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

Ruixuan Rachel Zhou\textsuperscript{1}, Liewei Wang\textsuperscript{2}, and Sihai Dave Zhao\textsuperscript{1}

\textsuperscript{1}Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820
\textsuperscript{2}Division of Clinical Pharmacology, Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN 55905

May 11, 2019

Abstract

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, as well as 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.

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 and 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 \emph{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 (Van de Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014, 2018). The sparsity level of the regression parameter is not typically known. Cai et al. (2017) discussed the construction of confidence intervals that can adapt to this unknown sparsity, and Zhu and Bradic (2018) proposed a test that avoids the sparsity requirement entirely. 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 and Vansteelandt (2014), but it may be difficult to achieve valid inference, for reasons that will be explained in Section 2.1.

Several semiparametric high-dimensional methods have recently been proposed in the causal inference literature, for the purposes of doing 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 at the end of Section 2.2. However, these approaches do not directly apply to estimating indirect effects in high-dimensional mediation models. Chen et al. (2015) and Huang and 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 non-methylation pathway, giving rise to potential direct effects. We illustrate another application in Section 5, 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 and 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 Section 5, 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

For the $i$th 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 centered to have zero mean. We consider the following linear mediation model:

$$Y_i = G_i^T \alpha_0 + S_i^T \alpha_1 + \epsilon_{1i}, \quad G_i^T \alpha_0 = S_i^T \beta_0 + \epsilon_{2i}, \quad (1)$$

where $\epsilon_1$ and $\epsilon_2$ are independent mean-zero subexponential random error terms, with respective variances $\sigma_1^2$ and $\sigma_2^2$, that are independent of $G_i$ and $S_i$. We can also include additional covariates such as interaction terms, demographic characteristics, or environmental variables. For example, in our data analysis in Section 5 we include principal components to adjust for population stratification. The second regression model is non-standard because the outcome $G_i^T \alpha_0$ is not directly observed, as it contains the unknown parameter $\alpha_0$. Instead, the $G_i$ are usually modeled as arising from the $S_i$ according to some multivariate linear regression model. The advantage of directly assuming model (1) is that the association between $G_i$ and $S_i$ need not be explicit.

We assume that our model is correctly specified and that there are no unobserved founders. Then the parameters $\alpha_1$, and $\beta_0$ can correspond to well-defined causal effects per unit change in $S_i$. We present these relationships under a counterfactual framework in Section A of the Appendix, following similar results for low-dimensional mediation models (Huang et al., 2014; VanderWeele and Vansteelandt, 2014).

Our interest is in performing inference for the indirect effect $\beta_0$ when the dimension of $G_i$ exceeds the sample size. To see why this is challenging, first suppose $\alpha_0$ were known. Then the best estimator for $\beta_0$ would be the ordinary least squares estimator $(\sum_i S_i S_i^T)^{-1} (\sum_i S_i G_i^T \alpha_0)$. It is therefore tempting to use this expression along with an estimate $\hat{\alpha}_0$ of $\alpha_0$ obtained from a debiased lasso estimator (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014), which satisfies $n^{1/2}(\hat{\alpha}_0 - \alpha_0) = W + \Delta$, where $W$ has a multivariate normal distribution and $\Delta$ is a remainder term. For this strategy to give an asymptotically normal estimator for $\beta_0$, we would need $(\sum_i S_i S_i^T)^{-1} \sum_i S_i G_i^T \Delta = o_P(1)$. However, existing theory only guarantees that each component of $\Delta$ is $o_P(1)$, whereas we would need a linear combination of the components to be $o_P(1)$. This may in fact be possible, assuming either that $p \log(p)/n^{1/2} \to 0$ or that the correlations between $S_i$ and $G_i$ are sparse. This is discussed in more detail in Section B of the Appendix. In the remainder of this paper, however, we will develop new estimators that can achieve valid inference as long as $\log(p)/n^{1/2} \to 0$ and without restrictions on the sparsity of the $S_i$ and $G_i$ correlations.

2.2 Inference for indirect effect under incomplete mediation

We first consider the incomplete mediation setting where both direct and indirect effects may exist. Below, let $| \cdot |_1$ denote the element-wise $\ell_1$ norm and $| \cdot |_{\infty}$ denote the element-wise $\ell_{\infty}$ norm of either a vector or a matrix. We first use the scaled lasso of Sun and Zhang (2012) to fit the first regression in model (1), which also gives a consistent estimate of the
residual error $\sigma_i^2$. Alternatively, we could leave $\alpha_1$ unpenalized, and this is further discussed in Section 6. Let $Y = (Y_1, \ldots, Y_n)^\top$, $G$ be the $n \times p$ matrix $(G_1, \ldots, G_n)^\top$, $S$ be the $n \times q$ matrix $(S_1, \ldots, S_n)^\top$, and $X$ be the $n \times (p + q)$ design matrix $(G, S)$. Then the estimate $\hat{\alpha} = (\hat{\alpha}_0^\top, \hat{\alpha}_1^\top)^\top$ obtained by the scaled lasso satisfies the Karush-Kuhn-Tucker conditions

$$\lambda_n \hat{\kappa} = n^{-1}X^\top(Y - X\hat{\alpha}) = n^{-1}X^\top\{X(\alpha - \hat{\alpha}) + \epsilon_1\} = \hat{\Sigma}_{XX}(\alpha - \hat{\alpha}) + n^{-1}X^\top\epsilon_1,$$

where $\lambda_n$ is the tuning parameter from the scaled lasso, $\hat{\Sigma}_{XX}$ is the sample covariance matrix of the $X$, and $\epsilon_1 = (\epsilon_{11}, \ldots, \epsilon_{1n})^\top$. The vector $\hat{\kappa}$ arises from the subdifferential of $|\alpha|_1$ and satisfies

$$||\hat{\kappa}||_{\infty} \leq 1 \quad \text{and} \quad \hat{\kappa}_j = \text{sign}(\hat{\alpha}_j) \quad \text{if} \quad \hat{\alpha}_j \neq 0.$$

Next, suppose for now that we can find a matrix $\hat{\Omega}_{XX}$ such that

$$\hat{\Omega}_{XX}\hat{\Sigma}_{XX} \approx \begin{pmatrix} \hat{\Sigma}_{SG} & 0 \\ 0 & \hat{\Sigma}_{SS} \end{pmatrix} \equiv \hat{D},$$

where $\hat{\Sigma}_{SG} = n^{-1}\sum_i S_iG_i^\top$ and $\hat{\Sigma}_{SS} = n^{-1}\sum_i S_iS_i^\top$. Then premultiplying both sides of the Karush-Kuhn-Tucker conditions by $\hat{\Omega}_{XX}$ gives

$$\lambda_n\hat{\Omega}_{XX}\hat{\kappa} + \hat{D}(\hat{\alpha} - \alpha) = n^{-1}\hat{\Omega}_{XX}X^\top\epsilon_1 + \hat{\delta}, \quad (2)$$

where the remainder term $\hat{\delta} = (\hat{D} - \hat{\Omega}_{XX}\hat{\Sigma}_{XX})(\hat{\alpha} - \alpha)$. The second equation of model (1) then implies that

$$\hat{D}\alpha = \begin{pmatrix} \hat{\Sigma}_{SG}\alpha_0 \\ \hat{\Sigma}_{SS}\alpha_1 \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}_{SS}\beta_0 + n^{-1}S^\top\epsilon_2 \\ \hat{\Sigma}_{SS}\alpha_1 \end{pmatrix},$$

where $\epsilon_2 = (\epsilon_{21}, \ldots, \epsilon_{2n})^\top$. Then if $I_2$ denotes the $2 \times 2$ identity matrix and $\otimes$ denotes the Kronecker product, premultiplying both sides of (2) by $I_2 \otimes \hat{\Sigma}_{SS}^{-1}$ implies

$$I_2 \otimes \hat{\Sigma}_{SS}^{-1}\lambda_n\hat{\Omega}_{XX}\hat{\kappa} + \begin{pmatrix} \hat{\Sigma}_{SS}\hat{\Sigma}_{SG}\alpha_0 \\ \hat{\Sigma}_{SS}\alpha_1 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \alpha_1 \end{pmatrix}$$

$$= \begin{pmatrix} n^{-1}\hat{\Sigma}_{SS}^{-1}S^\top\epsilon_2 \\ 0 \end{pmatrix} + (I_2 \otimes \hat{\Sigma}_{SS}^{-1})(n^{-1}\hat{\Omega}_{XX}X^\top\epsilon_1 + \hat{\delta}). \quad (3)$$

The terms on the right-hand side of (3) will be asymptotically normal if $\hat{\delta}$ is small.

We therefore propose using $\hat{b}$ in the following expression to estimate the indirect effect:

$$\begin{pmatrix} \hat{b} \\ \hat{a} \end{pmatrix} = (I_2 \otimes \hat{\Sigma}_{SS}^{-1})\lambda_n\hat{\Omega}_{XX}\hat{\kappa} + \begin{pmatrix} \hat{\Sigma}_{SS}\hat{\Sigma}_{SG}\alpha_0 \\ \hat{\Sigma}_{SS}\alpha_1 \end{pmatrix}. \quad (4)$$

We show in Section 3 that under certain conditions, $\hat{b}$ is indeed asymptotically normal and centered at the true $\beta_0$. 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, (4) also gives an estimate $\hat{a}$ for the direct effect, which we prove is also consistent and asymptotically normal. As pointed out by a referee, an alternative way to estimate the direct effect in the presence of high-dimensional
mediators is to subtract $\hat{b}$ from the ordinary least squares estimate of the total effect of $S_i$ on $Y_i$. We show in Section C.3 in the Appendix that these two approaches are asymptotically equivalent.

It remains to find a suitable matrix $\hat{\Omega}_{XX}$. This would be simple if $\hat{\Sigma}_{XX}$ were invertible, but this is not the case when $(p + q) > n$. Instead, we propose to obtain $\hat{\Omega}_{XX}$ using constrained $\ell_1$ optimization, similar to the precision matrix estimation procedure of Cai et al. (2011):

$$\min |\Omega|_1 \text{ subject to } |\hat{D} - \hat{\Omega}\hat{\Sigma}_{XX}|_\infty \leq \tau_n,$$

where $\tau_n$ is a tuning parameter. We will show in Section 3 that $\hat{\Omega}_{XX}$ will converge to the true $\Omega_{XX} = E(\hat{D})\Sigma_{XX}^{-1}$ under the condition that $\Omega_{XX}$ is sparse, which is reasonable in many genomics problems. This will allow us to establish convergence of our estimates of the asymptotic variances of $\hat{b}$ and $\hat{a}$.

In addition to (5), there are other ways to compute a matrix $\hat{\Omega}_{XX}$ that satisfies $\hat{\Omega}_{XX}\hat{\Sigma}_{XX} \approx \hat{D}$. One approach might be to minimize the variances of $\hat{b}$ and $\hat{a}$ while controlling the coherence parameter $|\hat{D} - \hat{\Omega}\hat{\Sigma}_{XX}|_\infty$. In the standard linear regression setting with high-dimensional covariates, Javanmard and 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, this strategy might also obviate the need for a sparse $\Omega_{XX}$ in the present mediation setting. This may indeed be possible, and is an important direction for future work. Section E.5 of the Appendix contains a detailed discussion and a simulation study exploring the robustness of our procedure to the accuracy of estimating $\Omega_{XX}$.

The central idea behind our approach is that we develop a debiased estimator not of $\alpha_0$, but of the quantity $E(S_i G_i^\top \alpha_0)$. This allows us to obtain an asymptotically normal estimator for $\beta_0$ by simply premultiplying by the low-dimensional quantity $(n^{-1} \sum_i S_i S_i^\top)^{-1}$. A related strategy was proposed by Athey et al. (2018) for inference for a causal effect in the presence of high-dimensional controls, where the goal was also to estimate a linear combination of a high-dimensional vector like $\alpha_0$. Both strategies proceed by obtaining an initial estimate $\hat{\alpha}$ of $\alpha_0$ using penalized regression, taking the desired linear combination using $\hat{\alpha}$, then debiasing the linear combination by subtracting a weighted sum of the residuals from the $\alpha_0$ estimation. 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 $(I_2 \otimes \hat{\Sigma}_{SS}^{-1})\hat{\Omega}_{XX}$. A major difference between our approaches is that the coefficients of the desired linear combination are known in the setting of Athey et al. (2018), while in our mediation setting they are equal to $E(S_i G_i^\top)$ and must be estimated. The method of Athey et al. (2018) is not directly applicable here but it would be interesting to apply their method to our mediation framework in the future.

### 2.3 Inference for indirect effect under complete mediation

In some applications, for example in the analysis of noncoding genetic variants, it may be known that exposure acts only indirectly on the outcome. We can make use of this extra information to develop a more efficient procedure for estimating $\beta_0$. As before, we fit the first regression model in (1) using the scaled lasso, though this time we set $\alpha_1 = 0$. The estimate $\tilde{\alpha}_0$ now satisfies new Karush-Kuhn-Tucker conditions

$$\lambda_n \hat{K} = \hat{\Sigma}_{GG}(\alpha_0 - \tilde{\alpha}_0) + n^{-1} G^\top \epsilon_1,$$

where $\hat{K}$ is the scaled lasso.

5
where $\hat{\Sigma}_{GG} = n^{-1} \sum_i G_i G_i^\top$. We define $\hat{\Omega}_{GG}$ to be the solution to

$$
\min |\Omega|_1 \text{ subject to } |\hat{\Sigma}_{SG} - \Omega \hat{\Sigma}_{GG}|_{\infty} \leq \tau_n',
$$

(6)

where $|\hat{\Sigma}_{SG} - \Omega \hat{\Sigma}_{GG}|_{\infty}$ is the new coherence parameter under complete mediation. Premultiplying both sides of the Karush-Kuhn-Tucker conditions by $\hat{\Omega}_{GG}$ gives

$$
\lambda_n \hat{\Omega}_{GG} \tilde{\kappa} + \hat{\Sigma}_{SS}^{-1} \hat{\Sigma}_{SG}(\tilde{\alpha}_0 - \alpha_0) = n^{-1} \hat{\Omega}_{GG} G^\top \epsilon_1 + \hat{\delta},
$$

where now $\hat{\delta} = (\hat{\Sigma}_{SG} - \hat{\Omega}_{GG} \hat{\Sigma}_{GG})(\tilde{\alpha}_0 - \alpha_0)$. Premultiplying by $\hat{\Sigma}_{SS}^{-1}$ and using the second equation of model (1) gives

$$
\hat{\Sigma}_{SS}^{-1}(\lambda_n \hat{\Omega}_{GG} \tilde{\kappa} + \hat{\Sigma}_{SG} \tilde{\alpha}_0) - \beta_0 = \hat{\Sigma}_{SS}^{-1}(n^{-1} \hat{\Omega}_{GG} G^\top \epsilon_1 + n^{-1} S^\top \epsilon_2 + \hat{\delta}).
$$

(7)

We therefore propose the following estimator for the indirect effect under complete mediation:

$$
\tilde{b} = \hat{\Sigma}_{SS}^{-1} \hat{\Sigma}_{SG} \tilde{\alpha}_0 + \hat{\Sigma}_{SS}^{-1} \lambda_n \hat{\Omega}_{GG} \tilde{\kappa}.
$$

(8)

We show in Section 3 that $\tilde{b}$ is asymptotically normal and centered at the true $\beta_0$, and provide estimates for its asymptotic variance.

Additionally, this estimator $\tilde{b}$ under complete mediation has an interesting efficiency property. When the indirect effect is equal to the total effect, $\beta_0$ 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 $\beta_0$ 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 and 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 $G_i^\top \tilde{\alpha}_0$ of its conditional expectation $G_i^\top \alpha_0$ and thus removing much of the variation from the error term $\epsilon_1$. However, this denoised outcome is biased, because of the bias in the lasso estimator $\tilde{\alpha}_0$. Removing this bias is the purpose of the second term in our proposed estimator (8). Alternatively, the incomplete mediation estimator $\hat{b}$ proposed in (4) can be used to estimate $\beta_0$ under complete mediation as well, but we also show that $\tilde{b}$ has a lower asymptotic variance compared to $\hat{b}$. In a closely related setting, Athey et al. (2016) found that leveraging surrogate outcomes can increase the efficiency of estimating treatment effects. These results are discussed in detail in Section 3.2.

2.4 Implementation

We first center the outcome, the gene expression values, and the genotype values. To apply the scaled lasso, we standardize all covariates to have unit variance and then choose the tuning parameter $\lambda_n$ using the quantile-based penalty procedure in the R package scalreg.

To estimate the asymptotic variances of our estimators, given in Theorems 2 and 4, we need estimates of the residual variances $\sigma_1^2$ and $\sigma_2^2$ from our mediation model (1). Sun and Zhang (2012) showed that the scaled lasso can provide a consistent estimate $\hat{\sigma}_1^2$ for $\sigma_1^2$. Since model (1) implies that $Y_i = S_i^\top (\beta_0 + \alpha_1) + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma_1^2 + \sigma_2^2)$, we can estimate $\sigma_2^2$ by first regressing $Y_i$ on $S_i$ to obtain the ordinary least squares residual variance estimator.
solving the constrained \( \ell_1 \) this estimation by taking advantage of fast algorithms from the Theorems 1 and 3 that this is equivalent to solving multiple D antzig selectors. We implement choosing the tuning parameter \( \tau \) the sparse matrix estimator (6) by choosing the difference of two precision matrices. Similarly, under complete mediation we implement XX under incomplete mediation. Define the vector \( \hat{\sigma}_i \) XX \( \hat{\sigma}_i \) is likely occurs when no mediators are associated with the outcome, i.e., \( \alpha_0 = 0 \), in which case \( \hat{\sigma}_2 \) indeed equals zero.

Under incomplete mediation, we implement the proposed sparse matrix estimate (5) by choosing the tuning parameter \( \tau_n \) to minimize the Akaike Information Criterion-type measure \( n|\tilde{D} - \hat{\Omega}_{XX}\hat{\Sigma}_{XX}|_\infty + 2|\hat{\Omega}_{XX}|_0 \), where \( |\hat{\Omega}_{XX}|_0 \) denotes the number of non-zero elements in \( \hat{\Omega}_{XX} \). This was motivated by the tuning criterion used by Zhao et al. (2014a) to estimate \( \hat{\sigma} \). This is sensible because \( \hat{\sigma} \) likely occurs when no mediators are associated with the outcome, i.e., \( \alpha_0 = 0 \), in which case \( \hat{\sigma}_2 \) indeed equals zero.

The time-consuming part of our method is estimating the sparse matrix \( \hat{\Omega}_{XX} \) or \( \hat{\Omega}_{GG} \) by solving the constrained \( \ell_1 \) optimization problems (5) or (6). It is justified in the proofs of Theorems 1 and 3 that this is equivalent to solving multiple Dantzig selectors. We implement XX \( \hat{\Omega}_{XX} \) or \( \hat{\Omega}_{GG} \) by solving the constrained \( \ell_1 \) optimization problems (5) or (6). It is justified in the proofs of Theorems 1 and 3 that this is equivalent to solving multiple Dantzig selectors. We implement XX \( \hat{\Omega}_{XX} \) or \( \hat{\Omega}_{GG} \) by solving the constrained \( \ell_1 \) optimization problems (5) or (6). It is justified in the proofs of Theorems 1 and 3 that this is equivalent to solving multiple Dantzig selectors. We implement XX \( \hat{\Omega}_{XX} \) or \( \hat{\Omega}_{GG} \) by solving the constrained \( \ell_1 \) optimization problems (5) or (6).

This section presents the theoretical properties of our proposed indirect effect inference procedure under incomplete mediation. Define the vector \( X_i = (G_{i1}, \ldots, G_{ip}, S_i)^T \). We will first need several sparsity assumptions.

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

This section presents the theoretical properties of our proposed indirect effect inference procedure under incomplete mediation. Define the vector \( X_i = (G_{i1}, \ldots, G_{ip}, S_i)^T \). We will first need several sparsity assumptions.

**Assumption 2** Let \( \Sigma_{SG} = E(S_i G_i^T) \), \( \Sigma_{GG} = E(G_i G_i^T) \), \( \Sigma_{GS} = E(G_i S_i^T) \), \( \Sigma_{SS} = E(S_i S_i^T) \) and \( \Sigma_{XX} = E(X_i X_i^T) \). Define

\[
D = \begin{pmatrix}
\Sigma_{SG} & 0 \\
0 & \Sigma_{SS}
\end{pmatrix}.
\]

There exist constants \( M_X \) and \( N_X \) such that \( \|\Sigma_{XX}^{-1}\|_{L_1} \leq M_X \) and \( \|(D \Sigma_{XX}^{-1})^T\|_{L_1} \leq N_X \). Furthermore if \( \omega_{ij} \) denotes the \( ij \)th entry of \( D \Sigma_{XX}^{-1} \), then \( \max_i \sum_j |\omega_{ij}|^\theta \leq s_0 \) 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 \( \|\Sigma_{XX}^{-1}\|_{L_1} \) essentially requires that none of the rows contain too many large entries. This is reasonable, as entries of the precision matrix are frequently used to model the conditional dependence 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 and Yu (2006), and is similar to requiring that \( S_i \) cannot be completely explained by \( G_i \).

**Theorem 1** Let \( \hat{\Omega}_{XX} \) solve (5) with tuning parameter \( \tau_n = (N_X + 1)C_1\{(\log(p + q))/n\}^{1/2} \) for \( C_1 = 2\eta^{-2}(2 + \tau + \eta^{-1}e^2K^2)^2 \), where \( K \) and \( \eta \) are from Assumption 2 and \( \tau > 0 \). Define the population-level matrix

\[
\Omega_{XX} \equiv D\Sigma_{XX}^{-1} = \left( \begin{array}{cc} \Sigma_{SG} & 0 \\ 0 & \Sigma_{SS} \end{array} \right) \left( \begin{array}{cc} \Sigma_{GG} & \Sigma_{GS} \\ \Sigma_{SG} & \Sigma_{SS} \end{array} \right)^{-1}.
\]

Then under Assumptions 1 and 2, \( |\hat{\Omega}_{XX} - \Omega_{XX}| \leq (4N_X + 2)C_1M_X\{(\log p)/n\}^{1/2} \) with probability greater than \( (1 - 4p^{-\tau}) \).

Theorem 1 shows that our \( \hat{\Omega}_{XX} \) (5) is a consistent estimate of the population-level matrix \( \Omega_{XX} \). As discussed in Section 2.2, in the standard linear regression setting, Javanmard and Montanari (2014) proposed a method for high-dimensional inference that does not require consistent estimation of precision matrices. Section E.5 in the Appendix discusses 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 behavior of our incomplete mediation estimators \( \hat{\beta}, \hat{\alpha} \) (4). We require additional assumptions necessary for the good performance of the scaled lasso of Sun and Zhang (2012).

**Theorem 2** Let \( \hat{\beta} \) and \( \hat{\alpha} \) be calculated such that both tuning parameters \( \lambda_n \) and \( \tau_n \) are of order \( O\{(n^{-1}\log p)^{1/2}\} \). Assume the model for \( Y_i \) in mediation model (1) satisfies the conditions of Theorem 2 of Sun and Zhang (2012) and that \( \alpha_0 \) has at most \( s_0 = o(n^{1/2}/\log p) \) non-zero components. Under Assumptions 1 and 2, if \( (\log p)/n^{1/2} \to 0 \) and \( \alpha_0 \) and \( \Sigma_{SG} \) are not both zero, and \( \Gamma \equiv \Sigma_{SS}^{-1}\Sigma_{SG}(\Sigma_{GG} - \Sigma_{GS}\Sigma_{SS}^{-1}\Sigma_{SG})^{-1}\Sigma_{GS}\Sigma_{SS}^{-1} \) converges to a positive-definite \( q \times q \) matrix, then

\[
n^{1/2} \left( \hat{\beta} - \beta_0 \right) \to N(0, V), \quad \text{where} \quad V = \left( \begin{array}{cc} \sigma^2\Gamma + \sigma^2\Sigma_{SS}^{-1} & -\sigma^2\Gamma \\ -\sigma^2\Gamma & \sigma^2(\Gamma + \Sigma_{SS}^{-1}) \end{array} \right).
\]

The ultra-sparsity assumption on \( s_0 \) in Theorem 2 is standard in the de-biased lasso literature (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014), though recently Zhu and Bradic (2018) proposed a method that does not need to assume ultra-sparsity. The choice of \( \tau_n \) controls the coherence parameter \( |\hat{\Omega}_{XX}\hat{\Sigma}_{XX} - \hat{D}|_\infty \) 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. Theorem 2 and equation (3) indicate that the asymptotic variance \( V \) can be consistently estimated using

\[
\hat{\sigma}^2(I_2 \otimes \hat{\Sigma}_{SS}^{-1})\hat{\Omega}_{XX}\hat{\Sigma}_{XX}\hat{\Omega}_{XX}^\top(I_2 \otimes \hat{\Sigma}_{SS}^{-1}) + \left( \begin{array}{cc} \hat{\sigma}^2\Sigma_{SS}^{-1} & 0 \\ 0 & 0 \end{array} \right).
\]
Consistency of the sample covariance estimates \( \hat{\Sigma}_{XX} \) and \( \hat{\Sigma}_{SS} \) is standard, and consistency of \( \hat{\Omega}_{XX} \) is given by Theorem 1. Estimation of \( \sigma_1^2 \) and \( \sigma_2^2 \) was discussed in Section 2.4.

We caution that Theorem 2 does not cover the setting where both \( \Sigma_{SG} = 0 \) and \( \alpha_0 = 0 \). This would cause terms on the right-hand side of (3) to exactly equal zero, which would invalidate the normal approximation in Theorem 2. A similar 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 is indeed reflected in our simulations in Section 4.2.

### 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.

**Assumption 3** There exist constants \( M_G \) and \( N_G \) such that \( \| \Sigma_{GG}^{-1} \|_1 \leq M_G \) and \( \| (\Sigma_{SG} \Sigma_{GG}^{-1})^\top \|_1 \leq N_G \). Furthermore, if \( \omega_{ij} \) denotes the \( ij \)th entry of \( \Sigma_{SG} \Sigma_{GG}^{-1} \), then \( \max_i \sum_j | \omega_{ij} |^\theta < s_0 \) for some \( \theta \in [0, 1) \).

Using these sparsity assumptions, we can show that \( \hat{\Omega}_{GG} \) calculated using (6) is a consistent estimate of \( \Omega_{GG} \).

**Theorem 3** Let \( \hat{\Omega}_{GG} \) solves (6) with tuning parameter \( \tau_n = (N_G + 1)C_1((\log p)/n)^{1/2} \). Define the population-level matrix \( \Omega_{GG} = \Sigma_{SG} \Sigma_{GG}^{-1} \). Then under Assumptions 1 and 3, \( \| \hat{\Omega}_{GG} - \Omega_{GG} \|_\infty \leq (4N_G + 2)C_1 M_G ((\log p)/n)^{1/2} \) with probability greater than \( (1 - 4p^{-\tau}) \), where \( C_1 \) and \( \tau \) are as in Theorem 1.

We can now characterize the asymptotic behavior of our complete mediation indirect effect estimator \( \tilde{b} \) (8). Equation (7) and Theorem 4 below indicate that the asymptotic variance of \( \tilde{b} \) can be consistently estimated by \( \sigma_1^2 \hat{\Sigma}_{SS}^{-1} \hat{\Omega}_{GG} \hat{\Sigma}_{GG} \hat{\Sigma}_{SS}^{-1} + \sigma_2^2 \hat{\Sigma}_{SS}^{-1} \). As in the incomplete mediation case discussed above, the sparsity assumptions on \( \Omega_{GG} \) in Assumption 3 are crucial to proving Theorem 4, and the requirement that \( \alpha_0 \) and \( \Sigma_{SG} \) are not both zero also arises here.

**Theorem 4** Let \( \tilde{b} \) be calculated such that both tuning parameters \( \lambda_n \) and \( \tau_n \) are of order \( O \left( (n^{-1} \log p)^{1/2} \right) \). Assume the model for \( Y \) in mediation model (1) has \( \alpha_1 = 0 \) but otherwise satisfies the conditions of Theorem 2 of Sun and Zhang (2012) and that \( \alpha_0 \) has at most \( s_0 = o(n^{1/2} / \log p) \) non-zero components. Under Assumptions 1 and 3, if \( (\log p)/n^{1/2} \to 0 \), \( \alpha_0 \) and \( \Sigma_{SG} \) are not both zero, and \( \Gamma \equiv \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS} \) converges to a positive-definite \( q \times q \) matrix as \( n, p \to \infty \), then

\[
n^{1/2} (\tilde{b} - \beta_0) \to N \left( 0, \sigma_1^2 \Sigma_{SS}^{-1} \Gamma \Sigma_{SS}^{-1} + \sigma_2^2 \Sigma_{SS}^{-1} \right).
\]
To prove Theorem 4 it is important to control the coherence parameter $|\hat{\Omega}_{GG}\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG}|_\infty$ at rate $\tau_n \asymp (n^{-1}\log p)^{1/2}$, so that the bias of our proposed $\tilde{b}$ goes to 0 when $n$ and $p$ go to infinity. As mentioned in Section 2.3, under complete mediation the indirect effect $\beta_0$ 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 $\tilde{b}$ with the ordinary least squares estimate of $\beta_0$.

**Proposition 1** In model (1), assume that the direct effect $\alpha_1 = 0$, so that the ordinary least squares estimator $\tilde{b}_{OLS} = (S^\top S)^{-1}S^\top Y$ is a consistent estimator of $\beta_0$. Then under the conditions of Theorem 4, $\text{var}\{n^{1/2}(\tilde{b}_{OLS} - \beta_0)\} - \text{var}\{n^{1/2}(\tilde{b} - \beta_0)\}$ converges to a positive semi-definite 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. In fact, it can be shown that under a slightly more restrictive model than model (1), our estimator $\tilde{b}$ achieves the minimum asymptotic variance among all asymptotic unbiased estimators of $\beta_0$; see Section D of the Appendix. The intuition behind this phenomenon was explained in Section 2.3. In practice, the Wald test based on $\tilde{b}$ can still be conservative when $\alpha_0$ and $\Sigma_{SG}$ are close to zero, for reasons discussed in Section 3.1, but simulation results in Section 4.3 show that our proposed $\tilde{b}$ can still have significant power gains compared to the ordinary least squares estimator over the majority of the parameter space.

Proposition 1 extends similar findings in low dimensions (Kenny and Judd, 2014; Zhao et al., 2014b; Loeys et al., 2015). In an apparently different setting, Athey et al. (2016) found that when estimating the causal effect of a treatment on a long-term outcome, leveraging intermediate outcomes can increase efficiency. This setting is closely related to complete mediation, and in Section C.2 of the Supplementary materials 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 power to detect important biological signals (Wang et al., 2012; Huang et al., 2015; Zhu et al., 2016).

On the other hand, the improved power guaranteed by Proposition 1 depends on the correct specification of both stages of the linear mediation model (1), and is subject to bias if either model is incorrect. 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 modeling assumptions. An important direction for future work is to develop nonparametric mediation models that allow for flexible relationships between the outcome and the mediators.

As mentioned in Section 2.3, under complete mediation the indirect effect $\beta_0$ can also be consistently estimated by our $\tilde{b}$ (4), proposed under incomplete mediation. We can now use Theorems 2 and 4 to analytically compare $\tilde{b}$ and $\tilde{b}$.

**Proposition 2** In model (1), assume that the direct effect $\alpha_1 = 0$. Under the conditions of Theorem 2 and Theorem 4, $\text{var}\{n^{1/2}(\tilde{b} - \beta_0)\} - \text{var}\{n^{1/2}(\tilde{b} - \beta_0)\}$ converges to a positive semi-definite matrix.

Proposition 2 shows that under complete mediation, $\tilde{b}$ is asymptotically more efficient. This makes sense because $\tilde{b}$ uses the extra information that $\alpha_1 = 0$ under complete mediation.
4 Numerical Results

4.1 Methods Compared

We implemented our proposed procedures as described in Section 2.4. To reduce the computational burden of parameter tuning in large-scale problems, we also experimented with setting $\tau_n$ to a fixed value. Simulation results in Section E.1 in the Appendix show that $\tau_n = \{(\log p)/n\}^{1/2}/3$ maintained type I error and gave good results.

We compared our proposed 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 et al., 2010), which possesses the oracle property. Appealing to this property, the method computes standard errors and $p$-values for each mediator’s effect and then Bonferroni-corrects for the number of retained mediators. Next, the method regresses each mediator on the exposure using ordinary least squares to obtain Bonferroni-adjusted $p$-values. We reject the null hypothesis of no indirect effect if there is at least one mediator whose adjusted $p$-values, for associations with the outcome and the exposure, are both less than the desired significance level.

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.

4.2 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, $\gamma$ was a $p \times 1$ coefficient vector, and $E_i \sim N(0, \Sigma_E)$. We generated $\Sigma_E$ following procedures in Danaher et al. (2014) such that $\Sigma_E^{-1}$ was sparse in the sense of Assumption 2 and had diagonal entries equal to 1. Finally, we generated the outcome according to $Y_i = G_i^T \alpha_0 + S_i^T \alpha_1 + \epsilon_{1i}$, where $\epsilon_{1i} \sim N(0, 5)$.

In Section E.4 of the Appendix, we show that our simulation results were similar even when $\epsilon_{1i}$ was not normally distributed. We let $\gamma$ have 15 non-zero components randomly generated between $[-1, 1]$, fixing $\gamma$ across replications, and let $\alpha_0$ have 15 non-zero components equal to one. We chose either one or five of these non-zero components to correspond to variables whose entries in $\gamma$ were also non-zero; these were the true mediators.

In these simulations, we used our proposed $\hat{b}$ as well as the method of Zhang et al. (2016) to test $H_0 : \beta_0 = 0$ at the $\alpha = 0.05$ significance level. Figure 1 reports average power curves over 200 replications and shows that both tests were able to maintain the nominal significance level. The power curves were not always symmetric because changing $\beta_0$ also changed the degree of multicollinearity between $G_i$ and $S_i$ in the regression model for $Y_i$ in model (1), affecting the power. The method of Zhang et al. (2016) was very powerful when $\beta_0$ was small but had counterintuitive behavior when $\beta_0$ is large. The conservative nature of our test for $\beta_0$ close to zero arises from the error of the normal approximation to the distribution of Wald-type test statistics for indirect effects under weak mediation, discussed in Section 3.1.
Figure 1: Average power curves for testing the indirect effect under incomplete mediation, at significance level $\alpha = 0.05$, over 200 replications. The direct effect was $\alpha_1 = 0.1$ in the upper panels and $\alpha_1 = 0.5$ in the lower panels. The number of true mediators was 1 in the left panels and 5 in the right panels. Proposed: $\hat{b}$ from (4); Zhang: method of Zhang et al. (2016).
Figure 2: Average power curves for testing the indirect effect under complete mediation, at significance level $\alpha = 0.05$, over 200 replications. Examples 1 and 2: 500 potential mediators with 1 or 5 true mediators, respectively; Examples 3 and 4: 1,000 potential mediators with 1 or 5 true mediators, respectively. Proposed: $\tilde{b}$ from (8); OLS: ordinary least squares estimate; Zhang: the method of Zhang et al. (2016).

4.3 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 in Section 4.2, where $\alpha_1 = 0$ in each setting. We generated either $p = 500$ or $p = 1,000$ potential mediators, with either 1 or 5 true mediators. We used our proposed $\tilde{b}$ and the method of Zhang et al. (2016) to test $H_0 : \beta_0 = 0$ at the $\alpha = 0.05$ significance level. Under complete mediation we can also use ordinary least squares to estimate and test $\beta_0$. Figure 2 compares the average power curves of our $\tilde{b}$, the method of Zhang et al. (2016), and ordinary least squares, in our four simulation examples, over 200 replications.

Our proposed method was always able to maintain the nominal significance level, and in every case had higher power than ordinary least squares for sufficiently large $\beta_0$, consistent with Proposition 1. Similar to the incomplete mediation setting, the method of Zhang et al. (2016) was very powerful when $\beta_0$ is small and counterintuitive behavior when $\beta_0$ is large. For the same reason as in the incomplete mediation setting, our proposed method was overly conservative for $\beta_0$ close to zero, but nevertheless remained valid. Furthermore, we show
below that in real data, our approach can indeed identify more significant noncoding single nucleotide polymorphisms compared to ordinary least squares and the method of Zhang et al. (2016).

5 Data analysis

5.1 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 lymphoblasticoid cell lines, which was also studied by Niu et al. (2012) and Hanson et al. (2016). Niu et al. (2010) collected genotype data on 1,362,849 single nucleotide polymorphisms and expression data on 54,613 probes, after preprocessing, 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).

5.2 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, resulting in a high-dimensional mediation analysis problem. Note that 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.

Here we applied our proposed inference procedure 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 \(\hat{b}\) (4), because the gene of interest may have a direct effect on the phenotype that does not proceed through the candidate gene set. For computational convenience, we chose a uniform tuning parameter of \(\tau_n = \{(\log p)/n\}^{1/2}/3\) to estimate \(\hat{\Omega}_{XX}\) (5), discussed in Section 4.1. As an illustrative example, we studied the effect 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 resource (Subramanian et al., 2005; Liberzon et al., 2011), then applied our proposed approach to test the indirect effect of TMED10 through each of the 4,436 candidates. Of these, 420 gene sets contained more genes than there were samples, making our high-dimensional approach indispensable. We could not compare to the method of Zhang et al. (2016) because their R implementation failed on the low-dimensional gene sets.

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

Table 1: Top 10 most significant gene sets through which the TMED10 gene may act on drug response. 95% CI: confidence intervals obtained from the proposed method under incomplete mediation (4); p-value: p-values using the proposed procedure to test the indirect effect.

Actually subgroups of larger sets. Table 1 reports the top 10 most significant ones, as ranked by their indirect effect p-values. Many of these gene sets 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.

5.3 Non-coding Variants Analysis

We next studied the effects of non-coding genetic variants on the response to docetaxel. The standard genome-wide association study approach of regressing docetaxel EC_{50} on each variant separately, controlling for the first five principal components, did not identify any significant variants after multiple testing correction. We were interested in whether a high-dimensional mediation analysis method could provide more power. We chose the top 1,000 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.

We next applied our complete mediation estimator \( \hat{b} \) (8), using \( \tau_n' = \{(\log p)/n\}^{1/2}/3 \) in the estimation of \( \hat{\Omega}_{GG} \) (6), 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 non-coding 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. Indeed, there is clinical evidence that the complement system is involved in the response to epirubicin/docetaxel treatment in breast cancer patients (Michlmayr et al., 2010), suggesting that rs11578000 and its mechanisms of action may be worth further study. Our \( \hat{b} \) provides novel findings that could not have been detected using standard approaches.
6 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 may be affected by the outcome, rather than cause the outcome. Our method’s findings should thus be further analyzed 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.

We have so far assumed that the residual errors $\epsilon_{1i}$ and $\epsilon_{2i}$ are independent of the exposure $S_i$ and mediator $G_i$ in model (1). Under heteroskedasticity, if the errors are dependent on either $S_i$ or $G_i$, our theoretical results will likely not hold because the terms on the right-hand sides of the Karush-Kuhn-Tucker conditions (3) and (7) will in general no longer be asymptotically normal. Extending our approach to heteroskedastic errors is an important research direction.

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 Section 2.2. We explored using $\hat{a}$ to test for the presence of a direct effect, and similar to Kenny and Judd (2014), we found that the power was relatively low. This may be because when calculating our estimators, we penalize the direct effect parameter $\alpha_1$ 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 $\alpha_1$ 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 $\alpha_1$ 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 $\hat{\alpha}_1$ 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 and Vansteelandt, 2010; VanderWeele, 2015).

Acknowledgements

The authors would like to 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 a grant awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The work of Dave Zhao was funded in part by the National Science Foundation.
References


A Causal interpretation of direct and indirect effect

Suppose in a mediation model we have scalar exposure $S$, scalar outcome $Y$, and $p$ potential mediators denoted by $G = (G_1, \ldots, G_p)^\top$. Let $G_s$ be the counterfactual value of $G$ if exposure $S$ were set to the value $s$ and let $Y_{sg}$ denote the value of outcome $Y$ if $S$ were set to the value $s$ and $G$ were set to $g$. The controlled direct effect on $Y$ of changing exposure $S$ from $s$ to $s^*$ is defined to be $Y_{sg} - Y_{s^*g}$, where the value of $G$ is kept at $g$. This measures the portion of the effect of $S$ on $Y$ that does not proceed through $G$. The natural direct effect of exposure $S$ on outcome $Y$ is defined as $Y_{sG_s} - Y_{s^*G_{s^*}}$. Finally, the natural indirect effect is defined as $Y_{sG_s} - Y_{s^*G_{s^*}}$. This measures the amount of the effect of $S$ on $Y$ that is due only to a change in $G$, while the value of $S$ itself is held constant at $s$.

Next, consider the model

$$
Y_i = G_i^\top \alpha_0 + S_i^\top \alpha_1 + \epsilon_i, \quad G_i = \gamma_0 S_i + E_i,
$$

where $\gamma_0$ is $p \times 1$ vector and $E_i$ is a $p \times 1$ mean-zero error vector with covariance matrix $\Sigma_E$. This implies mediation model (1) in the main text, and is essentially model (5) of VanderWeele and Vansteelandt (2014), assuming that there are no confounders. Then following the same derivation from Section 3.2 of VanderWeele and Vansteelandt (2014), it can be shown that

$$
E(Y_{sg} - Y_{s^*g}) = \alpha_1(s - s^*),
$$

$$
E(Y_{sG_s} - Y_{s^*G_{s^*}}) = \alpha_1(s - s^*),
$$

$$
E(Y_{sG_s} - Y_{s^*G_{s^*}}) = \gamma_0\alpha_0^\top(s - s^*).
$$

Thus our $\alpha_1$ can be interpreted as the average controlled and natural direct effects, and $\beta_0$ can be interpreted as the average natural indirect effect, of a one-unit change in the exposure $S$. The detailed proof can be found in the Appendix of VanderWeele and Vansteelandt (2014).

B Conditions for directly using a debiased estimator of $\alpha_0$

In Section 2.1 of the main text we argued that the reason we do not opt for plugging in a de-biased lasso estimator in $(\sum_i S_i S_i^\top)^{-1} (\sum_i S_i G_i^\top \alpha_0)$ is that then it will not be guaranteed that the bias of the estimator is $o_P(1)$. This may in fact be possible, assuming either that $p \log(p)/n^{1/2} \to 0$ or that the correlations between $S_i$ and $G_i$ are sparse, and here we give a detailed reasoning by look at the de-biased lasso estimator $\alpha_0$ in detail.

Recall that our model is
\[ Y = G\alpha_0 + \epsilon_1, \quad G\alpha_0 = S\beta_0 + \epsilon_2. \] (10)

If we directly used a de-biased estimator \( \hat{\alpha} \) for \( \alpha \), the estimator for the indirect effect \( \beta_0 \) would be \( \hat{\beta} = (S^T S)^{-1}S^T G\hat{\alpha} \), and we would have

\[
\hat{\beta} - \beta_0 = (S^T S)^{-1}S^T G(\hat{\alpha} - \alpha_0) + (S^T S)^{-1}S^T G\alpha_0 - \beta_0 \\
= \hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}(\hat{\alpha} - \alpha_0) + (S^T S)^{-1}S^T (S\beta_0 + \epsilon_2) - \beta_0,
\]

so

\[
n^{1/2}(\hat{\beta} - \beta_0) = n^{1/2}\hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}(\hat{\alpha} - \alpha_0) + n^{-1}\hat{\Sigma}_{SS}^{-1}S^T \epsilon_2.
\]

Based on properties of the debiased lasso estimator (Javanmard and Montanari, 2014),

\[
n^{1/2}(\hat{\alpha} - \alpha_0) = Z + \Delta,
\]

where \( Z = n^{1/2}MG^T \epsilon_1 \) and \( \Delta = n^{1/2}(M\hat{\Sigma}_{GG} - I)(\hat{\alpha} - \alpha_0) \), and \( M \) is chosen so that \( |M\hat{\Sigma}_{GG} - I|_\infty \lesssim (\log(p)/n)^{1/2} \). Since the debiased \( \hat{\alpha} \) satisfies \( |\hat{\alpha} - \alpha_0|_1 = O_P\{(\log(p)/n)^{1/2}\} \),

\[
|\Delta|_\infty = |n^{1/2}(M\hat{\Sigma}_{GG} - I)(\hat{\alpha} - \alpha_0)|_\infty \\
\leq n^{1/2}|M\hat{\Sigma}_{GG} - I|_\infty |\hat{\alpha} - \alpha_0|_1 \\
= O_P\{\log(p)/n^{1/2}\}.
\]

Then

\[
n^{1/2}(\hat{\beta} - \beta_0) = \hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}Z + n^{-1}\hat{\Sigma}_{SS}^{-1}S^T \epsilon_2 + \hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\Delta,
\]

and the first two terms can be shown to be asymptotically normal.

The difficulty arises when trying to show that the last term is \( o_P(1) \). The most straightforward analysis gives

\[
|\hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\Delta|_\infty \lesssim \|\hat{\Sigma}_{SS}^{-1}\|_1(\|\Sigma_{SG}\|_1|\Delta|_\infty + \|\hat{\Sigma}_{SG} - \Sigma_{SG}\|_1|\Delta|_\infty),
\]

where \( \| \cdot \|_1 \) denotes the maximum row-wise \( L_1 \) norm. Without any additional assumptions, \( \|\Sigma_{SG}\|_1 = O(p) \), so \( |\hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\Delta|_\infty = O_P\{p\log(p)/n^{1/2}\} \). Therefore, if \( p \) and \( n \) are such that \( p\log(p)/n^{1/2} \to 0 \), this term will be \( o_P(1) \). Alternatively, we might assume that each row of \( \Sigma_{SG} \) has only at most \( o\{n^{1/2}/\log(p)\} \) non-zero components, in which case the term will be \( o_P(1) \) as well.

C Relationship with other methods

C.1 Comparing with Athey et al. (2018)

Here we compare our method and the method of Athey et al. (2018), with both attempt to debias a linear combination of a sparse high-dimensional vector. First, Athey et al. (2018)
adopt the potential outcomes framework where they observe a treatment $W_i$, the covariates $X_i$, and outcome $Y_i^{\text{obs}}$ for each experimental unit $i$. We adopt a mediation framework with exposure $S_i$, mediators $G_i$, and outcome $Y_i$. Their treatment variable $W_i$ is analogous to our exposure $S_i$, though $W_i$ can only take values 0 or 1 while $S_i$ is under no restriction.

Second, Athey et al. (2018) assume that $Y_i^{\text{obs}}$ and $X_i$ are related by the linear model $E(Y_i(0) \mid X = x) = x^T \beta_c$ in the control group and $E(Y_i(1) \mid X = x) = x^T \beta_t$ in the treatment group. This is analogous to the first regression model $Y_i = G_i^T \alpha_0 + S_i^T \alpha_1 + \epsilon_i$ in our mediation model (1) in the main text. However, our second model $G_i^T \alpha_0 = S_i^T \beta_0 + \epsilon_2$ has no analog in the framework of Athey et al. (2018), as their $X_i$ are not related to their $W_i$.

Third, the objective of Athey et al. (2018) is to estimate the average treatment effect 

$$\tau = \frac{1}{n_t} \sum_{i:W_i=1} E(Y_i(1) - Y_i(0) \mid X_i),$$

where $n_t$ is the size of the treatment group. In contrast, we have a very different goal, which is to estimate an indirect effect.

Fourth, Athey et al. (2018) note that the average treatment effect can be written as 

$$\tau = X_t \beta_t - X_t \beta_c,$$

where $X_t = n_t^{-1} \sum_{i:W_i=1} X_i$. A major focus of their work is to estimate and do inference on the second term, which is a linear combination of the unknown high-dimensional vector $\beta_c$. A similar challenge lies at the heart of our work, which seeks to estimate an indirect effect $\beta_0$ that under (9) above can be written as $\beta_0 = \gamma_0^T \alpha_0$. This is also a linear combination of an unknown high-dimensional vector, namely $\alpha_0$. Directly using a penalized regression estimate of $\beta_c$ would result in a large bias for $\tau$, which is the same problem that we face with directly using a penalized regression estimate of $\alpha_0$; this was discussed in Section 2.1 in the main text. An additional difficulty that we have, due to our mediation framework, is that the coefficients $\gamma_0$ of our linear combination are unknown and must be estimated as well, in contrast to the known $X_t$ in Athey et al. (2018).

Finally, our solutions to this bias problem are similar in principle. Athey et al. (2018) proposed to estimate $X_t \beta_c$ using $X_t \beta_c + \sum_{i:W_i=0} \gamma_i(Y_i^{\text{obs}} - X_i \beta_c)$, where $\beta_c$ is the lasso estimate of $\beta_c$ and $\gamma_i$ are weights, chosen to minimize the estimation error. The term $X_t \beta_c$ corresponds to the first part $\sum_{SS}^{-1} \sum_{SG} \alpha_0$ of our estimator (8), and the term $\sum_{i:W_i=0} \gamma_i(Y_i^{\text{obs}} - X_i \beta_c)$ is analogous to the second part $\sum_{SS}^{-1} \lambda_n \Omega_{GG} \hat{\kappa}$ of our estimator, where $\lambda_n \hat{\kappa} = n^{-1} G^T(Y - G \hat{\alpha})$ by our Karush-Kuhn-Tucker conditions. In other words, our estimators have the same form and differ in how the weights of the penalized regression residuals are defined.

### C.2 Comparing with Athey et al. (2016)

Similar to our Proposition 1 in the main text, Athey et al. (2016) also found that leveraging surrogate outcomes can increase the efficiency of estimating average treatment effects. Here
we discuss the relationship between our results and Section 7 of Athey et al. (2016).

First, there are certain similarities between the surrogate outcome framework of Athey et al. (2016) and our mediation framework. In their single sample design, Athey et al. (2016) consider a binary treatment \( W \), a long-term outcome \( Y \), and intermediate outcome vector \( S \), and assume that \( S \) is a surrogate for \( Y \) in the sense that \( W \) and \( Y \) are independent conditional on \( S \). In the context of our mediation model (1) in the main text, their treatment \( W \) is analogous to our exposure \( S \), their surrogate \( S \) is analogous to our mediator \( G \), and their outcome \( Y \) is our outcome \( Y \) as well. Their surrogacy assumption is analogous to our assumption of complete mediation in Section 2.3 of the main text, because without a direct effect, \( Y \) and \( S \) are independent conditional in \( G \). On the other hand, Athey et al. (2016) consider a model-free approach while we posit a parametric linear mediation model.

The key to both our procedure and the procedure of Athey et al. (2016) is to assume that the surrogates/mediators are the only pathway through which the treatment/exposure can affect the outcome. When their surrogacy assumption holds, Athey et al. (2016) show that surrogates can be used to pool treated and untreated units to estimate a causal treatment effect. This increases the effective sample size and therefore the estimation efficiency. In contrast, our complete mediation method works by denoising our outcome \( Y \) using our mediators \( G \). We regress our denoised outcome on our exposure \( S \), which is more efficient than the regression of \( Y \) on \( S \) directly.

Both we and Athey et al. (2016) also characterize the theoretical efficiency gain provided by the surrogates/mediators, and our results share many qualitative similarities. Assuming homoskedasticity and no pretreatment variables or other covariates, Athey et al. (2016) show that the decrease in variance afforded by using surrogates can be expressed as

\[
E \left[ \frac{2\sigma^2}{p(1-p)} \cdot \{ p(1-p) - (r(S_i) - p)^2 \} \right],
\]

where \( p \) is the probability of receiving the treatment and \( r(S_i) \) is the conditional probability of having received the treatment given the value of the surrogates. The \( \sigma^2 \) in (11) denotes the variance of the outcome conditional on the surrogates and is thus analogous to our \( \sigma^2 \) from model (1). In our work, Proposition 1 shows that the difference of the asymptotic variances of the ordinary least squares estimator and our estimator can be represented as

\[
\sigma^2 \Sigma^{-1}_{SS} (\Sigma_{SS} - \Sigma_{SG} \Sigma^{-1}_{GG} \Sigma_{GS}) \Sigma^{-1}_{SS}.
\]

First, both (11) and (12) indicate that the efficiency gain is a linear function of the variance of the outcome conditional on the surrogates/mediators. This makes sense because both of our methods gain efficiency essentially by denoising the outcome using the surrogates/mediators. In Section E.3 below we confirmed this claim in simulations by varying \( \sigma^2 \).

Second, both analyses show that the relative efficiency gain is weakest when the treatment/exposure is perfectly correlated with the surrogates/mediators, as in the case the latter offer no additional information. Indeed, when treatment is perfectly correlated with the surrogates, \( r(S_i) = W \) in (12) and the efficiency gain is zero. When our exposure is perfectly correlated with our mediators, \( \Sigma_{SG} \Sigma^{-1}_{GG} \Sigma_{GS} = \Sigma_{SS} \), so (12) equals zero as well. Similarly, the efficiency gain is the greatest when the treatment/exposure is independent of the surrogates/mediators, when \( r(S_i) = p \) in (11) and \( \Sigma_{SG} = 0 \) in (12). Of course, in this case the
true treatment/indirect effect is also equal to zero, so this scenario is not interesting from the testing perspective. These results imply that when testing for a treatment/indirect effect, both methods are most useful when the treatment/exposure is weakly correlated with the surrogates/mediators.

A major difference is that our method assumes a parametric two-stage linear mediation model, whereas Athey et al. (2016) operate under a semi-parametric framework. Though the semi-parametric setting is more flexible, it also necessarily imposes restrictions on the support of the treatment/exposure: Athey et al. (2016) only allow for binary treatments, and extensions to more than two treatments groups or to continuous treatments are difficult. Furthermore, Athey et al. (2016) only proposed estimation methods, whereas we also provide distributional results that allow for valid inference. Nevertheless, relaxing our framework to semiparametric is an important next step and in the future, we hope to explore more in this direction.

C.3 Alternative way to estimate the direct effect under incomplete mediation

Under incomplete mediation, our mediation model is

\[ Y_i = G_i^T \alpha_0 + S_i^T \alpha_1 + \epsilon_{1i}, \quad G_i^T \alpha_0 = S_i^T \beta_0 + \epsilon_{2i}, \]

and we proposed estimators

\[
\begin{pmatrix}
\hat{b} \\
\hat{a}
\end{pmatrix} = 
\begin{pmatrix}
\hat{\Sigma}_{SS}^{-1}\hat{\Sigma}_{SG}\hat{\alpha}_0 \\
\hat{\alpha}_1
\end{pmatrix} + I_2 \otimes \hat{\Sigma}_{SS}^{-1}\lambda_n \hat{\Omega}_{XX}\hat{\kappa},
\]

where we have

\[
n^{1/2} \begin{pmatrix}
\hat{b} - \beta_0 \\
\hat{a} - \alpha_1
\end{pmatrix} \xrightarrow{\text{d}} N(0, V), \quad \text{where} \quad V = \begin{pmatrix}
\sigma^2 \Gamma + \sigma^2 \Sigma_{SS}^{-1} & -\sigma^2 \Gamma \\
-\sigma^2 \Gamma & \sigma^2 (\Gamma + \Sigma_{SS}^{-1})
\end{pmatrix},
\]

and \( \Gamma \equiv \Sigma_{SS}^{-1}\Sigma_{SG}(\Sigma_{GG} - \Sigma_{GS}\Sigma_{SS}^{-1}\Sigma_{SG})^{-1}\Sigma_{GS}\Sigma_{SS}^{-1} \). At the same time, our model implies that

\[ Y = S(\alpha_1 + \beta_0) + \epsilon_1 + \epsilon_2 = S\theta + \epsilon_1 + \epsilon_2, \]

where \( \theta = \alpha_1 + \beta_0 \) is the total effect. Define \( \hat{\theta}_{OLS} \) to be the ordinary least squares estimate of \( \theta \). As pointed out by a referee, an alternative way to estimate the direct effect \( \alpha_1 \) is \( \hat{\alpha}_1 = \hat{\theta}_{OLS} - \hat{b} \). We show here that this alternative is asymptotically equivalent to \( \hat{a} \).

First, it’s easy to see that \( \hat{\alpha}_1 \) is asymptotically normal and unbiased, because \( \text{E}(\hat{\theta}_{OLS} - \hat{b}) = \theta - \beta_0 = \alpha_1 \). To calculate its asymptotic variance, write

\[
\hat{\alpha}_1 - \alpha_1 = (\hat{\theta}_{OLS} - \hat{b}) - (\theta - \beta_0) = (\hat{\theta}_{OLS} - \theta) - (\hat{b} - \beta_0),
\]

so that

\[
n^{1/2}(\hat{\alpha}_1 - \alpha_1) = (1, -1)\{n^{1/2} \begin{pmatrix}
\hat{\theta}_{OLS} - \theta \\
\hat{b} - \beta_0
\end{pmatrix}\} = (1, -1) \begin{pmatrix}
\sigma^2 \Sigma_{SS}^{-1} S \epsilon_1 + \epsilon_2 \\
\sigma^2 \Sigma_{SS}^{-1} S \epsilon_2 + \Delta_2
\end{pmatrix},
\]

24
where $\Delta_2$ is same as in the proof of Theorem 2, where it is shown that $\Delta_2 = O_p(1)$.

Let

$$W = n^{-1/2} \sum_{i=1}^n \left( (\Sigma_{iSS}^{-1} S_i^\top \epsilon_{1i} + \Sigma_{iSS}^{-1} \sigma_i^\top \epsilon_{2i}) \right).$$

For each $i$ let

$$W_i = W_{1i} + W_{2i} = \left( (\Sigma_{iSS}^{-1} S_i^\top \epsilon_{1i}) \right) + \left( (\Sigma_{iSS}^{-1} \sigma_i^\top \epsilon_{2i}) \right).$$

Then $E(W_{1i}) = E(W_{2i}) = 0$, and

$$\text{var}(W_{1i}) = E(W_{1i}W_{1i}^\top) = E \left( \begin{pmatrix} \Sigma_{iSS}^{-1} \sigma_i^\top \epsilon_{1i} \epsilon_{1i}^\top S_i \Sigma_{iSS}^{-1} \\ (\Sigma_{iSS}^{-1} 0) \Omega_{XX} X_i^\top \epsilon_{1i} \epsilon_{1i}^\top S_i \Sigma_{iSS}^{-1} \\ (\Sigma_{iSS}^{-1} 0) \Omega_{XX} X_i^\top \epsilon_{1i} \epsilon_{1i}^\top \Omega_{XX}^\top \left( \Sigma_{iSS}^{-1} \right) \end{pmatrix} \right)$$

$$= \sigma_i^2 \Sigma_{iSS}^{-1} \begin{pmatrix} \sigma_i^2 \Sigma_{iSS}^{-1} \text{SSX} \Omega_{XX}^\top \left( \Sigma_{iSS}^{-1} \right) \\ \sigma_i^2 (\Sigma_{iSS}^{-1} 0) \Omega_{XX} \Sigma_{iSS} \Sigma_{iSS}^{-1} \sigma_i^2 (\Sigma_{iSS}^{-1} 0) \Omega_{XX} \Sigma_{iSS} \Omega_{XX}^\top \left( \Sigma_{iSS}^{-1} \right) \end{pmatrix}.$$
and
\[ \Sigma_{XX}^{-1} = \begin{pmatrix} J^{-1} & -\Sigma_{SS}^{-1} \Sigma_{SG} J^{-1} \\ \Sigma_{SG} & \Sigma_{SS} \end{pmatrix}, \]

where \( J = \Sigma_{GG} - \Sigma_{GS}^{-1} \Sigma_{SG} \), we have

\[ \Omega_{XX}^{\top} \left( \Sigma_{SS}^{-1} \right) = \begin{pmatrix} J^{-1} & -\Sigma_{SS}^{-1} \Sigma_{SG} J^{-1} \\ \Sigma_{SS} & \Sigma_{SS} \end{pmatrix} \begin{pmatrix} \Sigma_{GS} & \Sigma_{SS} \\ 0 & \Sigma_{SS} \end{pmatrix} = 0. \]

Similarly, it can be shown that

\[ \Sigma_{SX} \Omega_{XX}^{\top} \left( \Sigma_{SS}^{-1} \right) = \begin{pmatrix} \Sigma_{SG} & \Sigma_{SS} \\ \Sigma_{SS} & \Sigma_{SS} \end{pmatrix} \begin{pmatrix} J^{-1} & -\Sigma_{SS}^{-1} \Sigma_{SG} J^{-1} \\ \Sigma_{SS} & \Sigma_{SS} \end{pmatrix} = 0. \]

Therefore \( \tilde{\alpha}_1 \) and \( \hat{a} \) are asymptotically equivalent.

**D Optimal variance**

We show here that under certain conditions, the asymptotic variance of our proposed estimator for the indirect effect achieves the Cramér-Rao lower bound.

**Theorem 5** Suppose model (9) in the Appendix holds, and assume that \( \epsilon_1 \) and \( E \) are both normally distributed. Then among all asymptotically unbiased estimators of \( \beta_0 = \gamma_0^\top \alpha_0 \), our proposed estimator \( \hat{b} \) from Section 2.3 in the main text achieves the minimum variance bound.

**Proof.** Let \( \theta = (\gamma_0^\top, \alpha_0^\top)^\top \), so that \( \theta \) is a \( 2p \times 1 \) vector containing all parameters in model (9). Then \( \gamma_0 = V \theta \) where

\[
V = \begin{pmatrix}
1 & 0 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & & \ddots & \ddots & \vdots & & \vdots \\
0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0
\end{pmatrix} = (I_p \ O_p),
\]
where $I_p$ is a $p \times p$ identity matrix and $O_p$ is a $p \times p$ matrix of zeros. Similarly, $\alpha_0 = U \theta$, where $U = (O_p I_p)$. We can represent $\beta_0$ as a function of $\theta$, denoted as $g(\theta) = \theta^T V^T U \theta = \beta_0$.

Based on model (9) in the Appendix, the joint density function of a sample of $n$ observations is

$$f(Y, G, S) \propto f(Y \mid G, S)f(G \mid S) \propto \exp \left\{ -\frac{1}{2\sigma_1^2} \sum_{i=1}^{n} (Y_i - G_i \alpha_0)^2 \right\} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} (G_i - S_i \gamma_0^T) \Sigma_E^{-1} (G_i - S_i \gamma_0^T)^T \right\}.$$ 

Therefore the log-likelihood function equals

$$l(\theta; Y, G, S) \propto -\frac{1}{2\sigma_1^2} \sum_{i=1}^{n} G_i U \theta^T U^T G_i^T - \frac{1}{2} \sum_{i=1}^{n} S_i \theta^T V^T \Sigma_E^{-1} V \theta S_i + h(\theta)$$

where $h(\theta)$ only involves linear functions of $\theta$.

Then we have

$$-\frac{\partial^2 l(\theta; Y, G, S)}{\partial \theta^2} = -\sum_{i=1}^{n} \left( \frac{1}{\sigma_1^2} U^T G_i^T G_i U + S_i^2 V^T \Sigma_E^{-1} V \right),$$

so the expected Fisher information matrix $I_n(\theta)$ of $\theta$ is

$$-E\frac{\partial^2 l(\theta; Y, G, S)}{\partial \theta^2} = \frac{n}{\sigma_1^2} U^T E(G_i^T G_i) U + nE(S_i^T S_i) V^T \Sigma_E^{-1} V = n \begin{pmatrix} \Sigma_{SS} \Sigma_E^{-1} & 0 \\ 0 & \sigma_1^{-2} \Sigma_{GG} \end{pmatrix}$$

By the multivariate Cramér-Rao inequality, for any unbiased estimator $\delta$ of $g(\theta)$,

$$\text{cov}(\delta) \geq g'(\theta) I_n(\theta)^{-1} g'(\theta) = \frac{1}{n} \theta^T (V^T U + U^T V) \begin{pmatrix} \Sigma_{SS} \Sigma_E^{-1} & 0 \\ 0 & \sigma_1^2 \Sigma_{GG} \end{pmatrix} (V^T U + U^T V) \theta$$

$$= \frac{1}{n} \begin{pmatrix} \alpha_0^T & \gamma_0^T \end{pmatrix} \Sigma_{SS}^{-1} \Sigma_E^{-1} \begin{pmatrix} \alpha_0 \\ \gamma_0 \end{pmatrix}$$

$$= \frac{1}{n} \begin{pmatrix} \Sigma_{SS}^{-1} \alpha_0^T \Sigma_E \alpha_0 + \sigma_1^2 \gamma_0^T \Sigma_{GG} \gamma_0^T \end{pmatrix}$$

The residual error $\epsilon_{2i}$ from model (1) in the main text equals $E_i \alpha_0$ in the notation of model (9) from the Appendix, so $\sigma_2^2 = \alpha_0^T \Sigma_E \alpha_0$ and $\Sigma_{SG} = \text{cov}(S_i, G_i) = \Sigma_{SS} \gamma_0^T$. Therefore the minimal variance among all asymptotically unbiased estimators of $\beta_0$ is $n^{-1} (\sigma_2^2 \Sigma_{SS}^{-1} + \sigma_1^2 \Sigma_{SS}^{-1} \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS} \Sigma_{SS}^{-1})$.

$\square$

### E Additional simulation results

#### E.1 Uniform tuning

As mentioned in Section 4.1 of the main text, we also considered setting the parameter $\tau_n$ in the sparse matrix estimation step to always equal $\{(\log p)/n\}^{1/2}/3$, for computational convenience.
Figure 3: Average power curves for testing the indirect effect under complete mediation, at significance level $\alpha = 0.05$, over 200 replications, with uniform tuning parameter $\tau_n = \{(\log p)/n\}^{1/2}/3$ when estimating $\Omega_{GG}$. Examples 1 and 2: 500 potential mediators with 1 or 5 true mediators, respectively; Examples 3 and 4: 1,000 potential mediators with 1 or 5 true mediators, respectively. Proposed: $\tilde{b}$ from (8); OLS: ordinary least squares estimate.
Figure 3 presents the results of this uniform tuning procedure for complete mediation, under the same simulation settings as Section 4 of the main text. Our procedure still always maintains type I error, and importantly is still able to achieve higher power than the ordinary least squares estimator for sufficiently large $\beta_0$. Using uniform tuning under incomplete mediation also maintained type I error in all settings and we omit the presentation of the result.

### E.2 Risk

In the paper we mainly focused on the power of our method compared to ordinary least squares, under the complete mediation setting. Here, we also compare the risks of these two methods under squared error loss. The simulation setting we are using here are the same as those in Section 4.3 in the main text. Figure 4 compares both average power and risk curves of our $\tilde{b}$ and ordinary least squares, in our four simulation examples, over 200 replications. For most values of $\beta_0$ in most of the settings, our method has a smaller risk than ordinary least squares. The risk of our method increased with the number of true mediators, and exceeded the risk of ordinary least squares for certain $\beta_0$, but the difference was not large. The risk behavior in Example 4 of Figure 4 is interesting and may be worth future study.

### E.3 Effect of noise

Because we have a two-stage linear mediation model, we have two sources of noise: $\epsilon_1$ and $\epsilon_2$, from (1) in the main text. In our simulations we generated data following $G_i = c\gamma S_i + E_i$, where $E_i \sim N(0, \Sigma_E)$, so $\text{var}(\epsilon_2) = \sigma^2_2 = \alpha_0^\top \Sigma_E \alpha_0$. In this section we study the effect of varying $\Sigma_E$ and $\text{var}(\epsilon_1) = \sigma^2_1$ on the power of our proposed approach in the complete mediation setting. We used the same simulation settings as in Section 4.3.

Figure 5 reports power curves for different values of $\sigma_1$ when $\Sigma_E$ was fixed with diagonal entries equal to 1. As $\sigma_1$ increased, our method consistently outperformed ordinary least squares, even for very large $\sigma_1$. This is sensible because our method denoises the outcome using the observed mediators, and the advantage of doing so becomes more apparent the larger the noise. From another point of view, Proposition 1 in the main text states that the difference of asymptotic variance between ordinary least squares and our method converges to a positive semi-definite matrix $\sigma_1^2 \Sigma^{-1}_S (\Sigma^{-1}_S - \Sigma^{-1}_G \Sigma^{-1}_G \Sigma^{-1}_S) \Sigma^{-1}_S$, which increases in norm as $\sigma_1$ increases. Both of methods became less powerful as $\sigma_1$ increased, but the relative power advantage of our method compared to ordinary least squares remains large.

Figure 6 shows that when $\sigma_1$ was fixed and $\Sigma_E$ was varied, our method had the greatest advantage relative to ordinary least squares when $\Sigma_E$ was moderate. When $\Sigma_E$ was too large or too small, our procedure behaved more like ordinary least squares. This is consistent with Proposition 1 in the main text and makes intuitive sense as well. Our method should have the greatest relative advantage when $S$ and $G$ are correlated but only weakly so, or else $G$ would offer no additional information. This occurs when $\Sigma_E$ is not too large. When $\Sigma_E = 0$, the indirect effect becomes zero as well, so there is no effect to detect.
Figure 4: Average power and risk curves for testing the indirect effect under complete mediation at significance level $\alpha = 0.05$, over 200 replications. Examples 1 and 2: 500 potential mediators with 1 and 5 true mediators, respectively; Examples 3 and 4: 1,000 potential mediators with one and five true mediators, respectively. Proposed: $\tilde{b}$ from (8); OLS: ordinary least squares estimate.
Figure 5: Average power plots for testing the indirect effect under complete mediation at significance level $\alpha = 0.05$, over 200 replications, for different values of $\epsilon_{1i}$. All examples had 500 potential mediators with one true mediator and the diagonal entries of $\Sigma_E$ were fixed at 1. Proposed: $\tilde{b}$ from (8); OLS: ordinary least squares estimate.
Figure 6: Average power plots for testing the indirect effect under complete mediation at significance level $\alpha = 0.05$, over 200 replications, for different values of $\Sigma_E$. All examples had 500 potential mediators with one true mediator and $\sigma_1 = 5$. Proposed: $\hat{b}$ from (8); OLS: ordinary least squares estimate.
Figure 7: Average power plots for testing the indirect effect under complete mediation at significance level $\alpha = 0.05$, over 200 replications for $t_3$-distributed $\epsilon_{1i}$. All examples have 500 potential mediators with one true mediator, with $\sigma_1$ generated from $t_3 \cdot (25/3)^{1/2}$. The left and right panel denote implementing our method by using the Akaike information criterion-type measure or uniform parameter $\tau_n = \{(\log p)/n\}^{1/2}/3$ in estimating $\Omega_{GG}$. AIC: Akaike Information Criterion; Proposed: $\tilde{b}$ from (8); OLS: ordinary least squares estimate.

E.4 Non-Gaussian errors

In our simulations in Section 4 we generated $\epsilon_{1i}$ and $E_i$ from Gaussian distributions for simplicity, but as indicated in Theorem 2 and 4, we do not need the Gaussian assumptions. Here we report simulation results when the errors were $t$-distributed. Specifically, we adopt the same simulation settings from Example 1 in Section 4.3, but instead of generating $\epsilon_{1i} \sim N(0, 25)$, we now generate $\epsilon_{1i}$ from $t_3 \cdot (25/3)^{1/2}$, where $t_3$ denotes a $t$-distributed random variable with three degrees of freedom. The normalizing constant $(25/3)^{1/2}$ is such that $\text{var}(\epsilon_{1i})$ still equals 25.

Figure 7 reports the average power curves for our method and for ordinary least squares. With non-Gaussian noise, our proposed method was still able to maintain the nominal significance level, and had higher power than ordinary least squares for sufficiently large $\beta_0$ with either Akaike Information Criterion or uniform tuning, which is consistent with Proposition 1.

E.5 Robustness to matrix estimation accuracy

Currently, our proposed method uses a constrained $\ell_1$-minimization approach to estimate the matrices $\Omega_{XX}$ and $\Omega_{GG}$, defined in Theorems 1 and 3 in the main text. Under sparsity assumptions on various matrices, described in Assumptions 2 and 3, these estimates are consistent. On the other hand, Javanmard and Montanari (2014) showed that in the standard high-dimensional linear regression setting, asymptotically optimal inference is possible.
without consistent estimation or sparsity of the precision matrix of the covariates. In this section we explore whether consistent estimation is indeed necessary for our method. We considered the same complete mediation simulation settings as in Example 1 of Section 4.3 in the main text.

We first varied the accuracy with which $\Omega_{GG}$ was estimated by fixing the value of the tuning parameter $\tau_n$ of the estimation procedure (6) to 20 different values between 0.05 and 0.7, instead of choosing an optimal value. We measured the resulting estimation error using $\|\hat{\Omega}_{GG} - \Omega_{GG}\|_\infty$, and we evaluated the power of our procedure when $\beta_0$ equaled 1, 0.5 or 0. Figure 8 shows that for $\beta_0 \neq 0$, both very accurate and very inaccurate estimation resulted in diminished power. An estimation error of $\|\hat{\Omega}_{GG} - \Omega_{GG}\|_\infty \approx 0.11$ seemed to give the highest power.

We next examined whether the sparsity conditions in Assumptions 2 and 3 were necessary. We used the same simulations as above, but changed $\Sigma_E$ such that $\Sigma_E^{-1}$ was no longer sparse. Specifically, $\Sigma_E^{-1}$ was a compound symmetric matrix with diagonal entries equal to 1 and off-diagonal entries equal to 0.8. Figure 9 shows that our method could maintain type I error and have high power, for example when $\beta_0 = 1$, even when the sparsity assumptions were
Figure 9: Average power of our proposed method versus average estimation accuracy of $\Omega_{GG}$ for dense $\Sigma^{-1}_E$, with 500 potential mediators and one true mediator, at significance level $\alpha = 0.05$ over 200 replications.
Together, these results suggest that our procedure is indeed robust to sparsity and consistent estimation of $\Omega_{XX}$ and $\Omega_{GG}$. On the other hand, these properties are necessary for our current approach to establishing the asymptotic distributions of our estimators. Our general strategy is to apply the central limit theorem to the terms on the right-hand sides of Karush-Kuhn-Tucker conditions (3) and (7) in the main text, which involves $\hat{\Omega}_{XX}$ and $\hat{\Omega}_{GG}$. The behaviors of these random variables are difficult to characterize in general. We could consider conditioning on $G$ and $S$ so that $\hat{\Omega}_{XX}$ and $\hat{\Omega}_{GG}$ essentially become constants, but under our mediation model (1) this fixes the values of the residual errors $\epsilon_{2i}$, which leads to difficulties. We cannot condition on $S$ alone, because doing so induces dependencies between the terms in (3), as well as those in (7), and makes applying the central limit theorem difficult. We therefore marginalize over $G$ and $S$, but since $\hat{\Omega}_{XX}$ and $\hat{\Omega}_{GG}$ are functions of $G$ and $S$, we need to establish that they are close to well-defined population-level quantities so that we can apply an unconditional central limit theorem. The simplest way to do this is to show that $\hat{\Omega}_{XX}$ and $\hat{\Omega}_{GG}$ converge to $\Omega_{XX}$ and $\Omega_{GG}$, which requires sparsity assumptions. However, there may be other ways to show that $\hat{\Omega}_{XX}$ and $\hat{\Omega}_{GG}$ converge to well-defined population-level quantities. This would allow us to consider alternative estimators that are not based on constrained $\ell_1$-minimization, which may give improved finite-sample performance. For example, we can consider an extension of the approach of Javanmard and Montanari (2014) to our mediation analysis setting, as described in the main text in Section 2.2. In the future we hope to improve our proof strategies so that our theoretical results reflect the robustness demonstrated by our procedure in simulations.

**F Proofs of theorems**

**F.1 Proof of Theorem 1 and 3**

We first prove the result for the $\hat{\Omega}_{GG}$ that solves (6). First we show that $|\Omega_{GG}\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG}|_\infty \leq (N_G + 1)\{((\log p)/n)^{1/2}$ with probability going to 1. First we have $|\Omega_{GG}\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG}|_\infty \leq |\Omega_{GG}(\hat{\Sigma}_{GG} - \Sigma_{GG})|_\infty + |\Sigma_{SG} - \hat{\Sigma}_{SG}|_\infty$. Following the proofs of Theorems 1 and 4 from Cai et al. (2011), we have $|\hat{\Sigma}_{XX} - \Sigma_{XX}|_\infty \leq C_1\{(\log(p + q))/n\}^{1/2}$ with probability greater than $p_1 = 1 - 4p^{-\tau}$ under Assumption 1, where $C_1 = 2\eta^{-2}(2 + \tau + \eta^{-1}e^2K^2)^2$ and $\tau > 0$. Since $|\hat{\Sigma}_{GG} - \Sigma_{GG}|_\infty$ and $|\hat{\Sigma}_{SG} - \Sigma_{SG}|_\infty$ are both less than or equal to $|\hat{\Sigma}_{XX} - \Sigma_{XX}|_\infty$, and by Assumption 3 $\|\Omega_{GG}^T\|_{L_1} \leq N_G$, we have $|\Omega_{GG}\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG}|_\infty \leq \|\Omega_{GG}^T\|_{L_1}|\hat{\Sigma}_{GG} - \Sigma_{GG}|_\infty + |\Sigma_{SG} - \hat{\Sigma}_{SG}|_\infty \leq (N_G + 1)C_1\{(\log(p + q))/n\}^{1/2}$ with probability greater than $p_1$. Next, for $1 \leq i \leq q$, let $\hat{b}_i$ be the solution of the optimization problem: $\min |b|_1$ subject to $|b^T\hat{\Sigma}_{GG} - \hat{\Sigma}_{SG,i}|_\infty \leq \tau_n.$
where \( b \) is a vector in \( R^p \) and \( \Sigma_{SG,i} \) denote the \( i \)-th row of \( \Sigma_{SG} \). Let \( \hat{B} = (b_1^T, \ldots, b_q^T) \) and write \( \Omega = (\omega_1^T, \ldots, \omega_q^T) \). Then the constraint \( |\Omega \hat{\Sigma}_{SG} - \hat{\Sigma}_{SG}|_{\infty} \leq \tau_n \) from (6) is equivalent to
\[
|\omega_i^T \hat{\Sigma}_{GG} - \hat{\Sigma}_{SG,i}|_{\infty} \leq \tau_n \quad \text{for all} \quad 1 \leq i \leq q.
\]
Thus we have \( |\omega_i|_1 \geq |\beta_i|_1 \) for all \( 1 \leq i \leq q \). Since \( |B^T \hat{\Sigma}_{GG} - \hat{\Sigma}_{SG}|_{\infty} \leq \tau_n \), we have \( |\Omega|_1 \leq |\hat{B}|_1 \), thus \( \Omega = \hat{B} \). By the definition of \( b_i^T \), we can see that \( |b_i^T|_1 \leq \|\hat{\Omega}_{GG}\|_{L_1} \) for \( 1 \leq i \leq q \). Then we have \( \|\hat{\Omega}_{GG}^T\|_{L_1} \leq \|\Omega_{GG}^T\|_{L_1} \). Finally we have
\[
|\hat{\Omega}_{GG} - \Omega_{GG}|_{\infty} = |\hat{\Omega}_{GG}(\Sigma_{GG} - \hat{\Sigma}_{GG}) + \hat{\Omega}_{GG} \hat{\Sigma}_{GG} - \hat{\Sigma}_{SG} + \hat{\Sigma}_{SG} - \Sigma_{SG}|_{\infty}
\]
under the same event with probability greater than \( p_1 \). This gives us
\[
|\hat{\Omega}_{GG} - \Omega_{GG})\Sigma_{SG}|_{\infty} \leq 2N_G C_1 \{ \log (p + q)/n \}^{1/2}
\]
with probability greater than \( p_1 \).

The proof for \( \hat{\Omega}_{XX} \) can then be obtained by complete analogy. \( \square \)

F.2 Proof of Theorem 4

For simplicity let \( \Omega_0 = \Omega_{GG} \) and \( \hat{\Omega} = \hat{\Omega}_{GG} \). Based on Section 2.3 and Theorem 4, we have
\[
n^{1/2}(\hat{b} - \beta_0) = W + \Delta_0 + \Delta_1 + \Delta_2,
\]
where
\[
W = n^{-1/2} \Sigma_{SS}^{-1} \Sigma_{GG}^{-1} \Omega_0 G^T \epsilon_1 = n^{-1/2} \Sigma_{SS}^{-1} \Sigma_{GG}^{-1} \sum_{i=1}^n (S_i^T \epsilon_{2i} + \Omega_0 G_i^T \epsilon_{1i}),
\]
\[
\Delta_0 = n^{1/2} \Sigma_{SS}^{-1} \hat{\Omega}_{GG} \Sigma_{GG} - \hat{\Sigma}_{SG}) \hat{\alpha} - \alpha_0),
\]
\[
\Delta_1 = n^{-1/2} (\hat{\Sigma}_{SS}^{-1} \hat{\Omega}_{GG} - \Sigma_{SS}^{-1} \Omega_{GG}) G^T \epsilon_1,
\]
\[
\Delta_2 = n^{-1/2} (\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1}) S^T \epsilon_2.
\]
First we prove that \( W \rightarrow N(0, \sigma_1^2 \Sigma_{SS}^{-1} \Sigma_{GG}^{-1} \Sigma_{SG}^{-1} \Sigma_{SS}^{-1} + \sigma_2^2 \Sigma_{SS}^{-1}) \): For \( i = 1, \ldots, n \), let
\[
W_i = S_i^T \epsilon_{2i} + \Omega_0 G_i^T \epsilon_{1i}.
\]
Then we have
\[
E(W_i) = E\{E(S_i^T \epsilon_{2i} | S_i)\} + E\{E(\Omega_0 G_i^T \epsilon_{1i} | G_i)\} = 0,
\]
and
\[
\var(S_i^T \epsilon_{2i}) = \sigma_2^2 E(S_i^T S_i),
\]
\[
\var(\Omega_0 G_i^T \epsilon_{1i}) = \sigma_1^2 \Omega_0 E(G_i^T G_i) \Omega_0^T
\]
37
\[ \text{cov}(S_i^T \epsilon_{2i}, \Omega_0 G_i^T \epsilon_{1i}) = E\{E(S_i^T \epsilon_{2i} \epsilon_{1i}^T G_i \Omega_0^T \mid S_i, G_i)\} = 0, \]

which means that
\[ \text{var}(W_i) = \sigma_i^2 \Sigma_{SS} + \sigma_i^2 \Omega_0 \Sigma_{GG} \Omega_0^T. \]

Since \( W = n^{-1/2} \Sigma_{SS}^{-1} \sum_{i=1}^n W_i \), by CLT, when \( n \to \infty \),
\[ W \to N_q(0, \sigma_i^2 \Sigma_{SS}^{-1} \Sigma_{GG} \Sigma_{SS}^{-1} + \sigma_i^2 \Sigma_{SS}^{-1}). \]

Next we prove that \( |\Delta_0|_\infty = o_P(1), |\Delta_1|_\infty = o_P(1) \) and \( |\Delta_2|_\infty = o_P(1) \).

We start by showing
\[ |\Delta_1|_\infty = |n^{-1/2}(\hat{\Sigma}_{SS}^{-1} \hat{\Omega} - \Sigma_{SS}^{-1} \Omega_0)G^T \epsilon_1|_\infty = o_P(1). \]

Since \( |n^{-1/2}(\hat{\Sigma}_{SS}^{-1} \hat{\Omega} - \Sigma_{SS}^{-1} \Omega_0)G^T \epsilon_1|_\infty \)
\[ \leq |n^{-1/2}(\hat{\Sigma}_{SS}^{-1} \hat{\Omega} - \hat{\Sigma}_{SS}^{-1} \Omega_0)G^T \epsilon_1|_\infty + |n^{-1/2}(\hat{\Sigma}_{SS}^{-1} \Omega_0 - \Sigma_{SS}^{-1} \Omega_0)G^T \epsilon_1|_\infty \]
\[ \leq \|\hat{\Sigma}_{SS}^{-1}(\hat{\Omega}^T - \Omega_0^T)\|_{L_1} |n^{-1/2}G^T \epsilon_1|_\infty + \|\Omega_0^T(\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1})\|_{L_1} |n^{-1/2}G^T \epsilon_1|_\infty. \]

Let \( M_s = |\hat{\Sigma}_{SS}|_\infty \), then
\[ \|\hat{\Sigma}_{SS}^{-1}(\hat{\Omega}^T - \Omega_0^T)\|_{L_1} \leq q |\hat{\Sigma}_{SS}|_\infty \|\hat{\Omega}^T - \Omega_0^T\|_{L_1} \leq qM_s C_s 0 (\log p/n)^{1/2} \]

and
\[ \|\Omega_0^T(\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1})\|_{L_1} \leq q |\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1}|_\infty \|\Omega_0^T\|_{L_1} \leq qM (\log q/n)^{1/2}. \]

We have \( |G^T \epsilon_1|_\infty = \max_{j \in \{1, \ldots, p\}} \{\sum_{i=1}^n G_{ij} \epsilon_{1i}\} \), and \( G_{ij} \epsilon_{1i}, i = 1, \ldots, n \) are i.i.d random variables given each \( j \). Since each \( G_{ij} \) and \( \epsilon_{1i} \) are mean zero sub-exponential random variables for each \( i \) and \( j \), for all \( i \in \{1, \ldots, n\} \) and \( j \in \{1, \ldots, p\} \), we have
\[ E\{\exp(\frac{G_{ij} \epsilon_{1i}}{C})\} \leq E\{\exp(\frac{G_{ij}^2 + \epsilon_{1i}^2}{2C})\} = E\{\exp(\frac{G_{ij}^2}{2C'})\}E\{\exp(\frac{\epsilon_{1i}^2}{2C})\} \leq C' \]
for some constant \( C \) and \( C' \). So \( G_{ij} \epsilon_{1i} \) are also sub-exponential random variables for \( i \in \{1, \ldots, n\} \) and \( j \in \{1, \ldots, p\} \), also \( E(G_{ij} \epsilon_{1i}) = E(G_{ij})E(\epsilon_{1i}) = 0 \) since \( G_{ij} \perp \epsilon_{1i} \). Using Bernstein inequality,
\[ P\{|G^T \epsilon_1|_\infty \geq c(n \log p)^{1/2}\} = P\{\max_{j \in \{1, \ldots, p\}} (\sum_{i=1}^n G_{ij} \epsilon_{1i}) \geq c(n \log p)^{1/2}\} \]
\[ \leq p P\{\sum_{i=1}^n G_{ij} \epsilon_{1i} \geq c(n \log p)^{1/2}\} \]
\[ \leq p \exp\{-\frac{c' n \log p}{n + c(n \log p)^{1/2}}\} = \frac{1}{p^{c'-1}} \to 0 \]
when \( p > n \to \infty \) and \( c' > 1 \). This gives us \( P\{|n^{-1/2}G^T \epsilon_1|_\infty \geq c(\log p)^{1/2}\} \to 0 \), then we have \( P\{|n^{-1/2}(\hat{\Sigma}_{SS}^{-1} \hat{\Omega} - \Sigma_{SS}^{-1} \Omega_0)G^T \epsilon_1|_\infty \geq c_2(\log p)/n^{1/2}\} \to 0 \). And since \( (\log p)/n^{1/2} \to 0 \),
\[ P\{|n^{-1/2}(\hat{\Sigma}_{SS}^{-1} \hat{\Omega} - \Sigma_{SS}^{-1} \Omega_0)G^T \epsilon_1|_\infty = 0\} \to 1. \]
Next we show that
\[ |\Delta_2|_\infty = |n^{-1/2}(\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1})S^\top \epsilon_2|_\infty = o_P(1). \]
Similarly we have \( P\{|n^{-1/2}S^\top \epsilon_2|_\infty \geq c(\log q)^{1/2}\} \to 0. \) Combining \( |\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1}|_\infty \leq c/n^{1/2}, \)
we have when \( n \to \infty, \) \( P\{|n^{-1/2}(\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1})S^\top \epsilon_2|_\infty = 0\} \to 1. \)
Finally we show that
\[ |\Delta_0|_\infty = |n^{1/2}\Sigma_{SS}^{-1}(\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG})(\hat{\alpha} - \alpha_0)|_\infty = o_P(1). \]
Since
\[ |n^{1/2}\Sigma_{SS}^{-1}(\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG})(\hat{\alpha} - \alpha_0)|_\infty \leq |n^{1/2}\Sigma_{SS}^{-1}(\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG})|_\infty |\hat{\alpha} - \alpha_0|_1 \leq n^{1/2}|\Sigma_{SS}^{-1}L_1| |\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG}|_\infty |\hat{\alpha} - \alpha_0|_1 \]
From properties of scaled lasso we have \( |\hat{\alpha} - \alpha_0|_1 = O\{\log p/n^{1/2}\}, \)
and by definition \( \hat{\Omega} \)
solves (6) in the main text in, we have \( |\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG}|_\infty \leq \tau_n = O\{\log p/n^{1/2}\}. \) Then we have \( |\Delta_0|_\infty = O(\log p/n^{1/2}) = o_P(1). \)

**F.3 Proof of Theorem 2**

Since
\[ n^{1/2} \begin{pmatrix} \hat{b} - \beta_0 \\ \hat{\alpha} - \alpha_1 \end{pmatrix} = W + \Delta_0 + \Delta_1 + \begin{pmatrix} \Delta_2 \\ 0 \end{pmatrix}, \]
where
\[ W = n^{-1/2}I_2 \otimes \Sigma_{SS}^{-1} \Omega_{XX} X^\top \epsilon_1 + \begin{pmatrix} n^{-1/2}\Sigma_{SS}^{-1} S^\top \epsilon_2 \\ 0 \end{pmatrix}, \]
and
\[ \Delta_0 = n^{1/2}I_2 \otimes \Sigma_{SS}^{-1} (\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG})(\hat{\alpha} - \alpha), \]
\[ \Delta_1 = n^{-1/2}I_2 \otimes (\Sigma_{SS}^{-1}\hat{\Omega}_{XX} - \Sigma_{SS}^{-1}\Omega_{XX})X^\top \epsilon_1, \]
\[ \Delta_2 = n^{-1/2}(\hat{\Sigma}_{SS}^{-1} - \Sigma_{SS}^{-1})S^\top \epsilon_2. \]

\[ n^{1/2} \begin{pmatrix} \hat{b} - \beta_0 \\ \hat{\alpha} - \alpha_1 \end{pmatrix} = I_2 \otimes \Sigma_{SS}^{-1} \Omega_{XX} X^\top \epsilon_1 n^{1/2} + \Delta_1 + \begin{pmatrix} \Sigma_{SS}^{-1} S^\top \epsilon_2 n^{1/2} + \Delta_2 \\ 0 \end{pmatrix} + \Delta \]

Showing that \( |\Delta_2|_\infty = o_P(1) \) is exactly the same as in the proof of Theorem 4. By fitting a new \( \ell_1 \) optimization problem as in (3) instead of (4), \( |\Delta_0|_\infty = o_P(1) \) can be directly extended from \( |n^{1/2}\Sigma_{SS}^{-1}(\hat{\Sigma}_{SG} - \hat{\Omega}\hat{\Sigma}_{GG})(\hat{\alpha} - \alpha_0)|_\infty = o_P(1) \) and \( |\Delta_1|_\infty = o_P(1) \) can be extended from \( |n^{-1/2}(\hat{\Sigma}_{SS}^{-1}\hat{\Omega} - \Sigma_{SS}^{-1}\Omega_0)G^\top \epsilon_1|_\infty = o_P(1) \), which has already been proved in Theorem 4. So it remains to prove the asymptotic normality of \( W \):
\[ W = n^{-1/2}I_2 \otimes \Sigma_{SS}^{-1} \sum_{i=1}^n \{ \Omega_{XX} X_i^\top \epsilon_{1i} + \begin{pmatrix} S_i^\top \epsilon_{2i} \\ 0 \end{pmatrix} \}. \]
For each $i$ let

\[ W_i = W_{1i} + W_{2i} = \Omega_{XX}X_i^\top \epsilon_{1i} + \left( S_{i}^\top \epsilon_{2i} \right), \]

as before we have $E(W_{1i}) = E(W_{2i}) = 0$, and

\[ \text{var}(W_{1i}) = E(W_{1i}W_{1i}^\top) = \sigma_i^2 \Omega_{XX} E(X_i^\top X_i) \Omega_{XX}^\top = \sigma_i^2 \Omega_{XX} \Sigma_{XX} \Omega_{XX}^\top = \sigma_i^2 D \Sigma_{XX}^{-1} D^\top, \]

where $D = \begin{pmatrix} \Sigma_{SG} & 0 \\ \Sigma_{SS} & \Sigma_{SS} \end{pmatrix}$. By inversion of block matrix we have

\[ \Sigma_{XX}^{-1} = \begin{pmatrix} \Sigma_{GG} & \Sigma_{GS} \\ \Sigma_{SG} & \Sigma_{SS} \end{pmatrix}^{-1} = \begin{pmatrix} J^{-1} & -J^{-1} \Sigma_{SS} \Sigma_{GG}^{-1} \\ -J^{-1} \Sigma_{SS} J^{-1} \Sigma_{GG}^{-1} + \Sigma_{SS} \Sigma_{GS} J^{-1} \Sigma_{GG}^{-1} & \Sigma_{SS}^{-1} \end{pmatrix}, \]

where $J = \Sigma_{GG} - \Sigma_{GS} \Sigma_{SS}^{-1} \Sigma_{SG}$. Thus

\[ \text{var}(W_{1i}) = \sigma_i^2 \begin{pmatrix} \Sigma_{SG} J^{-1} \Sigma_{GS} & -\Sigma_{SG} J^{-1} \Sigma_{GS} \\ -\Sigma_{SG} J^{-1} \Sigma_{GS} + \Sigma_{SS} + \Sigma_{SG} J^{-1} \Sigma_{GS} \end{pmatrix}, \]

\[ \text{var}(W_{2i}) = E(W_{2i}W_{2i}^\top) = \sigma_2^2 \begin{pmatrix} \Sigma_{SS} & 0 \\ 0 & 0 \end{pmatrix}. \]

and

\[ \text{cov}(W_{1i}, W_{2i}) = E\{\Omega_{XX}X_i^\top \epsilon_{1i}(\epsilon_{2i}S_i, 0)\} = E\{E(\Omega_{XX}X_i^\top \epsilon_{1i}\epsilon_{2i}S_i, 0) | S_i, X_i\} = 0. \]

So

\[ \text{var}(W_i) = \text{var}(W_{1i}) + \text{var}(W_{2i}) = \sigma_i^2 \begin{pmatrix} \Sigma_{SG} J^{-1} \Sigma_{GS} & -\Sigma_{SG} J^{-1} \Sigma_{GS} \\ -\Sigma_{SG} J^{-1} \Sigma_{GS} + \Sigma_{SS} + \Sigma_{SG} J^{-1} \Sigma_{GS} \end{pmatrix} + \sigma_2^2 \begin{pmatrix} \Sigma_{SS}^{-1} & 0 \\ 0 & 0 \end{pmatrix}. \]

Since $W = n^{-1/2} I_2 \otimes \Sigma_{SS}^{-1} \sum_{i=1}^n W_i$, by CLT, when $n \to \infty$, we have

\[ W \to N_{2q} \{0, \begin{pmatrix} \sigma_i^2 \Gamma & -\sigma_i^2 \Gamma \\ -\sigma_i^2 \Gamma & \sigma_i^2 (\Gamma + \Sigma_{SS}^{-1}) \end{pmatrix} + \sigma_2^2 \begin{pmatrix} \Sigma_{SS}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \}, \]

where $\Gamma = \Sigma_{SS}^{-1} \Sigma_{SS} J^{-1} \Sigma_{SS} \Sigma_{SS}^{-1}$. $\square$

### F.4 Proof of Proposition 1 and 2

When $n \to \infty$, \( \text{var}(\hat{\beta}_a) \to (\sigma_1^2 + \sigma_2^2) \Sigma_{SS}^{-1} \) and \( \text{var}(\hat{\beta}) \to \sigma_1^2 \Sigma_{SS}^{-1} U \Sigma_{SS}^{-1} + \sigma_2^2 \Sigma_{SS}^{-1} \) where \( U = \Omega_0 \Sigma_{GG} \Omega_0^\top \) and \( \Omega_0 \Sigma_{GG} = \Sigma_{SG} \). So

\[ \text{var}(\hat{\beta}_a) - \text{var}(\hat{\beta}) \to \sigma_1^2 \Sigma_{SS}^{-1} (\Sigma_{SS} - U) \Sigma_{SS}^{-1} \]

and it suffices to show that \( \Sigma_{SS} - U \) is positive semi-definite. For any \( p \), the Schur complement of \( \Sigma_{GG} \) in \( D \) is \( \Sigma_{SS} - \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS} \). Since \( D \) is positive definite and symmetric, its Schur complement of \( \Sigma_{GG} \) in \( D \) is also positive definite. Let \( p \to \infty \) and we prove that \( \Sigma_{SS} - U = \Sigma_{SS} - \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS} \) is positive semi-definite.

Similarly, \( \text{var} \{n^{1/2}(\hat{\beta} - \beta_0) - n^{1/2}(\tilde{\beta} - \beta_0)\} \to PQP^\top \), where \( Q = (\Sigma_{SS} - \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS})^{-1} \) is positive semi-definite, and \( P = \Sigma_{SS}^{-1} \Sigma_{SG} \Sigma_{GG}^{-1} \Sigma_{GS} \). Thus the \( PQP^\top \) is positive semi-definite. $\square$