Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding *

Zijian Guo Domagoj Ćevid Peter Bühlmann

October 19, 2020

Abstract

Inferring causal relationships or related associations from observational data can be invalidated by the existence of hidden confounding. We focus on a high-dimensional linear regression setting, where the measured covariates are affected by hidden confounding and propose the *Doubly Debiased Lasso* estimator for individual components of the regression coefficient vector. Our advocated method simultaneously corrects both the bias due to estimation of high-dimensional parameters as well as the bias caused by the hidden confounding. We establish its asymptotic normality and also prove that it is efficient in the Gauss-Markov sense. The validity of our methodology relies on a dense confounding assumption, i.e. that every confounding variable affects many covariates. The finite sample performance is illustrated with an extensive simulation study and a genomic application.

Keywords: Causal Inference; Structural Equation Model; Dense Confounding; Linear Model; Spectral Deconfounding.

1 Introduction

Observational studies are often used to infer causal relationship in fields such as genetics, medicine, economics or finance. A major concern for confirmatory conclusions is the existence of hidden confounding [25, 40]. In this case, standard statistical methods can be severely biased, particularly for large-scale observational studies, where many measured covariates are possibly confounded.

To better address this problem, let us consider first the following linear Structural Equation Model (SEM) with a response Y_i, high-dimensional measured covariates $X_{i,\cdot} \in \mathbb{R}^p$ and hidden confounders $H_{i,\cdot} \in \mathbb{R}^q$:

$$Y_i \leftarrow \beta^\intercal X_{i,\cdot} + \phi^\intercal H_{i,\cdot} + e_i, \quad \text{and} \quad X_{i,\cdot} \leftarrow \Psi^\intercal H_{i,\cdot} + E_{i,\cdot} \quad \text{for } 1 \le i \le n, \tag{1}$$

*Z. Guo and D. Ćevid contributed equally to this work. The research of Z. Guo was supported in part by the NSF-DMS 1811857, 2015373 and NIH-1R01GM140463-01; Z. Guo also acknowledges financial support for visiting the Institute of Mathematical Research (FIM) at ETH Zurich. The research of D. Ćevid and P. Bühlmann was supported by the European Research Council under the Grant Agreement No 786461 (CausalStats - ERC-2017-ADG).

where the random error $e_i \in \mathbb{R}$ is independent of $X_{i,\cdot} \in \mathbb{R}^p$, $H_{i,\cdot} \in \mathbb{R}^q$ and $E_{i,\cdot} \in \mathbb{R}^p$ and the components of $E_{i,\cdot} \in \mathbb{R}^p$ are uncorrelated with the components of $H_{i,\cdot} \in \mathbb{R}^q$. The focus on a SEM as in (1) is not necessary and we relax this restriction in model (2) below. Such kind of models are used for e.g. biological studies to explore the effects of measured genetic variants on the disease risk factor, and the hidden confounders can be geographic information [44], data sources in mental analysis [46] or general population stratification in GWAS [41].

Our aim is to perform statistical inference for individual components β_j, $1 \leq j \leq p$, of the coefficient vector, where p can be large, in terms of obtaining confidence intervals or statistical tests. This inference problem is challenging due to high dimensionality of the model and the existence of hidden confounders. As a side remark, we mention that our proposed methodology can also be used for certain measurement error models, an important general topic in statistics and economics [9, 59].

1.1 Our Results and Contributions

We focus on a dense confounding model, where the hidden confounders $H_{i,\cdot}$ in (1) are associated with many measured covariates $X_{i,\cdot}$. Such dense confounding model seems reasonable in quite many practical applications, e.g. for addressing the problem of batch effects in biological studies [27, 31, 36].

We propose a two-step estimator for the regression coefficient β_j for $1 \leq j \leq p$ in the high-dimensional dense confounding setting, where a large number of covariates has possibly been affected by hidden confounding. In the first step, we construct a penalized spectral deconfounding estimator $\widehat{\beta}^{init}$ as in [10], where the standard squared error loss is replaced by a squared error loss after applying a certain spectral transformation to the design matrix X and the response Y. In the second step, for the regression coefficient of interest β_j, we estimate the high-dimensional nuisance parameters $\beta_{-j} = \{\beta_l;\ l \neq j\}$ by $\widehat{\beta}^{init}_{-j}$ and construct an approximately unbiased estimator $\widehat{\beta}_j$.

The main idea of the second step is to correct the bias from two sources, one from estimating the high-dimensional nuisance vector β_{-j} by $\widehat{\beta}^{init}_{-j}$ and the other arising from hidden confounding. In the standard high-dimensional regression setting with no hidden confounders, debiasing, desparsifying or Neyman's Orthogonalization were proposed for inference for β_j [61, 53, 30, 2, 13, 19, 12]. However, these methods, or some of its direct extensions, do not account for the bias arising from hidden confounding. In order to address this issue, we introduce a *Doubly Debiased Lasso* estimator which corrects both biases simultaneously. Specifically, we construct a spectral transformation $\mathcal{P}^{(j)} \in \mathbb{R}^{n \times n}$, which is applied to the nuisance design matrix X_{-j} when the parameter of interest is β_j. This spectral transformation is crucial to simultaneously correcting the two sources of bias.

We establish the asymptotic normality of the proposed *Doubly Debiased Lasso* estimator in Theorem 1. An efficiency result is also provided in Theorem 2 of Section 4.2.1, showing that the *Doubly Debiased Lasso* estimator retains the same Gauss-Markov efficiency bound as in standard high-dimensional linear regression with no hidden con-

founding [53, 29]. Our result is in sharp contrast to Instrumental Variables (IV) based methods, see Section 1.2, whose inflated variance is often of concern, especially with a limited amount of data [59, 4]. This remarkable efficiency result is possible by assuming denseness of confounding. Various intermediary results of independent interest are also derived in Section A of the Supplementary material. Finally, the performance of the proposed estimator is illustrated on simulated and real genomic data in Section 5.

To summarize, our main contribution is two-fold:

1. We propose a novel Doubly Debiased Lasso estimator for individual coefficients β_j and estimation of the corresponding standard error in a high-dimensional linear SEM with hidden confounding.

2. We show that the proposed estimator is asymptotically Gaussian and efficient in the Gauss-Markov sense. This implies the construction of asymptotically optimal confidence intervals for individual coefficients β_j.

1.2 Related Work

In econometrics, hidden confounding and measurement errors are unified under the framework of endogenous variables. Inference for treatment effects or corresponding regression parameters in presence of hidden confounders or measurement errors has been extensively studied in the literature with Instrumental Variables (IV) regression. The construction of IVs typically requires a lot of domain knowledge, and obtained IVs are often suspected of violating the main underlying assumptions [28, 59, 32, 6, 26, 58]. In high dimensions, the construction of IVs is even more challenging, since for identification of the causal effect, one has to construct as many IVs as the number of confounded covariates, which is the so-called "rank condition" [59]. Some recent work on the high-dimensional hidden confounding problem relying on the construction of IVs includes [21, 16, 38, 1, 62, 43, 23]. Another approach builds on directly estimating and adjusting with respect to latent factors [57].

A major distinction of the current work from the contributions above is that we consider a confounding model with a denseness assumption [11, 10, 50]. [10] consider point estimation of β in the high-dimensional hidden confounder model (1), whereas [50] deal with point estimation of the precision and covariance matrix of high-dimensional covariates, which are possibly confounded. The current paper is different in that it considers the challenging problem of confidence interval construction, which requires novel ideas for both methodology and theory.

The dense confounding model is also connected to the high-dimensional factor models [15, 35, 34, 18, 56]. The main difference is that the factor model literature focuses on accurately extracting the factors, while our method is essentially filtering them out in order to provide consistent estimators of regression coefficients, under much weaker requirements than for the identification of factors.

Another line of research [20, 52, 55] studies the latent confounder adjustment models but focuses on a different setting where many outcome variables can be possibly associated with a small number of observed covariates and several hidden confounders.

Notation. We use $X_j \in \mathbb{R}^n$ and $X_{-j} \in \mathbb{R}^{n\times(p-1)}$ to denote the $j-$th column of the matrix X and the sub-matrix of X excluding the $j-$th column, respectively; $X_{i,\cdot} \in \mathbb{R}^p$ is used to denote the $i-$th row of the matrix X (as a column vector); $X_{i,j}$ and $X_{i,-j}$ denote respectively the (i,j) entry of the matrix X and the sub-row of $X_{i,\cdot}$ excluding the j-th entry. Let $[p] = \{1, 2, \ldots, p\}$. For a subset $J \subset [p]$ and a vector $x \in \mathbb{R}^p$, x_J is the sub-vector of x with indices in J and x_{-J} is the sub-vector with indices in J^c. For a set S, $|S|$ denotes the cardinality of S. For a vector $x \in \mathbb{R}^p$, the ℓ_q norm of x is defined as $\|x\|_q = \left(\sum_{l=1}^p |x_l|^q\right)^{\frac{1}{q}}$ for $q \geq 0$ with $\|x\|_0 = |\{1 \leq l \leq p : x_l \neq 0\}|$ and $\|x\|_\infty = \max_{1\leq l\leq p} |x_l|$. We use e_i to denote the i-th standard basis vector in $\mathbb{R}^p$ and I_p to denote the identity matrix of size $p \times p$. We use c and C to denote generic positive constants that may vary from place to place. For a sequence of random variables X_n indexed by n, we use $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{d} X$ to represent that X_n converges to X in probability and in distribution, respectively. For a sequence of random variables X_n and numbers a_n, we define $X_n = o_p(a_n)$ if X_n/a_n converges to zero in probability. For two positive sequences a_n and b_n, $a_n \lesssim b_n$ means that $\exists C > 0$ such that $a_n \leq Cb_n$ for all n; $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$. We use $\lambda_j(M)$ to denote the j-th largest singular value of some matrix M, that is, $\lambda_1(M) \geq \lambda_2(M) \geq \ldots \geq \lambda_q(M) \geq 0$.

2 Hidden Confounding Model

We consider the Hidden Confounding Model for i.i.d. data $\{X_{i,\cdot}, Y_i\}_{1\leq i\leq n}$ and unobserved i.i.d. confounders $\{H_{i,\cdot}\}_{1\leq i\leq n}$, given by:

$$Y_i = \beta^\intercal X_{i,\cdot} + \phi^\intercal H_{i,\cdot} + e_i \quad \text{and} \quad X_{i,\cdot} = \Psi^\intercal H_{i,\cdot} + E_{i,\cdot}, \tag{2}$$

where $Y_i \in \mathbb{R}$ and $X_{i,\cdot} \in \mathbb{R}^p$ respectively denote the response and the measured covariates and $H_{i,\cdot} \in \mathbb{R}^q$ represents the hidden confounders. We assume that the random error $e_i \in \mathbb{R}$ is independent of $X_{i,\cdot} \in \mathbb{R}^p$, $H_{i,\cdot} \in \mathbb{R}^q$ and $E_{i,\cdot} \in \mathbb{R}^p$ and the components of $E_{i,\cdot} \in \mathbb{R}^p$ are uncorrelated with the components of $H_{i,\cdot} \in \mathbb{R}^q$.

The coefficient matrices $\Psi \in \mathbb{R}^{q\times p}$ and $\phi \in \mathbb{R}^{q\times 1}$ encode the linear effect of the hidden confounders $H_{i,\cdot}$ on the measured covariates $X_{i,\cdot}$ and the response Y_i. We consider the high-dimensional setting where p might be much larger than n. Throughout the paper it is assumed that the regression vector $\beta \in \mathbb{R}^p$ is sparse, with a small number k of nonzero components, and that the number q of confounding variables is small as well. However, both k and q are allowed to grow with n and p. We write Σ_E or Σ_X for the covariance matrices of $E_{i,\cdot}$ or $X_{i,\cdot}$, respectively. Without loss of generality, it is assumed that $\mathbb{E}X_{i,\cdot} = 0$, $\mathbb{E}H_{i,\cdot} = 0$, $\mathrm{Cov}(H_{i,\cdot}) = \mathrm{I}_q$ and hence $\Sigma_X = \Psi^\intercal\Psi + \Sigma_E$.

The probability model (2) is more general than the Structural Equation Model in (1). It only describes the observational distribution of the latent variable $H_{i,\cdot}$ and the observed data $(X_{i,\cdot}, Y_i)$, which possibly may be generated from the hidden confounding SEM (1).

Our goal is to construct confidence intervals for the components of β, which in the model (1) describes the causal effect of X on the response Y. The problem is challenging

due to the presence of unobserved confounding. In fact, the regression parameter β can not even be identified without additional assumptions. Our main condition addressing this issue is a denseness assumption that the rows $\Psi_{j,\cdot} \in \mathbb{R}^p$ are dense in a certain sense (see Condition (A2) in Section 4), i.e., many covariates of $X_{i,\cdot} \in \mathbb{R}^p$ are simultaneously affected by hidden confounders $H_{i,\cdot} \in \mathbb{R}^q$.

2.1 Representation as a Linear Model

The Hidden Confounding Model (2) can be represented as a linear model for the observed data $\{X_{i,\cdot}, Y_i\}_{1 \le i \le n}$:

$$Y_i = (\beta + b)^\intercal X_{i,\cdot} + \epsilon_i \quad \text{and} \quad X_{i,\cdot} = \Psi^\intercal H_{i,\cdot} + E_{i,\cdot}, \tag{3}$$

by writing

$$\epsilon_i = e_i + \phi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot} \quad \text{and} \quad b = \Sigma_X^{-1} \Psi^\intercal \phi.$$

As in (2) we assume that $E_{i,\cdot}$ is uncorrelated with $H_{i,\cdot}$ and, by construction of b, ϵ_i is uncorrelated with $X_{i,\cdot}$. With σ_e^2 denoting the variance of e_i, the variance of the error ϵ_i equals $\sigma_\epsilon^2 = \sigma_e^2 + \phi^\intercal \left(\mathrm{I}_q - \Psi \Sigma_X^{-1} \Psi^\intercal\right) \phi$. In model (3), the response is generated from a linear model where the sparse coefficient vector β has been perturbed by some perturbation vector $b \in \mathbb{R}^p$. This representation reveals how the parameter of interest β is not in general identifiable from observational data, where one can not easily differentiate it from the perturbed coefficient vector $\beta + b$, where the perturbation vector b is induced by hidden confounding. However, as shown in Lemma 2 in the supplement, b is dense and thus small for large p under the assumption of dense confounding, which enables us to identify β asymptotically. It is important to note that the term $b^\intercal X_{i,\cdot}$ induced by hidden confounders $H_{i,\cdot}$ is not necessarily small and hence cannot be simply ignored in model (3), but requires novel methodological approach.

Connection to measurement errors We briefly relate certain measurement error models to the Hidden Confounding Model (2). Consider a linear model for the outcome Y_i and covariates $X_{i\cdot}^0 \in \mathbb{R}^p$, where we only observe $X_{i,\cdot} \in \mathbb{R}^p$ with measurement error $W_{i,\cdot} \in \mathbb{R}^p$:

$$Y_i = \beta^\intercal X_{i\cdot}^0 + e_i \quad \text{and} \quad X_{i,\cdot} = X_{i,\cdot}^0 + W_{i,\cdot} \quad \text{for } 1 \le i \le n. \tag{4}$$

Here, e_i is a random error independent of $X_{i,\cdot}^0$ and $W_{i,\cdot}$, and $W_{i,\cdot}$ is the measurement error independent of X_i^0. We can then express a linear dependence of Y_i on the observed $X_{i,\cdot}$,

$$Y_i = \beta^\intercal X_{i,\cdot} + (e_i - \beta^\intercal W_{i,\cdot}) \quad \text{and} \quad X_{i,\cdot} = W_{i,\cdot} + X_{i,\cdot}^0$$

We further assume the following structure of the measurement error:

$$W_{i,\cdot} = \Psi^\intercal H_{i,\cdot},$$

i.e. there exist certain latent variables $H_{i,\cdot} \in \mathbb{R}^q$ that contribute independently and linearly to the measurement error, a conceivable assumption in some practical applications.

Combining this with the equation above we get

$$Y_i = \beta^\intercal X_{i,\cdot} + (e_i - \phi^\intercal H_{i,\cdot}) \quad \text{and} \quad X_{i,\cdot} = \Psi^\intercal H_{i,\cdot} + X^0_{i\cdot}, \tag{5}$$

where $\phi = \Psi\beta \in \mathbb{R}^q$. Therefore, the model (5) can be seen as a special case of the model (2), by identifying $X^0_{i\cdot}$ in (5) with $E_{i,\cdot}$ in (2).

3 Doubly Debiased Lasso Estimator

In this section, for a fixed index $j \in \{1, \ldots, p\}$, we propose an inference method for the regression coefficient β_j of the Hidden Confounding Model (2). The validity of the method is demonstrated by considering the equivalent model (3).

3.1 Double Debiasing

We denote by $\widehat{\beta}^{init}$ an initial estimator of β. We will use the spectral deconfounding estimator proposed in [10], described in detail in Section 3.4. We start from the following decomposition:

$$Y - X_{-j}\widehat{\beta}^{init}_{-j} = X_j\,(\beta_j + b_j) + X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j}) + X_{-j}b_{-j} + \epsilon \quad \text{for} \quad j \in \{1, \ldots, p\}. \tag{6}$$

The above decomposition reveals two sources of bias: the bias $X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j})$ due to the error of the initial estimator $\widehat{\beta}^{init}$ and the bias $X_{-j}b_{-j}$ induced by the perturbation vector b in the model (3), arising by marginalizing out the hidden confounding in (2). Note that the bias b_j is negligible in the dense confounding setting, see Lemma 2 in the supplement. The first bias, due to penalization, appears in the standard high-dimensional linear regression as well, and can be corrected with the debiasing methods proposed in [61, 53, 30] when assuming no hidden confounding. However, in presence of hidden confounders, methodological innovation is required for correcting both bias terms and conducting the resulting statistical inference. We propose a novel Doubly Debiased Lasso estimator for correcting both sources of bias simultaneously.

Denote by $\mathcal{P}^{(j)} \in \mathbb{R}^{n\times n}$ a symmetric spectral transformation matrix, which shrinks the singular values of the sub-design $X_{-j} \in \mathbb{R}^{n\times(p-1)}$. The detailed construction, together with some examples, is given in Section 3.3. We shall point out that the construction of the transformation matrix $\mathcal{P}^{(j)}$ depends on which coefficient β_j is our target and hence refer to $\mathcal{P}^{(j)}$ as the nuisance spectral transformation with respect to the coefficient β_j. Multiplying both sides of the decomposition (6) with the transformation $\mathcal{P}^{(j)}$ gives:

$$\mathcal{P}^{(j)}(Y - X_{-j}\widehat{\beta}^{init}_{-j}) = \mathcal{P}^{(j)}X_j\,(\beta_j + b_j) + \mathcal{P}^{(j)}X_{-j}(\beta_{-j} - \widehat{\beta}^{init}_{-j}) + \mathcal{P}^{(j)}X_{-j}b_{-j} + \mathcal{P}^{(j)}\epsilon. \tag{7}$$

The quantity of interest β_j appears on the RHS of the equation (7) next to the vector $\mathcal{P}^{(j)}X_j$, whereas the additional bias lies in the span of the columns of $\mathcal{P}^{(j)}X_{-j}$. For this reason, we construct a projection direction vector $\mathcal{P}^{(j)}Z_j \in \mathbb{R}^n$ as the transformed residuals of regressing X_j on X_{-j}:

$$Z_j = X_j - X_{-j}\widehat{\gamma}, \tag{8}$$

where the coefficients $\widehat{\gamma}$ are estimated with the Lasso for the transformed covariates using $\mathcal{P}^{(j)}$:

$$\widehat{\gamma} = \underset{\gamma \in \mathbb{R}^{p-1}}{\arg\min} \left\{ \frac{1}{2n} \|\mathcal{P}^{(j)} X_j - \mathcal{P}^{(j)} X_{-j} \gamma\|_2^2 + \lambda_j \sum_{l \neq j} \frac{\|\mathcal{P}^{(j)} X_{.,l}\|_2}{\sqrt{n}} |\gamma_l| \right\}, \tag{9}$$

with $\lambda_j = A\sigma_j \sqrt{\log p/n}$ for some positive constant $A > \sqrt{2}$ (for σ_j, see Section 4.1).

Finally, motivated by the equation (7), we propose the following estimator for β_j:

$$\widehat{\beta}_j = \frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} (Y - X_{-j} \widehat{\beta}^{init}_{-j})}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j}. \tag{10}$$

We refer to this estimator as the Doubly Debiased Lasso estimator as it simultaneously corrects the bias induced by $\widehat{\beta}^{init}$ and the confounding bias $X_{-j} b_{-j}$ by using the spectral transformation $\mathcal{P}^{(j)}$.

In the following, we briefly explain why the proposed estimator estimates β_j well. We start with the following error decomposition of $\widehat{\beta}_j$, derived from (7)

$$\widehat{\beta}_j - \beta_j = \underbrace{\frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} \epsilon}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j}}_{\text{Variance}} + \underbrace{\frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_{-j} (\beta_{-j} - \widehat{\beta}^{init}_{-j})}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j} + \frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_{-j} b_{-j}}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j} + b_j}_{\text{Remaining Bias}}. \tag{11}$$

In the above equation, the bias after correction consists of two components: the remaining bias due to the estimation error of $\widehat{\beta}^{init}_{-j}$ and the remaining confounding bias due to $X_{-j} b_{-j}$ and b_j. These two components can be shown to be negligible in comparison to the variance component under certain model assumptions, see Theorem 1 and its proof for details. Intuitively, the construction of the spectral transformation matrix $\mathcal{P}^{(j)}$ is essential for reducing the bias due to the hidden confounding. The term $\frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_{-j} b_{-j}}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j}$ in equation (11) is of a small order because $\mathcal{P}^{(j)}$ shrinks the leading singular values of X_{-j} and hence $\mathcal{P}^{(j)} X_{-j} b_{-j}$ is significantly smaller than $X_{-j} b_{-j}$. The induced bias $X_{-j} b_{-j}$ is not negligible since b_{-j} points in the direction of leading right singular vectors of X_{-j}, thus leading to $\|\frac{1}{\sqrt{n}} X_{-j} b_{-j}\|_2$ being of constant order. By applying a spectral transformation to shrink the leading singular values, one can show that $\|\frac{1}{\sqrt{n}} \mathcal{P}^{(j)} X_{-j} b_{-j}\|_2 = O_p(1/\sqrt{\min\{n, p\}})$.

Furthermore, the other remaining bias term $\frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_{-j} (\beta_{-j} - \widehat{\beta}^{init}_{-j})}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j}$ in (11) is small since the initial estimator $\widehat{\beta}^{init}$ is close to β in ℓ_1 norm and $\mathcal{P}^{(j)} Z_j$ and $\mathcal{P}^{(j)} X_{-j}$ are nearly orthogonal due to the construction of $\widehat{\gamma}$ in (9). This bias correction idea is analogous to the Debiased Lasso estimator introduced in [61] for the standard high-dimensional linear regression:

$$\widehat{\beta}^{DB}_j = \frac{(Z^{DB}_j)^\intercal (Y - X_{-j} \widehat{\beta}^{init}_{-j})}{(Z^{DB}_j)^\intercal X_j}, \tag{12}$$

where Z^{DB}_j is constructed similarly as in (8) and (9), but where $\mathcal{P}^{(j)}$ is the identity matrix. Therefore, the main difference between the estimator in (12) and our proposed

estimator (10) is that for its construction we additionally apply the nuisance spectral transformation $\mathcal{P}^{(j)}$.

We emphasize that the additional spectral transformation $\mathcal{P}^{(j)}$ is necessary even for just correcting the bias of $\widehat{\beta}_{-j}^{init}$ in presence of confounding (i.e., it is also needed for the first besides the second bias term in (11)). To see this, we define the best linear projection of $X_{1,j}$ to all other variables $X_{1,-j} \in \mathbb{R}^{p-1}$ with the coefficient vector $\gamma = [\mathbb{E}(X_{i,-j}X_{i,-j}^\intercal)]^{-1}\mathbb{E}(X_{i,-j}X_{i,j}) \in \mathbb{R}^{p-1}$ (which is then estimated by the Lasso in the standard construction of Z_j^{DB}). We notice that γ need not be sparse due to the fact that all covariates are affected by a common set of hidden confounders yielding spurious associations. Hence, the standard construction of Z_j^{DB} in (12) is not favorable in the current setting. In contrast, the proposed method with $\mathcal{P}^{(j)}$ works for two reasons: first, the application of $\mathcal{P}^{(j)}$ in (9) leads to a consistent estimator of the sparse component of γ, denoted as γ^M (see the expression of γ^M in Lemma 1); second, the application of $\mathcal{P}^{(j)}$ leads to a small prediction error $\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma} - \gamma^M)$. We illustrate in Section 5 how the application of $\mathcal{P}^{(j)}$ corrects the bias due to $\widehat{\beta}_{-j}^{init}$ and observe a better empirical coverage after applying $\mathcal{P}^{(j)}$ in comparison to the standard debiased Lasso in (12); see Figure 4.

3.2 Confidence Interval Construction

In Section 4, we establish the asymptotic normal limiting distribution of the proposed estimator $\widehat{\beta}_j$ under certain regularity conditions. Its standard deviation can be estimated by $\sqrt{\frac{\widehat{\sigma}_e^2 \cdot Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j}{[Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j]^2}}$ with $\widehat{\sigma}_e$ denoting a consistent estimator of σ_e. The detailed construction of $\widehat{\sigma}_e$ is described in Section 3.5. Therefore, a confidence interval (CI) with asymptotic coverage $1-\alpha$ can be obtained as

$$\mathrm{CI}(\beta_j) = \left(\widehat{\beta}_j - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\sigma}_e^2 \cdot Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j}{[Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j]^2}}, \widehat{\beta}_j + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\sigma}_e^2 \cdot Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j}{[Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j]^2}} \right), \tag{13}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal random variable.

3.3 Construction of Spectral Transformations

Construction of the spectral transformation $\mathcal{P}^{(j)} \in \mathbb{R}^{n\times n}$ is an essential step for the Doubly Debiased Lasso estimator (10). The transformation $\mathcal{P}^{(j)} \in \mathbb{R}^{n\times n}$ is a symmetric matrix shrinking the leading singular values of the design matrix $X_{-j} \in \mathbb{R}^{n\times(p-1)}$. Denote by $m = \min\{n, p-1\}$ and the SVD of the matrix X_{-j} by $X_{-j} = U(X_{-j})\Lambda(X_{-j})[V(X_{-j})]^\intercal$, where $U(X_{-j}) \in \mathbb{R}^{n\times m}$ and $V(X_{-j}) \in \mathbb{R}^{(p-1)\times m}$ have orthonormal columns and $\Lambda(X_{-j}) \in \mathbb{R}^{m\times m}$ is a diagonal matrix of singular values which are sorted in a decreasing order $\Lambda_{1,1}(X_{-j}) \geq \Lambda_{2,2}(X_{-j}) \geq \ldots \geq \Lambda_{m,m}(X_{-j}) \geq 0$. We then define the spectral transformation $\mathcal{P}^{(j)}$ for X_{-j} as $\mathcal{P}^{(j)} = U(X_{-j})S(X_{-j})[U(X_{-j})]^\intercal$, where $S(X_{-j}) \in \mathbb{R}^{m\times m}$ is a diagonal shrinkage matrix with $0 \leq S_{l,l}(X_{-j}) \leq 1$ for $1 \leq l \leq m$. The SVD for the complete design matrix X is defined analogously. We highlight the dependence of

the SVD decomposition on X_{-j}, but for simplicity it will be omitted when there is no confusion. Note that $\mathcal{P}^{(j)}X_{-j} = U\,(S\Lambda)\,V^\intercal$, so the spectral transformation shrinks the singular values $\{\Lambda_{l,l}\}_{1\le l\le m}$ to $\{S_{l,l}\Lambda_{l,l}\}_{1\le l\le m}$, where $\Lambda_{l,l} = \Lambda_{l,l}(X_{-j})$.

Trim transform For the rest of this paper, the spectral transformation that is used is the Trim transform [10]. It limits any singular value to be at most some threshold τ. This means that the shrinkage matrix S is given as: for $1 \le l \le m$,

$$S_{l,l} = \begin{cases} \tau/\Lambda_{l,l} & \text{if} \quad \Lambda_{l,l} > \tau \\ 1 & \text{otherwise} \end{cases}.$$

A good default choice for the threshold τ is the median singular value $\Lambda_{\lfloor m/2\rfloor,\lfloor m/2\rfloor}$, so only the top half of the singular values is shrunk to the bulk value $\Lambda_{\lfloor m/2\rfloor,\lfloor m/2\rfloor}$ and the bottom half is left intact. More generally, one can use any percentile $\rho_j \in (0,1)$ to shrink the top $(100\rho_j)\%$ singular values to the corresponding ρ_j-quantile $\Lambda_{\lfloor \rho_j m\rfloor,\lfloor \rho_j m\rfloor}$. We define the ρ_j-Trim transform $\mathcal{P}^{(j)}$ as

$$\mathcal{P}^{(j)} = U(X_{-j})S(X_{-j})[U(X_{-j})]^\intercal \text{ with } S_{l,l}(X_{-j}) = \begin{cases} \frac{\Lambda_{\lfloor \rho_j m\rfloor,\lfloor \rho_j m\rfloor}(X_{-j})}{\Lambda_{l,l}(X_{-j})} & \text{if } \; l \le \lfloor \rho_j m\rfloor \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

In Section 4 we investigate the dependence of the asymptotic efficiency of the resulting Doubly Debiased Lasso $\widehat{\beta}_j$ on the percentile choice $\rho_j = \rho_j(n)$. There is a certain trade-off in choosing ρ_j: a smaller value of ρ_j leads to a more efficient estimator, but one needs to be careful to keep $\rho_j m$ sufficiently large compared to the number of hidden confounders q, in order to ensure reduction of the confounding bias. In Section A.1 of the supplementary material, we describe the general conditions that the used spectral transformations need to satisfy to ensure good performance of the resulting estimator.

3.4 Initial Estimator $\widehat{\beta}^{init}$

For the Doubly Debiased Lasso (10), we use the spectral deconfounding estimator proposed in [10] as our initial estimator $\widehat{\beta}^{init}$. It uses a spectral transformation $\mathcal{Q} = \mathcal{Q}(X)$, constructed similarly as the transformation $\mathcal{P}^{(j)}$ described in Section 3.3, with the difference that instead of shrinking the singular values of X_{-j}, $\mathcal{Q}$ shrinks the leading singular values of the whole design matrix $X \in \mathbb{R}^{n\times p}$. Specifically, for any percentile $\rho \in (0,1)$, the ρ-Trim transform $\mathcal{Q}$ is given by

$$\mathcal{Q} = U(X)S(X)[U(X)]^\intercal \text{ with } S_{l,l}(X) = \begin{cases} \frac{\Lambda_{\lfloor \rho m\rfloor,\lfloor \rho m\rfloor}(X)}{\Lambda_{l,l}(X)} & \text{if } \; l \le \lfloor \rho m\rfloor \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

The estimator $\widehat{\beta}^{init}$ is computed by applying the Lasso to the transformed data $\mathcal{Q}X$ and $\mathcal{Q}Y$:

$$\widehat{\beta}^{init} = \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{2n}\|\mathcal{Q}\,(y - X\beta)\,\|_2^2 + \lambda\sum_{j=1}^{p}\frac{\|\mathcal{Q}X_{\cdot j}\|_2}{\sqrt{n}}|\beta_j|, \tag{16}$$

where $\lambda = A\sigma_e\sqrt{\log p/n}$ is a tuning parameter with $A > \sqrt{2}$.

The transformation $\mathcal{Q}$ reduces the effect of the confounding and thus helps for estimation of β. In Section A.3, the ℓ_1 and ℓ_2-error rates of $\widehat{\beta}^{init}$ are given, thereby extending the results of [10].

3.5 Noise Level Estimator

In addition to an initial estimator of β, we also require a consistent estimator $\widehat{\sigma}_e^2$ of the error variance $\sigma_e^2 = \mathbb{E}(e_i^2)$ for construction of confidence intervals. Choosing a noise level estimator which performs well for a wide range of settings is not easy to do in practice [47]. We propose using the following estimator:

$$\widehat{\sigma}_e^2 = \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}y - \mathcal{Q}X\widehat{\beta}^{init}\|_2^2, \tag{17}$$

where $\mathcal{Q}$ is the same spectral transformation as in (16).

The motivation for this estimator is based on the expression

$$\mathcal{Q}y - \mathcal{Q}X\widehat{\beta}^{init} = \mathcal{Q}\epsilon + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb, \tag{18}$$

which follows from the model (3). The consistency of the proposed noise level estimator, formally shown in Proposition 1, follows from the following observations: the initial spectral deconfounding estimator $\widehat{\beta}^{init}$ has a good rate of convergence for estimating β; the spectral transformation $\mathcal{Q}$ significantly reduces the additional error Xb induced by the hidden confounders; $\|\mathcal{Q}\epsilon\|_2^2/\mathrm{Tr}(\mathcal{Q}^2)$ consistently estimates σ_ϵ^2. Additionally, the dense confounding model is shown to lead to a small difference between the noise levels σ_ϵ^2 and σ_e^2, see Lemma 2 in the supplement. In Section 4 we show that variance estimator $\widehat{\sigma}_e^2$ defined in (17) is a consistent estimator of σ_e^2.

3.6 Method Overview and Choice of the Tuning Parameters

The Doubly Debiased Lasso needs specification of various tuning parameters. A good and theoretically justified rule of thumb is to use the Trim transform with $\rho = \rho_j = 1/2$, which shrinks the large singular values to the median singular value, see (14). Furthermore, similarly to the standard Debiased Lasso [61], our proposed method involves the regularization parameters λ in the Lasso regression for the initial estimator $\widehat{\beta}^{init}$ (see equation (16)) and λ_j in the Lasso regression for the projection direction $\mathcal{P}^{(j)}Z_j$ (see equation (9)). For choosing λ we use 10-fold cross-validation, whereas for λ_j, we increase slightly the penalty chosen by the 10-fold cross-validation, so that the variance of our estimator, which can be determined from the data up to a proportionality factor σ_e^2, increases by 25%, as proposed in [14].

The proposed Doubly Debiased Lasso method is summarized in Algorithm 1, which also highlights where each tuning parameter is used.

Algorithm 1 Doubly Debiased Lasso

Input: Data $X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n$; index j, tuning parameters $\rho, \rho_j \in (0,1)$ and $\lambda > 0$, $\lambda_j > 0$

Output: Point estimator $\widehat{\beta}_j$, standard error estimate $\widehat{\sigma}_e^2$ and confidence interval $\mathrm{CI}(\beta_j)$

$\mathcal{Q} \leftarrow \textsc{TrimTransform}(X, \rho)$ ▷ construct ρ-trim as in (15)
$\widehat{\beta}^{init} \leftarrow \textsc{Lasso}(\mathcal{Q}X,\, \mathcal{Q}Y,\, \lambda)$ ▷ Lasso regression with transformed data, see (16)

$\mathcal{P}^{(j)} \leftarrow \textsc{TrimTransform}(X_{-j}, \rho_j)$ ▷ construct ρ_j-trim as in (14)
$\widehat{\gamma} \leftarrow \textsc{Lasso}(\mathcal{P}^{(j)}X_{-j},\, \mathcal{P}^{(j)}X_j,\, \lambda_j)$ ▷ Lasso regression with transformed data, see (9)
$\mathcal{P}^{(j)}Z_j \leftarrow \mathcal{P}^{(j)}X_j - \mathcal{P}^{(j)}X_{-j}\widehat{\gamma}$ ▷ take the residuals as the projection direction
$\widehat{\beta}_j \leftarrow \textsc{DebiasedLasso}(\widehat{\beta}^{init}, \mathcal{P}^{(j)}X_{-j},\, \mathcal{P}^{(j)}X_j,\, \mathcal{P}^{(j)}Z_j)$ ▷ compute Doubly Debiased Lasso as in (12)

$\widehat{\sigma}_e^2 \leftarrow \textsc{NoiseLevel}(X, Y, \widehat{\beta}^{init}, \mathcal{Q})$ ▷ compute noise level as in (17)
$\mathrm{CI}(\beta_j) \leftarrow \textsc{ConfidenceInterval}(\widehat{\beta}_j, \mathcal{P}^{(j)}X_j, \mathcal{P}^{(j)}Z_j, \widehat{\sigma}_e^2, \alpha)$ ▷ compute the $(1-\alpha)$-CI as in (13)

4 Theoretical Justification

The current section provides theoretical justifications of the proposed method for the Hidden Confounding Model (2). The proof of the main result is presented in Section A of the supplementary materials together with several other technical results of independent interest.

4.1 Model assumptions

In the following, we fix the index $1 \le j \le p$ and introduce the model assumptions for establishing the asymptotic normality of our proposed estimator $\widehat{\beta}_j$ defined in (10). For the coefficient matrix $\Psi \in \mathbb{R}^{q\times p}$ in (3), we use $\Psi_j \in \mathbb{R}^q$ to denote the j-th column and $\Psi_{-j} \in \mathbb{R}^{q\times(p-1)}$ denotes the sub-matrix with the remaining $p-1$ columns. Furthermore, we write γ for the best linear approximation of $X_{1,j} \in \mathbb{R}$ by $X_{1,-j} \in \mathbb{R}^{p-1}$, that is $\gamma = \arg\min_{\gamma' \in \mathbb{R}^{p-1}} \mathbb{E}(X_{1,j} - X_{1,-j}\gamma')^2$, whose explicit expression is:

$$\gamma = [\mathbb{E}(X_{1,-j}X_{1,-j}^\intercal)]^{-1}\mathbb{E}(X_{1,-j}X_{1,j}).$$

We denote the corresponding residuals by $\eta_{i,j} = X_{i,j} - X_{i,-j}^\intercal\gamma$ for $1 \le i \le n$ and use σ_j to denote its standard error.

The first assumption is on the precision matrix of $E_{i,\cdot} \in \mathbb{R}^p$ in (2):

(A1) The precision matrix $\Omega_E = [\mathbb{E}(E_{i,\cdot}E_{i,\cdot}^\intercal)]^{-1}$ satisfies $c_0 \le \lambda_{\min}(\Omega_E) \le \lambda_{\max}(\Omega_E) \le C_0$ and $\|(\Omega_E)_{\cdot,j}\|_0 \le s$ where $1 \le j \le p$, $C_0 > 0$ and $c_0 > 0$ are some positive constants and s denotes the sparsity level which can grow with n and p.

Such assumptions on well-posedness and sparsity are commonly required for estimation of the precision matrix [42, 33, 60, 8] and are also used for confidence interval construction in the standard high-dimensional regression model without unmeasured confounding [53]. Here, the conditions are not directly imposed on the covariates $X_{i,\cdot}$, but rather on their unconfounded part $E_{i,\cdot}$.

The second assumption is about the coefficient matrix Ψ in (3), which describes the effect of the hidden confounding variables $H_{i,\cdot} \in \mathbb{R}^q$ on the measured variables $X_{i,\cdot} \in \mathbb{R}^p$:

(A2) The q-th singular value of the coefficient matrix $\Psi_{-j} \in \mathbb{R}^{q\times(p-1)}$ satisfies $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ and we have

$$\max\{\|\Psi(\Omega_E)_{\cdot,j}\|_2, \|\Psi_j\|_2, \|\Psi_{-j}(\Omega_E)_{-j,j}\|_2, \|\phi\|_2\} \lesssim \sqrt{q}(\log p)^c, \tag{19}$$

where Ψ and ϕ are defined in (2) and $0 < c \leq 1/4$ is some positive constant.

The condition (A2) is crucial for identifying the coefficient β_j in the high-dimensional Hidden Confounding Model (2). Condition (A2) is referred to as the dense confounding assumption. A few remarks are in order regarding when this identifiability condition holds.

Since all vectors $\Psi(\Omega_E)_{\cdot,j}$, Ψ_j, $\Psi_{-j}(\Omega_E)_{-j,j}$ and ϕ are q-dimensional, the upper bound condition (19) on their ℓ_2 norms are mild. If the vector $\phi \in \mathbb{R}^q$ has bounded entries and the vectors $\{\Psi_{\cdot,l}\}_{1\leq l\leq p} \in \mathbb{R}^q$ are independently generated with zero mean and bounded second moments, then the condition (19) holds with probability larger than $1-(\log p)^{-2c}$. A larger value $c > 1/4$ is possible: the condition then holds with even higher probability but makes the upper bounds for (31) in Lemma 1 and (32) in Lemma 2 slightly worse, which then requires more stringent conditions on k and q in Theorem 1, up to polynomial order of $\log p$.

In the factor model literature, cf. [17, 56], the spiked singular value condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ is quite common and holds under mild conditions. The Hidden Confounding Model is closely related to the factor model, where the hidden confounders $H_{i,\cdot}$ are the factors and the matrix Ψ describes how these factors affect the observed variables $X_{i,\cdot}$. The spiked singular value condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ can be shown to hold in certain dense confounding settings, that is, the hidden confounders affect a large number of covariates. The rigorous arguments are given in Section A.5 in the supplementary material. To illustrate this, consider first the special case with a single hidden confounder, that is, $q = 1$ and the effect matrix is reduced to a vector $\Psi \in \mathbb{R}^p$. In this case, $\lambda_1(\Psi_{-j}) = \|\Psi_{-j}\|_2$ and the denseness of the effect vector Ψ_{-j} leads to a large $\lambda_1(\Psi_{-j})$; by further assuming $\{\Psi_l\}_{1\leq l\leq p}$ are generated in an i.i.d. fashion, then $\lambda_1(\Psi_{-j})$ is at the scale of $\sqrt{p}$ with a high probability. Hence, if the only confounder affects many covariates, the condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ is automatically satisfied with a high probability. In the multiple hidden confounders setting, if the vectors $\{\Psi_l\}_{1\leq j\leq p}$ are generated as i.i.d. sub-Gaussian random vectors, which has an interpretation that all covariates are analogously affected by the confounders, then the spiked singular value condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ is satisfied with a high probability.

Furthermore, assumption (A2) holds in even more general settings: the condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ holds if $\{\Psi_l\}_{1\le l\le p}$ are independent sub-Gaussian random vectors and only a fixed proportion $r \in (0, 1]$ of these effects vectors is generated in an i.i.d. fashion; see Lemmas 4 and 5 in Section A.5 of the supplementary material for the statement. As a special case, this includes a practical setting where only a certain proportion of covariates is confounded. In Section 5.1, we also explore the numerical performance for different proportions $r \in (0, 1]$ of affected covariates and observe that the proposed method works well even if the hidden confounders only affect a small percentage of the covariates, say $r = 5\%$.

The condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ can in fact be empirically checked using the sample covariance matrix $\widehat{\Sigma}_X$. Since $\Sigma_X = \Psi^\intercal\Psi + \Sigma_E$, then the condition $\lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}$ implies that Σ_X has at least q spiked eigenvalues, which are of order p. If the population covariance matrix Σ_X has a few spikes, the corresponding sample covariance matrix will also have spiked eigenvalue structure with a high probability [56]. Hence, we can inspect the spectrum of the sample covariance matrix $\widehat{\Sigma}_X$ and informally check whether it has spiked singular values. See the left panel of Figure 2.

The third assumption is imposed on the distribution of various terms:

(A3) The random error e_i in (2) is assumed to be independent of $(X_{i,\cdot}^\intercal, H_{i,\cdot}^\intercal)^\intercal$ and the noise term $\eta_{i,j} = X_{i,j} - X_{i,-j}^\intercal\gamma$ is assumed to be independent of $X_{i,-j}$. Furthermore, $(E_{i,\cdot}^\intercal, e_i, \eta_{i,j})^\intercal$ is a sub-Gaussian random vector with sub-Gaussian norm $M_0 > 0$; for $1 \le j \le p$, $X_{i,j}$ is a sub-Gaussian random variable with sub-Gaussian norm $M_0 > 0$ (same constant M_0 is used for the ease of notation).

The independence assumption between the random error e_i and $(X_{i,\cdot}^\intercal, H_{i,\cdot}^\intercal)^\intercal$ in (1) is commonly assumed in SEMs, see for example [45]. This implies independence between e_i and $(X_{i,\cdot}^\intercal, H_{i,\cdot}^\intercal)^\intercal$ when the Hidden Confounder Model (2) is induced by the SEM (1). The independence assumption between $\eta_{i,j}$ and $X_{i,-j}$ holds automatically if $X_{i,\cdot}$ has a multivariate Gaussian distribution. As a remark, we assume individual components $X_{i,j}$ to be sub-Gaussian, instead of the whole vector $X_{i,\cdot} \in \mathbb{R}^p$.

The final assumption is that the restricted eigenvalue condition [3] for the transformed design matrices $\mathcal{Q}X$ and $\mathcal{P}^{(j)}X_{-j}$ is satisfied with high probability.

(A4) With probability at least $1 - \exp(-cn)$, we have

$$\mathrm{RE}\left(\tfrac{1}{n}X^\intercal\mathcal{Q}^2X\right) = \inf_{\substack{\mathcal{T}\subset[p]\\|\mathcal{T}|\le k}} \min_{\substack{\omega\in\mathbb{R}^p\\\|\omega_{\mathcal{T}^c}\|_1\le C\|\omega_{\mathcal{T}}\|_1}} \frac{\omega^\intercal\left(\frac{1}{n}X^\intercal\mathcal{Q}^2X\right)\omega}{\|\omega\|_2^2} \ge \tau_*; \tag{20}$$

$$\mathrm{RE}(\frac{1}{n}X_{-j}^\intercal(\mathcal{P}^{(j)})^2X_{-j}) = \inf_{\substack{\mathcal{T}_j\subset[p]\setminus\{j\}\\|\mathcal{T}_j|\le s}} \min_{\substack{\omega\in\mathbb{R}^{p-1}\\\|\omega_{\mathcal{T}_j^c}\|_1\le C\|\omega_{\mathcal{T}_j}\|_1}} \frac{\omega^\intercal(\frac{1}{n}X_{-j}^\intercal(\mathcal{P}^{(j)})^2X_{-j})\omega}{\|\omega\|_2^2} \ge \tau_*, \tag{21}$$

for some constants $c, C, \tau_* > 0$. (For ease of notation, we use the same constants τ_* and C in (20) and (21).)

Such assumptions are common in the high-dimensional statistics literature, cf. [5]. We show in Section A.6 of the supplementary material that the assumption (A4) holds for a broad class of random design matrices X: for moderately dimensional matrices with any sub-Gaussian distribution of the rows (see Proposition 5) and for high-dimensional Gaussian design matrices (see Proposition 6).

4.2 Main Results

In this section we present the most important properties of the proposed estimator (10). We always consider asymptotic expressions in the limit where both $n, p \to \infty$ and focus on the high-dimensional regime with $c^* = \lim p/n \in (0, \infty]$.

4.2.1 Asymptotic normality

We first present the limiting distribution of the proposed Doubly Debiased Lasso estimator.

Theorem 1. *Consider the Hidden Confounding Model* (2). *Suppose that conditions* (A1) − (A4) *hold and further assume that* $c^* = \lim p/n \in (0, \infty]$, $k \ll \sqrt{\min\{n, p/q\}}/\log p$, $s \ll n/\log p$, $q \ll \min\left\{\sqrt{n}/(\log p)^{3/4}, n/[s(\log p)^{3/2}], p/(n \log p)\right\}$, *and* $e_i \sim N(0, \sigma_e^2)$. *We recall that* $s = \|(\Omega_E)_{\cdot,j}\|_0$ *and* $k = \|\beta\|_0$. *Let the tuning parameters for* $\widehat{\beta}^{init}$ *in* (16) *and* $\widehat{\gamma}$ *in* (9) *respectively be* $\lambda \geq A\sigma_e\sqrt{\log p/n} + \sqrt{q \log p/p}$ *and* $\lambda_j \geq A\sigma_j\sqrt{\log p/n}$, *for some positive constant* $A > \sqrt{2}$. *Furthermore, let* $\mathcal{Q}$ *and* $\mathcal{P}^{(j)}$ *be the Trim transform* (14) *with* $\min\{\rho, \rho_j\} \geq (3q+1)/\min\{n, p-1\}$. *Then the Doubly Debiased Lasso estimator* (10) *satisfies*

$$\frac{1}{\sqrt{V}}\left(\widehat{\beta}_j - \beta_j\right) \xrightarrow{d} N(0, 1), \tag{22}$$

where

$$\mathrm{V} = \frac{\sigma_e^2 \cdot Z_j^\intercal (\mathcal{P}^{(j)})^4 Z_j}{[Z_j^\intercal (\mathcal{P}^{(j)})^2 X_j]^2} \quad \textit{and} \quad V^{-1}\frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} \xrightarrow{p} 1. \tag{23}$$

Remark 1. The Gaussianity of the random error e is mainly imposed to simplify the proof of asymptotic normality. We believe that this assumption is a technical condition and can be removed by applying more refined probability arguments as in [24] where the asymptotic normality of quadratic forms $(\mathcal{P}^{(j)}e)^\intercal \mathcal{P}^{(j)}e$ is established for the general sub-Gaussian case. The argument could be extended to obtain the asymptotic normality for $(\mathcal{P}^{(j)}\eta_j)^\intercal \mathcal{P}^{(j)}e$, which is essentially needed for the current result.

Remark 2. In constructing $\mathcal{Q}$ and $\mathcal{P}^{(j)}$, the main requirement is to trim the singular values enough in both cases, that is, $\min\{\rho, \rho_j\} \geq (3q+1)/\min\{n, p-1\}$. This condition is mild in the high-dimensional setting with a small number of hidden confounders. Our results are not limited to the proposed estimator which uses the Trim transform $\mathcal{P}^{(j)}$ in (14) and the penalized estimators $\widehat{\gamma}$ and $\widehat{\beta}^{init}$ in (9) and (16). All results hold for any transformation satisfying the conditions given in Section A.1 of the supplementary materials and any initial estimator satisfying the error rates presented in Section A.3 of the supplementary materials.

Remark 3. If we further assume the error ϵ_i in the model (3) to be independent of $X_{i,\cdot}$, then the conditions on the number of hidden confounders q and the sparsity k can be weakened to $q \ll \min\left\{\sqrt{n}/(\log p)^{3/4}, n/[s(\log p)^{3/2}]\right\}$ and $k \ll \sqrt{n}/\log p$.

There are three conditions on the parameters (s, q, k) imposed in the above Theorem 1. The most stringent one is the sparsity assumption $k \ll \sqrt{\min\{n, p/q\}}/\log p$. In standard high-dimensional sparse linear regression, a related sparsity assumption $k \ll \sqrt{n}/\log p$ has also been used for confidence interval construction [61, 53, 30] and has been established in [7] as a necessary condition for constructing adaptive confidence intervals. In the high-dimensional Hidden Confounder model with $p > nq$, the condition on the sparsity of β is then of the same asymptotic order as in standard high-dimensional regression with no hidden confounders. The condition on the sparsity of the precision matrix Ω_E, $s = \|(\Omega_E)_{\cdot,j}\|_0 \ll n/\log p$, is mild in the sense that it is the maximal sparsity level for identifying $(\Omega_E)_{\cdot,j}$. The condition that the number of hidden confounders q is small is fundamental for all reasonable factor or confounding models.

4.2.2 Efficiency

We investigate now the dependence of the asymptotic variance V in (23) on the choice of the spectral transformation $\mathcal{P}^{(j)}$. We further show that the proposed Doubly Debiased Lasso estimator (10) is efficient in the Gauss-Markov sense, with a careful construction of the transformation $\mathcal{P}^{(j)}$.

The Gauss-Markov theorem states that the smallest variance of any unbiased linear estimator of β_j in the standard low-dimensional regression setting (with no hidden confounders) is $\sigma_e^2/(n\sigma_j^2)$, which we use as a benchmark. The corresponding discussion on efficiency of the standard high-dimensional regression can be found in Section 2.3.3 of [53]. The asymptotic expression for the variance V of our proposed estimator (10) is given by $\frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]}$ (see Theorem 1). For the Trim transform defined in (14), which trims top $(100\rho_j)\%$ of the singular values, we have that

$$\frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} = \frac{\sigma_e^2}{\sigma_j^2} \cdot \frac{\sum_{j=1}^m S_{l,l}^4}{(\sum_{l=1}^m S_{l,l}^2)^2},$$

where we write $m = \min\{n, p-1\}$ and $S_{l,l} = S_{l,l}(X_{-j}) \in [0,1]$. Since $S_{l,l}^4 \le S_{l,l}^2$, $\sum_{l=1}^m S_{l,l}^2 \ge (1-\rho_j)m$ and $(\sum_{l=1}^m S_{l,l}^2)^2 \le m \cdot \sum_{j=1}^m S_{l,l}^4$, we obtain

$$\frac{\sigma_e^2}{\sigma_j^2 m} \le \frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} \le \frac{1}{1-\rho_j} \cdot \frac{\sigma_e^2}{\sigma_j^2 m}.$$

In the high-dimensional setting where $p - 1 \ge n$, we have $m = n$ and then

$$\frac{\sigma_e^2}{\sigma_j^2 n} \le \frac{\sigma_e^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]} \le \frac{1}{1-\rho_j} \cdot \frac{\sigma_e^2}{\sigma_j^2 n}. \tag{24}$$

Theorem 2. *Suppose that the assumptions of Theorem 1 hold. If* $p \geq n+1$ *and* $\rho_j = \rho_j(n) \to 0$*, then the Doubly Debiased Lasso estimator in* (10) *has asymptotic variance* $\frac{\sigma_e^2}{\sigma_j^2 n}$*, that is, it achieves the Gauss-Markov efficiency bound.*

The above theorem shows that in the regime $q \ll n$, the Doubly Debiased Lasso achieves the Gauss-Markov efficiency bound if $\rho_j = \rho_j(n) \to 0$ and $\min\{\rho, \rho_j\} \geq (3q+1)/n$ (which is a condition in Theorem 1). When using the median trim (that is, $\rho_j = 1/2$), the bound in (24) implies that the variance of this Doubly Debiased estimator is at most twice the size of the Gauss-Markov bound. In Section 5, we illustrate the finite-sample performance of the Doubly Debiased Lasso estimator for different values of ρ_j; see Figure 7.

In general for the high-dimensional setting $p/n \to c^* \in (0, \infty]$, the Asymptotic Relative Efficiency (ARE) of the proposed Doubly Debiased Lasso estimator with respect to the Gauss-Markov efficiency bound satisfies the following:

$$\text{ARE} \in \left[\frac{1}{\min\{c^*, 1\}}, \frac{1}{(1-\rho^*)\min\{c^*, 1\}} \right], \tag{25}$$

where $\rho^* = \lim_{n\to\infty} \rho_j(n) \in [0, 1)$. The equation (25) reveals how the efficiency of the Doubly Debiased Lasso is affected by the choice of the percentile $\rho_j = \rho_j(n)$ in transformation $\mathcal{P}^{(j)}$ and the dimensionality of the problem. Smaller ρ_j leads to a more efficient estimator, as long as the top few singular values are properly shrunk. Intuitively, a smaller percentile ρ_j means that less information in X_{-j} is trimmed out and hence the proposed estimator is more efficient. In addition, for the case $\rho^* = 0$, we have $\text{ARE} = \max\{1/c^*, 1\}$. With $\rho^* = 0$, a plot of ARE with respect to the ratio $c^* = \lim p/n$ is given in Figure 1. We see that for $c^* < 1$ (that is $p < n$), the relative efficiency of

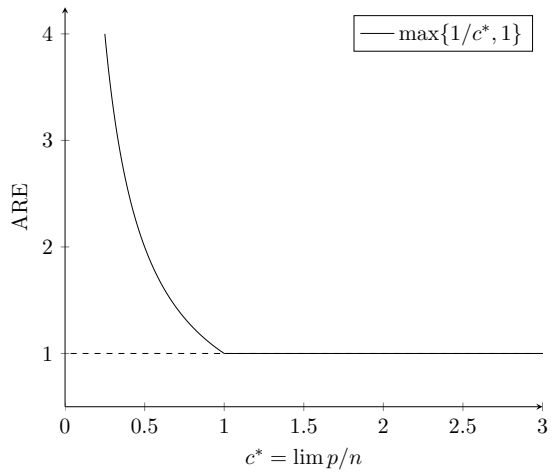

Figure 1: The plot of ARE versus $c^* = \lim p/n$, for the setting of $\rho^* = 0$.

the proposed estimator increases as the dimension p increases and when $c^* \geq 1$ (that

is $p \geq n$), we have that ARE $= 1$, saying that the Doubly Debiased Lasso achieves the efficiency bound in the Gauss-Markov sense.

The phenomenon that the efficiency is retained even in presence of hidden confounding is quite remarkable. For comparison, even in the classical low-dimensional setting, the most commonly used approach assumes availability of sufficiently many instrumental variables (IV) satisfying certain stringent conditions under which one can consistently estimate the effects in presence of hidden confounding. In Theorem 5.2 of [59], the popular IV estimator, two-stage-least-squares (2SLS), is shown to have variance strictly larger than the efficiency bound in the Gauss-Markov setting (with no unmeasured confounding). It has been also shown in Theorem 5.3 of [59] that the 2SLS estimator is efficient in the class of all linear instrumental variable estimators and thus, all linear instrumental variable estimators are strictly less efficient than our Doubly Debiased Lasso. On the other hand, our proposed method does not only avoid the difficult step of constructing a large number of valid instrumental variables, but also achieves the efficiency bound with a careful construction of the spectral transformation $\mathcal{P}^{(j)}$. This occurs due to a blessing of dimensionality and the assumption of dense confounding, where a large number of covariates are assumed to be affected by a small number of hidden confounders.

4.2.3 Asymptotic validity of confidence intervals

The asymptotic normal limiting distribution in Theorem 1 can be used for construction of confidence intervals for β_j. Consistently estimating the variance V of our estimator, defined in (23), requires a consistent estimator of the error variance σ_e^2. The following proposition establishes the rate of convergence of the estimator $\widehat{\sigma}_e^2$ proposed in (17):

Proposition 1. *Consider the Hidden Confounding Model* (2). *Suppose that conditions* (A1)-(A4) *hold. Suppose further that* $c^* = \lim p/n \in (0, \infty]$, $k \lesssim n/\log p$ *and* $q \ll n$. *Then with probability larger than* $1 - \exp(-ct^2) - \frac{1}{t^2} - c(\log p)^{-1/2} - n^{-c}$ *for some positive constant* $c > 0$ *and for any* $t > 0$, *we have*

$$\left|\widehat{\sigma}_e^2 - \sigma_e^2\right| \lesssim \frac{t}{\sqrt{n}}\left(1 + \sqrt{\frac{q \log p}{n}}\right) + k\frac{\log p}{\min\{n, \sqrt{np/q}\}} + \frac{q \log p}{\min\{n, p\}}.$$

As a remark, the estimation error $\left|\widehat{\sigma}_e^2 - \sigma_\epsilon^2\right|$ is of the same order of magnitude as $\left|\widehat{\sigma}_e^2 - \sigma_e^2\right|$ since the difference $\sigma_\epsilon^2 - \sigma_e^2$ is small in the dense confounding model, see Lemma 2 in the supplement. The above result, together with Theorem 1, implies the asymptotic coverage and precision properties of the proposed confidence interval $\mathrm{CI}(\beta_j)$, described in (13):

Corollary 1. *Suppose that the conditions of Theorem 1 hold, then the confidence interval defined in* (13) *satisfies the following properties:*

$$\liminf_{n,p\to\infty} \mathbb{P}\left(\beta_j \in \mathrm{CI}(\beta_j)\right) \geq 1 - \alpha, \tag{26}$$

$$\limsup_{n,p\to\infty} \mathbb{P}\left(\mathbf{L}\left(\mathrm{CI}(\beta_j)\right) \geq (2+c) z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_\varepsilon^2 \mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \mathrm{Tr}^2[(\mathcal{P}^{(j)})^2]}}\right) = 0, \tag{27}$$

for any positive constant $c > 0$, where $\mathbf{L}\left(\mathrm{CI}(\beta_j)\right)$ denotes the length of the proposed confidence interval.

Similarly to the efficiency results in Section 4.2.2, the exact length depends on the construction of the spectral transformation $\mathcal{P}^{(j)}$. Together with (24), the above proposition shows that the length of constructed confidence interval is shrinking at the rate of $n^{-1/2}$ for the Trim transform in the high-dimensional setting. Specifically, for the setting $p \geq n+1$, if we choose $\rho_j = \rho_j(n) \geq (3q+1)/n$ and $\rho_j(n) \to 0$, the constructed CI has asymptotically optimal length.

5 Empirical results

In this section we consider the practical aspects of Doubly Debiased Lasso methodology and illustrate its empirical performance on both real and simulated data. The overview of the method and the tuning parameters selection can be found in Section 3.6.

In order to investigate whether the given data set is potentially confounded, one can inspect the principal components of the design matrix X, or equivalently consider its SVD. Spiked singular value structure (see Figure 2) indicates the existence of hidden confounding, as much of the variance of our data can be explained by a small number of latent factors. This also serves as an informal check of the spiked singular value condition in the assumption (A2).

The scree plot can also be used for choosing the trimming thresholds, if one wants to depart from the default median rule (see Section 3.6). We have seen from the theoretical considerations in Section 4 that we can reduce the estimator variance by decreasing the trimming thresholds for the spectral transformations $\mathcal{P}^{(j)}$ and $\mathcal{Q}$. On the other hand, it is crucial to choose them so that the number of shrunk singular values is still sufficiently large compared to the number of confounders. However, exactly estimating the number of confounders, e.g. by detecting the elbow in the screen plot [56], is not necessary with our method, since the efficiency of our estimator decreases relatively slowly as we decrease the trimming threshold.

In what follows, we illustrate the empirical performance of the Doubly Debiased Lasso in practice. We compare the performance with the standard Debiased Lasso [61], even though it is not really a competitor for dealing with hidden confounding. Our goal is to illustrate and quantify the error and bias when using the naive and popular approach which ignores potential hidden confounding. We first investigate the performance of our method on simulated data for a range of data generating mechanisms and then investigate its behaviour on a gene expression dataset from the GTEx project [39].

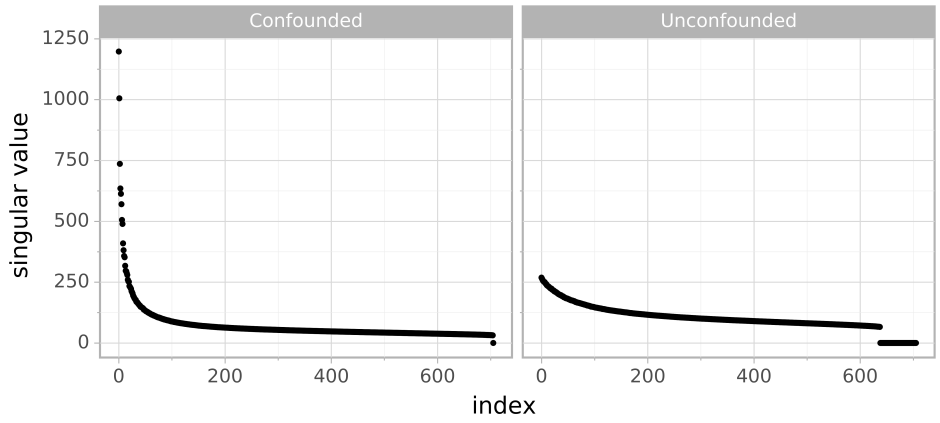

Figure 2: Left: Spiked singular values of the standardized gene expression matrix (see Section 5.2) indicate possible confounding. Right: Singular values after regressing out the $q = 65$ confounding proxies given in the dataset (thus labeled as "unconfounded"). The singular values in both plots are sorted decreasingly.

5.1 Simulations

In this section, we compare the Doubly Debiased Lasso with the standard Debiased Lasso in several different simulation settings for estimation of β_j and construction of the corresponding confidence intervals.

In order to make comparisons with the standard Debiased Lasso as fair as possible, we use the same procedure for constructing the standard Debiased Lasso, but with $\mathcal{Q} = \mathrm{I}_p$, $\mathcal{P}^{(j)} = \mathrm{I}_{p-1}$, whereas for the Doubly Debiased Lasso, $\mathcal{P}^{(j)}$, $\mathcal{Q}$ are taken to be median Trim transform matrices, unless specified otherwise. Finally, to investigate the usefulness of double debiasing, we additionally include the standard Debiased Lasso estimator with the same initial estimator $\widehat{\beta}^{init}$ as our proposed method, see Section 3.4. Therefore, this corresponds to the case where $\mathcal{Q}$ is the median Trim transform, whereas $\mathcal{P}^{(j)} = \mathrm{I}_{p-1}$.

We will compare the (scaled) bias and variance of the corresponding estimators. For a fixed index j, from the equation (11) we have

$$V^{-1/2}(\widehat{\beta}_j - \beta_j) = N(0,1) + B_\beta + B_b,$$

where the estimator variance V is defined in (23) and the bias terms B_β and B_b are given by

$$B_\beta = V^{-1/2}\frac{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}(\widehat{\beta}_{-j}^{init} - \beta_{-j})}{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j}, \qquad B_b = V^{-1/2}\frac{Z_j^\intercal(\mathcal{P}^{(j)})^2 Xb}{Z_j^\intercal(\mathcal{P}^{(j)})^2 X_j}.$$

Larger estimator variance makes the confidence intervals wider. However, large bias makes the confidence intervals inaccurate. We quantify this with the scaled bias terms B_β, which is due to the error in estimation of β, and B_b, which is due to the perturbation

b arising from the hidden confounding. Having small $|B_\beta|$ and $|B_b|$ is essential for having a correct coverage, since the construction of confidence intervals is based on the approximation $V^{-1/2}(\widehat{\beta}_j - \beta_j) \approx N(0,1)$. We investigate the validity of the confidence interval construction by measuring the coverage of the nominal 95% confidence interval.

Simulation parameters In all of the following simulations we fix $q = 3$, $s = 5$ and $\beta = (1,1,1,1,1,0,\dots 0)^\intercal$ and we target the coefficient $\beta_1 = 1$. The rows of the unconfounded design matrix E are generated from $N(0, \Sigma_E)$ distribution, where $\Sigma_E = \mathrm{I}_p$, as a default. The matrix of confounding variables H, the additive error e and the coefficient matrices Ψ and ϕ all have i.i.d. $N(0,1)$ entries, unless stated otherwise. Each simulation is averaged over $5,000$ independent repetitions.

Varying dimensions n and p In this simulation setting we investigate how the performance of our estimator depends on the dimensionality of the problem. The results can be seen in Figure 3. In the first scenario, shown in the top row, we have $n = 500$ and p varying from 50 to $2,000$, thus covering both low-dimensional and high-dimensional cases. In the second scenario, shown in the bottom row, the number of covariates is fixed at $p = 500$ and the sample size n varies from 100 to $2,000$.

We see that the absolute bias term $|B_b|$ due to confounding is substantially smaller for Doubly Debiased Lasso compared to the standard Debiased Lasso, regardless of which initial estimator is used. This is because $\mathcal{P}^{(j)}$ additionally removes bias by shrinking large principal components of X_{-j}. This spectral transformation helps also to make the absolute bias term $|B_\beta|$ smaller for the Doubly Debiased Lasso compared to the Debiased Lasso, even when using the same initial estimator $\widehat{\beta}^{init}$. This comes however at the expense of slightly larger variance, but we can see that the decrease in bias reflects positively on the validity of the constructed confidence intervals. Their coverage is significantly more accurate for Doubly Debiased Lasso, over a large range of n and p.

There are two challenging regimes for estimation under confounding. Firstly, when the dimension p is much larger than the sample size n, the coverage can be lower than 95%, since in this regime it is difficult to estimate β accurately and thus the term $|B_\beta|$ is fairly large, even after the bias correction step. We see that the absolute bias $|B_\beta|$ grows with p, but it is much smaller for the Doubly Debiased Lasso which positively impacts the coverage. Secondly, in the regime where p is relatively small compared to n, $|B_b|$ begins to dominate and leads to undercoverage of confidence intervals. B_b is caused by the hidden confounding and does not disappear when $n \to \infty$, while keeping p constant. The simulation results agree with the asymptotic analysis of the bias term in (59) in the Supplementary material, where the term $|B_b|$ decreases when increasing $m = \min\{n, p\}$ and not only the sample size n. In this regime $|B_b|$ will even grow since the bias becomes increasingly large compared to the estimator's variance. However, it is important to note that even in these difficult regimes, Doubly Debiased Lasso performs significantly better than the standard Debiased Lasso (irrespective of the initial estimator) as it manages to additionally decrease the estimator's bias.

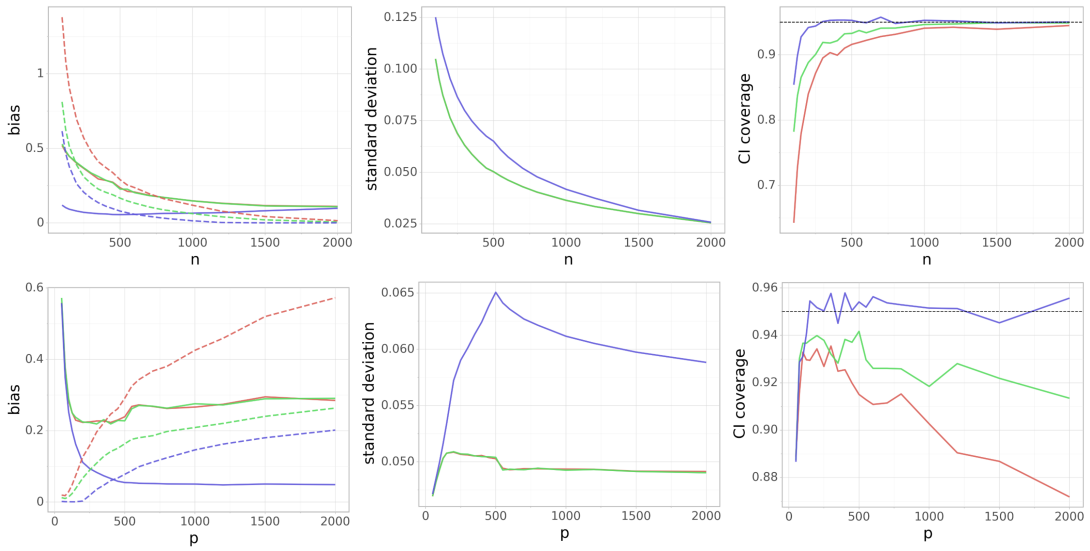

Figure 3: *(Varying dimensions)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of data points n (top row) and the number of covariates p (bottom row). On the left side, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. In the top row we fix $p = 500$, whereas in the bottom row we have $n = 500$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

No confounding bias We consider now the same simulation setting as in the previous case, where we fix $n = 500$ and vary p, but where in addition we remove the effect of the perturbation b that arises due to the confounding. We generate from the model (2), but then adjust for the confounding bias: $Y \leftarrow (Y - Xb)$, where b is the induced coefficient perturbation, as in Equation (3). In this way we still have a perturbed linear model, but where we have enforced $b = 0$ while keeping the same spiked covariance structure of X: $\Sigma_X = \Sigma_E + \Psi^\intercal \Psi$ as in (2). The results can be seen in Figure 4. We see that Doubly Debiased Lasso still has smaller absolute bias $|B_\beta|$, slightly higher variance and better coverage than the standard Debiased Lasso, even in absence of confounding. The bias term $B_b = 0$, since we have put $b = 0$. This shows that our method can provide us certain robustness against dense confounding: if there is such confounding, our proposed method is able to significantly reduce the bias caused by it; on the other hand, if there is no confounding, in comparison to the standard Debiased Lasso, our proposed method still has essentially as good performance, with a small increase in variance. In the above setting, there is even a decrease in estimation bias, most likely due to the fact that X has a spiked covariance structure.

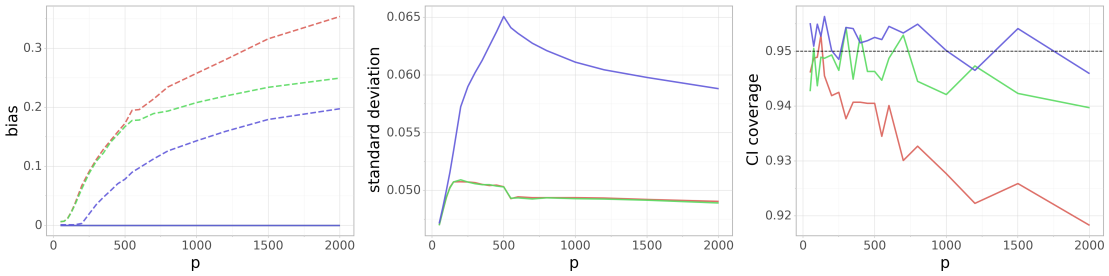

Figure 4: *(No confounding bias)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of covariates p, while keeping $n = 500$ fixed. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively, but $B_b = 0$ since we have enforced $b = 0$. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable B_b and V.

Toeplitz covariance structure Now we fix $n = 300$ and $p = 1,000$, but we generate the covariance matrix Σ_E of the unconfounded part of the design matrix X to have Toeplitz covariance structure: $(\Sigma_E)_{i,j} = \kappa^{|i-j|}$, where we vary κ across the interval $[0, 0.97]$. As we increase κ, the covariates $X_1, \ldots, X_5$ in the active set get more correlated, so it gets harder to distinguish their effects on the response and therefore to estimate β. Similarly, it gets as well harder to estimate γ in the regression of X_j on X_{-j}, since X_j can be explained well by many linear combinations of the other covariates that are correlated with X_j. In Figure 5 we can see that Doubly Debiased Lasso is much less affected by

correlated covariates. The (scaled) absolute bias terms $|B_b|$ and $|B_\beta|$ are much larger for standard Debiased Lasso, which causes the coverage to worsen significantly for values of κ that are closer to 1.

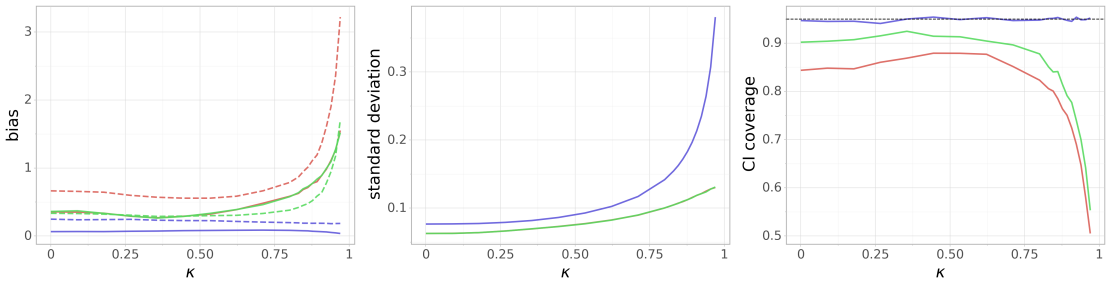

Figure 5: *(Toeplitz covariance)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the parameter κ of the Toeplitz covariance structure. $n = 300$ and $p = 1,000$ are fixed. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

Proportion of confounded covariates Now we again fix $n = 300$ and $p = 1,000$, but we change the proportion of covariates X_i that are affected by each confounding variable. We do this by setting to zero a desired proportion of entries in each row of the matrix $\Psi \in \mathbb{R}^{q \times p}$, which describes the effect of the confounding variables on each predictor. Its non-zero entries are still generated as $N(0, 1)$. We set once again $\Sigma_E = \mathrm{I}_p$ and we vary the proportion of nonzero entries from 5% to 100%. The results can be seen in Figure 6. We can see that Doubly Debiased Lasso performs well even when only a very small number (5%) of the covariates are affected by the confounding variables, which agrees with our theoretical discussion for assumption **(A2)**. We can also see that the coverage of the Debiased Lasso is poor even for a small number of affected variables and it worsens as the confounding variables affect more and more covariates. The coverage improves to some extent when we use a better initial estimator, but is still worse than our proposed method.

Trimming level We now investigate the dependence of the performance on the choice of the trimming threshold for the Trim transform (14), parametrized by the proportion of singular values ρ_j which we shrink. The spectral transformation $\mathcal{Q}$ used for the initial estimator $\widehat{\beta}^{init}$ is fixed to be the default choice of Trim transform with median rule. We fix $n = 300$ and $p = 1,000$ and consider the same setup as in Figure 3. We take $\tau = \Lambda_{\lfloor \rho_j m \rfloor, \lfloor \rho_j m \rfloor}$ to be the ρ_j-quantile of the set of singular values of the design matrix X, where we vary ρ_j across the interval $[0, 0.9]$. When $\rho_j = 0$, τ is the maximal singular

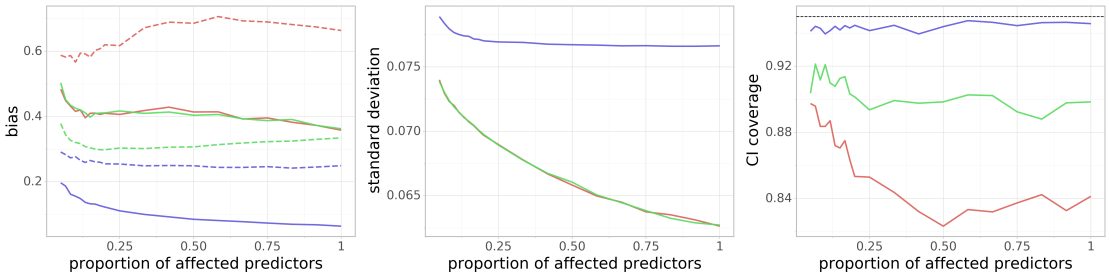

Figure 6: *(Proportion confounded)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on proportion of confounded covariates. $n = 300$ and $p = 1,000$ are fixed. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

value, so there is no shrinkage and our estimator reduces to the standard Debiased Lasso (with the initial estimator $\widehat{\beta}^{init}$). The results are displayed in Figure 7. We can see that Doubly Debiased Lasso is quite insensitive to the trimming level, as long as the number of shrunken singular values is large enough compared to the number of confounding variables q. In the simulation $q = 3$ and the (scaled) absolute bias terms $|B_b|$ and $|B_\beta|$ are still small when $\rho_j \approx 0.02$, corresponding to shrinking 6 largest singular values. We see that the standard deviation decreases as ρ_j decreases, i.e. as the trimming level τ increases, which matches our efficiency analysis in Section 4.2.1. However, we see that the default choice $\tau = \Lambda_{\lfloor m/2 \rfloor, \lfloor m/2 \rfloor}$ has decent performance as well.

Measurement error We now generate from the measurement error model (4), which can be viewed as a special case of our model (2). The measurement error $W = \Psi^\intercal H$ is generated by $q = 3$ latent variables $H_{i,\cdot} \in \mathbb{R}^q$ for $1 \leq i \leq n$. We fix the number of data points to be $n = 500$ and vary the number of covariates p from 50 to 1,000, as in Figure 3. The results are displayed in Figure 8, where we can see a similar pattern as before: Doubly Debiased Lasso decreases the bias at the expense of a slightly inflated variance, which in turn makes the inference much more accurate and the confidence intervals have significantly better coverage.

5.2 Real data

We investigate here the performance of Doubly Debiased Lasso on a genomic dataset. The data are obtained from the GTEx project [39], where the gene expression has been measured postmortem on samples coming from various tissue types. For our purposes, we use fully processed and normalized gene expression data for the skeletal muscle tissue.

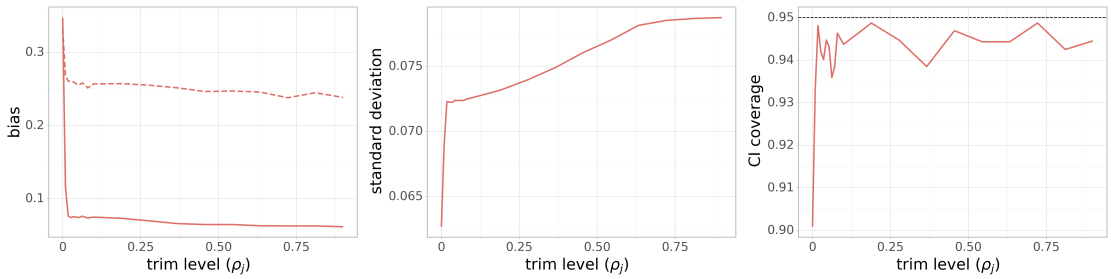

Figure 7: *(Trimming level)* Dependence of the (scaled) absolyte bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the trimming level ρ_j of the Trim transform (see Equation (14)). The sample size is fixed at $n = 300$ and the dimension at $p = 1,000$. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. The case $\rho_j = 0$ corresponds to Debiased Lasso with the spectral deconfounding initial estimator $\widehat{\beta}^{init}$, described in (16).

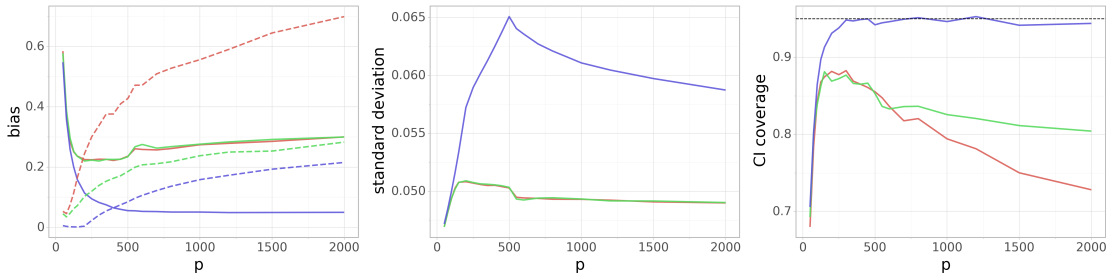

Figure 8: *(Measurement error)* Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of covariates p in the measurement error model (4). The sample size is fixed at $n = 500$. On the leftmost plot, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively. Blue color corresponds to the Doubly Debiased Lasso, red color represents the standard Debiased Lasso and green color corresponds also to the Debiased Lasso estimator, but with the same $\widehat{\beta}^{init}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.

The gene expression matrix X consists of measurements of expressions of $p = 12,646$ protein-coding genes for $n = 706$ individuals. Genomic datasets are particularly prone to confounding [37, 20, 22], and for our analysis we are provided with $q = 65$ proxies for hidden confounding, computed with genotyping principal components and PEER factors.

We investigate the associations between the expressions of different genes by regressing one target gene expression X_i on the expression of other genes X_{-i}. Since the expression of many genes is very correlated, researchers often use just $\sim 1,000$ carefully chosen landmark genes as representatives of the whole gene expression [51]. We will use several such landmark genes as the responses in our analysis.

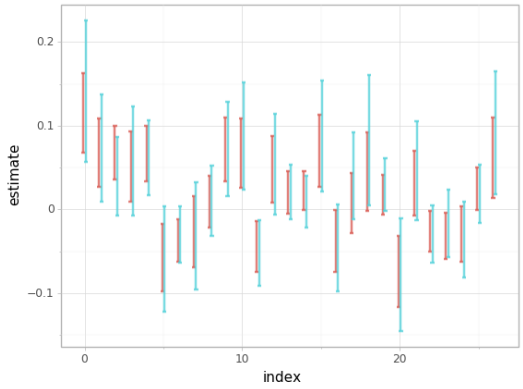

Figure 9: Comparison of 95% confidence intervals obtained by Doubly Debiased Lasso (blue) and Doubly Debiased Lasso (red) for regression of the expression of one target landmark gene on the other gene expressions.

In Figure 9 we can see a comparison of 95%-confidence intervals that are obtained from Doubly Debiased Lasso and standard Debiased Lasso. For a fixed response landmark gene X_i, we choose 25 predictor genes X_j where $j \neq i$ such that their corresponding coefficients of the Lasso estimator for regressing X_i on X_{-i} are non-zero. The covariates are ordered according to decreasing absolute values of their estimated Lasso coefficients. We notice that the confidence intervals follow a similar pattern, but that the Doubly Debiased Lasso, besides removing bias due to confounding, is more conservative as the resulting confidence intervals are wider.

This behavior becomes even more apparent in Figure 10, where we compare all p-values for a fixed response landmark gene. We see that Doubly Debiased Lasso is more conservative and it declares significantly less covariates significant than the standard Debiased Lasso. Even though the p-values of the two methods are correlated (see also Figure 12), we see that it can happen that one method declares a predictor significant, whereas the other does not.

Robustness against hidden confounding We now adjust the data matrix X by regressing out the $q = 65$ provided hidden confounding proxies. By regressing out these covariates, we obtain an estimate of the unconfounded gene expression matrix $\tilde{X}$. We

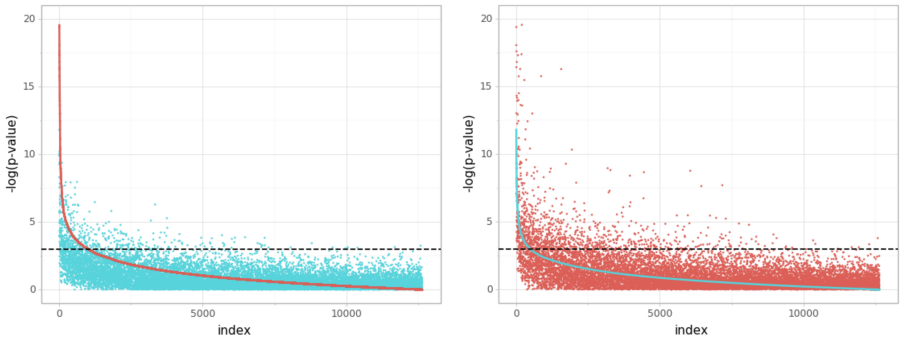

Figure 10: Comparison of p-values for two-sided test of the hypothesis $\beta_j = 0$, obtained by Doubly Debiased Lasso (red) and Doubly Debiased Lasso (blue) for regression of the expression of one target gene on the other gene expressions. The covariates are ordered by decreasing significance, either estimated by the Debiased Lasso (left) or by the Doubly Debiased Lasso (right). Black dotted line indicates the 5% significance level.

compare the estimates for the original gene expression matrix with the estimates obtained from the adjusted matrix.

For a fixed response landmark gene expression X_i, we can determine significance of the predictor genes by considering the p-values. One can perform variable screening by considering the set of most significant genes. For Doubly Debiased Lasso and the standard Lasso we compare the sets of most significant variables determined from the gene expression matrix X and the deconfounded matrix $\tilde{X}$. The difference of the chosen sets is measured by the Jaccard distance. A larger Jaccard distance indicates a larger difference between the chosen sets. The results can be seen in Figure 11. The results are averaged over 10 different response landmark genes. We see that the Doubly Debiased Lasso gives more similar sets for the large model size, indicating that the analysis conclusions obtained by using Doubly Debiased Lasso are more robust in presence of confounding variables. However, for small model size we do not see large gains. In this case the sets produced by any method are quite different, i.e. the Jaccard distance is very large. This indicates that the problem of determining the most significant covariates is quite difficult, since X and $\tilde{X}$ differ a lot.

In Figure 12 we can see the relationship between the p-values obtained by Doubly Debiased Lasso and the standard Debiased Lasso for the original gene expression matrix X and the deconfounded matrix $\tilde{X}$. The p-values are aggregated over 10 response landmark genes and are computed for all possible predictor genes. We can see from the left plot that the Doubly Debiased Lasso is much more conservative for the confounded data. The cloud of points is skewed upwards showing that the standard Debiased Lasso declares many more covariates significant in presence of the hidden confounding. On the other hand, in the right plot we can see that the p-values obtained by the two methods are much more similar for the unconfounded data and the point cloud is significantly less skewed upwards. The remaining deviation from the $y = x$ line might be due to the

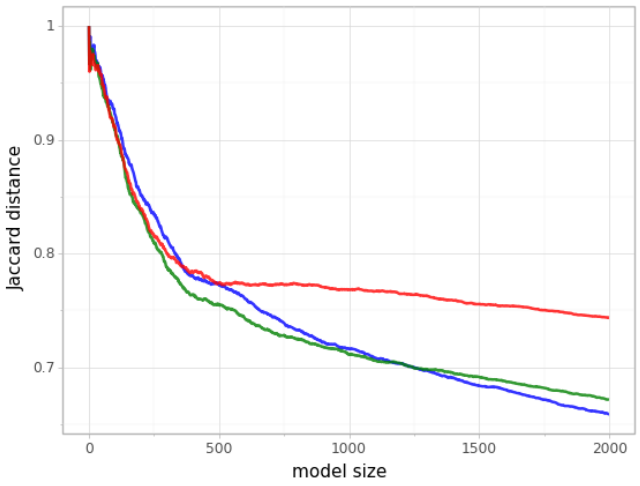

Figure 11: Comparison of the sets of the most significant covariates chosen based on the original expression matrix X and the deconfounded gene expression matrix $\tilde{X}$, for different cardinalities of the sets (model size). The set differences are measured by Jaccard distance. Red line represents the standard Debiased Lasso method, whereas the blue and green lines denote the Doubly Debiased Lasso that uses $\rho = 0.5$ and $\rho = 0.1$ for obtaining the trimming threshold, respectively; see Equation (14).

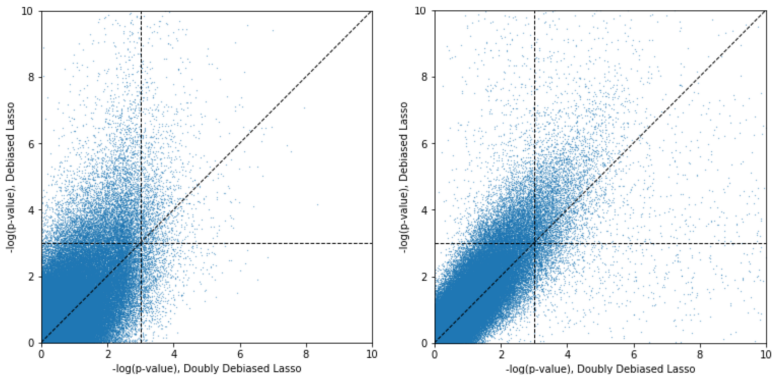

Figure 12: Comparison of p-values for two-sided test of the hypothesis $\beta_j = 0$, obtained by Doubly Debiased Lasso and standard Debiased Lasso for regression of the expression of one target gene on the other gene expressions. The points are aggregated over 10 landmark response genes. The p-values are either determined using the original gene expression matrix (left) or the matrix where we have regressed out the given $q = 65$ confounding proxies (right). Horizontal and vertical black dashed lines indicate the 5% significance level.

remaining confounding, not accounted for by regressing out the given confounder proxies.

6 Discussion

We propose the Doubly Debiased Lasso estimator for hypothesis testing and confidence interval construction for single regression coefficients in high-dimensional settings with "dense" confounding. We present theoretical and empirical justifications and argue that our double debiasing leads to robustness against hidden confounding. In case of no confounding, the price to be paid is (typically) small, with a small increase in variance but even a decrease in estimation bias, in comparison to the standard Debiased Lasso [61]; but there can be substantial gain when "dense" confounding is present.

It is ambitious to claim significance based on observational data. One always needs to make additional assumptions to guard against confounding. We believe that our robust Doubly Debiased Lasso is a clear improvement over the use of standard inferential high-dimensional techniques, yet it is simple and easy to implement, requiring two additional SVDs only, with no additional tuning parameters when using our default choice of trimming $\rho = \rho_j = 50\%$ of the singular values in Equations (14) and (15).

Acknowledgements

We thank Yuansi Chen for providing the script to preprocess the raw data from the GTEx project.

References

[1] Alexandre Belloni, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen. Program evaluation and causal inference with high-dimensional data. *Econometrica*, 85(1):233–298, 2017.

[2] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

[3] Peter J Bickel, Yaacov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[4] Anna GC Boef, Olaf M Dekkers, Jan P Vandenbroucke, and Saskia le Cessie. Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. *Journal of clinical Epidemiology*, 67(11):1258–1264, 2014.

[5] Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[6] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research*, 26(5):2333–2355, 2017.

[7] T Tony Cai and Zijian Guo. Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *The Annals of Statistics*, 45(2):615–646, 2017.

[8] Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[9] Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[10] Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding and perturbed sparse linear models. *arXiv preprint arXiv:1811.05352*, 2018.

[11] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[12] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[13] Victor Chernozhukov, Christian Hansen, and Martin Spindler. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688, 2015.

[14] Ruben Dezeure, Peter Bühlmann, and Cun-Hui Zhang. High-dimensional simultaneous inference with the bootstrap. *Test*, 26(4):685–719, 2017.

[15] Jianqing Fan, Yingying Fan, and Jinchi Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

[16] Jianqing Fan and Yuan Liao. Endogeneity in high dimensions. *The Annals of Statistics*, 42(3):872–917, 2014.

[17] Jianqing Fan, Yuan Liao, and Martina Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 75(4):603–680, 2013.

[18] Jianqing Fan, Yuan Liao, Weichen Wang, et al. Projected principal component analysis in factor models. *The Annals of Statistics*, 44(1):219–254, 2016.

[19] Max H Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

[20] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[21] Eric Gautier and Christiern Rose. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

[22] David Gerard and Matthew Stephens. Empirical bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics*, 21(1):15–32, 2020.

[23] David Gold, Johannes Lederer, and Jing Tao. Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217(1):79–111, 2020.

[24] Friedrich Götze and A Tikhomirov. Asymptotic distribution of quadratic forms and applications. *Journal of Theoretical Probability*, 15(2):423–475, 2002.

[25] Jason R Guertin, Elham Rahme, and Jacques LeLorier. Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *European journal of Clinical Pharmacology*, 72(12):1497–1505, 2016.

[26] Zijian Guo, Hyunseung Kang, T Tony Cai, and Dylan S Small. Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):793–815, 2018.

[27] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

[28] Chirok Han. Detecting invalid instruments using l1-gmm. *Economics Letters*, 101(3):285–287, 2008.

[29] Jana Jankova and Sara van de Geer. Semiparametric efficiency bounds for high-dimensional models. *The Annals of Statistics*, 46(5):2336–2359, 2018.

[30] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[31] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[32] Hyunseung Kang, Anru Zhang, T Tony Cai, and Dylan S Small. Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144, 2016.

[33] Clifford Lam, Jianqing Fan, et al. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278, 2009.

[34] Clifford Lam and Qiwei Yao. Factor modeling for high-dimensional time series: inference for the number of factors. *The Annals of Statistics*, 40(2):694–726, 2012.

[35] Clifford Lam, Qiwei Yao, and Neil Bathia. Estimation of latent factors for high-dimensional time series. *Biometrika*, 98(4):901–918, 2011.

[36] Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010.

[37] Jeffrey T Leek, John D Storey, et al. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genetics*, 3(9):1–12, 2007.

[38] Wei Lin, Rui Feng, and Hongzhe Li. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *Journal of the American Statistical Association*, 110(509):270–288, 2015.

[39] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, 2013.

[40] Kabir Manghnani, Adam Drake, Nathan Wan, and Imran Haque. Metcc: Metric learning for confounder control making distance matter in high dimensional biological analysis. *arXiv preprint arXiv:1812.03188*, 2018.

[41] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[42] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[43] Matey Neykov, Yang Ning, Jun S Liu, and Han Liu. A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statistical Science*, 33(3):427–443, 2018.

[44] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, and Matthew R Nelson. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[45] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[46] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

[47] Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

[48] Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

[49] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[50] Rajen D Shah, Benjamin Frot, Gian-Andrea Thanei, and Nicolai Meinshausen. Right singular vector projection graphs: fast high-dimensional covariance matrix estimation under latent confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82:361–389, 2020.

[51] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, and Jacob K Asiedu. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[52] Yunting Sun, Nancy R Zhang, Art B Owen, et al. Multiple hypothesis testing adjusted for latent variables, with an application to the agemap gene expression data. *The Annals of Applied Statistics*, 6(4):1664–1688, 2012.

[53] Sara van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[54] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina Eldar and Gitta Kutyniok, editors, *Compressed sensing: theory and applications*, pages 210–268. Cambridge University Press, 2012.

[55] Jingshu Wang, Qingyuan Zhao, Trevor Hastie, and Art B Owen. Confounder adjustment in multiple hypothesis testing. *Annals of statistics*, 45(5):1863–1894, 2017.

[56] Weichen Wang, Jianqing Fan, et al. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *The Annals of Statistics*, 45(3):1342–1374, 2017.

[57] Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528):1574–1596, 2019.

[58] Frank Windmeijer, Helmut Farbmacher, Neil Davies, and George Davey Smith. On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association*, 114(527):1339–1350, 2019.

[59] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

[60] Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286, 2010.

[61] Cun-Hui Zhang and Stephanie S Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[62] Ying Zhu. Sparse linear models and l1-regularized 2sls with high-dimensional endogenous regressors and instruments. *Journal of Econometrics*, 202(2):196–213, 2018.

A Intermediary Results and Proof of Theorem 1

In the following, we list three intermediary results in Sections A.1 to A.3 as the key components of proving our main result Theorem 1 and then provide the proof of Theorem 1 in Section A.4. We verify conditions (A2) and (A4) in Sections A.5 and A.6, respectively. The additional proofs are presented in Appendix B. All our theoretical derivations are done for the Hidden Confounding Model (2), but they additionally hold more generally for the perturbed linear model (3).

A.1 Valid spectral transformations

The first intermediary result is on the properties of the spectral transformation we use. We will show that the limiting distribution in Theorem 1 holds generally for the estimator (10) using any spectral transformations $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ that satisfy the following:

(P1) **Spectral Transformation Property.** $\mathcal{P}^{(j)} = U(X_{-j})S(X_{-j})U(X_{-j})^\intercal$ and $\mathcal{Q} = U(X)S(X)U(X)^\intercal$ satisfy

$$\frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}\|_2^2 \lesssim \max\left\{1, \frac{p}{n}\right\} \quad \text{and} \quad \frac{1}{n}\|\mathcal{Q}X\|_2^2 \lesssim \max\left\{1, \frac{p}{n}\right\} \tag{28}$$

$$\mathrm{Tr}[(\mathcal{P}^{(j)})^4] = \sum_{l=1}^{n}[S_{l,l}(X_{-j})]^4 \gtrsim m \quad \text{and} \quad \mathrm{Tr}(\mathcal{Q}^4) = \sum_{l=1}^{n}[S_{l,l}(X)]^4 \gtrsim m. \tag{29}$$

with $m = \min\{n, p-1\}$.

The first requirement means that $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ need to shrink the leading singular values of X_{-j} and X to a sufficiently small level, respectively. On the other hand, the second requirement says that the overall shrinkage of all singular values together is not too big.

For the proof of Theorem 1 and its intermediate results, we extensively use that our spectral transformations satisfy the property (P1). Therefore, we first need to show that the Trim transform $\mathcal{P}^{(j)}$ defined in (14) and $\mathcal{Q}$ defined in (15) satisfy the property (P1). Since $S_{l,l} = 1$ for $l > \lfloor \rho m \rfloor$, we have that at least $\lfloor (1-\rho)m \rfloor$ diagonal elements of S are equal to 1, which immediately gives us (29) for $\mathcal{Q}$ whenever $\rho < 1$. Similarly, (29) for $\mathcal{P}^{(j)}$ holds for any $\rho_j \in (0,1)$. However, in order to show the condition (28), we need to better understand the behaviour of the singular values of the random matrix X.

Proposition 2. *Suppose $E_{i,\cdot} \in \mathbb{R}^p$ is a sub-Gaussian random vector and $\lambda_{\max}(\Sigma_E) \leq C_0$, for some constant C_0, then with probability larger than $1 - \exp(-cn)$,*

$$\lambda_{l+3q}\left(\tfrac{1}{n}X^\intercal X\right) \leq \lambda_l\left(\tfrac{1}{n}E^\intercal E\right) \lesssim \max\{1, p/n\}, \quad \textit{for } 1 \leq l \leq n - 3q. \tag{30}$$

for some positive constant $c > 0$.

The above proposition is proved in the Section B.6 by applying the Cauchy interlacing law and it gives us that $\lambda_{1+3q}(X)$ will be smaller than $\sqrt{\max\{n,p\}}$. This now allows us to conclude that the Trim transform satisfies the property (P1):

Corollary 2. *Let $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ be the spectral transformation matrices obtained by applying the Trim transformation* (14) *and* (15)*, respectively. Suppose that the conditions of Proposition 2 hold and that* $\min\{\rho, \rho_j\} \geq (3q+1)/\min\{n, p-1\}$*. Then the Trim transformations $\mathcal{P}^{(j)}$ and $\mathcal{Q}$ satisfy* (P1)*.*

A.2 Approximate sparsity and perturbation size

The essential step of bias correction is to decouple the correlation between the variable of interest $X_{1,j}$ and other covariates $X_{1,-j} \in \mathbb{R}^{p-1}$. In order to get an informative projection direction $\mathcal{P}^{(j)}Z_j$, one needs to estimate the best linear approximation vector $\gamma = [\mathbb{E}(X_{1,-j}X_{1,-j}^\intercal)]^{-1}\mathbb{E}(X_{1,-j}X_{1,j}) \in \mathbb{R}^{p-1}$ well. Recall that the results for the standard Debiased Lasso [53] are based on the fact that the sparsity of the precision matrix Σ_X^{-1} gives sparsity of γ, thus justifying the estimation accuracy of the Lasso regression of $X_{1,j}$ on $X_{1,-j}$. However, even though the assumption (A1) ensures the sparsity of the precision matrix of the unconfounded part E of the design matrix X, γ will not be sparse, since the confounding variables H introduce additional correlations between the covariates.

However, Lemma 1 shows that in presence of confounding variables, the vector γ can be decomposed into a main sparse component γ^M and an additional small perturbation vector γ^A. The proof of the following Lemma is presented in Section B.2.

Lemma 1. *Suppose that the conditions* (A1) *and* (A2) *hold, then the vector* $\gamma = [\mathbb{E}(X_{1,-j}X_{1,-j}^\intercal)]^{-1}\mathbb{E}(X_{1,-j}X_{1,j})$ *defined as the minimizer of* $\mathbb{E}(X_{1,j} - X_{1,-j}^\intercal\gamma')^2$*, can be decomposed as* $\gamma = \gamma^M + \gamma^A$*, where* $\gamma^M = [\mathbb{E}(E_{1,-j}E_{1,-j}^\intercal)]^{-1}\mathbb{E}E_{1,j}E_{1,-j}$ *is a sparse vector with at most s non-zero components and*

$$\|\gamma^A\|_2 \leq \max_{1\leq l\leq q} \frac{C_0|\lambda_l(\Psi_{-j})|}{c_0\lambda_l^2(\Psi_{-j})+1}\|\Psi_j + \Psi_{-j}\gamma^M\|_2 \lesssim \sqrt{\frac{q}{p}}(\log p)^{1/4}. \tag{31}$$

The main component γ^M is fully determined by the covariance structure of $E_{i,\cdot}$. From the block matrix inverse formula, we get that γ^M is proportional to $(\Omega_E)_{j,-j} \in \mathbb{R}^{p-1}$ and therefore sparse with at most s non-zero components. Since the additional component γ^A converges to zero at the rate $\sqrt{q/p}(\log p)^{1/4}$, the regression vector γ is approximately sparse.

In a similar fashion, we will show that the perturbation b in (3), which is induced by the confounding variables, is dense and small as well.

Lemma 2. *Suppose that the conditions* (A1) *and* (A2) *hold, then* $|b_j| \lesssim q\sqrt{\log p}/p$*,* $\|b\|_2 \lesssim \sqrt{q/p}(\log p)^{1/4}$ *and*

$$\left|\sigma_\epsilon^2 - \sigma_e^2\right| = \left|\phi^\intercal\left(\mathrm{I}_q - \Psi\Sigma_X^{-1}\Psi^\intercal\right)\phi\right| \lesssim q\sqrt{\log p}/p. \tag{32}$$

The above lemma also shows that the variance of the error ϵ_i in (3) is close to that of the random error e_i. The proof of the above lemma is presented in Section B.3.

A.3 Error rates of $\widehat{\beta}^{init}$ and $\widehat{\gamma}$

In order to show the asymptotic normality of the proposed Doubly Debiased Lasso estimator (10), we need that the estimators $\widehat{\beta}^{init}$ and $\widehat{\gamma}$, which are used for construction of our estimator $\widehat{\beta}_j$, estimate the target values γ and β well. Theorem 1 can be shown to hold for the estimator (10) using any estimators $\widehat{\beta}^{init}$ and $\widehat{\gamma}$ that satisfy the following condition:

(P2) **Penalized Estimator Properties.** The estimators $\widehat{\beta}^{init}$ and $\widehat{\gamma}$ satisfy

$$\|\widehat{\beta}^{init}-\beta\|_1 \lesssim \frac{1}{\tau_*}k\sqrt{\frac{\log p}{\min\{n,p/q\}}}+\frac{q}{m}\sqrt{\min\left\{n,\frac{p}{q}\right\}}$$

$$\|\widehat{\beta}^{init}-\beta\|_2 \lesssim \frac{1}{\tau_*}\sqrt{\frac{k\log p}{\min\{n,p/q\}}}+\frac{q}{m}\sqrt{\min\left\{n,\frac{p}{q}\right\}}$$

$$\frac{1}{\sqrt{n}}\|\mathcal{Q}X(\widehat{\beta}^{init}-\beta)\|_2 \lesssim \frac{1}{\tau_*}\sqrt{\frac{k\log p}{\min\{n,p/q\}}}+\sqrt{\frac{q\log p}{m}}$$

$$\|\widehat{\gamma}-\gamma^M\|_1 \lesssim \frac{1}{\tau_*}s\sqrt{\frac{\log p}{n}}+\frac{q}{m}\sqrt{n}, \quad \|\widehat{\gamma}-\gamma^M\|_2 \lesssim \frac{1}{\tau_*}\sqrt{\frac{s\log p}{n}}+\frac{q}{m}\sqrt{n}$$

$$\frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M)\|_2 \lesssim \frac{1}{\tau_*}\sqrt{\frac{s\log p}{n}}+\sqrt{\frac{q\log p}{m}}$$

where $m=\min\{p-1,n\}$, q is the number of unmeasured confounders and $\tau_*>0$ is the lower bound for the restricted eigenvalue defined in (20) and (21).

In the following proposition, we show that the estimator $\widehat{\gamma}$ described in (9) accurately estimates γ and satisfies the property (P2) with a high probability. As we have seen in Section A.2, due to the confounding variables, γ can be decomposed as a sum of a sparse and a small perturbation vector. The proof of Proposition 3 is presented in Section B.7.

Proposition 3. *Suppose that the conditions* (A1) – (A4) *hold. If the spectral transformation* $\mathcal{P}^{(j)}$ *satisfies* (P1) *and the tuning parameter* λ_j *in* (9) *is chosen as* $\lambda_j \geq A\sigma_j\sqrt{\log p/n}$ *for some positive constant* $A>0$*, then with probability larger than* $1-e\cdot p^{1-c(A/M_0)^2}-\exp(-cn)$ *for some positive constant* $c>0$*, the estimator* $\widehat{\gamma}$ *proposed in* (9) *satisfies*

$$\|\widehat{\gamma}-\gamma^M\|_1 \leq \frac{1}{\tau_*}s\lambda_j+\frac{1}{\lambda_j}\left(\frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2}{\sqrt{n}}\right)^2 \lesssim \frac{1}{\tau_*}s\sqrt{\frac{\log p}{n}}+\frac{q}{m}\sqrt{n} \tag{33}$$

$$\|\widehat{\gamma}-\gamma^M\|_2 \leq \frac{1}{\tau_*}\sqrt{s}\lambda_j+\frac{1}{\lambda_j}\left(\frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2}{\sqrt{n}}\right)^2 \lesssim \frac{1}{\tau_*}\sqrt{s\frac{\log p}{n}}+\frac{q}{m}\sqrt{n} \tag{34}$$

$$\frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M)\|_2 \le \frac{1}{\tau_*}\sqrt{s}\lambda_j + \frac{\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2}{\sqrt{n}} \lesssim \frac{1}{\tau_*}\sqrt{s\frac{\log p}{n}} + \sqrt{\frac{q\log p}{m}}. \quad (35)$$

where $\tau_* > 0$ is the lower bound for the restricted eigenvalue defined in (21)*.*

As a remark, we combine (28) and (31) and establish

$$\frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2 \lesssim \max\{1,\frac{p}{n}\}\cdot\frac{q\sqrt{\log p}}{p} \le \frac{q\sqrt{\log p}}{\min\{n,p\}}, \quad (36)$$

which implies the second inequalities in (33), (34) and (35).

In addition, we show an analogous result that the initial spectral deconfounding estimator $\widehat{\beta}^{init}$ proposed in (16) satisfies the condition (P2) with a high probability:

Proposition 4. *Suppose that the conditions* (A1) – (A4) *hold. If the spectral transformation $\mathcal{Q}$ satisfies* (P1) *and the tuning parameter λ in* (16) *is chosen as $\lambda \ge A\sigma_e\sqrt{\log p/n} + \sqrt{q\log p/p}$ for some positive constant $A > 0$, then with probability larger than $1 - e\cdot p^{1-c(A/M_0)^2} - \exp(-cn) - (\log p)^{-1/2}$ for some positive constant $c > 0$, the estimator $\widehat{\beta}^{init}$ proposed in* (16) *satisfies*

$$\|\widehat{\beta}^{init}-\beta\|_1 \le \frac{1}{c*}k\lambda + \frac{1}{\lambda}\left(\frac{\|\mathcal{Q}Xb\|_2}{\sqrt{n}}\right)^2$$

$$\|\widehat{\beta}^{init}-\beta\|_2 \le \frac{1}{c*}\sqrt{k}\lambda + \frac{1}{\lambda}\left(\frac{\|\mathcal{Q}Xb\|_2}{\sqrt{n}}\right)^2$$

$$\frac{1}{\sqrt{n}}\|\mathcal{Q}X(\widehat{\beta}^{init}-\beta)\|_2 \le \frac{1}{c*}\sqrt{k}\lambda + \frac{\|\mathcal{Q}Xb\|_2}{\sqrt{n}}$$

where $\tau_* > 0$ is the lower bound for the restricted eigenvalue defined in (20)*.*

This extends the results in [10], where only the rate of convergence of $\|\widehat{\beta}^{init}-\beta\|_1$ has been established, but not of $\|\widehat{\beta}^{init}-\beta\|_2$ and $\frac{1}{\sqrt{n}}\|\mathcal{Q}X(\widehat{\beta}^{init}-\beta)\|_2$. The proof of Proposition 4 is presented in Section B.8. Similar to (36), we combine (28) and (31) and establish

$$\frac{1}{n}\|\mathcal{Q}Xb\|_2^2 \lesssim \max\{1,\frac{p}{n}\}\cdot\frac{q\sqrt{\log p}}{p} \le \frac{q\sqrt{\log p}}{\min\{n,p\}}, \quad (37)$$

Together with Proposition 4, We obtain that $\widehat{\beta}^{init}$ satisfies the property (P2) by taking $\lambda = A\sigma_e\sqrt{\log p/n} + \sqrt{q\log p/p}$. As a remark, if we further assume the error ϵ_i in the model (3) to be independent of $X_{i,\cdot}$, then we can take $\lambda = A\sigma_\epsilon\sqrt{\log p/n}$ and establish a slightly better rate of convergence.

A.4 Proof of Theorem 1

We write

$$V = \frac{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j\cdot\sigma_e^2}{(Z_j^\intercal(\mathcal{P}^{(j)})^2X_j)^2}$$

for the variance of the estimator $\widehat{\beta}_j$. We decompose

$$\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} = \frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}e}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} + \frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\Delta}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}$$

with $\Delta_i = \psi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot}$. Since e_i is Gaussian and independent of $X_{i,\cdot}$ and Z_j is a function of X, we establish

$$\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}e}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} \mid X \sim N(0,1). \tag{38}$$

It follows from Lemma 2 that

$$\frac{1}{n}\mathbb{E}\|\Delta\|_2^2 = \mathbb{E}|\Delta_i|^2 = \phi^\intercal\left(\mathrm{I}_q - \Psi\Sigma_X^{-1}\Psi^\intercal\right)\phi \lesssim q\sqrt{\log p}/p. \tag{39}$$

By Cauchy inequality, we have

$$\left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\Delta}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}\right| \leq \frac{1}{\sigma_e^2}\|\Delta\|_2.$$

Combined with (39), we establish that, with probability larger than $1-(\log p)^{-1/2}$,

$$\left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\Delta}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}\right| \lesssim \sqrt{\frac{nq\log p}{p}} \tag{40}$$

If $p \gg qn\log p$, we combine (38) and (40) and establish

$$\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} \xrightarrow{d} N(0,1). \tag{41}$$

From the equation (11), we have the following expression

$$\frac{1}{\sqrt{V}}(\widehat{\beta}_j - \beta_j) = \frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}\epsilon}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} + B_\beta + B_b, \tag{42}$$

where B_β and B_b the absolute value of the (scaled) bias terms defined as

$$B_\beta = \frac{Z_j^\intercal(\mathcal{P}^{(j)})^2X_{-j}(\widehat{\beta}_{-j}^{init} - \beta_{-j})}{\sqrt{Z_j^\intercal(\mathcal{P}^{(j)})^4Z_j\cdot\sigma_e^2}} \quad \text{and} \quad B_b = \frac{Z_j^\intercal(\mathcal{P}^{(j)})^2Xb}{\sqrt{Z_j^\intercal(\mathcal{P}^{(j)})^4Z_j\cdot\sigma_e^2}}.$$

We establish in the following lemma that B_b and B_β converges to 0 in probability under certain model conditions. The proof of this lemma is presented in Section B.1.

Lemma 3. *Suppose that the spectral transformation $\mathcal{P}^{(j)}$ satisfies* (P1)*, the initial estimators $\widehat{\gamma}$ and $\widehat{\beta}^{init}$ satisfy* (P2)*, and $\max\left\{\|\gamma-\gamma^M\|_2, \|b\|_2\right\} \lesssim \sqrt{p(\log p)^{1/2}/q}$ and $|b_j| \lesssim p(\log p)^{1/2}/q$. If $s \ll n/\log p$, $q \ll \min\left\{\sqrt{n}/(\log p)^{3/4}, n/[s(\log p)^{3/2}], p/\sqrt{n\log p}\right\}$, and $k \ll \sqrt{\min\{n, p/q\}}/\log p$, then we have*

$$\frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}{\mathrm{Tr}[(\mathcal{P}^{(j)})^2]\sigma_j^2} \xrightarrow{p} 1 \qquad \frac{Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j}{\mathrm{Tr}[(\mathcal{P}^{(j)})^4]\sigma_j^2} \xrightarrow{p} 1 \tag{43}$$

$$B_\beta \xrightarrow{p} 0 \qquad B_b \xrightarrow{p} 0. \tag{44}$$

For the above lemma, the conditions "spectral transformation $\mathcal{P}^{(j)}$ satisfies (P1)", "initial estimators $\widehat{\gamma}$ and $\widehat{\beta}^{init}$ satisfy (P2)", and "$\max\left\{\|\gamma-\gamma^M\|_2, \|b\|_2\right\} \lesssim \sqrt{p(\log p)^{1/2}/q}$ and $|b_j| \lesssim p(\log p)^{1/2}/q$" have been verified in Sections A.1, A.3 and A.2, respectively. By applying the decomposition (42) together with (41) and (44), we establish the limiting distribution in (22). The asymptotic expression of the variance V in (23) follows from (43).

A.5 Verification of Assumption (A2)

In the following two sections, we verify the conditions (A2) and (A4) for a general class of models. We now present the verification of condition (A2), whose proof can be found in Section B.4.

Lemma 4. *Suppose that $\{\Psi_{\cdot,l}\}_{1\le l\le p}$ are generated as i.i.d. q-dimensional sub-Gaussian random vectors with mean zero and covariance $\Sigma_\Psi \in \mathbb{R}^{q\times q}$. If $q \ll p$, $\lambda_{\max}(\Sigma_\Psi)/\lambda_{\min}(\Sigma_\Psi) \le C$ and $\|\phi\|_\infty/\lambda_{\min}(\Sigma_\Psi) \le C$ for some positive constant $C > 0$, then with probability larger than $1-(\log p)^{2c}$, we have*

$$\lambda_q(\Psi) \ge \lambda_q(\Psi_{-j}) \gtrsim \sqrt{p}\sqrt{\lambda_{\min}(\Sigma_\Psi)} \tag{45}$$

$$\max\left\{\|\Psi(\Omega_E)_{\cdot,j}\|_2, \|\Psi_j\|_2, \|\Psi_{-j}(\Omega_E)_{-j,j}\|_2, \|\phi\|_2\right\} \lesssim \sqrt{\lambda_{\max}(\Sigma_\Psi)}\cdot\sqrt{q}(\log p)^c, \tag{46}$$

where $c > 0$ is a positive constant.

The above lemma implies condition (A2) with $\lambda_{\min}(\Sigma_\Psi)$ and $\lambda_{\max}(\Sigma_\Psi)$ taken at a constant order. In fact, a slightly weaker version of condition (A2) is sufficient for our analysis, that is,

$$\frac{\|\Psi_{-j}(\Omega_E)_{-j,j}\|_2 + \|\Psi_j\|_2}{\lambda_q(\Psi_{-j})} \lesssim \sqrt{q/p}(\log p)^c, \quad \frac{\|\Psi(\Omega_E)_{\cdot,j}\|_2 + \|\phi\|_2}{\lambda_q(\Psi)} \lesssim \sqrt{q/p}(\log p)^c. \tag{47}$$

Condition (A2) directly implies (47). In comparison to Condition (A2), (47) is weaker in the sense that (47) still holds if the numerators $\|\Psi_{-j}(\Omega_E)_{-j,j}\|_2+\|\Psi_j\|_2$ and $\|\Psi(\Omega_E)_{\cdot,j}\|_2+\|\phi\|_2$ and the denominator $\lambda_q(\Psi_{-j})$ and $\lambda_q(\Psi)$ are rescaled by the same amount.

We can also apply Lemma 4 to directly verify (47). In verifying (47), the smallest eigenvalue of the covariance matrix Σ_Ψ is not required to be bounded away from zero but simply assume a type of "well-conditioning" assumption $\lambda_{\max}(\Sigma_\Psi)/\lambda_{\min}(\Sigma_\Psi) \leq C$. In another way, if we rescale the regression coefficient matrix Ψ and ϕ by the same amount, Lemma 4 still impies (47) .

The conclusion of Lemma 4 can be generalized to hold if a fixed proportion of columns of Ψ are i.i.d. sub-Gaussian in $\mathbb{R}^q$. This generalized result is stated in the following lemma, whose proof is presented in Section B.5:

Lemma 5. *Suppose that there exists a set $A \subset \{1, 2, \ldots, p\}$ such that $\{\Psi_{\cdot,l}\}_{l\in A}$ are generated as i.i.d sub-Gaussian random vector with mean zero and covariance $\Sigma_\Psi \in \mathbb{R}^{q\times q}$ and $\{\Psi_{\cdot,l}\}_{l\in A^c}$ are generated as independent q-dimensional sub-Gaussian random vectors with sub-Gaussian norm M_0. If $|A|/p \to r$ for some positive constant $r > 0$, $q \ll p$, $\max\{M_0, \lambda_{\max}(\Sigma_\Psi)\}/\lambda_{\min}(\Sigma_\Psi) \leq C$ and $\|\psi\|_\infty/\lambda_{\min}(\Sigma_\Psi) \leq C$ for some positive constant $C > 0$, then the assumption* (A2) *holds with probability larger than $1 - (\log p)^{2c}$.*

A.6 Verification of Assumption (A4)

The restricted eigenvalue condition (A4) is similar, but more complicated than the standard restricted eigenvalue condition introduced in [3]. The main complexity is that, rather than for the original design matrix, the restricted eigenvalue condition is imposed on the transformed design matrices $\mathcal{P}^{(j)}X_{-j}$ and $\mathcal{Q}X$, after applying the Trim transforms $\mathcal{P}^{(j)}$ and $\mathcal{Q}$, described in detail in Sections 3.3 and 3.4, respectively. We verify the restricted eigenvalue condition (A4) for $\frac{1}{n}X^\intercal\mathcal{Q}^2X$ and the argument can be extended to $\frac{1}{n}X_{-j}^\intercal(\mathcal{P}^{(j)})^2X_{-j}$.

In the following, we will verify the restricted in two special settings, even though we believe that the assumption (A4) is satisfied for any random design matrix X with i.i.d. sub-Gaussian rows, where $\lambda_{\min}(\Sigma_X)$ is bounded from below. The first setting is the moderately high-dimensional setting, where $n \geq cp$ for some $c > 1$ and the rows of X have sub-Gaussian distribution. In this case, the assumption (A4) is a direct consequence of the following proposition:

Proposition 5. *Suppose that $n > cp$ for some positive constants $c > 1$ and $X_{i,\cdot} = \Sigma_X^{1/2}Z_{i\cdot}$, where the entries of $Z_{i\cdot} \in \mathbb{R}^p$ are i.i.d sub-Gaussian random variables and the covariance matrix Σ_X satisfies $\lambda_{\min}(\Sigma_X) \geq \tau_0$ for some $\tau_0 > 0$. Then there exist positive constants $c_1 \in (0, 1)$ such that, with probability larger than $1 - c_1^n$, we have $\lambda_{\min}\left(\frac{1}{n}X^\intercal\mathcal{Q}^2X\right) \geq c_2$ for some positive constant $c_2 > 0$.*

The second setting is the more challenging high-dimensional setting $p \geq cn$ for some positive constant $c > 1$. We adapt the theoretical techniques developed in [50] to establish the restricted eigenvalue condition for Gaussian random designs:

Proposition 6. *Suppose that the rows of X are i.i.d. random vectors with $N(0, \Sigma_X)$ distribution. Assume further that $\lambda_{\max}(\Sigma_E) \lesssim p/(n \log p)$, $q \lesssim n/(\log p)$ and $p \geq cn$ for*

some $c > 1$. Then with probability larger than $1-\exp(-c_1 n) - p^{-c_1} - c_2^n$ for some positive constants $c_1 > 0$ and $c_2 \in (0,1)$, we have

$$\begin{aligned}\mathrm{RE}\left(\frac{1}{n}X^\intercal \mathcal{Q}^2 X\right) &\gtrsim \frac{p}{n}\lambda_{\min}(\Sigma_X)\cdot \mathrm{RE}(\frac{1}{n}VV^\intercal) \\ &\gtrsim \lambda_{\min}(\Sigma_X)\cdot\left[\min\left\{\frac{p}{n}, \lambda_{\min}(\Sigma_X)\right\} - k\sqrt{\frac{\log p}{n}}\|\Sigma_X\|_\infty\right].\end{aligned}$$

Note that in the Hidden Confounding Model (2) we have $\Sigma_X = \Psi^\intercal\Psi + \Sigma_E$. If we assume $\|\Psi_{\cdot,l}\|_2 \lesssim \sqrt{q}(\log p)^c$, then we have $\|\Sigma_X\|_\infty \le \|\Sigma_E\|_\infty + q(\log p)^{2c}$. Therefore, as long as

$$k\sqrt{\frac{\log p}{n}}\frac{\|\Sigma_E\|_\infty + q(\log p)^{2c}}{\lambda_{\min}(\Sigma_X)} \to 0, \tag{48}$$

the Proposition 6 shows that the restricted eigenvalue condition holds with $\mathrm{RE}(\frac{1}{n}X^\intercal\mathcal{Q}^2X) \gtrsim \lambda_{\min}(\Sigma_X)$.

B Additional Proofs

B.1 Proof of Lemma 3

We introduce the following lemma about the concentration of quadratic forms, which is Theorem 1.1 in [49].

Lemma 6. (Hanson-Wright inequality) *Let $\xi \in \mathbb{R}^n$ be a random vector with independent sub-Gaussian components ξ_i with zero mean and sub-Gaussian norm K. Let A be an $n \times n$ matrix. Then for every $t \ge 0$,*

$$\mathbf{P}\left(|\xi^\intercal A\xi - \mathbb{E}\xi^\intercal A\xi| > t\right) \le 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right)\right] \tag{49}$$

In the following, we control the two bias components. Note that

$$|B_b| = \left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j} + \frac{1}{\sqrt{V}}b_j\right| \le \frac{\left|(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}\right|}{\sqrt{\sigma_\epsilon^2\cdot Z_j^\intercal(\mathcal{P}^{(j)})^4Z_j}} + \left|\frac{1}{\sqrt{V}}b_j\right|$$

$$|B_\beta| = \left|\frac{1}{\sqrt{V}}\frac{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}(\beta_{-j} - \widehat{\beta}_{-j}^{init})}{(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_j}\right| \le \frac{\left|(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}(\beta_{-j} - \widehat{\beta}_{-j}^{init})\right|}{\sqrt{\sigma_\epsilon^2\cdot Z_j^\intercal(\mathcal{P}^{(j)})^4Z_j}}$$

The control of these bias components and also the limit of variance requires the following lemma, whose proof can be found at the end of the current subsection.

Lemma 7. *Suppose that the spectral deconfounding $\mathcal{P}^{(j)}$ satisfies* (P1)*, the initial estimators $\widehat{\gamma}$ and $\widehat{\beta}^{init}$ satisfy* (P2)*, $\max\{\|\gamma-\gamma^M\|_2, \|b\|_2\} \lesssim \sqrt{p/q}(\log p)^{1/4}$ and $|b_j| \lesssim p\sqrt{\log p}/q$, then with probability larger than $1-\exp(-ct^2)-\frac{1}{t^2}-p^{-c}-\exp(-cn)$,*

$$\begin{aligned}&\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j - \frac{\mathrm{Tr}[(\mathcal{P}^{(j)})^2]}{n}\sigma_j^2\right| \\ &\lesssim \frac{t}{\sqrt{n}}\left(\sqrt{\frac{m}{n}}+\sqrt{\frac{q(\log p)^{1/2}}{m}}\right)+\sqrt{\frac{(s+q(\log p)^{1/2})\log p}{n}}+\frac{q}{m}(\log p)^{1/2}\end{aligned} \tag{50}$$

$$\left|\frac{1}{n}Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j - \frac{\mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{n}\sigma_j^2\right| \lesssim \frac{t}{\sqrt{n}}\cdot\sqrt{\frac{m}{n}+\frac{q(\log p)^{1/2}}{m}}+\frac{s\log p}{n}+\frac{q\log p}{m} \tag{51}$$

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}b_{-j}\right| \lesssim \frac{t}{\sqrt{n}}\cdot\sqrt{\frac{q(\log p)^{1/2}}{m}}+\sqrt{\frac{sq(\log p)^{3/2}}{nm}}+\frac{q(\log p)^{3/4}}{m} \tag{52}$$

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})\right| \le \frac{k\log p}{\sqrt{n\min\{n,p/q\}}}+\frac{q}{m}\sqrt{\frac{\min\{n,p/q\}\log p}{n}} \tag{53}$$

For the high-dimensional setting where $p/n \to c^*$ for a positive constant $c^*>0$, then $m \asymp n$. We also note $\mathrm{Tr}[(\mathcal{P}^{(j)})^l] \asymp m$ for $m=\min\{n,p\}$ and $l=2,4,8$. Then we simplify (50) to (53) as

$$\left|\frac{\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j}{\frac{\mathrm{Tr}[(\mathcal{P}^{(j)})^2]}{n}\sigma_j^2}-1\right| \lesssim \frac{t}{\sqrt{n}}\left(1+\sqrt{\frac{q(\log p)^{1/2}}{n}}\right)+\sqrt{\frac{(s+q(\log p)^{1/2})\log p}{n}}+\frac{q}{n}(\log p)^{1/2} \tag{54}$$

$$\left|\frac{\frac{1}{n}Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j}{\frac{\mathrm{Tr}[(\mathcal{P}^{(j)})^4]}{n}\sigma_j^2}-1\right| \lesssim \frac{t}{\sqrt{n}}\left(1+\sqrt{\frac{q(\log p)^{1/2}}{n}}\right)+\frac{s\log p}{n}+\frac{q\log p}{n} \tag{55}$$

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}b_{-j}\right| \lesssim \frac{t}{\sqrt{n}}\cdot\sqrt{\frac{q(\log p)^{1/2}}{n}}+\frac{\sqrt{sq(\log p)^{3/2}}}{n}+\frac{q(\log p)^{3/4}}{n} \tag{56}$$

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})\right| \lesssim \frac{k\log p}{\sqrt{n\min\{n,p/q\}}}+\frac{q}{m}\sqrt{\frac{\min\{n,p/q\}\log p}{n}}. \tag{57}$$

Under the condition $s \ll n/\log p$ and $q \ll n/(\log p)^{3/2}$, we establish (43) by combining (54) and (55).

By (54) and (55) and the conditions $|b_j| \lesssim p\sqrt{\log p}/q$, $q \ll p/\sqrt{n\log p}$, we have

$$\frac{1}{\sqrt{V}}b_j \lesssim \sqrt{n}\frac{q\sqrt{\log p}}{p} \to 0 \tag{58}$$

By (55) and (56), we have

$$\frac{\left|(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)} X_{-j} b_{-j}\right|}{\sqrt{\sigma_e^2 \cdot Z_j^\intercal (\mathcal{P}^{(j)})^4 Z_j}} \lesssim t\sqrt{n}\sqrt{\frac{q(\log p)^{3/2}}{n}}\left(\sqrt{\frac{s}{n}}+\sqrt{\frac{q}{n}}\right) \tag{59}$$

By (55) and (57), we establish

$$\left|\frac{(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)} X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})}{\sqrt{\sigma_e^2 \cdot Z_j^\intercal (\mathcal{P}^{(j)})^4 Z_j}}\right| \lesssim \frac{q(\log p)^{1/2}}{\sqrt{n}} + \frac{k\log p}{\sqrt{\min\{n, p/q\}}} \tag{60}$$

Under the additional assumptions $q \ll \min\left\{\sqrt{n}/(\log p)^{3/4}, n/[s(\log p)^{3/2}], p/\sqrt{n\log p}\right\}$ and $k \ll \sqrt{\min\{n,p/q\}}/\log p$, then we establish (44) by combining (58), (59) and (60).

Now we present the proof of Lemma 7.

Proof of (50) Define $\eta_j = (\eta_{1,j}, \ldots, \eta_{n,j})^\intercal \in \mathbb{R}^n$. We decompose $\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j$ as

$$\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_j = \frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}\gamma - \frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M-\gamma^A))^\intercal \mathcal{P}^{(j)}\eta_j + \frac{1}{n}(\mathcal{P}^{(j)}\eta_j)^\intercal \mathcal{P}^{(j)}\eta_j$$

We control the right hand side of the above equation term by term. By applying (49) with $A = (\mathcal{P}^{(j)})^2$, then with probability larger than $1-2\exp(-ct^2)$,

$$\left|(\mathcal{P}^{(j)}\eta_j)^\intercal \mathcal{P}^{(j)}\eta_j - \mathrm{Tr}[(\mathcal{P}^{(j)})^2]\cdot\sigma_j^2\right| \lesssim t\cdot\sqrt{\mathrm{Tr}[(\mathcal{P}^{(j)})^4]} \lesssim t\sqrt{m}. \tag{61}$$

By the fact that $\|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}\|_\infty \lesssim \lambda_j$ with probability larger than $1-p^{-c}-\exp(-cn)$, we apply Hölder's inequality and establish

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M))^\intercal \mathcal{P}^{(j)}\eta_j\right| \le \|\widehat{\gamma}-\gamma^M\|_1 \|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}\|_\infty \lesssim \|\widehat{\gamma}-\gamma^M\|_1\lambda_j.$$

Following from the fact that $\widehat{\gamma}$ satisfies the property (P2), we have

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M))^\intercal \mathcal{P}^{(j)}\eta_j\right| \lesssim s\frac{\log p}{n} + \frac{q}{m}\sqrt{\log p}, \tag{62}$$

where the last inequality follows from (33). Since η_j is independent of X_{-j}, we show that $\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal \mathcal{P}^{(j)}\eta_j$ has mean zero and variance

$$\frac{\sigma_j^2}{n^2}(\gamma^A)^\intercal X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal \gamma^A \lesssim \frac{\|\gamma^A\|_2^2}{n}\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal\|_2 \lesssim \frac{1}{n}\cdot\frac{q(\log p)^{1/2}}{m},$$

where the last inequality follows from the control of $\|\gamma^A\|_2$ together with the property (P1). Then with probability larger than $1-\frac{1}{t^2}$,

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal \mathcal{P}^{(j)}\eta_j\right| \lesssim \frac{t}{\sqrt{n}}\cdot\sqrt{\frac{q(\log p)^{1/2}}{m}}. \tag{63}$$

By $\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}\gamma\right| \le \|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}\|_\infty \|\gamma\|_1$ and the KKT condition of (9), we have

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}\gamma\right| \le \|\gamma\|_1 \|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal \mathcal{P}^{(j)}X_{-j}\|_\infty \lesssim \lambda_j \|\gamma\|_1.$$

We then further control the right hand side as

$$\left(\|\gamma^M\|_1 + \|\gamma^A\|_1\right)\lambda_j \le \sqrt{s}\|\gamma^M\|_2\lambda_j + \sqrt{p}\|\gamma^A\|_2\lambda_j \lesssim \sqrt{\frac{(s+q(\log p)^{1/2})\log p}{n}} \tag{64}$$

Then a combination of (61), (62), (63) and (64) leads to (50).

Proof of (51) Note that

$$\frac{1}{n}Z_j^\intercal(\mathcal{P}^{(j)})^4 Z_j = \frac{1}{n}\|(\mathcal{P}^{(j)})^2 X_{-j}(\gamma^M - \widehat{\gamma} + \gamma^A)\|_2^2 + \frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^4\eta_j + 2\eta_j^\intercal(\mathcal{P}^{(j)})^4 X_{-j}(\gamma^M - \widehat{\gamma} + \gamma^A) \tag{65}$$

By applying (49) with $A = (\mathcal{P}^{(j)})^4$, then with probability larger than $1 - 2\exp(-ct^2)$,

$$\left|\eta_j^\intercal(\mathcal{P}^{(j)})^4\eta_j - \mathrm{Tr}[(\mathcal{P}^{(j)})^4]\cdot\sigma_j^2\right| \lesssim t\cdot\sqrt{\mathrm{Tr}[(\mathcal{P}^{(j)})^8]} \lesssim t\sqrt{m}. \tag{66}$$

By a similar argument as (62), we have

$$\left|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^4 X_{-j}(\widehat{\gamma} - \gamma^M)\right| \le \|\widehat{\gamma} - \gamma^M\|_1 \|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}\|_\infty \lesssim \frac{s\log p}{n} + \frac{q}{m}(\log p)^{1/2}. \tag{67}$$

In addition, $\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^4 X_{-j}\gamma^A$ has mean zero and variance

$$\frac{\sigma_j^2}{n^2}(\gamma^A)^\intercal X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal \gamma^A \lesssim \frac{\|\gamma^A\|_2^2}{n}\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal\|_2$$

and hence with probability larger than $1 - \frac{1}{t^2}$,

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal \mathcal{P}^{(j)}\eta_j\right| \lesssim \frac{t\|\gamma^A\|_2}{\sqrt{n}}\sqrt{\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4 X_{-j}^\intercal\|_2} \lesssim \frac{t}{\sqrt{n}}\cdot\sqrt{\frac{q(\log p)^{1/2}}{m}} \tag{68}$$

Note that

$$\begin{aligned}&\frac{1}{n}\|(\mathcal{P}^{(j)})^2 X_{-j}(\widehat{\gamma} - \gamma^M - \gamma^A)\|_2^2 \le \frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma} - \gamma^M - \gamma^A)\|_2^2 \\ &\lesssim \frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma} - \gamma^M)\|_2^2 + \frac{1}{n}\|\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2 \lesssim \frac{s\log p}{n} + \frac{q\log p}{m}\end{aligned} \tag{69}$$

where the last inequality follows from the control of $\|\gamma^A\|_2$ together with the properties (P1) and (P2). Then we establish (51) by combining (67), (68) and (69).

Proof of (52) We investigate $\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}$:

$$\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}=\frac{1}{n}(\mathcal{P}^{(j)}\eta_j)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}+\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M-\gamma^A))^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}$$

Note that $\frac{1}{n}(\mathcal{P}^{(j)}\eta_j)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}$ has mean zero and variance

$$\frac{\sigma_j^2}{n^2}(b_{-j})^\intercal X_{-j}(\mathcal{P}^{(j)})^4X_{-j}^\intercal b_{-j}\lesssim\frac{1}{n}\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4X_{-j}^\intercal\|_2\|b_{-j}\|_2^2,$$

and hence with probability larger than $1-\frac{1}{t^2}$,

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}b_{-j})^\intercal\mathcal{P}^{(j)}\eta_j\right|\lesssim t\frac{\|b_{-j}\|_2}{\sqrt{n}}\sqrt{\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^4X_{-j}^\intercal\|_2}\leq\frac{t}{\sqrt{n}}\sqrt{\frac{q(\log p)^{1/2}}{m}}\quad(70)$$

In addition, we note the following two inequalities

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}\gamma^A)^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}\right|\lesssim\|\gamma^A\|_2\|b_{-j}\|_2\|\frac{1}{n}X_{-j}(\mathcal{P}^{(j)})^2X_{-j}^\intercal\|_2\lesssim\frac{q(\log p)^{1/2}}{m}$$

$$\begin{aligned}\left|\frac{1}{n}(\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M))^\intercal\mathcal{P}^{(j)}X_{-j}b_{-j}\right|&\lesssim\frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^M)\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\|_2\|b_{-j}\|_2\\&\lesssim\sqrt{\frac{sq(\log p)^{3/2}}{nm}}+\frac{q(\log p)^{3/4}}{m}.\end{aligned}$$

Together with (70), we establish (52).

Proof of (53) We investigate $\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})$. It follows from Hölder's inequality and also the KKT condition of (9) that

$$\left|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}(\beta_{-j}-\widehat{\beta}_{-j}^{init})\right|\leq\|\beta_{-j}-\widehat{\beta}_{-j}^{init}\|_1\|\frac{1}{n}(\mathcal{P}^{(j)}Z_j)^\intercal\mathcal{P}^{(j)}X_{-j}\|_\infty\lesssim\lambda_j\|\beta_{-j}-\widehat{\beta}_{-j}^{init}\|_1$$

By the propoerty (P2), we establish (53).

B.2 Proof of Lemma 1

We express the model (2) as

$$X_{1,j}=\Psi_j^\intercal H_{1,\cdot}+E_{1,j},\quad X_{1,-j}=\Psi_{-j}^\intercal H_{1,\cdot}+E_{1,-j},$$

where $\Psi_j\in\mathbb{R}^q$ denotes the j-th column of Ψ_j and $\Psi_{-j}\in\mathbb{R}^{q\times(p-1)}$ denotes the sub-matrix of Ψ except for the j-th column. We define $B=\mathbb{E}E_{1,-j}E_{1,-j}^\intercal$. Since $\mathrm{Cov}(H_{i,\cdot})=\mathrm{I}_{q\times q}$ and the components of $H_{i,\cdot}$ are uncorrelated with the components of $\mathbb{E}_{i,\cdot}$, then we have

$$\gamma=[\mathbb{E}(X_{1,-j}X_{1,-j}^\intercal)]^{-1}\mathbb{E}(X_{1,-j}X_{1,j})=\left(\Psi_{-j}^\intercal\Psi_{-j}+B\right)^{-1}\left(\Psi_{-j}^\intercal\Psi_j+\mathbb{E}E_{1,j}E_{1,-j}\right).$$

We apply woodbury matrix identity and then have

$$\left(\Psi_{-j}^{\intercal}\Psi_{-j}+B\right)^{-1}=B^{-1}-B^{-1}\Psi_{-j}^{\intercal}\left(\mathrm{I}+\Psi_{-j}B^{-1}\Psi_{-j}^{\intercal}\right)^{-1}\Psi_{-j}B^{-1}.$$

We combine the above two equalities and establish the decomposition $\gamma=\gamma^M+\gamma^A$ with

$$\gamma^M=B^{-1}\mathbb{E}E_{1,j}E_{1,-j}$$

and

$$\gamma^A=\left(\Psi_{-j}^{\intercal}\Psi_{-j}+B\right)^{-1}\Psi_{-j}^{\intercal}\Psi_j-B^{-1}\Psi_{-j}^{\intercal}\left(\mathrm{I}+\Psi_{-j}B^{-1}\Psi_{-j}^{\intercal}\right)^{-1}\Psi_{-j}\gamma^M. \tag{71}$$

It remains to verify (31) for the approximation vector γ^A. We define $D=\Psi_{-j}B^{-\frac{1}{2}}\in\mathbb{R}^{q\times(p-1)}$ and hence the first component on the right hand side of (71) can be expressed as

$$\left(\Psi_{-j}^{\intercal}\Psi_{-j}+B\right)^{-1}\Psi_{-j}^{\intercal}\Psi_j=B^{-\frac{1}{2}}\left(D^{\intercal}D+\mathrm{I}\right)^{-1}D^{\intercal}\Psi_j.$$

By woodbury matrix identity, we have

$$\left(D^{\intercal}D+\mathrm{I}\right)^{-1}D^{\intercal}=\left(\mathrm{I}-D^{\intercal}(\mathrm{I}+DD^{\intercal})^{-1}D\right)D^{\intercal}=D^{\intercal}(\mathrm{I}+DD^{\intercal})^{-1}$$

and hence

$$\left(\Psi_{-j}^{\intercal}\Psi_{-j}+B\right)^{-1}\Psi_{-j}^{\intercal}\Psi_j=B^{-\frac{1}{2}}D^{\intercal}(\mathrm{I}+DD^{\intercal})^{-1}\Psi_j. \tag{72}$$

The second component on the right hand side of (71) can be expressed as

$$B^{-\frac{1}{2}}D^{\intercal}\left(\mathrm{I}+DD^{\intercal}\right)^{-1}\Psi_{-j}\gamma^M.$$

Together with (72), we simplify (71) as

$$\gamma^A=B^{-\frac{1}{2}}D^{\intercal}\left(\mathrm{I}+DD^{\intercal}\right)^{-1}\left(\Psi_j+\Psi_{-j}\gamma^M\right). \tag{73}$$

Under the assumption that $c_0\le\lambda_{\min}(\Omega_E)\le\lambda_{\max}(\Omega_E)\le C_0$, we introduce the SVD for D as $D=U(D)\Lambda(D)V(D)^{\intercal}$, where $U(D),\Lambda(D)\in\mathbb{R}^{q\times q}$ and $V(D)\in\mathbb{R}^{(p-1)\times q}$. Since

$$D^{\intercal}\left(\mathrm{I}+DD^{\intercal}\right)^{-1}=V(D)\Lambda(D)(\Lambda(D)^2+\mathrm{I})^{-1}U(D)^{\intercal},$$

it follows from (73) that

$$\|\gamma^A\|_2\le\|B^{-\frac{1}{2}}\|_2\max_{1\le l\le q}\frac{|\lambda_l(D)|}{\lambda_l^2(D)+1}\|\Psi_j+\Psi_{-j}\gamma^M\|_2, \tag{74}$$

where $\lambda_l(D)$ is the l-th largest singular value of D in absolute value. By the condition $c_0\le\lambda_{\min}(\Omega_E)\le\lambda_{\max}(\Omega_E)\le C_0$, we have $\frac{1}{C_0}\mathrm{I}\preceq B=\mathbb{E}E_{1,-j}E_{1,-j}^{\intercal}\preceq\frac{1}{c_0}\mathrm{I}$. We further have $c_0\lambda_l^2(\Psi_{-j})\le\lambda_l^2(D)\le C_0\lambda_l^2(\Psi_{-j})$ for $1\le l\le q$ and establish the first inequality of (31). The second inequality of (31) follows from condition (A2).

B.3 Proof of Lemma 2

The proof of this lemma is similar to Lemma 1 in terms of controlling $\|b\|_2$. We start with the exact expression of b

$$b = \Sigma_X^{-1}\Psi^\intercal\phi = (\Sigma_E + \Psi^\intercal\Psi)^{-1}\Psi^\intercal\phi.$$

By apply the woodbury matrix inverse formula, we have

$$b = \Sigma_E^{-1}\Psi^\intercal\left(\mathrm{I} + \Psi\Sigma_E^{-1}\Psi^\intercal\right)^{-1}\phi.$$

We define $D_E = \Psi\Sigma_E^{-1/2} \in \mathbb{R}^{q\times p}$ and hence we have

$$b = \Sigma_E^{-1/2}D_E^\intercal(\mathrm{I} + D_E D_E^\intercal)^{-1}\phi$$

and

$$b_j = (\Omega_E)_{\cdot,j}^\intercal\Psi^\intercal(\mathrm{I} + D_E D_E^\intercal)^{-1}\phi \tag{75}$$

Hence, we control $\|b\|_2$ as

$$\|b\|_2 \le \sqrt{C_0}\max_{1\le l\le q}\frac{\lambda_l(D_E)}{1+\lambda_l^2(D_E)}\|\phi\|_2 \lesssim \sqrt{q/p}(\log p)^{1/4}.$$

where the last inequality follows from the fact $c_0\lambda_j^2(\Psi) \le \lambda_j^2(D_E) \le C_0\lambda_j^2(\Psi)$ and the condition (A2). Similarly, we apply condition (A2) and control $|b_j|$ as

$$|b_j| \le \|\Psi(\Omega_E)_{\cdot,j}\|_2\frac{1}{1+\lambda_q^2(D_E)}\|\phi\|_2 \lesssim q\sqrt{\log p}/p.$$

It follows from woodbury matrix inverse formula that

$$\Psi\Sigma_X^{-1}\Psi^\intercal = \Psi^\intercal\Sigma_E^{-1}\Psi(\mathrm{I}_q + \Psi\Sigma_E^{-1}\Psi^\intercal)^{-1}$$

and hence

$$\phi^\intercal\left(\mathrm{I}_q - \Psi\Sigma_X^{-1}\Psi^\intercal\right)\phi = \phi^\intercal(\mathrm{I}_q + \Psi\Sigma_E^{-1}\Psi^\intercal)^{-1}\phi$$

Then (32) follows from the condition (A2) together with the same argument as that for $|b_j|$ (defined in (75)) with replacing the term $(\Omega_E)_{\cdot,j}^\intercal\Psi^\intercal$ with $\psi^\intercal$.

B.4 Proof of Lemma 4

We first control the lower bound of $\lambda_q(\Psi)$ and the argument for $\lambda_q(\Psi_{-j})$ is similar. Note that $\lambda_q^2(\Psi)$ is the smallest eigenvalue of $\Psi\Psi^\intercal = \sum_{l=1}^p \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal$. Since $\Psi_{\cdot,l} \in \mathbb{R}^q$ for $1 \le j \le p$ are i.i.d. sub-Gaussian random vectors, it follows from (5.26) in [54], with probability larger than $1 - p^{-c}$,

$$\|\frac{1}{p}\sum_{l=1}^p \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal - \Sigma_\Psi\|_2 \le C\lambda_{\max}(\Sigma_\Psi)\sqrt{\frac{q+\log p}{p}}$$

for some positive constants $c, C > 0$. This gives us that, with probability larger than $1 - p^{-c}$,

$$\lambda_q^2(\Psi) = \lambda_{\min}(\sum_{j=1}^{p} \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal) \gtrsim p\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}(\Sigma_\Psi)\sqrt{\frac{q+\log p}{p}}\right) \tag{76}$$

Similarly, we establish that, with probability larger than $1 - p^{-c}$,

$$\lambda_q^2(\Psi_{-j}) = \lambda_{\min}(\sum_{l\neq j}^{p} \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal) \gtrsim (p-1)\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}(\Sigma_\Psi)\sqrt{\frac{q+\log p}{p}}\right) \tag{77}$$

In the following, we control Ψa for $a \in \mathbb{R}^p$ by noting that $\mathbb{E}\|\Psi a\|_2^2 = \mathrm{Tr}(\Sigma_\Psi)\|a\|_2^2$. Hence, with probability larger than $1 - \frac{1}{t^2}$, we have

$$\|\Psi a\|_2^2 \leq t^2 \mathrm{Tr}(\Sigma_\Psi)\|a\|_2^2 \leq t^2 q \lambda_{\max}(\Sigma_\Psi)\|a\|_2^2. \tag{78}$$

By taking $a \in \mathbb{R}^p$ as $((\Omega_E)_{1,j}, \ldots, (\Omega_E)_{j-1,j}, 0, (\Omega_E)_{j+1,j}, \ldots, (\Omega_E)_{p,j})$, e_j and $(\Omega_E)_{j,\cdot}$, we establish that with probability larger than $1 - \frac{1}{t^2}$,

$$\|\Psi_{-j}(\Omega_E)_{-j,j}\|_2 \lesssim t\sqrt{q}\sqrt{\lambda_{\max}(\Sigma_\Psi)}\|(\Omega_E)_{-j,j}\|_2 \tag{79}$$

$$\|\Psi_j\|_2 \lesssim t\sqrt{q}\sqrt{\lambda_{\max}(\Sigma_\Psi)} \tag{80}$$

$$\|\Psi(\Omega_E)_{\cdot,j}\|_2 \lesssim t\sqrt{q}\sqrt{\lambda_{\max}(\Sigma_\Psi)}\|(\Omega_E)_{\cdot,j}\|_2 \tag{81}$$

The lemma follows from a combination of (76), (77), (79), (80) and (81).

B.5 Proof of Lemma 5

The proof is a generalization of that of Lemma 4 in Section B.4. Note that $\lambda_q^2(\Psi)$ is the smallest eigenvalue of $\Psi\Psi^\intercal = \sum_{l=1}^{p} \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal$ and $\sum_{l=1}^{p} \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal - \sum_{l\in A} \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal$ is a positive definite matrix. By the same argument for (76), we have

$$\begin{aligned}\lambda_q^2(\Psi) \geq \lambda_{\min}(\sum_{l\in A} \Psi_{\cdot,l}\Psi_{\cdot,l}^\intercal) &\gtrsim |A|\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}(\Sigma_\Psi)\sqrt{q/|A|}\right) \\ &\gtrsim p\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}(\Sigma_\Psi)\sqrt{q/p}\right)\end{aligned} \tag{82}$$

where the last inequality holds due to $|A|/p \to r$ for a positive constant $r > 0$. Similarly, we have

$$\lambda_q^2(\Psi_{-j}) \gtrsim p\left(\lambda_{\min}(\Sigma_\Psi) - \lambda_{\max}(\Sigma_\Psi)\sqrt{q/p}\right) \tag{83}$$

Similarly to (78), we establish that, with probability larger than $1 - \frac{1}{t^2}$,

$$\|\Psi a\|_2^2 \lesssim t^2 q \max\{\lambda_{\max}(\Sigma_\Psi), M_0\}\|a\|_2^2.$$

Then we can establish (79), (80) and (81) by replacing $\sqrt{\lambda_{\max}(\Sigma_\Psi)}$ with $\sqrt{\max\{\lambda_{\max}(\Sigma_\Psi), M_0\}}$. Combined with (82) and (83), we establish the lemma.

B.6 Proof of Proposition 2

The proof relies on the following version of the Cauchy Interlacing law.

Lemma 8. *Suppose that the symmetric matrix $A \in \mathbb{R}^{p\times p}$ has eigenvalues $\lambda_1(A) \geq \lambda_2(A) \geq \ldots \geq \lambda_p(A)$ and the perturbation matrix B has rank r, then we have*

$$\lambda_{l+2r}(A) \leq \lambda_{l+r}(A+B) \leq \lambda_l(A) \tag{84}$$

where $1 \leq l \leq p-2r$.

We express $\widehat{\Sigma}_X$ as

$$\widehat{\Sigma}_X = \frac{1}{n}\Psi^\intercal H^\intercal H\Psi + \frac{1}{n}E^\intercal E + \frac{1}{n}\Psi^\intercal H^\intercal E + \frac{1}{n}E^\intercal H\Psi \tag{85}$$

Note that $\frac{1}{n}\Psi^\intercal H^\intercal H\Psi$ has rank q and $\frac{1}{n}\Psi^\intercal H^\intercal E + \frac{1}{n}E^\intercal H\Psi$ has rank $2q$. By applying Lemma 8 with $r = 3q$, $A = \frac{1}{n}E^\intercal E$ and $B = \widehat{\Sigma}_X - \frac{1}{n}E^\intercal E$, then we establish

$$\lambda_{l+3q}(\widehat{\Sigma}_X) \leq \lambda_l\left(\frac{1}{n}E^\intercal E\right), \tag{86}$$

which is the first inequality in (30). The second inequality of (30) follows from Theorem 5.39 and equation (5.26) in [54], together with the condition that $\lambda_{\max}(\Sigma_E) \leq C_0$.

B.7 Proof of Proposition 3

For the vector $a \in \mathbb{R}^{p-1}$, we define the weighted ℓ_1 norm $\|a\|_{1,w} = \sum_{l\neq j} \frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}}|a_l|$ and define the event

$$\mathcal{A}_0 = \left\{ c \leq \frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}} \leq C \quad \text{for } 1 \leq l \leq p \right\},$$

for some positive constants $C > c > 0$. On the event $\mathcal{A}_0$, we have $\|a\|_1 \asymp \|a\|_{1,w}$. We now show that

$$\mathbb{P}(\mathcal{A}_0) \geq 1 - p^{-c} - \exp(-cn), \tag{87}$$

for some positive constant $c > 0$. By the construction of $\mathcal{P}^{(j)}$, we have $\frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}} \leq \frac{\|X_{\cdot,l}\|_2}{\sqrt{n}}$. Following from the fact that $X_{i,l}$ is sub-Gaussian random variable for $1 \leq l \leq p$ and Corollary 5.17 in [54], we establish that, with probability larger than $1 - p^{-c} - \exp(-cn)$,

$$\frac{\|X_{\cdot,l}\|_2}{\sqrt{n}} \lesssim \sqrt{\mathrm{Var}(X_{1,l})}(1 + \sqrt{\log p/n}) \lesssim C. \tag{88}$$

It follows from condition (A4) that

$$\min_{2\leq l\leq p} \frac{\|\mathcal{P}^{(j)}X_{\cdot,l}\|_2}{\sqrt{n}} \geq \sqrt{\tau_*}. \tag{89}$$

Define $\eta_j = (\eta_{1,j}, \ldots, \eta_{n,j})^\intercal \in \mathbb{R}^n$, $W \in \mathbb{R}^{p\times p}$ as a diagonal matrix with diagonal entries as $W_{l,l} = \|\mathcal{P}^{(j)} X_{\cdot,l}\|_2/\sqrt{n}$ for $1 \le l \le p$, and $\lambda_0 = \|\frac{1}{n}\eta_j^\intercal (\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_\infty$. By assuming $\eta_{i,j} = X_{i,j} - \gamma^\intercal X_{i,-j}$ to be sub-Gaussian and independent of $X_{i,-j}$, we apply Proposition 5.10 in [54] and the maximum inequality to establish

$$\mathbb{P}\left(\lambda_0 \ge A_0\sigma_j\sqrt{\log p/n}\right) \le e\cdot p^{1-c(A/M_0)^2} \tag{90}$$

for some positive constants $A_0 > 0$ and $c > 0$. We take $\lambda_j = A\sigma_j\sqrt{\log p/n}$ for $A = (1+c_1)A_0$ with c_1 denoting a small positive constant.

By the definition of the estimator $\widehat{\gamma}$, we have the following basic inequality,

$$\frac{1}{2n}\|\mathcal{P}^{(j)}(X_j - X_{-j}\widehat{\gamma})\|_2^2 + \lambda_j\|\widehat{\gamma}\|_{1,w} \le \frac{1}{2n}\|\mathcal{P}^{(j)}\left(X_j - X_{-j}\gamma^M\right)\|_2^2 + \lambda_j\|\gamma^M\|_{1,w}. \tag{91}$$

By decomposing $X_j - X_{-j}\widehat{\gamma} = X_{-j}\gamma^A + \eta_j + X_{-j}\left(\gamma^M - \widehat{\gamma}\right)$, we further decompose (91) as

$$\begin{aligned}&\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2^2 + \lambda_j\|\widehat{\gamma}\|_{1,w} \le \lambda_j\|\gamma^M\|_{1,w}\\ &- \frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}\left(\gamma^M - \widehat{\gamma}\right) - \frac{1}{n}\left(\mathcal{P}^{(j)}X_{-j}\gamma^A\right)^\intercal \mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right).\end{aligned} \tag{92}$$

Regarding the right hand side of the above inequality, we have

$$\left|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\right| \le \|\frac{1}{n}\eta_j^\intercal(\mathcal{P}^{(j)})^2 X_{-j}(W_{-j,-j})^{-1}\|_\infty \|W_{-j,-j}(\gamma^M - \widehat{\gamma})\|_1 = \lambda_0\|\gamma^M - \widehat{\gamma}\|_{1,w}$$

and

$$\left|\frac{1}{n}\left(\mathcal{P}^{(j)}X_{-j}\gamma^A\right)^\intercal \mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\right| \le \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2 \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2.$$

Then we further simply (92) as

$$\begin{aligned}\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2^2 + \lambda_j\|\widehat{\gamma}\|_{1,w} &\le \lambda_j\|\gamma^M\|_{1,w} + \lambda_0\|\gamma^M - \widehat{\gamma}\|_{1,w}\\ &+ \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2 \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2.\end{aligned}$$

Let $\mathcal{T}_j$ denote the support of γ^M. By the fact that $\|\gamma^M_{\mathcal{T}_j}\|_{1,w} - \|\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w} \le \|\gamma^M_{\mathcal{T}_j} - \widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}$ and $\|\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w} = \|\gamma^M_{\mathcal{T}_j^c} - \widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}$, then we establish

$$\begin{aligned}&\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2^2 + (\lambda_j - \lambda_0)\|\gamma^M_{\mathcal{T}_j^c} - \widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\\ &\le (\lambda_j + \lambda_0)\|\gamma^M_{\mathcal{T}_j} - \widehat{\gamma}_{\mathcal{T}_j}\|_{1,w} + \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2 \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2.\end{aligned} \tag{93}$$

The following analysis is based on (93) and divided into two cases depending on the dominating value on the right hand side of (93).

Case 1 We consider

$$(\lambda_j + \lambda_0)\|\gamma^M_{\mathcal{T}_j} - \widehat{\gamma}_{\mathcal{T}_j}\|_{1,w} \ge \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2 \|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^M - \widehat{\gamma}\right)\|_2$$

and then simplify (93) as

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2^2+(\lambda_j-\lambda_0)\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\leq(\lambda_j+\lambda_0)\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}. \quad (94)$$

It follows from (94) that $\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\leq\frac{\lambda_j+\lambda_0}{\lambda_j-\lambda_0}\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}$. By the choices of λ_j and λ_0, on the event $\mathcal{A}_0$, we establish $\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_1\leq C\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_1$ for some positive constant $C>0$. By the restricted eigenvalue condition (21), we have

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2^2\geq\frac{\tau_*}{2}\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2^2.$$

Then we have

$$\frac{\tau_*}{2}\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2^2\leq(\lambda_j+\lambda_0)\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\lesssim\sqrt{|\mathcal{T}_j|}\,(\lambda_j+\lambda_0)\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2,$$

which leads to

$$\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_2\lesssim\frac{1}{\tau_*}\sqrt{|\mathcal{T}_j|}\,(\lambda_j+\lambda_0).$$

Hence, we have

$$\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_1\lesssim\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\lesssim\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\lesssim\frac{1}{\tau_*}|\mathcal{T}_j|\,(\lambda_j+\lambda_0). \quad (95)$$

and

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2^2\lesssim\frac{1}{\tau_*}|\mathcal{T}_j|\,(\lambda_j+\lambda_0)^2. \quad (96)$$

We apply the restricted eigenvalue condition (21) again to establish

$$\|\gamma^M-\widehat{\gamma}\|_2\lesssim\sqrt{|\mathcal{T}_j|}\,(\lambda_j+\lambda_0). \quad (97)$$

Case 2 We consider

$$(\lambda_j+\lambda_0)\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\leq\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2$$

and then simplify (93) as

$$\frac{1}{2n}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2^2+(\lambda_j-\lambda_0)\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\leq\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2.$$

Then we derive

$$\frac{1}{\sqrt{n}}\|\mathcal{P}^{(j)}X_{-j}\left(\gamma^M-\widehat{\gamma}\right)\|_2\lesssim\|\frac{1}{\sqrt{n}}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2, \quad (98)$$

$$\|\gamma^M_{\mathcal{T}_j}-\widehat{\gamma}_{\mathcal{T}_j}\|_{1,w}\lesssim\frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{\lambda_j+\lambda_0}\text{ and }\|\gamma^M_{\mathcal{T}_j^c}-\widehat{\gamma}_{\mathcal{T}_j^c}\|_{1,w}\lesssim\frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{\lambda_j-\lambda_0}. \quad (99)$$

Then we also have

$$\|\gamma^M-\widehat{\gamma}\|_2\leq\|\gamma^M-\widehat{\gamma}\|_{1,w}\lesssim\frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{\lambda_j+\lambda_0}+\frac{\|\frac{1}{n}\mathcal{P}^{(j)}X_{-j}\gamma^A\|_2^2}{\lambda_j-\lambda_0}. \quad (100)$$

Finally, we establish (33) by combining (95), (99) and (36); establish (34) by combining (97), (100) and (36); establish (35) by combining (96), (98) and (36).

B.8 Proof of Proposition 4

The proof of Proposition 4 is similar to the proof of Proposition 3 in Section B.7. In the following, we prove Proposition 4 and mainly highlight its difference from the proof of Proposition 3 in Section B.7.

For $a \in \mathbb{R}^p$, we define $\|a\|_{1,w} = \sum_{l=1}^p \frac{\|\mathcal{Q}X_{\cdot,l}\|_2}{\sqrt{n}}|a_l|$ and define the event

$$\mathcal{A}_1 = \left\{ c \le \frac{\|\mathcal{Q}X_{\cdot,l}\|_2}{\sqrt{n}} \le C \quad \text{for } 1 \le l \le p \right\}.$$

for some positive constants $C > c > 0$. On the event $\mathcal{A}_1$, we have $\|a\|_1 \asymp \|a\|_{1,w}$. Similar to (87), we can show that $\mathbb{P}(\mathcal{A}_1) \ge 1 - p^{-c} - \exp(-cn)$ for some positive constant $c > 0$. Define $W \in \mathbb{R}^{p\times p}$ as a diagonal matrix with diagonal entries as $W_{l,l} = \|\mathcal{Q}X_{\cdot,l}\|_2/\sqrt{n}$ for $1 \le l \le p$.

The main part of the proof is to calculate the tuning parameter

$$\lambda = (1+c_0)\|\frac{1}{n}\epsilon^\intercal \mathcal{Q}^2 X W^{-1}\|_\infty$$

for a small positive constant $c_0 > 0$. Note that $\epsilon = e + \Delta$ with $\Delta_i = \psi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot}$. Since e_i is independent of $X_{i,\cdot}$, we apply Proposition 5.10 in [54] and the maximum inequality to establish

$$\mathbb{P}\left(\|\frac{1}{n} e^\intercal \mathcal{Q}^2 X W^{-1}\|_\infty \ge A_0 \sigma_e \sqrt{\log p/n} \right) \le p^{-c} \tag{101}$$

for some positive constants $c > 0$ and $A_0 > 0$. We then control the other part $\|\frac{1}{n}\Delta^\intercal \mathcal{Q}^2 X W^{-1}\|_\infty$ by the inequality

$$\|\frac{1}{n}\Delta^\intercal \mathcal{Q}^2 X W^{-1}\|_\infty \le \frac{1}{\sqrt{n}}\|\Delta\|_2$$

and the upper bound for $\frac{1}{n}\mathbb{E}\|\Delta\|_2^2$ in (39). As a consequence, we have

$$\mathbb{P}\left(\|\frac{1}{n}\Delta^\intercal \mathcal{Q}^2 X W^{-1}\|_\infty \ge \frac{1}{1+c}\sqrt{\frac{q\log p}{p}} \right) \lesssim (\log p)^{-1/2} \tag{102}$$

for a small positive constant $c > 0$. We then choose

$$\lambda_1 = A\sigma_e\sqrt{\log p/n} + \sqrt{q\log p/p} \quad \text{with} \quad A = (1+c_0)A_0$$

for some positive constant $c_0 > 0$. We combine (101) and (102) and establish that

$$\mathbb{P}\left((1+c_0)\|\frac{1}{n}\epsilon^\intercal \mathcal{Q}^2 X W^{-1}\|_\infty \le \lambda_1 \right) \ge 1 - (\log p)^{-1/2} - p^{-c}. \tag{103}$$

By the definition of $\widehat{\beta}^{init}$, we establish the basic inequality in a similar fashion to (91)

$$\frac{1}{2n}\|\mathcal{Q}(Y - X\widehat{\beta}^{init})\|_2^2 + \lambda_1\|\widehat{\beta}^{init}\|_{1,w} \le \frac{1}{2n}\|\mathcal{Q}\left(Y - X\beta\right)\|_2^2 + \lambda_1\|\beta\|_{1,w}. \tag{104}$$

We can apply the similar argument from (91) to (100) by replacing $\mathcal{P}^{(j)}$, X_j, X_{-j}, $\widehat{\gamma}$, γ^M, γ^A with $\mathcal{Q}$, Y, X, $\widehat{\beta}^{init}$, β, b, respectively. We replace the tuning parameters λ_j and λ_0 by λ and $\frac{1}{1+c_0}\lambda$, respectively. Then we establish Proposition 4.

B.9 Proof of Proposition 1

We note that

$$\mathcal{Q}y - \mathcal{Q}X\widehat{\beta}^{init} = \mathcal{Q}e + \mathcal{Q}\Delta + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb.$$

where $\Delta_i = \psi^\intercal H_{i,\cdot} - b^\intercal X_{i,\cdot}$ for $1 \le i \le n$.

Then we have

$$\begin{aligned}\widehat{\sigma}_e^2 - \sigma_e^2 = \frac{\|\mathcal{Q}e\|_2^2}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2 + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}\Delta + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb\|_2^2 \\ + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2\Delta + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\epsilon^\intercal\mathcal{Q}^2X(\beta - \widehat{\beta}^{init}) + \frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\epsilon^\intercal\mathcal{Q}^2Xb.\end{aligned} \tag{105}$$

The following analysis is to study the above decomposition term by term. First note that

$$\frac{\|\mathcal{Q}e\|_2^2}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2 = \frac{e^\intercal US^2U^\intercal e}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2.$$

By Lemma 6, we establish that with probability larger than $1 - \exp(-ct^2)$,

$$\left|\frac{e^\intercal US^2U^\intercal e}{\mathrm{Tr}(\mathcal{Q}^2)} - \sigma_e^2\right| \lesssim t\frac{\sqrt{\mathrm{Tr}(\mathcal{Q}^4)}}{\mathrm{Tr}(\mathcal{Q}^2)}. \tag{106}$$

By (39), we show that

$$\mathbf{P}\left(\frac{1}{n}\|\Delta\|_2^2 \lesssim q\log p/p\right) \ge 1 - (\log p)^{-1/2}. \tag{107}$$

Since e_i is independent of $X_{i,\cdot}$ and $H_{i,\cdot}$, the term $\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2\Delta$ is of mean zero and variance

$$\frac{1}{\mathrm{Tr}^2(\mathcal{Q}^2)}\sigma_e^2\|\mathcal{Q}^2\Delta\|_2^2 \lesssim \frac{\sigma_e^2}{n^2}\|\Delta\|_2^2 \lesssim \frac{q\log p}{np}\sigma_e^2.$$

where the first inequality follows from $\mathrm{Tr}(\mathcal{Q}^2) \asymp m \asymp n$ and$\|\mathcal{Q}^2\Delta\|_2 \le \|\Delta\|_2$ and the second inequality follows from (107). Hence, with probability larger than $1-(\log p)^{-1/2} - \frac{1}{t^2}$,

$$\left|\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2\Delta\right| \lesssim t\sqrt{\frac{q\log p}{np}}\sigma_e. \tag{108}$$

Since $\mathrm{Tr}(\mathcal{Q}^2) \asymp m \asymp n$ and$\|\mathcal{Q}\Delta\|_2 \le \|\Delta\|_2$, we have

$$\begin{aligned}\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}\|\mathcal{Q}\Delta + \mathcal{Q}X(\beta - \widehat{\beta}^{init}) + \mathcal{Q}Xb\|_2^2 &\lesssim \frac{1}{n}\|\mathcal{Q}\Delta\|_2^2 + \frac{1}{n}\|\mathcal{Q}X(\beta - \widehat{\beta}^{init})\|_2^2 + \frac{1}{n}\|\mathcal{Q}Xb\|_2^2 \\ &\lesssim \frac{q\log p}{p} + \frac{k\log p}{n} + \frac{1}{n}\|\mathcal{Q}Xb\|_2^2\end{aligned} \tag{109}$$

with probability larger than $1-(\log p)^{-1/2}$. We can establish that

$$\left|\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2X(\beta-\widehat{\beta}^{init})\right| \lesssim \|\frac{1}{n}e^\intercal\mathcal{Q}^2X\|_\infty\|\beta-\widehat{\beta}^{init}\|_1 \lesssim k\frac{\log p}{\min\{n,\sqrt{np/q}\}}+\left(\frac{\|\mathcal{Q}Xb\|_2}{\sqrt{n}}\right)^2. \tag{110}$$

Finally, we control $\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2Xb$, which has mean zero and variance

$$\mathbb{E}\left(\frac{1}{\mathrm{Tr}(\mathcal{Q}^2)}e^\intercal\mathcal{Q}^2Xb\right)^2 \lesssim \frac{1}{n^2}\sigma_e^2b^\intercal X^\intercal\mathcal{Q}^4Xb \leq \frac{\sigma_e^2}{n^2}\|X^\intercal\mathcal{Q}^2X\|_2\|b\|_2^2$$

and hence with probability larger than $1-\frac{1}{t^2}$,

$$\frac{1}{n}e^\intercal\mathcal{Q}^2Xb \lesssim \frac{t}{\sqrt{n}}\frac{1}{\sqrt{n}}\|\mathcal{Q}X\|_2\|b\|_2. \tag{111}$$

A combination of the decomposition (105) and the error bounds (106), (108), (109), (110), (111) and (37) leads to Proposition 1.

B.10 Proof of Proposition 5

We write $X=Z\Sigma_X^{\frac{1}{2}}$, where $X,Z\in\mathbb{R}^{n\times p}$ and $\Sigma_X^{\frac{1}{2}}\in\mathbb{R}^{p\times p}$ and the entries of Z are i.i.d Sub-gaussian random variables.

By Theorem 1.1 of [48], with probability larger than $1-c_1^n$ for a positive constant $c_1\in(0,1)$,

$$\lambda_{\min}\left(Z^\intercal Z\right) \gtrsim \left(\sqrt{n}-\sqrt{p}\right)^2. \tag{112}$$

Additionally, we have

$$\lambda_{\min}\left(Z^\intercal Z\right)=\lambda_{\min}\left(\Sigma_X^{-\frac{1}{2}}X^\intercal X\Sigma_X^{-\frac{1}{2}}\right) \leq \frac{1}{\lambda_{\min}(\Sigma_X)}\lambda_{\min}(X^\intercal X)$$

Combined with (112), we have

$$\lambda_{\min}(\frac{1}{n}X^\intercal X) \gtrsim \lambda_{\min}(\Sigma_X)\frac{\left(\sqrt{n}-\sqrt{p}\right)^2}{n} \gtrsim \lambda_{\min}(\Sigma_X).$$

By the definition of $\mathcal{Q}$ in (15), we have $\lambda_{\min}\left(\frac{1}{n}X^\intercal\mathcal{Q}^2X\right) \geq \lambda_{\min}(\frac{1}{n}X^\intercal X)$ and then establish the lower bound for $\lambda_{\min}\left(\frac{1}{n}X^\intercal\mathcal{Q}^2X\right)$.

B.11 Proof of Proposition 6

The proof relies on the results obtained in [50]. Specifically, we state the technical results adopted from [50] in the following two lemmas.

Lemma 9. *Suppose that $X_{i,\cdot}\in\mathbb{R}^p$ follows a Gaussian distribution with covariance matrix $\Sigma_X=\Gamma D^2\Gamma^\intercal$ for $1\leq i\leq n$. Then we have the eigen-decomposition $\mathbb{E}VV^\intercal\in\mathbb{R}^{p\times p}$ as $\Gamma C^2\Gamma^\intercal$ with*

$$\min_{1\leq l\leq p}C_{l,l}^2 \geq \min\left\{1,\frac{n}{p}\lambda_{\min}(\Sigma_X)\right\}. \tag{113}$$

This follows from Lemma 21 and equation (35) in [50] [1], together with the fact that $D_{l,l}^2 \geq \lambda_{\min}(\Sigma_X)$. The following Lemma is a statement of Lemma 14 in [50], in the terminology of the current paper.

Lemma 10. *Suppose that $X_{i,\cdot} \in \mathbb{R}^p$ follows a Gaussian distribution with covariance matrix $\Sigma_X = \Gamma D^2 \Gamma^\intercal$, $\lambda_{\max}(\Sigma_E) \lesssim p/(n \log p)$, $q \lesssim n/\log p$, and $p > cn$ for some $c > 1$. Then for any fixed $a, b \in \mathbb{R}^p$, we have that there exists $c_1, c_2 > 0$ such that*

$$\mathcal{P}\left(|a^\intercal V V^\intercal b - \mathbb{E} a^\intercal V V^\intercal b| \geq t\right) \lesssim \exp\left(-\frac{c_1 t^2 p^2}{\|\Sigma_X^{\frac{1}{2}} a\|_2^2 \|\Sigma_X^{\frac{1}{2}} b\|_2^2 n}\right) + e^{-c_2 n} + \frac{1}{np^3}. \tag{114}$$

In the following, we shall apply Lemmas 9 and 10 to verify the restricted eigenvalue conditions for $\frac{1}{n} V V^\intercal$. For any $\omega \in \mathbb{R}^p$, we have the decomposition

$$\omega^\intercal V V^\intercal \omega = \omega^\intercal (V V^\intercal - \mathbb{E} V V^\intercal)\omega + \omega^\intercal \mathbb{E} V V^\intercal \omega. \tag{115}$$

We note

$$\omega^\intercal \mathbb{E} V V^\intercal \omega \geq \min_{1 \leq l \leq p} C_{l,l} \|\omega\|_2^2 \gtrsim \min\left\{1, \frac{n}{p} \lambda_{\min}(\Sigma_X)\right\} \|\omega\|_2^2, \tag{116}$$

where the last inequality is a consequence of Lemma 9. We shall fix $\mathcal{T} \subset [p]$ and $|\mathcal{T}| \leq k$. Then we have

$$\begin{aligned}
&\max_{\|\omega_{\mathcal{T}^c}\|_1 \leq C\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2 = 1} \omega^\intercal (V V^\intercal - \mathbb{E} V V^\intercal)\omega \\
&\leq \|V V^\intercal - \mathbb{E} V V^\intercal\|_\infty \max_{\|\omega_{\mathcal{T}^c}\|_1 \leq C\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2 = 1} \|\omega\|_1^2 \\
&\leq \|V V^\intercal - \mathbb{E} V V^\intercal\|_\infty (1 + C)^2 \max_{\|\omega\|_2 = 1} \|\omega_{\mathcal{T}}\|_1^2 \\
&\leq (1 + C)^2 k \|V V^\intercal - \mathbb{E} V V^\intercal\|_\infty.
\end{aligned} \tag{117}$$

Define

$$\mathrm{RE}\left(\frac{1}{n} V V^\intercal\right) = \inf_{\substack{\mathcal{T} \subset [p] \\ |\mathcal{T}| \leq k}} \min_{\substack{\omega \in \mathbb{R}^p \\ \|\omega_{\mathcal{T}^c}\|_1 \leq C\|\omega_{\mathcal{T}}\|_1}} \frac{\omega^\intercal \left(\frac{1}{n} V V^\intercal\right) \omega}{\|\omega\|_2^2}.$$

Then we combine (116) and (117) and establish

$$\mathrm{RE}\left(\frac{1}{n} V V^\intercal\right) \geq \min\left\{1, \frac{n}{p} \lambda_{\min}(\Sigma_X)\right\} - (1 + C)^2 k \|V V^\intercal - \mathbb{E} V V^\intercal\|_\infty \tag{118}$$

By applying (114), we take $t = C \frac{n}{p} \sqrt{\frac{\log p}{n}} \|\Sigma_X\|_\infty$, then with probability larger than $1 - p^{-c_1} - \exp(-c_1 n)$ for some positive constant $c_1 > 0$,

$$\|V V^\intercal - \mathbb{E} V V^\intercal\|_\infty \lesssim \frac{n}{p} \sqrt{\frac{\log p}{n}} \|\Sigma_X\|_\infty.$$

[1] In the current proof, the reference numbers of lemmas and equations of [50] correspond to those in the version `https://arxiv.org/pdf/1811.01076v3.pdf`.

Combined with (117), we establish that

$$\max_{\|\omega_{\mathcal{T}^c}\|_1 \leq C\|\omega_{\mathcal{T}}\|_1, \|\omega\|_2=1} \omega^\intercal (VV^\intercal - \mathbb{E}VV^\intercal)\omega \lesssim k\frac{n}{p}\sqrt{\frac{\log p}{n}}\|\Sigma_X\|_\infty.$$

Together with (118), we establish that, with probability larger than $1-p^{-c_1}-\exp(-c_1 n)$,

$$\mathrm{RE}\left(\frac{1}{n}VV^\intercal\right) \geq \min\left\{1, \frac{n}{p}\lambda_{\min}(\Sigma_X)\right\} - k\frac{n}{p}\sqrt{\frac{\log p}{n}}\|\Sigma_X\|_\infty. \tag{119}$$

In the following, we connect $\mathrm{RE}\left(\frac{1}{n}X^\intercal \mathcal{Q}^2 X\right)$ to $\mathrm{RE}\left(\frac{1}{n}VV^\intercal\right)$. Since

$$\omega^\intercal \left(\tfrac{1}{n}X^\intercal \mathcal{Q}^2 X\right)\omega = \omega^\intercal \left(\tfrac{1}{n}VS^2\Lambda^2 V^\intercal\right)\omega \geq \min_{1\leq l\leq n}(S_{l,l}\Lambda_{l,l})^2 \frac{1}{n}\|V^\intercal\omega\|_2^2$$

we have

$$\mathrm{RE}\left(\tfrac{1}{n}X^\intercal \mathcal{Q}^2 X\right) \geq \min_{1\leq l\leq n}(S_{l,l}\Lambda_{l,l})^2 \cdot \mathrm{RE}\left(\frac{1}{n}VV^\intercal\right) \tag{120}$$

We note that

$$\min_{1\leq l\leq n}(S_{l,l}\Lambda_{l,l})^2 \geq \lambda_n(\frac{1}{n}X^\intercal X) = \lambda_{\min}(\frac{1}{n}XX^\intercal). \tag{121}$$

With $Z = X\Sigma_X^{-\frac{1}{2}}$, we have

$$\lambda_{\min}\left(ZZ^\intercal\right) = \lambda_{\min}\left(X\Sigma_X^{-1}X^\intercal\right) \leq \frac{1}{\lambda_{\min}(\Sigma_X)}\lambda_{\min}(XX^\intercal)$$

By Theorem 1.1 of [48], with probability larger than $1-c_2^n-\exp(-c_3 p)$ for positive constants $c_2 \in (0,1)$ and $c_3 > 0$, $\lambda_{\min}\left(ZZ^\intercal\right) \gtrsim \left(\sqrt{p}-\sqrt{n}\right)^2$. Hence, we have

$$\lambda_n(\frac{1}{n}XX^\intercal) \gtrsim \frac{p}{n}\lambda_{\min}(\Sigma_X).$$

Together with (120) and (121), we establish

$$\mathrm{RE}\left(\frac{1}{n}X^\intercal \mathcal{Q}^2 X\right) \gtrsim \frac{p}{n}\lambda_{\min}(\Sigma_X)\cdot \mathrm{RE}(\frac{1}{n}VV^\intercal).$$

Combined with (119), we establish the proposition.