach indispensable.

Our procedure found 257 gene sets with significant indirect effects that passed Bonferroni correction. One reason for the large number of significant findings is that many gene sets are subgroups of larger sets. Table 1 reports the top ten most significant ones, as ranked by their

Table 1. *Top ten most significant gene sets through which the TMED10 gene may act on drug response*

Gene set	95% CI	p-value
Regulation of heart rate	-0.83 ± 0.22	6.3×10^{-14}
Synaptic vesicle cycle	-0.60 ± 0.17	2.6×10^{-12}
Regulation of vasoconstriction	-1.08 ± 0.30	3.1×10^{-12}
Negative regulation of transporter activity	-0.64 ± 0.18	5.3×10^{-12}
Negative regulation of cation transmembrane transport	-0.70 ± 0.20	5.3×10^{-12}
Positive regulation of blood circulation	-1.07 ± 0.31	1.1×10^{-11}
Negative regulation of transmembrane transport	-0.73 ± 0.21	2.2×10^{-11}
Regulation of cardiac muscle contraction	-0.60 ± 0.18	6.8×10^{-11}
Neurotransmitter transport	-0.58 ± 0.18	1.4×10^{-10}
Regulation of oxidoreductase activity	-0.73 ± 0.22	1.5×10^{-10}

95% CI, confidence intervals obtained from the proposed method under incomplete mediation (6); p-value, raw p-values obtained from the proposed procedure.

indirect effect p-values. Many of these are involved in transmembrane transport, which suggests that the role of TMED10 in the response to docetaxel may be to move small molecules into and out of cells. Our proposed method can thus generate useful exploratory results for further downstream analysis. We also implemented the method of Zhang et al. (2016), which found no significant gene sets.

6.3. *Noncoding variants analysis*

We next studied the effects of noncoding genetic variants on the response to docetaxel. We first performed a standard genome-wide association study and regressed docetaxel EC_{50} on each variant separately, controlling for the first five principal components of the genotype data in order to control for population stratification (Price et al., 2006). This approach did not identify any significant variants after multiple testing correction. We were then interested in whether a high-dimensional mediation analysis method could provide more power. We chose the top 1000 expression probes with the largest variances as potential mediators and controlled for the first five principal components. We first applied the method of Zhang et al. (2016), but it did not detect any significant variants that passed Bonferroni correction.

It is known that noncoding variants likely do not have a direct effect on the phenotype. This justifies application of our complete mediation estimator $\tilde{b}$ (9) to test for noncoding variances associated with EC_{50}. We use $\tau_n' = \{(\log p)/n\}^{1/2}/3$ when estimating $\hat{\Omega}_{\text{C}}$ (8), and controlled for the first five principal components in all of our analyses. Our new procedure was indeed able to identify one significant variant that passed Bonferroni correction for all noncoding variants: the single nucleotide polymorphism rs11578000, with an estimated indirect effect of $\hat{b} = -0.0777 \pm 0.0186$ and a p-value of 2.8×10^{-16}. Interestingly, the Genotype Tissue Expression Project (Lonsdale et al., 2013) found that in heart and muscle tissue, rs11578000 regulated the expression of the gene SUSD4, which has been found to inhibit the complement system (Holmquist et al., 2013), a system of proteins involved in innate immunity that may be involved in the response to epirubicin/docetaxel treatment in breast cancer patients (Michlmayr et al., 2010). Our $\tilde{b}$ provides novel findings that could not have been detected using standard approaches.

7. Discussion

Our methods require that the directions of causality in mediation model (1) be correctly specified. In practice this causal pathway may be complex, as some genes react to the outcome, rather than cause the outcome. Our method's findings should thus be further analysed to verify that the causal directions are indeed of interest. One potential solution to this issue is to use recently developed methods for high-dimensional causal inference (Bühlmann et al., 2014) to first screen out reactive genes before applying our proposed procedures.

Though we focused on testing the indirect effect in this paper, our incomplete mediation method also provides $\hat{a}$, a natural estimate of the direct effect, as discussed in § 2.3. We explored using $\hat{a}$ to test for the presence of a direct effect, and similar to Kenny & Judd (2014), we found that the power was relatively low. This may be because when calculating our estimators we penalize the direct effect parameter α_1 when we fit the scaled lasso. This makes sense if the direct effect is expected to be zero, which is sensible in our integrative genomics applications, but an alternative is to leave α_1 unpenalized. This may give a more powerful test for the direct effect, and more work is required to derive the asymptotic distribution of the resulting estimator. Inference for the direct effect in high dimensions could also be achieved by applying debiased lasso methods to test α_1 in the regression of Y_i on G_i and S_i in our model (1). Based on some simulations, we found that our estimator $\hat{a}$ is always smaller in absolute value, and usually had smaller variance, compared to the debiased estimator of Van de Geer et al. (2014).

Finally, we have so far only considered linear mediation models for continuous outcomes. It is possible to extend our methods to generalized linear models for the outcome Y_i in mediation model (1). However, the causal interpretation of these nonlinear models requires special care (VanderWeele & Vansteelandt, 2010; VanderWeele, 2015). Also, we have so far assumed that the residual errors ϵ_{1i} and E_i are independent of the exposure S_i and mediator G_i in model (1). Under heteroscedasticity, if the errors are dependent on either S_i or G_i, our theoretical results will likely not hold, and extending our approach to this setting is an important research direction.

Acknowledgement

The authors thank the reviewers and the associate editor for their extremely useful comments, as well as Casey Hanson for his help with data processing. This work was funded in part by the Mayo Clinic-UIUC Alliance and by the trans-NIH Big Data to Knowledge initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Zhao was funded in part by the National Science Foundation and is also affiliated with the Carl R. Woese Institute for Genomic Biology at the University of Illinois.

Supplementary material

Supplementary material available at *Biometrika* online contains further comparisons, additional simulation results and proofs of theorems and propositions.

References

Athey, S., Chetty, R., Imbens, G. & Kang, H. (2020). Estimating treatment effects using multiple surrogates: The role of the surrogate score and the surrogate index. *arXiv:*1603.09326v3.

Athey, S., Imbens, G. W. & Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Statist. Soc.* B **80**, 597–623.

Barfield, R., Shen, J., Just, A. C., Vokonas, P. S., Schwartz, J., Baccarelli, A. A., VanderWeele, T. J. & Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genet. Epidem.* **41**, 824–33.

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85**, 233–98.

Bühlmann, P., Kalisch, M. & Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Ann. Rev. Statist. Appl.* **1**, 255–78.

Cai, T., Liu, W. & Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.

Cai, T. T. & Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45**, 615–46.

Chen, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D. & Lindquist, M. A. (2015). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19**, 121–36.

Danaher, P., Wang, P. & Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *J. R. Statist. Soc.* B **76**, 373–97.

Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, *p*-values and R-software HDI. *Statist. Sci.* **30**, 533–58.

Hanson, C., Cairns, J., Wang, L. & Sinha, S. (2016). Computational discovery of transcription factors associated with drug response. *Pharmacogenomics J.* **16**, 573–82.

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York: Guilford Press.

Hirshberg, D. A. & Wager, S. (2019). Augmented minimax linear estimation. *arXiv:*1712.00038v5.

Holmquist, E., Okroj, M., Nodin, B., Jirström, K. & Blom, A. M. (2013). Sushi domain-containing protein 4 (SUSD4) inhibits complement by disrupting the formation of the classical C3 convertase. *FASEB J.* **27**, 2355–66.

Huang, Y.-T., Liang, L., Moffatt, M. F., Cookson, W. O. & Lin, X. (2015). IGWAS: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genet. Epidem.* **39**, 347–56.

Huang, Y.-T. & Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics* **72**, 402–13.

Huang, Y.-T., VanderWeele, T. J. & Lin, X. (2014). Joint analysis of SNP and gene expression data in genetic association studies of complex diseases. *Ann. Appl. Statist.* **8**, 352–76.

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–909.

Javanmard, A. & Montanari, A. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *Ann. Statist.* **46**, 2593–622.

Kenny, D. A. & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychol. Sci.* **25**, 334–9.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. & Mesirov, J. P. (2011). Molecular signatures database (MSIGDB) 3.0. *Bioinformatics* **27**, 1739–40.

Loeys, T., Moerkerke, B. & Vansteelandt, S. (2015). A cautionary note on the power of the test for the indirect effect in mediation analysis. *Front. Psychol.* **5**, 1549.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N. et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Gene.* **45**, 580–85.

MacKinnon, D. P. (2008). *Introduction to Statistical Mediation Analysis*. London: Routledge.

Michlmayr, A., Bachleitner-Hofmann, T., Baumann, S., Marchetti-Deschmann, M., Rechweichselbraun, I., Burghuber, C., Pluschnig, U., Bartsch, R., Graf, A., Greil, R. et al. (2010). Modulation of plasma complement by the initial dose of epirubicin/docetaxel therapy in breast cancer and its predictive value. *Br. J. Cancer* **103**, 1201–8.

Niu, N., Schaid, D. J., Abo, R. P., Kalari, K., Fridley, B. L., Feng, Q., Jenkins, G., Batzler, A., Brisbin, A. G., Cunningham, J. M. et al. (2012). Genetic association with overall survival of taxane-treated lung cancer patients: A genome-wide association study in human lymphoblastoid cell lines followed by a clinical association study. *BMC Cancer* **12**, 422.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Gene.* **38**, 904–9.

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol. Methodol.* **13**, 290–312.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Nat. Acad. Sci.* **102**, 15545–50.

Sun, T. & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–98.

Van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.

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

VanderWeele, T. J. & Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *Am. J. Epidemiol.* **172**, 1339–48.

VanderWeele, T. J. & Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiol. Meth.* **2**, 95–115.

Wang, W., Baladandayuthapani, V., Morris, J. S., Broom, B. M., Manyam, G. & Do, K.-A. (2012). iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* **29**, 149–59.

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

Zhang, C.-H. & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–42.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E. et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32**, 3150–54.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

Zhao, S. D., Cai, T. T. & Li, H. (2014a). Direct estimation of differential networks. *Biometrika* **101**, 253–68.

Zhao, S. D., Cai, T. T. & Li, H. (2014b). More powerful genetic association testing via a new statistical framework for integrative genomics. *Biometrics* **70**, 881–90.

Zhu, Y. & Bradic, J. (2018). Linear hypothesis testing in dense high-dimensional linear models. *J. Am. Statist. Assoc.* **113**, 1583–600.

[*Received on* 14 *December* 2017. *Editorial decision on* 17 *October* 2019]