The Annals of Statistics
2022, Vol. 50, No. 3, 1320–1347
https://doi.org/10.1214/21-AOS2152

DOUBLY DEBIASED LASSO: HIGH-DIMENSIONAL INFERENCE UNDER HIDDEN CONFOUNDING

BY ZIJIAN GUO[1,a], DOMAGOJ ĆEVID[2,b] AND PETER BÜHLMANN[2,c]

[1]*Department of Statistics, Rutgers University,* [a]*zijguo@stat.rutgers.edu*
[2]*Seminar for Statistics, ETH Zürich,* [b]*cevid@stat.math.ethz.ch,* [c]*buhlmann@stat.math.ethz.ch*

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, that is, that every confounding variable affects many covariates. The finite sample performance is illustrated with an extensive simulation study and a genomic application. The method is implemented by the `DDL` package available from CRAN.

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 [28, 45]. 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 $H_{i,\cdot} \in \mathbb{R}^q$:

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

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 example, in biological studies to explore the effects of measured genetic variants on the disease risk factor, and the hidden confounders can be geographic information [49], data sources in mental analysis [52] or general population stratification in GWAS [46].

Our aim is to perform statistical inference for individual components β_j, $1 \le j \le p$, 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 [11, 64].

Received October 2020; revised November 2021.
MSC2020 subject classifications. Primary 62E20, 62F12; secondary 62J07.
Key words and phrases. Causal inference, structural equation model, dense confounding, linear model, spectral deconfounding.

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, for example, for addressing the problem of batch effects in biological studies [31, 36, 41].

We propose a two-step estimator for the regression coefficient β_j for $1 \le j \le p$ in the high-dimensional dense confounding setting, where a large number of covariates have possibly been affected by hidden confounding. In the first step, we construct a penalized spectral deconfounding estimator $\widehat{\beta}^{\text{init}}$ as in [12], 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 $\beta_{-j} = \{\beta_l; l \ne j\}$ by $\widehat{\beta}^{\text{init}}_{-j}$ 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 $\widehat{\beta}^{\text{init}}_{-j}$ and the other arising from hidden confounding. In the standard high-dimensional regression setting with no hidden confounding, debiasing, desparsifying or Neyman's orthogonalization were proposed for inference for β_j [4, 14, 16, 22, 34, 58, 66]. 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 confounding [33, 58]. 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 [6, 64]. 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 [29]. Finally, the performance of the proposed estimator is illustrated on simulated and real genomic data in Section 5.

To summarize, our main contribution is twofold:

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 [8, 30, 32, 37, 63, 64]. 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" [64]. Some recent work on the high-dimensional hidden confounding problem relying on the construction of IVs includes [3, 19, 24, 26, 43, 48, 68].

Another approach builds on directly estimating and adjusting with respect to latent factors [62].

A major distinction of the current work from the contributions above is that we consider a confounding model with a denseness assumption [12, 13, 55]. [12] consider point estimation of β in the high-dimensional hidden confounding model (1), whereas [55] 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 [18, 21, 39, 40, 61]. 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 [23, 57, 60] 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 jth column of the matrix X and the submatrix of X excluding the jth column, respectively; $X_{i,\cdot} \in \mathbb{R}^p$ is used to denote the ith 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 subrow of $X_{i,\cdot}$ excluding the jth entry. Let $[p] = \{1, 2, \ldots, p\}$. For a subset $J \subseteq [p]$ and a vector $x \in \mathbb{R}^p$, x_J is the subvector of x with indices in J and x_{-J} is the subvector 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 $\|x\|_q = (\sum_{l=1}^p |x_l|^q)^{\frac{1}{q}}$ 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 sub-Gaussian random variable X, we use $\|X\|_{\psi_2}$ to denote its sub-Gaussian norm; see definitions 5.7 and 5.22 in [59]. 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 $\limsup_{n\to\infty} a_n/b_n = 0$. For a matrix M, we use $\|M\|_F$, $\|M\|_2$ and $\|M\|_\infty$ to denote its Frobenius norm, spectral norm and elementwise maximum norm, respectively. We use $\lambda_j(M)$ to denote the jth largest singular value of some matrix M, that is, $\lambda_1(M) \geq \lambda_2(M) \geq \cdots \geq \lambda_q(M) \geq 0$. For a symmetric matrix A, we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote its maximum and minimum eigenvalues, respectively.

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, respectively. 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 a small positive integer. 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 cannot 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), that is, 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 $\sigma_\epsilon^2 = \sigma_e^2 + \phi^\intercal(\mathrm{I}_q - \Psi\Sigma_X^{-1}\Psi^\intercal)\phi$. 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 $\beta + b$, with the perturbation vector b induced by hidden confounding. However, as shown in Lemma 2 in the Supplementary Material [29], b is dense and $\|b\|_2$ is 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:

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

that is, 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

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

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 $j \in \{1, \dots, p\}$, 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 $\widehat{\beta}^{\text{init}}$ an initial estimator of β. We will use the spectral deconfounding estimator proposed in [12], described in detail in Section 3.4. We start from the following decomposition:

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

The above decomposition reveals two sources of bias: the bias $X_{-j}(\beta_{-j} - \widehat{\beta}^{\text{init}}_{-j})$ due to the error of the initial estimator $\widehat{\beta}^{\text{init}}$ 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 Supplementary Material [29]. 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 [34, 58, 66] 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 subdesign $X_{-j} \in \mathbb{R}^{n\times(p-1)}$. 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

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

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 $\mathcal{P}^{(j)}$:

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

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:

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

We refer to this estimator as the doubly debiased lasso estimator as it simultaneously corrects the bias induced by $\widehat{\beta}^{\text{init}}$ 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):

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

In the above equation, the bias after correction consists of two components: the remaining bias due to the estimation error of $\widehat{\beta}^{\text{init}}_{-j}$ 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 $\frac{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_{-j} (\beta_{-j} - \widehat{\beta}^{\text{init}}_{-j})}{(\mathcal{P}^{(j)} Z_j)^\intercal \mathcal{P}^{(j)} X_j}$ in (11) is small since the initial estimator $\widehat{\beta}^{\text{init}}$ 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 [66] for the standard high-dimensional linear regression:

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

where Z^{DB}_j is constructed similarly as in (8) and (9) with setting $\mathcal{P}^{(j)}$ as the identity matrix. Therefore, the main difference between the estimator in (12) and our proposed estimator (10) is that 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 $\widehat{\beta}^{\text{init}}_{-j}$ 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^{DB}_j). 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^{DB}_j 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^E (see the expression of γ^E in Lemma 1); second, the application of $\mathcal{P}^{(j)}$ leads to a small prediction error $\mathcal{P}^{(j)}X_{-j}(\widehat{\gamma}-\gamma^E)$. We illustrate in Section 5 how the application of $\mathcal{P}^{(j)}$ corrects the bias due to $\widehat{\beta}_{-j}^{\text{init}}$ and observe a better empirical coverage after applying $\mathcal{P}^{(j)}$ in comparison to the standard debiased lasso in (12); see Figure 7.

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 $\sqrt{\frac{\widehat{\sigma}_e^2\cdot Z_j^\top(\mathcal{P}^{(j)})^4Z_j}{[Z_j^\top(\mathcal{P}^{(j)})^2X_j]^2}}$ 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

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

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 $X_{-j}=U(X_{-j})\Lambda(X_{-j})[V(X_{-j})]^\top$, 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 $\Lambda_{1,1}(X_{-j})\geq\Lambda_{2,2}(X_{-j})\geq\cdots\geq\Lambda_{m,m}(X_{-j})\geq 0$. We then define the spectral transformation $\mathcal{P}^{(j)}$ for X_{-j} as $\mathcal{P}^{(j)}=U(X_{-j})S(X_{-j})[U(X_{-j})]^\top$, 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 $\mathcal{P}^{(j)}X_{-j}=U(S\Lambda)V^\top$, so the spectral transformation shrinks the singular values $\{\Lambda_{l,l}\}_{1\leq l\leq m}$ to $\{S_{l,l}\Lambda_{l,l}\}_{1\leq l\leq m}$, 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 [12]. It limits any singular value to be at most some threshold τ. This means that the shrinkage matrix S is given as: for $1\leq l\leq m$,

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

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

$$\begin{aligned}\mathcal{P}^{(j)}&=U(X_{-j})S(X_{-j})\big[U(X_{-j})\big]^\top \quad\text{with}\\ S_{l,l}(X_{-j})&=\begin{cases}\dfrac{\Lambda_{\lfloor\rho_j m\rfloor,\lfloor\rho_j m\rfloor}(X_{-j})}{\Lambda_{l,l}(X_{-j})} & \text{if } l\leq\lfloor\rho_j m\rfloor,\\ 1 & \text{otherwise.}\end{cases}\end{aligned} \tag{14}$$

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 [29], we describe the general conditions that the used spectral transformations need to satisfy to ensure good performance of the resulting estimator.

Other constructions of spectral transformations include the spectral transformation induced by the LAVA estimator [15], the Puffer transformation [35], and the PCA adjustment [50]. See more detailed discussions in Section 3.2.1 in [12].

3.4. *Initial estimator $\widehat{\beta}^{init}$.* For the doubly debiased lasso (10), we use the spectral deconfounding estimator proposed in [12] as our initial estimator $\widehat{\beta}^{\text{init}}$. 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

$$\mathcal{Q} = U(X)S(X)\big[U(X)\big]^{\mathsf{T}} \quad \text{with } S_{l,l}(X) = \begin{cases} \dfrac{\Lambda_{\lfloor \rho m\rfloor,\lfloor \rho m\rfloor}(X)}{\Lambda_{l,l}(X)} & \text{if } l \le \lfloor \rho m\rfloor, \\ 1 & \text{otherwise.} \end{cases} \tag{15}$$

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

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

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 $\widehat{\beta}^{\text{init}}$ are given, thereby extending the results of [12].

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 [54]. We propose using the following estimator:

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

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

The motivation for this estimator is based on the expression

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

which follows from the model (3). The consistency of the proposed noise level estimator, formally shown in Proposition 2, follows from the following observations: the initial spectral deconfounding estimator $\widehat{\beta}^{\text{init}}$ 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; $\|\mathcal{Q}\epsilon\|_2^2/\operatorname{Tr}(\mathcal{Q}^2)$ 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 Supplementary Material [29]. In Section 4, we show that variance estimator $\widehat{\sigma}_e^2$ defined in (17) is a consistent estimator of σ_e^2.

Algorithm 1 Doubly debiased lasso

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

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

$\mathcal{Q} \leftarrow \textsc{TrimTransform}(X, \rho)$ ▷ construct ρ-trim as in (15)

$\widehat{\beta}^{\mathrm{init}} \leftarrow \textsc{Lasso}(\mathcal{Q}X, \mathcal{Q}Y, \lambda)$ ▷ 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

$\widehat{\beta}_j \leftarrow \textsc{DebiasedLasso}(\widehat{\beta}^{\mathrm{init}}, \mathcal{P}^{(j)}X_{-j}, \mathcal{P}^{(j)}X_j, \mathcal{P}^{(j)}Z_j)$ ▷ compute doubly debiased lasso as in (12)

$\widehat{\sigma}_\varepsilon^2 \leftarrow \textsc{NoiseLevel}(X, Y, \widehat{\beta}^{\mathrm{init}}, \mathcal{Q})$ ▷ compute noise level as in (17)

$\mathrm{CI}(\beta_j) \leftarrow \textsc{ConfidenceInterval}(\widehat{\beta}_j, \mathcal{P}^{(j)}X_j, \mathcal{P}^{(j)}Z_j, \widehat{\sigma}_\varepsilon^2, \alpha)$ ▷ compute the $(1-\alpha)$-CI as in (13)

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, similar to the standard debiased lasso [66], our proposed method involves the regularization parameters λ in the lasso regression for the initial estimator $\widehat{\beta}^{\mathrm{init}}$ (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 tenfold cross-validation, whereas for λ_j, we increase slightly the penalty chosen by the tenfold cross-validation, so that the variance of our estimator, which can be determined from the data up to a proportionality factor σ_ε^2, increases by 25%, as proposed in [17].

The proposed doubly debiased lasso method is summarized in Algorithm 1, which also highlights where each tuning parameter is used.

4. Theoretical justification. This section provides theoretical justifications of the proposed method for the hidden confounding model (2). The proofs of the main results are presented in Sections A and B in the Supplementary Material [29] 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 jth column and $\Psi_{-j} \in \mathbb{R}^{q\times(p-1)}$ denotes the submatrix 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

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

For ease of notation, we do not explicitly express the dependence of γ on j. Similarly, define

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

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

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

(A1) The precision matrix $\Omega_E = [\mathbb{E}(E_{i,\cdot}E_{i,\cdot}^\top)]^{-1}$ 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 $C_0 > 0$ 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 [9, 38, 47, 65] and are also used for confidence interval construction in the standard high-dimensional regression model without unmeasured confounding [58]. Here, the conditions are not directly imposed on the covariates $X_{i,\cdot}$, but rather on their unconfounded part $E_{i,\cdot}$. In the high-dimensional linear model without hidden confounders, the sparse precision matrix assumption can be relaxed using the technique in [34]. However, it is unclear whether the method in [34] can be generalized to our model due to the additional hidden confounding bias as in (11).

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 qth singular value of the coefficient matrix $\Psi_{-j} \in \mathbb{R}^{q\times(p-1)}$ satisfies

$$\lambda_q(\Psi_{-j}) \gg l(n,p,q) := \max\left\{M\sqrt{\frac{qp}{n}}(\log p)^{3/4}, \sqrt{Mq}\,p^{1/4}(\log p)^{3/8}, \sqrt{qn\log p}\right\}, \tag{19}$$

where M is the sub-Gaussian norm for components of $X_{i,\cdot}$, as defined in Assumption (A3). Furthermore, we have

$$\max\{\|\Psi(\Omega_E)_{\cdot,j}\|_2, \|\Psi_j\|_2, \|\Psi_{-j}(\Omega_E)_{-j,j}\|_2, \|\phi\|_2\} \lesssim \sqrt{q}(\log p)^c, \tag{20}$$

where Ψ and ϕ are defined in (2) and $0 < c \le 1/4$ 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 (20) on their ℓ_2 norms is mild. If the vector $\phi \in \mathbb{R}^q$ has bounded entries and the vectors $\{\Psi_{\cdot,l}\}_{1\le l\le p} \in \mathbb{R}^q$ are independently generated with zero mean and bounded second moments, then the condition (20) holds with probability larger than $1-(\log p)^{-2c}$, where c is defined in (20). A larger value $c > 1/4$ is possible: the condition then holds with even higher probability, but makes the upper bounds for (32) in Lemma 1 and (35) in Lemma 2 in the Supplementary Material [29] slightly worse, which then requires more stringent conditions on $\lambda_q(\Psi_{-j})$ in Theorem 1, up to polynomial order of $\log p$.

In the factor model literature [20, 61], the spiked singular value condition $\lambda_q(\Psi) \asymp \sqrt{p}$ 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 $X_{i,\cdot}$. However, for our analysis, our assumption on $\lambda_q(\Psi_{-j})$ in (19) can be much weaker than the classical factor assumption $\lambda_q(\Psi_{-j}) \asymp \sqrt{p}$, especially for a range of dimensionality where $p \gg n$. In certain dense confounding settings, we can show that condition (19) holds with high probability. 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 $\lambda_1(\Psi_{-j})$. The condition (19) can be satisfied even if only a certain proportion of covariates is affected by hidden confounding. When $q = 1$, if we assume that there exists a set $A \subseteq \{1,2,\dots,p\}$ such that $\{\Psi_l\}_{l\in A}$ are i.i.d. and $|A| \gg l(n,p,q)^2$, where $l(n,p,q)$ is defined in (19), then with high probability $\lambda_q(\Psi) \gtrsim \sqrt{|A|} \gg l(n,p,q)$. In the multiple hidden confounders setting, if the vectors $\{\Psi_l\}_{l\in A}$ 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 (19) is satisfied with high

probability as well. See Lemmas 4 and 5 in Section A.5 of the Supplementary Material [29] for the exact statement. In Section 5.1, we also explore the numerical performance of the method when different proportions of the covariates are affected and observe that the proposed method works well even if the hidden confounders only affect a small percentage of the covariates, say 5%.

Under the model (2), if the entries of Ψ are assumed to be i.i.d. sub-Gaussian with zero mean and variance σ_Ψ^2, then we have $\lambda_q(\Psi_{-j}) \asymp \sqrt{p}\sigma_\Psi$ with high probability. Together with (19), this requires

$$\sigma_\Psi \gg \max\left\{M\sqrt{\frac{q}{n}}(\log p)^{3/4}, \sqrt{\frac{qn\log p}{p}}, \frac{\sqrt{qM(\log p)^{3/4}}}{p^{1/4}}\right\}.$$

So if $p \gg qn\log p$ and $\min\{n, p\} \gg q^3(\log p)^{3/2}M^2$, then the required effect size σ_Ψ of the hidden confounder $H_{i,\cdot}$ on an individual covariate $X_{i,j}$ can diminish to zero fairly quickly.

The condition (19) can in fact be empirically checked using the sample covariance matrix $\widehat{\Sigma}_X$. Since $\Sigma_X = \Psi^\top\Psi + \Sigma_E$, then the condition (19) implies that Σ_X has at least q spiked eigenvalues. 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 [61]. 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 for an illustration.

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 $(X_{i,\cdot}^\top, H_{i,\cdot}^\top)^\top$, the error vector $E_{i,\cdot}$ is assumed to be independent of the hidden confounder $H_{i,\cdot}$, and the noise term $v_{i,j} = E_{i,j} - E_{i,-j}^\top\gamma^E$ is assumed to be independent of $E_{i,-j}$. Furthermore, $E_{i,\cdot}$ is a sub-Gaussian random vector and e_i and $v_{i,j}$ are sub-Gaussian random variables, whose sub-Gaussian norms satisfy $\max\{\|E_{i,\cdot}\|_{\psi_2}, \|e_i\|_{\psi_2}, \max_{1\le l\le p}\|v_{i,l}\|_2\} \le C$, where $C > 0$ is a positive constant independent of n and p. For $1 \le l \le p$, $X_{i,l}$ are sub-Gaussian random variables whose sub-Gaussian norms satisfy $\max_{1\le l\le p}\|X_{i,l}\|_{\psi_2} \le M$, where $1 \lesssim M \lesssim \sqrt{q\log p}$.

The independence assumption between the random error e_i and $(X_{i,\cdot}^\top, H_{i,\cdot}^\top)^\top$ is commonly assumed for the SEM (1), and thus it holds in the induced hidden confounding model (2) as well; see, for example, [51]. Analogously, when modeling $X_{i,\cdot}$ as a SEM where the hidden variables $H_{i,\cdot}$ are directly influencing $X_{i,\cdot}$, that is, they are parents of the $X_{i,\cdot}$'s, the independence of $E_{i,\cdot}$ from $H_{i,\cdot}$ is a standard assumption. The independence assumption between $v_{i,j}$ and $E_{i,-j}$ holds automatically if $E_{i,\cdot}$ has a multivariate Gaussian distribution (but $X_{i,\cdot}$ is still allowed to be non-Gaussian, e.g., due to non-Gaussian confounders). Additionally, the independence assumption between $v_{i,j}$ and $E_{i,-j}$ holds if all elements of $E_{i,\cdot}$ are independent, but not necessarily Gaussian. In Appendix D, we explore the robustness of our proposed doubly debiased lasso estimator to the violation of this independence assumption; see Figure A2 for details.

We emphasize that the individual components $X_{i,j}$ are assumed to be sub-Gaussian, instead of the whole vector $X_{i,\cdot} \in \mathbb{R}^p$. The sub-Gaussian norm M is allowed to grow with q and p. Particularly, if we assume $H_{i,\cdot}$ to be a sub-Gaussian vector, then condition (20) implies that $M \lesssim \sqrt{q}(\log p)^c\|H_{i,\cdot}\|_{\psi_2}$. Furthermore, our theoretical analysis also covers the case when the sub-Gaussian norm M is of constant order. This happens, for example, when the entries of Ψ are of order $1/\sqrt{q}$, since $M \asymp \max_{l=1,\dots,p}\|\Psi_l\|_2$.

The final assumption is that the restricted eigenvalue condition [5] 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

$$\text{RE}\Big(\frac{1}{n}X^\intercal\mathcal{Q}^2X\Big)=\inf_{\substack{\mathcal{T}\subseteq[p]\\|\mathcal{T}|\le k}}\min_{\substack{\omega\in\mathbb{R}^p\\\|\omega_{\mathcal{T}^c}\|_1\le CM\|\omega_{\mathcal{T}}\|_1}}\frac{\omega^\intercal(\frac{1}{n}X^\intercal\mathcal{Q}^2X)\omega}{\|\omega\|_2^2}\ge\tau_*; \tag{21}$$

$$\text{RE}\Big(\frac{1}{n}X_{-j}^\intercal(\mathcal{P}^{(j)})^2X_{-j}\Big)=\inf_{\substack{\mathcal{T}\subseteq[p]\setminus\{j\}\\|\mathcal{T}|\le s}}\min_{\substack{\omega\in\mathbb{R}^{p-1}\\\|\omega_{\mathcal{T}^c}\|_1\le CM\|\omega_{\mathcal{T}}\|_1}}\frac{\omega^\intercal(\frac{1}{n}X_{-j}^\intercal(\mathcal{P}^{(j)})^2X_{-j})\omega}{\|\omega\|_2^2}\ge\tau_*, \tag{22}$$

where $c, C, \tau_* > 0$ are positive constants independent of n and p and M is the sub-Gaussian norm for components of $X_{i,\cdot}$, as defined in Assumption (A3). For ease of notation, the same constants τ_* and C are used in (21) and (22).

Such assumptions are common in the high-dimensional statistics literature; see [7]. The restricted eigenvalue condition (A4) is similar, but more complicated than the standard restricted eigenvalue condition introduced in [5]. 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. In the following, we verify the restricted eigenvalue condition (A4) for $\frac{1}{n}X^\intercal\mathcal{Q}^2X$ and the argument can be extended to $\frac{1}{n}X_{-j}^\intercal(\mathcal{P}^{(j)})^2X_{-j}$.

PROPOSITION 1. *Suppose that assumptions* (A1) *and* (A3) *hold*, $H_{i,\cdot}$ *is a sub-Gaussian random vector*, $q+\log p\lesssim\sqrt{n}$ *and* $k=\|\beta\|_0$ *satisfies* $M^2kq^2\log p\log n/n\to 0$. *Assume further that the loading matrix* $\Psi\in\mathbb{R}^{q\times p}$ *satisfies* $\|\Psi\|_\infty\lesssim\sqrt{\log(qp)}$, $\lambda_1(\Psi)/\lambda_q(\Psi)\lesssim 1$ *and that*

$$\lambda_q(\Psi)\gg\frac{\sqrt{Mp}\max\{k^{1/4}q^{5/4},1\}\log(np)}{\min\{n,p\}^{1/4}}. \tag{23}$$

If $\lambda_{\lfloor\rho m\rfloor}(\frac{1}{n}XX^\intercal)\ge c\max\{1,p/n\}$ *for* ρ *defined in* (15) *and some positive constant* $c>0$ *independent of* n *and* p, *then there exist positive constants* $c_1, c_2>0$ *such that, with probability larger than* $1-p^{-c_2}-\exp(-c_2n)$, *we have* $\text{RE}(\frac{1}{n}X^\intercal\mathcal{Q}^2X)\ge c_1\lambda_{\min}(\Sigma_X)$.

An important condition for establishing Proposition 1 is the condition (23). Under the commonly assumed spiked singular value condition $\lambda_q(\Psi)\asymp\sqrt{p}$ [1, 2, 20, 61], the condition (23) is reduced to $k\ll\min\{n,p\}/(M^2q^5\log(np)^4)$. As a comparison, for the standard high-dimensional regression model with no hidden confounders, [53, 67] verified the restricted eigenvalue condition under the sparsity condition $k\ll n/\log p$. That is, if $\lambda_q(\Psi)\asymp\sqrt{p}$, then the sparsity requirement in Proposition 1 is the same as that for the high-dimensional regression model with no hidden confounders, up to a polynomial order of q and $\log(np)$,

In comparison to the condition (19) in (A2), (23) can be slightly stronger for a range of dimensionality where $p\gg n^{3/2}$. However, Proposition 1 does not require the strong spiked singular value condition $\lambda_q(\Psi)\asymp\sqrt{p}$. The proof of Proposition 1 is presented in Section B in the Supplementary Material [29]. The condition $\lambda_{\lfloor\rho m\rfloor}(\frac{1}{n}XX^\intercal)\ge c\max\{1,p/n\}$ can be empirically verified from the data. In Section B.1 in the Supplementary Material [29], further theoretical justification for this condition is provided, under mild assumptions.

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 $c^*=\lim p/n\in(0,\infty]$. We mention here that we also give some new results on point estimation of the initial estimator $\widehat{\beta}^{\text{init}}$ defined in (16) in Section A.3 in the Supplementary Material [29], as they are established under more general conditions than in [12].

4.2.1. *Asymptotic normality.* We first present the limiting distribution of the proposed doubly debiased lasso estimator. The proof of Theorem 1 and important intermediary results for establishing Theorem 1 are presented in Section A in the Supplementary Material [29].

THEOREM 1. *Consider the hidden confounding model* (2). *Suppose that conditions* (A1)–(A4) *hold and further assume that* $c^* = \lim p/n \in (0, \infty]$, $k := \|\beta\|_0 \ll \sqrt{n}/(M^3 \times \log p)$, $s := \|(\Omega_E)_{\cdot,j}\|_0 \ll n/(M^2 \log p)$ *and* $e_i \sim N(0, \sigma_e^2)$. *Let the tuning parameters for* $\widehat{\beta}^{\text{init}}$ *in* (16) *and* $\widehat{\gamma}$ *in* (9), *respectively, be* $\lambda \asymp \sigma_e\sqrt{\log p/n} + \sqrt{q \log p/\lambda_q^2(\Psi)}$ *and* $\lambda_j \asymp \sigma_j\sqrt{\log p/n} + \sqrt{q \log p/\lambda_q^2(\Psi_{-j})}$. *Furthermore, let* $\mathcal{Q}$ *and* $\mathcal{P}^{(j)}$ *be the Trim transform* (14) *with* $\min\{\rho, \rho_j\} \geq (q+1)/\min\{n, p-1\}$ *and* $\max\{\rho, \rho_j\} < 1$. *Then the doubly debiased lasso estimator* (10) *satisfies*

$$\frac{1}{\sqrt{V}}(\widehat{\beta}_j - \beta_j) \xrightarrow{d} N(0, 1), \tag{24}$$

where

$$\mathrm{V} = \frac{\sigma_e^2 Z_j^\intercal (\mathcal{P}^{(j)})^4 Z_j}{[Z_j^\intercal (\mathcal{P}^{(j)})^2 X_j]^2} \quad \text{and} \quad V^{-1} \frac{\sigma_e^2 \operatorname{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \operatorname{Tr}^2[(\mathcal{P}^{(j)})^2]} \xrightarrow{p} 1. \tag{25}$$

REMARK 1. The Gaussianity of the random error e_i 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 [27], 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. For constructing $\mathcal{Q}$ and $\mathcal{P}^{(j)}$, the main requirement is to trim the singular values enough in both cases, that is, $\min\{\rho, \rho_j\} \geq (q+1)/\min\{n, p-1\}$. 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 $\widehat{\beta}^{\text{init}}$ in (9) and (16), but hold for any transformation satisfying the conditions given in Section A.1 of the Supplementary Material [29] and any initial estimator satisfying the error rates presented in Section A.3 of the Supplementary Material [29].

REMARK 3. If we further assume the error ϵ_i in the model (3) to be independent of $X_{i,\cdot}$, then the requirement (19) of the condition (A2) can be relaxed to

$$\lambda_q(\Psi_{-j}) \gg \max\left\{ M\sqrt{\frac{qp}{n}}(\log p)^{3/4}, \sqrt{qM}\, p^{1/4}(\log p)^{3/8}, \sqrt{(sM^2 + k\sqrt{n}M^3)q \log p} \right\}.$$

Note that the factor model implies the upper bound $\lambda_q(\Psi_{-j}) \lesssim \sqrt{p}$. Even if $n \geq p$, the above condition on $\lambda_q(\Psi_{-j})$ can still hold if $p \gg kqM^3 \log p\sqrt{n}$. On the other hand, the condition (19) together with $\lambda_q(\Psi_{-j}) \lesssim \sqrt{p}$ imply that $p \gg qn \log p$, which excludes the setting $n \geq p$.

There are three conditions on the parameters s, q, k imposed in the Theorem 1 above. The most stringent one is the sparsity assumption $k \ll \sqrt{n}/[M^3 \log p]$. 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 [34, 58, 66] and has been established in [10] as a necessary condition for constructing adaptive confidence intervals. In the high-dimensional hidden confounding model with $M \asymp 1$, the condition on the sparsity of β is then of the same

asymptotic order as in the standard high-dimensional regression with no hidden confounding. The condition on the sparsity of the precision matrix, $s = \|(\Omega_E)_{\cdot,j}\|_0 \ll n/(M^2 \log p)$, is mild in the sense that, for $M \asymp 1$, it is the maximal sparsity level for identifying $(\Omega_E)_{\cdot,j}$. Implied by (19), 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 (25) 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 confounding) 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 [58]. The expression for the asymptotic variance V of our proposed estimator (10) is given by $\frac{\sigma_e^2 \operatorname{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \operatorname{Tr}^2[(\mathcal{P}^{(j)})^2]}$ (see Theorem 1). For the Trim transform defined in (14), which trims top $(100\rho_j)\%$ of the singular values, we have that

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

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

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

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

$$\frac{\sigma_e^2}{\sigma_j^2 n} \le \frac{\sigma_e^2 \operatorname{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \operatorname{Tr}^2[(\mathcal{P}^{(j)})^2]} \le \frac{1}{1-\rho_j} \cdot \frac{\sigma_e^2}{\sigma_j^2 n}. \tag{26}$$

THEOREM 2. *Suppose that the assumptions of Theorem 1 hold. If $p \ge n+1$ 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 $q \ll n$ regime, the doubly debiased lasso achieves the Gauss–Markov efficiency bound if $\rho_j = \rho_j(n) \to 0$ and $\min\{\rho, \rho_j\} \ge (q+1)/n$ (which is also a condition of Theorem 1). When using the median Trim transform, that is, $\rho_j = 1/2$, the bound in (26) implies that the variance of the resulting 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 6.

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:

$$\text{ARE} \in \left[\frac{1}{\min\{c^*, 1\}}, \frac{1}{(1-\rho^*)\min\{c^*, 1\}}\right], \tag{27}$$

where $\rho^* = \lim_{n\to\infty} \rho_j(n) \in [0, 1)$. The equation (27) reveals how the efficiency of the doubly debiased lasso is affected by the choice of the percentile $\rho_j = \rho_j(n)$ in transformation

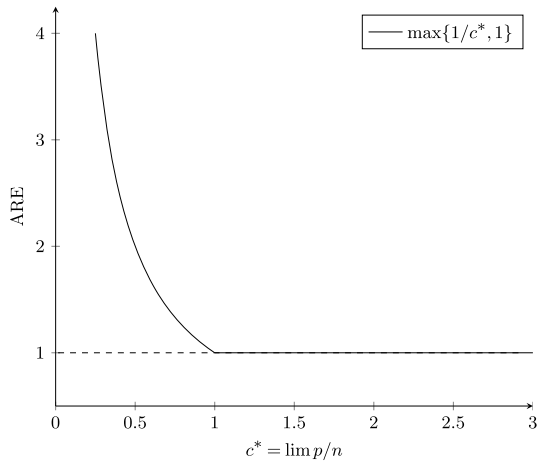

FIG. 1. *The plot of* ARE *versus* $c^* = \lim p/n$, *for the setting of* $\rho^* = 0$.

$\mathcal{P}^{(j)}$ 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 $c^* < 1$ (i.e., $p < n$), the relative efficiency of the proposed estimator increases as the dimension p increases and when $c^* \geq 1$ (i.e., $p \geq n$), we have that $\text{ARE} = 1$, 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 [64], 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 [64] 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 not only avoids the difficult step of coming up with 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 (25), 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 2. *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* $q \ll \min\{n, p/\log p\}$. *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* $0 < t \le \sqrt{n}$, *we have*

$$|\widehat{\sigma}_e^2 - \sigma_e^2| \lesssim \frac{t}{\sqrt{n}} + M^2 k \frac{\log p}{n} + \frac{q \log p}{p} + \frac{pq\sqrt{\log p}/n + M^2 kq \log p}{\lambda_q^2(\Psi)},$$

where M *is the sub-Gaussian norm for components of* $X_{i,\cdot}$ *defined in Assumption* (A3).

Together with (19) of the condition (A2), we apply the above proposition and establish $\widehat{\sigma}_e^2 - \sigma_e^2 \xrightarrow{p} 0$. As a remark, the estimation error $|\widehat{\sigma}_e^2 - \sigma_\epsilon^2|$ is of the same order of magnitude as $|\widehat{\sigma}_e^2 - \sigma_e^2|$ since the difference $\sigma_\epsilon^2 - \sigma_e^2$ is small in the dense confounding model; see Lemma 2 in the Supplementary Material [29].

Proposition 2, together with Theorem 1, imply 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*:

$$\liminf_{n,p\to\infty} \mathbb{P}\big(\beta_j \in \mathrm{CI}(\beta_j)\big) \ge 1 - \alpha, \tag{28}$$

$$\limsup_{n,p\to\infty} \mathbb{P}\left(\mathbf{L}\big(\mathrm{CI}(\beta_j)\big) \ge (2+c) z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_e^2 \operatorname{Tr}[(\mathcal{P}^{(j)})^4]}{\sigma_j^2 \operatorname{Tr}^2[(\mathcal{P}^{(j)})^2]}}\right) = 0, \tag{29}$$

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

Similar 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 (26), 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 $p \ge n + 1$, if we choose $\rho_j = \rho_j(n) \ge (q+1)/n$ and $\rho_j(n) \to 0$, the constructed confidence interval 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 transformation $\mathcal{P}^{(j)}$. 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, for example, by detecting the elbow in the scree plot [61], 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 [66], even though it

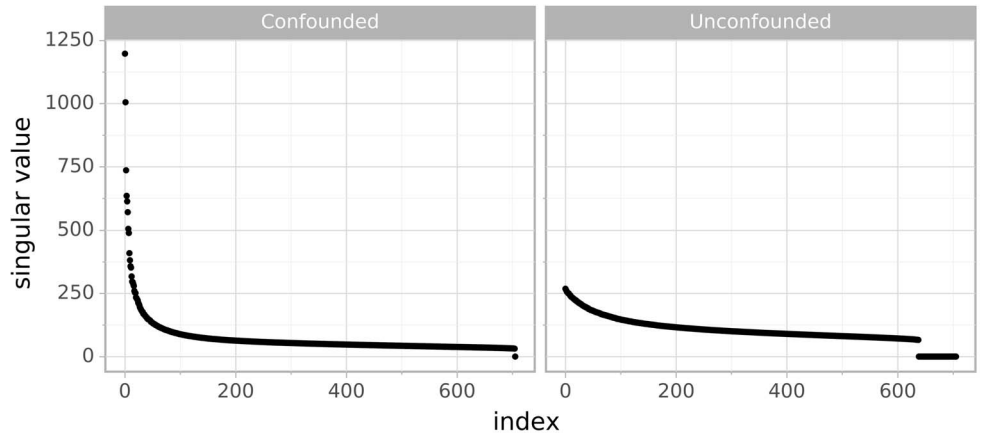

FIG. 2. *Left*: *Spiked singular values of the standardized gene expression matrix* (*see Section* 5.2) *indicate possible confounding*. *Right*: *Singular values after regressing out the* $q = 65$ *confounding proxies given in the data set* (*thus labeled as "unconfounded"*). *The singular values in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

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 behavior on a gene expression data set from the GTEx project [44].

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 $\widehat{\beta}^{\text{init}}$ 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 (25) and the bias terms B_β and B_b are given by

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

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. We present here a wide range of simulations settings and further simulations can be found in the Appendix D.

Simulation parameters. Unless specified otherwise, in all simulations we fix $q = 3$, $s = 5$ and $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)^\mathsf{T}$ 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 $p = 500$ and n 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 sample size is fixed at $n = 500$ and the number of covariates p varies from 100 to 2,000. We provide analogous simulations in Appendix D, where both the random variables and the model parameters are generated from non-Gaussian distributions.

We see that the absolute bias term $|B_b|$ due to confounding is substantially smaller for the 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 $\widehat{\beta}^{\text{init}}$. 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 the doubly debiased lasso, over a large range of n and p.

There are two challenging regimes for estimation under confounding. First, when the dimension p is much larger than the sample size n, the coverage can be lower than 95%, since

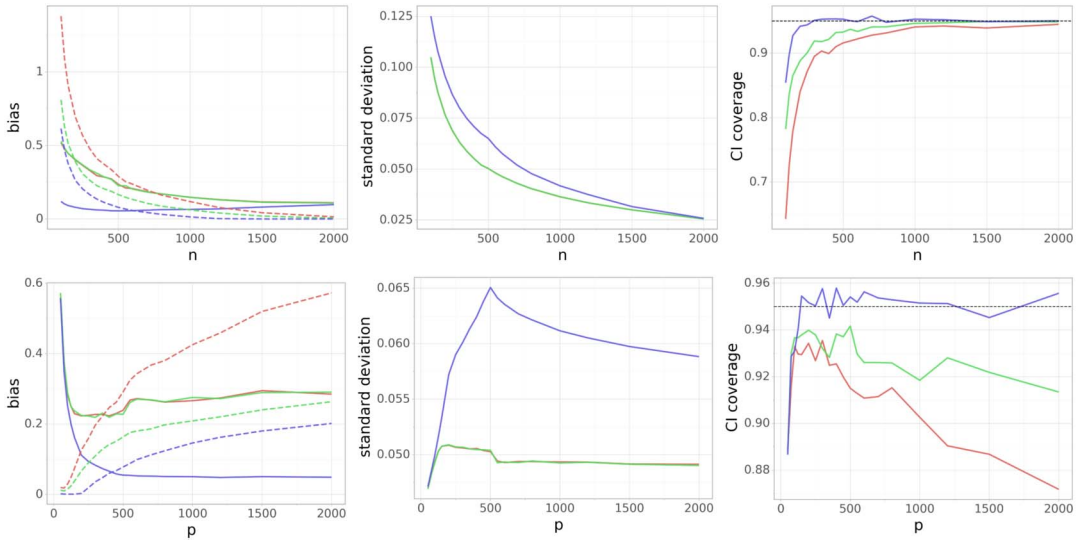

FIG. 3. *(Varying dimensions) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of data points n (top row) and the number of covariates p (bottom row). On the left 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 $\widehat{\beta}^{\text{init}}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.*

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. Second, 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 (52) in the Supplementary Material [29], where the term $|B_b|$ vanishes as $\lambda_q(\Psi)$ increases, in addition to increasing the sample size n. In the regime considered in this simulation, $|B_b|$ can 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, the 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.

Toeplitz covariance structure for Σ_E. Now we fix $n = 300$, $p = 1{,}000$, 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 4, we can see that the 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.

Proportion of confounded covariates. In order to investigate how the confounding denseness affects the performance of our method, we now 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 nonzero 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 of Ψ from 5% to 100%. The results can be seen in Figure 5. We can see that the 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 standard debiased lasso is poor even for a small number of affected variables and it worsens as the confounding

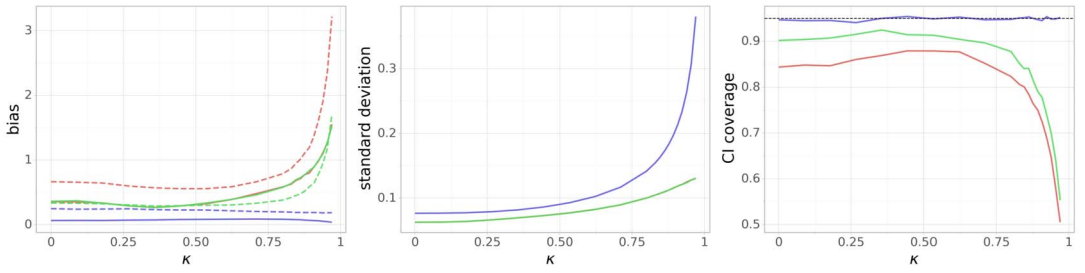

FIG. 4. *(Toeplitz covariance for* Σ_E*) Dependence of the (scaled) absolute bias terms* $|B_\beta|$ *and* $|B_b|$ *(left), standard deviation* $V^{1/2}$ *(middle) and the coverage of the 95% confidence interval (right) on the parameter* κ *of the Toeplitz covariance structure.* $n = 300$ *and* $p = 1{,}000$ *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* $\widehat{\beta}^{\text{init}}$ *as our proposed method. Note that the last two methods have almost indistinguishable* $|B_b|$ *and* V.

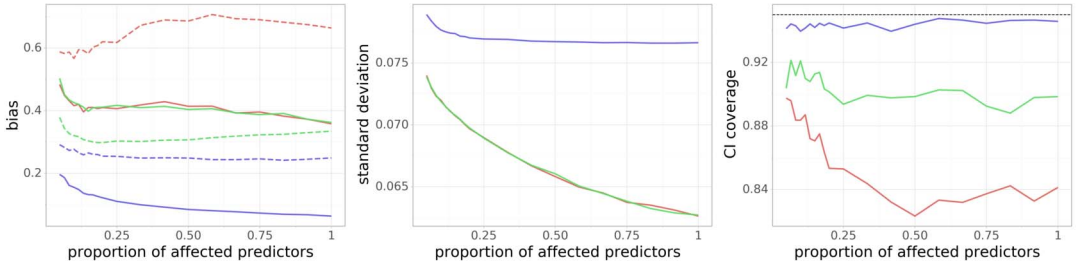

FIG. 5. *(Proportion confounded) Dependence of the* (*scaled*) *absolute bias terms* $|B_\beta|$ *and* $|B_b|$ (*left*), *standard deviation* $V^{1/2}$ (*middle*) *and the coverage of the* 95% *confidence interval* (*right*) *on proportion of confounded covariates.* $n = 300$ *and* $p = 1{,}000$ *are 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* $\widehat{\beta}^{\text{init}}$ *as our proposed method. Note that the last two methods have almost indistinguishable* $|B_b|$ *and* V.

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.

In Appendix D, we also show how the performance changes with the strength of confounding, by gradually decreasing the size of the entries of the loading matrix Ψ.

Trimming level. We investigate here 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 $\widehat{\beta}^{\text{init}}$ 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 value, so there is no shrinkage and our estimator reduces to the standard debiased lasso (with the initial estimator $\widehat{\beta}^{\text{init}}$). The results are displayed in Figure 6. 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, that is, 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. In Appendix D, we also explore whether the choice of spectral transformation significantly af-

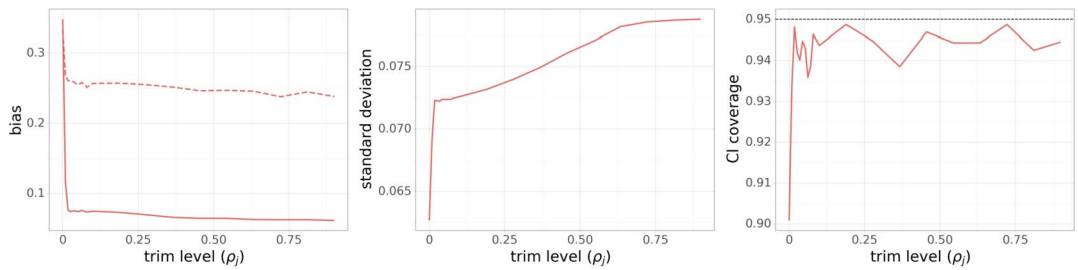

FIG. 6. *(Trimming level) Dependence of the* (*scaled*) *absolyte bias terms* $|B_\beta|$ *and* $|B_b|$ (*left*), *standard deviation* $V^{1/2}$ (*middle*) *and the coverage of the* 95% *confidence interval* (*right*) *on the trimming level* ρ_j *of the Trim transform* (*see equation* (14)). *The sample 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* $\widehat{\beta}^{\text{init}}$, *described in* (16).

fects the performance, with a focus on the PCA adjustment, which maps first several singular values to 0, while keeping the others intact.

No confounding bias. We consider now the same simulation setting as in Figure 3, 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 the top row of Figure 7. 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 equals 0, since we have put $b = 0$. We can even observe a decrease in estimation bias for large p, and thus an improvement in the confidence interval coverage. This is due to the fact that X has a spiked covariance structure and trimming the large singular values reduces the correlations between the predictors. This phenomenon is also illustrated in the additional simulations in the Appendix D, where we set $q = 0$ and put E to have either Toeplitz or equicorrelation covariance structure with varying degree of spikiness (by varying the correlation parameters).

In the bottom row of Figure 7, we repeat the same simulation, but where we set $q = 0$ and take $\Sigma_X = \Sigma_E = I$ in order to investigate the performance of the method in the setting without confounding, but where the covariance matrix of the predictors is not spiked. We see that there is not much difference in the bias and only a slight increase in the variance of our estimator and thus also there is not much difference in the coverage of the confidence intervals. We conclude that our method can provide 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

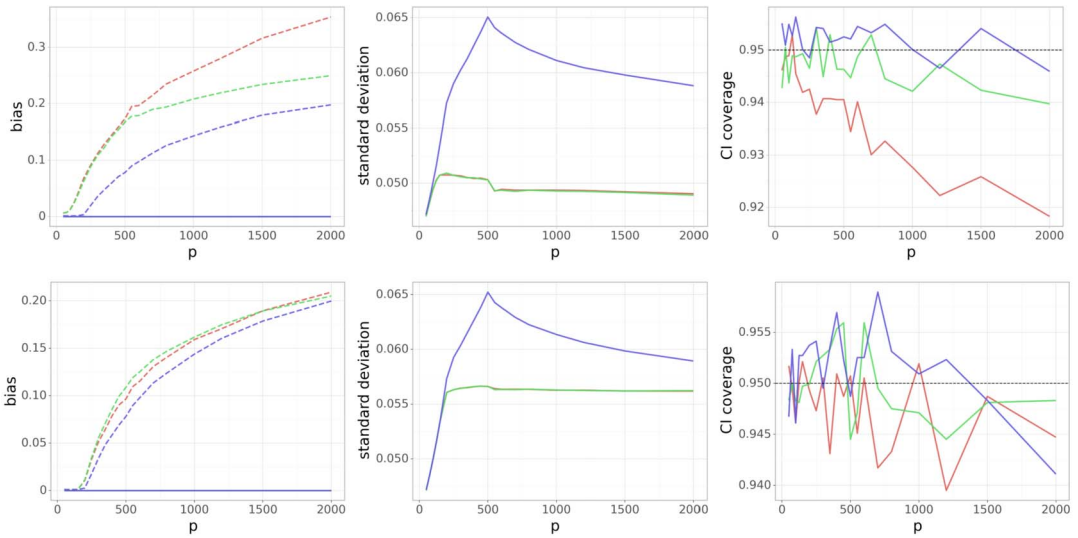

FIG. 7. *(No confounding bias) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of covariates p, while keeping $n = 500$ fixed. In the plots on the left, $|B_\beta|$ and $|B_b|$ are denoted by a dashed and a solid line, respectively, but $B_b = 0$ since we have enforced $b = 0$. Top row corresponds to the spiked covariance case $\Sigma_X = \Psi^T \Psi + I$, whereas for the bottom row we set $\Sigma_X = I$. 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 $\widehat{\beta}^{\text{init}}$ as our proposed method. Note that the last two methods have almost indistinguishable V.*

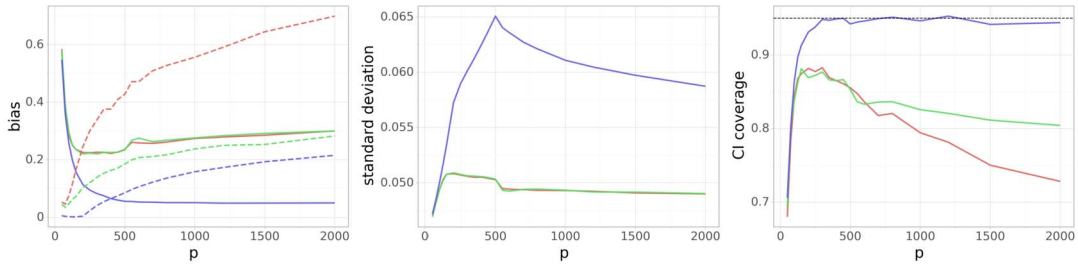

FIG. 8. *(Measurement error) Dependence of the (scaled) absolute bias terms $|B_\beta|$ and $|B_b|$ (left), standard deviation $V^{1/2}$ (middle) and the coverage of the 95% confidence interval (right) on the number of covariates p in the measurement error model (4). The sample size is fixed at $n = 500$. 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 $\widehat{\beta}^{\text{init}}$ as our proposed method. Note that the last two methods have almost indistinguishable $|B_b|$ and V.*

lasso, our proposed method still has essentially as good performance, with a small increase in variance.

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 $1 \le i \le n$. 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: The 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 (with a default trimming level of 50%) on a genomic data set. The data are obtained from the GTEx project [44], 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. The gene expression matrix X consists of measurements of expressions of $p = 12{,}646$ protein-coding genes for $n = 706$ individuals. Genomic data sets are particularly prone to confounding [23, 25, 42], 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 [56]. We will use several such landmark genes as the responses in our analysis.

In Figure 9, we can see a comparison of 95%-confidence intervals that are obtained from the doubly debiased lasso and standard debiased lasso. For a fixed response, landmark gene X_i, we choose 25 predictor genes X_j where $j \ne i$ such that their corresponding coefficients of the lasso estimator for regressing X_i on X_{-i} are nonzero. 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

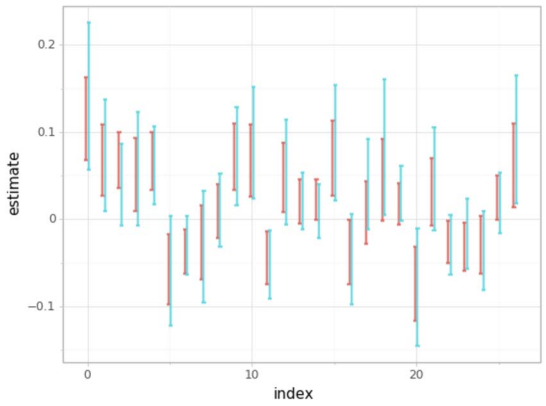

FIG. 9. *Comparison of* 95% *confidence intervals obtained by the doubly debiased lasso* (*blue*) *and doubly debiased lasso* (*red*) *for regression of the expression of one target landmark gene on the other gene expressions.*

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 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 the 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 the doubly debiased lasso are more robust in presence of confounding variables. However, for small model size

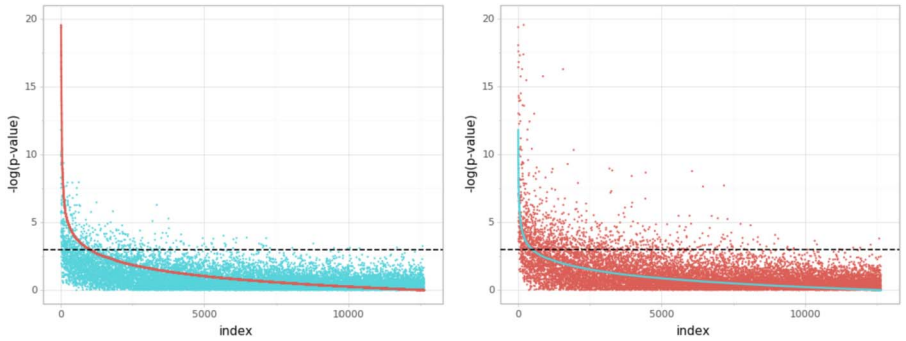

FIG. 10. *Comparison of p-values for two-sided test of the hypothesis* $\beta_j = 0$, *obtained by doubly debiased lasso* (*red*) *and doubly debiased lasso* (*blue*) *for regression of the expression of one target gene on the other gene expressions. The covariates are ordered by decreasing significance,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

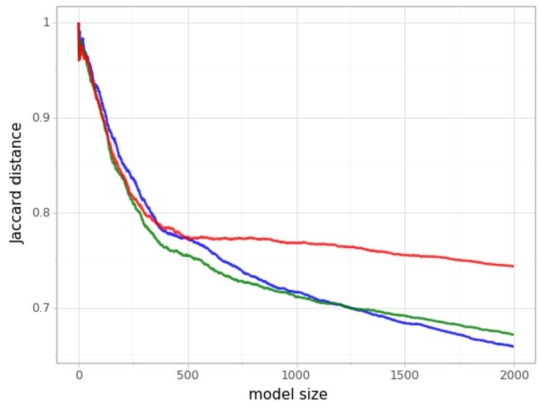

FIG. 11. *Comparison of the sets of the most significant covariates chosen based on the original expression matrix X and the deconfounded gene expression matrix $\tilde{X}$, for different cardinalities of the sets (model size). The set differences are measured by Jaccard distance. Red line represents 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).*

we do not see large gains. In this case, the sets produced by any method are quite different, that is, 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 the 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 remaining confounding, not accounted for by regressing out the given confounder proxies.

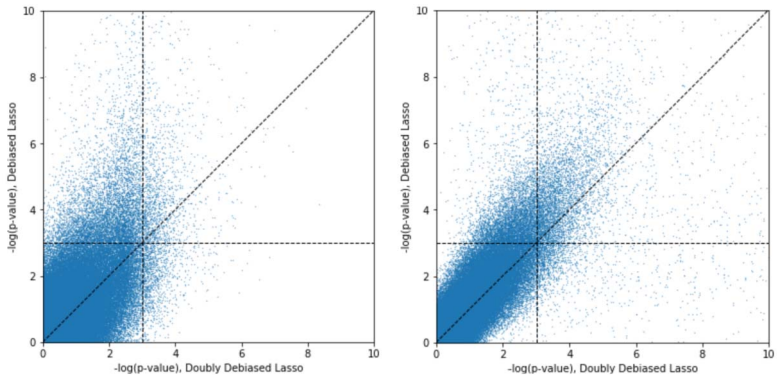

FIG. 12. *Comparison of p-values for two-sided test of the hypothesis $\beta_j = 0$, obtained by doubly debiased lasso and standard debiased lasso for regression of the expression of one target gene on the other gene expressions. The points are aggregated over 10 landmark response genes. The p-values 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.*

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 [66]; 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).

Acknowledgments. We thank Yuansi Chen for providing the code to preprocess the raw data from the GTEx project. We also thank Matthias Löffler for his help and useful discussions about random matrix theory.

Z. Guo and D. Ćevid contributed equally to this work.

Funding. The research of Z. Guo was supported in part by the NSF Grants DMS-1811857, 2015373 and NIH-1R01GM140463-01; Z. Guo also acknowledges financial support for visiting the Institute of Mathematical Research (FIM) at ETH Zurich.

D. Ćevid and P. Bühlmann received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 786461).

SUPPLEMENTARY MATERIAL

Supplement to "Doubly debiased lasso: High-dimensional inference under hidden confounding" (DOI: 10.1214/21-AOS2152SUPP; .pdf). Contains all proofs and additional simulation results.

REFERENCES

[1] BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica* **71** 135–171. MR1956857 https://doi.org/10.1111/1468-0262.00392

[2] BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259 https://doi.org/10.1111/1468-0262.00273

[3] BELLONI, A., CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and HANSEN, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica* **85** 233–298. MR3611771 https://doi.org/10.3982/ECTA12723

[4] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

[5] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[6] BOEF, A. G., DEKKERS, O. M., VANDENBROUCKE, J. P. and LE CESSIE, S. (2014). Sample size importantly limits the usefulness of instrumental variable methods, depending on instrument strength and level of confounding. *J. Clin. Epidemiol.* **67** 1258–1264.

[7] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. *Springer Series in Statistics*. Springer, Heidelberg. MR2807761 https://doi.org/10.1007/978-3-642-20192-9

[8] BURGESS, S., SMALL, D. S. and THOMPSON, S. G. (2017). A review of instrumental variable estimators for Mendelian randomization. *Stat. Methods Med. Res.* **26** 2333–2355. MR3712236 https://doi.org/10.1177/0962280215597579

[9] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973 https://doi.org/10.1198/jasa.2011.tm10155

[10] CAI, T. T. and GUO, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Ann. Statist.* **45** 615–646. MR3650395 https://doi.org/10.1214/16-AOS1461

[11] CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. *Monographs on Statistics and Applied Probability* **105**. CRC Press/CRC, Boca Raton, FL. MR2243417 https://doi.org/10.1201/9781420010138

[12] ĆEVID, D., BÜHLMANN, P. and MEINSHAUSEN, N. (2020). Spectral deconfounding via perturbed sparse linear models. *J. Mach. Learn. Res.* **21** Paper No. 232, 41. MR4209518 https://doi.org/10.22405/2226-8383-2020-21-1-221-232

[13] CHANDRASEKARAN, V., PARRILO, P. A. and WILLSKY, A. S. (2012). Latent variable graphical model selection via convex optimization. *Ann. Statist.* **40** 1935–1967. MR3059067 https://doi.org/10.1214/11-AOS949

[14] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[15] CHERNOZHUKOV, V., HANSEN, C. and LIAO, Y. (2017). A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.* **45** 39–76. MR3611486 https://doi.org/10.1214/16-AOS1434

[16] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Ann. Rev. Econ.* **7** 649–688.

[17] DEZEURE, R., BÜHLMANN, P. and ZHANG, C.-H. (2017). High-dimensional simultaneous inference with the bootstrap. *TEST* **26** 685–719. MR3713586 https://doi.org/10.1007/s11749-017-0554-2

[18] FAN, J., FAN, Y. and LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *J. Econometrics* **147** 186–197. MR2472991 https://doi.org/10.1016/j.jeconom.2008.09.017

[19] FAN, J. and LIAO, Y. (2014). Endogeneity in high dimensions. *Ann. Statist.* **42** 872–917. MR3210990 https://doi.org/10.1214/13-AOS1202

[20] FAN, J., LIAO, Y. and MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva. MR3091653 https://doi.org/10.1111/rssb.12016

[21] FAN, J., LIAO, Y. and WANG, W. (2016). Projected principal component analysis in factor models. *Ann. Statist.* **44** 219–254. MR3449767 https://doi.org/10.1214/15-AOS1364

[22] FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23. MR3397349 https://doi.org/10.1016/j.jeconom.2015.06.017

[23] GAGNON-BARTSCH, J. A. and SPEED, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13** 539–552.

[24] GAUTIER, E. and ROSE, C. (2011). High-dimensional instrumental variables regression and confidence sets. ArXiv preprint. Available at arXiv:1105.2454.

[25] GERARD, D. and STEPHENS, M. (2020). Empirical Bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *Biostatistics* **21** 15–32. MR4043843 https://doi.org/10.1093/biostatistics/kxy029

[26] GOLD, D., LEDERER, J. and TAO, J. (2020). Inference for high-dimensional instrumental variables regression. *J. Econometrics* **217** 79–111. MR4093746 https://doi.org/10.1016/j.jeconom.2019.09.009

[27] GÖTZE, F. and TIKHOMIROV, A. (2002). Asymptotic distribution of quadratic forms and applications. *J. Theoret. Probab.* **15** 423–475. MR1898815 https://doi.org/10.1023/A:1014867011101

[28] GUERTIN, J. R., RAHME, E. and LELORIER, J. (2016). Performance of the high-dimensional propensity score in adjusting for unmeasured confounders. *Eur. J. Clin. Pharmacol.* **72** 1497–1505. https://doi.org/10.1007/s00228-016-2118-x

[29] GUO, Z., ĆEVID, D. and BÜHLMANN, P. (2022). Supplement to "Doubly debiased lasso: High-dimensional inference under hidden confounding." https://doi.org/10.1214/21-AOS2152SUPP

[30] GUO, Z., KANG, H., CAI, T. T. and SMALL, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 793–815. MR3849344 https://doi.org/10.1111/rssb.12275

[31] HAGHVERDI, L., LUN, A. T. L., MORGAN, M. D. and MARIONI, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36** 421–427. https://doi.org/10.1038/nbt.4091

[32] HAN, C. (2008). Detecting invalid instruments using L_1-GMM. *Econom. Lett.* **101** 285–287. MR2477476 https://doi.org/10.1016/j.econlet.2008.09.004

[33] JANKOVÁ, J. and VAN DE GEER, S. (2018). Semiparametric efficiency bounds for high-dimensional models. *Ann. Statist.* **46** 2336–2359. MR3845020 https://doi.org/10.1214/17-AOS1622

[34] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[35] JIA, J. and ROHE, K. (2015). Preconditioning the lasso for sign consistency. *Electron. J. Statist.* **9** 1150–2015. MR3354334 https://doi.org/10.1214/15-EJS1029

[36] JOHNSON, W. E., LI, C. and RABINOVIC, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8** 118–127.

[37] KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *J. Amer. Statist. Assoc.* **111** 132–144. MR3494648 https://doi.org/10.1080/01621459.2014.994705

[38] LAM, C. and FAN, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254–4278. MR2572459 https://doi.org/10.1214/09-AOS720

[39] LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. MR2933663 https://doi.org/10.1214/12-AOS970

[40] LAM, C., YAO, Q. and BATHIA, N. (2011). Estimation of latent factors for high-dimensional time series. *Biometrika* **98** 901–918. MR2860332 https://doi.org/10.1093/biomet/asr048

[41] LEEK, J. T., SCHARPF, R. B., BRAVO, H. C., SIMCHA, D., LANGMEAD, B., JOHNSON, W. E., GEMAN, D., BAGGERLY, K. and IRIZARRY, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11** 733–739.

[42] LEEK, J. T., STOREY, J. D.(2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3** 1–12.

[43] LIN, W., FENG, R. and LI, H. (2015). Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. *J. Amer. Statist. Assoc.* **110** 270–288. MR3338502 https://doi.org/10.1080/01621459.2014.908125

[44] LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F. et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* **45** 580–585.

[45] MANGHNANI, K., DRAKE, A., WAN, N. and HAQUE, I. (2018). METCC: METric learning for confounder control making distance matter in high dimensional biological analysis. ArXiv preprint. Available at arXiv:1812.03188.

[46] MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. and HIRSCHHORN, J. N. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9** 356–369.

[47] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[48] NEYKOV, M., NING, Y., LIU, J. S. and LIU, H. (2018). A unified theory of confidence regions and testing for high-dimensional estimating equations. *Statist. Sci.* **33** 427–443. MR3843384 https://doi.org/10.1214/18-STS661

[49] NOVEMBRE, J., JOHNSON, T., BRYC, K., KUTALIK, Z., BOYKO, A. R., AUTON, A., INDAP, A., KING, K. S., BERGMANN, S. et al. (2008). Genes mirror geography within Europe. *Nature* **456** 98–101.

[50] NOVEMBRE, J. and STEPHENS, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40** 646–649. https://doi.org/10.1038/ng.139

[51] PEARL, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge Univ. Press, Cambridge. MR2548166 https://doi.org/10.1017/CBO9780511803161

[52] PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. and REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38** 904–909.

[53] RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res.* **11** 2241–2259. MR2719855

[54] REID, S., TIBSHIRANI, R. and FRIEDMAN, J. (2016). A study of error variance estimation in Lasso regression. *Statist. Sinica* **26** 35–67. MR3468344

[55] SHAH, R. D., FROT, B., THANEI, G.-A. and MEINSHAUSEN, N. (2020). Right singular vector projection graphs: Fast high dimensional covariance matrix estimation under latent confounding. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 361–389. MR4084168 https://doi.org/10.1111/rssb.12359

[56] SUBRAMANIAN, A., NARAYAN, R., CORSELLO, S. M., PECK, D. D., NATOLI, T. E., LU, X., GOULD, J., DAVIS, J. F., TUBELLI, A. A. et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171** 1437–1452.

[57] SUN, Y., ZHANG, N. R. and OWEN, A. B. (2012). Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. *Ann. Appl. Stat.* **6** 1664–1688. MR3058679 https://doi.org/10.1214/12-AOAS561

[58] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[59] VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

[60] WANG, J., ZHAO, Q., HASTIE, T. and OWEN, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Ann. Statist.* **45** 1863–1894. MR3718155 https://doi.org/10.1214/16-AOS1511

[61] WANG, W. and FAN, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Ann. Statist.* **45** 1342–1374. MR3662457 https://doi.org/10.1214/16-AOS1487

[62] WANG, Y. and BLEI, D. M. (2019). The blessings of multiple causes. *J. Amer. Statist. Assoc.* **114** 1574–1596. MR4047282 https://doi.org/10.1080/01621459.2019.1686987

[63] WINDMEIJER, F., FARBMACHER, H., DAVIES, N. and SMITH, G. D. (2019). On the use of the Lasso for instrumental variables estimation with some invalid instruments. *J. Amer. Statist. Assoc.* **114** 1339–1350. MR4011783 https://doi.org/10.1080/01621459.2018.1498346

[64] WOOLDRIDGE, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*, 2nd ed. MIT Press, Cambridge, MA. MR2768559

[65] YUAN, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* **11** 2261–2286. MR2719856

[66] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[67] ZHOU, S. (2009). Restricted eigenvalue conditions on subgaussian random matrices. ArXiv preprint. Available at arXiv:0912.4045.

[68] ZHU, Y. (2018). Sparse linear models and l_1-regularized 2SLS with high-dimensional endogenous regressors and instruments. *J. Econometrics* **202** 196–213. MR3778835 https://doi.org/10.1016/j.jeconom.2017.10.002