Math. Stat. Learn. 7 (2024), 189–220
DOI 10.4171/MSL/47

A consistent estimator for confounding strength

Luca Rendsburg, Leena Chennuru Vankadara, Debarghya Ghoshdastidar,
and Ulrike von Luxburg

Abstract. Regression on observational data can fail to capture a causal relationship in the presence of unobserved confounding. Confounding strength measures this mismatch, but estimating it requires itself additional assumptions. A common assumption is the independence of causal mechanisms, which relies on concentration phenomena in high dimensions. While high dimensions enable the estimation of confounding strength, they also necessitate adapted estimators. In this paper, we derive the asymptotic behavior of the confounding strength estimator by Janzing and Schölkopf (2018) and show that it is generally not consistent. We then use tools from random matrix theory to derive an adapted, consistent estimator.

1. Introduction

A common machine learning task is to learn the influence of features x on a target variable y from a set of observations $\{(x_i, y_i)\}_{i=1}^n$. In many applications, we are not only interested in the statistical problem of predicting y after *observing* x; instead, we ask the causal question of how y changes after *intervening* on x. Unfortunately, the causal dependence structure between x and y is in general not identifiable from their statistical dependencies [31]. Simply regressing y on x attributes all dependencies to direct causal influence and is therefore only appropriate when x causes y without hidden confounders. However, this solution can be grossly misleading in the other possible cases where y causes x or both are caused by a common confounder [35].

For example, we might want to predict how increasing the height x from which an object is dropped affects its momentum y when it hits the ground. In this case, height clearly causes momentum and not vice versa. However, if this experiment is merely observed, we cannot rule out the possibility that both are influenced by some hidden confounders. Consider fruits falling from a tree: the type of fruit influences the height of the tree (type causes height) and the mass of the fruit, which in turn influences the momentum (type causes momentum). This makes the type a confounding factor.

Mathematics Subject Classification 2020: 62J99 (primary); 62D20 (secondary).
Keywords: confounding, observational data, high-dimensional linear regression, random matrix theory.

Disregarding confounding can lead to counter-intuitive findings, such as Simpson's paradox: for apples and oranges, where orange trees are slightly shorter but bear much heavier fruit, we would observe a negative correlation between height and momentum, even though the causal influence of height on momentum is positive.

A priori, it is generally unclear to what extent an observed statistical dependence is due to direct causal influence or due to confounding factors. This fundamental non-identifiability issue of causal structure from observational distribution can be addressed in different ways. One way is access to additional data, such as data from different environments [13, 33] or instrumental variables [4, 16], which reduces the causal learning problem to a statistical one. Alternatively, one can assume that the underlying causal model follows a certain data-generating process, such as additive noise models [14, 25, 43]. This reduces the number of causal models which can explain a given observational structure and therefore mitigates the non-identifiability. A more abstract approach to choosing a causal model among those compatible with an observational structure is to postulate certain information-theoretic properties of the causal model. For example, the causal directions are those that maximize conditional entropies or the causal factorization of the joint distribution is the one with minimal Kolmogorov complexity [3, 19, 29, 38].

In this paper, we theoretically analyze the confounding strength estimator by [21]. This estimator assumes that x causes y and aims to estimate the strength of unobserved confounding based on observational data $\{(x_i, y_i)\}_{i=1}^n$. Here, the confounding strength is defined as the discrepancy between the causal effect of x on y and the statistical regression vector. To mitigate the non-identifiability, the estimator considers a linear Gaussian causal model under the assumption of independent causal mechanisms, a common assumption in causal learning [19, 27, 34]. Abstractly, this principle states that the different causal mechanisms share no information. In our example of fruits falling from trees, these mechanisms are the physical mechanisms of gravity (height → momentum), inertia (fruit type → momentum) and the biological mechanism that determines the growth of a tree (fruit type → height). Arguably, understanding one of those mechanisms does not inform about the others. While the task of confounding strength estimation remains ill posed in finite dimensions, it becomes solvable in the high-dimensional limit due to concentration of measure phenomena. Crucially, this approach therefore requires large dimension d to reduce the non-identifiability error, but at the same time requires an even larger number of samples $n \gg d$ to reduce the finite-sample error. This is because it uses the empirical covariance matrix and regression vector in an intermediate step to estimate the corresponding population quantities, which is only consistent for $n \gg d$. It is therefore not guaranteed that this estimator is consistent in the high-dimensional regime. We address this issue by analyzing this estimator, from here on referred to as the plug-in estimator, in the proportional asymptotic regime $n, d \to \infty$ with $d/n \to \gamma \in [0, 1)$

and make the following contributions:

- We derive the asymptotic behavior of the plug-in estimator for confounding strength from [21] in the proportional asymptotic regime and show that it is not generally consistent. We also show that the approach based on population instead of finite-sample quantities is consistent.
- We derive a consistent estimator for confounding strength by correcting the above estimator with tools from random matrix theory.

The paper is structured as follows. Section 2 gives an overview of related work on causal inference under unobserved confounding. Section 3 introduces the confounded causal model, the measure of confounding strength, and basic notions from random matrix theory which are needed for the analysis. Section 4 describes the general approach of [21] and shows that it is consistent based on population quantities in Section 4.1, but generally biased based on plug-in quantities in Section 4.2. A corrected, consistent estimator for confounding strength is then derived in Section 5. Section 6 concludes with a discussion.

2. Related work

Learning causal relationships under the presence of unobserved confounding has been investigated by multiple works. Reference [15] detects the causal direction in linear non-Gaussian models based on the structure of the mixing matrix and [18] does so for non-linear additive noise models. Reference [22] detects low-complexity confounding based on a purity criterion for conditional distributions. Reference [23] decides whether a causal model is confounded based on the algorithmic Markov condition. Reference [5] considers the stability of the regression vectors under different environments as an indication of causal influence.

Our paper falls into another line of work that detects confounding based on the assumption of independent causal mechanisms. This assumption induces certain nongeneric alignments between the coefficients of the observational distribution, which can be used to identify confounding. Reference [2] uses this assumption to learn a sparse causal DAG under dense confounding. Reference [20] introduces the notion of confounding strength and estimate it under scalar confounding. Their method is based on the observation that a weighted spectral measure of the covariance matrix concentrates in high dimensions. Reference [28] builds on this idea by moving from the spectral measure to its first moment. Reference [21] extends this setting to multivariate confounding, which is the setting of our work. Reference [17] considers a subsequent task of learning a causal model with ridge regression. It uses an estimate of confounding strength to choose an appropriate regularization parameter, which is motivated by an analogy between finite sample error and confounding. Reference [6] generalizes the notion of confounding strength beyond independent causal mechanisms and characterizes the relationship between confounding strength and the causal risk of ridge regression in the high-dimensional limit.

Another related field is sensitivity analysis for treatment-effect studies based on observational data. Sensitivity analysis aims to quantify how sensitive causal conclusions are to potential unobserved confounding [7]. Since this task suffers from the same non-identifiability issue as described above, early work relies on assumptions about the unobserved confounder [11, 40]. A more recent, popular approach without assumptions gives bounds based on two (unknown) sensitivity parameters for how strong confounding would need to be in order to explain away any observed statistical associations between treatment and effect [9, 32, 37]. The region of sensitivity parameters that explain away associations can be condensed into a single E-value, which acts as a measure of confounding strength and can be computed from observational data [41, 42].

3. Preliminaries

This preliminary section introduces our confounded causal model and a notion of confounding strength in Section 3.1, as well as basic tools from random matrix theory needed for analysis in Section 3.2.

3.1. The confounded causal model

We first describe the problem setup and introduce basic quantities. We consider a confounded causal model with linear conditionals and Gaussian distributions. The model depends on a set of hyperparameters $\alpha \in \mathbb{R}^l$, $\beta \in \mathbb{R}^d$, $M \in \mathbb{R}^{d \times l}$ with dimensions $l \geq d$ and noise $\sigma^2 \geq 0$. $I_l \in \mathbb{R}^{l \times l}$ denotes the identity matrix. Specifically, we define the causal model in terms of its structural equations

$$\begin{aligned} \boldsymbol{z} &\sim \mathcal{N}(0, I_l), \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(0, \sigma^2), \\ \boldsymbol{x} &= M\boldsymbol{z}, \\ \boldsymbol{y} &= \boldsymbol{x}^T\beta + \boldsymbol{z}^T\alpha + \boldsymbol{\varepsilon}. \end{aligned} \tag{3.1}$$

Figure 1 shows the corresponding directed acyclic graph (DAG). All random variables $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ have mean 0 and the covariance of the features $\boldsymbol{x}$ is given by $\Sigma := MM^T \in \mathbb{R}^{d \times d}$. We additionally assume that M has full rank d such that Σ is invertible. We use the notation $\|x\|_\Sigma^2 := x^T \Sigma x$ for the generalized norm, M^+ for the pseudo-inverse of M, and $M^{+T} := (M^+)^T$ as shorthand. Random variables are boldfaced.

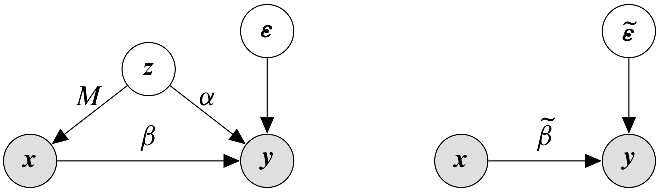

Figure 1. *Left:* DAG corresponding to the causal model (3.1). *Right:* corresponding observational model as in Lemma 1. Observed variables are shaded.

By construction, β describes the causal influence of $\boldsymbol{x}$ on $\boldsymbol{y}$. This is formally captured by the interventional distribution of the *do*-calculus [30] for any $x_0 \in \mathbb{R}^d$, under which $\boldsymbol{y} = x_0^T \beta + \boldsymbol{z}^T \alpha + \boldsymbol{\varepsilon}$ is only a random variable in $\boldsymbol{z}, \boldsymbol{\varepsilon}$ and has expectation $x_0^T \beta$. However, we do not assume access to interventional data; instead, we only observe values $(\boldsymbol{x}, \boldsymbol{y})$. The corresponding statistical dependencies between $\boldsymbol{x}$ and $\boldsymbol{y}$ are captured by the usual conditional distribution.

Lemma 1 (Observational distribution). *For the causal model* (3.1)*, the observational distribution of* $\boldsymbol{y}$ *given* $\boldsymbol{x}$ *is* $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}^T \tilde{\beta}, \tilde{\sigma}^2)$*, where* $\tilde{\beta} = \beta + M^{+T}\alpha$ *and* $\tilde{\sigma}^2 = \sigma^2 + \|\alpha\|^2_{I_l - M^+ M}$.

Proof. Since $\boldsymbol{z} \sim \mathcal{N}(0, I_l)$ is Gaussian and $\boldsymbol{x} = M\boldsymbol{z}$ is a linear map, it is a standard result that $\boldsymbol{z}^T|\boldsymbol{x}$ is Gaussian again with parameters $\boldsymbol{z}^T|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}^T M^{+T}, I - M^+ M)$. Subsequently, we have $\boldsymbol{z}^T \alpha|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}^T M^{+T}\alpha, \|\alpha\|^2_{I-M^+M})$. With $\boldsymbol{y} = \boldsymbol{x}^T \beta + \boldsymbol{z}^T \alpha + \boldsymbol{\varepsilon}$, we arrive at

$$\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}^T(\beta + M^{+T}\alpha), \sigma^2 + \|\alpha\|^2_{I-M^+M}) = \mathcal{N}(\boldsymbol{x}^T \tilde{\beta}, \tilde{\sigma}^2). \qquad \blacksquare$$

The statistical parameter $\tilde{\beta}$ can also be viewed as the result of regressing $\boldsymbol{y}$ on $\boldsymbol{x}$ on the population level. Notice that $\tilde{\beta}$ is equal to the causal parameter β up to an error term $M^{+T}\alpha$, which results from the influence of the confounder $\boldsymbol{z}$ on $\boldsymbol{y}$. This error term cannot be identified even if we have access to the full joint distribution $\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})}$, which demonstrates the fundamental non-identifiability issue of causal learning. To quantify the error of incorrectly treating $\tilde{\beta}$ as the causal parameter, [20] proposes the following measure of confounding strength.

Definition 2 (Measure of confounding strength [20]). The *confounding strength* ζ for the causal model (3.1) is defined as the relative error between statistical parameter $\tilde{\beta}$ and causal parameter β via

$$\zeta := \frac{\|\tilde{\beta} - \beta\|^2}{\|\beta\|^2 + \|\tilde{\beta} - \beta\|^2}. \tag{3.2}$$

The confounding strength ζ takes values in $[0,1]$, where $\zeta = 0$ describes the unconfounded case $\alpha = 0$ for which $\widetilde{\beta} = \beta$ and $\zeta = 1$ describes the purely confounded case $\beta = 0$. A larger confounding strength implies that the statistical parameter is further away from the causal parameter.

The goal of this paper is to estimate the confounding strength based on finite samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ from the observational distribution $\mathbb{P}_{(\boldsymbol{x},\boldsymbol{y})}$, which we compactly write as $\boldsymbol{X} \in \mathbb{R}^{d\times n}$ and $\boldsymbol{Y} \in \mathbb{R}^n$. We define two quantities which are central to the following estimators, namely, the sample covariance matrix $\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T$ and the result of regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$, $\widehat{\boldsymbol{\beta}} := (\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T)^+\frac{1}{n}\boldsymbol{X}\boldsymbol{Y}$.

3.2. Basic tools from random matrix theory

To make statements about the confounding strength, we need to control the behavior of the sample covariance matrix. However, the high-dimensional regime inhibits an entry-wise control, because we do not have many more samples than dimensions. Fortunately, the confounding strength only depends on a low-dimensional aspect of this matrix, which is the distribution of its eigenvalues. Controlling this distribution is one of the main objectives of random matrix theory. We therefore briefly recap some standard tools and results from random matrix theory to analyze the following estimators for confounding strength in the high-dimensional regime. The analysis is based on the Stieltjes transform of the empirical spectral distribution, an alternative description of the spectrum that is easier to handle.

Definition 3 (Empirical spectral distribution and Stieltjes transform). Let $\Sigma \in \mathbb{R}^{d\times d}$ be a symmetric matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$. The *empirical spectral distribution* of Σ is defined as the normalized counting measure of its eigenvalues $\mu_\Sigma := \frac{1}{d}\sum_{i=1}^d \delta_{\lambda_i}$. The corresponding *Stieltjes transform* of this measure is defined as the function $m_\Sigma(z) := \sum_{i=1}^d \frac{1}{\lambda_i - z}$ for $z \in \mathbb{C} \setminus \{\lambda_1, \ldots, \lambda_d\}$.

We need to characterize the spectra of the empirical covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{d\times d}$$

and the closely related empirical kernel matrix $\widehat{\boldsymbol{K}} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{n\times n}$. One might guess that they simply tend towards the spectrum of the population matrix, but their relation is more complicated in the asymptotic regime. The following standard result relates their limiting spectra to the spectrum of the population covariance in terms of Stieltjes transforms.

Theorem 4 (Asymptotics of the sample covariance matrix [36]). *Let $n, d \to \infty$ such that $d/n \to \gamma \in (0, \infty)$, and assume that the sequence of covariance matrices $\Sigma = \Sigma_d$*

has bounded operator norm $\limsup_{d\to\infty}\|\Sigma\| < \infty$. *Further, assume that the empirical spectral distribution of* Σ *converges, that is,* $\mu_\Sigma \to \nu$ *with bounded support and corresponding Stieltjes transform* m_ν. *Then, it holds that* $\boldsymbol{\mu}_{\widehat{\boldsymbol{\Sigma}}} \xrightarrow{a.s.} \mu$ *and* $\boldsymbol{\mu}_{\widehat{\boldsymbol{K}}} \xrightarrow{a.s.} \tilde{\mu}$ *as* $d \to \infty$, *where* $\mu, \tilde{\mu}$ *are the unique measures having Stieltjes transforms* $m(z)$ *and* $\tilde{m}(z)$, *respectively. For* $z \in \mathbb{C} \setminus \mathbb{R}_+$, *they satisfy*

$$m(z) = \frac{1}{\gamma}\tilde{m}(z) + \frac{1-\gamma}{\gamma z}, \tag{3.3}$$

$$m_\nu\left(-\frac{1}{\tilde{m}(z)}\right) = -zm(z)\tilde{m}(z). \tag{3.4}$$

Equation (3.3) also holds in finite dimensions for the Stieltjes transforms of $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{K}}$,

$$\boldsymbol{m}_{\widehat{\boldsymbol{\Sigma}}}(z) = \frac{1}{\frac{d}{n}}\boldsymbol{m}_{\widehat{\boldsymbol{K}}}(z) + \frac{1-\frac{d}{n}}{\frac{d}{n}z},$$

and simply reflects the fact that $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{K}}$ share the same eigenvalues up to the eigenvalue 0 with multiplicity $|n-d|$. Equation (3.4) is the main result that connects the limiting Stieltjes transforms of the empirical matrices $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{K}}$ to the limiting Stieltjes transform of the population covariance Σ. The solution m to this equation remains implicitly defined in all but the simplest case $\Sigma = I_d$, where m is the Stieltjes transform of a Marčenko–Pastur distribution.

4. Asymptotic behavior of the population and plug-in estimators for confounding strength

In this section, we describe the general approach for estimating confounding strength based on the assumption of independent causal mechanisms [21]. We show that the estimator is consistent based on population quantities in Section 4.1 but is generally biased for $n \not\gg d$ based on sample (plug-in) quantities in Section 4.2.

The main ingredient to tackle the non-identifiability of the causal model is the assumption of independent causal mechanisms, a common assumption in causal learning [19]. This abstract principle states that the physical mechanisms of a causal model that transfers causes to effect share no information. A possible translation for the causal model (3.1) is the assumption that the mechanisms $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are drawn from independent rotationally invariant distributions. Specifically, we assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are independent with $\boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2 I_l)$ and $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 I_d)$ for unknown hyperparameters $\sigma_\alpha^2, \sigma_\beta^2 \geq 0$. From here on out, we treat $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as random and reflect this by boldfacing them. Intuitively, this assumption facilitates estimation because it implies a certain alignment between the covariance matrix $\Sigma = MM^T$ and the regression vector $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha}$: for large confounding σ_α^2, the error term $M^{+T}\boldsymbol{\alpha}$ is aligned with small singular value directions of M. Correspondingly, $\widetilde{\boldsymbol{\beta}}$ is aligned with small eigendirections of Σ.

Assumption 5. We make the following assumptions about the (sequence of) causal models.

(A1) The parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ of model (3.1) are independently sampled with $\boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2 I_l)$ and $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 I_d)$ for hyperparameters $\sigma_\alpha^2, \sigma_\beta^2 \geq 0$.

(A2) The number of samples n, data dimension d, and latent confounder dimension l are in the proportional asymptotic regime, that is, $n, d, l \to \infty$ such that $d/n \to \gamma \in (0, 1)$ and $l/d \to \widetilde{\gamma} \geq 1$.

(A3) The covariance $\Sigma = \Sigma_d$ has bounded operator norm $\limsup_{d\to\infty} \|\Sigma\| < \infty$ and as $d \to \infty$ its empirical spectral distribution μ_Σ converges to a distribution ν with bounded support, that is, $\operatorname{supp}(\nu) \subseteq [h_1, h_2]$ with $0 < h_1 \leq h_2 < \infty$.

Assumption (A1) is the assumption of independent causal mechanisms. Assumption (A2) captures that this approach to confounding strength estimation requires high dimensions so that concentration effects can mitigate the non-identifiability issue. We exclude the case $\gamma \geq 1$ because the estimation of the term $\frac{1}{d}\operatorname{Tr}(\Sigma^{-1})$ (which later turns out to be relevant) is hard; see [8, Remark 2.11] for a discussion. The restriction on the latent dimensions $\widetilde{\gamma} \geq 1$ ensures that $l \geq d$ so that the population covariance $\Sigma = MM^T$ with $M \in \mathbb{R}^{d\times l}$ can be full rank, which is necessary for Assumption (A3).

Remark 6. Even if M and the full observational distribution are known, this only amounts to knowing the statistical parameter

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha},$$

which does not uniquely determine the multivariate causal mechanisms $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The assumption of independent causal mechanisms does not resolve this non-identifiability issue, but it does enable the estimation of certain *scalar* functions of these parameters: norms $\|\boldsymbol{\alpha}\|^2/d$ and more generally certain quadratic forms $\boldsymbol{\alpha}^T A\boldsymbol{\alpha}/d$ concentrate in high dimensions. As we will see, this enables the estimation of the confounding strength, which only depends on quadratic forms of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

The following key lemma states that random quadratic forms can concentrate around their trace.

Lemma 7 (Quadratic-form-close-to-the-trace [1, Lemma B.26]). *Let*

$$\boldsymbol{x} = (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^d) \in \mathbb{R}^d$$

have independent entries $\boldsymbol{x}^i$ of zero mean, unit variance, and $\boldsymbol{E}\,[|\boldsymbol{x}^i|^K] \le \nu_K$ for some $K \ge 1$. Then, for $A \in \mathbb{R}^{d\times d}$ and $k \ge 1$,

$$\boldsymbol{E}\left[|\boldsymbol{x}^T A\boldsymbol{x} - \operatorname{Tr} A|^k\right] \le C_k\left[\left(\nu_4 \operatorname{Tr}\left(AA^T\right)\right)^{k/2} + \nu_{2k} \operatorname{Tr}\left(AA^T\right)^{k/2}\right]$$

for some constant $C_k > 0$ independent of d. In particular, if the operator norm of A satisfies $\|A\| \le 1$ and the entries of x have bounded eighth-order moment, then

$$\boldsymbol{E}\left[\left(\boldsymbol{x}^T A\boldsymbol{x} - \operatorname{Tr} A\right)^4\right] \le Cd^2$$

for some $C > 0$ independent of d, and consequently,

$$\frac{1}{d}\boldsymbol{x}^T A\boldsymbol{x} - \frac{1}{d}\operatorname{Tr} A \xrightarrow[d\to\infty]{a.s.} 0.$$

Using this lemma, we directly obtain concentration of the confounding strength.

Corollary 8 (Confounding strength concentrates). *Under Assumption 5,*

$$\zeta - \frac{\tau^{\text{pop}}\cdot\theta^*}{1+\tau^{\text{pop}}\cdot\theta^*} \xrightarrow{a.s.} 0, \tag{4.1}$$

where $\tau^{\text{pop}} := \frac{1}{d}\operatorname{Tr}(\Sigma^{-1})$ and $\theta^* := \sigma_\alpha^2/\sigma_\beta^2$.

Proof. Using $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha}$, we can rewrite the confounding strength from equation (3.2) in terms of the hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, M$:

$$\zeta = \frac{\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2}{\|\boldsymbol{\beta}\|^2 + \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2} = \frac{\frac{1}{d}\boldsymbol{\alpha}^T M^+ M^{+T}\boldsymbol{\alpha}}{\frac{1}{d}\boldsymbol{\beta}^T I_d \boldsymbol{\beta} + \frac{1}{d}\boldsymbol{\alpha}^T M^+ M^{+T}\boldsymbol{\alpha}}.$$

This expression consists only of quadratic terms that can be controlled by Lemma 7, which yields

$$\begin{aligned}\zeta &= \frac{\frac{1}{d}\boldsymbol{\alpha}^T M^+ M^{+T}\boldsymbol{\alpha}}{\frac{1}{d}\boldsymbol{\beta}^T I_d \boldsymbol{\beta} + \frac{1}{d}\boldsymbol{\alpha}^T M^+ M^{+T}\boldsymbol{\alpha}}\\ &\overset{a.s.}{\approx} \frac{\frac{1}{d}\operatorname{Tr}(M^+ M^{+T})\sigma_\alpha^2}{\frac{1}{d}\operatorname{Tr}(I_d)\sigma_\beta^2 + \frac{1}{d}\operatorname{Tr}(M^+ M^{+T})\sigma_\alpha^2}\\ &= \frac{\tau^{\text{pop}}\cdot\theta^*}{1+\tau^{\text{pop}}\cdot\theta^*}.\end{aligned}$$

■

It only remains to estimate the trace term τ^{pop} and the ratio θ^*. In the following, we distinguish between three different kinds of estimators for various quantities: estimators based on the population quantities Σ, $\widetilde{\boldsymbol{\beta}}$, based on the plug-in quantities $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\beta}}$, and consistent estimators derived by random matrix theory. For example, we write τ^{pop}, $\boldsymbol{\tau}^{\text{plg}}$, or $\boldsymbol{\tau}^{\text{RMT}}$.

4.1. The population estimator for confounding strength is consistent

First, we consider estimation based on the population quantities Σ and $\widetilde{\boldsymbol{\beta}}$, which basically assumes that there are no finite-sample issues. In this case, $\tau^{\mathrm{pop}} = \frac{1}{d}\operatorname{Tr}(\Sigma^{-1})$ is known and does not need to be estimated. To estimate $\theta^* = \sigma_\alpha^2/\sigma_\beta^2$, observe that Assumption 5 (A1) on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ implies $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma_\beta^2 + \sigma_\alpha^2\Sigma^{-1})$. With respect to the uniform distribution on the sphere S^{d-1}, the distribution of the normalized vector $\widetilde{\boldsymbol{\beta}}/\|\widetilde{\boldsymbol{\beta}}\|$ has the log density

$$\log p_{\theta^*}(v) = -.5(\log\det(\Sigma + \theta^*) + d\log\langle v, \Sigma(\Sigma+\theta^*)^{-1}v\rangle - \log\det\Sigma),$$

where $v \in S^{d-1}$. Correspondingly, θ^* can then be estimated via maximum likelihood estimation as[1]

$$\begin{aligned} \boldsymbol{\theta}^{\mathrm{pop}} &= \operatorname*{arg\,min}_{\theta\geq 0} \boldsymbol{f}^{\mathrm{pop}}(\theta), \\ \text{where} \quad \boldsymbol{f}^{\mathrm{pop}}(\theta) &= \frac{1}{d}\log\det(\Sigma+\theta) + \log\left\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle. \end{aligned} \tag{4.2}$$

In summary, we consider the following population estimator for confounding strength.

Definition 9 (Population estimator for confounding strength). Given Σ and $\widetilde{\boldsymbol{\beta}}$, the *population estimator* for confounding strength $\boldsymbol{\zeta}^{\mathrm{pop}}$ is defined as

$$\boldsymbol{\zeta}^{\mathrm{pop}} = \frac{\tau^{\mathrm{pop}}\cdot\boldsymbol{\theta}^{\mathrm{pop}}}{1+\tau^{\mathrm{pop}}\cdot\boldsymbol{\theta}^{\mathrm{pop}}}, \tag{4.3}$$

where $\tau^{\mathrm{pop}} = \frac{1}{d}\operatorname{Tr}(\Sigma^{-1})$ and $\boldsymbol{\theta}^{\mathrm{pop}}$ is given by equation (4.2).

We now analyze this estimator by analyzing the asymptotic behavior of $\boldsymbol{\theta}^{\mathrm{pop}}$ from equation (4.2). Since $\boldsymbol{\theta}^{\mathrm{pop}}$ is implicitly defined as the minimizer of the function $\boldsymbol{f}^{\mathrm{pop}}$, we first derive the asymptotic behavior of $\boldsymbol{f}^{\mathrm{pop}}$ as an intermediate step. Specifically, we consider its derivative, which is given by

$$\partial_\theta \boldsymbol{f}^{\mathrm{pop}}(\theta) = m_\Sigma(-\theta) - \frac{\left\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle}{\left\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle}. \tag{4.4}$$

This idea is realized in the next theorem, which shows that the confounding strength estimator based on population quantities is consistent as $n, d \to \infty$, $d/n \to \gamma \in (0,1)$.

[1] Maximum likelihood estimation on the density of $\widetilde{\boldsymbol{\beta}}$ directly leads to the same optimality condition for $\boldsymbol{\theta}^{\mathrm{pop}}$.

Theorem 10 (Population estimator is consistent). *Under Assumption 5 with $\theta^* > 0$.*

(1) *For every $\theta \geq 0$, the derivative of the function from equations (4.2) satisfies*

$$\partial_\theta \boldsymbol{f}_d^{\text{pop}}(\theta) \xrightarrow{a.s.} (\theta - \theta^*) \operatorname{Var}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right] \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}\right]^{-1}. \tag{4.5}$$

(2) *For some $C > \theta^*$ and every $d \in \mathbb{N}$, let $\boldsymbol{\theta}_d^{\text{pop}}$ be a root of $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$ in $[0, C]$ if it exists or 0 otherwise. Additionally, assume that ν is not a point mass. Then, the sequence $\{\boldsymbol{\theta}_d^{\text{pop}}\}$ converges to θ^* almost surely.*

Proof. We just present a proof sketch here; the full proof is deferred to Appendix A. For the first statement about the population function $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$, we treat the three terms in equation (4.4) separately. The first term $m_\Sigma(-\theta)$ converges to $m_\nu(-\theta)$ by Assumption 5 (A3). The two quadratic forms are handled by Lemma 7 after rewriting

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha} = \begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d\end{pmatrix}\boldsymbol{u}$$

for some $\boldsymbol{u} \sim \mathcal{N}(0, I_{l+d})$. Plugging everything together and simplifying yield the result.

We prove the second statement by first upgrading the convergence of equation (4.5) to uniform convergence on $[0, C]$ using Vitali's convergence theorem [39] and then conclude that the roots converge to the unique root θ^* of the limiting function using Hurwitz's theorem [39]. ∎

This theorem shows that the approach of minimizing the log probability based on population quantities in equation (4.2) correctly estimates θ^* in the limit. Therefore, equation (4.3) leads to a consistent estimator for confounding strength. For the second statement, it is necessary to assume that the limiting spectral distribution ν of Σ is not a point mass because otherwise $\operatorname{Var}_{\boldsymbol{\lambda}\sim\nu}[1/(\boldsymbol{\lambda}+\theta)] = 0$. In this case, equation (4.5) states that the derivative $\partial_\theta \boldsymbol{f}^{\text{pop}}$ converges to the constant 0 function, which contains no information about θ^*. This is perfectly in line with the intuition presented for this approach: estimation of confounding strength is made possible by an alignment of $\widetilde{\boldsymbol{\beta}}$ with small eigendirections of Σ, but if Σ is a multiple of the identity (or, equivalently, the distribution of eigenvalues ν is a point mass), there is no particular small eigendirection.

4.2. The plug-in estimator for confounding strength is generally biased

The population estimator considered above crucially relies on the population quantities Σ and $\widetilde{\boldsymbol{\beta}}$, which are not directly available. In practice, we only have access to the corresponding empirical quantities $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\beta}}$ based on samples X, Y. This section considers the resulting plug-in estimator for confounding strength as introduced by [21] and shows in a similar asymptotic analysis that this estimator is generally biased. Formally, the plug-in estimator follows the same structure as Definition 9 but replaces the population quantities Σ, $\widetilde{\boldsymbol{\beta}}$ with the empirical quantities $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\beta}}$.

Definition 11 (Plug-in estimator for confounding strength [21]). The *plug-in estimator* for confounding strength $\boldsymbol{\zeta}^{\mathbf{plg}}$ is defined as

$$\boldsymbol{\zeta}^{\mathbf{plg}} = \frac{\boldsymbol{\tau}^{\mathrm{plg}} \cdot \boldsymbol{\theta}^{\mathrm{plg}}}{1 + \boldsymbol{\tau}^{\mathrm{plg}} \cdot \boldsymbol{\theta}^{\mathrm{plg}}}, \tag{4.6}$$

where $\boldsymbol{\tau}^{\mathrm{plg}} = \frac{1}{d}\operatorname{Tr}(\widehat{\boldsymbol{\Sigma}}^{-1})$ and $\boldsymbol{\theta}^{\mathrm{plg}}$ is given by

$$\begin{aligned} \boldsymbol{\theta}^{\mathrm{plg}} &= \arg\min_{\theta \geq 0} \boldsymbol{f}^{\mathrm{plg}}(\theta), \\ \text{where} \quad \boldsymbol{f}^{\mathrm{plg}}(\theta) &= \frac{1}{d}\log\det(\widehat{\boldsymbol{\Sigma}} + \theta) + \log\left\langle \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|}, (\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}} + \theta)^{-1})\frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|} \right\rangle. \end{aligned} \tag{4.7}$$

The main issue with the plug-in estimator in the proportional asymptotic regime is that $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\beta}}$ are not consistent estimators for Σ and $\widetilde{\boldsymbol{\beta}}$. Any subsequent estimators are therefore also not guaranteed to be consistent. The first example of such behavior is given by the plug-in estimator $\boldsymbol{\tau}^{\mathrm{plg}} = \frac{1}{d}\operatorname{Tr}(\widehat{\boldsymbol{\Sigma}}^{-1})$ for $\tau^{\mathrm{pop}} = \frac{1}{d}\operatorname{Tr}(\Sigma^{-1})$, one of the two quantities which need to be estimated in equation (4.1).

Proposition 12 (Asymptotic trace of inverse covariance). *Under Assumption 5, it holds*

$$\boldsymbol{\tau}^{\mathrm{plg}} - (1-\gamma)^{-1}\tau^{\mathrm{pop}} \xrightarrow[d\to\infty]{a.s.} 0.$$

Proof. In terms of Stieltjes transforms, the statement reads

$$(1-\gamma)\boldsymbol{m}_{\widehat{\boldsymbol{\Sigma}}}(0) - m_{\Sigma}(0) \xrightarrow[d\to\infty]{a.s.} 0.$$

The limiting empirical and population Stieltjes transforms are given by $\boldsymbol{m}_{\widehat{\boldsymbol{\Sigma}}}(z) \xrightarrow{a.s.} m(z)$ and $m_{\Sigma}(z) \xrightarrow{a.s.} m_{\nu}(z)$ as $d \to \infty$, so it remains to relate $m(0)$ to $m_{\nu}(0)$. By combining equations (3.3) and (3.4) from Theorem 4, we get

$$m_{\nu}\left(-\frac{1}{\tilde{m}(z)}\right) = (1-\gamma-zm(z))m(z).$$

Taking $z \to 0$, it is $1/\tilde{m}(z) \to 0$, and therefore, we get by continuity that $m_{\nu}(0) = (1-\gamma)m(0)$. ■

This result shows that the plug-in estimator for the trace of the inverse covariance matrix is off by a factor of $(1-\gamma)$. This factor is negligible in the case $n \gg d$, where $\gamma = d/n \approx 0$, but becomes increasingly relevant as γ grows.

Next, we treat the plug-in estimator $\boldsymbol{\theta}^{\text{plg}}$ similarly as $\boldsymbol{\theta}^{\text{pop}}$ in Theorem 10 and show that it is generally biased. Here, $\partial_\theta \boldsymbol{f}^{\text{plg}}$ is given analogously to equation (4.4).

Theorem 13 (Plug-in estimator is generally biased). *Under Assumption 5 with $\theta^* > 0$, the following hold.*

(1) *For all $\theta \geq 0$, the derivative of the function from equation (4.7) satisfies*

$$\partial_\theta \boldsymbol{f}_d^{\text{plg}}(\theta) \xrightarrow{a.s.} \\ \left[\theta - (1+\gamma\widetilde{\gamma})\theta^* + \gamma\theta^*(1-\theta m(-\theta))\left(1 + \frac{M(-\theta)}{M(-\theta) - m(-\theta)^2}\right)\right] h(\theta), \tag{4.8}$$

with

$$h(\theta) = (M(-\theta) - m(-\theta)^2) \\ \cdot (1 - \theta m(-\theta) + (1 - 2\gamma + \gamma\widetilde{\gamma})\theta^* m(-\theta) + \gamma\theta\theta^* m(-\theta)^2)^{-1},$$

$m(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}[1/(\boldsymbol{\lambda} + \theta)]$*, and* $M(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}[1/(\boldsymbol{\lambda} + \theta)^2]$.

(2) *For every $d \in \mathbb{N}$, let $\boldsymbol{\theta}_d^{\text{plg}}$ be a root of $\partial_\theta \boldsymbol{f}_d^{\text{plg}}$ if it exists or 0 otherwise. Additionally, assume that $\widetilde{\gamma}$ does not satisfy*

$$\widetilde{\gamma} = (1 - \theta^* m(-\theta^*))\left(1 + \frac{M(-\theta^*)}{M(-\theta^*) - m(-\theta^*)^2}\right). \tag{4.9}$$

Then, the sequence $\{\boldsymbol{\theta}_d^{\text{plg}}\}$ almost surely does not converge to θ^*.

Proof. We again only sketch the proof here; the full proof is deferred to Appendix B. The proof for the first statement follows the same strategy as in Theorem 10 but now deals with the sample quantities $\widehat{\boldsymbol{\Sigma}}$, $\widehat{\boldsymbol{\beta}}$ in place of the population quantities Σ, $\widetilde{\boldsymbol{\beta}}$. Similarly as for $\widetilde{\boldsymbol{\beta}}$, we treat $\widehat{\boldsymbol{\beta}}$ by combining the equations

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{Y}, \quad \boldsymbol{Y} = \boldsymbol{X}^T\widetilde{\boldsymbol{\beta}} + \boldsymbol{E}$$

for $\boldsymbol{E} \sim \mathcal{N}(0, \widetilde{\boldsymbol{\sigma}}^2 I_n)$, and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha}$ to obtain

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d & \widetilde{\boldsymbol{\sigma}}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\end{pmatrix}\boldsymbol{v}$$

for some $\boldsymbol{v} \sim \mathcal{N}(0, I_{l+d+n})$. Additional complications arise because $\widehat{\boldsymbol{\beta}}$ depends on both the population term M and the empirical quantities. This produces mixed terms $\mathrm{Tr}[(\widehat{\boldsymbol{\Sigma}} + \theta)^{-1}\widehat{\boldsymbol{\Sigma}}\,\Sigma^+]$ for $k \in \{1, 2\}$, which need to be treated with a separate result by [26] in Lemma 21.

For the second statement, we use similar arguments as in the proof of Theorem 10 to show that the convergence $\boldsymbol{\theta}_d^{\text{plg}} \to \theta^*$ implies that θ^* is a root of the right-hand side in equation (4.8). This is equivalent to equation (4.9), which does not hold by assumption. ∎

The limiting derivative for the plug-in estimator in equation (4.8) is phrased in terms of the limiting sample distribution μ instead of the limiting population distribution ν. The main structural difference to equation (4.5) is the existence of an additional term $\gamma\theta^*(1-\theta m(-\theta))(1+M(-\theta)/(M(-\theta)-m(-\theta)^2))$, which prevents a closed-form expression for the corresponding roots $\boldsymbol{\theta}^{\mathrm{plg}}$ of this function. We therefore cannot directly exclude the possibility that θ^* is a root, in which case the plug-in estimator would be consistent. However, by simply plugging in θ^* in the limiting derivative, we see that θ^* being a root is equivalent to the condition in equation (4.9). This condition generally does not hold because the limiting ratio of dimensions $\tilde{\gamma} = \lim_{d,l\to\infty} l/d$ on the left-hand side stands in no special relationship to the terms on the right-hand side. Therefore, the plug-in estimator $\boldsymbol{\theta}^{\mathrm{plg}}$ is generally a biased estimator for θ^*. This means that the resulting plug-in estimator for confounding strength $\boldsymbol{\zeta}^{\mathbf{plg}}$ is generally a biased estimator for the true confounding strength $\boldsymbol{\zeta}$.

5. A consistent estimator for confounding strength

In this section, we derive a novel estimator for confounding strength using tools from random matrix theory. We show that this estimator consistently recovers the true confounding strength in the high-dimensional asymptotic limit ($n, d \to \infty, d/n \to \gamma \in (0,1)$). To this end, we can derive a consistent estimator of $\boldsymbol{\theta}^{\mathrm{RMT}}$ by first consistently estimating $\boldsymbol{f}^{\mathrm{pop}}(\theta)$ and then finding the minimizer of this function. While this procedure indeed yields a consistent estimator, it is stochastic, which can adversely affect the optimization algorithm at finite d. Therefore, we also provide a consistent estimator based on finding the zeros of $\partial_\theta \boldsymbol{f}^{\mathrm{pop}}(\theta)$ which is deterministic given a fixed sample. Coupled with the consistent estimator for τ^{pop} in Proposition 12, we arrive at a consistent estimator for confounding strength.

5.1. A consistent estimator for $\boldsymbol{f}^{\mathbf{pop}}(\boldsymbol{\theta})$

Recall from equation (4.2) that maximum likelihood estimation of θ^* is equivalent to the optimization problem

$$\boldsymbol{\theta}^{\mathrm{pop}} = \operatorname*{arg\,min}_{\theta\geq 0} \boldsymbol{f}^{\mathrm{pop}}(\theta),$$

$$\text{where} \quad \boldsymbol{f}^{\mathrm{pop}}(\theta) = \frac{1}{d}\log\det(\Sigma+\theta) + \log\left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle.$$

To consistently estimate $\boldsymbol{f}^{\mathrm{pop}}(\theta)$, it suffices to consistently estimate the two quantities $\frac{1}{d}\log\det(\Sigma+\theta)$ and $\log\langle\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\rangle$. We derive such estimators in

Theorems 14 and 15 using tools from random matrix theory. The main results are included here, and we defer the proofs to Appendix C.

Theorem 14 (A consistent estimator for log determinant [24]). *For any $\theta \in \mathbb{R}^+$, let $\boldsymbol{W} = \boldsymbol{X} + \sqrt{\theta}\boldsymbol{E}$, where $\boldsymbol{E} \in \mathbb{R}^{d\times n}$ is a random matrix with standard normal entries. Then, as $d, n \to \infty$ such that $d/n \to \gamma \in (0,1)$,*

$$\log\theta + \frac{1}{d}\log\det\frac{1}{n\theta}\boldsymbol{W}\boldsymbol{W}^T + (1-\gamma)\log\frac{\gamma-1}{\gamma} + 1 - \frac{1}{d}\log\det(\Sigma+\theta) \xrightarrow{a.s.} 0.$$

In other words, the function

$$\boldsymbol{g}_1(\theta) = \log\theta + \frac{1}{d}\log\det\frac{1}{n\theta}\boldsymbol{W}\boldsymbol{W}^T + (1-\gamma)\log((\gamma-1)/\gamma) + 1$$

is a consistent estimator of $\log\det(\Sigma+\theta)$.

Proposition 15 (A consistent estimator for the quadform). *Under Assumption 5, for any $\theta \in \mathbb{R}^+$, let η be the unique solution in $\mathbb{R}^-$ satisfying $\tilde{m}(\eta) = 1/\theta$. Then, as $d, n \to \infty$ such that $d/n \to \gamma \in (0,1)$,*

$$\frac{\frac{1}{d}\langle\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widehat{\boldsymbol{\beta}}\rangle - \frac{\boldsymbol{S}}{\theta} - \frac{\boldsymbol{S}(1-\gamma)}{\eta}}{\frac{1}{d}\|\widehat{\boldsymbol{\beta}}\|^2 - \boldsymbol{S}\gamma m(0)} - \left\langle\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle \xrightarrow{a.s.} 0,$$

where $\boldsymbol{S} = (1-\gamma)^{-1}\|\boldsymbol{Y}\|^2_{I-\boldsymbol{X}^+\boldsymbol{X}}/(nd)$.

In other words, the function

$$\boldsymbol{g}_2(\theta) = \log\frac{\frac{1}{d}\langle\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widehat{\boldsymbol{\beta}}\rangle - \frac{\boldsymbol{S}}{\theta} - \frac{\boldsymbol{S}(1-\gamma)}{\eta}}{\frac{1}{d}\|\widehat{\boldsymbol{\beta}}\|^2 - \boldsymbol{S}\gamma m(0)}$$

is a consistent estimator of $\log\langle\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\rangle$. Thereby, for every $\theta \in \mathbb{R}^+$, as $n, d \to \infty$ as $d/n \to \gamma \in (0,1)$,

$$\boldsymbol{g}_1(\theta) + \boldsymbol{g}_2(\theta) - \boldsymbol{f}^{\text{pop}}(\theta) \xrightarrow{a.s.} 0 \tag{5.1}$$

Therefore, a consistent estimator of $\boldsymbol{f}^{\text{pop}}(\theta)$ is given by $\boldsymbol{f}^{\text{RMT}}(\theta) := \boldsymbol{g}_1(\theta) + \boldsymbol{g}_2(\theta)$.

Stochasticity of the estimation. Observe that the estimator for the log determinant given by $\boldsymbol{g}_1(\theta)$ is not a deterministic function of a given sample $\boldsymbol{X}, \boldsymbol{Y}$ since the matrix $\boldsymbol{W}$ is stochastic. Following arguments similar to the proof of Theorems 10 and 13 [2],

[2] With an additional argument to deal with the stochasticity of the log det estimator.

we can indeed obtain an asymptotically consistent estimator for confounding strength. However, at finite d, our experiments suggest that the stochasticity can adversely affect the optimization step. Furthermore, the dependence of $\boldsymbol{g}_1(\theta)$ on θ is highly non-linear. Iterative optimization procedures require multiple evaluations (and therefore estimation of) $\boldsymbol{g}_1(\theta)$ which considerably increases the computation complexity. To overcome these limitations, we also provide a deterministic and consistent estimator of θ by first consistently estimating the function $\partial_\theta \boldsymbol{f}^{\text{pop}}(\theta)$ for any $\theta \in \mathbb{R}^+$ and showing that the roots of the estimating function asymptotically converge to θ^*.

5.2. A consistent estimator for $\partial_\theta f^{\text{pop}}(\theta)$

As derived in equation (4.4), the derivative of $\boldsymbol{f}^{\text{pop}}(\theta)$ is given by

$$\partial_\theta \boldsymbol{f}^{\text{pop}}(\theta) = \frac{\left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle \cdot m_\Sigma(-\theta) - \left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle}{\left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle}.$$

In order to consistently estimate $\partial_\theta \boldsymbol{f}^{\text{pop}}(\theta)$, it suffices to consistently estimate the three quantities $m_\Sigma(-\theta)$, $\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\rangle$, and $\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\rangle$. Proposition 15 provides us with a consistent estimator for the quantity

$$\left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle.$$

In Propositions 16 and 17, we derive estimators for the remaining quantities.

Proposition 16 (Estimation of Stieltjes transform). *Under the assumptions of Theorem 4, for any $\theta \in \mathbb{R}^+$, let η be the unique solution in $\mathbb{R}^-$ satisfying $\tilde{m}(\eta) = 1/\theta$. Then, as $d, n \to \infty$ such that $d/n \to \gamma \in (0,1)$,*

$$-\frac{1}{\gamma\theta}\left(\frac{\eta}{\theta} - \gamma + 1\right) - m_\Sigma(-\theta) \to 0.$$

Proof. From Theorem 4, we have $m_\nu(-\frac{1}{\tilde{m}(z)}) = (1-\gamma - zm(z))m(z)$ for any $z \in \mathbb{C}/\mathbb{R}^+$. Letting $\eta \in \mathbb{R}^-$ such that $\tilde{m}(\eta) = 1/\theta$, we arrive at the estimator. ■

Now, we present a consistent estimator of the quadratic form

$$\left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle.$$

From Proposition 15, we know that, for any $\theta \in \mathbb{R}^+$, $\boldsymbol{g}_2(\theta)$ is a consistent estimator of $\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\rangle$. To derive an estimator of the quadratic form

$$\left\langle \frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\tilde{\boldsymbol{\beta}}}{\|\tilde{\boldsymbol{\beta}}\|}\right\rangle,$$

we utilize the so-called derivative trick [10, 12]. First, observe that

$$\langle \widetilde{\boldsymbol{\beta}}, \Sigma(\Sigma+\theta)^{-2}\widetilde{\boldsymbol{\beta}}\rangle = -\partial_\theta\big(\langle \widetilde{\boldsymbol{\beta}}, \Sigma(\Sigma+\theta)^{-1}\widetilde{\boldsymbol{\beta}}\rangle\big).$$

Furthermore, for every fixed $\theta \in \mathbb{R}^+$, we know that as $n, d \to \infty$ and $d/n \to \gamma \in (0,1)$,

$$\boldsymbol{g}_2(\theta) - \left\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle \xrightarrow{a.s.} 0.$$

It is also easy to verify that $\boldsymbol{g}_2(\theta) - \langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\rangle$ is analytic and uniformly bounded in θ in the domain $\mathbb{R}^+$. Therefore, we can apply Vitali's convergence theorem to show that the limit of the derivatives converges to the derivative of the limit. Therefore, a consistent estimator for the quadratic form $\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\rangle$ is given by $-\partial_\theta \boldsymbol{g}_2(\theta)$ and is formally presented in Theorem 17.

Proposition 17 (Consistent estimator for quadratic form). *For any $\theta \in \mathbb{R}^+$, let η be the unique solution in $\mathbb{R}^-$ satisfying $\tilde{m}(\eta) = 1/\theta$, and let $\eta' = 1/(\theta^2\tilde{m}'(\eta))$. As $d, n \to \infty$ such that $d/n \to \gamma \in (0,1)$,*

$$\frac{\frac{\eta'}{d}\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-2}\widehat{\boldsymbol{\beta}}\rangle - \frac{\boldsymbol{S}}{\theta^2} + \frac{\boldsymbol{S}\eta'(1-\gamma)}{\eta^2}}{\frac{1}{d}\|\widehat{\boldsymbol{\beta}}\|^2 - \boldsymbol{S}\gamma m(0)} - \left\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma+\theta)^{-2}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle \xrightarrow{a.s.} 0,$$

where

$$\boldsymbol{S} = \frac{1}{(1-\gamma)nd}\|\boldsymbol{Y}\|^2_{I-\boldsymbol{X}^+\boldsymbol{X}}/(nd).$$

From Propositions 15, 16, and 17, for any $\theta \in \mathbb{R}^+$, a consistent estimator of $\partial_\theta \boldsymbol{f}^{\text{pop}}(\theta)$ is given by

$$\boldsymbol{h}_{\text{RMT}}(\theta) := \frac{\frac{\boldsymbol{g}_2(\theta)}{\gamma\theta}(\gamma - 1 - \frac{\eta}{\theta}) - \partial_\theta \boldsymbol{g}_2(\theta)}{\boldsymbol{g}_2(\theta)}.$$

The RMT estimator for confounding strength is then naturally defined via the roots of $\boldsymbol{h}_{\text{RMT}}(\theta)$ and RMT-corrected estimate of τ^{pop} as formally presented in Definition 18 which consistently estimates the true confounding strength ζ.

Definition 18 (RMT estimator for confounding strength). The *RMT estimator* for confounding strength $\boldsymbol{\zeta}^{\text{RMT}}$ can then be defined as

$$\boldsymbol{\zeta}^{\text{RMT}} = \frac{\boldsymbol{\tau}^{\text{RMT}} \cdot \boldsymbol{\theta}^{\text{RMT}}}{1 + \boldsymbol{\tau}^{\text{RMT}} \cdot \boldsymbol{\theta}^{\text{RMT}}}, \tag{5.2}$$

where $\boldsymbol{\tau}^{\text{RMT}} = (1-\gamma)\boldsymbol{\tau}^{\text{plg}}$ and $\boldsymbol{\theta}^{\text{RMT}}$ is a root of $\boldsymbol{h}_{\text{RMT}}(\theta)$ if it exists and 0 otherwise.

Theorem 19 (RMT estimator is consistent). *Let $\boldsymbol{\theta}_d^{\mathrm{RMT}}$ be defined as a root of $\boldsymbol{h}_{\mathrm{RMT}}(\theta)$ in some $[0, C]$ for some $C < \infty$ if it exists or 0 otherwise. Additionally, assume that ν is not a point mass. Then, under Assumption 5 with $\theta^* > 0$, the sequence $\{\boldsymbol{\theta}_d^{\mathrm{RMT}}\}$ converges a.s. to θ^*.*

6. Discussion and future work

We analyze the asymptotic behavior of the confounding strength estimator by [21] in the high-dimensional proportional regime. While the approach is consistent under population quantities, the corresponding plug-in estimator is generally biased. We correct for this bias and present a consistent estimator using tools from random matrix theory. High dimensions can help to identify the causal model, but they also warrant adapted estimators if the number of samples does not grow even faster than the dimensions. More generally, our work highlights the inherent complexity of the causal estimation problem since it requires careful consideration of statistical estimation in conjunction with the problem of causal non-identifiability, particularly in high-dimensional settings.

In this work, we focus on developing estimators that consistently estimate the true confounding strength in the proportional asymptotic regime. An important direction for future work involves establishing non-asymptotic guarantees for the convergence of the RMT estimator, $\boldsymbol{\zeta}^{\mathrm{RMT}}$. Furthermore, determining convergence rates would further enhance the applicability of the RMT estimator. It would also be of considerable interest to define and estimate the strength of confounding for broader classes of causal models, such as those in the Reproducing Kernel Hilbert Space (RKHS). From a practical perspective, estimating the strength of confounding under less stringent assumptions—for instance, when additional proxy variables are available—may also prove beneficial. We leave these for future work.

Faithful estimation of confounding strength can indeed facilitate causal learning from observational data, for instance, via regularization. This has been empirically demonstrated in [17] and under the same model setting as ours, precisely characterized in [6]. However, it is important to practice caution in applying such techniques more generally since causal learning or even estimation of confounding strength is a very hard problem and does require strong assumptions.

A. Proof of Theorem 10

This section gives the full proof of Theorem 10 for the asymptotic behavior of the population estimator for confounding strength. We state the theorem here again for reference.

Theorem 10 (Population estimator is consistent). *Under Assumption 5 with $\theta^* > 0$.*

(1) *For every $\theta \geq 0$, the derivative of the function from equations (4.2) satisfies*

$$\partial_\theta \boldsymbol{f}_d^{\text{pop}}(\theta) \xrightarrow{a.s.} (\theta - \theta^*) \operatorname{Var}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right] \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}\right]^{-1}. \tag{4.5}$$

(2) *For some $C > \theta^*$ and every $d \in \mathbb{N}$, let $\boldsymbol{\theta}_d^{\text{pop}}$ be a root of $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$ in $[0, C]$ if it exists or 0 otherwise. Additionally, assume that ν is not a point mass. Then, the sequence $\{\boldsymbol{\theta}_d^{\text{pop}}\}$ converges to θ^* almost surely.*

Proof. We first show equation (4.5). According to equation (4.4), the function is given by $\partial_\theta \boldsymbol{f}^{\text{pop}}(\theta) = m_\Sigma(-\theta) - \frac{1}{d}\widetilde{\boldsymbol{\beta}}^T\Sigma(\Sigma+\theta)^{-2}\widetilde{\boldsymbol{\beta}} / \frac{1}{d}\widetilde{\boldsymbol{\beta}}^T\Sigma(\Sigma+\theta)^{-1}\widetilde{\boldsymbol{\beta}}$. The first term $m_\Sigma(-\theta)$ converges to $m_\nu(-\theta)$ by assumption. The two quadratic forms are handled by Lemma 7 after rewriting $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha} = \begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d\end{pmatrix}\boldsymbol{u}$ for some $\boldsymbol{u} \sim \mathcal{N}(0, I_{l+d})$, which is possible because by assumption $\boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2 I_l)$ and $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_\beta^2 I_d)$ are independent. We have

$$\begin{aligned}
\frac{1}{d}\widetilde{\boldsymbol{\beta}}^T\Sigma(\Sigma+\theta)^{-1}\widetilde{\boldsymbol{\beta}} &= \frac{1}{d}\boldsymbol{u}^T\begin{pmatrix}\sigma_\alpha M^+\\ \sigma_\beta I_d\end{pmatrix}\Sigma(\Sigma+\theta)^{-1}\begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d\end{pmatrix}\boldsymbol{u}\\
&\overset{a.s.}{\approx} \frac{1}{d}\operatorname{Tr}\left[\begin{pmatrix}\sigma_\alpha M^+\\ \sigma_\beta I_d\end{pmatrix}\Sigma(\Sigma+\theta)^{-1}\begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d\end{pmatrix}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Lemma 7)}\\
&= \frac{1}{d}\operatorname{Tr}\left[\Sigma(\Sigma+\theta)^{-1}\begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d\end{pmatrix}\begin{pmatrix}\sigma_\alpha M^+\\ \sigma_\beta I_d\end{pmatrix}\right]\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\operatorname{Tr}(A\cdot B) = \operatorname{Tr}(B\cdot A))\\
&= \frac{1}{d}\operatorname{Tr}\left[\Sigma(\Sigma+\theta)^{-1}(\sigma_\alpha^2\Sigma^{-1} + \sigma_\beta^2 I_d)\right] \qquad (\Sigma = MM^T)\\
&= \frac{\sigma_\beta^2}{d}\operatorname{Tr}\left[(\Sigma+\theta)^{-1}(\Sigma+\theta^*)\right] \qquad (\theta^* = \sigma_\alpha^2/\sigma_\beta^2)\\
&\to \sigma_\beta^2\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}\right]. \qquad (\mu_\Sigma \to \nu)
\end{aligned}$$

Similarly, we get $\frac{1}{d}\widetilde{\boldsymbol{\beta}}^T\Sigma(\Sigma+\theta)^{-2}\widetilde{\boldsymbol{\beta}} \xrightarrow[d\to\infty]{a.s.} \sigma_\beta^2\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}[\frac{\boldsymbol{\lambda}+\theta^*}{(\boldsymbol{\lambda}+\theta)^2}]$. Plugging everything together yields

$$\begin{aligned}
\partial_\theta \boldsymbol{f}^{\text{pop}}(\theta) &\xrightarrow[d\to\infty]{a.s.} m_\nu(-\theta) - \frac{\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}[\frac{\boldsymbol{\lambda}+\theta^*}{(\boldsymbol{\lambda}+\theta)^2}]}{\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}]}\\
&= \left(m_\nu(-\theta)\cdot\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}\right] - \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{(\boldsymbol{\lambda}+\theta)^2}\right]\right)\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}\right]^{-1}.
\end{aligned}$$

Using $m_\nu(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}[\frac{1}{\boldsymbol{\lambda}+\theta}]$ and the identity $\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta} = 1 - (\theta-\theta^*)\frac{1}{1+\boldsymbol{\lambda}}$, we can simplify the first factor as follows.

$$\begin{aligned}
&m_\nu(-\theta)\cdot \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{\boldsymbol{\lambda}+\theta}\right] - \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{\boldsymbol{\lambda}+\theta^*}{(\boldsymbol{\lambda}+\theta)^2}\right]\\
&\quad= \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right]\left(1-(\theta-\theta^*)\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right]\right) - \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right]\\
&\qquad+(\theta-\theta^*)\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{(\boldsymbol{\lambda}+\theta)^2}\right]\\
&\quad= (\theta-\theta^*)\left(\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{(\boldsymbol{\lambda}+\theta)^2}\right] - \boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right]^2\right)\\
&\quad= (\theta-\theta^*)\operatorname{Var}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right].
\end{aligned}$$

This concludes the first part of the proof.

For the second statement, first, observe that the almost sure convergence in equation (4.5) for each $\theta \geq 0$ implies that this convergence also holds almost surely on a countable set, such as $[0, C] \cap \mathbb{Q}$. Since each function $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$ is analytic and bounded on $[0, C]$, we can further upgrade equation (4.5) to almost surely uniform convergence on $[0, C]$ by Vitali's convergence theorem. Now, let $(\boldsymbol{\theta}_d^{\text{pop}})_{d\in\mathbb{N}}$ be a sequence of roots as described in the theorem, and let $F^{\text{pop}}(\theta)$ denote the function on the right-hand side of equation (4.5). First, note that the functions $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$ eventually have a root $\boldsymbol{\theta}_d^{\text{pop}}$ in $[0, C]$ with probability 1: since $\theta^* < C$, there exist θ_-, θ_+ with $0 < \theta_- < \theta^* < \theta_+ < C$ with $F^{\text{pop}}(\theta_-) < 0$ and $F^{\text{pop}}(\theta_+) > 0$. The convergence of the functions $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$ then implies that $\partial_\theta \boldsymbol{f}_d^{\text{pop}}(\theta_-) < 0$ and $\partial_\theta \boldsymbol{f}_d^{\text{pop}}(\theta_+) > 0$ eventually. Since $\partial_\theta \boldsymbol{f}_d^{\text{pop}}$ is continuous, the intermediate value theorem then implies the existence of a root in $(\theta_-, \theta_+) \subset [0, C]$. The proof is concluded with Hurwitz's theorem, which states that the sequence of roots $(\boldsymbol{\theta}_d^{\text{pop}})_{d\in\mathbb{N}}$ of analytic functions converges to the unique root θ^* of the limiting function. ∎

B. Proof of Theorem 13

For the proof of Theorem 13 about the asymptotic behavior of the plug-in estimator, we require additional technical statements. The first characterizes the asymptotic behavior of the statistical noise for our causal model.

Lemma 20 (Asymptotics of the statistical noise). *Under Assumption 5, the statistical noise $\widetilde{\boldsymbol{\sigma}}^2$ concentrates as*

$$\frac{\widetilde{\boldsymbol{\sigma}}^2}{d} - (\widetilde{\gamma}-1)\sigma_\alpha^2 \xrightarrow[d\to\infty]{a.s.} 0.$$

Proof. According to Lemma 1, the statistical noise is given by

$$\widetilde{\boldsymbol{\sigma}}^2 = \sigma^2 + \|\boldsymbol{\alpha}\|^2_{I_l - M^+M}.$$

The term σ^2 is assumed to be constant, but the quadratic form $\|\boldsymbol{\alpha}\|^2_{I_l - M^+M}$ grows with d and is controlled by Lemma 7 as

$$\begin{aligned}\frac{\widetilde{\boldsymbol{\sigma}}^2}{d} &= \frac{\sigma^2}{d} + \frac{1}{d}\boldsymbol{\alpha}^T(I_l - M^+M)\boldsymbol{\alpha} \overset{a.s.}{\approx} \frac{\mathrm{Tr}(I_l - M^+M)}{d}\sigma^2_\alpha \\ &= \frac{(l - \mathrm{Tr}(MM^+))}{d}\sigma^2_\alpha \\ &= \frac{l-d}{d}\sigma^2_\alpha \\ &= (\widetilde{\gamma} - 1)\sigma^2_\alpha. \end{aligned}$$

■

The second technical lemma covers the asymptotic behavior of traces that involve both the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ and the population covariance matrix Σ.

Lemma 21 (Asymptotics of mixed terms). *Under Assumption 5, it holds for any $\theta \geq 0$ that*

$$\frac{1}{d}\,\mathrm{Tr}\left[(\widehat{\boldsymbol{\Sigma}} + \theta)^{-1}\widehat{\boldsymbol{\Sigma}}\Sigma^+\right] \xrightarrow[d\to\infty]{a.s.} \gamma\theta m(-\theta)^2 + (1-\gamma)m(-\theta)$$

and

$$\begin{aligned}\frac{1}{d}\,\mathrm{Tr}\left[(\widehat{\boldsymbol{\Sigma}} + \theta)^{-2}\widehat{\boldsymbol{\Sigma}}\Sigma^+\right] \xrightarrow[d\to\infty]{a.s.} & -\gamma m(-\theta)^2 + 2\gamma\theta m(-\theta)M(-\theta) \\ & + (1-\gamma)M(-\theta),\end{aligned}$$

where $m(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}[\frac{1}{\boldsymbol{\lambda}+\theta}]$ and $M(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}[\frac{1}{(\boldsymbol{\lambda}+\theta)^2}]$.

Proof. The asymptotic behavior of these quadratic forms is not covered by Theorem 4 because the dependencies between $\widehat{\boldsymbol{\Sigma}}$ and Σ create complications. To treat these, we require an additional result by [26] combined with Vitali's convergence theorem which, in our notation, states that

$$\frac{1}{d}\,\mathrm{Tr}\left((\widehat{\boldsymbol{\Sigma}} - z)^{-1}g(\Sigma)\right) \xrightarrow[d\to\infty]{a.s.} -\frac{1}{z}\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[\frac{g(\boldsymbol{\lambda})}{\widetilde{m}(z)\boldsymbol{\lambda} + 1}\right].$$

We first use this result to obtain the limit for $\frac{1}{d}\,\mathrm{Tr}((\widehat{\boldsymbol{\Sigma}} - z)^{-1}\Sigma^+)$ by considering $g(\boldsymbol{\lambda}) = 1/\boldsymbol{\lambda}$ and the identity

$$-\frac{1}{z\boldsymbol{\lambda}}\frac{1}{\widetilde{m}(z)\boldsymbol{\lambda}+1} = \frac{1}{z}\left(\frac{1}{\boldsymbol{\lambda} - \left(-\frac{1}{\widetilde{m}(z)}\right)} - \frac{1}{\boldsymbol{\lambda}}\right),$$

which yields

$$\frac{1}{d}\operatorname{Tr}\left((\widehat{\boldsymbol{\Sigma}}-z)^{-1}\Sigma^{+}\right)\xrightarrow[d\to\infty]{a.s.}\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}\left[-\frac{1}{z\boldsymbol{\lambda}}\frac{1}{\tilde{m}(z)\boldsymbol{\lambda}+1}\right]$$
$$=\frac{1}{z}m_{\nu}\left(-\frac{1}{\tilde{m}(z)}\right)-\frac{1}{z}m_{\nu}(0),$$

where we recall that $m_\nu(z)=\boldsymbol{E}_{\boldsymbol{\lambda}\sim\nu}[\frac{1}{\boldsymbol{\lambda}-z}]$. To relate the population Stieltjes transform m_ν back to the sample Stieltjes transforms m and $\tilde{m}$, we can use the identities from Theorem 4 to obtain

$$\frac{1}{d}\operatorname{Tr}\left((\widehat{\boldsymbol{\Sigma}}-z)^{-1}\Sigma^{+}\right)\xrightarrow[d\to\infty]{a.s.}-\gamma m(z)\tilde{m}(z)-\frac{1}{z}m_{\nu}(0)\qquad\text{(equation (3.4))}$$
$$=-\gamma m(z)^{2}+\frac{1-\gamma}{z}m(z)-\frac{1}{z}m_{\nu}(0).\qquad\text{(equation (3.3))}$$

Evaluating the above expression at $z=-\theta$ then yields

$$\frac{1}{d}\operatorname{Tr}\left((\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\Sigma^{+}\right)\xrightarrow[d\to\infty]{a.s.}-\gamma m(-\theta)^{2}-\frac{1-\gamma}{\theta}m(-\theta)+\frac{1}{\theta}m_{\nu}(0).$$

All that remains is to relate $(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\Sigma^{+}$ to the terms we are interested in. Using the identity $(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\widehat{\boldsymbol{\Sigma}}=I-\theta(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}$, we get the first statement of this lemma:

$$\frac{1}{d}\operatorname{Tr}\left[(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\widehat{\boldsymbol{\Sigma}}\Sigma^{+}\right]=\frac{1}{d}\operatorname{Tr}\left[\Sigma^{+}\right]-\theta\frac{1}{d}\operatorname{Tr}\left[\left(\widehat{\boldsymbol{\Sigma}}+\theta\right)^{-1}\Sigma^{+}\right]$$
$$\xrightarrow[d\to\infty]{a.s.}m_{\nu}(0)-\theta\left(-\gamma m(-\theta)^{2}-\frac{1-\gamma}{\theta}m(-\theta)+\frac{1}{\theta}m_{\nu}(0)\right)$$
$$=\gamma\theta m(-\theta)^{2}+(1-\gamma)m(-\theta).$$

The second statement of this lemma also follows directly by taking the derivative, which can be exchanged with the limit $d\to\infty$ using similar arguments as in the main paper after Proposition 16, to obtain

$$\frac{1}{d}\operatorname{Tr}\left[(\widehat{\boldsymbol{\Sigma}}+\theta)^{-2}\widehat{\boldsymbol{\Sigma}}\Sigma^{+}\right]=-\partial_{\theta}\frac{1}{d}\operatorname{Tr}\left[(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\widehat{\boldsymbol{\Sigma}}\Sigma^{+}\right]$$
$$\xrightarrow[d\to\infty]{a.s.}-\partial_{\theta}\left(\gamma\theta m(-\theta)^{2}+(1-\gamma)m(-\theta)\right)$$
$$=-\gamma m(-\theta)^{2}+2\gamma\theta m(-\theta)M(-\theta)+(1-\gamma)M(-\theta),$$

where the last step used $\partial_\theta m(-\theta)=M(-\theta)$. ■

We are now ready to give the full proof of Theorem 13.

Theorem 13 (Plug-in estimator is generally biased). *Under Assumption 5 with $\theta^*>0$, the following hold.*

(1) *For all $\theta \geq 0$, the derivative of the function from equation (4.7) satisfies*

$$\partial_\theta \boldsymbol{f}_d^{\mathrm{plg}}(\theta) \xrightarrow{a.s.} \left[\theta - (1+\gamma\widetilde{\gamma})\theta^* + \gamma\theta^*(1-\theta m(-\theta))\left(1 + \frac{M(-\theta)}{M(-\theta)-m(-\theta)^2}\right)\right] h(\theta), \tag{4.8}$$

with

$$h(\theta) = (M(-\theta) - m(-\theta)^2) \cdot (1 - \theta m(-\theta) + (1-2\gamma+\gamma\widetilde{\gamma})\theta^* m(-\theta) + \gamma\theta\theta^* m(-\theta)^2)^{-1},$$

$m(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}[1/(\boldsymbol{\lambda}+\theta)]$*, and* $M(-\theta) = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}[1/(\boldsymbol{\lambda}+\theta)^2]$.

(2) *For every $d \in \mathbb{N}$, let $\boldsymbol{\theta}_d^{\mathrm{plg}}$ be a root of $\partial_\theta \boldsymbol{f}_d^{\mathrm{plg}}$ if it exists or* 0 *otherwise. Additionally, assume that $\widetilde{\gamma}$ does not satisfy*

$$\widetilde{\gamma} = (1-\theta^* m(-\theta^*))\left(1 + \frac{M(-\theta^*)}{M(-\theta^*)-m(-\theta^*)^2}\right). \tag{4.9}$$

Then, the sequence $\{\boldsymbol{\theta}_d^{\mathrm{plg}}\}$ almost surely does not converge to θ^*.

Proof. We first show equation (4.8). This proof for the plug-in quantities $\widehat{\boldsymbol{\Sigma}}, \widehat{\boldsymbol{\beta}}$ follows the same strategy as the proof of Theorem 10 for $\Sigma, \widetilde{\boldsymbol{\beta}}$, but additional complications arise because $\widehat{\boldsymbol{\beta}}$ asymptotically depends on both the population term M and the empirical quantities. Similarly, as for $\widetilde{\boldsymbol{\beta}}$, we treat $\widehat{\boldsymbol{\beta}}$ by combining the equations $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{Y}$, $\boldsymbol{Y} = \boldsymbol{X}^T\widetilde{\boldsymbol{\beta}} + \boldsymbol{E}$ for $\boldsymbol{E} \sim \mathcal{N}(0, \widetilde{\boldsymbol{\sigma}}^2 I_n)$, and $\widetilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + M^{+T}\boldsymbol{\alpha}$ to obtain

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d & \widetilde{\boldsymbol{\sigma}}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\end{pmatrix}\boldsymbol{v} \quad \text{for some } \boldsymbol{v} \sim \mathcal{N}(0, I_{l+d+n}).$$

As before, we get for $k \in \{1,2\}$ that

$$\begin{aligned}
&\frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\widehat{\boldsymbol{\beta}} \\
&= \frac{1}{d}\boldsymbol{v}^T\begin{pmatrix}\sigma_\alpha M^+ \\ \sigma_\beta I_d \\ \widetilde{\boldsymbol{\sigma}}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\end{pmatrix}\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d & \widetilde{\boldsymbol{\sigma}}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\end{pmatrix}\boldsymbol{v} \\
&\overset{a.s.}{\approx} \frac{1}{d}\operatorname{Tr}\left[\begin{pmatrix}\sigma_\alpha M^+ \\ \sigma_\beta I_d \\ \widetilde{\boldsymbol{\sigma}}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\end{pmatrix}\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d & \widetilde{\boldsymbol{\sigma}}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\end{pmatrix}\right] && \text{(Lemma 7)}
\end{aligned}$$

$$= \frac{1}{d}\operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\begin{pmatrix}\sigma_\alpha M^{+T} & \sigma_\beta I_d & \widetilde{\boldsymbol{\sigma}}(XX^T)^+X\end{pmatrix}\begin{pmatrix}\sigma_\alpha M^+\\ \sigma_\beta I_d\\ \widetilde{\boldsymbol{\sigma}}X^T(XX^T)^+\end{pmatrix}\right]$$

$$(\operatorname{Tr}(A\cdot B)=\operatorname{Tr}(B\cdot A))$$

$$= \frac{1}{d}\operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\left(\sigma_\alpha^2\Sigma^+ + \sigma_\beta^2 I_d + \frac{\widetilde{\boldsymbol{\sigma}}^2}{n}\widehat{\boldsymbol{\Sigma}}^{-1}\right)\right]$$

$$= \frac{1}{d}\operatorname{Tr}\left[\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}(\sigma_\alpha^2\Sigma^+ + \sigma_\beta^2 I_d + \gamma(\widetilde{\gamma}-1)\sigma_\alpha^2\widehat{\boldsymbol{\Sigma}}^{-1})\right] \qquad \text{(Lemma 20)}$$

$$= \frac{\sigma_\beta^2}{d}\operatorname{Tr}\left[(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}(\widehat{\boldsymbol{\Sigma}}+\gamma(\widetilde{\gamma}-1)\theta^*)\right] + \theta^*\frac{\sigma_\beta^2}{d}\operatorname{Tr}\left[(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\widehat{\boldsymbol{\Sigma}}\Sigma^+\right].$$

The second term contains both the population term Σ and the sample term $\widehat{\boldsymbol{\Sigma}}$, which is treated separately in Lemma 21. For readability, we use the shorthand notation

$$m = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}\left[\frac{1}{\boldsymbol{\lambda}+\theta}\right] \quad\text{and}\quad M = \boldsymbol{E}_{\boldsymbol{\lambda}\sim\mu}\left[\frac{1}{(\boldsymbol{\lambda}+\theta)^2}\right],$$

under which the limit for the first term is given by

$$\frac{1}{d}\operatorname{Tr}\left[(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}(\widehat{\boldsymbol{\Sigma}}+\gamma(\widetilde{\gamma}-1)\theta^*)\right]\xrightarrow[d\to\infty]{a.s.}\begin{cases}1-\theta m+\gamma(\widetilde{\gamma}-1)\theta^* m & \text{for } k=1,\\ m-\theta M+\gamma(\widetilde{\gamma}-1)\theta^* M & \text{for } k=2.\end{cases}$$

Combined with Lemma 21, this yields

$$\begin{aligned}&\frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-k}\widehat{\boldsymbol{\beta}}\\ &\xrightarrow[d\to\infty]{a.s.}\begin{cases}1-\theta m+\theta^*(\gamma\theta m^2+(1-2\gamma+\gamma\widetilde{\gamma})m), & k=1,\\ m-\theta M+\theta^*(-\gamma m^2+2\gamma\theta mM+(1-2\gamma+\gamma\widetilde{\gamma})M), & k=2.\end{cases}\end{aligned}$$

Together with $m_{\widehat{\boldsymbol{\Sigma}}}(-\theta)\xrightarrow[d\to\infty]{a.s.} m$, this covers the individual components of

$$\partial_\theta \boldsymbol{f}^{\text{plg}}(\theta) = \boldsymbol{m}_{\widehat{\boldsymbol{\Sigma}}}(-\theta) - \frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-2}\widehat{\boldsymbol{\beta}}/\frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\widehat{\boldsymbol{\beta}}.$$

It remains to plug everything in, which we do after factoring out the denominator $\frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\widehat{\boldsymbol{\beta}}$ to obtain

$$\begin{aligned}&\boldsymbol{m}_{\widehat{\boldsymbol{\Sigma}}}(-\theta)\cdot\frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-1}\widehat{\boldsymbol{\beta}} - \frac{1}{d}\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}+\theta)^{-2}\widehat{\boldsymbol{\beta}}\\ &\xrightarrow[d\to\infty]{a.s.} m\cdot\left[1-\theta m+\theta^*(\gamma\theta m^2+(1-2\gamma+\gamma\widetilde{\gamma})m)\right]\\ &\qquad -\left[m-\theta M+\theta^*(-\gamma m^2+2\gamma\theta mM+(1-2\gamma+\gamma\widetilde{\gamma})M)\right]\\ &\qquad = (\theta-(1-2\gamma+\gamma\widetilde{\gamma})\theta^*)\cdot(M-m^2)+\gamma\theta^*\left(\theta m^3+m^2-2\theta mM\right)\end{aligned}$$

$$
\begin{aligned}
&= (\theta - (1 - 2\gamma + \gamma\widetilde{\gamma})\theta^*) \cdot (M - m^2) \\
&\quad + \gamma\theta^*(2m^2 - 2M - (1 - \theta m)m^2 + 2(1 - \theta m)M) \\
&= (\theta - (1 + \gamma\widetilde{\gamma})\theta^*) \cdot (M - m^2) + \gamma\theta^*(1 - \theta m)(2M - m^2) \\
&= \left[\theta - (1 + \gamma\widetilde{\gamma})\theta^* + \gamma\theta^*(1 - \theta m)\left(1 + \frac{M}{M - m^2}\right)\right] \cdot (M - m^2),
\end{aligned}
$$

which concludes the first part of the proof.

For the second statement, observe that equation (4.9) is equivalent to $F^{\text{plg}}(\theta^*) = 0$, where F^{plg} is the function on the right-hand side of equation (4.8). The assumption in this theorem therefore states that $F^{\text{plg}}(\theta^*) \neq 0$. Let $(\boldsymbol{\theta}_d^{\text{plg}})_{d\in\mathbb{N}}$ be the sequence described in the theorem. In the case where $\partial_\theta \boldsymbol{f}_d^{\text{plg}}$ does not have a root infinitely often, we have $\boldsymbol{\theta}_d^{\text{plg}} = 0$ infinitely often and therefore $\boldsymbol{\theta}_d^{\text{plg}} \not\to \theta^*$ as $d \to \infty$ since $\theta^* \neq 0$. Therefore, now assume that $\boldsymbol{\theta}_d^{\text{plg}}$ is a root of $\partial_\theta \boldsymbol{f}_d^{\text{plg}}$ eventually. Assume that the claim is false, that is, $\boldsymbol{\theta}_d^{\text{plg}} \xrightarrow[d\to\infty]{} \theta^*$ with positive probability. Similarly to the proof of Theorem 10, we get that the convergence in equation (4.8) holds almost surely uniformly on $[0, C]$ for some $C > \theta^*$. The convergence $\boldsymbol{\theta}_d^{\text{plg}} \to \theta^*$ also implies that $\boldsymbol{\theta}_d^{\text{plg}} \in [0, C]$ eventually. Putting everything together, we get for sufficiently large d that

$$
\begin{aligned}
|F^{\text{plg}}(\theta^*)| &= |F^{\text{plg}}(\theta^*) - \partial_\theta \boldsymbol{f}_d^{\text{plg}}(\boldsymbol{\theta}_d^{\text{plg}})| \qquad (\partial_\theta \boldsymbol{f}_d^{\text{plg}}(\boldsymbol{\theta}_d^{\text{plg}}) = 0) \\
&\leq |\partial_\theta \boldsymbol{f}_d^{\text{plg}}(\boldsymbol{\theta}_d^{\text{plg}}) - F^{\text{plg}}(\boldsymbol{\theta}_d^{\text{plg}})| + |F^{\text{plg}}(\boldsymbol{\theta}_d^{\text{plg}}) - F^{\text{plg}}(\theta^*)| \\
&\leq \sup_{\theta\in[0,C]} |\partial_\theta \boldsymbol{f}_d^{\text{plg}}(\theta) - F^{\text{plg}}(\theta)| + |F^{\text{plg}}(\boldsymbol{\theta}_d^{\text{plg}}) - F^{\text{plg}}(\theta^*)| \\
&\xrightarrow{a.s.} 0,
\end{aligned}
$$

where the first summand goes to 0 by uniform convergence and the second summand goes to 0 by continuity of F^{plg} and $\boldsymbol{\theta}_d^{\text{plg}} \to \theta^*$. This implies $F^{\text{plg}}(\theta^*) = 0$, which is a contradiction. ■

C. RMT consistent estimators for quantities of interest

Theorem 22 (Consistent estimation of statistical noise). *Under the model in equation (3.1),*

$$
\frac{1}{1-\gamma}\frac{\|\boldsymbol{Y}\|^2_{I-\boldsymbol{X}+\boldsymbol{X}}}{nd} - \frac{\widetilde{\boldsymbol{\sigma}}^2}{d} \xrightarrow{a.s.} 0.
$$

Proof. We have

$$
\frac{1}{nd}\|\boldsymbol{Y}\|^2 = \frac{1}{nd}\|\boldsymbol{X}^T\widetilde{\boldsymbol{\beta}} + \boldsymbol{E}\|^2 = \frac{1}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\widetilde{\boldsymbol{\beta}} + \frac{1}{nd}\boldsymbol{E}^T\boldsymbol{E} + \frac{2}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{E}.
$$

We know that the minimum l_2 norm estimator admits a following closed form solution given by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{Y} = (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}(\boldsymbol{X}^T\widetilde{\boldsymbol{\beta}} + \boldsymbol{E}) \overset{w.h.p.}{=} \widetilde{\boldsymbol{\beta}} + (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E},$$

where we used the fact that $\operatorname{rank}(\boldsymbol{X}\boldsymbol{X}^T) = d$ with high probability (w.h.p.) to arrive at the last equality. Letting $\boldsymbol{\kappa} = (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E}$, we have

$$\begin{aligned}\frac{1}{nd}\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\widehat{\boldsymbol{\beta}} &= \frac{1}{nd}(\widetilde{\boldsymbol{\beta}} + \boldsymbol{\kappa})^T\boldsymbol{X}\boldsymbol{X}^T(\widetilde{\boldsymbol{\beta}} + \boldsymbol{\kappa}),\\ &= \frac{1}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\widetilde{\boldsymbol{\beta}} + \frac{1}{nd}\boldsymbol{\kappa}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\kappa} + \frac{2}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\kappa}.\end{aligned}$$

From the closed form expression for $\widehat{\boldsymbol{\beta}}$, we have

$$\begin{aligned}\frac{1}{nd}\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\widehat{\boldsymbol{\beta}} &= \frac{1}{nd}\boldsymbol{Y}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{Y},\\ &= \frac{1}{nd}\boldsymbol{Y}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{Y},\\ &= \frac{1}{nd}\boldsymbol{Y}^T\boldsymbol{X}^+\boldsymbol{X}\boldsymbol{Y}.\end{aligned}$$

Similarly, substituting $\boldsymbol{\kappa} = (\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E}$, we have

$$\begin{aligned}\frac{1}{nd}\boldsymbol{\kappa}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\kappa} &= \frac{1}{nd}\boldsymbol{E}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E},\\ &= \frac{1}{nd}\boldsymbol{E}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E},\\ &= \frac{1}{nd}\boldsymbol{E}^T\boldsymbol{X}^+\boldsymbol{X}\boldsymbol{E},\\ &= \frac{\gamma\widetilde{\boldsymbol{\sigma}}^2}{d} + \mathcal{O}(1/\sqrt{d}).\end{aligned}$$

To derive the last equality, we show first that

$$\frac{1}{nd}\boldsymbol{E}^T\boldsymbol{X}^+\boldsymbol{X}\boldsymbol{E} = \frac{\widetilde{\boldsymbol{\sigma}}^2}{nd}\operatorname{Tr}[\boldsymbol{X}^+\boldsymbol{X}] + \mathcal{O}(1/\sqrt{p})$$

with Lemma 7. The equality follows using $\operatorname{Tr}[AA^+] = \operatorname{rank}(A)$ for any $A \in \mathbb{R}^{n\times d}$ and

$$\frac{1}{nd}\boldsymbol{E}^T\boldsymbol{X}^+\boldsymbol{X}\boldsymbol{E} = \frac{\gamma\widetilde{\boldsymbol{\sigma}}^2}{d} + \mathcal{O}(1/\sqrt{d}).$$

Now, let us consider the term $\frac{2}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\kappa}$:

$$\begin{aligned}\frac{2}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\kappa} &= \frac{2}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E},\\ &= \frac{2}{nd}\widetilde{\boldsymbol{\beta}}^T\boldsymbol{X}\boldsymbol{E} \xrightarrow[d\to\infty]{a.s.} \infty. \qquad \text{(Strong law of large numbers)}\end{aligned}$$

Following similar arguments, we have

$$\frac{1}{nd}\boldsymbol{E}^T\boldsymbol{E} = \frac{\widetilde{\boldsymbol{\sigma}}^2}{d} + \mathcal{O}\left(\frac{1}{d\sqrt{n}}\right).$$

Putting everything together, we have

$$\begin{aligned}\frac{1}{nd}\|\boldsymbol{Y}\|^2 &= \frac{1}{nd}\boldsymbol{Y}^T\boldsymbol{X}^+\boldsymbol{X}\boldsymbol{Y} - \frac{\gamma\widetilde{\boldsymbol{\sigma}}^2}{d} + \frac{\widetilde{\boldsymbol{\sigma}}^2}{d} + \mathcal{O}(1/\sqrt{d})\\ \frac{\widetilde{\boldsymbol{\sigma}}^2}{d} &= \frac{1}{(1-\gamma)nd}\|\boldsymbol{Y}\|^2_{\boldsymbol{I}-\boldsymbol{X}^+\boldsymbol{X}} + \mathcal{O}(1/\sqrt{d}).\end{aligned}$$ ■

Lemma 23 (Asymptotics of quadratic form with a deterministic sequence). *For any $\theta \in \mathbb{R}^+$, let η be the unique solution in $\mathbb{R}^-$ satisfying*

$$\tilde{m}(\eta) = 1/\theta.$$

Then, for any deterministic sequence of vectors $\{v_d\}$ with uniformly bounded (Euclidean) norm, as $d, n \to \infty$ such that $d/n \to \gamma \in (0,1)$, we have

$$\langle v_d, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}} - \eta)^{-1}v_d\rangle - \langle v_d, \Sigma(\Sigma + \theta)^{-1}v_d\rangle \xrightarrow{a.s.} 0.$$

Proof. Observe that, for any $\eta < 0$, we have

$$\langle v_d, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}} - \eta)^{-1}v_d\rangle = \|v_d\|^2 - \langle v_d, (\widehat{\boldsymbol{\Sigma}} - \eta)^{-1}v_d\rangle.$$

The result follows from the generalized Marchenko-Pastur theorem [36], which states that, for any $\theta \in \mathbb{R}^+$,

$$\langle v_d, (\widehat{\boldsymbol{\Sigma}} - \eta)^{-1}v_d\rangle - \langle v_d, (\Sigma + \theta)^{-1}v_d\rangle \xrightarrow{a.s.} 0.$$ ■

Proposition 15 (A consistent estimator for the quadform). *Under Assumption 5, for any $\theta \in \mathbb{R}^+$, let η be the unique solution in $\mathbb{R}^-$ satisfying $\tilde{m}(\eta) = 1/\theta$. Then, as $d, n \to \infty$ such that $d/n \to \gamma \in (0,1)$,*

$$\frac{\frac{1}{d}\langle\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}} - \eta)^{-1}\widehat{\boldsymbol{\beta}}\rangle - \frac{\boldsymbol{S}}{\theta} - \frac{\boldsymbol{S}(1-\gamma)}{\eta}}{\frac{1}{d}\|\widehat{\boldsymbol{\beta}}\|^2 - \boldsymbol{S}\gamma m(0)} - \left\langle\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma + \theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle \xrightarrow{a.s.} 0,$$

where $\boldsymbol{S} = (1-\gamma)^{-1}\|\boldsymbol{Y}\|^2_{\boldsymbol{I}-\boldsymbol{X}^+\boldsymbol{X}}/(nd)$.

Proof. Let η be the unique solution in $\mathbb{R}^-$ satisfying $\tilde{m}(\eta) = 1/\theta$. From Lemma 23, we have, for any $\theta \in \mathbb{R}^+$ as $n, d \to \infty$ such that $d/n \to \gamma \in (0,1)$,

$$\left\langle\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}} - \eta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle - \left\langle\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \Sigma(\Sigma + \theta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\right\rangle \xrightarrow{a.s.} 0 \qquad \text{(C.1)}$$

Therefore, it suffices to consistently estimate $\langle \frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\frac{\widetilde{\boldsymbol{\beta}}}{\|\widetilde{\boldsymbol{\beta}}\|}\rangle$. First, we characterize the asymptotic behavior of $\frac{1}{d}\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widehat{\boldsymbol{\beta}}\rangle$, where

$$\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{\sigma}}^2(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E},$$

where $\boldsymbol{E} \sim \mathcal{N}(0, I_n)$. We have

$$\begin{aligned}\frac{1}{d}\langle \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widehat{\boldsymbol{\beta}}\rangle &= \frac{1}{d}\langle \widetilde{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widetilde{\boldsymbol{\beta}}\rangle + \frac{2\widetilde{\boldsymbol{\sigma}}^2}{d}\widetilde{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E}\\ &\quad + \frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\boldsymbol{E}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E}.\end{aligned}$$

The first term in the expansion resembles the quantity of interest.

For the second term, notice that, since $\boldsymbol{E} \sim \mathcal{N}(0, I_n)$,

$$\frac{2\widetilde{\boldsymbol{\sigma}}^2}{d}\widetilde{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E} \sim \mathcal{N}\left(0, \left\|\frac{2\widetilde{\boldsymbol{\sigma}}^2}{d}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widetilde{\boldsymbol{\beta}}\right\|^2\right),$$

where

$$\begin{aligned}&\left\|\frac{2\widetilde{\boldsymbol{\sigma}}^2}{d}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widetilde{\boldsymbol{\beta}}\right\|^2\\ &\quad = \frac{4\widetilde{\boldsymbol{\sigma}}^2}{d^2}\widetilde{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widetilde{\boldsymbol{\beta}}\\ &\quad = \frac{4\widetilde{\boldsymbol{\sigma}}^2}{d^2 n}\widetilde{\boldsymbol{\beta}}^T\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widehat{\boldsymbol{\Sigma}}^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widetilde{\boldsymbol{\beta}}\\ &\quad \xrightarrow{a.s.} 0.\end{aligned}$$

Therefore, the second term vanishes. For the last expression,

$$\begin{aligned}&\frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\boldsymbol{E}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E}\\ &\quad = \frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\frac{1}{n^2}\boldsymbol{E}^T\boldsymbol{X}^T\widehat{\boldsymbol{\Sigma}}^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\widehat{\boldsymbol{\Sigma}}^+\boldsymbol{X}\boldsymbol{E}\\ &\quad \xrightarrow{a.s.} \frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\frac{1}{n}\operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}\right) \quad \text{(Lemma 7 applied to } E \text{ conditioned on } X)\\ &\quad \xrightarrow{a.s.} \gamma\frac{\widetilde{\boldsymbol{\sigma}}^2}{d}m(\eta).\end{aligned}$$

From Theorem 4, we know that

$$m(\eta) = \frac{1}{\gamma}\left(\tilde{m}(\eta) + \frac{1-\gamma}{\eta}\right) = \frac{1}{\gamma}\left(\frac{1}{\theta} + \frac{1-\gamma}{\eta}\right).$$

Therefore,

$$\frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\boldsymbol{E}^T\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^+\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\Sigma}}-\eta)^{-1}(\boldsymbol{X}\boldsymbol{X}^T)^+\boldsymbol{X}\boldsymbol{E}-\frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\left(\frac{1}{\theta}+\frac{1-\gamma}{\eta}\right)\xrightarrow{a.s.}0.$$

Following the same arguments, it is easy to verify that

$$\frac{1}{d}\|\widehat{\boldsymbol{\beta}}\|^2-\frac{\widetilde{\boldsymbol{\sigma}}^2}{d}\gamma m(0)-\frac{1}{d}\|\widetilde{\boldsymbol{\beta}}\|^2\xrightarrow{a.s.}0.$$

Combining the estimators with the result from Theorem 22, we have the desired result. ■

Theorem 19 (RMT estimator is consistent). *Let $\boldsymbol{\theta}_d^{\mathrm{RMT}}$ be defined as a root of $\boldsymbol{h}_{\mathrm{RMT}}(\theta)$ in some $[0, C]$ for some $C < \infty$ if it exists or 0 otherwise. Additionally, assume that ν is not a point mass. Then, under Assumption 5 with $\theta^* > 0$, the sequence $\{\boldsymbol{\theta}_d^{\mathrm{RMT}}\}$ converges a.s. to θ^*.*

Proof. The proof follows the same arguments as the proof of 10. ■

Acknowledgments. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Leena Chennuru Vankadara and Luca Rendsburg.

Funding. This work has been supported by the German Research Foundation through the Cluster of Excellence “Machine Learning – New Perspectives for Science” (EXC 2064/1 no. 390727645) and the BMBF Tübingen AI Center (FKZ: 01IS18039A).

References

[1] Z. Bai and J. W. Silverstein, *Spectral analysis of large dimensional random matrices*. 2nd edn., Springer Ser. Statist., Springer, New York, 2010 Zbl 1301.60002 MR 2567175

[2] A. Bellot and M. van der Schaar, Linear deconfounded score method: scoring DAGs with dense unobserved confounding. *IEEE Trans. Neural Netw. Learn. Syst.* **35** (2024), no. 4, 4948–4962 MR 4734179

[3] P. Bloebaum, D. Janzing, T. Washio, S. Shimizu, and B. Schoelkopf, Cause-effect inference by comparing regression errors. In *Proc. int. conf. artif. intell. stat. (aistats)*, pp. 900–909, 2018

[4] R. J. Bowden and D. A. Turkington, *Instrumental variables*. Econom. Soc. Monogr. Quant. Econ., Cambridge University Press, Cambridge, 1990 Zbl 0744.62149 MR 1113481

[5] Y.-L. Chen, L. Minorics, and D. Janzing, Correcting confounding via random selection of background variables. 2022, arXiv:2202.02150

[6] L. Chennuru Vankadara, L. Rendsburg, U. Luxburg, and D. Ghoshdastidar, Interpolation and regularization for causal learning. In *Adv. neural inf. process. syst. (neurips)*, pp. 36627–36639, 2022

[7] J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder, Smoking and lung cancer: Recent evidence and a discussion of some questions*. *Int. J. Epidemiol.* **38** (2009), no. 5, 1175–1191

[8] R. Couillet and Z. Liao, *Random matrix methods for machine learning*. Cambridge University Press, 2022 Zbl 1493.68001

[9] P. Ding and T. J. VanderWeele, Sensitivity analysis without assumptions. *Epidemiology* **27** (2016), no. 3, article no. 368

[10] E. Dobriban and S. Wager, High-dimensional asymptotics of prediction: Ridge regression and classification. *Ann. Statist.* **46** (2018), no. 1, 247–279 Zbl 1428.62307 MR 3766952

[11] W. D. Flanders and M. J. Khoury, Indirect assessment of confounding: Graphic description and limits on effect of adjusting for covariates. *Epidemiology* **1** (1990), no. 3, 239–246

[12] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, Surprises in high-dimensional ridgeless least squares interpolation. *Ann. Statist.* **50** (2022), no. 2, 949–986 Zbl 1486.62202 MR 4404925

[13] C. Heinze-Deml, J. Peters, and N. Meinshausen, Invariant causal prediction for nonlinear models. *J. Causal Inference* **6** (2018), no. 2, article no. 20170016 MR 4335430

[14] P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf, Nonlinear causal discovery with additive noise models. In *Adv. neural inf. process. syst. (neurips)*, pp. 689–696, 2008

[15] P. O. Hoyer, S. Shimizu, A. J. Kerminen, and M. Palviainen, Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *Internat. J. Approx. Reason.* **49** (2008), no. 2, 362–378 Zbl 1184.62224 MR 2460274

[16] G. W. Imbens and J. D. Angrist, Identification and estimation of local average treatment effects. *Econometrica* **62** (1994), no. 2, 467–475 Zbl 0800.90648

[17] D. Janzing, Causal regularization. In *Adv. neural inf. process. syst. (neurips)*, pp. 12704–1271, 2019

[18] D. Janzing, J. Peters, J. Mooij, and B. Schölkopf, Identifying confounders using additive noise models. In *Proc. conf. uncertainty artif. intell. (uai)*, pp. 249–257, 2009

[19] D. Janzing and B. Schölkopf, Causal inference using the algorithmic Markov condition. *IEEE Trans. Inform. Theory* **56** (2010), no. 10, 5168–5194 Zbl 1366.62020 MR 2808671

[20] D. Janzing and B. Schölkopf, Detecting confounding in multivariate linear models via spectral analysis. *J. Causal Inference* **6** (2018), no. 1, article no. 20170013 MR 4351485

[21] D. Janzing and B. Schölkopf, Detecting non-causal artifacts in multivariate linear regression models. In *Proc. int. conf. mach. learn. (icml)*, pp. 2245–2253, 2018

[22] D. Janzing, E. Sgouritsa, O. Stegle, J. Peters, and B. Schölkopf, Detecting low-complexity unobserved causes. In *Proc. conf. uncertainty artif. intell. (uai)*, pp. 383–391, 2011

[23] D. Kaltenpoth and J. Vreeken, We are not your real parents: Telling causal from confounded using mdl. In *Proc. siam int. conf. data min.*, pp. 199–207, 2019

[24] A. Kammoun, R. Couillet, J. Najim, and M. Debbah, Performance of capacity inference methods under colored interference. *IEEE Trans. Inf. Theory* (2011)

[25] Y. Kano et al., Causal inference using nonnormality. In *Int. symp. sci. model. 30th anniv. inf. criterion*, pp. 261–270, 2003

[26] O. Ledoit and S. Péché, Eigenvectors of some large sample covariance matrix ensembles. *Probab. Theory Related Fields* **151** (2011), no. 1-2, 233–264 Zbl 1229.60009 MR 2834718

[27] J. Lemeire and D. Janzing, Replacing causal faithfulness with algorithmic independence of conditionals. *Minds Mach.* **23** (2013), no. 2, 227–249

[28] F. Liu and L. Chan, Confounder detection in high-dimensional linear models using first moments of spectral measures. *Neural Comput.* **30** (2018), no. 8, 2284–2318 Zbl 1475.62242 MR 3865940

[29] A. Marx and J. Vreeken, Telling cause from effect by local and global regression. *Knowl. Inf. Syst.* **60** (2019), no. 3, 1277–1305

[30] J. Pearl, Causal inference in statistics: An overview. *Stat. Surv.* **3** (2009), 96–146 Zbl 1300.62013 MR 2545291

[31] J. Pearl, *Causality*. 2nd edn., Cambridge University Press, Cambridge, 2009 Zbl 1188.68291 MR 2548166

[32] J. M. Peña, Simple yet sharp sensitivity analysis for unmeasured confounding. *J. Causal Inference* **10** (2022), no. 1, 1–17 MR 4393183

[33] J. Peters, P. Bühlmann, and N. Meinshausen, Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** (2016), no. 5, 947–1012 Zbl 1414.62297 MR 3557186

[34] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: Foundations and learning algorithms*. Adapt. Comput. Mach. Learn., MIT Press, Cambridge, MA, 2017 Zbl 1416.62012 MR 3822088

[35] H. Reichenbach, *The direction of time*. 65, University of California Press, 1956

[36] J. W. Silverstein and Z. D. Bai, On the empirical distribution of eigenvalues of a class of large-dimensional random matrices. *J. Multivariate Anal.* **54** (1995), no. 2, 175–192 Zbl 0833.60038 MR 1345534

[37] A. Sjölander, A note on a sensitivity analysis for unmeasured confounding, and the related E-value. *J. Causal Inference* **8** (2020), no. 1, 229–248 MR 4350084

[38] X. Sun, D. Janzing, and B. Schölkopf, Causal inference by choosing graphs with most plausible markov kernels. In *Proc. int. symp. artif. intell. math.*, pp. 1–11, 2006

[39] E. C. Titchmarsh, *The theory of functions*. 2nd edn., Oxford University Press, Oxford, 1939 Zbl 65.0302.01 MR 3728294

[40] T. J. VanderWeele and O. A. Arah, Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology* **22** (2011), no. 1, 42–52

[41] T. J. VanderWeele and P. Ding, Sensitivity analysis in observational research: Introducing the e-value. *Ann. Intern. Med.* **167** (2017), no. 4, 268–274

[42] T. J. VanderWeele, P. Ding, and M. Mathur, Technical considerations in the use of the E-value. *J. Causal Inference* **7** (2019), no. 2, article no. 20180007 MR 4350069

[43] K. Zhang and A. Hyvärinen, On the identifiability of the post-nonlinear causal model. In *Proc. conf. uncertainty artif. intell. (uai)*, pp. 647–655, 2009

Received 9 December 2022; revised 5 June 2024.

Luca Rendsburg
Department of Computer Science and Tübingen AI Center, University of Tübingen,
Maria-von-Linden-Straße 6, Baden-Württemberg, 72076 Tübingen, Germany;
luca.rendsburg@uni-tuebingen.de

Leena Chennuru Vankadara
Department of Computer Science and Tübingen AI Center, University of Tübingen,
Maria-von-Linden-Straße 6, Baden-Württemberg, 72076 Tübingen, Germany;
leena.chennuru-vankadara@uni-tuebingen.de

Debarghya Ghoshdastidar
School of Computation, Information and Technology and Munich Data Science Institute,
Technical University of Munich, Boltzmannstraße 3, Bavaria, 85748 Garching, Germany;
ghoshdas@cit.tum.de

Ulrike von Luxburg
Department of Computer Science and Tübingen AI Center, University of Tübingen,
Maria-von-Linden-Straße 6, Baden-Württemberg, 72076 Tübingen, Germany;
ulrike.luxburg@uni-tuebingen.de