Unmeasured Confounding in High-Dimensional Data

Per August Moen

Contents

1 Introduction

In many statistical applications, one is interested in modeling the relationship between a dependent variable Y and some covariates X. If there is some unmeasured variable H related to both Y and X, the variable might cause spurious relationships and severely bias estimates of the dependency between Y and X. In the context of Causal Inference, if H causes both X and Y, we say that H is an *unmeasured confounder*. Even if we are not considering Causal Inference, we will still call H an unmeasured confounder whenever H is dependent on both Y and X.

Unmeasured confounding is particularly troublesome when doing Causal Inference on observational studies. In an observational study, we cannot intervene on the covariates of interest, as opposed to a randomized control study. Instead, the sample $(Y_i, X_i, H_i)_{i=1}^n$ is generated from a Data Generating Process (DGP) that we cannot control. Suppose the DGP is a linear Structural Equation Model (SEM) satisfying the causal graph in Figure 1. Regressing Y on the observed covariates X while ignoring the confounders H can lead to a significant bias in the estimated causal effects, and the bias does not vanish asymptotically. A classical remedy, stemming from the Economics literature, is the use of Instrumental Variables (IV).

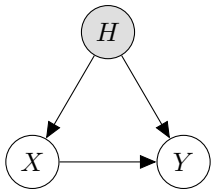

Figure 1: Causal Graph. Grey corresponds to an unmeasured variable.

Modern datasets are often High-Dimensional. What this means is that there are far more variables of interest than there are observations. Using an IV approach to adjust for confounders in High-Dimensional data is practically infeasible since the rank and order conditions (see Wooldridge 2010) will require us to have at least as many instruments as confounded covariates. With High-Dimensional datasets, estimation is difficult even without unmeasured confounders, although several methods have been proposed. For instance, in a Linear SEM as in Figure 1 without any confounders H, the Lasso (see Bühlmann and van de Geer 2011) can consistently estimate the Causal effect of X on Y under some sparsity conditions.

With the presence of unmeasured confounders, a major advantage of High-Dimensional data is that we can "learn" about the unmeasured confounders from the data under some circumstances. Intuitively, if we observe a large number of covariates X and we assume that there is only a small number of confounders, we may be able to retrieve some information about the confounders by analyzing the dependencies between the covariates. Indeed, if we assume the covariates X and confounders H are related through a factor model, the factor structure will allow us to model and estimate the relationship between X and H.

In the following, we will give a short review of factor models, proving useful and providing some intuition for the material to come. We will thereafter review some recently proposed methods to adjust for unmeasured confounders in High-Dimensional Linear Models. We will also investigate some of their properties via numerical simulations, and discuss possible extensions to non-linear models.

2 Preliminaries

2.1 Notation

We will use the following notation throughout the essay. For $n \in \mathbb{N}$, we let $[n] \coloneqq \{1, \dots, n\}$. For a matrix A, A_i will always mean the i-th column, while $A_{i,(\cdot)}$ will mean the i-th row (as a row vector), and A_{-i} will mean the matrix A with the i-th column removed. For any subset $\mathcal{C}$ of indices, we let $A_{\mathcal{C}}$ be the matrix whose columns are $(A_i \ : \ i \in \mathcal{C})$, and we let $A_{\mathcal{C},(\cdot)}$ be the matrix whose whose rows are $(A_{i,(\cdot)} \ : \ i \in \mathcal{C})$. We let $\|A\|_\infty$ denote the sup-norm of A and $\|A\|$ be the operator norm of A induced by the Euclidian 2-norm. We define $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to be the largest and smallest singular values of A, respectively, and let $\text{cond}(A)$ denote the condition number $\lambda_{\max}(A)/\lambda_{\min}(A)$ of A. For a vector v, stochastic or not, we will let v_i denote the i-th element, and let v_{-i} denote the vector v with the i-th element removed. For any subset $\mathcal{C}$ of indices, we let $v_{\mathcal{C}}$ be the vector $(v_i \ : \ i \in \mathcal{C})$. For $p > 0$, we let $\|v\|_p$ denote the Euclidian p-norm of v, and let $\|v\|_0$ denote the number of non-zero elements in v. We denote the standard Euclidian basis vectors by e_j. For any sequence of random variables X_n and any sequence of numbers a_n, we define $X_n = \text{o}_\text{p}(a_n)$ to mean that $X_n/a_n \xrightarrow{\text{p}} 0$ and $X_n = \mathcal{O}_\text{p}(a_n)$ to mean that X_n is bounded by a_n in probability. For any two positive sequences a_n, b_n, $a_n \lesssim b_n$ will mean that $\exists C > 0$ such that $a_n \leq C b_n$. We will write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$, and $a_n \ll b_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$.

In order to align notation as much as possible, we will in subsection 3.1 write $a_n \lesssim b_n$ to mean $a_n = \mathcal{O}(b_n)$, $X_n \lesssim_\text{p} (a_n)$ to mean $X_n = \mathcal{O}_\text{p}(b_n)$, and $X_n \gtrsim_\text{p} (a_n)$ to mean $b_n/X_n = \mathcal{O}_\text{p}(1)$.

2.2 Factor models

A *factor model* describes the dependence between random variables in terms of some unobserved latent random variables called *factors*. A quintessential example of such a model is the Linear Factor Model, studied in Chamberlain and Rothschild (1983) and Harman (1976). Let $X = (X_1, \dots, X_p)$ and $H = (H_1, \dots, H_q)$ be random row vectors, which we both may take without loss of generality (WLOG) to be mean-zero. We say that X follows a *Linear Factor Model* with latent factors H if

$$X_j = \sum_{k=1}^{q} \Gamma_{k,j} H_k + E_j, \tag{1}$$

or in matrix notation,

$$X = H\Gamma + E,$$

for some mean zero random variables $\{E_j\}_{j=1}^p$ independent of H and a matrix of *factor loadings* $\{\Gamma_{k,j}\}_{k,j=1}^{q,p}$. We may WLOG take the covariance matrix of H to be I_q. If the covariance matrix of $E = (E_1, \dots, E_p)$ is Σ, the covariance matrix of X is given by

$$\Sigma_X \coloneqq \text{Cov}(X) = \Sigma + \Gamma^\intercal \Gamma \ .$$

If $\Sigma = I_p$, the Linear Factor Model can be seen as a generalization of the Spiked Covariance Model (see Bacallado and Shah 2020) with spikes in multiple directions.

There is a vast literature on factor analysis, dating back as far as the early 20th century (see Harman 1976). Commonly, one wishes to estimate the factor loadings Γ and the idiosyncratic covariance matrix Σ from n i.i.d. samples of X, both when n is moderately large compared to p, and also in the High-Dimensional case $p > n$. As an example of the latter case, Bai and Li (2012) consider estimation of both Γ and Σ in the Linear Factor Model (1) through quasi-maximum likelihood estimation. They consider an asymptotic regime where n and p are allowed to grow to infinity, while the number of

factors q is fixed. In subsection 3.3 we will use the said estimation technique to adjust for confounders in a multiple testing setting.

One may also be solely interested in estimating the idiosyncratic covariance matrix Σ in (1), for instance, when considering the H's as confounders. Drawing intuition from the Spiked Covariance Model, if the eigenvalues of $\Gamma^\mathsf{T}\Gamma$ are larger than the eigenvalues of Σ, one should expect that the top q eigenvectors of Σ_X span most of the column space of $\Gamma^\mathsf{T}\Gamma$ (and, importantly, they should not span much of the column space of Σ). Thus, if we let $\tilde{\Sigma}$ be the matrix Σ_X with the top q eigenvectors of Σ_X removed (i.e. we set the top q eigenvalues of Σ_X to zero in the eigendecomposition), we should expect $\tilde{\Sigma}$ to approximate the idiosyncratic covariance matrix Σ well. It turns out that under some conditions, this intuition carries through. We have the following Proposition from Shah et al. (2018). To align notation with the rest of the essay, we state the Proposition in terms of singular values instead of eigenvalues (which is equivalent).

Proposition 2.1. *Let Π_Γ denote the orthogonal projection matrix onto the column space of Γ. Let $\rho_1 = \|\Pi_\Gamma \Sigma\|$ and $\rho_2 = \max_j \|\Pi_\Gamma e_j\|_2$. Suppose $\lambda_{\min}(\Sigma)$ is bounded away from zero and $\lambda_{\min}(\Gamma)^2 > c\lambda_{\max}(\Sigma)$ for some $c > 1$. Then*

$$\|\tilde{\Sigma} - \Sigma\|_\infty \lesssim \rho_1\rho_2 + \lambda_{\max}(\Gamma)^2\rho_1^2/(\lambda_{\min}(\Gamma))^4.$$

For instance, if $\lambda_{\max}(\Sigma)$ and $\mathrm{cond}(\Gamma^\mathsf{T}\Gamma)$ are bounded from above, and $\lambda_{\min}(\Gamma) \to \infty$ and $\rho_2 \to 0$ as $p, q \to \infty$, we have that $\|\tilde{\Sigma} - \Sigma\|_\infty \to 0$. In such a case, removing the top q eigenvectors from Σ_X asymptotically retrieves the idiosyncratic covariance matrix Σ. Shah et al. (2018) also propose an estimator of Σ based on i.i.d. samples of X in the High-Dimensional setting called the *Right Singular Vector Projection* (RSVP). Instead of removing the top q eigenvectors of the empirical covariance matrix $\hat{\Sigma}_X$, the RSVP estimator sets all eigenvalues of $\hat{\Sigma}_X$ to 1, and will consistenly estimate Σ up to a constant factor under a variety of asymptotic regimes.

In subsections 3.1 and 3.2, we will see that the intuition on removing the top eigenvectors of the empirical covariace matrix will carry on into linear regression if the confounders and covariates are related through a linear factor model.

3 Inference in confounded linear models

In this section, we will consider inference in confounded linear models. The first two subsections will deal with point estimation and single parameter confidence intervals, and the last subsection will deal with multiple testing.

3.1 Point estimation

In this subsection, we will follow Ćevid et al. (2018), and all results will follow from the paper unless otherwise is stated. Throughout will consider an asymptotic regime where both p and n are allowed to grow, although we will suppress this in the notation.

Consider the *Confounded Linear Model*:

$$Y = X\beta + H\delta + \nu, \tag{2}$$
$$X = H\Gamma + E, \tag{3}$$

where $Y \in \mathbb{R}^{n\times 1}$ is a vector of responses, $X \in \mathbb{R}^{n\times p}$ is a design matrix with i.i.d. rows, and $H \in \mathbb{R}^{n\times q}$ is a matrix of unobserved confounding variables with i.i.d. rows. The vectors $\beta \in \mathbb{R}^p$ and $\delta \in \mathbb{R}^q$ are coefficient vectors, and $\Gamma \in \mathbb{R}^{q\times p}$ are factor loadings. We allow q to grow with n and p. We will assume that ν is a vector of mean-zero independent sub-Gaussian variables with fixed variance σ_ν^2, that each row $H_{i,(\cdot)}$ of H has distribution $\mathrm{N}_q(0, \mathrm{I}_q)$, and that each row $E_{i,(\cdot)}$ of E has distribution $\mathrm{N}_p(0, \Sigma_{\mathrm{E}})$[1], with ν being independent of E and H. We will also assume $E \perp\!\!\!\perp H$.

The Confounded Linear Model has the form of an ordinary Linear Model with the additional property that X is generated from a factor model with latent variables H and factor loadings Γ, and that H also linearly relates to Y. We thus model the confounding explicitly. Just as in the discussion on factor models above, we have that the covariance matrix Σ of any row of X satisfies $\Sigma \coloneqq \mathrm{Cov}\,(X_{1,(\cdot)}) = \Gamma^{\mathsf{T}}\Gamma + \Sigma_E$.

Remark 3.1. In a Causal Inference setting, the Confounded Linear Model (2), (3) can represent the distributional realization of a linear SEM, where, for each row i, $H_{i,(\cdot)}$ is a parent of $X_{i,(\cdot)}$ and Y_i, and $X_{i,(\cdot)}$ is a parent of Y_i, as in Figure 1. In such a case, the coefficient β is interpreted as the causal effect of $X_{i,(\cdot)}$ on Y_i. Ignoring the confounders H can lead to a large bias in the estimate of the causal effect β.

In the analysis to follow, it will be helpful to consider the Confounded Linear Model (2), (3) in its marginal form. By splitting up $H\delta$ into a part that correlates with X and another part that does not, we may write the model marginally as

$$Y = X(\beta + b) + \varepsilon, \tag{4}$$

where $b \coloneqq \Sigma^{-1}\Gamma^{\mathsf{T}}\delta$ and $\varepsilon \coloneqq H\delta - Xb + \nu$. The error term ε (not to be confused with ν) in (4) is a vector of sub-Gaussian random variables with expectation zero and variance σ^2 bounded by $\sigma_\nu^2 + \|\delta\|^2$. Due to the Gaussianity (and independence) of X and H, the error term ε is independent of X.

We are interested in estimating the coefficient vector β. As is typical (see Bühlmann and van de Geer 2011), we will assume that the true underlying β is sparse, with $s = \|\beta\|_0$ number of non-zero entries at indices S. We allow s to grow with n and p. Under some conditions on Γ and Σ_E, one can show that b must become small asymptotically:

[1] The assumption on Gaussianity simplifies the exposition because the (joint) Gaussianity of X and H imply that ε in (4) is independent of X. The Gaussianity assumption may be replaced by sub-Gaussianity, however, in which case ε is only uncorrelated with X. Results analogous to Theorem 3.7 and Corollary 3.9, when X, H are sub-Gaussians, are provided in Guo et al. (2020) (Proposition 4).

Lemma 3.2. *Assume that the Confounded Linear Model (2), (3) satisfies* $\lambda_{\min}(\Gamma) \gtrsim \sqrt{p}$ *and cond*$(\Sigma_E) \lesssim 1$. *Then,*

$$\|b\|_2^2 \lesssim \frac{\sigma^2}{p}.$$

Under the assumptions of Lemma 3.2, if $\frac{\sigma^2}{p} \to 0$, then β is asymptotically identifiable (recall that σ^2 is the variance of the error term inn the marginal Confouded Linear Model, which, unlike σ_ν, depends on p and q). Motivated by Lemma 3.2 and equation (4), one might be tempted to regress Y on X by using the Lasso. The problem with this approach is that, even though β is sparse, $\beta + b$ may very well not be, and the usual results on the convergence of the Lasso cannot be immediately applied, as they require sparsity of the coefficient vector (see Bühlmann and van de Geer 2011).

To break through this impasse, we will take advantage of the factor model structure of X. As discussed in Section 2, the covariance matrix of the rows of X should have some spiked eigenvalues, at least when the eigenvalues of Σ_E are reasonably small. By removing the top singular vectors from X (equivalently, remove the top eigenvectors from the empirical covariance $X^\mathsf{T}X/n$), we can, in some sense, filter out the part of X that depends on the confounding variables H. This can be done by so-called *spectral linear transformations* which modify the singular values of X. For now, we will consider an arbitrary linear transformation F. After transforming both sides of equation (4) by F, we have

$$\tilde{Y} = \tilde{X}\beta + \tilde{X}b + \tilde{\varepsilon}, \tag{5}$$

where $Y \coloneqq FY$, $X \coloneqq FX$ and $\tilde{\varepsilon} \coloneqq F\varepsilon$. We will estimate β by applying the Lasso to equation (5), that is,

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\}. \tag{6}$$

We have the following Lemma on the ℓ_1 error of the Lasso (6), which holds for sub-Gaussian models of the form (4) under arbitrary linear transformations F. Recall that s is the number of non-zero elements of β, and S is their indices.

Lemma 3.3. *Let F be a linear transformation and let $\tilde{\cdot}$ denote transformation by F. Consider a model of the form (4) where ε is a vector of independent sub-Gaussian variables with variance σ^2, $X \perp\!\!\!\perp \varepsilon$, and X has independent sub-Gaussian rows with covariance Σ. Suppose $\max_i \Sigma_{i,i} \lesssim 1$ and let $\tilde{\Sigma} \coloneqq \frac{1}{n}\tilde{X}^\mathsf{T}\tilde{X}$.*

Let $A > 0$ be fixed. For the Lasso estimator as in (6) and $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}\lambda_{\max}(F)^2$, with probability at least $1 - 2p^{1-A^2/(32\max_i(\Sigma_{i,i}))} - pe^{-n/136}$, we have that

$$\|\hat{\beta} - \beta\|_1 \leq C_1\frac{s\lambda}{\varphi_{\tilde{\Sigma}}^2} + C_2\frac{\|\tilde{X}b\|_2^2}{n\lambda}, \tag{7}$$

where

$$\varphi_{\tilde{\Sigma}}^2 \coloneqq \inf_{\alpha:\|\alpha\|_1 \leq 5\|\alpha_S\|_1} \frac{\sqrt{\alpha^\mathsf{T}\tilde{\Sigma}\alpha}}{\frac{1}{\sqrt{s}}\|\alpha_S\|_1},$$

for some constants C_1, $C_2 > 0$.

Remark 3.4. Lemma 3.3 is called Theorem 2 in Ćevid et al. (2018). The proof follows the same strategy as the usual proofs of the convergence rate of the Lasso under no confounding. Indeed, to

prove Lemma 3.3, one shows that (7) holds on the event $\Omega := \left\{ \frac{1}{n} \|\tilde{X}^\mathsf{T} \tilde{\varepsilon}\|_\infty \leq h \right\}$ (for h suitably chosen) and that Ω has high probability. The proof that Ω has high probability is found in Ćevid et al. (2018) (Lemma 8), but is unfortunately unsatisfactory. In the proof, the authors treat the standard error σ_ν of ν as if it were the sub-Gaussian parameter of ν, and similarly for the entries of X. The variance and the sub-Gaussian parameter of a sub-Gaussian variable are not in general equal, and the sub-Gaussian parameter is never smaller than the variance. This ambiguity can immediately be resolved by taking σ_ν to be the sub-Gaussian parameter of ν and assuming that X has mean-zero Gaussian rows with covariance Σ (then $\sqrt{\Sigma_{j,j}}$ is the sub-Gaussian parameter of $X_{i,j}$ $\forall i, j$). Gaussianity of X is in fact assumed in the Linear Confounded Model, and we therefore ignore the ambiguity of the proof and take Lemma 3.3 as given.

Remark 3.5. Taking F to be the identity matrix corresponds to not performing any spectral transformation at all. In such a case, the term $\frac{\|\tilde{X}b\|_2^2}{n\lambda}$ in equation (7) may be large, even if b becomes small. This observation is in line with our discussion above, where we noted that it is not obvious that the Lasso should perform well under confounding. However, if we are ready to assume that b is sufficiently small, the Theorem above will provide consistency of the Lasso. In numerical experiments, we will see that the Lasso can perform better under confounding than one might expect.

To apply Lemma 3.3 and improve upon the performance of the Lasso under confounding, we need to pick F such that $\|\tilde{X}b\|_2^2 \leq \lambda_{\max}(\tilde{X})\|b\|_2^2$ becomes small. With Proposition 2.1, we saw that removing the top eigenvalues of Σ (equivalently, the singular values) in some sense reduced the covariance of X stemming from the factor term. Can we do something similar with our observed covariates X, which are assumed to follow a factor model?

Let $X = UDV^\mathsf{T}$ be the singular value decomposition of X, where $D = \operatorname{diag}(d_1, \ldots, d_r)^\mathsf{T}$ are the ordered singular values of X and $r = \min(n, p)$. We will let F be the transformation that trims all singular values of X to be at most τ, that is,

$$\begin{aligned} \tilde{d}_i &:= \min(d_1, \tau), \ \forall i \in [r], \\ F &:= U \cdot \operatorname{diag}(\tilde{d}_1/d_1, \ldots, \tilde{d}_r/d_r) \cdot U^\mathsf{T}. \end{aligned} \tag{8}$$

The transformation F in (8) is called the *Trim transform*, see Ćevid et al. (2018). When applied to X, the Trim transform reduces the magnitude of the singular vectors of X corresponding to the largest singular values. The Trim transform has the nice property that we do not need to know how many singular vectors to remove. In practice, a good choice of τ is the median largest singular value of the observed matrix X (see discussion in Ćevid et al. 2018). More generally, we can also choose $\tau = d_{\lfloor tr \rfloor}$ for any $t \in (0, 1)$. From now on, we will focus our attention on the Trim transform. With F being the Trim transform, we will call the solution to (6) the *Trimmed Lasso*.

The following Lemma bounds the rate of $\lambda_{\max}(\tilde{X}) = d_{\lfloor tn \rfloor}$ to $\sqrt{p}$ when F is the Trim transform.

Lemma 3.6. *Consider a model of the form (4), where $X \in \mathbb{R}^{n \times p}$ has i.i.d. sub-Gaussian rows with covariance matrix Σ and sub-Gaussian norm [2] $M \lesssim 1$. Let $d_1, \ldots, d_r \geq 0$ be the singular values of X. Assume that $\operatorname{Tr}(\Sigma) \asymp p$, $\sqrt{\log(p)/n} \to 0$ and that $p > n$. Then for any $t \in (0, 1)$,*

$$d_{\lfloor tn \rfloor} \lesssim_p \sqrt{p}.$$

Proof.
Note: The following proof substantially elaborates the proof of Lemma 2 in Ćevid et al. (2018). In the original proof, the tail bound (9) is not shown. It is also not shown that (10) implies (11) and that (11)

[2] Meaning that $\forall u \in \mathbb{R}^p$ such that $\|u\|_2 = 1$, the variable $u^\mathsf{T} X_{i,(\cdot)}$ is sub-Gaussian with parameter M.

implies (12). I also add an extra hypothesis to the Lemma ($M \lesssim 1$) to avoid the ambiguity discussed in Remark 3.4.

By a sub-Gaussian tail bound, we have that

$$\|\frac{1}{n}X^\mathsf{T}X - \Sigma\|_\infty \lesssim_\mathrm{p} \sqrt{\frac{\log p}{n}}. \tag{9}$$

Indeed, since the rows of X are sub-Gaussian with sub-Gaussian norm M, for all l and j, the product $X_{i,j}X_{i,l}$ satisfies Bernstein's condition (see Bacallado and Shah 2020) with parameters $(8M^2, 4M^2)$. By Bernstein's inequality, we have that for any $t > 0$,

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^n X_{i,j}X_{i,l} - \Sigma_{j,l}| > tM^2\right) \leq 2\exp\left(-\frac{nt^2}{2(8^2 - 4t)}\right).$$

By a union bound,

$$\begin{aligned}\mathbb{P}\left(\|\frac{1}{n}X^\mathsf{T}X - \Sigma\|_\infty > tM^2\right) &\leq 2p^2\exp\left(-\frac{nt^2}{2(8^2 - 4t)}\right) \\ &= 2\exp\left(-\frac{nt^2}{2(8^2 - 4t)} + 2\log p\right).\end{aligned}$$

Since $\sqrt{\log(p)/n} \to 0$, taking $t^2 = C\frac{\log p}{n}$ with $C = (4 \cdot 8^2 + 1)$ gives that

$$\begin{aligned}\mathbb{P}\left(\|\frac{1}{n}X^\mathsf{T}X - \Sigma\|_\infty > M^2\sqrt{C\frac{\log p}{n}}\right) &\leq 2\exp\left(-\log(p)\frac{C}{2\left(8^2 - 4\sqrt{C\frac{\log(p)}{n}}\right)} + 2\log p\right) \\ &\to 0.\end{aligned}$$

Since $M^2 \lesssim 1$, we establish (9). By the uniform bound (9) it follows immediately that

$$\left|\frac{1}{n}\operatorname{tr}\big(X^\mathsf{T}X\big) - \operatorname{tr}(\Sigma)\right| \lesssim_\mathrm{p} p\sqrt{\frac{\log p}{n}}. \tag{10}$$

Noticing that $\operatorname{tr}\big(X^\mathsf{T}X\big) = \sum_{i=1}^n d_i^2$ (where we sum up to n since $\operatorname{rank}(X^\mathsf{T}X) \leq n$), we get that

$$\frac{1}{n}\sum_{i=1}^n d_i^2 = \operatorname{tr}(\Sigma) + \mathcal{O}_\mathrm{p}\left(p\sqrt{\frac{\log p}{n}}\right).$$

Since $\operatorname{tr}(\Sigma) \asymp p$, we also have that $\operatorname{tr}(\Sigma) \gtrsim p$. Thus,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n d_i^2 &= \operatorname{tr}(\Sigma)\left(1 + \frac{\mathcal{O}_\mathrm{p}(p\sqrt{\frac{\log p}{n}})}{\operatorname{tr}(\Sigma)}\right) \\ &= \operatorname{tr}(\Sigma)(1 + \mathrm{o}_\mathrm{p}(1)),\end{aligned} \tag{11}$$

where the last inequality follows from the fact that $\sqrt{\log(p)/n} \to 0$ and $\operatorname{tr}(\Sigma) \gtrsim p$. Now,

$$d_{\lfloor tn \rfloor}^2 \leq d_{\lfloor tn \rfloor}^2 + \sum_{i=\lfloor tn \rfloor}^n d_i^2$$

$$
\begin{aligned}
&\lesssim \frac{n - \lfloor tn \rfloor}{n} d^2_{\lfloor tn \rfloor} + \frac{1}{n} \sum_{i=\lfloor tn \rfloor}^{n} d_i^2 \\
&\leq \frac{1}{n} \sum_{i=1}^{n} d_i^2 \\
&= \mathrm{tr}(\Sigma)(1 + \mathrm{o_p}(1)).
\end{aligned}
$$

Since $\mathrm{tr}(\Sigma) \asymp p$ also implies that $\mathrm{tr}(\Sigma) \lesssim p$, we thus get that

$$
\frac{d^2_{\lfloor tn \rfloor}}{p} \lesssim (1 + \mathrm{o_p}(1)).
$$

One may check using the defintion of boundedness in probabiltity that this implies

$$
\frac{d^2_{\lfloor tn \rfloor}}{p} \lesssim_{\mathrm{p}} 1.
$$

Thus,

$$
d_{\lfloor tn \rfloor} \lesssim_{\mathrm{p}} \sqrt{p}, \tag{12}
$$

and we are done. □

The following Theorem makes clear why the properties of the Trim transform fit nicely with Theorem 3.3. Notice in particular that the Trim transform satisfies condition **(A2)** by Lemma 3.6. The reader is reminded that we assume the rows of X are Gaussian with covariance matrix Σ.

Theorem 3.7. *Consider the Confounded Linear Model (2) and (3). Let F be any linear transformation such that $\lambda_{\max}(F)$ is uniformly bounded from above and below. Suppose $\max_i \Sigma_{i,i} \lesssim 1$, $cond(\Sigma_E) \lesssim 1$ and that $\lambda_{\min}(\Sigma)$ is uniformly bounded from below. Assume that the following conditions hold*

(A1) $\lambda_{\min}(\Gamma) \gtrsim \sqrt{p}$,

(A2) $\lambda_{\max}(\tilde{X}) \lesssim_p \sqrt{p}$,

(A3) $\varphi^2_{\tilde{\Sigma}} \gtrsim_p \lambda_{\min}(\Sigma)$.

Then for $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$,

$$
\|\hat{\beta} - \beta\|_1 \lesssim_p \sigma s \sqrt{\frac{\log p}{n}}. \tag{13}
$$

Proof.
Note: The proof follows the proof of a slightly more general result (Corollary 1) in Ćevid et al. (2018) (the only thing that differs is the assumption (A1)). I provide more intermediary calculations and explanations than in the original proof, most importantly why we can rescale λ and why the boundedness in probability (13) holds.

Assume first that $\lambda = A\sigma\sqrt{\frac{\log(p)}{n}}\lambda_{\max}(F)^2$. By Lemma 3.3, with probability at least $1 - 2p^{1-A^2/(32\max_i(\Sigma_{i,i}))} - pe^{-n/136}$, we have that

$$
\|\hat{\beta} - \beta\|_1 \leq C_1 \frac{s\lambda}{\varphi^2_{\tilde{\Sigma}}} + C_2 \frac{\|\tilde{X}b\|_2^2}{n\lambda},
$$

for some constants C_1, C_2. Pick A sufficiently large so that $1 - 2p^{1-A^2/(32\max_i(\Sigma_{i,i}))} - pe^{-n/136} \to 0$. Then, with probability tending to 1, we have that

$$\|\hat{\beta} - \beta\|_1 \lesssim \frac{s\lambda}{\varphi^2_{\tilde{\Sigma}}} + \frac{\|\tilde{X}b\|_2^2}{n\lambda}. \tag{14}$$

Since $\lambda_{\max}(F)$ is uniformly bounded from above and below and A is a constant, the inequality (14) still holds on a set with probability tending to 1 if we redefine λ to be $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$. By Lemma 3.2 and assumption **(A2)**, we have that

$$\begin{aligned}
\frac{\|\tilde{X}b\|_2^2}{n\lambda} &\leq \underbrace{\lambda_{\max}(\tilde{X})}_{\lesssim_{\mathrm{p}} p} \underbrace{\|\beta\|_2^2}_{\lesssim \frac{\sigma^2}{p}} \frac{1}{n\lambda} \\
&\lesssim_{\mathrm{p}} \frac{\sigma^2}{n\lambda} \\
&\leq \frac{s\log(p)\sigma^2}{n\lambda},
\end{aligned}$$

where we in the last equality simply multiplied by $s\log p \geq 1$. Taking $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$, we get that

$$\frac{\|\tilde{X}b\|_2^2}{n\lambda} \lesssim_{\mathrm{p}} s\sigma\sqrt{\frac{\log p}{n}}.$$

By assumption **(A2)**, we also have that

$$\begin{aligned}
\frac{s\lambda}{\varphi^2_{\tilde{\Sigma}}} &\lesssim s\sigma\sqrt{\frac{\log p}{n}}\frac{1}{\varphi^2_{\tilde{\Sigma}}} \\
&\lesssim_{\mathrm{p}} s\sigma\sqrt{\frac{\log p}{n}}\frac{1}{\lambda_{\min}(\Sigma)}.
\end{aligned}$$

Since (14) holds on a set with probability tending to 1, by a union bound we have that

$$\|\hat{\beta} - \beta\|_1 \lesssim_{\mathrm{p}} C_1 s\sigma\sqrt{\frac{\log p}{n}}\frac{1}{\lambda_{\min}(\Sigma)} + C_2 s\sigma\sqrt{\frac{\log p}{n}}.$$

Since $\lambda_{\min}(\Sigma)$ is uniformly bounded from below, we are done. □

Theorem 3.7 has an intuitive (although not obvious) explanation. The following explanation is complements that in Ćevid et al. (2018). By assuming that the error term E in 3 satisfies $\mathrm{cond}(\Sigma_E) \lesssim 1$, we essentially ensure that E does not have spiked singular values. Recall from our discussion on factor models (subsection 2.2) that we should expect the design matrix X to hve some "spiked" singular values due to the factor structure. As these spikes do not originate from E, they must come from the term $H\Gamma$. If the number of confounders q is relatively small, the number "spiked" singular values should be small as well. By the definition of b, Xb is the part of $H\Gamma$ that correlates with X (i.e., Xb is the projection of $H\Gamma$ onto X). Therefore, Xb should approximately be contained in the span of the top (left) singular vectors of X. Since we assume that F reduces the magnitude of these top singular vectors, the Trim transform reduces the magnitude of Xb. Since β is sparse, $X\beta$ is a linear combination of only a few columns of X, which are increasingly unlikely to be in the span of top singular vectors of X. Thus, F reduces the magnitude of Xb much more than it does to $X\beta$, and hence reducing the confounding effect Xb without reducing the "signal" $X\beta$ too much.

We call the assumption **(A1)** the *dense confounding assumption* since it ensures that sufficiently many covariates (columns of X) are sufficiently correlated with the confounders for the design matrix X to have spiked singular vectors. The dense confounding assumption can be shown to hold with high probability if, for instance, the entries of Γ is generated i.i.d. from a standard Normal distribution, $p \gg q$, and $\|\delta\|_\infty$ is uniformly bounded above. See Lemma 4 and Lemma 5 in Guo et al. (2020) for more general statements. The condition **(A3)** can also be shown to hold with high probability, under some conditions we do not state here. See Lemma 4 in Ćevid et al. (2018) for details.

Remark 3.8. The assumption **(A3)** closely resembles the so-called Compatability Condition, which is a common assumption for the Lasso even for an unconfounded linear model (see Bühlmann and van de Geer 2011).

By combining Theorem 3.7 and Lemma 3.6, we have the following immediate Corollary on the convergence rate of the Trimmed Lasso.

Corollary 3.9. *Consider the Confounded Linear Model (2) and (3). Let F be the Trim transform. Suppose that $\max_i \Sigma_{i,i} \lesssim 1$, $cond(\Sigma_E) \lesssim 1$ and that $\lambda_{\min}(\Sigma)$ is bounded uniformly from below. Under conditions **(A1)** and **(A3)** in Theorem 3.7, if $p > n$ and $\sigma s\sqrt{\frac{\log p}{n}} \to 0$, then for $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$, we have that*

$$\|\hat{\beta} - \beta\|_1 \xrightarrow{p} 0.$$

Proof.
Note: This Corollary is my own, but as the proof shows, it is (almost) immediate.
Since X has i.i.d. Gaussian rows with covariance matrix Σ with $\max_i \Sigma_{i,i} \lesssim 1$, the hypothesis of Lemma 3.6 is satisfied. Thus, the assumption **(A2)** in Theorem 3.7 holds. Since $\sigma s\sqrt{\frac{\log p}{n}} \to 0$, it follows from Theorem 3.7 that $\|\hat{\beta} - \beta\|_1 \xrightarrow{\text{p}} 0$. □

Corollary 3.9 shows that, under the dense confounding assumption, the Trimmed Lasso has the same convergence rate as the standard Lasso has under no confounding. We will see in the simulations that the Trimmed Lasso can improve the estimation error of Lasso even when $\Gamma = 0$ (i.e., no confounding), as the Trimmed Lasso also adjusts for correlation in the covariates.

Remark 3.10. A caveat of Theorem 3.7 and Corollary 3.9 is that the dependence of the assumptions on the number of confounders q is not explicit. Furthermore, the conditions on Σ and σ depend on the other parameters in a somewhat roundabout way, although making assumptions on Σ and σ directly makes the proofs simpler. In the next subsection, we will see some refined conditions which involve q explicitly and do not put conditions on Σ and σ directly.

In practice, the optimal value of the penalty parameter λ (i.e., minimizing $\|\hat{\beta} - \beta\|_1$) is unknown because it depends on unknown quantities. A reasonable way to choose λ in the unconfounded case is to minimize the estimated squared prediction error via cross-validation (CV), see Bühlmann and van de Geer (2011). We will see in our simulations that, the stronger the presence of confounding, the more CV will underestimate the optimal λ, although choosing λ by CV works fairly good even in the confounded case. For a more detailed discussion on the topic of CV for the Trimmed Lasso, see Ćevid et al. (2018). In our simulations, we will also see that the performance of the standard Lasso under confounding can be improved by choosing a larger value of λ than what CV suggests.

3.2 Single parameter confidence intervals

In the last subsection, we saw that the regression coefficient β in the Confounded Linear Model can be consistently estimated by the Trimmed Lasso. In this subsection, we will see how we can extend these ideas to obtain confidence intervals for individual components of the coefficient vector β. We will continue to let F denote the Trim transform (8) and let $\tilde{\cdot}$ denote transformation by F. We will follow Guo et al. (2020), and all results in this subsection are results from the paper unless otherwise is stated. We begin with a few remarks.

Guo et al. (2020) extend the results from the previous subsection with some minor changes in the setup. We will still consider the Confounded Linear Model (2), (3), where ν is independent of X and H. We will no longer assume that the rows of X and H are jointly Gaussian. Instead, we will require only that E and H are element-wise uncorrelated, and that each element of X is sub-Gaussian. As before, the rows of ν, X and H are assumed to be i.i.d.. Additionally, the Lasso defined as in (6) does not take into account that the columns of X may have different scales. It is, therefore, reasonable to scale the columns of $\tilde{X}$ to have 2-norm $\sqrt{n}$ before applying the Lasso, or equivalently, redefine the Trimmed Lasso to be

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda \sum_{j=1}^{p} \frac{\|\tilde{X}_j\|_2}{\sqrt{n}} |\beta_j| \right\}. \tag{15}$$

We mention in passing that this weighting of the penalties does not make much practical difference. Indeed, it can be shown that the weighted penalties are of the same order as non-weighted penalties. See, for instance, the proof of Proposition 4 in Guo et al. (2020).

Results similar to Lemma 3.6 and Theorem 3.7 will hold in our new setting, with some slightly refined (and more technical) conditions, which can be found listed below. In particular, the compatibility condition (A3) is replaced by a Restricted Eigenvalue assumption. Also, since X and H are no longer assumed to be jointly Gaussian, the error term ε in the marginal Confounded Linear Model (4) is only uncorrelated with X, and a slightly larger penalty term λ is needed. For more details, see Corollary 2 and Proposition 4 in Guo et al. (2020).

For simplicity, we will from now on assume that each element of ν is i.i.d. Gaussian with mean zero and variance σ_ν^2. [3]

Suppose we seek a confidence interval for β_j for some $j \in [p]$. After computing $\hat{\beta}$ by the Trimmed Lasso, we may rewrite the marginal Confounded Linear Model (4) to

$$Y - X_{-j}\hat{\beta}_{-j} = X_j(\beta_j + b_j) + X_{-j}(\beta_{-j} - \hat{\beta}_{-j}) + X_{-j}b_{-j} + \varepsilon. \tag{16}$$

If we were in the unconfounded case, i.e. $\delta = 0$, then $b = 0$, and the above equation would simplify to

$$Y - X_{-j}\hat{\beta}_{-j} = X_j(\beta_j) + X_{-j}(\beta_{-j} - \hat{\beta}_{-j}) + \nu.$$

Then, if we could construct a $v \in \mathbb{R}^n$ such that v is approximately orthogonal to X_{-j} and $v^\mathsf{T} X_j \approx 1$, we could right multiply by v^T to get

$$\begin{aligned} v^\mathsf{T}(Y - X_{-j}\hat{\beta}_{-j}) &\approx \beta_j + \underbrace{v^\mathsf{T} X_{-j}(\beta_{-j} - \hat{\beta}_{-j})}_{\approx 0} + v^\mathsf{T}\nu \\ &\overset{\mathrm{d}}{\approx} \mathrm{N}(\beta_j, \|v\|_2^2 \sigma^2), \end{aligned}$$

[3] See discussion in Guo et al. (2020) for a possible relaxation of this assumption.

to get an asymptotically normally distributed estimator of β_j. This method of *debiasing* the Lasso stems from C.-H. Zhang and S. S. Zhang (2014), where v is chosen to be proportional to the residuals from regressing X_j onto X_{-j} by using the Lasso. For conditions under which the above can be made precise, see C.-H. Zhang and S. S. Zhang (2014). In particular, one crucial condition is that the linear regression model $X_j = X_{-j}\gamma + \kappa_j$ satisfies that γ is sparse (i.e., X_j is conditionally independent of most of the columns of X_{-j}).

Returning to our Confounded Linear Model and the decomposition (16), we see immediately that the debiasing method from C.-H. Zhang and S. S. Zhang (2014) cannot be applied verbatim. The presence of confounders can cause spurious correlations between elements of $X_{i,(\cdot)}$, which violate the assumptions in C.-H. Zhang and S. S. Zhang (2014). Moreover, the penultimate term in (16) need not be approximately orthogonal to v since the bias term $X_{-j}\beta_{-j}$ can be large. Recalling that the Trim transform reduces both the bias term and correlations between covariates, we see that these two issues may be addressed by applying a Trim transformation to (16).

Let $X_{-j} = U^{(j)}D^{(j)}(V^{(j)})^\mathsf{T}$ be the singular value decomposition of X_{-j}, with $r = \min(n, p-1)$. Let τ_j be the $\lfloor rt \rfloor$-th largest singular value of X_{-j} and let

$$F^{(j)} \coloneqq U^{(j)} \cdot \operatorname{diag}\left(\min(\tau_j, D^{(j)}_{1,1})/D^{(j)}_{1,1}, \ldots, \min(\tau_j, D^{(j)}_{n-1,n-1})/D^{(j)}_{n-1,n-1}\right) \cdot (U^{(j)})^\mathsf{T}. \tag{19}$$

Then $F^{(j)}$ is the Trim transform corresponding to X_{-j}. Borrowing intuition from the previous subsection, we should expect $X_{-j}b_{-j}$ to be small, and also that regressing $F^{(j)}X_j$ onto $F^{(j)}X_{-j}$ using the Lasso should yield a consistent estimator of the conditional linear dependence between E_j and E_{-j} if the underlying relationship is sparse. This is indeed the case under conditions similar to Theorem 3.7, see Guo et al. (2020), Proposition 3.

Transforming equation (16) by $F^{(j)}$, we have that

$$F^{(j)}(Y - X_{-j}\hat{\beta}_{-j}) = F^{(j)}X_j(\beta_j + b_j) + F^{(j)}X_{-j}(\beta_{-j} - \hat{\beta}_{-j}) + F^{(j)}X_{-j}b_{-j} + F^{(j)}\varepsilon. \tag{20}$$

Let $\hat{\gamma}$ be the estimated coefficients from Lasso regressing $F^{(j)}X_j$ onto $F^{(j)}X_{-j}$:

$$\hat{\gamma} \coloneqq \arg\min_{\gamma} \left\{ \frac{1}{2n}\|F^{(j)}X_j - F^{(j)}X_{-j}\gamma\|_2^2 + \lambda \sum_{j=1}^{p-1} \frac{\|F^{(j)}X_j\|_2}{\sqrt{n}}|\gamma_j| \right\}. \tag{21}$$

Let Z_j be the residuals of the regression (21), i.e., $Z_j = X_{-j} - \hat{\gamma}X_{-j}$. We take our vector v to be the transformed and rescaled residuals,

$$v \coloneqq F^{(j)}Z_j \frac{1}{(F^{(j)}Z_j)^\mathsf{T}F^{(j)}X_j}.$$

Noting that $v^\mathsf{T}F^{(j)}X_j(\beta_j + b_j) = \beta_j + b_j$, we get from equation (20) that

$$v^\mathsf{T}F^{(j)}(Y - X_{-j}\hat{\beta}_{-j}) - \beta_j = b_j + v^\mathsf{T}F^{(j)}X_{-j}(\beta_{-j} - \hat{\beta}_{-j}) + v^\mathsf{T}F^{(j)}X_{-j}b_{-j} + v^\mathsf{T}F^{(j)}\varepsilon. \tag{22}$$

We will in Theorem 3.12 see that $v^\mathsf{T}F^{(j)}\varepsilon$ dominates all the other terms on the right hand side (RHS) of (22) under appropriate conditions. The *Doubly Debiased Lasso* estimator of β_j is therefore defined by

$$\hat{\beta}_j^{\mathrm{DD}} \coloneqq v^\mathsf{T}F^{(j)}(Y - X_{-j}\hat{\beta}_{-j}). \tag{23}$$

The Doubly Debiased Lasso (proposed by Guo et al. 2020) has its name because it adjusts for both the inherent bias of the Lasso and the bias originating from the confounding. Thus, it is "doubly debiased."

The variance of $\hat{\beta}_j^{\mathrm{DD}}$ is equal to the variance of the RHS of equation (23). Recall that the variance term ε from the marginal formulation (4) of the Confounded Linear Model is $\varepsilon = \nu + \Delta$, where $\Delta \coloneqq H\delta - Xb$, and ν is the error term in the non-marginal Confounded Linear Model. It will turn out that Δ will be negligible compared to ν under certain conditions. With this in mind, and noting that v and $F^{(j)}$ only depend on X, we define

$$\begin{aligned} V &\coloneqq \mathrm{Var}\left(v^{\mathsf{T}} F^{(j)} \varepsilon | X\right) \\ &= \sigma_\nu^2 v^{\mathsf{T}} (F^{(j)})^2 v. \end{aligned} \tag{24}$$

Note that we suppress the dependence on j in the definition of V. Bringing everything together, and recalling that $\nu \sim \mathrm{N}(0, \sigma_\nu^2 I)$, we end up with

$$\frac{1}{\sqrt{V}}(\hat{\beta}_j^{\mathrm{DD}} - \beta_j) \sim \mathrm{N}(0,1) + o_{\mathrm{p}}(1),$$

conditionally on X. Before we formalize the above in a Theorem, we need to lay out some conditions for the above arguments to hold. We keep j fixed throughout.

Let γ be the "true" regression coefficient of $X_{1,j}$ onto $X_{1,-j}$, i.e., $\gamma = (\mathrm{Cov}(X_{1,-j}))^{-1}\,\mathbb{E}(X_{1,-j}X_{1,j})$, and let $\eta_{i,j}$ be the corresponding residuals, $\eta_{i,j} = X_{i,j} - X_{i,-j}^{\mathsf{T}}\gamma$, with variance denoted by σ_j^2. Consider the following assumptions.

(A1*) The covariance matrix Σ_E of each row of E is invertible and there are positive constants c_0, c_1 such that $c_0 \leq \lambda_{\min}(\Sigma_E) \leq \lambda_{\max}(\Sigma_E) \leq c_1$. Furthermore, the j-th column of the precision matrix Σ_E^{-1} is sparse, i.e., for some $k \in \mathbb{N}$, $\|(\Sigma_E^{-1})_{j,(\cdot)}\|_0 \leq k$.

(A2*) The q-th largest singular value of Γ_{-j} satisfies $\lambda_q(\Gamma_{-j}) \gtrsim \sqrt{p}$, and $\max(\|\Gamma(\Sigma_E^{-1})_{j,(\cdot)}\|_2, \|\Gamma_j\|_2, \|\Gamma_{-j}(\Sigma_E^{-1})_{-j,j}\|_2, \|\delta\|_2) \lesssim \sqrt{q}(\log(p))^c$, for some $0 < c \leq 1/4$.

(A3*) The noise term ν in the Linear Confounded Model (2) satisfies $\nu \sim \mathrm{N}(0, \sigma_\nu^2 I)$ and is independent of X and H. The noise term $\eta_{i,j}$ is independent of $X_{i,-j}$. Moreover, $((E_{i,(\cdot)})^{\mathsf{T}}, \nu_i, \eta_{i,j})$ is a sub-Gaussian random vector with sub-Gaussian norm $M_0 > 0$, and each entry $X_{i,j}$ of X is sub-Gaussian with parameter M_0.

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

$$\begin{aligned} \mathrm{RE}(\frac{1}{n}X^{\mathsf{T}}F^2X^{\mathsf{T}}) &= \inf_{\mathcal{T}\subset[p], |\mathcal{T}|\leq s\omega\in\mathbb{R}^p, \|\omega_{\mathcal{T}^c}\|_1 \leq C\|\omega_{\mathcal{T}}\|} \frac{\omega^{\mathsf{T}}(\frac{1}{n}X^{\mathsf{T}}F^2X)\omega}{\|\omega\|_2^2} \geq \tau_*, \\ \mathrm{RE}(\frac{1}{n}X_{-j}^{\mathsf{T}}(F^{(j)})^2X_{-j}^{\mathsf{T}}) &= \inf_{\mathcal{T}_j\subset[p]-\{j\}, |\mathcal{T}_j|\leq k\omega\in\mathbb{R}^{p-1}, \|\omega_{\mathcal{T}_j^c}\|_1 \leq C\|\omega_{\mathcal{T}_j}\|} \frac{\omega^{\mathsf{T}}(\frac{1}{n}X_{-j}^{\mathsf{T}}(F^j)^2X_{-j})\omega}{\|\omega\|_2^2} \geq \tau_*. \end{aligned}$$

Remark 3.11. The first assumption **(A1*)** ensures that if we remove the factor term in the definition of X in (3), then each covariate $X_{i,j}$ only conditionally correlates with a sparse subset of the other covariates $\{X_{i,k} : k \neq j\}$. Thus, the assumption states that the covariates are sparse linear combinations of each other plus a common term depending on the same factors. Assumption **(A1*)** also ensures that Σ_E does not have spiked eigenvalues. The assumption **(A2*)** is a refinement of our original dense confounding assumption and is extended so that it also applies to the Trimmed Lasso for regressing $X_{i,j}$ onto the remaining covariates. Assumption **(A2*)** also ensures that δ is not too large. Assumption **(A4*)** replaces our previous Compatibility Condition with a similar Restricted

Eigenvalue assumption. Assumption **(A4*)** can be shown to hold with high probability under some additional conditions on the growth rate of p, see Proposition 5 in Guo et al. (2020) for details. The assumptions **(A1*)** - **(A4*)** also ensure that results very similar the ones in the previous subsection hold as well. We will not prove nor state these similar results, and instead we refer to Guo et al. (2020).

We formalize the arguments above in the following Theorem.

Theorem 3.12. *Consider the Confounded Linear Model (2), (3) under assumptions **(A1*)** - **(A4*)**. Assume that* $\lim p/n > 0$, $s \ll \sqrt{\min(n, p/q)}/\log(p)$, $k \ll n/\log(p)$, *and* $q \ll \min\{\sqrt{n}/(\log p)^{3/4}, n/[k(\log p)^{3/2}], p/[n \log p]\}$. *Let the penalty terms for the Trimmed Lasso regressions (15) and (21) be* $\lambda \geq A\sigma_\nu\sqrt{\log(p)/n} + \sqrt{q\log(p)/p}$ *and* $\lambda_j \geq A\sigma_j\sqrt{\log(p)/n}$, *respectively, where A is some sufficiently large constant. Suppose the trim levels τ,τ_j for the Trim transforms (8), (19) satisfy* $\min(\tau, \tau_j) \geq (3q+1)/\min(n, p-1)$.

Then, with the Doubly Debiased Lasso $\hat{\beta}_j^{DD}$ *as in (23), and V as in (24), we have that*

$$\frac{1}{\sqrt{V}}(\hat{\beta}_j^{DD} - \beta_j) \xrightarrow{d} N(0,1). \tag{25}$$

Proof.
Note: The proof follows the proof of Theorem 1 in Guo et al. (2020). I have fixed some typos in the original proof (B_b is incorrectly defined and σ_ν is missing some places) and I provide more intermediary calculations and explanations. In particular, the application of Markov's inequality is not shown (not even stated).

Recall from (22) that we may write

$$\frac{1}{\sqrt{V}}(\hat{\beta}_j^{\mathrm{DD}} - \beta_j) = \underbrace{\frac{1}{\sqrt{V}}b_j + \frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}X_{-j}b_{-j}}_{:=B_b} + \underbrace{\frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}X_{-j}(\beta_{-j} - \hat{\beta}_{-j})}_{:=B_\beta} + \frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\varepsilon. \tag{26}$$

It can be shown that the bias terms B_b and B_β both tend to zero in probability, (roughly) because $b_j \to 0$ and that v is approximately orthogonal to the columns of $F^{(j)}X_{-j}$. For details, see Lemma 3 and Lemma 7 in Guo et al. (2020). Thus, we only need to consider the right-most term in equation (26). Recalling that $\varepsilon = \nu + \Delta$, we have that

$$\frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\varepsilon = \frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\nu + \frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\Delta.$$

Since the term $v^\mathsf{T}F^{(j)}\nu$ is Gaussian conditional on X and V is its conditional variance, it immediately follows that $\frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\nu \overset{\mathrm{d}}{=} \mathrm{N}(0,1)$ conditional on X. Since the limiting distribution does not depend on X, we conclude that it holds marginally as well. Furthermore, we have that

$$\left|\frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\Delta\right| = \frac{1}{\sigma_\nu}\left|\frac{v^\mathsf{T}F^{(j)}\Delta}{\|F^{(j)}v\|_2}\right|,$$

and by the (non-stochastic) Cauchy-Schwarz inequality,

$$\leq \frac{\|\Delta\|_2}{\sigma_\nu}.$$

We will apply Markov's inequality to show that $\frac{\|\Delta\|_2}{\sigma_\nu} = \mathrm{o_p}(1)$, which combined with the convergence of B_b and B_β implies that $\frac{1}{\sqrt{V}}(\hat{\beta}_j^{\mathrm{DD}} - \beta_j) = Z + \mathrm{o_p}(1)$, where $Z \sim \mathrm{N}(0,1)$, and the conclusion (25)

follows. By the independence between individual terms of Δ,

$$\frac{1}{n}\,\mathbb{E}\|\Delta\|_2^2 = \mathbb{E}(\Delta_i^2).$$

Recall that $\Delta_i = H_{i,(\cdot)}\delta - X_{i,(\cdot)}b = H_{i,(\cdot)}\delta - (E_{i,(\cdot)} + H_{i,(\cdot)}\Gamma)b$, and that $E_{i,(\cdot)}$ and $H_{i,(\cdot)}$ are mean zero and independent with respective covariances Σ_E and I_q. Since $\Sigma = \Sigma_E + \Gamma^\mathsf{T}\Gamma$ and $b = \Sigma^{-1}\Gamma^{\mathsf{T}}\delta$, we get that

$$\begin{aligned}
\frac{1}{n}\,\mathbb{E}\|\Delta\|_2^2 &= \mathrm{Var}(H_{i,(\cdot)}\delta - X_{i,(\cdot)}b) \\
&= \mathrm{Var}\left(H_{i,(\cdot)}(\delta - \Gamma b) - E_{i,(\cdot)}b\right) \\
&= (\delta - \Gamma b)^\mathsf{T} I\,(\delta - \Gamma b) + b^\mathsf{T}\Sigma_E b \\
&= \delta^\mathsf{T}\delta - \delta^\mathsf{T}\Gamma\Sigma{-}1\Gamma^{\mathsf{T}}\delta \\
&= \delta^\mathsf{T}\left(I - \Gamma\Sigma^{-1}\Gamma^\mathsf{T}\right)\delta.
\end{aligned}$$

It can be shown that $\delta^\mathsf{T}\left(I - \Gamma\Sigma^{-1}\Gamma^\mathsf{T}\right)\delta \lesssim q\sqrt{\log p}/p$ by an application of Woodbury's matrix inversion formula. See Lemma 2 in Guo et al. (2020) for details. By Markov's inequality, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}(\|\Delta\|_2/\sigma_\nu \geq t) &\leq \frac{\mathbb{E}\|\Delta\|_2^2}{t^2\sigma_\nu^2} \\
&= \frac{n\delta^\mathsf{T}\left(I - \Gamma\Sigma^{-1}\Gamma^\mathsf{T}\right)\delta}{t^2\sigma_\nu^2} \\
&\lesssim \frac{nq\sqrt{\log(p)}\,p}{t^2\sigma_\nu^2}.
\end{aligned}$$

Taking $t = \sqrt{\frac{nq\log(p)}{p\sigma_\nu^2}}$ gives that with probability larger than $1 - (\log p)^{-1/2}$,

$$\left|\frac{1}{\sqrt{V}}v^\mathsf{T}F^{(j)}\Delta\right| \lesssim \sqrt{\frac{nq\log(p)}{p\sigma_\nu^2}}. \tag{27}$$

Since we have assumed that $q \ll \frac{p}{n\log p}$ and σ_ν is fixed, the RHS of (27) tends to zero, and we are done. □

As hinted to at the beginning of this subsection, in Theorem 3.12 we need to increase the penalty term for the Trimmed Lasso (21) slightly since we are no longer assuming that $\varepsilon \perp\!\!\!\perp X$. The bounds on the growth rates of the model parameters in Theorem 3.12 ensure that β and γ are sufficiently sparse and that the number of confounding variables q is sufficiently small. The lower bounds on the Lasso penalty terms ensure that the Lasso regressions (15) and (21) are sufficiently sparse and thus consistent. The lower bounds on τ,τ_j ensures that the Trim transforms do not reduce too many large singular values of X and X_j.

Remark 3.13. van de Geer et al. (2013) propose an alternative version of the Debiased Lasso, which estimates the whole coefficient vector β instead of an individual entry and has a limiting multivariate normality (under no confounding). This alternative Debiased Lasso can be used to construct asymptotic confidence regions. A natural question to ask is whether we can extend the Doubly Debiased Lasso in the same way (*note: this is my own idea*). Unfortunately, it is not immediately obvious how one should approach this, since a crucial part of the Doubly Debiased Lasso is to transform by the Trim transform $F^{(j)}$, which depends on j. To the author's knowledge, no such extension has been proposed in the literature.

Remark 3.14. Recall that the standard debiasing of the Lasso proposed by C.-H. Zhang and S. S. Zhang (2014) requires that the j-th covariate $X_{i,j}$ conditionally correlates with only a sparse subset of other covariates $\{X_{i,k} : k \neq j\}$. This assumption may, in practice, be unrealistic. However, the Doubly Debiased Lasso relaxes this assumption by allowing the covariates to also correlate with some common unobserved factors (adjusting for the factors we still require sparse conditional correlation between the covariates). Therefore, it can be reasonable to use the Doubly Debiased Lasso even when we know that confounding is not present, as it is robust to correlations among the covariates caused by common factors. This additional robustness property of the Doubly Debiased Lasso is not explicitly mentioned in Guo et al. (2020) (*i.e., this is my own observation*).

In order to use Theorem 3.12 to construct an asymptotic confidence interval for β_j, we will need to estimate the variance σ_ν^2 of the error term ν_i in the Confounded Linear Model (2), as σ_ν^2 appears in the asymptotic variance of $\hat{\beta}_j^{\mathrm{DD}}$. However, our *observed* model is the marginal Confounded Linear Model (4), which has error term $\varepsilon = \nu + \Delta$. By the independence between ν and $\{X, H\}$, we have that

$$\begin{aligned}
\sigma^2 &= \mathrm{Var}(\varepsilon_i) \\
&= \mathrm{Var}(\nu_i + \Delta_i) \\
&= \mathrm{Var}(\nu_i) + \mathrm{Var}(\Delta_i) \\
&= \mathrm{Var}(\nu_i) + \frac{1}{n}\mathbb{E}\|\Delta\|_2^2 \\
&= \sigma_\nu^2 + \delta^\mathsf{T}\left(I - \Gamma\Sigma^{-1}\Gamma^\mathsf{T}\right)\delta,
\end{aligned}$$

and, thus,

$$\sigma^2 - \sigma_\nu^2 = \delta^\mathsf{T}\left(I - \Gamma\Sigma^{-1}\Gamma^\mathsf{T}\right)\delta. \tag{33}$$

As already mentioned in the proof of Theorem 3.12, we have by Lemma 2 in Guo et al. (2020) that the RHS of (33) tends to zero under the conditions of Theorem 3.12. It is, therefore, reasonable to estimate the variance of the observed error term ε_i as a proxy estimator of σ_ν^2.

Recalling the Trim transformed marginal Confounded Linear Model (5), and letting $\hat{\beta}$ be the Trimmed Lasso as in (15), we have that

$$\tilde{Y} - \tilde{X}\hat{\beta} = \tilde{\varepsilon} + \tilde{X}(\beta - \hat{\beta}) + \tilde{X}b.$$

Since $\hat{\beta}$ can consistently estimate β, we should expect $(\beta - \hat{\beta})$ to be small. In the proof of Theorem 3.7, we saw that $\tilde{X}b$ is small as well. Thus, we should expect

$$\|\tilde{Y} - \tilde{X}\hat{\beta}\|_2^2 \approx \|\tilde{\varepsilon}\|_2^2.$$

Taking an expectation, we get that

$$\begin{aligned}
\mathbb{E}\|\tilde{Y} - \tilde{X}\hat{\beta}\|_2^2 &\approx \mathbb{E}\|\tilde{\varepsilon}\|_2^2 \\
&= \mathbb{E}\|F\varepsilon\|_2^2 \\
&= \mathrm{Tr}(\mathrm{Cov}(F\varepsilon)) \\
&= \sigma^2\,\mathrm{Tr}\big(F^2\big).
\end{aligned}$$

This suggests using

$$\hat{\sigma}^2 := \frac{\|\tilde{Y} - \tilde{X}\hat{\beta}\|_2^2}{\mathrm{Tr}(F^2)} \tag{34}$$

as an estimator for σ^2 (and hence σ_ν^2). We have the following Proposition, which is a simplification of Proposition 1 in Guo et al. (2020).

Proposition 3.15. *Consider the Confounded Linear Model (2). With $\hat{\sigma}^2$ as in (34), and under the assumptions of Theorem 3.12, we have that*

$$\hat{\sigma}^2 \xrightarrow{p} \sigma_\nu^2.$$

The proof of Proposition 3.15 formalizes our heuristic arguments above, and so we omit the proof and refer to Guo et al. (2020) for details.

By combining Theorem 3.12 and Proposition 3.15, we can construct an asymptotically valid confidence interval for β_j. Let $\hat{V}$ be the estimated version of V, that is,

$$\begin{aligned} \hat{V} &\coloneqq \hat{\sigma}_\nu^2 v^\mathsf{T}(F^{(j)})^2 v \\ &= \hat{\sigma}_\nu^2 \frac{Z_j^\mathsf{T}(F^{(j)})^4 Z_j}{\left[Z_j^\mathsf{T}(F^{(j)})^2 X_j\right]}. \end{aligned} \tag{35}$$

Then, a confidence interval for β_j with asymptotic coverage $1-\alpha$ is given by

$$\left(\hat{\beta}_j^{\mathrm{DD}} - z_{1-\alpha/2}\sqrt{\hat{V}}, \hat{\beta}_j^{\mathrm{DD}} + z_{1-\alpha/2}\sqrt{\hat{V}}\right), \tag{36}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

In practice, choosing the penalty term λ_j minimizing the squared prediction error via cross validation (as we also do for λ) works reasonably well, and is what we do in the simulations. We also remark in passing that, in the high-dimensional case with some mild assumptions on the trimming value τ being large enough, the above confidence interval is efficient in the Gauss-Markov sense. See Theorem 2 in Guo et al. (2020) for details.

Although it follows from Theorem 3.12 and Proposition 3.15 that $\frac{1}{\hat{V}}(\hat{\beta}_j^{\mathrm{DD}} - \beta_j) \xrightarrow{\mathrm{d}} \mathrm{N}(0,1)$ and thus that the confidence interval (36) is asymptotically valid (and even efficient), we will see in our simulations that the convergence can be somewhat slow. In particular, the estimator $\hat{\sigma}$ has a heavy left tail and often tends to underestimate σ_ν (and σ) even for moderate sample sizes ($n = 200$). This property is highly unfavorable when constructing confidence intervals, as we typically wish to guard against falsely inferring $\beta_j \neq 0$ when the opposite is true. This motivates us to search for another estimator $\hat{\sigma}$ that is less likely to underestimate σ_ν.

Estimating the variance term in a high-dimensional linear regression model is a difficult task even without the presence of confounding. Reid et al. (2013) study several proposed variance estimators, and through a numerical study, they find that underestimating the variance is a common trait among most of them.

I propose the following simple and intuitive estimator, inspired by Fan et al. (2010), which does not tend to underestimate σ_ν. Define the K folds of the observed data by the partition $\mathcal{D}_1, \ldots, \mathcal{D}_K \subset [n]$. Let

$$\hat{\sigma}_{\mathrm{fold}}^2 \coloneqq \min_\lambda \sum_{k=1}^K \sum_{i \in \mathcal{D}_i} \frac{\|\tilde{Y} - \tilde{X}\hat{\beta}_\lambda^{(-k)}\|_2^2}{\mathrm{Tr}(F^2)}, \tag{37}$$

where F is our usual Trim transform, $\tilde{\cdot}$ denotes transformation by F, and $\hat{\beta}_\lambda^{(-k)}$ is the Trimmed Lasso (15) when the k-th fold is omitted and with penalization term λ.

Simulations show that the proposed estimator $\hat{\sigma}_{\mathrm{fold}}$ tends to overestimate σ_ν slightly. Importantly, though, it seldom underestimates σ_ν, and also exhibits less variability than the estimator $\hat{\sigma}$ from (34)

proposed by Guo et al. (2020). Additionally, the Doubly Debiased Lasso estimator $\hat{\beta}_j^{\text{DD}}$ has a finite sample bias. Therefore, using a conservative estimator like $\hat{\sigma}_{\text{fold}}$ is reasonable, as it guards against too low coverage of confidence intervals. Indeed, simulations show that using $\hat{\sigma}_{\text{fold}}$ in the place of $\hat{\sigma}$ in the confidence interval (36) can make the finite sample coverage closer to level α.

3.3 Multiple testing

In this subsection, we will move our attention to multiple hypothesis testing. We will consider a slightly different model than before, and we will employ a different confounder adjustment technique. We will still model confounding through a factor model, but this time we will estimate the factor loadings directly in order to adjust for the confounding. See the end of this subsection for a more detailed comparison between the confounder adjustment techniques employed in this subsection and the Trimmed and Doubly Debiased Lasso. We will follow J. Wang et al. (2017), and all results in this subsection are results from the paper unless otherwise is stated.

A highly cited paper by Leek and Storey (2008) suggests that the following factor model can represent multiple hypothesis testing based on linear regression. Let

$$Y = X\beta^{\mathsf{T}} + H\,\Gamma^{\mathsf{T}} + E, \tag{38}$$

where $Y \in \mathbb{R}^{n\times p}$, $X \in \mathbb{R}^n$, $H \in \mathbb{R}^{n\times q}$ and $E \in \mathbb{R}^{n\times p}$ all have i.i.d. rows, $\beta \in \mathbb{R}^p$ is the parameter of interest, $\Gamma^{\mathsf{T}} \in \mathbb{R}^{q\times p}$ are factor loadings, and $E \perp\!\!\!\perp \{H, X\}$. The primary variable of interest is X.

We emphasize that the model (38) is not the same as the Confounded Linear Model from the previous subsections. Notice that each side of (38) has p columns (Y is a matrix). Informally, each column represents regressing a column of Y onto the vector X. Therefore, the model (38) represents p separate linear regressions, which is not the same as the Confounded Linear Model (which is a multiple linear regression model).

Remark 3.16. The variables X and Y in (38) do not necessarily have the same interpretations as they usually have in linear models (such as the Confounded Linear Model), where Y is the "response" variable and X are the "predictors"/"covariates." In some situations, the variable X in model (38) can represent what we usually think of as the "response." For instance, for a medical study, X can be the prevalence of a particular disease, while Y can be different gene expressions, i.e., what we usually think of as "predictors". In other cases, the variable X in (38) can represent what we usually think of as a "predictor", and Y can contain many "responses." For instance, for another medical study, X can represent a medical treatment while the columns of Y are measurements of multiple different symptoms.

The term $H\Gamma^{\mathsf{T}}$ in the model (38) essentially captures the dependence between testing individual components of β that arises from performing individual tests based on the same data (the same X). However, the model (38) does not capture dependence arising from confounding. Therefore, we additionally assume that the H's are confounders:

$$H = X\alpha^{\mathsf{T}} + W, \tag{39}$$

where $\alpha \in \mathbb{R}^q$ is an unknown parameter (column) vector and $W \in \mathbb{R}^{n\times q}$ is a random error matrix. As for the distributions of the random variables, we will assume that X_i is i.i.d. with mean 0 and variance 1 for all i. Furthermore, we will assume that each entry of W is i.i.d. $\text{N}(0,1)$, and that each row of E is i.i.d. $\text{N}_p(0,\Sigma)$ with $\Sigma = \text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$. Lastly, we will assume that W is independent of X (recalling that we already assumed that E is independent of X and H). We will see later that we will depend heavily on the normality assumptions.

We will consider an asymptotic regime where n and p are allowed to grow while the number q of confounders stays fixed. Bai and Ng (2002) show that q may be consistently estimated under the assumptions we will impose later (and some additional technical assumptions). We will, therefore, assume that q is known.

Remark 3.17. In contrast with the Confounded Linear Model, it is not obvious how H in the model (38) and (39) can represent a confounder in the language of Causal Inference. Indeed, the mapping $X \mapsto X\alpha^\mathsf{T}$ is generally not invertible, and thus, X cannot generally be written as a function of H and an independent error term. Therefore, the model given by (38) and (39) cannot arise from a non-parametric structural equations model (see Pearl 2009) where H is a parent of X. We therefore think of H as a confounder in a non-causal sense.

We will use the following Lemma, which simplifies more general results found in Bai and Li (2012), Theorem 5.2, 5.4, and identifying condition IC3. We use tildes to avoid a mix-up with the model (39).

Lemma 3.18. *Suppose the random matrix $\tilde{Y} \in \mathbb{R}^{n\times p}$ originates from the following factor model with non-random factors $\tilde{H}$.*

$$\tilde{Y} = \tilde{H}\tilde{\Gamma}^\mathsf{T} + \tilde{E},$$

where $\tilde{H} \in \mathbb{R}^{n\times q}$ is a matrix of fixed real numbers whose sample variance is I_q, $\tilde{\Gamma}^\mathsf{T} \in \mathbb{R}^{q\times p}$ are factor loadings, and $\tilde{E} \in \mathbb{R}^{n\times p}$ has i.i.d. $N(0,\tilde{\Sigma})$ rows with $\tilde{\Sigma} = diag(\tilde{\sigma}_1^2,\ldots,\tilde{\sigma}_p^2)$.

Letting S denote the sample covariance matrix of the rows in $\tilde{Y}$, define the quasi-log likelihood ℓ_{quasi} by

$$\ell_{quasi}(\tilde{\Sigma},\tilde{\Gamma};\tilde{Y}) = -\frac{1}{2p}\log\det\left(\tilde{\Gamma}\tilde{\Gamma}^\mathsf{T}+\tilde{\Sigma}\right) - \frac{1}{2p}\operatorname{tr}\left(S\left[\tilde{\Gamma}\,\tilde{\Gamma}^\mathsf{T}+\tilde{\Sigma}\right]^{-1}\right). \tag{40}$$

Moreover, assume that the following conditions hold.

> *(B1) The limits $\lim_{p\to\infty}\frac{1}{p}\tilde{\Gamma}^\mathsf{T}\tilde{\Sigma}^{-1}\tilde{\Gamma}$ and $\lim_{p\to\infty}\sum_{j=1}^p \tilde{\sigma}_j^{-4}((\tilde{\Gamma}_{j,(\cdot)})^\mathsf{T}\otimes(\tilde{\Gamma}_{j,(\cdot)})^\mathsf{T})(\tilde{\Gamma}_{j,(\cdot)}\otimes\tilde{\Gamma}_{j,(\cdot)})$ exists and are positive definite matrices.*
>
> *(B2) There is a constant $C>0$ such that $\|\Gamma_{j,(\cdot)}\|_2 \le C$, and $C^{-2} \le \tilde{\sigma}_j^2$, $\hat{\tilde{\sigma}}_j \le C^2$ for all j, where $\hat{\tilde{\sigma}}_j$ are the estimated variances.*

Assume also that $\frac{1}{p}\tilde{\Gamma}^\mathsf{T}\tilde{\Sigma}^{-1}\tilde{\Gamma}$ is diagonal. Then, as $n,p\to\infty$, with q fixed,

$$\sqrt{n}(\hat{\Gamma}_{j,(\cdot)} - \Gamma_{j,(\cdot)}) \xrightarrow{d} N(0,\sigma_j^2 I_q),$$

and

$$\sqrt{n}(\hat{\sigma}_j^2 - \sigma_j^2) \xrightarrow{d} N(0,2\sigma_j^4).$$

Lemma 3.18 gives us the (marginal) asymptotic distributions of the columns in $\hat{\tilde{\Gamma}}$ and variances $\hat{\tilde{\sigma}}_j^2$ when the estimators are the maximizers of the quasi-log likelihood. Notice that each column of $\tilde{\Gamma}$ represents the dependence of each column of $\tilde{Y}$ on the factors.

Let us return to our confounded multiple testing model given by (38) and (39). To decouple the estimation of β and α from Γ and Σ, we will rotate (38) by an appropriately defined rotation matrix.

The reason for doing so will become clear shortly. The reader is reminded that for a matrix A, A_j denotes the j-th column of A, while $A_{j,(\cdot)}$ is the j-th row *as a row vector*. Letting Q^T be the Householder matrix that maps X to $\|X\|_2 e_1$ [4] and left-multiplying equation (38) by Q^T, we get that

$$\tilde{Y} \coloneqq Q^\mathsf{T} Y = \|X\|_2 e_1 \beta^\mathsf{T} + \tilde{H}\Gamma^\mathsf{T} + \tilde{E},$$

where

$$\tilde{H} \coloneqq Q^\mathsf{T} H = \|X\|_2 e_1 \alpha^\mathsf{T} + \tilde{W}.$$

Since Q^T is unitary and E,W have i.i.d Gaussian rows, we have that $\tilde{E} \overset{\mathrm{d}}{=} E$ and $\tilde{W} \overset{\mathrm{d}}{=} W$ (unconditionally on X). The first row of $\tilde{Y}$ is given by

$$\tilde{Y}_{1,(\cdot)} = \|X\|_2 \beta^\mathsf{T} + \tilde{H}_{1,(\cdot)}\Gamma^\mathsf{T} + \tilde{E}_{1,(\cdot)} \ \sim_{|X} \mathrm{N}(\|X\|_2(\beta + \Gamma\alpha)^\mathsf{T}, \Sigma + \Gamma\Gamma^\mathsf{T}), \tag{41}$$

conditional on X, and the rest of the rows are given by

$$\tilde{Y}_{-1,(\cdot)} = \tilde{H}_{-1,(\cdot)}\Gamma^\mathsf{T} + \tilde{E}_{-1,(\cdot)}, \tag{42}$$

whose rows are i.i.d. $\mathrm{N}(0, \Sigma + \Gamma\Gamma^\mathsf{T})$, conditional on X.

We now see why the choice of Q^T is sensible; equation (42) is almost the type of model that Lemma 3.18 applies to, and equation (42) does not depend on α or β. This suggests estimating Γ and Σ by applying Lemma 3.18 to equation (42), plugging the resulting estimates into equation (41) and estimate α and β. There are two obstacles we need to overcome. Firstly, the parameters α and β not identifiable with our current assumptions. Secondly, the conditions of Lemma 3.18 are nearly, but not entirely, satisfied. We will deal with these two obstacles in order.

For the first obstacle, J. Wang et al. (2017) present two different assumptions that make α and β identifiable; sparsity of β, or negative controls. We will focus our attention on the latter because it is the simplest. To begin, J. Wang et al. (2017) give some mild conditions on the factor loadings Γ^T making them identifiable up to a rotation (see Lemma 2.1):

> (B3) If any row of Γ is removed, there remain two disjoint sub-matrices of Γ of rank q,
>
> (B4) $\Gamma^\mathsf{T}\Sigma^{-1}\Gamma/p$ is diagonal, and the diagonal elements are distinct, positive and arranged in decreasing order.

Using the identifiability of Γ^T (up to a rotation), J. Wang et al. (2017) show that the following assumption on α ensures identifiability: There exists a set of *negative controls* $\mathcal{C}$ such that $\beta_\mathcal{C} = 0$ and $\mathrm{rank}(\Gamma_{C,(\cdot)}) = q$. See J. Wang et al. (2017), Proposition 2.1, for a short and elegant proof of this fact. We will use these negative controls directly in our estimation procedure later.

The second obstacle requires a bit more work than the first one. We cannot immediately apply Lemma 3.18 to equation (42) in order to estimate Γ and Σ. The reason is that $\tilde{H}_{-1,(\cdot)}$ and Γ^T do not satisfy the hypothesis of the Lemma. However, can estimate Σ and Γ when Γ is estimated up to a linear transformation by the following argument. Choose an invertible matrix $R \in \mathbb{R}^{q\times q}$ such that $\frac{1}{n-1}(\tilde{H}_{-1,(\cdot)}(R^{-1})^\mathsf{T})^\mathsf{T}(\tilde{H}_{-1,(\cdot)}(R^{-1})^\mathsf{T}) = I_q$ and $\frac{1}{p}(\Gamma R)^\mathsf{T}\Sigma^{-1}(\Gamma R)$ is diagonal[5] . It then follows that $RR^\mathsf{T} = \frac{1}{n-1}\tilde{H}^\mathsf{T}_{-1,(\cdot)}\tilde{H}_{-1,(\cdot)}$. We emphasize that R is an unobserved random matrix depending on

[4] The explicit expression for the relevant Householder matrix is $Q^\mathsf{T} = I_n - 2vv^\mathsf{T}$, where $v = \frac{X - \|X\|_2 e_1}{\|X - \|X\|_2 e_1\|_2}$.

[5] J. Wang et al. (2017) call the matrix R a rotation matrix, which is ambiguous since we do not in general have $R^\mathsf{T} = R^{-1}$. Furthermore, the existence of R is not explicitly shown in J. Wang et al. (2017), and we will in this essay take the existence as given.

$\tilde{H}_{-1,(\cdot)}$ and X. Defining $\tilde{H}^{(0)}_{-1,(\cdot)} \coloneqq \tilde{H}_{-1,(\cdot)}(R^{-1})^\mathsf{T}$ and $\Gamma^{(0)} \coloneqq \Gamma R$, we may rewrite equation (42) in the following way:

$$\begin{aligned}\tilde{Y}_{-1,(\cdot)} &= \tilde{H}_{-1,(\cdot)}\Gamma^\mathsf{T} + \tilde{E}_{-1,(\cdot)} \\ &= \tilde{H}_{-1,(\cdot)}(R^{-1})^\mathsf{T}R^\mathsf{T}\Gamma^\mathsf{T} + \tilde{E}_{-1,(\cdot)} \\ &= \tilde{H}^{(0)}_{-1,(\cdot)}(\Gamma^{(0)})^\mathsf{T} + \tilde{E}_{-1,(\cdot)}. \end{aligned} \tag{45}$$

The hypotheses of Lemma 3.18 is satisfied with our choice of $\Gamma^{(0)}$, $\tilde{H}^{(0)}_{-1,(\cdot)}$ and $\tilde{Y}_{-1,(\cdot)}$. Letting $\hat{\Gamma}$ and $\hat{\Sigma} = \text{diag}(\hat{\sigma_1^2}, \dots, \hat{\sigma_p^2})$ be the maximizers of (40) applied to $\tilde{Y}_{-1,(\cdot)}$ in equation (45), we have that as $n, p \to \infty$, $\sqrt{n}(\hat{\Gamma}_{j,(\cdot)} - \Gamma^{(0)}_{j,(\cdot)}) \xrightarrow{\text{d}} \text{N}(0, \sigma_j^2 I_q)$ and $\sqrt{n}(\hat{\sigma}_j^2 - \sigma_j^2) \xrightarrow{\text{d}} \text{N}(0, 2\sigma_j^4)$, conditional on $\tilde{H}_{-1,(\cdot)}$ and X. We will later see that $R \xrightarrow{\text{a.s.}} I_q$, which implies that $\Gamma^{(0)} \xrightarrow{\text{a.s.}} \Gamma$. Although the Lemma applies to *non-random* factors H, the limiting distribution does not depend on H, so it holds when the H's are random as well.

Now that we know the properties of the quasi-ML estimators of Γ (up to a transformation) and Σ, we return to the estimation of α and β in equation (41). First, we rewrite the first row of $\tilde{Y}$ so that we are dealing with $\Gamma^{(0)}$ instead of Γ. We have

$$\begin{aligned}\tilde{Y}_{1,(\cdot)} &= \|X\|_2\beta^\mathsf{T} + \tilde{H}_{1,(\cdot)}\Gamma^\mathsf{T} + \tilde{E}_{1,(\cdot)} \\ &= \|X\|_2(\beta^\mathsf{T} + \alpha^\mathsf{T}\Gamma^\mathsf{T}) + \tilde{W}_{1,(\cdot)}\Gamma^\mathsf{T} + \tilde{E}_{1,(\cdot)} \\ &= \|X\|_2\beta^\mathsf{T} + (\|X\|_2\alpha^\mathsf{T} + \tilde{W}_{1,(\cdot)})\Gamma^\mathsf{T} + \tilde{E}_{1,(\cdot)} \\ &= \|X\|_2\beta^\mathsf{T} + (\|X\|_2\alpha^\mathsf{T} + \tilde{W}_{1,(\cdot)})RR^\mathsf{T}\Gamma^\mathsf{T} + \tilde{E}_{1,(\cdot)}. \end{aligned}$$

Defining $\alpha^{(0)} \coloneqq R^\mathsf{T}(\alpha + \tilde{W}_{1,(\cdot)}/\|X\|_2)$, we get that

$$\tilde{Y}^\mathsf{T}_{1,(\cdot)}/\|X\|_2 = \beta + \Gamma^{(0)}\alpha^{(0)} + \tilde{E}^\mathsf{T}_{1,(\cdot)}/\|X\|_2. \tag{46}$$

Recall that we assumed the existence of a set $\mathcal{C}$ of "negative controls" such that $\beta_\mathcal{C} = 0$. Splitting up equation (46) according to $\mathcal{C}$, we have that

$$\tilde{Y}^\mathsf{T}_{1,\mathcal{C}}/\|X\|_2 = \Gamma^{(0)}_{\mathcal{C},(\cdot)}\alpha^{(0)} + \tilde{E}^\mathsf{T}_{1,\mathcal{C}}/\|X\|_2, \tag{47}$$

$$\tilde{Y}^\mathsf{T}_{1,-\mathcal{C}}/\|X\|_2 = \beta_{-\mathcal{C}} + \Gamma^{(0)}_{-\mathcal{C},(\cdot)}\alpha^{(0)} + \tilde{E}^\mathsf{T}_{1,-\mathcal{C}}/\|X\|_2. \tag{48}$$

Notice that equation (47) only depends on $\alpha^{(0)}$, which we can use to our advantage. We summarize the outlined estimation procedure for β here. First, estimate $\Gamma^{(0)}$ and Σ by the quasi-ML estimators $\hat{\Gamma}$ and $\hat{\Sigma}$. Then, insert the estimates into (47) and estimate $\alpha^{(0)}$ by $\hat{\alpha}$ using generalized least-squares. Lastly, insert $\hat{\Gamma}$,$\hat{\Sigma}$ and $\hat{\alpha}$ into equation (48) and estimate $\beta_{-\mathcal{C}}$ by $\hat{\beta}_{-\mathcal{C}}$. That is,

$$(\hat{\Gamma}, \hat{\Sigma}) \coloneqq \underset{(\Gamma,\Sigma)}{\text{argmax}}\ \ell_{\text{quasi}}(\tilde{Y}_{-1,(\cdot)}),$$

$$\hat{\alpha} \coloneqq \left(\hat{\Gamma}^\mathsf{T}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_\mathcal{C}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}^\mathsf{T}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_\mathcal{C}\tilde{Y}^\mathsf{T}_{1,\mathcal{C}}/\|X\|_2, \tag{49}$$

$$\hat{\beta}_{-\mathcal{C}} \coloneqq \tilde{Y}^\mathsf{T}_{1,-\mathcal{C}}/\|X\|_2 - \hat{\Gamma}_{-\mathcal{C},(\cdot)}\hat{\alpha} \tag{50}$$

where we, for convenience, adapt the notation that $\Sigma_\mathcal{C} \coloneqq \Sigma_{\mathcal{C},\mathcal{C}}$ and likewise for $\hat{\Sigma}$.

The following Theorem gives us the limiting distribution of $\hat{\beta}$.

Theorem 3.19. *Assume that the conditions (B1), (B2) in Lemma 3.18 apply to Γ and Σ, in addition to the identifying conditions (B3) and (B4). Assume also that $\lim_{p\to\infty} |\mathcal{C}|^{-1}\Gamma_{\mathcal{C},(\cdot)}^{\mathsf{T}}\Sigma_{\mathcal{C}}^{-1}\Gamma_{\mathcal{C},(\cdot)}$ exists and is positive definite. Let $\mathcal{S}$ be any fixed set of indices that is disjoint with $\mathcal{C}$ and of finite cardinality. Then, if $n, p \to \infty$ and $p/n^k \to 0$ for some $k > 0$, we have*

$$\sqrt{n}(\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}) \xrightarrow{d} N\big(0, (1 + \|\alpha\|_2^2)(\Sigma_{\mathcal{S}} + \Delta_{\mathcal{S}})\big)\,, \tag{51}$$

where $\Delta_{\mathcal{S}} \coloneqq \lim_{p\to\infty} \Gamma_{\mathcal{S},(\cdot)}(\Gamma_{\mathcal{C},(\cdot)}^{\mathsf{T}}\Sigma_{\mathcal{C}}^{-1}\Gamma_{\mathcal{C},(\cdot)})^{-1}\Gamma_{\mathcal{S},(\cdot)}^{\mathsf{T}}$. If in addition $|\mathcal{C}| \to \infty$, then

$$\sqrt{n}(\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}) \xrightarrow{d} N\big(0, (1 + \|\alpha\|_2^2)\Sigma_{\mathcal{S}}\big)\,. \tag{52}$$

Proof.
Note: The proof follows the proof of Theorem 3.1 in J. Wang et al. (2017). I provide more intermediary calculations and try to clarify the argument.

We will show that the limiting distribution (51) holds conditional on W, X, and H. Since the limiting distribution does not depend on either of these quantities, the result also holds unconditionally.

By using $p/n^k \to 0$, one can extend Lemma 3.18 and show that

$$\sqrt{n}\left(\hat{\Gamma}_{\mathcal{V},(\cdot)} - \Gamma_{\mathcal{V},(\cdot)}^{(0)}\right) \xrightarrow{\mathrm{d}} Z, \tag{53}$$

for any fixed index set $\mathcal{V}$, where $Z \in \mathbb{R}^{|\mathcal{V}|\times q}$ is a random matrix with i.i.d. columns distributed as $\mathrm{N}(0, \Sigma_{\mathcal{V}})$. One can also show that the entries of $\hat{\Sigma} - \Sigma$ and $\Gamma - \Gamma^{(0)}$ converge *uniformly* to zero in probability. The proof of these two facts builds upon the proof of Lemma 3.18 in Bai and Li (2012) and is omitted. See Lemma A.1 in J. Wang et al. (2017) and its proof for details.

By the Strong Law of Large Numbers (SLLN),

$$\frac{1}{\sqrt{n}}\|X\|_2 \xrightarrow{\text{a.s.}} \sqrt{\mathbb{E}\left(X_i^2\right)} = 1.$$

Recalling that $RR^{\mathsf{T}} = \frac{1}{n-1}\tilde{H}_{-1,(\cdot)}^{\mathsf{T}}\tilde{H}_{-1,(\cdot)}$ and $\tilde{H}_{-1,(\cdot)}$ has i.i.d. $\mathrm{N}(0, I_q)$ rows, SLLN also ensures that

$$RR^{\mathsf{T}} \xrightarrow{\text{a.s.}} I_q.$$

Lemma (A.1) in Bai and Li (2012) then gives that $R \xrightarrow{\text{a.s.}} I_q$ as well (indeed, insert R into the Lemma and pass the limit, which we can do on a set of probability 1). The rest of the proof is relatively simple but somewhat messy. We explicitly show intermediary steps for the reader to follow more easily.

Using the definition of $\hat{\beta}_{\mathcal{S}}$ and $\hat{\alpha_{\mathcal{S}}}$ from equations (49) and (50), we have that

$$\begin{aligned}
\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}} &= -\beta_{\mathcal{S}} + \tilde{Y}_{1,\mathcal{S}}^{\mathsf{T}}/\|X\|_2 - \hat{\Gamma}_{\mathcal{S},(\cdot)}\hat{\alpha} \\
&= \beta_{\mathcal{S}} - \beta_{\mathcal{S}} + \Gamma_{\mathcal{S},(\cdot)}^{(0)}\alpha^{(0)} + \tilde{E}_{1,\mathcal{S}}^{\mathsf{T}}/\|X\|_2 - \hat{\Gamma}_{\mathcal{S},(\cdot)}\hat{\alpha} \\
&= \tilde{E}_{1,\mathcal{S}}^{\mathsf{T}}/\|X\|_2 + \Gamma_{\mathcal{S},(\cdot)}^{(0)}\alpha^{(0)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\hat{\alpha} \\
&= \tilde{E}_{1,\mathcal{S}}^{\mathsf{T}}/\|X\|_2 + \Gamma_{\mathcal{S},(\cdot)}^{(0)}\alpha^{(0)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}_{\mathcal{C},(\cdot)}^{\mathsf{T}}\hat{\Sigma}_{\mathcal{C}}^{-1}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}_{\mathcal{C},(\cdot)}^{\mathsf{T}}\hat{\Sigma}_{\mathcal{C}}^{-1}\tilde{Y}_{1,\mathcal{C}}^{\mathsf{T}}/\|X\|_2 \\
&= \tilde{E}_{1,\mathcal{S}}^{\mathsf{T}}/\|X\|_2 + \Gamma_{\mathcal{S},(\cdot)}^{(0)}\alpha^{(0)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}_{\mathcal{C},(\cdot)}^{\mathsf{T}}\hat{\Sigma}_{\mathcal{C}}^{-1}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}_{\mathcal{C},(\cdot)}^{\mathsf{T}}\hat{\Sigma}_{\mathcal{C}}^{-1}\left(\Gamma_{\mathcal{C},(\cdot)}^{(0)}\alpha^{(0)} + \tilde{E}_{1,\mathcal{C}}^{\mathsf{T}}/\|X\|_2,\right) \\
&= \tilde{E}_{1,\mathcal{S}}^{\mathsf{T}}/\|X\|_2 - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}_{\mathcal{C},(\cdot)}^{\mathsf{T}}\hat{\Sigma}_{\mathcal{C}}^{-1}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}_{\mathcal{C},(\cdot)}^{\mathsf{T}}\hat{\Sigma}_{\mathcal{C}}^{-1}\tilde{E}_{1,\mathcal{C}}^{\mathsf{T}}/\|X\|_2
\end{aligned}$$

$$
\begin{aligned}
&\quad + \Gamma^{(0)}_{\mathcal{S},(\cdot)}\alpha^{(0)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\Gamma^{(0)}_{\mathcal{C}(\cdot)}\alpha^{(0)} \\
&= \tilde{E}^{\intercal}_{1,\mathcal{S}}/\|X\|_2 - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\tilde{E}^{\intercal}_{1,\mathcal{C}}/\|X\|_2 \\
&\quad + \Gamma^{(0)}_{\mathcal{S},(\cdot)}\alpha^{(0)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}(-\hat{\Gamma}_{\mathcal{C},(\cdot)} + (\Gamma^{(0)}_{\mathcal{C},(\cdot)} - \hat{\Gamma}_{\mathcal{C},(\cdot)}))\alpha^{(0)} \\
&= \tilde{E}^{\intercal}_{1,\mathcal{S}}/\|X\|_2 - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\tilde{E}^{\intercal}_{1,\mathcal{C}}/\|X\|_2 \\
&\quad + \left(\Gamma^{(0)}_{\mathcal{S},(\cdot)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\right)\alpha^{(0)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\left(\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}((\Gamma^{(0)}_{\mathcal{C},(\cdot)} - \hat{\Gamma}_{\mathcal{C},(\cdot)}))\alpha^{(0)}.
\end{aligned}
$$

Thus, by multiplying by $\sqrt{n}$ on both sides, we have

$$
\begin{aligned}
\sqrt{n}\left(\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\right) = \tilde{E}^{\intercal}_{1,\mathcal{S}}\underbrace{\frac{\sqrt{n}}{\|X\|_2}}_{\xrightarrow{\text{a.s.}}1} - &\underbrace{\hat{\Gamma}_{\mathcal{S},(\cdot)}}_{\Gamma_{S,(\cdot)}+o_p(1)}\underbrace{\left(\frac{1}{|\mathcal{C}|}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}}_{(\frac{1}{|\mathcal{C}|}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)})^{-1}+o_p(1)}\underbrace{\frac{1}{|\mathcal{C}|}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\tilde{E}^{\intercal}_{1,\mathcal{C}}}_{\frac{1}{|\mathcal{C}|}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\tilde{E}^{\intercal}_{1,\mathcal{C}}+o_p(1)}\underbrace{\frac{\sqrt{n}}{\|X\|_2}}_{\xrightarrow{\text{a.s.}}1} \\
&+ \sqrt{n}\left(\Gamma^{(0)}_{\mathcal{S},(\cdot)} - \hat{\Gamma}_{\mathcal{S},(\cdot)}\right)\underbrace{\alpha^{(0)}}_{\xrightarrow{p}\alpha} \\
&- \underbrace{\hat{\Gamma}_{\mathcal{S},(\cdot)}}_{\Gamma_{S,(\cdot)}+o_p(1)}\underbrace{\left(\frac{1}{|\mathcal{C}|}\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\hat{\Gamma}_{\mathcal{C},(\cdot)}\right)^{-1}}_{(\frac{1}{|\mathcal{C}|}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)})^{-1}+o_p(1)}\underbrace{\hat{\Gamma}^{\intercal}_{\mathcal{C},(\cdot)}\hat{\Sigma}^{-1}_{\mathcal{C}}\sqrt{n}(\Gamma^{(0)}_{\mathcal{C},(\cdot)} - \hat{\Gamma}_{\mathcal{C},(\cdot)})}_{\frac{1}{|\mathcal{C}|}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\sqrt{n}(\Gamma^{(0)}_{\mathcal{C},(\cdot)}-\hat{\Gamma}_{\mathcal{C},(\cdot)})+o_p(1)}\underbrace{\alpha^{(0)}}_{\xrightarrow{p}\alpha},
\end{aligned}
$$

where all underbraces follow from the fact that $R \xrightarrow{\text{a.s.}} I_q$, the uniform convergence of $\hat{\Gamma}$ and $\hat{\Sigma}$, and the Continuous Mapping Theorem. Since $\tilde{E} \overset{\text{d}}{=} E$ (recall that each entry of E are independent due to normality) and $\mathcal{C} \cap \mathcal{S} = \varnothing$, we have that $\tilde{E}_{1,(\cdot)} \perp\!\!\!\perp \hat{\Gamma}$ and $\tilde{E}_{C,(\cdot)} \perp\!\!\!\perp \tilde{E}_{S,(\cdot)}$. By the above equation, the result in equation (53) and Slutsky's Theorem, we get that

$$
\sqrt{n}\left(\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}\right) \xrightarrow{\text{d}} Z_1 + Z_2 + Z_3 + Z_4,
$$

where $\{Z_j : j \in [4]\}$ are mutually independent random vectors with distributions

- $Z_1 \sim \text{N}(0, \Sigma_{\mathcal{S}})$,
- $Z_2 \sim \text{N}(0, \lim_{p\to\infty}\Gamma_{\mathcal{S},(\cdot)}(\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)})^{-1}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Sigma_{\mathcal{C}}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)}(\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)})^{-1}\Gamma^{\intercal}_{\mathcal{S},(\cdot)})$,
- $Z_3 \sim \text{N}(0, \Sigma_{\mathcal{S}}\|\alpha\|_2^2)$,
- $Z_4 \sim \text{N}(0, \lim_{p\to\infty}\Gamma_{\mathcal{S},(\cdot)}(\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)})^{-1}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Sigma_{\mathcal{C}}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)}(\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)})^{-1}\Gamma^{\intercal}_{\mathcal{S},(\cdot)}\|\alpha\|_2^2)$.

Canceling terms and inserting for $\Delta_{\mathcal{S}}$ in the above we get our desired result,

$$
\sqrt{n}(\hat{\beta}_{\mathcal{S}} - \beta_{\mathcal{S}}) \xrightarrow{\text{d}} \text{N}\left(0, (1 + \|\alpha\|_2^2)(\Sigma_{\mathcal{S}} + \Delta_{\mathcal{S}})\right).
$$

The final statement in the Theorem follows from the fact that since $\lim_{p\to\infty}|\mathcal{C}|^{-1}\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)}$ exists and is positive definite, the minimum eigenvalue of $\Gamma^{\intercal}_{\mathcal{C},(\cdot)}\Sigma^{-1}_{\mathcal{C}}\Gamma_{\mathcal{C},(\cdot)}$ tends to infinity as $|\mathcal{C}| \to \infty$. Then the entries of $\Delta_{\mathcal{S}}$ tend uniformly to 0, and the result follows.

□

Remark 3.20. The asymptotic variance of $\hat{\beta}_{\mathcal{S}}$ in (52) is the same as the variance of the "oracle" estimator that observes (X, Y, H) and estimates $\beta_{\mathcal{S}}$ by least-squares according to equations (39) and (38). This remarkable fact shows the power of Theorem 3.19 since the estimation of $\beta_{\mathcal{S}}$ by $\hat{\beta}_{\mathcal{S}}$ is done without observing the confounders H.

Remark 3.21. J. Wang et al. (2017) prove similar results under the alternative identifying assumption of sparsity of β. Moreover, they also show that results similar to Theorem 3.19 hold when we include nuisance covariates to the model (38) (i.e., additional covariates whose coefficients are not of interest).

Unfortunately, the limiting distribution in Theorem 3.19 only holds for fixed index sets $\mathcal{S}$ of finite cardinality. The following Theorem shows that, despite this, we can control both the average type-I error and the family-wise error rate (FWER) when testing each entry of β simultaneously. We state the Theorem without proof, noting in passing that the Theorem does not follow trivially from Theorem 52. We emphasize that the Theorem gives us a way to perform multiple hypothesis testing in the presence of confounders in a theoretically grounded way.

Theorem 3.22. *Consider simultaneously testing $H_{0,j} : \beta_j = 0$ vs $H_{1,j} : \beta_j \neq 0$ for $j \in [p]$ using t-statistics*

$$t_j = \frac{\|X\|_2 \hat{\beta}_j}{\hat{\sigma}_j \sqrt{1 + \|\hat{\alpha}\|_2^2}}.$$

Let $\mathcal{N}_p \coloneqq \{j \in [p] \ : \ \beta_j = 0\}$ be the index set of true null hypotheses. Under the assumptions of Theorem 3.19, as $n, p, |\mathcal{C}| \to \infty$,

$$\frac{1}{\mathcal{N}_p} \sum_{j \in \mathcal{N}_p} I\left(|t_j| > z_{1-\gamma/2}\right) \xrightarrow{p} \gamma,$$

and

$$\limsup \mathbb{P}\left(\sum_{j \in \mathcal{N}_p} I(|t_j| > z_{1-\gamma/(2p)}) \geq 1 \right) \leq \gamma,$$

where $z_{1-\gamma/2}$ is the $1-\gamma/2$ quantile of the standard normal distribution and $\gamma \in (0,1)$ is the test level.

We end this subsection with a short comparison between the deconfouning approach taken in this subsection and one taken in the previous two subsections. Recall that the Trimmed Lasso adjusted for confounding effects in a rather rough way; it "trims" the singular values of the design matrix. In this subsection, we have seen a more refined approach by estimating the factor loadings, although at the price of stronger parametric assumptions, notably Gaussianity. In this subsection, we also fixed the number of confounding variables and assumed the existence of negative controls for identifiability, noting that a sparsity assumption would work as well. The conditions in this subsection are not all stronger than the assumptions for the Trimmed Lasso and the Doubly Debiased Lasso, though. Identifiability in subsection 3.1 followed immediately from assumptions implying that the confounding bias vanishes in the limit, which we do not require here. Furthermore, we did not require any "dense confounding assumption" in this subsection, although we did impose identifiability conditions (B3) and (B4) on the factor loadings Γ.

4 Discussion: deconfounding methods for non-linear models

In section 3, we had to make strong assumptions to remove confounding effects successfully in estimation. For instance, in subsections 3.1 and 3.2, we relied upon linearity, sparsity, he "dense confounding" assumption and the assumption that a factor model can describe the relationship between the covariates and the confounders. In subsection 3.3, we relied upon similar assumptions except that the sparsity assumption was replaced with a "negative control" assumption. We did not need the "dense confounding" assumption either (although we had to impose some other conditions on the factor loadings).

A natural question to ask is whether we can generalize the methods in section 3. For instance, we may wish to modify the Confounded Linear Model (2) from subsection 3.1 and allow the response Y to follow a Generalized Linear Model (GLM) with linear predictor $X\beta + H\delta$, while letting the covariates X and the confounders H follow a linear factor model as in equation (3). There is a literature on the asymptotic properties of the Lasso extended to GLMs, see, for instance, Bühlmann and van de Geer (2011) chapter 6. Moreover, van de Geer et al. (2013) show that the debiased Lasso (not to be confused with the Doubly Debiased Lasso) can be extended to unconfounded GLMs. The deconfounding techniques presented in subsections 3.1 and 3.2 do not, however, immediately carry over into the GLM case. The reason is that both the Trimmed Lasso and the Doubly Debiased Lasso rely on the Trim transform, which is a linear transformation. Therefore, novel approaches are needed in order to extend the Trimmed Lasso and the Doubly Debiased Lasso to the more complicated case of a GLM. To the author's knowledge, no such extensions of the Trimmed Lasso or the Doubly Debiased Lasso have been proposed (*i.e., these are my own ideas*).

The deconfounding approach taken in subsection 3.3 relied on the high-dimensional factor loading and variance estimation techniques from Bai and Li (2012). There has recently been some progress on the estimation of non-linear factor models. For instance, F. Wang (2020) estimate factor loadings for non-linear high-dimensional factor models by a maximum likelihood approach (similar to Bai and Li 2012) and prove limiting normality under some strengthened conditions. This motivates modifying the method from 3.3 to build upon these recent developments and allow the relationship (38) to be non-linear. However, there is no obvious way to do this, as the deconfounding approach in 3.3 relied on linearity; recall that we decoupled estimation of the parameters of interest from the factor loadings and variance terms by rotating equation (38) (which is a linear operation). To the author's knowledge, no non-linear extensions of J. Wang et al. (2017) have been proposed (*i.e., these are my own ideas*).

A notable paper on non-linear confounder adjustment by Y. Wang and Blei (2019) has been subject to considerable attention. Y. Wang and Blei (2019) consider confounder adjustment in a potential outcomes framework (see Imbens and Rubin 2015) and propose an intuitive algorithm called the Deconfounder. We present the algorithm heuristically. Say we are interested in estimating $\mathbb{E}(Y_i(x))$, where Y_i is the potential outcome of a response variable Y_i for a treatment $x \in \mathbb{R}^p$. The variable $Y_i = Y(X_i)$ is the observed outcome, and X_i is random and possibly correlated with $Y_i(x)$. Estimating $\mathbb{E}(Y_i(x))$ a fundamentally difficult problem and is at the heart of Causal Inference. Suppose there is an unmeasured confounder H_i such that $Y_i(x) \perp\!\!\!\perp X_i \mid H_i$. That is, H_i renders the potential outcome function $Y_i(x)$ independent of the realized treatment X_i. Assume further that the dependence between the observed treatments X_i and the confounders H_i are related through a probabilistic factor model,

$$\begin{aligned} H_i &\sim p_H(\,\cdot\,|\alpha), \text{ for } i \in [n], \\ X_{i,j} \mid H_i &\sim p_X(\,\cdot\,|h_i, \theta_j), \text{ for } i \in [n],\ j \in [p], \end{aligned} \tag{54}$$

where p_H and p_X are densities. Under some assumptions such as "no observed single-cause confounders", pinpointability and positivity (and a few more), Y. Wang and Blei (2019) argue that we can estimate a *substitute confounder* $\hat{H}_i$ from the factor model (54) and obtain estimates of $\mathbb{E}(Y_i(x))$ by conditioning on the substitute confounder. That is, they claim that estimating $\mathbb{E}(Y_i \mid X_i, H_i = \hat{H}_i)$

should estimate the causal effect $\mathbb{E}(Y_i(x))$. This is *the Deconfounder* algorithm.

We can draw a parallel between the Deconfounder algorithm and the deconfounding methods from section 3. Recall that the Trimmed Lasso and the Doubly Debiased Lasso relied on a linear factor model to crudely "trim away" the confounding effects. In subsection 3.3, we also relied upon a linear factor model, but we used the estimated factor loadings in order to adjust for the confounding effects[6]. The Deconfounder also depends on a factor model assumption but relaxes the linearity assumption. The Deconfounded takes advantage of the factor model even more directly than we did in subsection 3.3; it estimates the confounder itself (via the factor model) and uses it to estimate the causal effects of interest.

Compared to the methods in section 3, some theoretical properties of the Deconfounder have not been thoroughly proven. In section 3, the identifiability of the parameters of interest was evident. Identifiability in the Deconfounder framework is, however, not clear. Even though a subsequent paper by Y. Wang and Blei (2020) tries to clarify the validity of their claims, Ogburn et al. (2020), Ogburn et al. (2019), and D'Amour (2019) show that identification with the Deconfounder is not always possible by providing counterexamples. Moreover, other aspects of the Deconfounder algorithm have also been criticized. Grimmer et al. (2020) show that stronger conditions than stated in Y. Wang and Blei (2019) are sometimes needed for asymptotic unbiasedness of the Deconfounder. Importantly, they argue that the necessary assumptions also imply asymptotic unbiasedness of a "naive" regression that ignores the confounding altogether.

5 Simulation study

In the following, we will review some numerical simulations to examine the properties of the Trimmed Lasso and the Doubly Debiased Lasso. The simulations have exclusively been performed by the author. The R code can be found here: https://github.com/partiiistudent/essay.

We generate data from the Confounded Linear Model (2), (3). We let $s = 5$ and $\beta^\mathsf{T} = (1, 1, 1, 1, 1, 0, \dots, 0)$. We sample each element of Γ and δ independently from a standard normal distribution for each individual simulation. Furthermore, we let $\nu \sim \mathrm{N}(0, \sigma_\nu^2)$, let each row $E_{j,(\cdot)}$ of E have distribution $\mathrm{N}(0, \sigma_E^2 I_p)$ and let each row $H_{j,(\cdot)}$ of H have distribution $\mathrm{N}(0, I_q)$. Unless otherwise is stated, we let $\sigma_\nu = 1$, $\sigma_E = 2$, $n = 200$ and $p = 600$. These simulation parameters are the same as the ones found in Ćevid et al. (2018). For the Trim transform, we let the trimming levels τ and τ_j (τ_j is the trimming level used by the Doubly Debiased Lasso) be the median singular values of X and X_{-j}, respectively. Unless otherwise stated, the penalty parameters λ and λ_j are chosen by 10-fold cross-validation (CV) minimizing the squared prediction error. All results are based on $N = 2000$ independent simulations.

Trimmed Lasso vs the standard Lasso.
In Figure 2, we see comparisons of the average ℓ_1 error rates of the Trimmed Lasso (red) and the standard Lasso (black) under various values of n. We plot the (average) ℓ_1 norm of the bias term b caused by confounding to get an idea of how strong the confounding is. We notice that under moderate to small confounding (top-left plot), the Trimmed Lasso performs noticeably better than the standard Lasso. Furthermore, the ℓ_1 error of the standard Lasso gets close to $\|b\|_1$ as n increases. When we increase the confounding strength (increasing q to 100, top-right plot), we see that $\|b\|_1$ increases. Furthermore, the error of the standard Lasso increases, but not as much as $\|b\|_1$, since many entries of b may be small, and the standard Lasso encourages sparsity in the coefficient estimates. We also see

[6] We remark again, though, that it is not obvious how the framework of subsection 3.3 could represent a causal model, and is thus not directly comparable to the Decofounder.

that the Trimmed Lasso has a considerably lower ℓ_1 error rate for $n \geq 200$.

When there is no confounding present ($\delta = 0$), we observe that the Trimmed Lasso and the standard Lasso have very similar ℓ_1 error rates, both when X is generated from a factor model ($\Gamma \neq 0$) and when the rows of X are uncorrelated $\Gamma = 0$. This suggests that the Trimmed Lasso introduces confounding robustness to a low cost. If we keep the confounding effects at zero ($\delta = 0$) and multiply each entry of Γ by 6 (or any other number larger than 1), we increase the correlation between elements in each row of X. In this case (Figure 3), we see that the Trimmed Lasso has a considerably lower error rate than the standard Lasso. Thus, the Trimmed Lasso can have better performance than the Lasso even when there is no confounding, as the Trimmed Lasso adjusts for correlations between the covariates.

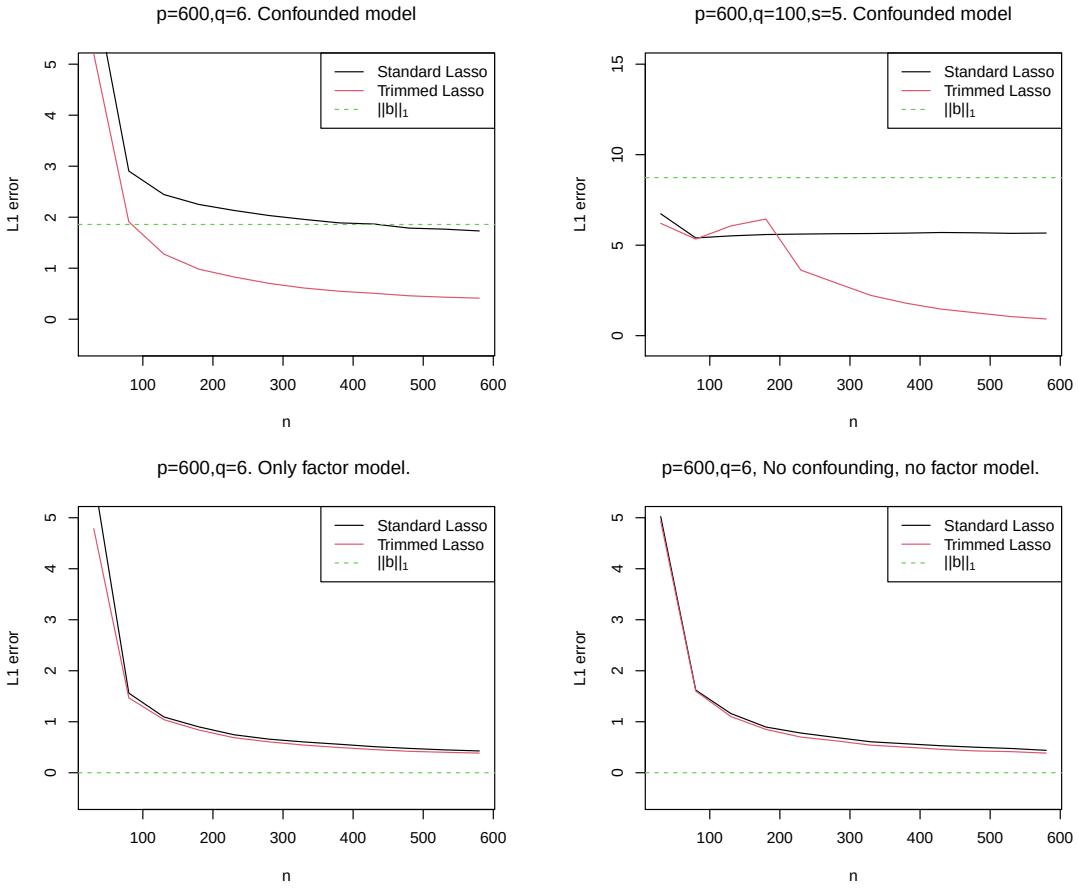

Figure 2: ℓ_1 error rates of the Trimmed Lasso and the standard Lasso for different values of n. We keep $p = 600$ fixed, and consider four different scenarios: 1. moderate confounding (top-left, title "Confounded model"), 2. strong confounding (top-right, title "Confounded model"), 3. no confounding and $X's$ generated from factor model (bottom-left, title "Only factor model"), and 4. no confounding and uncorrelated X's (bottom-right, title "No confounding, no factor model"). The vertical green bar is the average ℓ_1 norm of the bias term b.

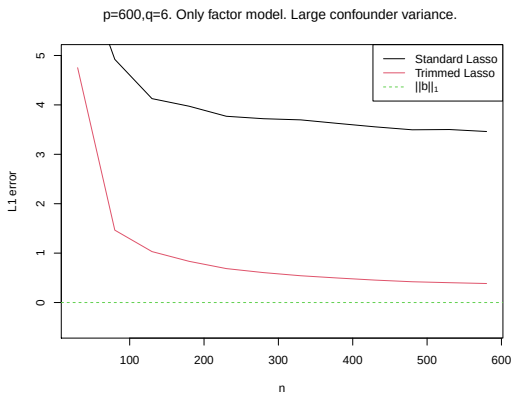

Figure 3: ℓ_1 error rates of the Trimmed Lasso and the standard Lasso for different values of n. Here there is no confounding (i.e., $\delta = 0$), and the X's are generated from a factor model as in (3), with each entry of Γ sampled independently from a $\mathrm{N}(0, 6^2)$ distribution.

Cross-validation choice of the tuning parameter under confounding.
A practical way to choose the penalty parameter λ for the Trimmed Lasso and the standard Lasso is by CV. Under confounding, however, the value of λ chosen by CV (denoted λ_{CV}) will tend to be too small. In Figure 4, we plot the average ℓ_1 errors of the Trimmed Lasso and the standard Lasso after scaling λ_{CV} by different values of $\varphi \in \mathbb{R}$, when $n = 300$. λ_{CV} is chosen by 10-fold CV. Under moderate confounding (left), we see that CV chooses a slightly smaller value of λ than is optimal (in terms of ℓ_1 estimation error). This tendency is more pronounced for the standard Lasso, where the CV choice of λ is noticeably smaller than the optimal value. We also remark that we can improve upon the ℓ_1 estimation error of the standard Lasso under confounding by choosing a penalty λ larger than λ_{CV}.

Under stronger confounding (i.e., a larger number of confounders q), on the right of Figure 4, the tendency of CV to choose a smaller than optimal penalty term is more prominent. Under strong confounding, the optimal value of λ for the Trimmed Lasso is approximately $e \approx 2.7$ times larger than λ_{CV}. For the standard Lasso, the same number is approximately $e^3 \approx 20$. Intuitively, the number of observations n and the number of covariates p are not large enough for the Trimmed Lasso to effectively remove the confounding, which we also see in Figure 3. Thus, as the standard Lasso, the Trimmed Lasso over-fits, although considerably less than the standard Lasso.

Figure 4 also shows that there is no obvious principled way to increase the value of λ_{CV} to prevent minimize the ℓ_1 estimation error. Indeed, the optimal increase in λ_{CV} depends on the confounding strength, which is unknown in practice. Ćevid et al. (2018) suggest increasing λ_{CV} "slightly", but this is perhaps too vague to be of much value in practice. In the following simulations, we increase λ_{CV} by the (somewhat arbitrarily chosen) value 1.2.

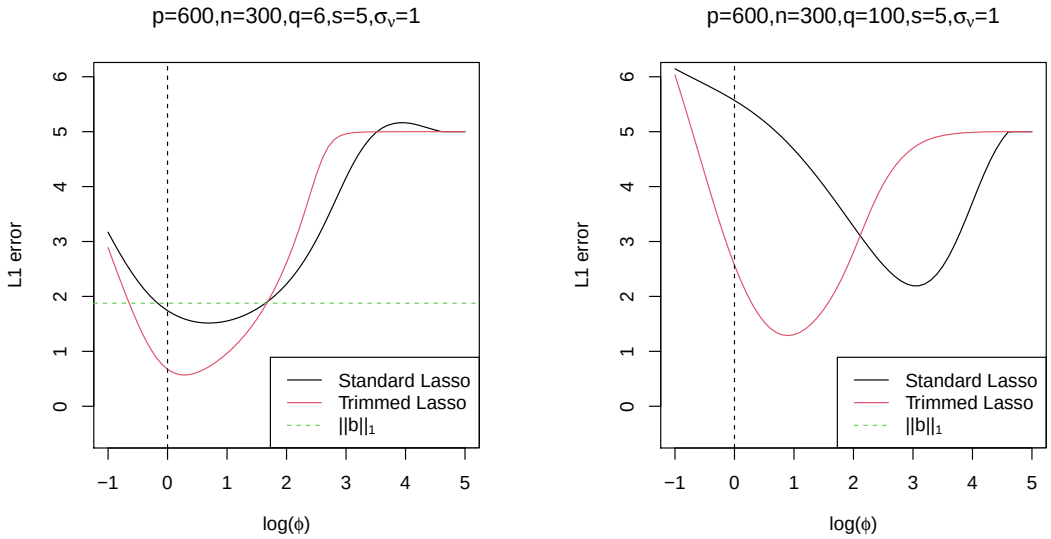

Figure 4: ℓ_1 error of the Trimmed Lasso and the standard Lasso when choosing penalty parameter $\lambda = \varphi\lambda_{\text{CV}}$, plotted against $\log(\varphi)$. To the left, the confounding is moderate ($q = 6$). To the right, the confounding is strong ($q = 30$). In the case $q = 30$, the ℓ_1 norm $\|b\|_1 \approx 8.7$, and is, therefore, not visible in the right plot.

The Doubly Debiased Lasso with two different variance estimators.
Figure 5 shows the empirical distributions of $\hat{\sigma}$ from (34) and my proposed estimator $\hat{\sigma}_{\text{fold}}$ from (37) for $q = 30$ and $\sigma_E = 1$. In Figure 5, the same (randomly sampled) δ and Γ are used for all simulations. The solid vertical line is at $\sigma_\nu = 1$, and the dotted vertical line is at σ, the standard deviation of the error term in the marginal Confounded Linear Model. We see that the estimator $\hat{\sigma}$ is volatile and has a heavy left tail. Among the $N = 2000$ sampled values of $\hat{\sigma}$, a small (although not negligible) proportion is below 0.2. The sampled values of $\hat{\sigma}_{\text{fold}}$ display somewhat opposite properties. None of sampled values are below σ_ν, and we see clearly that $\hat{\sigma}_{\text{fold}}$ tends to overestimate σ_ν. We also notice that $\hat{\sigma}_{\text{fold}}$ is less volatile than $\hat{\sigma}$ and does not have a heavy left tail.

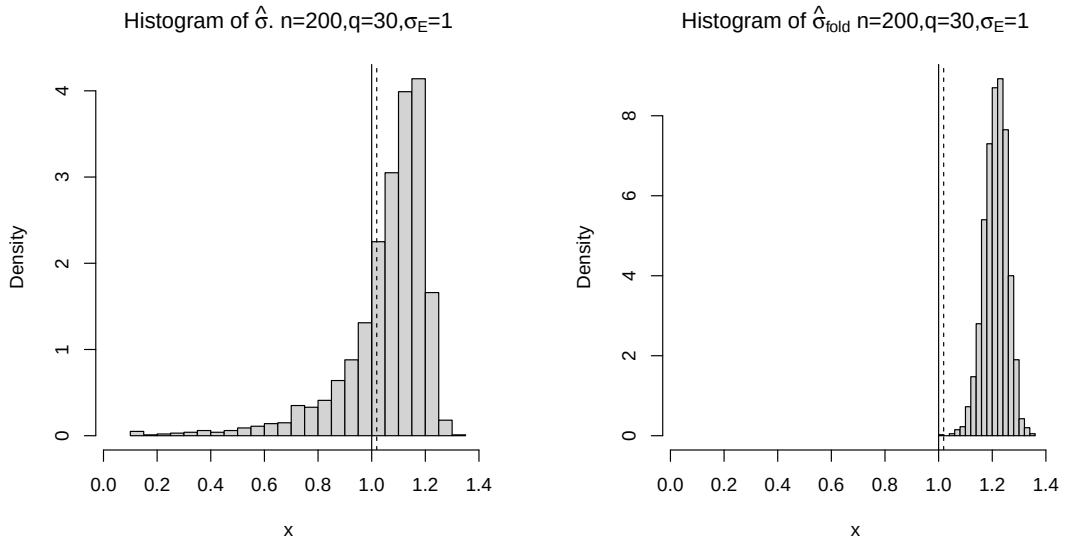

Figure 5: Empirical distribution of $\hat{\sigma}$ (left) and $\hat{\sigma}_{\text{fold}}$ (right), for $q = 30$ and $\sigma_E = 1$.

Recall that we can estimate the standard error of $\hat{\beta}_j^{\text{DD}}$ by $\hat{V}$ defined in (35), an estimator which depends on $\hat{\sigma}$. The heavy left tail of the finite-sample distribution of $\hat{\sigma}$ causes the empirical distribution of $\hat{V}$ to have a heavy left tail as well. Let $\hat{V}_{\text{fold}}$ be the analogous estimator where we replace $\hat{\sigma}$ by my proposed estimator $\hat{\sigma}_{\text{fold}}$. Figure 6 shows histograms of the resulting t-statistics for β_1; $(\hat{\beta}_1^{\text{DD}} - \beta_1)/\hat{V}$ (top), and $(\hat{\beta}_1^{\text{DD}} - \beta_1)/\hat{V}_{\text{fold}}$ (bottom), along with the $N(0, 1)$ density and a kernel density estimate. The values of δ and Γ are the same for all samples. We choose the first index of β simply because the effect of underestimating σ_ν is the most pronounced for coefficients that are truly non-zero.

We make two observations. Firstly, we see from both plots that the mean of $\hat{\beta}_1^{\text{DD}}$ is slightly below the true value due to the finite-sample of the Doubly Debiased Lasso. Remarkably, though, the bias is relatively small since the t-statistics are almost perfectly centered around zero. This could, of course, have been differently for another randomly sampled pair of Γ and δ. Secondly, we observe that the t-statistics using the estimator $\hat{V}$ are heavy-tailed. Indeed, in Figure 6, we see that some samples of the t-statistics are below -10 and that the proportion of such outliers is not negligible. On the other hand, the t-statistics using the estimator $\hat{V}_{\text{fold}}$ do not have such outliers. Furthermore, ignoring the slight bias in the t-statistics, we see that the empirical distribution of the t-statistics based on $\hat{V}_{\text{fold}}$ is closer to the $\text{N}(0, 1)$ distribution than the t-statistics based on $\hat{V}$.

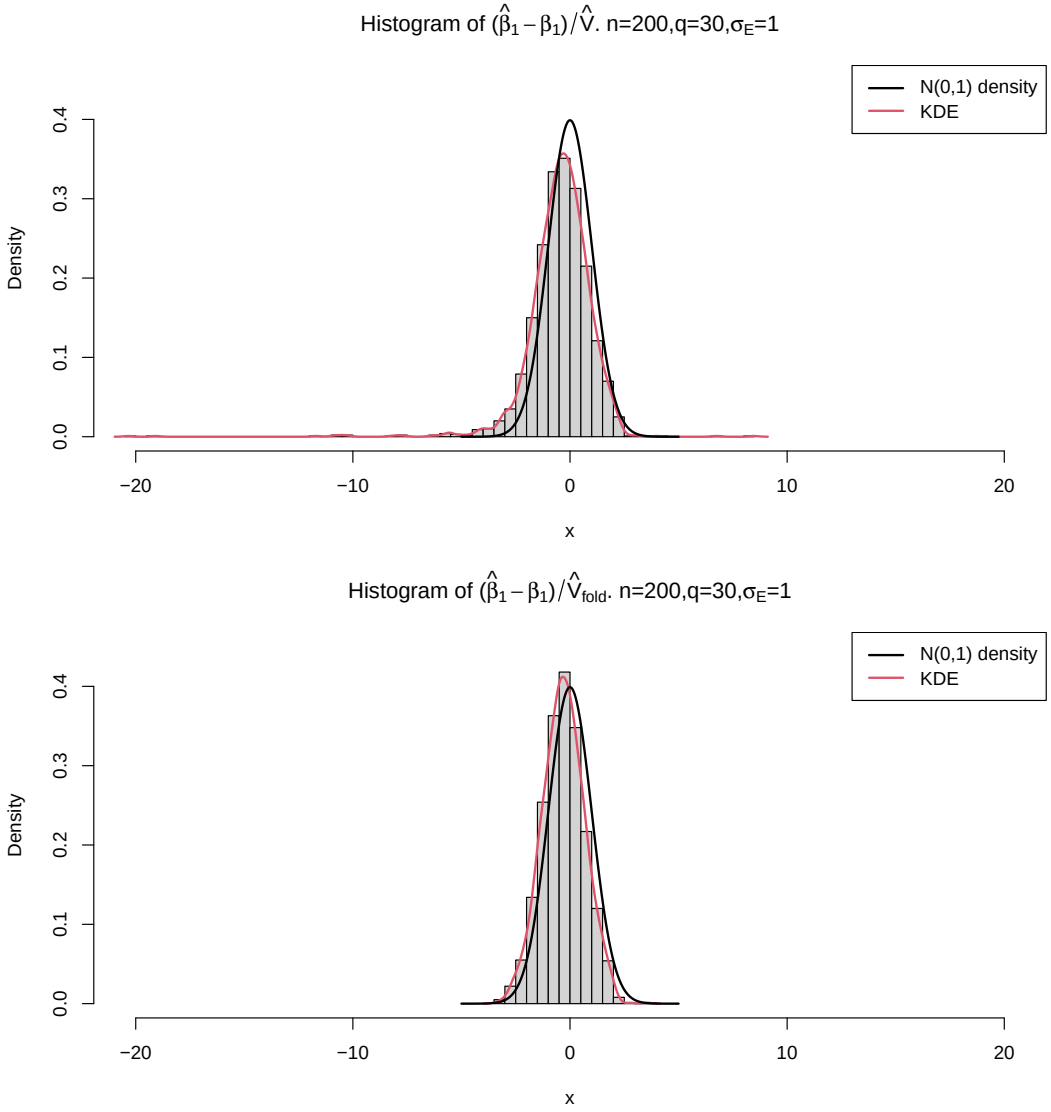

Figure 6: Top: Empirical distribution of the t-statistics $(\hat{\beta}_1^{\mathrm{DD}} - \beta_1)/\hat{V}$, along with the $\mathrm{N}(0,1)$ density (black) and a kernel density estimate (red). Bottom: Empirical distribution of the t-statistics $(\hat{\beta}_1^{\mathrm{DD}} - \beta_1)/\hat{V}_{\mathrm{fold}}$, along with the $\mathrm{N}(0,1)$ density (black) and a kernel density estimate (red). For the definition of $\hat{V}_{\mathrm{fold}}$, see the text above the figure.

In Figure 7, we plot the coverage of level 0.95 confidence intervals (CI's) for $\beta_1 = 1$ (left) and $\beta_{s+1} = 0$ (right) for $n = 200$ fixed. We do not longer fix Γ and δ and instead sample them independently for each iteration. We consider three different types of confidence intervals: (1) CI based on the "standard" Debiased Lasso (see C.-H. Zhang and S. S. Zhang 2014) (black), (2) the CI defined in (36) based on the Doubly Debiased Lasso with variance estimator $\hat{\sigma}^2$ (red), and (3) the CI defined in (36) based on the Doubly Debiased Lasso with variance estimator $\hat{\sigma}^2_{\mathrm{fold}}$ (green). The dashed horizontal line is at the desired confidence level $\alpha = 0.95$. For β_1 we see that all of the three CI's have coverage less than 0.95. The Doubly Debiased CI for β_1 based on $\hat{\sigma}_{\mathrm{fold}}$ is consistently close to level 0.95, with an average coverage (over the p's) of 0.93. The Doubly Debiased CI for β_1 based on $\hat{\sigma}$ is not as close, with an

average coverage of 0.88. The CI for β_1 based on the standard Debiased Lasso is nowhere close to level 0.95, with an average coverage of only 0.33. The three CI's for β_{s+1} display the same tendency, although all three CI's have coverage close to the level 0.95. The CI based on $\hat{\sigma}_{\text{fold}}$ is nonetheless the only of the three CI's to have a coverage of 0.95 or more. We also remark that the Doubly Debiased CI based on $\hat{\sigma}$ is closer to a coverage of 0.95 than the CI based on the Debiased Lasso. This is not surprising due to confounding being present. What is somewhat surprising, however, is the high coverage the CI of β_{s+1} based on the standard Debiased Lasso. This may be because $\beta_{s+1} = 0$ and that the standard Lasso encourages sparsity. Since the confounding is only moderately large in this simulation, the standard Lasso is reasonably able to infer $\beta_{s+1} = 0$. We remark that this could have been different if the confounding effects were stronger.

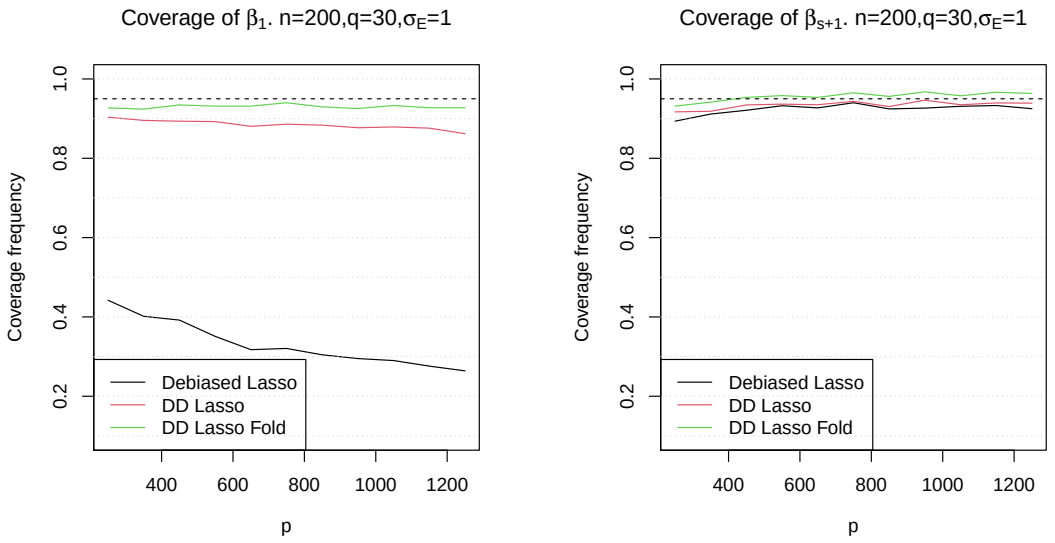

Figure 7: Coverage frequencies of three different confidence intervals for β_1 (left) and for β_{s+1} (right), as a function of p, fixing $n = 200$. The confidence intervals are: (1) CI based on the Debiased Lasso (black), (2) CI based on the Doubly Debiased Lasso and $\hat{\sigma}$ (red), and (3) CI based on the Doubly Debiased Lasso and $\hat{\sigma}_{\text{fold}}$ (green).

We emphasize that the two confidence intervals based on the Doubly Debiased Lasso are too narrow only in finite samples. In Figure 8, we plot the coverage frequencies of the same CIs as in Figure 7, fixing $p = 600$ and letting n vary. We see that, as n increases, the Doubly Debiased CIs based on $\hat{\sigma}$ and $\hat{\sigma}_{\text{fold}}$ approach each other. This suggests that the CI based on $\hat{\sigma}_{\text{fold}}$ is better than the CI based on $\hat{\sigma}$ for smaller samples and not any worse for large samples. We lastly remark that the standard Debiased Lasso is not even close to having coverage 0.95 for β_1 even as n increases.

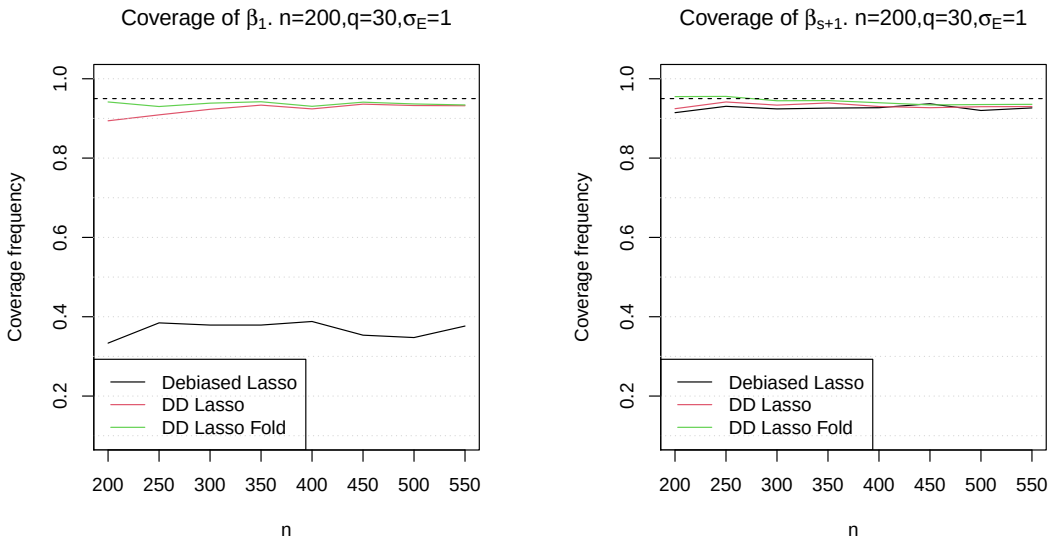

Figure 8: Coverage frequencies of three different confidence intervals for β_1 (left) and for β_{s+1} (right), as a function of n, fixing $p = 600$. The confidence intervals are: (1) CI based on the Debiased Lasso (black), (2) CI based on the Doubly Debiased Lasso and $\hat{\sigma}$ (red), and (3) CI based on the Doubly Debiased Lasso and $\hat{\sigma}_{\text{fold}}$ (green).

6 Concluding remarks

In this essay, we have reviewed different techniques for confounder adjustment. We first considered the Trimmed Lasso, which relied upon "dense confounding" through a factor model. We showed that the Trimmed Lasso could efficiently estimate the model coefficients despite the precense of confounders. In numerical experiments, we saw that the Trimmed Lasso is robust as it performs comparably to or even better than the standard Lasso when there is no confounding. We then applied de-biasing techniques to the Trimmed Lasso and showed that the Doubly Debiased Lasso could provide an asymptotically normally distributed estimator and thus provide asymptotically valid confidence intervals for individual coefficients. We remarked that the limiting normality of the Doubly Debiased Lasso is only shown marginally for each individual coefficient and questioned whether joint limiting normality could be obtained by refining the Doubly Debiased Lasso. We also proposed an alternative estimator for the variance of the noise term in the model. We saw in simulations that the proposed estimator of the noise term improved upon the coverage of confidence intervals constructed with the Doubly Debiased Lasso. We then considered a general model for linear multiple hypothesis testing under confouding. We saw that we could use estimation procedures from factor analysis to deconfound test statistics and get limiting normality of coefficient estimates, and control the familywise error rate. We then discussed possible extensions of the above methods to non-linear models and noted that little work has been done on this topic, calling for more research. Lastly, we briefly reviewed the Deconfounder algorithm and pointed to some similarities with the other deconfounder methods in this essay. We remarked that some of the claimed theoretical properties of the Deconfounder have been rejected. We conclude that there is much to be done on the topic of confounder adjustment for non-linear models.

7 Acknowledgements

I thank Dr. Rajen Shah and Dr. Qingyuan Zhao for allowing to write an essay on this interesting topic, and for their friendly guidance along the way. I also thank Emil Aas Stoltenberg, Nils Lid Hjort and Domagoj Ćevid for helpful discussions.

References

Bacallado, Segio and Rajen Shah (2020). "Modern Statistical Methods, Lecture notes".

Bai, Jushan and Kunpeng Li (2012). "Statistical analysis of factor models of high dimension". In: *The Annals of Statistics* 40.1, pp. 436–465. DOI: 10.1214/11-AOS966. URL: https://doi.org/10.1214/11-AOS966.

Bai, Jushan and Serena Ng (2002). "Determining the Number of Factors in Approximate Factor Models". In: *Econometrica* 70.1, pp. 191–221. DOI: https://doi.org/10.1111/1468-0262.00273. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00273. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00273.

Bühlmann, Peter and Sara van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 1st. Springer Publishing Company, Incorporated. ISBN: 3642201911.

Ćevid, Domagoj, Peter Bühlmann, and Nicolai Meinshausen (Nov. 2018). "Spectral Deconfounding via Perturbed Sparse Linear Models". In: arXiv: 1811.05352 [stat.ME].

Chamberlain, Gary and Michael Rothschild (1983). "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets". In: *Econometrica* 51.5, pp. 1281–1304. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912275.

D'Amour, Alexander (2019). *On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives*. arXiv: 1902.10286 [stat.ML].

Fan, Jianqing, Shaojun Guo, and Ning Hao (2010). *Variance Estimation Using Refitted Cross-validation in Ultrahigh Dimensional Regression*. arXiv: 1004.5178 [stat.ME].

Grimmer, Justin, Dean Knox, and Brandon M. Stewart (2020). *Naive regression requires weaker assumptions than factor models to adjust for multiple cause confounding*. arXiv: 2007.12702 [stat.ME].

Guo, Zijian, Domagoj Ćevid, and Peter Bühlmann (Apr. 2020). "Doubly Debiased Lasso: High-Dimensional Inference under Hidden Confounding". In: arXiv: 2004.03758 [stat.ME].

Harman, H.H. (1976). *Modern Factor Analysis*. University of Chicago Press. ISBN: 9780226316529. URL: https://books.google.no/books?id=e-vMN68C3M4C.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press. DOI: 10.1017/CBO9781139025751.

Leek, Jeffrey T. and John D. Storey (2008). "A general framework for multiple testing dependence". In: *Proceedings of the National Academy of Sciences* 105.48, pp. 18718–18723. ISSN: 0027-8424. DOI: 10.1073/pnas.0808709105. eprint: https://www.pnas.org/content/105/48/18718.full.pdf. URL: https://www.pnas.org/content/105/48/18718.

Ogburn, Elizabeth L., Ilya Shpitser, and Eric J. Tchetgen Tchetgen (2019). *Comment on "Blessings of Multiple Causes"*. arXiv: 1910.05438 [stat.ME].

— (2020). *Counterexamples to "The Blessings of Multiple Causes" by Wang and Blei*. arXiv: 2001.06555 [stat.ME].

Pearl, Judea (2009). *Causality*. 2nd ed. Cambridge University Press. DOI: 10.1017/CBO9780511803161.

Reid, Stephen, Robert Tibshirani, and Jerome Friedman (Nov. 2013). "A Study of Error Variance Estimation in Lasso Regression". In: arXiv: 1311.5274 [stat.ME].

Shah, Rajen et al. (Nov. 2018). "RSVP-graphs: Fast High-dimensional Covariance Matrix Estimation under Latent Confounding". In: arXiv: 1811.01076 [stat.ME].

van de Geer, Sara et al. (Mar. 2013). "On asymptotically optimal confidence regions and tests for high-dimensional models". In: *Annals of Statistics 2014, Vol. 42, No. 3, 1166-1202*. DOI: 10.1214/14-AOS1221. arXiv: 1303.0518 [math.ST].

Wang, Fa (2020). "Maximum likelihood estimation and inference for high dimensional generalized factor models with application to factor-augmented regressions". In: *Journal of Econometrics*. ISSN: 0304-4076. DOI: https://doi.org/10.1016/j.jeconom.2020.11.002. URL: https://www.sciencedirect.com/science/article/pii/S0304407620303894.

Wang, Jingshu et al. (2017). "Confounder adjustment in multiple hypothesis testing". English. In: *The Annals of Statistics* 45.5, pp. 1863–1894. ISSN: 0090-5364; 2168-8966/e.

Wang, Yixin and David M. Blei (2019). "The Blessings of Multiple Causes". In: *Journal of the American Statistical Association* 114.528, pp. 1574–1596. DOI: 10.1080/01621459.2019.1686987. eprint: https://doi.org/10.1080/01621459.2019.1686987. URL: https://doi.org/10.1080/01621459.2019.1686987.

— (2020). *Towards Clarifying the Theory of the Deconfounder*. arXiv: 2003.04948 [stat.ML].

Wooldridge, Jefrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press. ISBN: 9780262232586. URL: http://www.jstor.org/stable/j.ctt5hhcfr.

Zhang, Cun-Hui and Stephanie S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 217–242. DOI: https://doi.org/10.1111/rssb.12026. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12026. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12026.