Journal of Machine Learning Research 23 (2022) X-XX Submitted 8/21; Revised 3/23; Published X/XX

Naïve regression requires weaker assumptions than factor models to adjust for multiple cause confounding *

Justin Grimmer JGRIMMER@STANFORD.EDU
Professor, Department of Political Science Stanford University
Senior Fellow at the Hoover Institution.

Dean Knox DCKNOX@UPENN.EDU
Faculty Fellow of Analytics at Wharton
Assistant Professor, Operations, Information and Decisions Department,
Wharton School, University of Pennsylvania

Brandon M. Stewart BMS4@PRINCETON.EDU
Associate Professor, Department of Sociology
Office of Population Research, Princeton University.

Editor: XXXX

Abstract

The empirical practice of using factor models to adjust for shared, unobserved confounders, $\boldsymbol{Z}$, in observational settings with multiple treatments, $\boldsymbol{A}$, is widespread in fields including genetics, networks, medicine, and politics. Wang and Blei (2019, WB) generalize these procedures to develop the "deconfounder," a causal inference method using factor models of $\boldsymbol{A}$ to estimate "substitute confounders," $\hat{\boldsymbol{Z}}$, then estimating treatment effects—regressing the outcome, $\boldsymbol{Y}$, on part of $\boldsymbol{A}$ while adjusting for $\hat{\boldsymbol{Z}}$. WB claim the deconfounder is unbiased when (among other assumptions) there are no single-cause confounders and $\hat{\boldsymbol{Z}}$ is "pinpointed." We clarify pinpointing requires each confounder to affect *infinitely* many treatments. We prove that when the conditions hold for the deconfounder to be asymptotically unbiased, a naïve semiparametric regression of $\boldsymbol{Y}$ on $\boldsymbol{A}$ which ignores confounding is also asymptotically unbiased. We provide bias formulas for finite numbers of treatments and show that different deconfounders exhibit different kinds of bias. We replicate every deconfounder analysis with available data and find that neither the naïve regression nor the deconfounder consistently outperform the other. In practice, the deconfounder produces implausible estimates in WB's case study to movie earnings: estimates suggest comic author Stan Lee's cameo appearances causally contributed $15.5 billion, most of Marvel movie revenue. We conclude neither approach is a viable substitute for careful research design in real-world applications.

Keywords: causal inference, deconfounder, machine learning, unmeasured confounding

*. For helpful discussion and feedback we thank David Blei, Matias Cattaneo, Guilherme Duarte, Chris Felton, Sandy Handan-Nader, Gary King, Apoorva Lal, Christopher Lucas, Ian Lundberg, Jason Mian Luo, Jonathan Mummolo, Clayton Nall, Elizabeth Ogburn, Marc Ratkovic, Fredrik Sävje, Ilya Shpitser, Eric Tchetgen Tchetgen, Rocío Titiunik, Matt Tyler, Daniel Thompson, Yixin Wang, Sean Westwood, Yiqing Xu and three anonymous reviewers. Research reported in this publication was supported by Analytics at Wharton, the Carnegie Corporation of New York, and The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879. The statements made and views expressed are solely the responsibility of the authors.

1. Introduction

Machine learning methods are increasingly applied across statistics, social science, and industry to improve causal inference by flexibly adjusting for background confounders. In a recent important instance, Wang and Blei (2019) (WB) introduces the *deconfounder*, an approach for causal inference on multiple treatments, $\boldsymbol{A}$, that affect an outcome, $\boldsymbol{Y}$, in observational settings where shared, unobserved confounders, $\boldsymbol{Z}$, affect both $\boldsymbol{A}$ and $\boldsymbol{Y}$. The deconfounder fits a factor model to the treatments to estimate a *substitute confounder*, $\hat{\boldsymbol{Z}} = f(\boldsymbol{A})$, a function of the observed treatments; it then estimates treatment effects with a regression of $\boldsymbol{Y}$ on $\hat{\boldsymbol{Z}}$ and some (or all) treatments in $\boldsymbol{A}$. This procedure generalizes a popular genetics estimator (Price et al., 2006) and is closely related to other work that provides conditions for credible estimation even when some confounders are unmeasured (Ćevid et al., 2020).

This paper makes three primary contributions, demonstrating the limits of methods like the deconfounder to adjust for unmeasured confounding. First, we clarify how the assumptions that justify the deconfounder relate to empirical practice. Building on D'Amour (2019b) we show that the conditions justifying the deconfounder require an infinite number of treatments, all of which are affected by a small set of confounders. Second, we show that if the conditions hold for the deconfounder to be unbiased, then a suitably flexible naïve regression that ignores the confounding will be unbiased too. If the deconfounder's conditions fail to hold, we show the analyst would need strong, context-specific information about treatment effects and confounding to determine which approach has less bias. Third, we show that the quality of evidence used to support the deconfounder has been overstated. We reexamine existing simulations and real-world evidence, and we construct our own simulations. We find that neither the deconfounder nor naïve regression consistently outperforms the other; indeed, we find that neither consistently performs well at all.

Our results may be surprising in light of recent work claiming that the deconfounder is a viable tool for many causal inference problems. WB argue that adjusting for the substitute confounder will provide unbiased estimates of causal effects under certain conditions: "the theory finds confounders that are *effectively observed*, even if not explicitly so, and embedded in the multiplicity of the causes" (p. 7). This is the basic intuition underlying the use of factor models for causal inference: *if the following assumptions are satisfied*, multi-cause confounding has observable implications for the joint distribution of the treatments, which the analyst can capture with dimension reduction of $\boldsymbol{A}$ and then adjust for. These are (A.1) there are no confounders that affect only one treatment, (A.2) there is a multi-cause confounder $\boldsymbol{Z}$ satisfying weak unconfoundedness, (A.3) the substitute confounder $\hat{\boldsymbol{Z}} = f(\boldsymbol{A})$ *pinpoints* the confounder $\boldsymbol{Z}$ as defined in Section 2.1, and (A.4) $\boldsymbol{A}$ follow a probabilistic factor model and are conditionally independent given $\boldsymbol{Z}$.[1]

We show that these assumptions lead to a taxing set of conditions that empirical researchers must satisfy in order to use the deconfounder. While it is true that multiple treatments can provide information about unobserved confounding, the deconfounder requires strong conditions that are unlikely to hold even approximately in many settings. For

1. Wang and Blei (2020) condenses into two assumptions: (1) there is a $\boldsymbol{Z}$ which is a smallest σ-algebra such that $P(\boldsymbol{A}|\boldsymbol{Z}) = \prod_{j=1}^{J} P(\boldsymbol{A}_j|\boldsymbol{Z})$ and $P(\boldsymbol{A}_j|\boldsymbol{Z})$ is not a point mass for any value of j. Conditioning on $\boldsymbol{Z}$ must yield $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{Y}(\mathbf{a})|\boldsymbol{Z}$. (2) $\boldsymbol{Z}$ is pinpointed by $\boldsymbol{A}$ s.t. $\mathrm{P}(\boldsymbol{Z}|\boldsymbol{A}) = \delta_{f(\boldsymbol{A})}$. These are stronger than ours due to the minimal σ-algebra condition. Ogburn et al. (2019) gives alternatives.

example, we clarify that pinpointing (A.3) requires *strong infinite confounding*—every confounder must asymptotically affect an infinite number of treatments, and each confounder's effects on those treatments must not go to zero too quickly (D'Amour, 2019b). Thus, beyond the "no single-cause confounding" assumption (A.1), consistent causal inference further requires *no finite-cause confounding*. Theoretically, we show that in the absence of infinite causes, deconfounder methods cannot consistently estimate causal effects—even as the number of observations goes to infinity.[2]

Under the conditions where the deconfounder is consistent, other estimators that ignore confounding will also be consistent. Our main theoretical result, Theorem 1, demonstrates that what WB term *naïve* regressions—the class of (possibly semiparametric) regressions that directly estimate $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{A}]$, ignoring confounding—asymptotically approach consistency as the number of treatments, m, grows large. In other words, everything but the multiplicity of treatments is unnecessary: the deconfounder's added factor-model machinery is not needed to achieve theoretical identification of causal effects.

While the same conditions that ensure consistency of the deconfounder also imply consistency of naïve regression, the two approaches suffer from different kinds of biases when these conditions are violated. To help build intuition, we derive analytic bias expressions for a specific family of data-generating processes in which confounders, treatments, and outcome follow normal linear models. We show that the bias of the naïve regression comes from omitted variable bias due to the unmeasured confounder, which goes to zero under strong infinite confounding. We further show that different types of deconfounders suffer from different kinds of biases. The *subset deconfounder*—a variant popular in genetics (Price et al., 2006), which regresses $\boldsymbol{Y}$ on only $\hat{\boldsymbol{Z}}$ and a *subset* of $\boldsymbol{A}$—purges bias from the confounder, at the expense of inducing a different kind of omitted variable bias due to the excluded treatments. We show that the bias of the subset deconfounder goes to zero if (i) the causal effects of the excluded treatments go to zero and (ii) the confounding satisfies strong infinite confounding or a related condition. The *penalized full deconfounder*, which regresses the outcome on all treatments and the substitute confounder using a penalized regression, has biases that go to zero under strong infinite confounding, balancing the biases from the naïve regression and the subset deconfounder.

Simply put, the strong conditions for either the naïve or deconfounder estimators to be consistent are unlikely to hold in practice. We reach this conclusion after investigating the finite-sample performance of both estimators using all six simulation designs used across deconfounder papers and several simulations of our own design. In line with our theoretical results, we find that neither the naïve nor deconfounder estimators consistently outperform each other. Practically speaking, our analysis also reveals that the factor model machinery introduces new estimation complexity that is avoidable using the naïve regressions.

Nor do we find evidence that either the naïve nor deconfounder is a viable general method for observational causal inference. We revisit WB's main case study, the effects of actor appearances on movie revenue. We show both the deconfounder and naïve regression produce results that are implausible. For example, both WB's model and naïve regression estimate that Stan Lee—the legendary comic writer who appears for 200 seconds in Marvel

2. In principle, consistent estimation can also be achieved with only a finite set of treatments, but only when confounders are so strong that they dominate all noise and perfectly determine the value of the treatments. We view this as a highly unrealistic setting.

superhero films—caused over an eight-fold increase in movie revenue with his cameos, more than any of the 900 other actors analyzed, and contributed $15.5 billion of revenue. We show simply adjusting for film budget yields far more reasonable results, even though budget is a quintessential example of the multi-cause confounding that factor modeling is intended to address.

While machine learning methods do allow for more flexible specifications, they do not alter the basic assumptions needed for unbiased causal inference. In contrast, our results show thoughtful research designs *do* weaken the assumptions needed—such as when researchers use proxy variables associated with confounders, but conditionally independent of treatments and outcome (e.g. Miao et al., 2018, 2022). While multiple treatments can reveal the imprint of an underlying confounder, we show this requires numerous conditions: strong infinite confounding; a large number of associated but not causally ordered treatments; and some guarantee, such as the existence of a factor model (A.4), that the confounders' imprint is recoverable. Given the implausibility of these conditions jointly holding in applied settings, deconfounder-type algorithms are no substitute for explicitly measuring confounders and then adjusting. Machine learning methods are useful, but the foundational challenges of causal inference remain.

We provide background for the deconfounder in Section 2. Section 3 characterizes the asymptotic behavior of the deconfounder and the naïve regression, followed by a discussion of finite-sample implications and performance in Section 4. We discuss implications for real-world studies in Section 5, then conclude. Supplements include a notation guide (Supplement A), proofs (Supplement B), and additional empirical details including replications of every deconfounder simulation and case study with available data (Supplements C–G). All replication code and data is available on Code Ocean.[3]

2. The Deconfounder and Multiple Causal Inference

In this section, we offer background on the deconfounder and preview our results. Throughout, we use $\boldsymbol{A}$ to denote the set of m treatments for which we have n observations, $\boldsymbol{Z}$ is the true shared unobserved confounder, and $\hat{\boldsymbol{Z}}$ is the estimated substitute confounder.

2.1 The Deconfounder and Related Work

WB formalizes and provides statistical theory for a procedure that has been used extensively in genetics, social science, operations research, network science, medicine, and industry (e.g. Price et al., 2006)—estimating a factor model of treatments with the goal of adjusting for shared confounding among those treatments. WB seeks to "develop the deconfounder algorithm, prove that it is unbiased, and show that it requires weaker assumptions than traditional causal inference" (p. 1574). The main algorithm of theirs we consider fits a factor model to the treatments, checks its fit, and then runs a regression of the outcome on all treatments and low-dimensional representations of each unit extracted from the factor model. WB use two new assumptions for developing the statistical theory of the deconfounder: "no single cause confounders" and "consistency of the substitute confounder" (called "pinpointing" in subsequent papers; Wang and Blei, 2020). The latter assumption

3. Each simulation and application has its own replicable capsule at the following links:

requires that the substitute confounder, $\hat{\boldsymbol{Z}}$—which is a deterministic function of the observed treatments—is a bijective transformation of $\boldsymbol{Z}$. While the pinpointing assumption is stated as an exact equality, any method to consistently estimate $\boldsymbol{Z}$ requires asymptotics in n and m. In Definition 2 of Supplement B.1.2, we offer a redefinition of pinpointing as an asymptotic property. We analyze the deconfounder in the most favorable asymptotic regime, where $m \to \infty$ and for each treatment, $n \to \infty$.

Prior to the final publication of WB, D'Amour (2019b) showed that general nonparametric identification for the deconfounder is impossible. The core problem is that the factor model and the "no single cause confounders" assumptions only partially constrain the observed data distribution. To address this WB offer three special cases (their Theorems 6–8) which leverage parametric assumptions (Theorem 6) and limitations on what can be estimated (Theorems 7–8) to achieve identification.[4] Commentaries, published alongside WB, clarified the implications of the assumptions powering these theorems. Imai and Jiang (2019) notes that $\hat{\boldsymbol{Z}}$ converges to a function of the observed treatments rather than the true $\boldsymbol{Z}$, a random variable (Imai and Jiang, 2019, 1607). This is problematic because the adjustment criteria implicitly assumes the support of $p(\hat{\boldsymbol{Z}}_i|\boldsymbol{A}_i = \mathbf{a})$ is the same as that of $p(\hat{\boldsymbol{Z}}_i)$—which cannot be true because pinpointing implies that $p(\hat{\boldsymbol{Z}}_i|\boldsymbol{A}_i = \mathbf{a})$ is degenerate. Ogburn et al. (2019) and D'Amour (2019b) emphasize that pinpointing generally requires m going to infinity.

WB's version of the deconfounder is one instance of a broader strategy for addressing unmeasured confounding. A related line of papers considers spectral deconfounding for the linear model under a dense confounding assumption that is related to the "no single cause confounders" assumption (Ćevid et al., 2020; Bühlmann and Ćevid, 2020; Guo et al., 2022). Spectral deconfounding is closely connected to the penalized deconfounder that we explore below. Broadly speaking, these papers discuss scenarios where confounding disappears asymptotically (Chernozhukov et al., 2017). Throughout this paper, we focus on WB's formulation for the sake of simplicity.

Finally, we note that other papers have studied the multiple-treatment setting using partial identification (Zheng et al., 2021) or proxies for nonparametric identification (Miao et al., 2022).

2.2 Takeaways and Contributions

Because the substitute confounder is a function of the observed treatments, $\hat{\boldsymbol{Z}} = f(\boldsymbol{A})$, the deconfounder estimates $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{A}, \hat{\boldsymbol{Z}}] = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{A}, f(\boldsymbol{A})]$, which reduces to $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{A}]$. In other words, the deconfounder is only a method to learn a transformation of the treatments. There are several important restrictions implicit in the deconfounder assumptions including, notably, that the treatments, $\boldsymbol{A}$, cannot causally depend on each other. We maintain these assumptions and return to them in Section 5.2.

Our contribution is twofold. First, we present Propositions 1–6 and Theorem 1, showing that in a simplified linear model and under the deconfounder assumptions, naïve regression

4. Theorem 6 provides identification of average treatment effects by making parametric assumptions, including that the substitute confounder is piecewise constant and there can be no confounder/cause interactions. Theorem 7 identifies the average treatment effects of a subset of the treatments. Finally, Theorem 8 restricts estimation to only those treatments which map to the same value of the substitute confounder.

is asymptotically (in both n and m) unbiased and that every variant deconfounder estimator only also achieves asymptotic unbiasedness if it uses the same information as a naïve regression. When the deconfounder uses only a subset of the treatments, additional untestable and strong assumptions must be made about the treatment effects. Second, we show through extensive empirical evaluation that the theoretical concerns raised here and in prior papers make the deconfounder and naïve regression unsuitable for current real-world applications.

3. Asymptotic Theory Justifies the Naïve Regression Whenever The Deconfounder Can Be Used

The most basic DGP for multi-cause confounding is a linear factor model and a linear outcome model—a case we call the *linear-linear* model. This is defined in WB in Equations 8, 9 and 20 and implied in other applications of the deconfounder (Price et al., 2006). We first develop asymptotic theory for this simple setting before generalizing to nonlinear factor and outcome models. In Sections 3–4 we assume a simplified best-case setting for the deconfounder where we know the true DGP including the number of confounders, treatment effects are constant, and there is separable confounding. We return to these assumptions and accompanying real-world concerns in Section 5.

3.1 The Linear-Linear Model and Strong Infinite Confounding

Consider n observations drawn i.i.d. from the following DGP:

$$k \text{ unobserved confounders:} \qquad \boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}); \tag{1}$$

$$m \geq k \text{ observed treatments:} \qquad \boldsymbol{A}_i \sim \mathcal{N}(\boldsymbol{Z}_i^\top \boldsymbol{\theta}, \sigma^2 \boldsymbol{I}); \tag{2}$$

$$\text{scalar outcome:} \qquad Y_i \sim \mathcal{N}(\boldsymbol{A}_i^\top \boldsymbol{\beta} + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}, \omega^2); \tag{3}$$

We assume elements of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are finite and σ^2 is nonzero. Our goal is to estimate $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]^\top$, the causal effects of increasing the corresponding $[A_{i,1}, \ldots, A_{i,m}]$ by one unit; following WB, effects are assumed constant. Results are collected in $\boldsymbol{Z}$, $\boldsymbol{A}$, and $\boldsymbol{Y}$.[5]

The variable $\boldsymbol{Z}$ is unobserved and therefore confounds our inferences about the causal effect of $\boldsymbol{A}$ when both $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ are nonzero. However, if the analyst could observe $\boldsymbol{Z}$ and adjust for it, they would have the *oracle* estimator,

$$\left[\hat{\boldsymbol{\beta}}^{\text{oracle}\top}, \hat{\boldsymbol{\gamma}}^{\text{oracle}\top} \right]^\top \equiv \left([\boldsymbol{A}, \boldsymbol{Z}]^\top [\boldsymbol{A}, \boldsymbol{Z}] \right)^{-1} [\boldsymbol{A}, \boldsymbol{Z}]^\top \boldsymbol{Y}. \tag{4}$$

It follows directly from the properties of ordinary least squares that the oracle is an unbiased and consistent estimator of treatment effects for any m.

No other estimator that we will consider is consistent for finite m. Both here and in the general nonlinear case, we will therefore consider asymptotics for sequences of DGPs in which the distribution of the confounder $\boldsymbol{Z}$, its dimensionality k, and the outcome-model confounding parameters $\boldsymbol{\gamma}$ are held fixed. We begin with a one-dimensional ($m = 1$) treatment $\boldsymbol{A}$, generated by factor-model parameters $\boldsymbol{\theta} = \theta_1$, and an outcome model with

5. We occasionally denote simultaneous sampling of all n observations with $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ or similar.

treatment-effect parameters $\boldsymbol{\beta} = \beta_1$. We then consider an $m = 2$ DGP with $\boldsymbol{\theta} = [\theta_1, \theta_2]$ and $\boldsymbol{\beta} = [\beta_1, \beta_2]$, then an $m = 3$ DGP with $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$ and $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3]$, and so on. We will then consider asymptotics as m goes to infinity. In other words, we will consider sequences of DGPs in which new treatments are iteratively added, holding fixed all of the following: the distribution and effects of latent confounders, the relationship between old treatments and confounders, and the effects of old treatments.[6] (For compactness, we will omit the sequence index; all results refer to the behavior of estimators as applied to the m-th DGP.) We will show that within this broad set of DGP sequences, for the subset satisfying what we refer to as "strong infinite confounding," the naïve regression will approach consistency.[7]

Definition 1. *(Strong infinite confounding under the linear-linear model.) A sequence of linear-linear data-generating processes with a fixed number of confounders, k, and growing number of causes, m, is said to be* ***strongly infinitely confounded*** *if as $m \to \infty$, all diagonal elements of $\boldsymbol{\theta}\boldsymbol{\theta}^\top$ tend to infinity.*

The j-th diagonal element of $\boldsymbol{\theta}\boldsymbol{\theta}^\top$ contains the sum of the squared coefficients relating confounder j to each of the m treatments.[8] Intuitively, strong infinite confounding says that as m grows large, the finite k confounders continue to strongly affect a growing number of treatments. We discuss the practical implication of this condition for finite samples in Section 4. In Supplement B.1.2, we build on an example from D'Amour (2019a) to show an example of "weak" infinite confounding, where the number of treatments grows but the diagonal elements of $\boldsymbol{\theta}\boldsymbol{\theta}^\top$ do not tend towards infinity.

3.2 The Naïve Regression in the Linear-Linear Setting

As a baseline for the deconfounder, WB present the *naïve* estimator, which simply ignores $\boldsymbol{Z}$. In Proposition 1, we characterize the asymptotic properties of naïve regression with finite m and, perhaps surprisingly, establish it is asymptotically unbiased for the linear-linear model as both n and m go to infinity under strong infinite confounding.

Proposition 1. *(Asymptotic Bias of the Naïve Regression in the Linear-Linear Model.) Under the linear-linear model, the asymptotic bias of the naïve estimator, $\hat{\boldsymbol{\beta}}^{\text{naïve}} \equiv \left(\boldsymbol{A}^\top \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{Y}$, follows $p\text{-lim}_{n\to\infty} \hat{\boldsymbol{\beta}}^{\text{naïve}} - \boldsymbol{\beta} = (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma}$, indicating that the naïve estimator is inconsistent. However, when applied to a sequence of data-generating processes with growing m which satisfy strong infinite confounding, $\lim_{m\to\infty} p\text{-lim}_{n\to\infty} \hat{\boldsymbol{\beta}}^{\text{naïve}} - \boldsymbol{\beta} = \mathbf{0}$.*

Intuitively, the naïve regression is unbiased as the number of treatments grow, because linear regression adjusts along the eigenvectors of the covariance matrix of the treatments,

6. This formulation generalizes straightforwardly to additively separable settings in which the m-th DGP obeys outcome model $\mathbb{E}[Y_i|\boldsymbol{A}_i, \boldsymbol{Z}_i] = \sum_{j=1}^{m} f_j(A_{i,j}) + g_Y(\boldsymbol{Z}_i)$, allowing for nonlinear effects. Note that the requirement that prior effects remain unchanged when adding a new treatment will generally preclude interactive effects.
7. Guo et al. (2022), defines a related "dense confounding" condition.
8. Lemma 1 of Supplement B proves strong infinite confounding is necessary for pinpointing, and Lemma 5 connects this to the conditions for unbiased estimation of a naïve regression.

and in this setting the most consequential eigenvectors are the confounders. This connects to the core (but incomplete) intuition of the deconfounder and a prior literature in genetics—under deconfounder assumptions, shared confounding leaves an imprint on the observed data distribution (Price et al., 2006; Wang and Blei, 2019, 2020)—and, as it turns out, this imprint is useful for the naïve regression.

The asymptotic regime we use implies that as the number of treatments increases, the relative size of $\boldsymbol{\gamma}$ (the outcome coefficients on the confounder) necessarily decreases. As a result, the amount of confounding, relative to the effect of the treatments, decreases—that is, confounding vanishes. If we scaled $\boldsymbol{\gamma}$ to increase with the number of treatments, then Proposition 1 would not necessarily hold, because the additional explanatory information from the adding treatments could no longer reduce the bias more than the increase in the confounding from the new treatments. We also emphasize that the asymptotic results focus on the element-wise convergence of the treatment effects, or the components of the $\boldsymbol{\beta}$. Importantly, this does not imply that the entire vector converges to the true value of $\boldsymbol{\beta}$ simultaneously. This is because if we expand our focus to more than a finite subset of coefficients, those coefficients can be combined to create errors that fail to disappear as m gets larger. This is a direct consequence of our asymptotic regime, which grows the number of treatments and therefore the size of $\boldsymbol{\beta}$. These concerns arise in the linear-linear model as a consequence of the more general overlap problem (see D'Amour, 2019b).

3.3 Deconfounder Under the Linear-Linear Model

In place of the naïve estimator, WB recommend the *full deconfounder*, which under the linear-linear model proceeds in three steps: (1) take the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, (2) extract the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$, and (3) adjust using

$$\left[\ \hat{\boldsymbol{\beta}}^{\text{full}\top},\ \hat{\boldsymbol{\gamma}}^{\text{full}\top}\ \right]^\top \equiv \left(\left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]\right)^{-1} \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y}. \tag{5}$$

But there is a problem: adding these new terms to the naïve regression renders Equation (5) inestimable. The substitute confounder is merely a linear transformation of the original treatments, $\hat{\boldsymbol{Z}} = \sqrt{n}\boldsymbol{A}\boldsymbol{V}_{1:k}\boldsymbol{D}_{1:k}^{-1}$, meaning that $\left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]$ is always rank-deficient. This perfect collinearity is a consequence of that fact that the inclusion of $\hat{\boldsymbol{Z}}$ brings no new information beyond that contained in the original treatments, $\boldsymbol{A}$.

The deconfounder papers and tutorials deploy variants of the linear-linear model throughout their simulations and empirical examples. To render the model estimable, these examples use two modifications to the full deconfounder to break the perfect collinearity between the treatments and the substitute confounder: (1) the *penalized full deconfounder*, which uses penalized outcome models to estimate treatment effects and adjust for the substitute confounder, and (2) the *posterior full deconfounder*, which adds variation to the substitute confounder $\hat{\boldsymbol{Z}}$ by sampling it from an approximate posterior. The genetics literature uses an approach called the *subset deconfounder* which adjusts for only a subset of treatments in the outcome regression. All three of these methods are ways of approaching the fundamental challenge of perfect collinearity. We analyze the specific type of bias that results from each approach.

3.3.1 Asymptotic Theory for Full Deconfounder Variants

We analyze the *penalized full confounder*, an estimator used in Wang and Blei (2019) and Zhang et al. (2019) through their use of normal priors on regression coefficients. We analyze the frequentist version of this estimator, which uses a ridge penalty. In Supplement B.3, we prove Proposition 2, which gives the asymptotic bias of the penalized full confounder.

Proposition 2. *(Asymptotic Bias of the Penalized Full Deconfounder.)*
The penalized deconfounder estimator, as implemented in WB, is

$$\left[\ \hat{\boldsymbol{\beta}}^{\text{penalty}\top},\ \hat{\boldsymbol{\gamma}}^{\text{penalty}\top}\ \right]^\top \equiv \left(\left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right] + \lambda(n)\boldsymbol{I}\right)^{-1} \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y},$$

where $\hat{\boldsymbol{Z}}$ is obtained by taking the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ and extracting the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$, and $\lambda(n)$ is a ridge penalty assumed to be sublinear in n. Under the linear-linear model, the asymptotic bias of this estimator is given by

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} - \boldsymbol{\beta} = \overbrace{-\boldsymbol{Q}_{1:k}\ \text{diag}_j\left(\frac{1}{\sigma^2 + \Lambda_j + 1}\right)\boldsymbol{Q}_{1:k}^\top\boldsymbol{\beta}}^{\textit{Regularization of treatment effect estimation}} \underbrace{+\boldsymbol{Q}_{1:k}\ \text{diag}_j\left(\frac{\Lambda_j}{\sigma^2 + \Lambda_j + 1}\right)\boldsymbol{Q}_{1:k}^\top\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\gamma}}_{\textit{Regularization of confounder adjustment}},$$

where $\boldsymbol{Q}$ and $\boldsymbol{\Lambda} = [\Lambda_1, \ldots, \Lambda_k, 0, \ldots]$ are respectively eigenvectors and eigenvalues obtained from decomposition of $\boldsymbol{\theta}^\top\boldsymbol{\theta}$. Under strong infinite confounding,

$$\lim_{m\to\infty}\ \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} - \boldsymbol{\beta} = \boldsymbol{0}.$$

Proposition 2 shows that the finite-m bias in $\hat{\boldsymbol{\beta}}^{\text{penalty}}$ comes from two sources: regularization of coefficients and omitted-variable bias from excluding the true confounders, $\boldsymbol{Z}$. Under strong infinite confounding, as m and n grow large, both regularization bias and omitted-variable bias go to zero; the latter is true because $(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}$ goes to $\boldsymbol{0}$. Therefore, like a naïve regression, the $\hat{\boldsymbol{\beta}}^{\text{penalty}}$ estimator is asymptotically consistent in m, but only because it effectively nests the naïve regression.

Briefly, the proof proceeds by examining the singular value decomposition of the augmented data matrix, $\left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right] = \boldsymbol{U}^*\boldsymbol{D}^*\boldsymbol{V}^{*\top}$, and using the facts that (1) the first m components of $\boldsymbol{U}^*$ remain unchanged from $\boldsymbol{U}$, since $\hat{\boldsymbol{Z}}$ are merely rescaled versions of the original left-singular vectors; (2) the first k diagonal elements of $\boldsymbol{D}^{*2}$ are equal to $\boldsymbol{D}^2 + n\boldsymbol{I}$ due to the additional variance of $\hat{\boldsymbol{Z}}$; and (3) the last k diagonal elements of $\boldsymbol{D}^*$ are zero.

The second strategy employed by the deconfounder papers to address perfect collinearity in the linear-linear case is to integrate over an approximate posterior. This renders the deconfounder estimable, since the resulting samples are no longer perfectly collinear with

$\boldsymbol{A}$. Proposition 3 gives the asymptotic bias (proof in Supplement B.6) and shows that sampling the substitute confounder from a posterior is inconsistent with a finite m but converges to a naïve regression as m grows large.[9]

Proposition 3. *(Asymptotic Bias of the Posterior-Mean Deconfounder.)*
The posterior-mean deconfounder estimator is

$$\left[\ \hat{\boldsymbol{\beta}}^{\text{pm}\top},\ \hat{\boldsymbol{\gamma}}^{\text{pm}\top}\ \right]^\top \equiv \int \left([\boldsymbol{A},\boldsymbol{z}]^\top [\boldsymbol{A},\boldsymbol{z}]\right)^{-1} [\boldsymbol{A},\boldsymbol{z}]\, \boldsymbol{Y}\ f(\boldsymbol{z}|\boldsymbol{A})\, \mathrm{d}\boldsymbol{z}\,,$$

where $f(\boldsymbol{z}|\boldsymbol{A})$ is a posterior obtained from Bayesian principal component analysis.[a] *Under the linear-linear model, the asymptotic bias of this estimator is given by*

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{pm}} - \boldsymbol{\beta} = (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma},$$

and under strong infinite confounding,

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{pm}} - \boldsymbol{\beta} = \boldsymbol{0}$$

a. While the regression cannot be estimated when $\boldsymbol{z} = \mathbb{E}[\boldsymbol{Z}|\boldsymbol{A}]$, it is almost surely estimable for samples $\boldsymbol{z}^* \sim f(\boldsymbol{z}|\boldsymbol{A})$ due to posterior uncertainty, which eliminates perfect collinearity with $\boldsymbol{A}$. The posterior-mean implementation of WB evaluates the integral by Monte Carlo methods and thus is able to compute the regression coefficients for each sample.

3.3.2 Asymptotic Theory for the Subset Deconfounder

The genetics literature (Price et al., 2006) uses an alternate version of the deconfounder—the subset deconfounder—which estimates the effect of some treatments, ignores others, and adjusts for the substitute confounder (this formulation is captured by Theorem 7 of WB). After extracting a substitute confounder, $\hat{\boldsymbol{Z}}$, this estimator designates a finite number, m_F, of the m treatments as "focal" (we denote this column subset as $\boldsymbol{A}_F$) and sets aside the remaining m_N "non-focal" treatments ($\boldsymbol{A}_N$). It then regresses the outcome, $\boldsymbol{Y}$, on only $\boldsymbol{A}_F$ and $\hat{\boldsymbol{Z}}$. The subset confounder avoids the collinearity issue if $m_F + k < m$. In the genetics literature this is used to estimate the effect of one treatment at a time.

In Proposition 4, we show that the asymptotic bias of the subset deconfounder remains non-zero *even under strong infinite confounding*. To approach consistency, the subset deconfounder requires additional strong conditions on the effects of the non-focal treatments.

Proposition 4. *(Asymptotic Bias of the Subset Deconfounder.)*
The subset deconfounder estimator, based on Theorem 7 from WB, is

$$\left[\ \hat{\boldsymbol{\beta}}_F^{\text{subset}\top},\ \hat{\boldsymbol{\gamma}}^{\text{subset}\top}\ \right]^\top \equiv \left(\left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]\right)^{-1} \left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y}. \tag{6}$$

9. In Supplement B.5 we also offer a proof of Proposition B.5, that estimators adding a fixed amount of white noise remain asymptotically inconsistent as m grows, even under strong infinite confounding.

where the column subsets $\boldsymbol{A}_F$ and $\boldsymbol{A}_N$ respectively partition $\boldsymbol{A}$ into a finite number of focal causes of interest and non-focal causes. The substitute confounder, $\hat{\boldsymbol{Z}}$, is obtained by taking the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ and extracting the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$. Under the linear-linear model, the asymptotic bias of this estimator is given by

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{subset}} - \boldsymbol{\beta}_F = -\left(\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right)^{-1}\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N,$$

with $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_N$ indicating the column subsets of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{A}_F$ and $\boldsymbol{A}_N$, respectively. The subset deconfounder is unbiased for $\boldsymbol{\beta}_F$ (i) if $\boldsymbol{\theta}_F = \boldsymbol{0}$, (ii) if $\lim_{m\to\infty}\boldsymbol{\theta}_N\boldsymbol{\beta}_N = \boldsymbol{0}$ and $\lim_{m\to\infty}\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]^{-1}$ is convergent, or (iii) if both strong infinite confounding holds and $(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N$ goes to $\boldsymbol{0}$ as $m \to \infty$. If one of these additional conditions hold,

$$\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{subset}} - \boldsymbol{\beta}_F = \boldsymbol{0}$$

A proof is given in Supplement B.7; we interpret conditions (i–iii) below. Intuitively, the subset deconfounder is a biased estimator of the effect of $\boldsymbol{A}_F$ because of mismodeling of the dependence structure among the causes. Though $\boldsymbol{A}_F$ and $\boldsymbol{A}_N$ would be conditionally independent if the true $\boldsymbol{Z}$ could be observed and adjusted for, Lemma 4 shows that they are not conditionally independent given $\hat{\boldsymbol{Z}}$. This mismodeled dependence leads to omitted variable bias when excluding $\boldsymbol{A}_N$. But, unlike naïve regression, strong infinite confounding does not resolve this omitted variable bias for the subset deconfounder. In Supplement B.8 we provide further intuition for this result, leveraging properties of PCA regression to show the subset confounder is effectively a regularized regression that only adjusts along the first k eigenvectors (Hastie et al., 2013).

The subset deconfounder and similar estimators in genetics (e.g. Price et al., 2006) rely on more than the intuition that there is shared confounding.[10] Proposition 4 provides three stronger conditions, of which at least one is required for the subset deconfounder to approach unbiasedness. Condition (i), $\boldsymbol{\theta}_F = \boldsymbol{0}$, states that the focal treatment is unconfounded, and therefore no adjustment is needed. Condition (ii) is $\lim_{m\to\infty}\boldsymbol{\theta}_N\boldsymbol{\beta}_N = \boldsymbol{0}$ and $\lim_{m\to\infty}\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]^{-1}$ is convergent. This will hold if, for example, each element of $\boldsymbol{\beta}_N$, the treatment-outcome effects, and each column of $\boldsymbol{\theta}_N$, the confounder-treatment relationships, are drawn from zero-expectation distributions with finite variance and zero covariance between $\boldsymbol{\beta}_N$ and $\boldsymbol{\theta}_N$. Condition (iii) says that the subset deconfounder will hold if strong infinite confounding holds and $\lim_{m\to\infty}\left(\boldsymbol{\theta}\boldsymbol{\theta}'\right)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N = 0$. This could be satisfied, if, for example, the infinite sum $\sum_{j=1}^m \theta_{N,k',j}\beta_{N,j}$ is convergent for all latent confounders $k' \in \{1,\ldots,k\}$. A necessary (but not sufficient) condition for this to hold is if $\left(\theta_{N,k',j}\ \beta_{N,j}\right)_{j=1}^m$ converges to zero for each k' as $m \to \infty$. For example, analysts might assume that treatment effects in the non-focal set, $(\beta_{N,j})_{j=1}^m$, go to zero fast enough to ensure the infinite series converge. All of these conditions are extremely strong—even more than strong infinite confounding—because they either require assuming characteristics about the

10. In Supplement B.8 we provide an example where strong infinite confounding holds, but the subset deconfounder has bias that cannot be eliminated by adding more treatments.

treatment effects (when the very reason for estimation is that they are unknown) or assuming there is no confounding at all.

Takeaways In the linear-linear setting, the default full deconfounder is rank deficient because the substitute confounder is a linear projection of the treatments. There are three ways of addressing this issue: penalizing the outcome regression, sampling the substitute confounder from its posterior distribution and analyzing a subset of the causes. The penalized and posterior full deconfounders converge asymptotically in n and m under strong infinite confounding to the correct solution only by virtue of the fact that they contain the naïve regression. By contrast, the subset deconfounder requires strong and unverifiable assumptions about the values of the treatment effects or the nonexistence of confounding.

3.4 Extensions to Separable Nonlinear Settings

In this section, we extend our linear-linear results to settings with nonlinear factor and outcome regression models under separable confounding and strong infinite confounding. The perfect collinearity of the linear-linear case manifests as a general problem of a lack of overlap implied by the pinpointing assumptions (Imai and Jiang, 2019; D'Amour, 2019b). We set this point aside until Section **??** and focus on constant treatment effect setting.

Under constant treatment effects, we connect the deconfounder with a partially linear regression (Robinson, 1988) to demonstrate that a semi-parametric naïve regression approaches asymptotic unbiasedness for the same reason that the deconfounder does. Following Wang and Blei (2019, 2020) and Zhang et al. (2019), we study constant-effects outcome models of the form $\mathbb{E}[Y_i|\boldsymbol{A}_i, \boldsymbol{Z}_i] = \boldsymbol{A}_i^\top \boldsymbol{\beta} + g_Y(\boldsymbol{Z}_i)$. Theorem 1 shows any consistent deconfounder converges to a flexible naïve regression, which is also consistent.

Theorem 1. *(Deconfounder-Naïve Convergence under Strong Infinite Confounding.)*
Consider all data-generating processes in which (i) treatments and confounders are drawn from a factor model with continuous density $f(\boldsymbol{z}, \boldsymbol{a})$; (ii) $\boldsymbol{Z}$ is pinpointed, i.e. the conditional entropy $H(\boldsymbol{Z}_i|\boldsymbol{A}_i) = 0$; and (iii) the outcome model is of the form $\mathbb{E}[Y_i|\boldsymbol{A}_i, \boldsymbol{Z}_i] = \boldsymbol{A}_i^\top \boldsymbol{\beta} + g_Y(\boldsymbol{Z}_i)$, i.e. contains constant treatment effects and additively separable confounding. Any consistent deconfounder converges to a naïve estimator for any finite subset of treatment effects as n grows large.

The proof, given in Supplement B.9, is constructive: we provide a concrete naïve estimator, based on partially linear models and treatment-set partitioning, that is asymptotically equivalent to any consistent deconfounder. It proceeds by noting that these conditions imply that there are an infinite number of treatments and that the stochastic one-to-many function mapping $\boldsymbol{Z}_i$ to $\boldsymbol{A}_i$, $g_{\boldsymbol{A}}(\cdot)$, is asymptotically recoverable and invertible. This means that the deconfounder proceeds in three steps: (i) it estimates the factor model $\hat{g}_{\boldsymbol{A}}(\cdot)$, then uses this to obtain an estimate of the confounder $\hat{Z}_i = \hat{g}_{\boldsymbol{A}}^{-1}(\boldsymbol{A}_i)$; and (ii) it estimates and adjusts for a nonlinear function of $\hat{Z}_i$, $\hat{g}_Y^{\text{deconf}}(\cdot)$, while also estimating a m-dimensional vector of constant treatment effects, $\hat{\boldsymbol{\beta}}^{\text{deconf}}$. In other words, the nonlinear deconfounder

involves the following partially linear regression:

$$\left(\hat{\boldsymbol{\beta}}^{\text{deconf}}, \hat{g}_Y^{\text{deconf}}\right) = \underset{\boldsymbol{\beta}^*, g_Y^*}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{A}_i^\top \boldsymbol{\beta}^* - g_Y^*(\hat{g}_{\boldsymbol{A}}^{-1}(\boldsymbol{A}_i))\right)^2. \tag{7}$$

In general, if the factor-model function $g_{\boldsymbol{A}}^{-1}(\cdot)$ is unknown and must also be estimated from data, then the problem would become unidentifiable. This is because an arbitrarily flexible $g_Y^*({g_{\boldsymbol{A}}^*}^{-1}(\boldsymbol{A}_i))$ could not be distinguished from the constant treatment effects. That is, any part of $\boldsymbol{A}_i^\top \boldsymbol{\beta}^*$ could be shifted into this function while still achieving the same squared error. However, because the functional form of a nonlinear $g_{\boldsymbol{A}}^{-1}(\cdot)$ is specified in advance, then (7) is consistent for $\boldsymbol{\beta}$ (Robinson, 1988). Intuitively, this shows that what the deconfounder buys is the inner part of the function mapping the treatments to the outcome, $\hat{g}_Y(\hat{g}_{\boldsymbol{A}}^{-1}(\cdot))$, leaving only the outer $\hat{g}_Y(\cdot)$ to be estimated from data.

We will now show that instead of the deconfounder procedure described above, in which $\hat{\boldsymbol{Z}}$ is estimated with a factor model and then incorporated it into (7), we can instead directly fit a different partially linear regression that subsumes the entire $g_Y^*(\hat{g}_{\boldsymbol{A}}^{-1}(\boldsymbol{A}_i))$ term in (7). As in Section B.2, we partition the m causes, $\boldsymbol{A}_i$, into finite m_F focal causes of interest, $\boldsymbol{A}_{i,F}$, and m_N nonfocal causes, $\boldsymbol{A}_{i,N}$. We will show that for any choice of focal treatments, the estimates of $\boldsymbol{\beta}_F$ produced by this naïve procedure will be asymptotically equal to the corresponding deconfounder estimates, $\hat{\boldsymbol{\beta}}_F^{\text{deconf}}$. Then the conditional expectation function of the outcome can be rewritten $\mathbb{E}[Y_i|\boldsymbol{A}, \boldsymbol{Z}] = \boldsymbol{A}_{i,F}^\top \boldsymbol{\beta}_F + \boldsymbol{A}_{i,N}^\top \boldsymbol{\beta}_N + g_Y(\boldsymbol{Z}_i)$.

The semiparametric naïve regression proceeds in three steps: (i) it residualizes the focal treatments, $\boldsymbol{A}_{i,F}$, by removing any part that can be explained with the nonfocal treatments, $\boldsymbol{A}_{i,N}$; (ii) it similarly removes any part of the outcome Y_i that can be explained with the nonfocal treatments; and (iii) it conducts a regression of the residualized outcome, $\tilde{Y}$, on the residualized focal treatments, $\tilde{\boldsymbol{A}}_F$. Formally, it computes

$$\hat{h}_{\boldsymbol{A}_F}^{\text{naïve}} = \underset{h_{\boldsymbol{A}_F}^*}{\arg\min} \sum_{i=1}^{n} \left\|\boldsymbol{A}_{i,F} - h_{\boldsymbol{A}_F}^*(\boldsymbol{A}_{i,N})\right\|_F^2 \tag{8}$$

$$\hat{h}_Y^{\text{naïve}} = \underset{h_Y^*}{\arg\min} \sum_{i=1}^{n} \left(Y_i - h_Y^*(\boldsymbol{A}_{i,N})\right)^2 \tag{9}$$

and uses these to obtain $\tilde{\boldsymbol{A}}_F^{\text{naïve}} \equiv \boldsymbol{A}_{i,F} - \hat{h}_{\boldsymbol{A}_F}^{\text{naïve}}(\boldsymbol{A}_{i,N})$ and $\tilde{\boldsymbol{Y}}^{\text{naïve}} \equiv Y_i - \hat{h}_Y^{\text{naïve}}(\boldsymbol{A}_{i,N})$. Finally, treatment-effect estimates are obtained by

$$\hat{\boldsymbol{\beta}}_F^{\text{naïve}} = \left(\tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{A}}_F^{\text{naïve}}\right)^{-1} \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{Y}}^{\text{naïve}} \tag{10}$$

This procedure, like other "naïve" approaches, estimates $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{A}]$ directly, rather than first estimating $\boldsymbol{Z}$ and then attempting to utilize it.

These two approaches asymptotically converge to one another. In short, the key problem is estimating the composite function $g_Y(\hat{g}_{\boldsymbol{A}}^{-1}(\cdot))$. This can be done directly by a flexible naïve regression without the added complication or factor-model functional form assumptions of the deconfounder. If an analyst knows the precise functional form of the factor model, then the incorporation of this information into the deconfounding procedure is likely to improve efficiency; conversely, misspecified factor models are likely to distort inferences in unpredictable ways. It is important to note that when this information is available, it can also be used to improve the efficiency of the naïve estimator, as we show in Section 4.2.

4. Deconfounder Does Not Consistently Outperform Naïve Regression in Finite Samples

Having established that the deconfounder offers no gains over naïve regression in asymptotic bias, we now reconsider the simulation evidence for finite sample performance. Results demonstrate the deconfounder cannot in general improve over naïve regression. We note that these findings conflict with the positive simulation evidence presented in the deconfounder papers. The divergence in our findings largely stems from (i) improvements that we make in estimation, including substantial gains in stability, and (ii) our extension of simulations to more thoroughly probe changes in key parameters. Supplement C.1 provides a thorough discussion of these and other deviations. This section concludes with a brief overview of additional simulation evidence in the supplement.

4.1 Linear-Linear Deconfounders Only Help When Biases Cancel

In the linear-linear setting, the substitute confounder is a linear function of the treatments—information already captured by including the treatments in the linear outcome model. We show that in the linear-linear setting, deconfounder can sometimes outperform the naïve regression in subsets of the parameter space where differing naïve and deconfounder biases align in the right way. However, these situations always rely on parameters that would be unknown to the analyst. We also show that given estimation instability in near-collinear estimators, the full deconfounder with a linear factor model is never appropriate.

4.1.1 Medical Deconfounder

Zhang et al. (2019) presents an application of the deconfounder to the analysis of electronic health records. The first simulation study presented in the paper considers a situation where there are two treatments, of which only one has a true non-zero coefficient.The true data generating process draws $n = 1,000$ patients from a linear-linear model.[11] They estimate a one-dimensional substitute confounder using probabilistic principal component analysis.[12] We introduce a faster and more accurate variant deconfounder which performs PCA, extracts the top component, and then runs ridge regression with a penalty chosen by cross-validation. Zhang et al. (2019) report a single sample. We repeat this process 1,000 times, assessing bias, variance and root mean squared error (RMSE) in Table 1.

The original deconfounder performs poorly, with higher bias and variance than the naïve estimator due to near collinearity driving estimation instability. Our PCA+CV-Ridge deconfounder appears to perform better than naïve, but this does not hold across the parameter space. Under the original data generating process, the effect of A_1 is zero and therefore, the ridge penalty drives the coefficient of A_1 towards the truth. This results in apparent good performance for the simplified deconfounder variant. In the last two rows of Table 1, we repeat the same simulation switching the true effect of A_1 to -0.3: the simplified deconfounder now performs slightly worse than the naïve regression.

11. This process is $Z_i \sim \mathcal{N}(0,1)$, $A_{i,1} \sim \mathcal{N}(0.3Z_i, 1)$, $A_{i,2} \sim \mathcal{N}(.4Z_i, 1)$, $Y_i \sim \mathcal{N}(0.5Z_i + 0.3A_{i,2}, 1)$.
12. Zhang et al. (2019) uses black box variational inference (Ranganath et al., 2014), then estimates the outcome model with automatic differentiation variational inference (Kucukelbir et al., 2017). We use `Stan` code for probabilistic PCA and the outcome model. Further details are in Supplement C.2.1.

		Bias		Std. Dev.		RMSE	
	Model	β_1	β_2	β_1	β_2	β_1	β_2
	Naïve	0.120	0.160	0.033	0.033	0.125	0.164
Orig. simulation	Oracle	0.000	0.000	0.032	0.033	0.032	0.033
($\beta_1 = 0$, $\beta_2 = 0.3$)	Deconfounder	0.145	0.189	0.675	0.877	0.690	0.897
	PCA+CV-Ridge	0.028	-0.146	0.014	0.029	0.031	0.149
Our simulation	Deconfounder	0.150	0.197	0.685	0.893	0.701	0.914
($\beta_1 = -0.3$, $\beta_2 = 0.3$)	PCA+CV-Ridge	0.188	-0.099	0.028	0.037	0.191	0.106

Table 1: **Simulation Study 1 of the Medical Deconfounder**. The main "Deconfounder" estimation procedure is from Zhang et al. (2019) and uses probabilistic PCA and Bayesian linear regression. "PCA + CV-Ridge" is an improved deconfounder estimator we developed. "Naïve"and "Oracle" estimators are as described in the main text.

4.1.2 Subset Deconfounder

Proposition 4 shows strong infinite confounding is insufficient for the subset deconfounder to provide unbiased estimates of treatment effects, even when naïve regression can achieve oracle-like performance. Only under strong assumptions about treatment effects will the subset deconfounder be unbiased. We design a simulation to demonstrate this fact for different sequences of treatment effects, β, even when strong infinite confounding is satisfied ($\theta_j = 10 \; \forall \; j$). The full simulation is included in Supplement C.5.

Table 2 provides the average RMSE for treatment effect estimates. When treatment effects are constant ($\beta_j = 10$ or $\beta_j = 100$) the subset deconfounder's performance fails to improve as more treatments are added. This is true even though naïve regression's average RMSE converges on the oracle's performance. Similarly, we see that when $\beta_j \sim \mathcal{N}(1, 2)$, the subset deconfounder converges on an average bias of 1—the mean of β_j (see Case ii from Proposition 4). In Table 2 the subset deconfounder is unbiased only when the sequence of treatment effects converge to 0 as more treatments are added (e.g., if $\beta_j = \frac{1}{j}$).

Takeaways In finite-sample linear-linear settings, the deconfounder cannot improve performance over naïve regression. Due to estimation instability, variants of the linear factor full deconfounder are never preferable to naïve regressions. Subset deconfounders only outperform naïve regressions under strong assumptions about treatment effects.

4.2 Nonlinear Deconfounder and Naïve Approaches Can Sometimes Exploit Parametric Information

The best-case scenario for the deconfounder exploits known parametric information about a nonlinear data generating process. We examine one such case, drawing on a simulation posted on the `blei-lab` GitHub page (Wang, 2019). Even in this ideal setting, the deconfounder is outperformed by a correctly specified nonlinear naïve regression.

	Method	m=3	m=10	m=50	m=100	m=200
$\beta_j = 10$	Oracle	0.010	0.010	0.010	0.010	0.010
	Naïve	0.333	0.100	0.022	0.014	0.011
	Deconfounder	10.000	10.000	10.000	10.000	10.000
$\beta_j = 100$	Oracle	0.010	0.010	0.010	0.010	0.010
	Naïve	0.333	0.100	0.022	0.014	0.011
	Deconfounder	100.000	100.000	100.000	100.000	100.000
$\beta_j \sim \mathcal{N}(1, 2)$	Oracle	0.010	0.010	0.010	0.010	0.010
	Naïve	0.333	0.101	0.022	0.014	0.011
	Deconfounder	1.465	1.283	1.362	1.195	1.026
$\beta_j = \frac{1}{m}$	Oracle	0.011	0.010	0.010	0.010	0.010
	Naïve	0.333	0.101	0.022	0.014	0.011
	Deconfounder	0.611	0.293	0.091	0.054	0.033

Table 2: **The Subset Deconfounder is Unbiased Only under Strong Assumptions about Treatment Effects** As the number of treatments grow (columns moving from left to right), both the naïve regression converges on the oracle's average RMSE (entries in table average over RMSE of each treatment effect), while the subset deconfounder's performance depends on the treatment's effects. The subset deconfounder is unbiased only when the sequence of treatments converges to zero. Even when the treatments are random draws from a normal distribution, the bias of the subset deconfounder converges on the average treatment effect.

We simulate $n = 10,000$ draws from the following data-generating process,

$$\begin{bmatrix} A_{i,1} \\ A_{i,2} \\ Z_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \right)$$

$$Y_i \sim \mathcal{N}(0.4 + 0.2A_{i,1}^2 + 1A_{i,2}^2 + 0.9Z_i^2,\ 1) \tag{11}$$

for $\rho = 0.4$. As before, these are collected in $\boldsymbol{Z}$, $\boldsymbol{A}$, and $\boldsymbol{Y}$. A substitute confounder is obtained by taking the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ and extracting the first component, $\hat{\boldsymbol{Z}} \equiv \boldsymbol{A}\boldsymbol{V}_1$, so for $\hat{Z}_i = \frac{\sqrt{n}}{D_1}(V_{11}A_{i1} + V_{21}A_{i2})$. Following Wang (2019), we then estimate a linear regression of $\boldsymbol{Y}$ on three predictors: $\boldsymbol{A}_1^2$, $\boldsymbol{A}_2^2$, and $\hat{\boldsymbol{Z}}^2$.

The deconfounder captures the interaction of the treatments, as can be seen by expanding the polynomial $\hat{Z}_i^2 = \left(\frac{\sqrt{n}}{D_1}(V_{11}A_{i1} + V_{21}A_{i2})\right)^2 = \frac{n}{D_1^2}\left(V_{11}^2A_{i1}^2 + V_{21}^2A_{i2}^2 + V_{11}V_{21}A_{i1}A_{i2}\right)$. The expansion is not a linear combination of $A_{i,1}^2$ and $A_{i,2}^2$ due to the inclusion of the interaction $A_{i,1}A_{i,2}$. However, the deconfounder only incorporates partial information about the true functional form of the outcome model, Equation (11). By using the same information more carefully, a better parametric naïve estimator can be derived. By properties of the multivariate normal distribution, $\frac{1}{\rho^2}\mathbb{E}[Z_i|\boldsymbol{A}_i]^2 = A_{i,1}^2 + A_{i,2}^2 + 2A_{i,1}A_{i,2}$. Therefore, the causal effects can also be estimated by a regression of $\boldsymbol{Y}$ on $\boldsymbol{A}_1^2$, $\boldsymbol{A}_2^2$, and $(\boldsymbol{A}_1^2 + \boldsymbol{A}_2^2 + 2\boldsymbol{A}_1 \circ \boldsymbol{A}_2)$, where $\circ$ denotes the elementwise product. We refer to this approach as the "parametric" alternative; it is also a naïve regression, as it does not seek to estimate $\boldsymbol{Z}$. As Figure 1 (left panel) shows, this substantially improves over the deconfounder for negative ρ, in terms

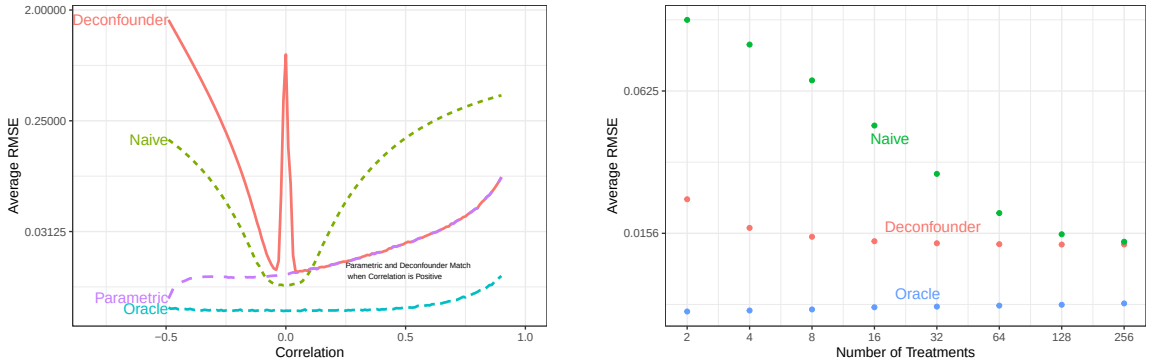

Figure 1: **Under Strong Infinite Confounding, a Parametric Naïve Model Outperforms the Deconfounder Over the Parameter Space (left); Naïve Regression Converges to the Same Performance as the Number of Treatments Increases (right).** The left plot shows average RMSE for varying ρ, the off-diagonal covariances in Equation (11). The right plot shows as m increases, deconfounder and naïve RMSEs converge.

of average root mean squared error, while capturing the same information for positive ρ. Moreover, when the latent confounder satisfies strong infinite confounding, the naïve regression will approach the deconfounder's performance as the number of treatments grows. The right panel demonstrates this point for the original setting of $\rho = 0.4$.

While the quadratic example shows it may be theoretically possible to design a simulation where nonlinear information helps the deconfounder improve over an incorrectly specified naïve regression, it is difficult. In this tutorial simulation, the factor model performs poorly for negative correlations while the parametric model does well. In more complex simulations involving factor model nonlinearity, reported gains are often modest. Wang and Blei (2019) presents a simulation based on Genome Wide Association Study (GWAS) data which show a maximum RMSE reduction of merely 3% for the deconfounder over the naïve. (In our replication in Supplement C.4, we show that in fact, the deconfounder actually performs worse than naïve regression for the non-zero coefficients—an unfortunate pattern for applied genetics, where the primary interest is detecting nonzero coefficients and assessing their size.) Similarly, a second simulation study in Zhang et al. (2019) report deconfounder RMSE that is only slightly better than naïve RMSE (our simulations point to the opposite conclusion, that the deconfounder does slightly *worse*; see Supplement C.2.2).

Takeaways If the deconfounder learns the correct nonlinearity, it is possible to improve over incorrectly specified naïve regression in finite samples. However, in practice, making use of this fact is impossible without extensive knowledge of the data generating process. Moreover, even the deconfounder papers show, at best, marginal improvements. When such extensive knowledge is available, a well-specified naïve regression making use of that knowledge will also perform well. We conjecture that if the high-dimensional treatments lie on a low-dimensional manifold *and* the correct factor model specification is known, it might be more efficient to model the relationship between $\hat{\boldsymbol{Z}}$ and $\boldsymbol{Y}$ semiparametrically (as in the

deconfounder) rather than directly modeling high-dimensional $\boldsymbol{A}$ and $\boldsymbol{Y}$ semiparametrically (as in naïve regressions). However, this has yet to be demonstrated in any simulation.

4.3 Takeaways and Additional Empirical Results

We have replicated every simulation across the deconfounder papers for which data is available, and we find no evidence that the deconfounder consistently outperforms the naïve regression. Several simulations—like the medical deconfounder and quadratic examples highlighted above—perform better at some parameter values but worse at others. Full details of all six replicated simulations are in the supplement.

Separately, Wang and Blei (2019) use posterior predictive checks (PPCs) of the factor model, arguing these will assess when the deconfounder can improve estimates. If this claim were true, it would allow highly flexible density estimation to be used, even when the true parametric form of the factor model was unknown—as is always the case in practice. However, it is not. Theoretically, Proposition 4 proves that for subset deconfounders, this is impossible because the performance depends on untestable assumptions about the treatment effects, not the factor model. And for the full deconfounder, PPCs are ill-suited to evaluating conditional independence of $\boldsymbol{A}$ given $\hat{\boldsymbol{Z}}$, perhaps the most relevant observable property of the factor model (Imai and Jiang, 2019). Empirically, we further present a new simulation in Supplement F demonstrating that the PPC does not reliably indicate whether a deconfounder will perform well, either in absolute terms or relative to naïve regression.

Our asymptotic results suggested that the deconfounder would not outperform the naïve regression, and simulations have shown this to hold in finite samples. In nonlinear settings it is possible to exploit parametric information in the factor model, but it is both difficult to do in practice and can also be used to comparably perform the naïve regression. It remains possible that a simulation could establish a particular data generating process where the deconfounder performs better than naïve regression, but this has yet to be demonstrated.

5. Neither Naïve Regression nor Deconfounder is Currently Suitable For Real World Applications

After deriving the deconfounder's properties, WB recommend it for empirical research in in social science, neuroscience, and medicine. Zhang et al. (2019) propose the deconfounder as a solution for assessing drug treatment effects using observational health records, writing that the deconfounder is "guaranteed to capture all multi-medication confounders, both observed and unobserved" (Zhang et al., 2019, p. 2).

As we have shown, however, this is not a property of the deconfounder. In practice, both the deconfounder and naïve regression will fail to capture confounders unless many other strong and unverifiable assumptions are satisfied. Therefore, we cannot recommend either current approach for use in real-world applications. This is true even though we have shown that (i) under the deconfounder's assumptions, the naïve estimator is asymptotically unbiased across many settings given strong infinite confounding; and (ii) the full deconfounder inherits this property because it includes the same information as naïve regression. We emphasize that the required assumptions are *exceedingly strong*. To demonstrate the consequences of their violation, we now investigate the real-world case study in WB—of actors' effects on box office revenue—and show that both produce implausible estimates.

We then highlight some of the explicit and implicit assumptions of the deconfounder (and naïve regression) which lead us to be skeptical that credible applications can be found.

5.1 Actor Case Study Reveals Limitations of the Deconfounder

WB's case study investigates how the cast of a movie causally affects movie revenue. The deconfounder is applied to the TMDB 5,000 data set, estimating how much each of the $m = 901$ actors affected the revenue of $n = 2,828$ movies. WB presents results from a full deconfounder in which substitute confounders are estimated using the leading $k = 50$ dimensions of a poisson matrix factorization (PMF) of the binary matrix of movie-actor appearance indicators. A linear regression of log revenue on actor appearance and substitute confounders is used to estimate what is described as the causal effect of the cast.

We replicate this analysis in Table 3, using cached PMF output from WB to ensure that our conclusions are unaffected by random seed. The five largest estimates from this model as well as several alternatives are reported in Table 3 (expanded results, including appearance-weighted log-scale coefficients,[13] are in Supplement G).

According to WB's results, the single most valuable actor is Stan Lee—whose appearances are cameos in movies based primarily on his Marvel comic books[14] totaling 200 seconds of screen time. These unreported estimates suggest that with his casting, Marvel Cinematic Universe (MCU) producers causally increased their movies' revenue by 831%—more than nine times the box-office haul of their counterfactual Stan-less versions, a total of $15.5 billion in additional earnings. The subset deconfounder's estimates are similarly implausible, suggesting that Jess Harnell causally increases a movie's revenue by 1,128%. This is driven by his appearance as a voice actor in the high-budget "Transformers" series, as his credits are otherwise in peripheral roles not included in this data set, such as a supporting role in the animated series *Doc McStuffins*. WB's subset deconfounder suggests that his appearances collectively increased revenue by $2.5 billion.

The deconfounder produces implausible estimates because it fails to capture important multi-cause confounders. This is clearest when we explicitly adjust for a movie's budget—the quintessential multi-cause confounder, enabling the casting of big-name stars and also reflecting the studio's underlying belief in the viability of the film.[15] This simple adjustment produces dramatically different assessments of actor value that are far more reasonable in scale, though likely still overstated. The deconfounder claims to capture all multi-cause confounding—not only from budget but also genre, series, directors, writers, language, and release season. WB explicitly argue that including observed covariates with the deconfounder is not necessary—yet this example shows that it is.

13. WB rank actors by multiplying each actor's log-scale coefficients by their number of movie appearances. This transformation is difficult to interpret substantively but produces similar results.
14. And a 40 second appearance in "Mallrats."
15. In Wang and Blei (2019) and replication code generously shared with us, actor analyses did not condition on any observed covariates. After we shared our draft with Wang and Blei in July 2020, a reference implementation conditioning on budget and runtime was posted.

Table 3: **The Deconfounder Estimates Implausible Effects for Actors.** Estimated causal effect of each actor's casting on movie revenue. Following WB, estimates are computed by linear regression of log revenue on actor indicators and additional covariates. Each row reports a different specification (for example, "deconfounder" rows each adjust for a 50-dimensional substitute confounder, and the "controls" row adjusts for budget, a multi-cause confounder). The top panel contains estimators that analyze all actors simultaneously, including the full deconfounder; the bottom panel contains estimators that analyze each actor j in isolation, including the subset deconfounder and the univariate naïve estimator $\hat{\beta}_j = \text{Cov}(\boldsymbol{A}_j, \boldsymbol{Y})/\text{Var}(\boldsymbol{A}_j)$. Two versions of each deconfounder estimator are used, one relying on a cached poisson matrix factorization (PMF) provided by WB and another using a re-estimated PMF. For each estimator, the top five actors and associated estimates are presented in the form "Actor ($\times e^{\hat{\beta}_j}$)," indicating an estimate that Actor causally modifies revenue by a multiplicative factor of $e^{\hat{\beta}_j}$.

Estimating all actor effects simultaneously (full deconfounder)		
Naïve	Standard	Stan Lee (×9.31), John Ratzenberger (×9.26), Sacha Baron Cohen (×7.09), Leonardo DiCaprio (×5.50), Josh Hutcherson (×5.19)
Deconfounder	Cached PMF	Stan Lee (×9.29), John Ratzenberger (×8.29), Sacha Baron Cohen (×8.12), Josh Hutcherson (×5.02), Corey Burton (×4.91)
Deconfounder	Rerun PMF	Courteney Cox (×15.32), Tom Cruise (×14.74), John Ratzenberger (×11.13), Vera Farmiga (×9.86), Sacha Baron Cohen (×9.83)
Controls	Adjusting for budget	Sacha Baron Cohen (×6.93), Brian Doyle Murray (×4.08), Conrad Vernon (×4.04), Julie Andrews (×3.84), Tomas Arana (×3.83)
Estimating effects one actor at a time (subset deconfounder)		
Naïve	Univariate	Jess Harnell (×12.28), Ava Acres (×10.16), Warwick Davis (×10.09), Stan Lee (×9.85), Orlando Bloom (×9.50)
Deconfounder	Cached PMF	Jess Harnell (×13.49), Ava Acres (×10.49), Chris Miller (×9.19), Orlando Bloom (×9.02), Stan Lee (×8.77)
Deconfounder	Rerun PMF	Lasco Atkins (×5.64), Sacha Baron Cohen (×4.24), John Ratzenberger (×4.01), Desmond Llewelyn (×3.91), Will Smith (×3.64)

5.2 Strong Assumptions Rule Out Other Applications

The deconfounder's exceedingly strong assumptions often make it unsuitable for many uses. Although there are other embedded assumptions—see Ogburn et al. (2019) for more—we focus on four. First, the deconfounder requires that treatments arise from a factor model with a low-dimensional confounder. Practically speaking, analysts must know enough about the functional form of this factor model to feasibly estimate it. Second, the deconfounder requires treatments to be independently drawn given $\boldsymbol{Z}$, ruling out settings where treatments cause other treatments. This implies that casting one actor cannot influence whether another actor is cast later. This alone excludes many realistic settings—except perhaps genetics, where many of these ideas originated. Third, pinpointing confounders with a factor model requires strong infinite confounding. In practice, this means analysts must record a very large number of treatments, which are contaminated by comparatively few confounders. In the actor setting, this would be violated by producers who regularly work with the same sets of actors. Furthermore, the mere fact that all movies have finite casts implies limits to the information learned about confounding. Fourth, even when a pinpointable factor model of the proper class exists, parametric assumptions such as separable confounding or constant treatment effects are used in many proofs, both here and in WB. Often these conditions help to address failures of positivity by leveraging functional form assumptions. Yet, particularly in social and medical problems, causal heterogeneity is the rule, not the exception.

It is unknown how sensitive the naïve and deconfounder families of methods are to slight violations of these assumptions. Until there is a way to relax these assumptions or otherwise evaluate the severity of the consequences of violating them, we cannot recommend either the naïve regression or the deconfounder for real-world applications.

5.3 Takeaways

Assumptions used to prove deconfounder properties, like pinpointing, are extremely strong and unlikely to hold in real applications. While analysts cannot know whether the deconfounder estimates are accurate, results from the actor case study are highly implausible. Given the high-stakes nature of many proposed applications, we think a great deal more evidence is warranted before these methods are put into practice.

6. Discussion

A widespread practice across several fields of inqury examines the ability of the latent structure of treatments to adjust for shared confounding. To assess these deconfounder methods, we have re-examined the theory for every variant deconfounder estimator, as well as every empirical application and simulation for which data are available. We prove new results showing that for any finite m, the deconfounder is inconsistent. As $m \to \infty$, under strong infinite confounding, the naïve regression and full deconfounder both approach asymptotic unbiasedness, but the subset deconfounder requires strong additional assumptions. We also examined finite-sample properties through simulation, finding no evidence that the deconfounder systematically outperforms the naïve regression—or that analysts could possibly identify when it might. Finally, we show that the deconfounder's estimates in existing

real-world case studies are not credible and highlight the strong assumptions embedded in the deconfounder framework. In every simulation and empirical study in Wang and Blei (2019), Zhang et al. (2019) and Wang (2019) for which data was available, our replications show—as predicted by our theory—that no deconfounder consistently improves over the naïve regression across the parameter space.

Every estimator considered in this paper without access to the latent confounders is inconsistent for finite m. The deconfounder's pinpointing assumption requires *strong infinite confounding*—an asymptotic regime for m that is a very stringent assumption. For the subset deconfounder, *strong infinite confounding* is insufficient and requires further strong assumptions about the treatment effects and/or confounding. When the deconfounder does work (is estimable and satisfies conditions for asymptotic unbiasedness), we prove that a suitably flexible naïve regression converges to the deconfounder asymptotically in m. Thus the deconfounder works in limited settings. When the deconfounder works, the factor-model machinery of the deconfounder is unnecessary, because a naïve regression asymptotically produces the same result.

We note that all theory in this paper is in an asymptotic regime where $n \to \infty$ for each treatment. This is helpful for clarifying the strong assumptions necessary for the deconfounder and naïve regression to hold, but because n does not grow as a function of m, empirical practice in high-dimensional settings likely requires even stronger assumptions.

Collectively, our findings suggest that if assumptions hold, there is no reason to prefer the deconfounder to the naïve regression. However, we ultimately think that the strength of the assumptions is such that neither method should be used in practice.

References

Peter Bühlmann and Domagoj Ćevid. Deconfounding and causal regularisation for stability and external validity. *International Statistical Review*, 88:S114–S134, 2020.

Domagoj Ćevid, Peter Bühlmann, and Nicolai Meinshausen. Spectral deconfounding via perturbed sparse linear models. *The Journal of Machine Learning Research*, 21(1):9442–9482, 2020.

Victor Chernozhukov, Christian Hansen, and Yuan Liao. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76, 2017.

Alexander D'Amour. Comment: Reflections on the deconfounder. *Journal of the American Statistical Association*, 114(528):1597–1601, 2019a.

Alexander D'Amour. On multi-cause approaches to causal inference with unobserved counfounding: Two cautionary failure cases and a promising alternative. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3478–3486, 2019b.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Zijian Guo, Domagoj Ćevid, and Peter Bühlmann. Doubly debiased lasso: High-dimensional inference under hidden confounding. *Annals of statistics*, 50(3):1320, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2013.

Kosuke Imai and Zhichao Jiang. Comment: The challenges of multiple causes. *Journal of the American Statistical Association*, 114(528):1605–1610, 2019. doi: 10.1080/01621459.2019.1689137.

Kosuke Imai and David A Van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99 (467):854–866, 2004.

Elizabeth Johnson, Francesca Dominici, Michael Griswold, and Scott L Zeger. Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey. *Journal of Econometrics*, 112(1):135–151, 2003.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(1):430–474, 2017.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Wang Miao, Wenjie Hu, Elizabeth L Ogburn, and Xiao-Hua Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, pages 1–15, 2022.

Elizabeth L Ogburn, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Comment on "blessings of multiple causes". *Journal of the American Statistical Association*, 114(528):1611–1615, 2019.

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

Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771, 2015.

P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988. ISSN 00129682, 14680262.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Yixin Wang. The blessings of multiple causes: A tutorial. GitHub Repository, August 2019. Commit 7ba553b5d669bb90b3cdd7cf2be42e686328dce1.

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

Yixin Wang and David M Blei. Towards clarifying the theory of the deconfounder. *arXiv preprint arXiv:2003.04948*, 2020.

Linying Zhang, Yixin Wang, Anna Ostropolets, Jami J. Mulgrave, David M. Blei, and George Hripcsak. The medical deconfounder: Assessing treatment effects with electronic health records. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 490–512. PMLR, 09–10 Aug 2019.

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *arXiv preprint arXiv:2102.09412*, 2021.

Supplemental Information

Appendix A. Notation

Table 4: **Notation reference.** Table of notation for indices, parameters, random variables, and estimators used throughout the manuscript (continued on following page).

Indices:	
n	number of observations
k	dimensionality of latent confounders
m	dimensionality of treatments
m_F, m_N	dimensionality of focal and non-focal treatments, respectively
Random variables:	
$\underset{n\times k}{\boldsymbol{Z}}$	unobserved confounders
$\underset{n\times m}{\boldsymbol{\nu}}$	random component of treatments
$\underset{n\times m}{\boldsymbol{A}}$	observed treatment
$\underset{n\times m_F}{\boldsymbol{A}_F}$, $\underset{n\times m_N}{\boldsymbol{A}_N}$	observed subsets of focal and non-focal treatments
$\underset{n\times 1}{\boldsymbol{\epsilon}}$	random component of outcome
$\underset{n\times 1}{\boldsymbol{Y}}$	observed outcome
Parameters:	
$\boldsymbol{\theta}$	coefficients linearly mapping confounders to treatments
$\boldsymbol{\gamma}$	coefficients linearly mapping confounders to outcome
$\boldsymbol{\beta}$	coefficients linearly mapping treatments to outcome
$\boldsymbol{\beta}_F, \boldsymbol{\beta}_N$	subset of $\boldsymbol{\beta}$ corresponding to focal and non-focal treatments
σ^2	conditional variance of treatments
ω^2	conditional variance of outcome

Table 5: **Notation reference (continued).** Table of notation for indices, parameters, random variables, and estimators used throughout the manuscript (continued from previous page).

Estimators:	
$\hat{\boldsymbol{Z}}$	substitute confounder estimated by factor modeling of treatments
$\hat{\boldsymbol{\theta}}$	implicitly learned mapping from substitute confounder to outcome
$\hat{\boldsymbol{\beta}}^{\text{oracle}}$	infeasible oracle regression estimator using unobserved latent confounder
$\hat{\boldsymbol{\beta}}^{\text{full}}$	infeasible full deconfounder estimator using collinear substitute confounder
$\hat{\boldsymbol{\beta}}^{\text{subset}}$	subset deconfounder estimator using substitute confounder and focal treatments only
$\hat{\boldsymbol{\beta}}^{\text{naïve}}$	naïve regression of outcome on treatments, ignoring confounding
$\hat{\boldsymbol{\beta}}^{\text{penalty}}$	ridge deconfounder estimator using collinear substitute confounder and penalization
$\hat{\boldsymbol{\beta}}^{\text{nonlinear}}$	nonlinear ridge deconfounder estimator using orthogonal polynomials of substitute confounder and penalization
$\hat{\boldsymbol{\beta}}^{\text{wn}}$	white-noise deconfounder estimator using collinear substitute confounder with generated noise
$\hat{\boldsymbol{\beta}}^{\text{pm}}$	posterior-mean deconfounder estimator using probabilistic principal components
$\boldsymbol{S}$	generated noise (posterior noise) used in white-noise (posterior-mean) deconfounder
$\boldsymbol{M}^*$	annihilator matrix corresponding to any regression estimator $*$
Derived variables:	
$\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$	output of singular-value decomposition of $\boldsymbol{A}$
$\boldsymbol{Q}, \boldsymbol{\Lambda}$	output of eigendecomposition $\boldsymbol{\theta}^\top\boldsymbol{\theta} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$

Appendix B. Derivations

B.1 Minor Asymptotic Results

In this section, we collect a number of minor results that will be useful in bias derivations. We begin by reviewing properties of probabilistic principal component analysis, then present lemmas relating to the asymptotic behavior of the deconfounder and naïve estimators.

B.1.1 Properties of Probabilistic Principal Component Analysis

For convenience, we review properties of probabilistic principal component analysis used in the remainder of this section. The generative model is

$$\begin{aligned}\boldsymbol{Z} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\\ \boldsymbol{A} &\sim \mathcal{N}(\boldsymbol{Z\theta}, \sigma^2\boldsymbol{I}).\end{aligned}$$

For compactness, we use $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ to denote independently sampling of $\boldsymbol{Z}_i$ from a normal distribution centered on the i-th row of the mean matrix and the given covariance matrix, then collecting samples in $\boldsymbol{Z}$. Tipping and Bishop (1999) show that this data-generating process implies

$$(\boldsymbol{Z}|\boldsymbol{A}) \sim \mathcal{N}\left(\boldsymbol{A\theta}^\top\left(\boldsymbol{\theta\theta}^\top + \sigma^2\boldsymbol{I}\right)^{-1},\ \sigma^2\left(\boldsymbol{\theta\theta}^\top + \sigma^2\boldsymbol{I}\right)^{-1}\right). \tag{12}$$

We now examine asymptotic relationships between the singular value decomposition $\boldsymbol{A} = \boldsymbol{UDV}^\top$ and the eigendecomposition $\boldsymbol{\theta}^\top\boldsymbol{\theta} = \boldsymbol{Q\Lambda Q}^\top = \boldsymbol{Q}_{1:k}\boldsymbol{\Lambda}_{1:k}\boldsymbol{Q}_{1:k}^\top$; note that the trailing $m-k$ eigenvalues are zero. (Subscripts of the form $\boldsymbol{X}_{i:j}$ generally indicate column subsets of matrix $\boldsymbol{X}$ from column i to column j, except when indexing diagonal matrices where they indicate the corresponding diagonal element.)

$$\begin{aligned}p\text{-}\lim_{n\to\infty}\frac{1}{\sqrt{n}}\boldsymbol{D}_{1:k} &= \left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}}\\ p\text{-}\lim_{n\to\infty}\frac{1}{\sqrt{n}}\boldsymbol{D}_{(k+1):m} &= \sigma\boldsymbol{I}\\ p\text{-}\lim_{n\to\infty}\boldsymbol{V}_{1:k} &= \boldsymbol{Q}_{1:k}\\ p\text{-}\lim_{n\to\infty}\boldsymbol{V}_{(k+1):m}\boldsymbol{V}_{(k+1):m}^\top &= \boldsymbol{Q}_{(k+1):m}\boldsymbol{Q}_{(k+1):m}^\top\end{aligned}$$

where the last equality follows from $p\text{-}\lim_{n\to\infty}\boldsymbol{VV}^\top = \boldsymbol{QQ}^\top = \boldsymbol{I}$.

B.1.2 Consistency and Inconsistency Results for the Deconfounder

In this section, we present minor results relating to the consistency of the deconfounder. Consider n observations drawn from a data-generating process with k unobserved confounders, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $m \geq k$ observed treatments, $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{Z\theta}, \sigma^2\boldsymbol{I})$.

Definition 2. *(Pinpointedness of the substitute confounder.) A substitute confounder, $\boldsymbol{Z}$ is said to be pinpointed if its posterior distribution, $f(\boldsymbol{z}|\boldsymbol{A})$, collapses to a Dirac delta, $\delta(g(\boldsymbol{Z}))$, where $g(\boldsymbol{Z})$ is a bijective transformation of $\boldsymbol{Z}$.*

Specifically, pinpointedness (Wang and Blei, 2019, previously referred to as "consistency of the substitute confounder") does not require convergence of $\hat{\boldsymbol{Z}}$ to $\boldsymbol{Z}$, as consistency; for example, convergence to a rotation or rescaling will suffice. Equivalently, pinpointing requires that the conditional entropy $H(\boldsymbol{Z}|\boldsymbol{A}) = 0$. Below, we show pinpointing requires an infinite number of stochastic causes for smooth factor models.

Lemma 1. *In linear factor models, strong infinite confounding is necessary for $\hat{\boldsymbol{Z}}$ to asymptotically pinpoint $\boldsymbol{Z}$ as the number of causes goes to infinity.*

Proof of Lemma 1. Tipping and Bishop (1999) show that under the probabilistic principal components analysis model (see Supplement B.1.1), the posterior of the confounder, $f(\boldsymbol{z}|\boldsymbol{A})$, follows (12). The substitute confounder is a summary statistic, such as the mode, of this posterior. We examine the best-case scenario in which $\boldsymbol{\theta}$ and σ^2 are known. In this setting, the posterior variance of the k'-th confounder is $\text{Var}\left(Z_{i,k'}|\boldsymbol{A}_i, \boldsymbol{\theta}, \sigma^2\right) = \sigma^2\left(\boldsymbol{\theta}_{k'}\boldsymbol{\theta}_{k'}^\top + \sigma^2\right)^{-1}$. Because we assume $\sigma^2 > 0$, if the variance goes to zero, which pinpointing implies, then each $\boldsymbol{\theta}_{k'}\boldsymbol{\theta}_{k'}$, the k'-th diagonal element of $\boldsymbol{\theta}\boldsymbol{\theta}^\top$, must go to infinity. (We rule out the case of $\sigma^2 = 0$, by the assumption that the causes are a nondeterministic function of the latent confounders.) Thus, pinpointing of $\boldsymbol{Z}_i$ implies strong infinite confounding. □

Lemma 2 states that the infinite-m requirement generalizes to all factor models with continuous density, i.e. models with continuous $f(\boldsymbol{z}_i)$ and continuous $f(\boldsymbol{a}_i|\boldsymbol{z}_i) = \prod_{j=1}^m f(\boldsymbol{a}_{i,j}|\boldsymbol{z}_i)$.

Lemma 2. *In continuous factor models, an infinite number of causes are necessary for $\hat{\boldsymbol{Z}}$ to pinpoint $\boldsymbol{Z}$.*

Proof of Lemma 2. Pinpointing of $\boldsymbol{Z}$ is equivalent to the statement that $H(\boldsymbol{Z}|\boldsymbol{A}) = \iint f(\boldsymbol{z},\boldsymbol{a}) \log \frac{f(\boldsymbol{z},\boldsymbol{a})}{f(\boldsymbol{a})} \,\mathrm{d}\boldsymbol{z}\,\mathrm{d}\boldsymbol{a} = 0$. Consider all $\boldsymbol{a}$ and $\boldsymbol{z}$ with $f(\boldsymbol{z},\boldsymbol{a}) > 0$. Pinpointing requires $f(\boldsymbol{a}|\boldsymbol{z})f(\boldsymbol{z}) = \int f(\boldsymbol{a}|\boldsymbol{z}^*)f(\boldsymbol{z}^*)\,\mathrm{d}\boldsymbol{z}^*$. In other words, $f(\boldsymbol{a}|\boldsymbol{z}^*) = 0$ for all $\boldsymbol{z}^* \neq \boldsymbol{z}$ with $f(\boldsymbol{z}^*) > 0$, i.e., there cannot be a $\boldsymbol{z}$ and a $\boldsymbol{z}^*$ that are both consistent with an observed $\boldsymbol{a}$. However, we will now show that for any candidate $\boldsymbol{z}$ that is consistent with the observed $\boldsymbol{a}$, there always exists a slight perturbation of $\boldsymbol{z}$ that is also consistent with $\boldsymbol{a}$. With finite m, continuity implies that for any $0 < \epsilon < f(\boldsymbol{a}|\boldsymbol{z})$, there exists a $\boldsymbol{\delta} \in \mathbb{R}^m$ such that $0 < f(\boldsymbol{a}|\boldsymbol{z}) - \epsilon < f(\boldsymbol{a}|\boldsymbol{z}+\boldsymbol{\delta})$, so there are always multiple candidate confounder values consistent with any observed treatment vector. □

Next, we present results relating to the inconsistency of various components of the deconfounder in the linear-linear setting. To review, the deconfounder proceeds as follows. It takes the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, then extracts the first k components to form $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$. It then computes $\hat{\mathbb{E}}[\boldsymbol{A}|\hat{\boldsymbol{Z}}] = \hat{\boldsymbol{Z}}\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}} \equiv \frac{1}{\sqrt{n}}\boldsymbol{D}_{1:k}\boldsymbol{V}_{1:k}^\top$.

Lemma 3. *(Inconsistency of $\hat{\boldsymbol{\theta}}$.) The asymptotic behavior of $\hat{\boldsymbol{\theta}}$ is governed by*

$$\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\theta}} = \left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}} \boldsymbol{\Lambda}_{1:k}^{-\frac{1}{2}}\boldsymbol{R}^\top\boldsymbol{\theta},$$

where $\boldsymbol{R}$ and $\boldsymbol{\Lambda}_{1:k}$ are given by the eigendecomposition $\boldsymbol{\theta}\boldsymbol{\theta}^\top = \boldsymbol{R}\boldsymbol{\Lambda}_{1:k}\boldsymbol{R}^\top$.

Proof We begin with the eigendecomposition $\boldsymbol{\theta}^\top\boldsymbol{\theta} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top = \boldsymbol{Q}_{1:k}\boldsymbol{\Lambda}_{1:k}\boldsymbol{Q}_{1:k}^\top$, where the last step follows from the fact that the trailing $m-k$ diagonal entries of $\boldsymbol{\Lambda}$ are zero.

We now turn to $\hat{\boldsymbol{\theta}} = \frac{1}{\sqrt{n}}\boldsymbol{D}_{1:k}\boldsymbol{V}_{1:k}^\top$. By the properties of probabilistic PCA, $p\text{-lim}_{n\to\infty}\frac{1}{\sqrt{n}}\boldsymbol{D}_{1:k} = \left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}}$ and $p\text{-lim}_{n\to\infty}\boldsymbol{V}_{1:k} = \boldsymbol{Q}_{1:k}$ (Tipping and Bishop, 1999). The lemma then follows from the singular value decomposition $\boldsymbol{\theta} = \boldsymbol{R}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{Q}^\top = \boldsymbol{R}\boldsymbol{\Lambda}_{1:k}^{\frac{1}{2}}\boldsymbol{Q}_{1:k}^\top$ by solving for $\boldsymbol{Q}_{1:k}$ and substituting.

It will be the case that the white-noised and subset deconfounder implicitly rely on $\hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}}$ to adjust for dependence between causes. In Lemma 4, we show that a consequence of Lemma 3 (inconsistency of $\hat{\boldsymbol{\theta}}$) is that $\hat{\text{Cov}}(\boldsymbol{A}) \equiv \hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}}$ is a poor estimator of the covariance of $\boldsymbol{A}$; the dependence will be incorrectly modeled even as n goes to infinity.

Lemma 4. *(Mismodeled dependence structure in $\boldsymbol{A}$.) When $\widehat{\text{Cov}}(\boldsymbol{A}) = \hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}}$ is used as an estimator for $\text{Cov}(\boldsymbol{A})$, the unmodeled residual dependence among causes is asymptotically equal to $\sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}\right]$.*

When the number of causes is finite, this residual covariance is nonspherical. In contrast, the true conditional dependence is $\text{Cov}(\boldsymbol{A}|\boldsymbol{Z}) = \sigma^2\boldsymbol{I}$.
Proof.

$$\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\,\text{Cov}(\boldsymbol{A}) - \widehat{\text{Cov}}(\boldsymbol{A}) &= \underset{n\to\infty}{p\text{-lim}}\,\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} - \hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}} \\
&= \boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} - \boldsymbol{\theta}^\top\boldsymbol{R}\left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)\boldsymbol{\Lambda}_{1:k}^{-1}\boldsymbol{R}^\top\boldsymbol{\theta} \\
&= \sigma^2\boldsymbol{I} - \sigma^2\boldsymbol{\theta}^\top\boldsymbol{R}\boldsymbol{\Lambda}_{1:k}^{-1}\boldsymbol{R}^\top\boldsymbol{\theta} \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}\right]
\end{aligned}$$

□

Under the strong infinite confounding assumption,

$$\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\,\text{Cov}(\boldsymbol{A}) - \widehat{\text{Cov}}(\boldsymbol{A}) = \sigma^2\boldsymbol{I}$$

An Example When Strong Infinite Confounding Fails. Here we consider an example where the number of treatments increases, but strong infinite confounding does not hold. This builds on an idea found in D'Amour (2019a). Suppose $Z_i \sim \mathcal{N}(0,\sigma^2)$. We will suppose that $A_{i,m} = \frac{1}{m^2}Z_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$. In this example, as $\lim_{m\to\infty}\boldsymbol{\theta}\boldsymbol{\theta}^\top = \sum_{m=1}^{\infty}\frac{1}{m^4} = \frac{\pi^4}{90}$. Therefore, strong infinite confounding fails. In general, in the one-dimensional case, the infinite series must diverge for strong infinite confounding to hold.

We now provide a minor result that helps characterize the behavior of the naïve estimator, (13), when applied to sequences of data-generating processes satisfying *strong infinite confounding* (Definition 1). In Supplement B.2 (proof of Proposition 1), we will show the conditions for asymptotic unbiasedness of the naïve estimator as n and m go to infinity. Lemma 5 states that under the assumption of strong infinite confounding, this condition is asymptotically satisfied as m grows.

B.1.3 Behavior of the Naïve Estimator

Lemma 5. *(Naïve convergence under strong infinite confounding.) A sequence of strongly infinitely confounded data-generating processes satisfies*

$$\lim_{m\to\infty} \boldsymbol{\theta}\left(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\theta}^\top = \boldsymbol{I}$$

Proof. By the Woodbury matrix identity,

$$\begin{aligned}
\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\frac{1}{\sigma^4}\boldsymbol{I} + \frac{1}{\sigma^2}\boldsymbol{I}\right)^{-1} &= \sigma^2\boldsymbol{I} - \sigma^4\boldsymbol{I}\boldsymbol{\theta}(\sigma^4\boldsymbol{I} + \sigma^2\boldsymbol{I}\boldsymbol{\theta}^\top\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^\top \\
\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\frac{1}{\sigma^2}\boldsymbol{I} + \boldsymbol{I}\right)^{-1}\sigma^2 &= \sigma^2\boldsymbol{I} - \sigma^2\boldsymbol{\theta}(\sigma^2\boldsymbol{I} + \boldsymbol{\theta}^\top\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^\top \\
\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\frac{1}{\sigma^2}\boldsymbol{I} + \boldsymbol{I}\right)^{-1} &= \boldsymbol{I} - \boldsymbol{\theta}(\sigma^2\boldsymbol{I} + \boldsymbol{\theta}^\top\boldsymbol{\theta})^{-1}\boldsymbol{\theta}^\top
\end{aligned}$$

for any m. Because both the entries and number of columns of $\boldsymbol{\theta}$ are finite, the strong infinite confounding condition requires that the diagonal elements of $\boldsymbol{\theta}\boldsymbol{\theta}^\top$ also tend to infinity as m grows large. Therefore $\lim_{m\to\infty}\left(\frac{1}{\sigma^2}\boldsymbol{\theta}\boldsymbol{\theta}^\top + \boldsymbol{I}\right)^{-1} = \boldsymbol{0}$, and $\lim_{m\to\infty}\boldsymbol{\theta}\left(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\theta}^\top = \boldsymbol{I}$. □

Supplement B.7 (proof of Proposition 4) shows that unbiasedness of the subset deconfounder estimator requires $\lim_{m\to\infty}\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\right)^{-1} = \boldsymbol{0}$, which is trivially satisfied for strongly infinitely confounded sequences of data-generating processes.

B.2 Bias of the Naïve Estimator

For convenience, we reiterate the data-generating process, naïve estimation procedure. We will suppose, without loss of generality, that the m causes, $\boldsymbol{A}$, are divided into m_F focal causes of interest, the column subset $\boldsymbol{A}_F$, and m_N nonfocal causes, $\boldsymbol{A}_N$. As before, we

consider n observations drawn i.i.d. as follows.

$$\begin{aligned}
\underset{n\times k}{\boldsymbol{Z}} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\
\underset{n\times m_F}{\boldsymbol{\nu}_F} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m_N}{\boldsymbol{\nu}_N} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m_F}{\boldsymbol{A}_F} &= \boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F \\
\underset{n\times m_N}{\boldsymbol{A}_N} &= \boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N \\
\underset{n\times 1}{\boldsymbol{\epsilon}} &\sim \mathcal{N}(\mathbf{0}, \omega^2) \\
\underset{n\times 1}{\boldsymbol{Y}} &= \boldsymbol{A}_F\boldsymbol{\beta}_F + \boldsymbol{A}_N\boldsymbol{\beta}_N + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

The naïve estimator estimates the treatment effects by conducting a regression of the outcome on both focal and nonfocal causes, producing estimates for both focal and nonfocal effects, then discarding the latter. The full regression coefficients are

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_F^{\text{naïve}} \\ \hat{\boldsymbol{\beta}}_N^{\text{naïve}} \end{bmatrix} \equiv \left([\boldsymbol{A}_F, \boldsymbol{A}_N]^\top [\boldsymbol{A}_F, \boldsymbol{A}_N] \right)^{-1} [\boldsymbol{A}_F, \boldsymbol{A}_N]^\top \boldsymbol{Y}. \tag{13}$$

We will prove Proposition 5, a generalization of Proposition 1 that distinguishes between focal and non-focal treatment cases. The proof of Proposition 1 is by reduction to the special case in which all causes are of interest, so that $\boldsymbol{A}_F = \boldsymbol{A}$ and $\boldsymbol{A}_N$ is empty.

Proposition 5. *(Asymptotic Bias of the Naïve Regression under Strong Infinite Confounding.)*

Suppose that the m causes, $\boldsymbol{A}$, are divided into focal causes of interest, the column subset $\boldsymbol{A}_F$, and nonfocal causes, $\boldsymbol{A}_N$, without loss of generality. The bias of the naïve estimator, (13)*, for the corresponding focal effects, $\boldsymbol{\beta}_F$, is given by*

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{naïve}} - \boldsymbol{\beta}_F = \left[\boldsymbol{\theta}_F^\top\boldsymbol{\theta}_F + \sigma^2\boldsymbol{I} - \boldsymbol{\theta}_F^\top\boldsymbol{\Omega}\boldsymbol{\theta}_F \right]^{-1} \left[\boldsymbol{\theta}_F^\top - \boldsymbol{\theta}_F^\top\boldsymbol{\Omega} \right] \boldsymbol{\gamma},$$

where $\boldsymbol{\Omega} = \boldsymbol{\theta}_N \left(\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_N + \sigma^2\boldsymbol{I} \right)^{-1} \boldsymbol{\theta}_N^\top$ and $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_N$ are the corresponding column subsets of $\boldsymbol{\theta}$. Under the assumptions of the linear-linear model,

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{naïve}} - \boldsymbol{\beta}_F = \mathbf{0}.$$

Proof of Proposition 5. By the Frisch-Waugh-Lovell theorem, $\hat{\boldsymbol{\beta}}_F$ can be re-expressed in terms of the portion of $\boldsymbol{A}_F$ not explained by $\boldsymbol{A}_N$. Without loss of generality we suppose that the set of focal treatments, $\boldsymbol{A}_F$ are fixed at the outset and are finite. We denote the

residualized focal treatments as $\tilde{\boldsymbol{A}}_F^{\text{naïve}} = \boldsymbol{A}_F - \hat{\boldsymbol{A}}_F^{\text{naïve}}$, where

$$\begin{aligned}
\hat{\boldsymbol{A}}_F^{\text{naïve}} &= \hat{\mathbb{E}}[\boldsymbol{A}_F | \boldsymbol{A}_N] = \boldsymbol{A}_N \hat{\boldsymbol{\zeta}}, \\
\hat{\boldsymbol{\zeta}} &= (\boldsymbol{A}_N^\top \boldsymbol{A}_N)^{-1} \boldsymbol{A}_N^\top \boldsymbol{A}_F, \text{ and} \\
\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\zeta}} &\equiv \boldsymbol{\zeta} = \left(\boldsymbol{\theta}_N^\top \boldsymbol{\theta}_N + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{\theta}_N^\top \boldsymbol{\theta}_F.
\end{aligned}$$

The naïve estimator is then rewritten as follows:

$$\hat{\boldsymbol{\beta}}_F^{\text{naïve}} = \left(\frac{1}{n} \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{A}}_F^{\text{naïve}}\right)^{-1} \frac{1}{n} \tilde{\boldsymbol{A}}_F^{\text{naïve}} \boldsymbol{Y}$$

We now characterize the asymptotic bias of this estimator by examining the behavior of $\frac{1}{n} \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{A}}_F^{\text{naïve}}$ and $\frac{1}{n} \tilde{\boldsymbol{A}}_F^{\text{naïve}} \boldsymbol{Y}$ in turn. Beginning with the residual variance of the focal causes,

$$\begin{aligned}
\frac{1}{n} \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{A}}_F^{\text{naïve}} &= \frac{1}{n} \left(\boldsymbol{A}_F - \hat{\boldsymbol{A}}_F^{\text{naïve}}\right)^\top \left(\boldsymbol{A}_F - \hat{\boldsymbol{A}}_F^{\text{naïve}}\right) \\
&= \frac{1}{n} \left(\boldsymbol{A}_F^\top \boldsymbol{A}_F + \hat{\boldsymbol{A}}_F^{\text{naïve}\top} \hat{\boldsymbol{A}}_F^{\text{naïve}} - \boldsymbol{A}_F^\top \hat{\boldsymbol{A}}_F^{\text{naïve}} - \hat{\boldsymbol{A}}_F^{\text{naïve}\top} \boldsymbol{A}_F\right) \\
&= \frac{1}{n} (\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F)^\top (\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F) \\
&\quad + \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top (\boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N)^\top (\boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N) \hat{\boldsymbol{\zeta}} \\
&\quad - \frac{1}{n} (\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F)^\top (\boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N) \hat{\boldsymbol{\zeta}} \\
&\quad - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top (\boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N)^\top (\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F) \\
\underset{n\to\infty}{p\text{-lim}}\, \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{A}}_F^{\text{naïve}} &= \underset{n\to\infty}{p\text{-lim}}\, \boldsymbol{\theta}_F^\top \boldsymbol{\theta}_F + \sigma^2 \boldsymbol{I} + \hat{\boldsymbol{\zeta}}^\top (\boldsymbol{\theta}_N^\top \boldsymbol{\theta}_N + \sigma^2 \boldsymbol{I}) \hat{\boldsymbol{\zeta}} - \boldsymbol{\theta}_F^\top \boldsymbol{\theta}_N \hat{\boldsymbol{\zeta}} - \hat{\boldsymbol{\zeta}}^\top \boldsymbol{\theta}_N^\top \boldsymbol{\theta}_F \\
&= \boldsymbol{\theta}_F^\top \boldsymbol{\theta}_F + \sigma^2 \boldsymbol{I} - \boldsymbol{\theta}_F^\top \boldsymbol{\theta}_N \left(\boldsymbol{\theta}_N^\top \boldsymbol{\theta}_N + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{\theta}_N^\top \boldsymbol{\theta}_F, \quad \text{and} \qquad (14) \\
\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \tilde{\boldsymbol{A}}_F^{\text{naïve}} &= \sigma^2 \boldsymbol{I} \quad \text{under the infinite confounding assumption.} \qquad (15)
\end{aligned}$$

Turning to the residual covariance between the focal causes and the outcome,

$$\begin{aligned}
\frac{1}{n} \tilde{\boldsymbol{A}}_F^{\text{naïve}\top} \boldsymbol{Y} &= \frac{1}{n} \left(\boldsymbol{A}_F - \hat{\boldsymbol{A}}_N \hat{\boldsymbol{\zeta}}\right)^\top (\boldsymbol{A}_F \boldsymbol{\beta}_F + \boldsymbol{A}_N \boldsymbol{\beta}_N + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\
&= \frac{1}{n} \boldsymbol{A}_F^\top \boldsymbol{A}_F \boldsymbol{\beta}_F + \frac{1}{n} \boldsymbol{A}_F^\top \boldsymbol{A}_N \boldsymbol{\beta}_N + \frac{1}{n} \boldsymbol{A}_F^\top \boldsymbol{Z}\boldsymbol{\gamma} + \frac{1}{n} \boldsymbol{A}_F^\top \boldsymbol{\epsilon} \\
&\quad - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top \boldsymbol{A}_N^\top \boldsymbol{A}_F \boldsymbol{\beta}_F - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top \boldsymbol{A}_N^\top \boldsymbol{A}_N \boldsymbol{\beta}_N - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top \boldsymbol{A}_N^\top \boldsymbol{Z}\boldsymbol{\gamma} - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top \boldsymbol{A}_N^\top \boldsymbol{\epsilon} \\
&= \frac{1}{n} (\boldsymbol{\theta}_F^\top \boldsymbol{Z}^\top + \boldsymbol{\nu}_F^\top)(\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F) \boldsymbol{\beta}_F + \frac{1}{n} (\boldsymbol{\theta}_F^\top \boldsymbol{Z}^\top + \boldsymbol{\nu}_F^\top)(\boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N) \boldsymbol{\beta}_N \\
&\quad + \frac{1}{n} (\boldsymbol{\theta}_F^\top \boldsymbol{Z}^\top + \boldsymbol{\nu}_F^\top) \boldsymbol{Z}\boldsymbol{\gamma} + \frac{1}{n} \boldsymbol{A}_F^\top \boldsymbol{\epsilon} - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top (\boldsymbol{\theta}_N^\top \boldsymbol{Z}^\top + \boldsymbol{\nu}_N^\top)(\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F) \boldsymbol{\beta}_F \\
&\quad - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top (\boldsymbol{\theta}_N^\top \boldsymbol{Z}^\top + \boldsymbol{\nu}_N^\top)(\boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N) \boldsymbol{\beta}_N - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top (\boldsymbol{\theta}_N^\top \boldsymbol{Z}^\top + \boldsymbol{\nu}_N^\top) \boldsymbol{Z}\boldsymbol{\gamma} - \frac{1}{n} \hat{\boldsymbol{\zeta}}^\top \boldsymbol{A}_N^\top \boldsymbol{\epsilon}
\end{aligned}$$

Taking limits,

$$\underset{n\to\infty}{p\text{-lim}}\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{naïve}\top}\boldsymbol{Y} = \left[\boldsymbol{\theta}_F^\top\boldsymbol{\theta}_F + \sigma^2 - \boldsymbol{\theta}_F^\top\boldsymbol{\theta}_N\left(\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_N + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F + \left[\boldsymbol{\theta}_F^\top - \boldsymbol{\theta}_F^\top\boldsymbol{\theta}_N\left(\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_N + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\theta}_N^\top\right]\boldsymbol{\gamma}, \quad \text{and} \tag{16}$$

$$\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{naïve}\top}\boldsymbol{Y} = \sigma^2\boldsymbol{\beta}_F \quad \text{under the infinite confounding assumption.} \tag{17}$$

Combining (14) and (16) and applying Lemma 3 to $\boldsymbol{\theta}_N$,

$$\underset{n\to\infty}{p\text{-lim}}\hat{\boldsymbol{\beta}}_F^{\text{naïve}} = \boldsymbol{\beta}_F + \left[\boldsymbol{\theta}_F^\top\boldsymbol{\theta}_F + \sigma^2\boldsymbol{I} - \boldsymbol{\theta}_F^\top\boldsymbol{\theta}_N\left(\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_N + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_F\right]^{-1} \cdot \left[\boldsymbol{\theta}_F^\top - \boldsymbol{\theta}_F^\top\boldsymbol{\theta}_N\left(\boldsymbol{\theta}_N^\top\boldsymbol{\theta}_N + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\theta}_N^\top\right]\boldsymbol{\gamma},$$

and under the infinite confounding assumption, (15) and (17) yield

$$\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\hat{\boldsymbol{\beta}}_F^{\text{naïve}} = \boldsymbol{\beta}_F.$$

When all effects are of interest, the above reduces to

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{naïve}} &\equiv \left(\boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{Y} \\
\underset{n\to\infty}{p\text{-lim}}\frac{1}{n}\boldsymbol{A}^\top\boldsymbol{A} &= \boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} \\
\underset{n\to\infty}{p\text{-lim}}\frac{1}{n}\boldsymbol{A}^\top\boldsymbol{Y} &= \underset{n\to\infty}{p\text{-lim}}\frac{1}{n}(\boldsymbol{\theta}^\top\boldsymbol{Z}^\top + \boldsymbol{\nu}^\top)(\boldsymbol{Z}\boldsymbol{\theta}\boldsymbol{\beta} + \boldsymbol{\nu}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\theta}^\top\boldsymbol{\theta}\boldsymbol{\beta} + \boldsymbol{\theta}^\top\boldsymbol{\gamma} + \sigma^2\boldsymbol{I}\boldsymbol{\beta} \\
\underset{n\to\infty}{p\text{-lim}}\hat{\boldsymbol{\beta}}^{\text{naïve}} &= \boldsymbol{\beta} + (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma} \\
\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\hat{\boldsymbol{\beta}}^{\text{naïve}} &= \boldsymbol{\beta} \quad \square
\end{aligned}$$

B.3 Bias of the Penalized Deconfounder Estimator

For convenience, we reiterate the data-generating process and penalized deconfounder estimation procedure here, along with identities that will be useful in the proof of Proposition 2. As before, we consider n observations drawn i.i.d. as follows.

$$\begin{aligned}
\underset{n\times k}{\boldsymbol{Z}} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{\nu}} &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{A}} &= \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\nu} \\
\underset{n\times 1}{\boldsymbol{\epsilon}} &\sim \mathcal{N}(\boldsymbol{0}, \omega^2) \\
\underset{n\times 1}{\boldsymbol{Y}} &= \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

The penalized deconfounder estimator (1) takes the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$; (2) extracts the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$; and (3) estimates the focal effects by computing

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{\text{penalty}} \\ \hat{\boldsymbol{\gamma}}^{\text{penalty}} \end{bmatrix} \equiv \left(\left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right] + \lambda(n)\boldsymbol{I} \right)^{-1} \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y}$$

and discarding $\hat{\boldsymbol{\gamma}}^{\text{penalty}}$. The $\lambda(n)$ term indicates the strength of the ridge penalty; we allow this term to scale with n for full generality. Note that identification is purely from this term ridge penalty—because $\hat{\boldsymbol{Z}}$ is merely a linear transformation of $\boldsymbol{A}$, the above is non-estimable when $\lambda(n) = 0$.

We now restate Proposition 2 for convenience.

Proposition 2. *(Asymptotic Bias of the Penalized Full Deconfounder.)*

Consider the linear-linear data-generating process, in which n observations are sampled i.i.d. by drawing k unobserved confounders, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$; these generate $m \geq k$ observed treatments, $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{Z}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$; and a scalar outcome is drawn from $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}, \omega^2)$. The penalized deconfounder estimator, as implemented in WB, is

$$\left[\ \hat{\boldsymbol{\beta}}^{\text{penalty}\top},\ \hat{\boldsymbol{\gamma}}^{\text{penalty}\top}\ \right]^\top \equiv \left(\left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right] + \lambda(n)\boldsymbol{I} \right)^{-1} \left[\boldsymbol{A}, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y},$$

where $\hat{\boldsymbol{Z}}$ is obtained by taking the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ and extracting the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$, and $\lambda(n)$ is a ridge penalty that is assumed to be sublinear in n. The asymptotic bias of this estimator is given by

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} - \boldsymbol{\beta} = \overbrace{-\boldsymbol{Q}_{1:k}\ \text{diag}_j\left(\frac{1}{\sigma^2 + \Lambda_j + 1}\right)\boldsymbol{Q}_{1:k}^\top\boldsymbol{\beta}}^{\textit{Regularization of treatment effect estimation}} \underbrace{+\boldsymbol{Q}_{1:k}\ \text{diag}_j\left(\frac{\Lambda_j}{\sigma^2 + \Lambda_j + 1}\right)\boldsymbol{Q}_{1:k}^\top\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\gamma}}_{\textit{Regularization of confounder adjustment}},$$

where $\boldsymbol{Q}$ and $\boldsymbol{\Lambda} = [\Lambda_1, \ldots, \Lambda_k, 0, \ldots]$ are respectively eigenvectors and eigenvalues obtained from decomposition of $\boldsymbol{\theta}^\top\boldsymbol{\theta}$. Under strong infinite confounding,

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} - \boldsymbol{\beta} = \boldsymbol{0}.$$

In what follows, we prove Proposition 2 by relating the asymptotic behavior of the penalized deconfounder to the eigendecomposition $\boldsymbol{\theta}^\top\boldsymbol{\theta} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top = \boldsymbol{Q}_{1:k}\boldsymbol{\Lambda}_{1:k}\boldsymbol{Q}_{1:k}^\top$. We will rely on the singular value decompositions of $\boldsymbol{A}$ and $[\boldsymbol{A}, \hat{\boldsymbol{Z}}]$. To distinguish these, for this section only, we denote the former as $\boldsymbol{A} = \boldsymbol{U}_A\boldsymbol{D}_A\boldsymbol{V}_A^\top$ and the latter as $[\boldsymbol{A}, \hat{\boldsymbol{Z}}] = \boldsymbol{U}_{AZ}\boldsymbol{D}_{AZ}\boldsymbol{V}_{AZ}^\top$. Lemma 6 characterizes the relationship between these.

Lemma 6. *For any n, the singular value decomposition $[\boldsymbol{A}, \hat{\boldsymbol{Z}}] = \boldsymbol{U}_{AZ}\boldsymbol{D}_{AZ}\boldsymbol{V}_{AZ}^\top$ obeys*

$$\boldsymbol{U}_{AZ} = [\boldsymbol{U}_A,\ *]$$

$$\boldsymbol{D}_{AZ} = \begin{bmatrix} \left(\boldsymbol{D}_{A,1:k}^2 + n\boldsymbol{I}\right)^{\frac{1}{2}}, & \boldsymbol{0}, & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{D}_{A,(k+1):m}, & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{0}, & \boldsymbol{0} \end{bmatrix}$$

$$\boldsymbol{V}_{AZ}^\top = \begin{bmatrix} \left(\boldsymbol{D}_{A,1:k}^2 + n\boldsymbol{I}\right)^{-\frac{1}{2}} \boldsymbol{D}_{A,1:k}\boldsymbol{V}_{A,1:k}^\top, & \sqrt{n}\left(\boldsymbol{D}_{A,1:k}^2 + n\boldsymbol{I}\right)^{-\frac{1}{2}} \\ \boldsymbol{V}_{A,(k+1):m}^\top, & \boldsymbol{0} \\ * & * \end{bmatrix},$$

where $*$ indicates irrelevant normalizing columns in $\boldsymbol{U}_{AZ}$ and $\boldsymbol{V}_{AZ}$.

Proof of Lemma 6. The first equality follows from the fact that the newly appended $\hat{\boldsymbol{Z}}$ columns are merely linear transformations of $\boldsymbol{A}$, so that the leading m left singular vectors remain unchanged.

Of the unchanged left singular vectors, each of the first k is directly proportional to the corresponding column of $\hat{\boldsymbol{Z}}$. Because $\hat{\boldsymbol{Z}}$ is standardized by construction, the variance explained by each of the first k left singular vector increases by one; the $(k+1)$-th through m-th left singular vectors are orthogonal to the newly appended $\hat{\boldsymbol{Z}}$ and so their singular values remain unchanged. This yields the second equality.

The third equality can verified by $\boldsymbol{U}_{AZ}\boldsymbol{D}_{AZ}\boldsymbol{V}_{AZ}^\top = [\boldsymbol{U}_A\boldsymbol{D}_A\boldsymbol{V}_A^\top, \sqrt{n}\boldsymbol{U}_{A,1:k}] = [\boldsymbol{A}, \hat{\boldsymbol{Z}}]$ and $\boldsymbol{V}_{AZ,1:m}^\top\boldsymbol{V}_{AZ,1:m} = \boldsymbol{I}$. □

We now examine the asymptotic behavior of the penalized deconfounder estimator.

Proof of Proposition 2.

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{\text{penalty}} \\ \hat{\boldsymbol{\gamma}}^{\text{penalty}} \end{bmatrix} = \left([\boldsymbol{A},\ \hat{\boldsymbol{Z}}]^\top[\boldsymbol{A},\ \hat{\boldsymbol{Z}}] + \lambda(n)\boldsymbol{I}\right)^{-1}[\boldsymbol{A},\ \hat{\boldsymbol{Z}}]^\top\ \boldsymbol{Y} \tag{18}$$

$$= \boldsymbol{V}_{AZ}\left(\boldsymbol{D}_{AZ}^2 + \lambda(n)\boldsymbol{I}\right)^{-1}\boldsymbol{D}_{AZ}\boldsymbol{U}_{AZ}^\top\boldsymbol{Y} \tag{19}$$

By Lemma 6,

$$= \begin{bmatrix} \boldsymbol{V}_A\boldsymbol{D}_A, & * \\ \sqrt{n}\boldsymbol{I}, & * \end{bmatrix} \begin{bmatrix} \left(\boldsymbol{D}_A^2 + \lambda(n)\boldsymbol{I} + n\cdot\text{diag}_j 1\{j \le k\}\right)^{-1}, & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{0} \end{bmatrix} [\boldsymbol{U}_A,\ *]^\top\boldsymbol{Y}$$

where asterisks denote irrelevant blocks, eliminated below.

$$= \begin{bmatrix} \boldsymbol{V}_A\boldsymbol{D}_A \\ \sqrt{n}\boldsymbol{I} \end{bmatrix} \left(\boldsymbol{D}_A^2 + \lambda(n)\boldsymbol{I} + n\cdot\text{diag}_j 1\{j \le k\}\right)^{-1}\boldsymbol{U}_A^\top(\ \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})$$

We now subset to $\hat{\boldsymbol{\beta}}^{\text{penalty}}$, then substitute $\boldsymbol{A} = \boldsymbol{U}_A \boldsymbol{D}_A \boldsymbol{V}_A^\top$, $\boldsymbol{U}_A^\top = \boldsymbol{D}_A^{-1} \boldsymbol{V}_A^\top \boldsymbol{A}^\top$ and $\boldsymbol{Z} = (\boldsymbol{A} - \boldsymbol{\nu})\boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}$,

$$
\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} &= \underset{n\to\infty}{p\text{-lim}}\, \boldsymbol{V}_A \boldsymbol{D}_A \left(\boldsymbol{D}_A^2 + \lambda(n)\boldsymbol{I} + n \cdot \text{diag}_j 1\{j \le k\}\right)^{-1} \boldsymbol{D}_A \boldsymbol{V}_A^\top \boldsymbol{\beta} \\
&\quad + \boldsymbol{V}_A \boldsymbol{D}_A \left(\boldsymbol{D}_A^2 + \lambda(n)\boldsymbol{I} + n \cdot \text{diag}_j 1\{j \le k\}\right)^{-1} \boldsymbol{D}_A \boldsymbol{V}_A^\top \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma} \\
&\quad - \boldsymbol{V}_A \left(\boldsymbol{D}_A^2 + \lambda(n)\boldsymbol{I} + n \cdot \text{diag}_j 1\{j \le k\}\right)^{-1} \boldsymbol{V}_A^\top \boldsymbol{A}^\top \boldsymbol{\nu} \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma} \\
&= \underset{n\to\infty}{p\text{-lim}}\, \boldsymbol{V}_A \, \text{diag}_j \left(\frac{\sigma^2 + 1\{j \le k\}\Lambda_j}{\sigma^2 + \lambda(n)/n + 1\{j \le k\}(\Lambda_j + 1)} \right) \boldsymbol{V}_A^\top \boldsymbol{\beta} \\
&\quad + \boldsymbol{V}_A \, \text{diag}_j \left(\frac{\sigma^2 + 1\{j \le k\}\Lambda_j}{\sigma^2 + \lambda(n)/n + 1\{j \le k\}(\Lambda_j + 1)} \right) \boldsymbol{V}_A^\top \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma} \\
&\quad - \boldsymbol{V}_A \, \text{diag}_j \left(\frac{\sigma^2}{\sigma^2 + \lambda(n)/n + 1\{j \le k\}(\Lambda_j + 1)} \right) \boldsymbol{V}_A^\top \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma}
\end{aligned} \tag{20}
$$

By properties of PPCA (Tipping and Bishop, 1999),

$$
\begin{aligned}
&= \underset{n\to\infty}{p\text{-lim}}\, \boldsymbol{\beta} - \boldsymbol{V}_A \, \text{diag}_j \left(\frac{\lambda(n)/n + 1\{j \le k\}}{\sigma^2 + \lambda(n)/n + 1\{j \le k\}(\Lambda_j + 1)} \right) \boldsymbol{V}_A^\top \boldsymbol{\beta} \\
&\quad + \boldsymbol{V}_A \, \text{diag}_j \left(\frac{1\{j \le k\}\Lambda_j}{\sigma^2 + \lambda(n)/n + 1\{j \le k\}(\Lambda_j + 1)} \right) \boldsymbol{V}_A^\top \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma} \\
&= \boldsymbol{\beta} - \boldsymbol{Q}_{1:k} \, \text{diag}_j \left(\frac{\lambda(n)/n + 1}{\sigma^2 + \lambda(n)/n + \Lambda_j + 1} \right) \boldsymbol{Q}_{1:k}^\top \boldsymbol{\beta} \\
&\quad - \left(\frac{\lambda(n)/n}{\sigma^2 + \lambda(n)/n} \right) \boldsymbol{Q}_{(k+1):m} \boldsymbol{Q}_{(k+1):m}^\top \boldsymbol{\beta} \\
&\quad + \boldsymbol{Q}_{1:k} \, \text{diag}_j \left(\frac{\Lambda_j}{\sigma^2 + \lambda(n)/n + \Lambda_j + 1} \right) \boldsymbol{Q}_{1:k}^\top \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma}
\end{aligned}
$$

When $\lambda(n)$ is sublinear,

$$
\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} &= \boldsymbol{\beta} - \boldsymbol{Q}_{1:k} \, \text{diag}_j \left(\frac{1}{\sigma^2 + \Lambda_j + 1} \right) \boldsymbol{Q}_{1:k}^\top \boldsymbol{\beta} \\
&\quad + \boldsymbol{Q}_{1:k} \, \text{diag}_j \left(\frac{\Lambda_j}{\sigma^2 + \Lambda_j + 1} \right) \boldsymbol{Q}_{1:k}^\top \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma}.
\end{aligned}
$$

Under strong infinite confounding, it can be seen that $\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{penalty}} = \boldsymbol{\beta}$. This follows from $\lim_{m\to\infty} \boldsymbol{\Lambda}_{1:k}^{-1} = \boldsymbol{0}$ and $\lim_{m\to\infty} \boldsymbol{\theta}^\top (\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \boldsymbol{\gamma} = \boldsymbol{0}$. □

B.4 Bias of the Penalized Deconfounder Estimator under Nonlinear Confounding

In this section, we evaluate the behavior of the deconfounder in a more general setting. We consider n observations drawn i.i.d. from the below data-generating process.

$$\begin{aligned}
\underset{n\times k}{\boldsymbol{Z}} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{\nu}} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{A}} &= \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\nu} \\
\underset{n\times 1}{\boldsymbol{\epsilon}} &\sim \mathcal{N}(\mathbf{0}, \omega^2)
\end{aligned}$$

However, we relax the outcome model to allow for arbitrary additively separable confounding,

$$\underset{n\times 1}{\boldsymbol{Y}} = \left[\boldsymbol{A}_i^\top\boldsymbol{\beta} + g_Y(\boldsymbol{Z}_i) + \boldsymbol{\epsilon}_i\right] \tag{21}$$

The linear-linear model defined in Section 3 is a special case of the linear-separable model. Note that in empirical settings, the specific functional form of $g_Y(\cdot)$ is rarely known except when analyzing simple physical systems. However, any $g_Y(\cdot)$ can be approximated to arbitrary degree d as follows. First, denote the polynomial basis expansion of $\boldsymbol{Z}_i$ as $h(\boldsymbol{Z}_i) \equiv \left[\prod_{k'=1}^{k} Z_{i,k'}^{d'_{k'}}\right]_{\sum_{k'=1}^{k} d'_{k'} \leq d}$ and collect these in rows of $h(\boldsymbol{Z}) = [h(\boldsymbol{Z}_i)]$. Then, (21) can be rewritten by Taylor expansion as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + h(\boldsymbol{Z})\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{22}$$

with approximation error that grows arbitrarily small as d grows large. We will choose d sufficiently to fully capture $g_Y(\cdot)$. Let $\boldsymbol{W}$ be the orthogonal higher-order polynomials of $\boldsymbol{Z}$. Then, (21) can be rewritten yet again as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{\beta} + \gamma\boldsymbol{Z} + \delta\boldsymbol{W} + \boldsymbol{\epsilon}. \tag{23}$$

As before, the effects $\boldsymbol{\beta}$ are the causal quantities of interest. Note that the confounding, $g_Y(\boldsymbol{Z})$, is a nuisance term, so there is no need to reconstruct it from its expansion. We will assume that $g_Y(\boldsymbol{Z})$ is zero-mean for convenience; this assumption is trivial to relax using an added intercept.

We will derive the asymptotic behavior of the flexible penalized deconfounder, which generalizes the penalized full deconfounder of Supplement B.3 for all additively separable forms of confounding. The flexible penalized deconfounder estimator consists of the following procedure: (1) take the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$; and (2) extract the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$. To allow for nonlinear confounding, (3) compute $h(\hat{\boldsymbol{Z}})$ and take its QR decomposition, $h(\hat{\boldsymbol{Z}}) = \boldsymbol{Q}_Z\boldsymbol{R}_Z = \left[\frac{1}{\sqrt{n}}\hat{\boldsymbol{Z}},\ \frac{1}{\sqrt{n}}\hat{\boldsymbol{W}}\right]\boldsymbol{R}_Z$.[16] Finally, (4) estimate $\boldsymbol{\beta}$ by a ridge regression of the form

$$[\hat{\boldsymbol{\beta}}^{\text{nonlinear}\top},\ \hat{\boldsymbol{\gamma}}^{\text{nonlinear}\top},\ \hat{\boldsymbol{\delta}}^{\text{nonlinear}\top}]^\top = \left([\boldsymbol{A},\ \hat{\boldsymbol{Z}},\ \hat{\boldsymbol{W}}]^\top[\boldsymbol{A},\ \hat{\boldsymbol{Z}},\ \hat{\boldsymbol{W}}] + \lambda(n)\boldsymbol{I}\right)^{-1}[\boldsymbol{A},\ \hat{\boldsymbol{Z}},\ \hat{\boldsymbol{W}}]^\top\boldsymbol{Y}.$$

16. The invariance of $\hat{\boldsymbol{Z}}$ follows from its orthonormality.

As in Supplement B.3, the $\lambda(n)$ term indicates the strength of the ridge penalty and is allowed to scale sublinearly in n. Again, identification is purely from this ridge penalty, because $\hat{\boldsymbol{Z}}$ is merely a linear transformation of $\boldsymbol{A}$ and thus the matrix is non-invertible without regularization.

Proposition 6. *(Asymptotic Bias of the Flexible Penalized Deconfounder under Additively Separable Confounding.)*

For all data-generating processes containing a linear factor model and additively separable confounding, the asymptotic bias of the flexible ridge deconfounder is given by

$$\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\beta}}^{\text{nonlinear}} - \boldsymbol{\beta} = -\boldsymbol{Q}_{1:k}\,\text{diag}_j\left(\frac{1}{\sigma^2+\Lambda_j+1}\right)\boldsymbol{Q}_{1:k}^\top\boldsymbol{\beta} + \boldsymbol{Q}_{1:k}\,\text{diag}_j\left(\frac{\Lambda_j}{\sigma^2+\Lambda_j+1}\right)\boldsymbol{Q}_{1:k}^\top\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\gamma},$$

where $\boldsymbol{Q}$ and $\boldsymbol{\Lambda} = [\Lambda_1, \ldots, \Lambda_k, 0, \ldots]$ are respectively eigenvectors and eigenvalues obtained from decomposition of $\boldsymbol{\theta}^\top\boldsymbol{\theta}$. Under strong infinite confounding,

$$\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\beta}}^{\text{nonlinear}} = \boldsymbol{\beta}.$$

We briefly offer intuition for the form of this bias before proceeding to the proof. The bias expressions in Proposition 6 are identical to those of Proposition 2, though the interpretation diverges slightly due to the flexible nature of the confounding function, $g_Y(\boldsymbol{Z})$. The term $\boldsymbol{\gamma}$ represents the portion of the confounding due to the linear trend in $g_Y(\boldsymbol{Z})$, which induces bias as described above. In contrast, $\boldsymbol{\delta}$ represents the nonlinear portion of the confounding that remains after eliminating the main linear trend. Because this part of $g_Y(\boldsymbol{Z})$ is by construction orthogonal to $\boldsymbol{Z}$ (and therefore to $\boldsymbol{A}$, due to the linear nature of the factor model) it cannot induce bias in $\hat{\boldsymbol{\beta}}^{\text{nonlinear}}$.

Proof of Proposition 6. In what follows, we will relate the asymptotic behavior of the flexible penalized deconfounder to the eigendecomposition $\boldsymbol{\theta}^\top\boldsymbol{\theta} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top = \boldsymbol{Q}_{1:k}\boldsymbol{\Lambda}_{1:k}\boldsymbol{Q}_{1:k}^\top$. To do so, we will rely on the singular value decompositions of $\boldsymbol{A}$, $[\boldsymbol{A}, \hat{\boldsymbol{Z}}]$, and $[\boldsymbol{A}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{W}}]$. For this section only, we respectively denote these as $\boldsymbol{A} = \boldsymbol{U}_A\boldsymbol{D}_A\boldsymbol{V}_A^\top$, $[\boldsymbol{A}, \hat{\boldsymbol{Z}}] = \boldsymbol{U}_{AZ}\boldsymbol{D}_{AZ}\boldsymbol{V}_{AZ}^\top$, and $[\boldsymbol{A}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{W}}] = \boldsymbol{U}_{AZW}\boldsymbol{D}_{AZW}\boldsymbol{V}_{AZW}^\top$. Lemma 6 characterizes the relationship between the first two; we now describe the latter.

For any n, the singular value decomposition $[\boldsymbol{A}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{W}}] = \boldsymbol{U}_{AZW}\boldsymbol{D}_{AZW}\boldsymbol{V}_{AZW}^{\top}$ can be seen to obey

$$\boldsymbol{U}_{AZW} = \left[\frac{1}{\sqrt{n}}\hat{\boldsymbol{Z}},\ \boldsymbol{U}_{A,(k+1):m},\ *,\ \frac{1}{\sqrt{n}}\hat{\boldsymbol{W}}\right]$$

$$\boldsymbol{D}_{AZW} = \begin{bmatrix} \left(\boldsymbol{D}_{A,1:k}^2 + n\boldsymbol{I}\right)^{\frac{1}{2}}, & \mathbf{0}, & \mathbf{0}, & \mathbf{0} \\ \mathbf{0}, & \boldsymbol{D}_{A,(k+1):m}, & \mathbf{0}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{0}, & \mathbf{0}, & \mathbf{0} \\ \mathbf{0}, & \mathbf{0}, & \mathbf{0}, & \sqrt{n}\boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{V}_{AZW}^{\top} = \begin{bmatrix} \left(\boldsymbol{D}_{A,1:k}^2 + n\boldsymbol{I}\right)^{-\frac{1}{2}}\boldsymbol{D}_{A,1:k}\boldsymbol{V}_{A,1:k}^{\top}, & \sqrt{n}\left(\boldsymbol{D}_{A,1:k}^2 + n\boldsymbol{I}\right)^{-\frac{1}{2}}, & \mathbf{0} \\ \boldsymbol{V}_{A,(k+1):m}^{\top}, & \mathbf{0}, & \mathbf{0} \\ *, & *, & * \\ \mathbf{0}, & \mathbf{0}, & \boldsymbol{I} \end{bmatrix},$$

where $*$ indicates irrelevant normalizing columns in $\boldsymbol{U}_{AZW}$ and $\boldsymbol{V}_{AZW}$. The above is due to Lemma 6 for the first $k+m$ columns. The behavior of the trailing columns follows from the fact that $\hat{\boldsymbol{W}}$ is normalized and orthogonal to $\hat{\boldsymbol{Z}}$ (and therefore to $\boldsymbol{A}$) by construction, and therefore remains invariant in the decomposition.

We now substitute the singular value decomposition of $[\boldsymbol{A}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{W}}]$ into the ridge estimator.

$$\begin{aligned} \begin{bmatrix} \hat{\boldsymbol{\beta}}^{\text{nonlinear}} \\ \hat{\boldsymbol{\gamma}}^{\text{nonlinear}} \\ \hat{\boldsymbol{\delta}}^{\text{nonlinear}} \end{bmatrix} &= \left([\boldsymbol{A},\ \hat{\boldsymbol{Z}},\ \hat{\boldsymbol{W}}]^{\top}[\boldsymbol{A},\ \hat{\boldsymbol{Z}},\ \hat{\boldsymbol{W}}] + \lambda(n)\boldsymbol{I}\right)^{-1}[\boldsymbol{A},\ \hat{\boldsymbol{Z}},\ \hat{\boldsymbol{W}}]^{\top}\boldsymbol{Y} \\ &= \boldsymbol{V}_{AZW}\left(\boldsymbol{D}_{AZW}^2 + \lambda(n)\boldsymbol{I}\right)^{-1}\boldsymbol{D}_{AZW}\boldsymbol{U}_{AZW}^{\top}\boldsymbol{Y} \end{aligned}$$

Eliminating dimensions with zero singular values and subsetting to $\hat{\boldsymbol{\beta}}^{\text{nonlinear}}$, we obtain

$$= \boldsymbol{V}_A\boldsymbol{D}_A\left(\boldsymbol{D}_A^2 + \lambda(n)\boldsymbol{I} + n \cdot \text{diag}_j 1\{j \leq k\}\right)^{-1}\boldsymbol{U}_A^{\top}\boldsymbol{Y}$$

and go to (20) in the proof of Proposition 2. □

B.5 Bias of the White-noised Deconfounder Estimator

In one of the tutorial simulations in Wang (2019), gaussian noise is added to the substitute confounder to render it estimable. This simulation and our reanalysis is discussed in Supplement C.3.1. Here we prove properties of this general strategy.

For convenience, we reiterate the data-generating process and white-noised deconfounder estimation procedure here. As before, we consider n observations drawn i.i.d. as follows.

$$\begin{aligned}
\underset{n\times k}{\boldsymbol{Z}} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{\nu}} &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{A}} &= \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\nu}_F \\
\underset{n\times 1}{\boldsymbol{\epsilon}} &\sim \mathcal{N}(\boldsymbol{0}, \omega^2) \\
\underset{n\times 1}{\boldsymbol{Y}} &= \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

The white-noised deconfounder estimator (1) takes the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$; (2) extracts the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$ and accompanying $\hat{\boldsymbol{\theta}} \equiv \frac{1}{\sqrt{n}}\boldsymbol{D}_{1:k}\boldsymbol{V}_{1:k}^\top$; adds noise $\boldsymbol{S} \sim \mathcal{N}(\boldsymbol{0}, \psi^2\boldsymbol{I})$ to $\hat{\boldsymbol{Z}}$ to break perfect collinearity with $\boldsymbol{A}$; and (4) estimates effects by computing

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{\text{wn}} \\ \hat{\boldsymbol{\gamma}}^{\text{wn}} \end{bmatrix} \equiv \left(\left[\boldsymbol{A}, \hat{\boldsymbol{Z}} + \boldsymbol{S}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}} + \boldsymbol{S}\right] \right)^{-1} \left[\boldsymbol{A}, \hat{\boldsymbol{Z}} + \boldsymbol{S}\right]^\top \boldsymbol{Y}. \tag{24}$$

We now restate Proposition B.5 before proceeding to the proof.

Proposition B.5. *(Asymptotic Bias of the White-noised Deconfounder.)*
Consider n observations drawn from a data-generating process with k unobserved confounders, $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$; $m \geq k$ observed treatments, $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{Z}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$; and outcome $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}, \omega^2)$. The white-noised deconfounder estimator, is

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{\text{wn}} \\ \hat{\boldsymbol{\gamma}}^{\text{wn}} \end{bmatrix} \equiv \left(\left[\boldsymbol{A}, \hat{\boldsymbol{Z}} + \boldsymbol{S}\right]^\top \left[\boldsymbol{A}, \hat{\boldsymbol{Z}} + \boldsymbol{S}\right] \right)^{-1} \left[\boldsymbol{A}, \hat{\boldsymbol{Z}} + \boldsymbol{S}\right] \boldsymbol{Y},$$

where $\hat{\boldsymbol{Z}}$ is obtained by taking the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ and extracting the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$; the addition of white noise, $\boldsymbol{S} \sim \mathcal{N}(0, \psi^2\boldsymbol{I})$, makes this regression estimable. The asymptotic bias of this estimator is given by

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{wn}} - \boldsymbol{\beta} = \left\{ \boldsymbol{\theta}^\top \left[\boldsymbol{I} - \frac{\sigma^2}{\psi^2}(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \right] \boldsymbol{\theta} + \frac{\sigma^2}{\psi^2}(1+\psi^2)\boldsymbol{I} \right\}^{-1} \boldsymbol{\theta}^\top\boldsymbol{\gamma},$$

and under strong infinite confounding,

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{wn}} - \boldsymbol{\beta} = \left[\boldsymbol{\theta}^\top\boldsymbol{\theta} + \frac{\sigma^2}{\psi^2}(1+\psi^2)\boldsymbol{I} \right]^{-1} \boldsymbol{\theta}^\top\boldsymbol{\gamma}$$

Proof of Proposition B.5.

After subsetting (24) to the treatment effects, the estimator can be rewritten as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\text{wn}} &= (\boldsymbol{A}^\top\boldsymbol{M}^{\text{wn}}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{M}^{\text{wn}}\boldsymbol{Y}, \quad \text{where} \\
\boldsymbol{M}^{\text{wn}} &\equiv \boldsymbol{I} - (\hat{\boldsymbol{Z}} + \boldsymbol{S}) \left[(\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top (\hat{\boldsymbol{Z}} + \boldsymbol{S}) \right]^{-1} (\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top.
\end{aligned}$$

Note that

$$\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\beta}} &= \underset{n\to\infty}{p\text{-lim}}(\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{Y} \\
&= \underset{n\to\infty}{p\text{-lim}}(\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}(\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta} + \underset{n\to\infty}{p\text{-lim}}(\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{Z}\boldsymbol{\gamma}.
\end{aligned}$$

We will proceed by first examining $p\text{-lim}_{n\to\infty} \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{A}$ and then $p\text{-lim}_{n\to\infty} \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{Z}$.

$$\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{A} &= \underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \left\{\boldsymbol{I} - (\hat{\boldsymbol{Z}} + \boldsymbol{S})\left[(\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top(\hat{\boldsymbol{Z}} + \boldsymbol{S})\right]^{-1}(\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top\right\}\boldsymbol{A} \\
&= \underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \left[\boldsymbol{I} - \frac{1}{n(1+\psi^2)}(\hat{\boldsymbol{Z}} + \boldsymbol{S})(\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top\right]\boldsymbol{A} \\
&= \underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \boldsymbol{A} - \frac{1}{1+\psi^2}\left(\frac{1}{n}\boldsymbol{A}^\top \hat{\boldsymbol{Z}}\right)\left(\frac{1}{n}\hat{\boldsymbol{Z}}^\top \boldsymbol{A}\right) \\
&= \underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \boldsymbol{A} - \frac{1}{1+\psi^2}\hat{\boldsymbol{\theta}}^\top\hat{\boldsymbol{\theta}} \\
&= \frac{\psi^2}{1+\psi^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} - \frac{\sigma^2}{1+\psi^2}\boldsymbol{\theta}^\top\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\right)^{-1}\boldsymbol{\theta} \quad \text{by Lemma 4} \\
&= \frac{\psi^2}{1+\psi^2}\boldsymbol{\theta}^\top\left[\boldsymbol{I} - \frac{\sigma^2}{\psi^2}(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\right]\boldsymbol{\theta} + \sigma^2\boldsymbol{I} \\
\underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}*}\boldsymbol{Z} &= \underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\boldsymbol{A}^\top \left\{\boldsymbol{I} - (\hat{\boldsymbol{Z}} + \boldsymbol{S})\left[(\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top(\hat{\boldsymbol{Z}} + \boldsymbol{S})\right]^{-1}(\hat{\boldsymbol{Z}} + \boldsymbol{S})^\top\right\}\boldsymbol{Z} \\
&= \underset{n\to\infty}{p\text{-lim}}\,\boldsymbol{\theta}^\top - \frac{1}{1+\psi^2}\left(\frac{1}{n}\boldsymbol{A}^\top \hat{\boldsymbol{Z}}\right)\left(\frac{1}{n}\hat{\boldsymbol{Z}}^\top \boldsymbol{Z}\right) \\
&= \underset{n\to\infty}{p\text{-lim}}\,\boldsymbol{\theta}^\top - \frac{1}{1+\psi^2}\boldsymbol{\theta}^\top(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\left[\frac{1}{n}\boldsymbol{A}^\top(\boldsymbol{A} - \boldsymbol{\nu})\right]\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1} \\
&= \underset{n\to\infty}{p\text{-lim}}\,\boldsymbol{\theta}^\top - \frac{1}{1+\psi^2}\boldsymbol{\theta}^\top(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\boldsymbol{\theta}^\top\boldsymbol{\theta}\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}
\end{aligned}$$

By Lemma 3,

$$\begin{aligned}
&= \underset{n\to\infty}{p\text{-lim}}\,\boldsymbol{\theta}^\top - \frac{1}{1+\psi^2}\boldsymbol{\theta}^\top\boldsymbol{R}\boldsymbol{\Lambda}_{1:k}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1} \\
&\qquad\qquad \cdot \hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top\left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{-\frac{1}{2}}\boldsymbol{\Lambda}_{1:k}^{\frac{1}{2}}\boldsymbol{R}^\top \\
&= \frac{\psi^2}{1+\psi^2}\boldsymbol{\theta}^\top \\
\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\beta}}^{\text{wn}} - \boldsymbol{\beta} &= \underset{n\to\infty}{p\text{-lim}}(\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{A})^{-1}\boldsymbol{A}^\top \boldsymbol{M}^{\text{wn}}\boldsymbol{Z}\boldsymbol{\gamma} \\
&= \left\{\boldsymbol{\theta}^\top\left[\boldsymbol{I} - \frac{\sigma^2}{\psi^2}(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\right]\boldsymbol{\theta} + \frac{\sigma^2}{\psi^2}(1+\psi^2)\boldsymbol{I}\right\}^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma} \quad \text{and} \\
\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\beta}}^{\text{wn}} - \boldsymbol{\beta} &= \left[\boldsymbol{\theta}^\top\boldsymbol{\theta} + \frac{\sigma^2}{\psi^2}(1+\psi^2)\boldsymbol{I}\right]^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma} \quad \square
\end{aligned}$$

B.6 Bias of the Posterior-mean Deconfounder Estimator

For convenience, we reiterate the data-generating process and posterior-mean deconfounder estimation procedure here. As before, we consider n observations drawn i.i.d. as follows.

$$\begin{aligned}
\underset{n\times k}{\boldsymbol{Z}} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{\nu}} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \\
\underset{n\times m}{\boldsymbol{A}} &= \boldsymbol{Z}\boldsymbol{\theta} + \boldsymbol{\nu}_F \\
\underset{n\times 1}{\boldsymbol{\epsilon}} &\sim \mathcal{N}(\mathbf{0}, \omega^2) \\
\underset{n\times 1}{\boldsymbol{Y}} &= \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

The posterior-mean deconfounder estimator is an approximate Bayesian procedure, in the sense that it obtains an estimate for the effects $\boldsymbol{\beta}$ by integrating over an approximation to the full joint posterior, $f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$, as follows.

First, the full posterior is factorized as $f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z}) f(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$. Then, $f(\boldsymbol{z}|\boldsymbol{A})$ is obtained by a Bayesian principal components analysis of $\boldsymbol{A}$ alone—i.e., ignoring information from $\boldsymbol{Y}$—and used as an approximation to $f(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$. A Bayesian linear regression of $\boldsymbol{Y}$ on $\boldsymbol{A}$ and $\boldsymbol{z}$ is used to obtain the conditional posterior $f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z})$, and finally $\boldsymbol{z}$ is integrated out. The posterior-mean deconfounder estimator is thus

$$\left[\hat{\boldsymbol{\beta}}^{\mathrm{pm}\top}, \hat{\boldsymbol{\gamma}}^{\mathrm{pm}\top}\right]^\top \equiv \int f(\boldsymbol{z}|\boldsymbol{A})\ \mathbb{E}\left\{\left[\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top\right]^\top \mid \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z}\right\} \mathrm{d}\boldsymbol{z}\,.$$

We leave priors for all parameters unspecified; by the Bernstein-von Mises theorem, our results are invariant to the choice of any prior with positive density on the true parameters.

We now restate Proposition 3 before proceeding to the proof.

Proposition 3. *(Asymptotic Bias of the Posterior-Mean Deconfounder.)*
Consider n observations drawn from a data-generating process with k unobserved confounders, $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$; $m \geq k$ observed treatments, $\boldsymbol{A} \sim \mathcal{N}(\boldsymbol{Z}\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$; and outcome $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma}, \omega^2)$, following WB. The posterior-mean deconfounder estimator is

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}^{\mathrm{pm}} \\ \hat{\boldsymbol{\gamma}}^{\mathrm{pm}} \end{bmatrix} \equiv \int \left([\boldsymbol{A}, \boldsymbol{z}]^\top [\boldsymbol{A}, \boldsymbol{z}]\right)^{-1} [\boldsymbol{A}, \boldsymbol{z}]\, \boldsymbol{Y}\ f(\boldsymbol{z}|\boldsymbol{A})\, \mathrm{d}\boldsymbol{z}\,,$$

where $f(\boldsymbol{z}|\boldsymbol{A})$ is a posterior obtained from Bayesian principal component analysis.[a] The asymptotic bias of this estimator is given by

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\mathrm{pm}} - \boldsymbol{\beta} = (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma},$$

and under strong infinite confounding,

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\mathrm{pm}} - \boldsymbol{\beta} = \mathbf{0}$$

a. While the regression cannot be estimated when $\boldsymbol{z} = \mathbb{E}[\boldsymbol{Z}|\boldsymbol{A}]$, it is almost surely estimable for samples $\boldsymbol{z}^* \sim f(\boldsymbol{z}|\boldsymbol{A})$ due to posterior uncertainty, which eliminates perfect collinearity with $\boldsymbol{A}$. The posterior-mean implementation of WB evaluates the integral by Monte Carlo methods and thus is able to compute the regression coefficients for each sample.

Proof of Proposition 3.

Under the Bayesian principal components generative model,

$$\begin{bmatrix} \boldsymbol{Z}_i \\ \boldsymbol{A}_i \end{bmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{I}, & \boldsymbol{\theta} \\ \boldsymbol{\theta}^\top, & \boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} \end{bmatrix}\right)$$

and by properties of the multivariate normal,

$$f(\boldsymbol{z}_i|\boldsymbol{A}_i, \boldsymbol{\theta}, \sigma^2) = \phi\left(\boldsymbol{z}_i;\ \boldsymbol{\theta}(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{A}_i,\ \boldsymbol{I} - \boldsymbol{\theta}(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top\right).$$

We will decompose the conditional posterior over confounders as $f(\boldsymbol{z}|\boldsymbol{A}) = f(\boldsymbol{z}|\boldsymbol{\theta}, \sigma^2, \boldsymbol{A})f(\boldsymbol{\theta}, \sigma^2|\boldsymbol{A})$. A sample $\boldsymbol{z}_i^*$ can be drawn from the Bayesian principal component posterior by first sampling $\boldsymbol{\theta}^*$ and σ^{*2} from $f(\boldsymbol{\theta}, \sigma^2|\boldsymbol{A})$, deterministically constructing $\hat{\boldsymbol{Z}}_i^* \equiv \mathbb{E}[\boldsymbol{Z}_i|\boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{A}_i] = \boldsymbol{\theta}^*(\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* + \sigma^{*2}\boldsymbol{I})^{-1}\boldsymbol{A}_i$, sampling $\boldsymbol{s}_i^*$ from $f(\boldsymbol{s}_i|\boldsymbol{\theta}^*, \sigma^{*2}) = \phi\left(\boldsymbol{s}_i;\ \boldsymbol{0},\ \boldsymbol{I} - \boldsymbol{\theta}^*(\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* + \sigma^{*2}\boldsymbol{I})^{-1}\boldsymbol{\theta}^{*\top}\right)$, and deterministically taking $\boldsymbol{z}_i^* = \hat{\boldsymbol{Z}}_i^* + \boldsymbol{s}_i^*$.

We can then rewrite

$$\left[\hat{\boldsymbol{\beta}}^{\text{pm}\top}, \hat{\boldsymbol{\gamma}}^{\text{pm}\top}\right]^\top = \int f(\boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{s}^*|\boldsymbol{A})\ \mathbb{E}\left\{\left[\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top\right]^\top \mid \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{s}^*\right\} \mathrm{d}\boldsymbol{\theta}^*\, \mathrm{d}\sigma^{*2}\, \mathrm{d}\boldsymbol{s}^*$$

where

$$\begin{aligned}
&\mathbb{E}\left\{\left[\boldsymbol{\beta}^{*\top}, \boldsymbol{\gamma}^{*\top}\right]^\top \mid \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{s}^*\right\} \\
&\quad = \left(\left[\boldsymbol{A},\ (\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\right]^\top \left[\boldsymbol{A},\ (\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\right]\right)^{-1} \left[\boldsymbol{A},\ (\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\right]^\top \boldsymbol{Y}. \qquad (25)
\end{aligned}$$

Note that the posterior $f(\boldsymbol{\theta}, \sigma^2|\boldsymbol{A})$ concentrates on true σ^2 and $\boldsymbol{\theta}$ (up to a rotation). Thus, candidate $\boldsymbol{\theta}^*$ and σ^{*2} values that fail to satisfy $\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* = \boldsymbol{\theta}^\top\boldsymbol{\theta}$ and $\sigma^{*2} = \sigma^2$ grow vanishingly unlikely as n grows large. We examine the asymptotic behavior of the conditional estimator, (25), in this region and show that the bias is constant. Thus, the estimator remains asymptotically biased after integrating over all possible rotations of $\boldsymbol{\theta}^*$.

After subsetting (25) to the treatment effects, the conditional estimator can be rewritten as $\hat{\boldsymbol{\beta}}^* \equiv \mathbb{E}\left[\boldsymbol{\beta}^* \mid \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{s}^*\right] = (\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{Y}$, where $\boldsymbol{M}^{\text{pm}*}$ denotes the conditional annihilator, $\boldsymbol{I} - (\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\left[(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\right]^{-1}(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top$. Note that

$$\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^* &= \underset{n\to\infty}{p\text{-lim}} (\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{Y} \\
&= \underset{n\to\infty}{p\text{-lim}} (\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}(\boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{\beta} + \underset{n\to\infty}{p\text{-lim}} (\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{M}^{\text{pm}*}\boldsymbol{Z}\boldsymbol{\gamma}.
\end{aligned}$$

for any $\boldsymbol{\theta}^*$ and σ^{*2}. We will proceed by first examining $p\text{-lim}_{n\to\infty} \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{pm}*}\boldsymbol{A}$ and then $p\text{-lim}_{n\to\infty} \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{pm}*}\boldsymbol{Z}$.

$$
\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{pm}*}\boldsymbol{A} &= \underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \left\{\boldsymbol{I} - (\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\left[(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\right]^{-1}(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top\right\}\boldsymbol{A} \\
&= \underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \left[\boldsymbol{I} - \frac{1}{n}(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top\right]\boldsymbol{A} \\
&= \underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{A} - \left(\frac{1}{n}\boldsymbol{A}^\top \hat{\boldsymbol{Z}}^*\right)\left(\frac{1}{n}\hat{\boldsymbol{Z}}^{*\top}\boldsymbol{A}\right) \\
&= \underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{A} - (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})(\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* + \sigma^{*2}\boldsymbol{I})^{-1}\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* \\
&\qquad\qquad \cdot (\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* + \sigma^{*2}\boldsymbol{I})^{-1}(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I}) \\
&= \boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} - \boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* \\
&= \sigma^2\boldsymbol{I} \\
\underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \boldsymbol{M}^{\text{pm}*}\boldsymbol{Z} &= \underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}\boldsymbol{A}^\top \left\{\boldsymbol{I} - (\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\left[(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)\right]^{-1}(\hat{\boldsymbol{Z}}^* + \boldsymbol{s}^*)^\top\right\}\boldsymbol{Z} \\
&= \underset{n\to\infty}{p\text{-lim}}\, \frac{1}{n}(\boldsymbol{\theta}^\top\boldsymbol{Z}^\top + \boldsymbol{\nu}^\top)\boldsymbol{Z} - (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})(\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* + \sigma^{*2}\boldsymbol{I})^{-1}\boldsymbol{\theta}^{*\top}\left(\frac{1}{n}\hat{\boldsymbol{Z}}^{*\top}\boldsymbol{Z}\right) \\
&= \underset{n\to\infty}{p\text{-lim}}\, \boldsymbol{\theta}^\top - \boldsymbol{\theta}^{*\top}\left(\frac{1}{n}\boldsymbol{\theta}^*(\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^* + \sigma^{*2}\boldsymbol{I})^{-1}\boldsymbol{A}^\top\boldsymbol{Z}\right) \\
&= \sigma^2(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top
\end{aligned}
$$

Note that both expressions depend only on $\boldsymbol{\theta}^{*\top}\boldsymbol{\theta}^*$, not $\boldsymbol{\theta}^*$ alone. Thus, the bias is constant over the entire asymptotic posterior (i.e., all rotations) of $\boldsymbol{\theta}^*$.

$$
\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{pm}} - \boldsymbol{\beta} &= \underset{n\to\infty}{p\text{-lim}} \int f(\boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{s}^*|\boldsymbol{A})\ \mathbb{E}\left[\boldsymbol{\beta}^* - \boldsymbol{\beta}|\ \boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{\theta}^*, \sigma^{*2}, \boldsymbol{s}^*\right] \mathrm{d}\boldsymbol{\theta}^*\, \mathrm{d}\sigma^{*2}\, \mathrm{d}\boldsymbol{s}^* \\
&= (\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\theta}^\top\boldsymbol{\gamma}
\end{aligned}
$$

and under strong infinite confounding,

$$
\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}^{\text{pm}} - \boldsymbol{\beta} = \boldsymbol{0} \quad \square
$$

B.7 Bias of the Subset Deconfounder

For convenience, we reiterate the data-generating process and subset deconfounder estimation procedure here. We will suppose, without loss of generality, that the m causes, $\boldsymbol{A}$, are divided into m_F focal causes of interest, the column subset and m_N nonfocal causes, $\boldsymbol{A}_N$.

As before, we consider n observations drawn i.i.d. as follows.

$$\begin{aligned}
\underset{n\times k}{\boldsymbol{Z}} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \\
\underset{n\times m_F}{\boldsymbol{\nu}_F} &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m_N}{\boldsymbol{\nu}_N} &\sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}) \\
\underset{n\times m_F}{\boldsymbol{A}_F} &= \boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F \\
\underset{n\times m_N}{\boldsymbol{A}_N} &= \boldsymbol{Z}\boldsymbol{\theta}_N + \boldsymbol{\nu}_N \\
\underset{n\times 1}{\boldsymbol{\epsilon}} &\sim \mathcal{N}(\boldsymbol{0}, \omega^2) \\
\underset{n\times 1}{\boldsymbol{Y}} &= \boldsymbol{A}_F\boldsymbol{\beta}_F + \boldsymbol{A}_N\boldsymbol{\beta}_N + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

The subset deconfounder estimator (1) takes the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$; (2) extracts the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$ and accompanying $\hat{\boldsymbol{\theta}} \equiv \frac{1}{\sqrt{n}}\boldsymbol{D}_{1:k}\boldsymbol{V}_{1:k}^\top$; and (3) estimates the focal effects by computing

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_F^{\text{subset}} \\ \hat{\boldsymbol{\gamma}}^{\text{subset}} \end{bmatrix} \equiv \left(\left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right] \right)^{-1} \left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y}$$

and discarding $\hat{\boldsymbol{\gamma}}^{\text{subset}}$.

We now restate Proposition 4 before proceeding to the proof.

Proposition 4. (Asymptotic Bias of the Subset Deconfounder.)
The subset deconfounder estimator, based on Theorem 7 from WB, is

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_F^{\text{subset}} \\ \hat{\boldsymbol{\gamma}}^{\text{subset}} \end{bmatrix} \equiv \left(\left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]^\top \left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right] \right)^{-1} \left[\boldsymbol{A}_F, \hat{\boldsymbol{Z}}\right]^\top \boldsymbol{Y}. \tag{26}$$

where the column subsets $\boldsymbol{A}_F$ and $\boldsymbol{A}_N$ respectively partition $\boldsymbol{A}$ into a finite number of focal causes of interest and non-focal causes. The substitute confounder, $\hat{\boldsymbol{Z}}$, is obtained by taking the singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ and extracting the first k components, $\hat{\boldsymbol{Z}} \equiv \sqrt{n}\boldsymbol{U}_{1:k}$. Under the linear-linear model, the asymptotic bias of this estimator is given by

$$\underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{subset}} - \boldsymbol{\beta}_F = -\left(\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right)^{-1} \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N,$$

with $\boldsymbol{\theta}_F$ and $\boldsymbol{\theta}_N$ indicating the column subsets of $\boldsymbol{\theta}$ corresponding to $\boldsymbol{A}_F$ and $\boldsymbol{A}_N$, respectively. The subset deconfounder is unbiased for $\boldsymbol{\beta}_F$ (i) if $\boldsymbol{\theta}_F = \mathbf{0}$, (ii) if $\lim_{m\to\infty} \boldsymbol{\theta}_N\boldsymbol{\beta}_N = \mathbf{0}$ and $\lim_{m\to\infty}\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]^{-1}$ is convergent or (iii) if both strong infinite confounding holds and $(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N$ goes to $\mathbf{0}$, as $m \to \infty$. If one of these additional conditions hold,

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{subset}} - \boldsymbol{\beta}_F = \mathbf{0}$$

Proof of Proposition 4. By the Frisch-Waugh-Lovell theorem, $\hat{\boldsymbol{\beta}}_F^{\text{subset}}$ can be re-expressed in terms of the portion of $\boldsymbol{A}_F$ not explained by $\hat{\boldsymbol{Z}}$. We denote the residualized focal treatments as $\tilde{\boldsymbol{A}}_F^{\text{subset}} = \boldsymbol{A}_F - \hat{\boldsymbol{A}}_F^{\text{subset}} = \boldsymbol{A}_F^{\text{subset}} - \hat{\boldsymbol{Z}}\hat{\boldsymbol{\theta}}_F = \boldsymbol{U}_{(k+1):m}\boldsymbol{D}_{(k+1):m}\boldsymbol{V}_{(k+1):m,F}^\top$. The subset deconfounder estimator is then rewritten as follows:

$$\hat{\boldsymbol{\beta}}_F^{\text{subset}} = \left(\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\tilde{\boldsymbol{A}}_F^{\text{subset}}\right)^{-1}\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}}\boldsymbol{Y}$$

We now characterize the asymptotic bias of this estimator by examining the behavior of $\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\tilde{\boldsymbol{A}}_F^{\text{subset}}$ and $\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}}\boldsymbol{Y}$ in turn. Beginning with the residual variance of the focal causes,

$$\begin{aligned}
\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\tilde{\boldsymbol{A}}_F^{\text{subset}} &= \frac{1}{n}\left(\boldsymbol{A}_F^\top - \hat{\boldsymbol{A}}_F^{\text{subset}\top}\right)\left(\boldsymbol{A}_F - \hat{\boldsymbol{A}}_F^{\text{subset}}\right)\\
&= \frac{1}{n}\left(\boldsymbol{A}_F^\top\boldsymbol{A}_F + \hat{\boldsymbol{A}}_F^{\text{subset}\top}\hat{\boldsymbol{A}}_F^{\text{subset}} - \boldsymbol{A}_F^\top\hat{\boldsymbol{A}}_F^{\text{subset}} - \hat{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{A}_F\right)\\
&= \frac{1}{n}\left(\boldsymbol{\theta}_F^\top\boldsymbol{Z}^\top + \boldsymbol{\nu}_F^\top\right)(\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F) + \frac{1}{n}\hat{\boldsymbol{\theta}}_F^\top\hat{\boldsymbol{Z}}^\top\hat{\boldsymbol{Z}}\hat{\boldsymbol{\theta}}_F\\
&\quad - \frac{1}{n}\left(\boldsymbol{V}_{1:k,F}\boldsymbol{D}_{1:k}\boldsymbol{U}_{1:k}^\top + \boldsymbol{V}_{(k+1):m,F}\boldsymbol{D}_{(k+1):m}\boldsymbol{U}_{(k+1):m}^\top\right)\boldsymbol{U}_{1:k}\boldsymbol{D}_{1:k}\boldsymbol{V}_{1:k,F}^\top\\
&\quad - \frac{1}{n}\boldsymbol{V}_{1:k,F}\boldsymbol{D}_{1:k}\boldsymbol{U}_{1:k}^\top\left(\boldsymbol{U}_{1:k}\boldsymbol{D}_{1:k}\boldsymbol{V}_{1:k,F}^\top + \boldsymbol{U}_{(k+1):m}\boldsymbol{D}_{(k+1):m}\boldsymbol{V}_{(k+1):m,F}^\top\right)\\
&= \frac{1}{n}\left(\boldsymbol{\theta}_F^\top\boldsymbol{Z}^\top + \boldsymbol{\nu}_F^\top\right)(\boldsymbol{Z}\boldsymbol{\theta}_F + \boldsymbol{\nu}_F) + \hat{\boldsymbol{\theta}}_F^\top\hat{\boldsymbol{\theta}}_F - \frac{2}{n}\boldsymbol{V}_{1:k,F}\boldsymbol{D}_{1:k}^2\boldsymbol{V}_{1:k,F}^\top \qquad (27)
\end{aligned}$$

Taking limits,

$$\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\tilde{\boldsymbol{A}}_F^{\text{subset}} &= \boldsymbol{\theta}_F^\top\boldsymbol{\theta}_F + \sigma^2\boldsymbol{I} - \hat{\boldsymbol{\theta}}_F^\top\hat{\boldsymbol{\theta}}_F\\
&= \sigma^2\boldsymbol{I} - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F \quad \text{by Lemma 4, and} \qquad (28)\\
\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\tilde{\boldsymbol{A}}_F^{\text{subset}} &= \sigma^2\boldsymbol{I} \quad \text{under strong infinite confounding.} \qquad (29)
\end{aligned}$$

Turning to the residual covariance between the focal causes and the outcome,

$$\begin{aligned}
\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{Y} &= \frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\left(\boldsymbol{A}_F\boldsymbol{\beta}_F + \boldsymbol{A}_N\boldsymbol{\beta}_N + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}\right) \\
&= \frac{1}{n}\left[\left(\boldsymbol{A}_F^\top - \hat{\boldsymbol{A}}_F^{\text{subset}\top}\right)\boldsymbol{A}_F\boldsymbol{\beta}_F + \tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{A}_N\boldsymbol{\beta}_N + \tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{Z}\boldsymbol{\gamma} + \tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{\epsilon}\right] \\
&= \frac{1}{n}(\boldsymbol{A}_F^\top - \hat{\boldsymbol{A}}_F^{\text{subset}\top})\boldsymbol{A}_F\boldsymbol{\beta}_F + \frac{1}{n}\boldsymbol{V}_{(k+1):m,F}\boldsymbol{D}_{(k+1):m}\boldsymbol{U}_{(k+1):m}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}_N^\top\boldsymbol{\beta}_N \\
&\quad + \frac{1}{\sqrt{n}}\boldsymbol{V}_{(k+1):m,F}\boldsymbol{D}_{(k+1):m}\boldsymbol{U}_{(k+1):m}^\top\boldsymbol{U}_{1:k}\boldsymbol{\gamma} + \frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{Z}\boldsymbol{\gamma} + \frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{\epsilon} \\
&= \frac{1}{n}(\boldsymbol{A}_F^\top - \hat{\boldsymbol{A}}_F^{\text{subset}\top})\boldsymbol{A}_F\boldsymbol{\beta}_F + \frac{1}{n}\boldsymbol{V}_{(k+1):m,F}\boldsymbol{D}_{(k+1):m}^2\boldsymbol{V}_{(k+1):m,N}^\top\boldsymbol{\beta}_N \\
&\quad + \frac{1}{n}(\boldsymbol{\theta}_F^\top\boldsymbol{Z}^\top + \boldsymbol{\nu}_F^\top - \hat{\boldsymbol{\theta}}_F^\top\hat{\boldsymbol{Z}}^\top)\boldsymbol{Z}\boldsymbol{\gamma} + \frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{\epsilon}
\end{aligned}$$

We will proceed by reducing $(\boldsymbol{A}_F^\top - \hat{\boldsymbol{A}}_F^{\text{subset}\top})\boldsymbol{A}_F$ as above, substituting

$$\hat{\boldsymbol{Z}}^\top\boldsymbol{Z} = \left[(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\boldsymbol{A}^\top\right]\left[(\boldsymbol{A}-\boldsymbol{\nu})\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\right]$$

and invoking Lemmas 3 and 4.

$$\begin{aligned}
\underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{Y} &= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N \\
&\quad + \boldsymbol{\theta}_F^\top\boldsymbol{\gamma} - \underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\hat{\boldsymbol{\theta}}_F^\top(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\boldsymbol{A}^\top(\boldsymbol{A}-\boldsymbol{\nu})\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\gamma} \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N \\
&\quad + \boldsymbol{\theta}_F^\top\boldsymbol{\gamma} - \hat{\boldsymbol{\theta}}_F^\top(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^\top\boldsymbol{\theta} + \sigma^2\boldsymbol{I} - \sigma^2\boldsymbol{I}\right)\boldsymbol{\theta}^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\gamma} \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N \\
&\quad + \boldsymbol{\theta}_F^\top\boldsymbol{\gamma} - \hat{\boldsymbol{\theta}}_F^\top(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\boldsymbol{\theta}^\top\boldsymbol{\gamma} \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N \\
&\quad + \boldsymbol{\theta}_F^\top\boldsymbol{\gamma} - \hat{\boldsymbol{\theta}}_F^\top(\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top)^{-1}\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^\top\left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{-\frac{1}{2}}\boldsymbol{\Lambda}_{1:k}^{\frac{1}{2}}\boldsymbol{R}^\top\boldsymbol{\gamma} \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N \\
&\quad + \boldsymbol{\theta}_F^\top\boldsymbol{\gamma} - \boldsymbol{\theta}_F^\top\boldsymbol{R}\left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{\frac{1}{2}}\boldsymbol{\Lambda}_{1:k}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}_{1:k} + \sigma^2\boldsymbol{I}\right)^{-\frac{1}{2}}\boldsymbol{\Lambda}_{1:k}^{\frac{1}{2}}\boldsymbol{R}^\top\boldsymbol{\gamma} \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]\boldsymbol{\beta}_F - \sigma^2\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N \qquad (30)
\end{aligned}$$

and under strong infinite confounding,

$$\lim_{m\to\infty}\underset{n\to\infty}{p\text{-lim}}\,\frac{1}{n}\tilde{\boldsymbol{A}}_F^{\text{subset}\top}\boldsymbol{Y} = \sigma^2\boldsymbol{\beta}_F. \qquad (31)$$

Combining (28) and (30),

$$\underset{n\to\infty}{p\text{-lim}}\,\hat{\boldsymbol{\beta}}_F^{\text{subset}} = \boldsymbol{\beta}_F - \left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]^{-1}\boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N,$$

Consider the additional conditions. (i) If $\boldsymbol{\theta}_F = \mathbf{0}$ then convergence is immediate. (ii) If $\lim_{m\to\infty} \boldsymbol{\theta}_N\boldsymbol{\beta}_N$ and $\lim_{m\to\infty} \left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]^{-1}$ is convergent then $\lim_{m\to\infty}(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N = \mathbf{0}$ so the product of the limits is $\mathbf{0}$. (iii) If $\lim_{m\to\infty}\left(\boldsymbol{\theta}\boldsymbol{\theta}^\top\right)^{-1}\boldsymbol{\theta}_N\boldsymbol{\beta}_N = \mathbf{0}$ and strong infinite confounding holds then $\lim_{m\to\infty}\left[\boldsymbol{I} - \boldsymbol{\theta}_F^\top(\boldsymbol{\theta}\boldsymbol{\theta}^\top)^{-1}\boldsymbol{\theta}_F\right]^{-1} = \boldsymbol{I}$ so the bias term also goes to zero. Therefore, combining Equations (29) and (31) if one of these conditions holds yields

$$\lim_{m\to\infty} \underset{n\to\infty}{p\text{-lim}}\, \hat{\boldsymbol{\beta}}_F^{\text{subset}} = \boldsymbol{\beta}_F. \quad \square$$

Proposition 5 demonstrates that the naïve regression estimator is an unbiased estimator for focal treatments; it is a generalization of Proposition 1. The proof is given in Section B.2.

B.8 Subset Deconfounder Requires Assumptions about Treatment Effects

To see why strong infinite confounding is insufficient, consider a simple example. Using the linear-linear data-generating process, consider the case of $k = 1$. For the sake of this example, we will suppose that for each treatment j that $\theta_j = \bar{\theta}$ and that $\beta_j = \bar{\beta}$. This clearly satisfies strong infinite confounding, PCA will be a consistent estimator of the substitute confounder, and naïve regression will be a consistent estimator of the treatment effects. But this is not the case for the subset deconfounder. Using Proposition 4, the bias for arbitrary focal treatment j as $m \to \infty$ is given by:

$$\begin{aligned}\lim_{m\to\infty} \mathbb{E}[\hat{\beta}_j] - \beta_j &= \lim_{m\to\infty} \left(1 - \frac{\bar{\theta}^2}{m\bar{\theta}^2}\right)^{-1} \frac{(m-1)\bar{\theta}^2\bar{\beta}}{m\theta^2} \\ &= \bar{\beta}\end{aligned}$$

So long as $\bar{\beta} \neq 0$ then the bias is non-zero, regardless of how many treatments are present. The intuition is that as we add more treatments we are having two countervailing effects on our estimator. Additional treatments are giving us a better estimate of the substitute confounder, which is reducing the correlation between the focal and non-focal treatments, which on its own would reduce bias. But at the same time, we're adding additional omitted variable bias. That additional omitted variable bias causes the subset deconfounder's bias to not decrease as the treatments are added.

Subset Deconfounder as Regularization Connections between the subset deconfounder and naïve regression can be seen through the following straightforward argument. Consider first naïve regression. By the Frisch-Waugh-Lovell theorem, for any treatment $\boldsymbol{A}_j$, its estimated effect is related to the residualized treatment, $\tilde{\boldsymbol{A}}_j^{\text{naïve}} = \boldsymbol{A}_j - \hat{\boldsymbol{A}}_j^{\text{naïve}}$, where $\hat{\boldsymbol{A}}_j^{\text{naïve}} = \left(\boldsymbol{A}_{\setminus j}^\top\boldsymbol{A}_{\setminus j}\right)^{-1}\boldsymbol{A}_{\setminus j}^\top\boldsymbol{A}_j$ is the part of $\boldsymbol{A}_j$ that can be predicted from the other treatments, $\boldsymbol{A}_{\setminus j}$. Specifically, $\hat{\beta}_j^{\text{naïve}} = \frac{\text{Cov}(\boldsymbol{Y}, \tilde{\boldsymbol{A}}_j^{\text{naïve}})}{\text{Var}(\tilde{\boldsymbol{A}}_j^{\text{naïve}})}$. Denoting the SVD of $\boldsymbol{A}_{\setminus j}$ as $\boldsymbol{U}_{\setminus j}\boldsymbol{D}_{\setminus j}\boldsymbol{V}_{\setminus j}^\top$, then $\hat{\boldsymbol{A}}_j^{\text{naïve}} = \boldsymbol{U}_{\setminus j}\boldsymbol{U}_{\setminus j}^\top\boldsymbol{A}_j$. As $m, n \to \infty$ then under linear-linear confounding this approaches $\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{A}_j$.

Now consider the subset deconfounder. Also by Frisch-Waugh-Lovell, for any single treatment $\boldsymbol{A}_j$, its estimated effect is $\hat{\beta}_j^{\text{subset}} = \frac{\text{Cov}(\boldsymbol{Y}, \tilde{\boldsymbol{A}}_j^{\text{subset}})}{\text{Var}(\tilde{\boldsymbol{A}}_j^{\text{subset}})}$ and $\tilde{\boldsymbol{A}}_j^{\text{subset}} = \boldsymbol{A}_j^{\text{subset}} - \hat{\boldsymbol{A}}_j^{\text{subset}}$. Denoting the SVD of $\boldsymbol{A}$ as $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, it can be seen that $\hat{\boldsymbol{A}}_j^{\text{subset}} = \boldsymbol{U}_{1:k}\boldsymbol{U}_{1:k}^\top\boldsymbol{A}_j$.

This makes clear that the naïve regression is adjusting along the same eigenvectors as the subset deconfounder. Further, it shows that the subset deconfounder is performing the same regression as the naïve regression but with a particular form of regularization. Specifically, the deconfounder imposes no penalty on the first k eigenvectors, but then regularizes by suppressing the influence of dimensions $k+1$ to m.

B.9 Convergence of Deconfounder and Naïve Estimators

In this section, we extend our previous results on asymptotic equivalence between the deconfounder and naïve analyses to a broad class of nonlinear-nonlinear data-generating processes. We consider factor models in which the distributions of $\boldsymbol{Z}_i$ and $\boldsymbol{A}_i|\boldsymbol{Z}_i$ have continuous density. Following the deconfounder papers, we analyze the case in which pinpointedness holds and this requires strong infinite confounding and infinite m.

We also follow the deconfounder papers in restricting attention to outcome models with constant treatment effects and finite variance. That is, we study outcome models satisfying $\mathbb{E}[Y_i|\boldsymbol{A},\boldsymbol{Z}] = \boldsymbol{A}_i^\top\boldsymbol{\beta} + g_Y(\boldsymbol{Z}_i)$, allowing for arbitrarily complex nonlinear confounding. This family is more restrictive than the outcome models considered in Section B.10 but nevertheless nests all data-generating processes studied in prior deconfounder papers and our paper.

For convenience, we restate Theorem 1 here.

Theorem 1. *(Deconfounder-Naïve Convergence under Strong Infinite Confounding.) Consider all data-generating processes in which (i) treatments and confounders are drawn from a factor model with continuous density $f(\boldsymbol{z},\boldsymbol{a})$; (ii) $\boldsymbol{Z}$ is pinpointed, i.e. the conditional entropy $H(\boldsymbol{Z}_i|\boldsymbol{A}_i) = 0$; and (iii) the outcome model is of the form $\mathbb{E}[Y_i|\boldsymbol{A}_i,\boldsymbol{Z}_i] = \boldsymbol{A}_i^\top\boldsymbol{\beta} + g_Y(\boldsymbol{Z}_i)$, i.e. contains constant treatment effects and additively separable confounding. Any consistent deconfounder converges to a naïve estimator for any finite subset of treatment effects as n grows large.*

Proof. We begin with some preliminaries. As in Section B.2, we partition the m causes, $\boldsymbol{A}_i$, into finite m_F focal causes of interest, $\boldsymbol{A}_{i,F}$, and m_N nonfocal causes, $\boldsymbol{A}_{i,N}$. Then the conditional expectation function of the outcome can be rewritten $\mathbb{E}[Y_i|\boldsymbol{A},\boldsymbol{Z}] = \boldsymbol{A}_{i,F}^\top\boldsymbol{\beta}_F + \boldsymbol{A}_{i,N}^\top\boldsymbol{\beta}_N + g_Y(\boldsymbol{Z}_i)$. In what follows, we will also use the conditional expectation function $g_{\boldsymbol{A}}(\boldsymbol{z}) \equiv \mathbb{E}[\boldsymbol{A}_i|\boldsymbol{Z}_i = \boldsymbol{z}]$, as well as its partitioned counterparts, $g_{\boldsymbol{A}_F}(\boldsymbol{z}) \equiv \mathbb{E}[\boldsymbol{A}_{i,F}|\boldsymbol{Z}_i = \boldsymbol{z}]$ and $g_{\boldsymbol{A}_N}(\boldsymbol{z}) \equiv \mathbb{E}[\boldsymbol{A}_{i,N}|\boldsymbol{Z}_i = \boldsymbol{z}]$.

Pinpointedness implies that $g_{\boldsymbol{A}}(\boldsymbol{z})$ must be invertible and consistently estimable; it also requires that both n and m go to infinity. When these conditions hold, $p\text{-lim}_{n\to\infty}\hat{g}_{\boldsymbol{A}}^{-1}(\boldsymbol{A}_i) = g_{\boldsymbol{A}}^{-1}(\boldsymbol{A}_i) = \boldsymbol{Z}_i$, up to symmetries such as rotation invariance. The deconfounder estimator then reduces to the partially linear regression

$$\left(\hat{\boldsymbol{\beta}}_F^{\text{deconf}}, \hat{\boldsymbol{\beta}}_N^{\text{deconf}}, \hat{g}_Y^{\text{deconf}}\right) = \underset{\boldsymbol{\beta}_F^*,\boldsymbol{\beta}_N^*,g_Y^*}{\arg\min} \sum_{i=1}^{n}\left(Y_i - \boldsymbol{A}_{i,F}^\top\boldsymbol{\beta}_F^* - \boldsymbol{A}_{i,N}^\top\boldsymbol{\beta}_N^* - g_Y^*(\hat{g}_{\boldsymbol{A}}^{-1}(\boldsymbol{A}_i))\right)^2$$

which is consistent for $\boldsymbol{\beta}_F$ (Robinson, 1988).

Now consider the following alternative estimator, the partially linear naïve regression

$$\hat{h}^{\text{naïve}}_{\boldsymbol{A}_F} = \underset{h^*_{\boldsymbol{A}_F}}{\arg\min} \sum_{i=1}^{n} \left\| \boldsymbol{A}_{i,F} - h^*_{\boldsymbol{A}_F}(\boldsymbol{A}_{i,N}) \right\|_F^2 \tag{32}$$

$$\hat{h}^{\text{naïve}}_{Y} = \underset{h^*_{Y}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - h^*_{Y}(\boldsymbol{A}_{i,N}) \right)^2 \tag{33}$$

$$\hat{\boldsymbol{\beta}}^{\text{naïve}}_F = \left(\tilde{\boldsymbol{A}}^{\text{naïve}\top}_F \tilde{\boldsymbol{A}}^{\text{naïve}}_F \right)^{-1} \tilde{\boldsymbol{A}}^{\text{naïve}\top}_F \tilde{\boldsymbol{Y}}^{\text{naïve}} \tag{34}$$

where $\tilde{\boldsymbol{A}}^{\text{naïve}}_F$ collects $\boldsymbol{A}_{i,F} - \hat{h}^{\text{naïve}}_{\boldsymbol{A}_F}(\boldsymbol{A}_{i,N})$ and $\tilde{\boldsymbol{Y}}^{\text{naïve}}$ collects $Y_i - \hat{h}^{\text{naïve}}_{Y}(\boldsymbol{A}_{i,N})$. Like its linear-linear analogue, (13), this generalized naïve estimator models the outcome in terms of the treatments only, ignoring the existence of confounding. It can be seen that (8) and (9) flexibly estimate the conditional means of $\boldsymbol{A}_{i,F}$ and Y_i, respectively, given $\boldsymbol{A}_{i,N}$.

We now note that because pinpointedness only holds with infinite m, it must also hold with $m - m_F$ treatments. That is, if $g^{-1}_{\boldsymbol{A}}(\boldsymbol{A}_i) = \boldsymbol{Z}_i$, then also $g^{-1}_{\boldsymbol{A}_N}(\boldsymbol{A}_{i,N}) = \boldsymbol{Z}_i$. Next, because $g_{\boldsymbol{A}_F}(\cdot)$ by definition minimizes the conditional variance in the focal treatments, $g_{\boldsymbol{A}_F}(g^{-1}_{\boldsymbol{A}_N}(\boldsymbol{A}_{i,N}))$ is asymptotically the minimizer of (8). Similarly, it is easy to see that $\boldsymbol{A}^{\top}_{i,N}\boldsymbol{\beta}_N + g_Y(g^{-1}_{\boldsymbol{A}_N}(\boldsymbol{A}_{i,N}))$ asymptotically solves (9). As a result (10) identifies using $\boldsymbol{A}_{i,F} - \mathbb{E}[\boldsymbol{A}_{i,F}|\boldsymbol{Z}_i]$, the component of the focal treatments which is uncontaminated by confounding. Thus, $\hat{\boldsymbol{\beta}}^{\text{deconf}}_F$ and $\hat{\boldsymbol{\beta}}^{\text{naïve}}_F$ both converge in probability to $\boldsymbol{\beta}_F$, and hence to each other. □

B.10 Inconsistency of the Deconfounder in Nonlinear Settings

We now generalize our previous results to a broad class of nonlinear-nonlinear data-generating processes. We consider all factor models in which the distributions of $\boldsymbol{Z}_i$ and $\boldsymbol{A}_i|\boldsymbol{Z}_i$ have continuous density. We restrict our attention to the class of outcome models with additively separable confounding, a family that nests all data-generating processes studied in our paper and in applications of the deconfounder. That is, we study outcome models satisfying $\mathbb{E}[Y_i|\boldsymbol{A}, \boldsymbol{Z}] = f(\boldsymbol{A}_i) + g_Y(\boldsymbol{Z}_i)$, allowing for arbitrarily complex confounding and arbitrarily complex, interactive treatment effects.

We offer Theorem 2.

Theorem 2. *(Inconsistency of the Deconfounder in Nonlinear Settings.)*
Consider all data-generating processes in which finite treatments are drawn from a factor model with continuous density. The deconfounder is inconsistent for any outcome model with additively separable confounding.

This result follows from a relatively simple premise, building on Lemma 2. There is always noise in $\boldsymbol{A}_i$, which means the analyst can never recover the exact value of the $\boldsymbol{Z}_i$ when there are a finite number of treatments drawn from a continuous factor model. This is because there are many candidate values of the latent confounder that are consistent with the observed treatments. Let $\hat{\boldsymbol{Z}}_i = \hat{g}^{+}(\boldsymbol{A}_i)$ be the output of the deconfounder factor-model stage, a pseudo-inverse function that attempts to recover $\boldsymbol{Z}_i$ from $\boldsymbol{A}_i$. The error, $\tilde{\boldsymbol{Z}}_i = \boldsymbol{Z}_i - \hat{\boldsymbol{Z}}_i = \boldsymbol{Z}_i - \hat{g}^{+}(\boldsymbol{A}_i)$, depends on $\boldsymbol{A}_i$. Moreover, the outcome Y_i is in part dependent on $\tilde{\boldsymbol{Z}}_i$ (put differently Y_i is dependent on $\boldsymbol{Z}_i$, which is additively composed of both $\hat{\boldsymbol{Z}}_i$ and

$\tilde{\boldsymbol{Z}}_i$). Therefore, an outcome analysis that neglects the unobserved mismeasurement, $\tilde{\boldsymbol{Z}}_i$, will necessarily be affected by omitted variable bias

Proof. By Lemma 2, $\hat{\boldsymbol{Z}}$ generically fails to pinpoint $\boldsymbol{Z}$ for finite m. The deconfounder thus attempts to learn $\mathbb{E}[Y_i|\boldsymbol{A}_i, \boldsymbol{Z}_i] = f(\boldsymbol{A}_i) + g_Y(\boldsymbol{Z}_i)$ using only Y_i, $\boldsymbol{A}_i$, and $\hat{\boldsymbol{Z}}_i$. That is, it estimates $\hat{f}(\boldsymbol{A}_i) + \hat{g}_Y(\hat{\boldsymbol{Z}}_i)$. In general, unbiased estimation of $f(\boldsymbol{A}_i)$ requires the ignorability of omitted variables. However, the omitted variable in this deconfounder outcome-regression stage, $\tilde{\boldsymbol{Z}}_i = \boldsymbol{Z}_i - \hat{\boldsymbol{Z}}_i = \boldsymbol{Z}_i - \hat{g}^+(\boldsymbol{A}_i)$, is not only associated with the treatments—it is actually a function of them. Because $\boldsymbol{Z}_i - \hat{g}^+(\boldsymbol{A}_i) \not\perp\!\!\!\perp \boldsymbol{A}_i$, failures of pinpointing thus result in biased inference on $f(\boldsymbol{A}_i)$. □

Appendix C. Simulation Details

Our simulation code is based on replication materials for Wang and Blei (2019) that were provided to us by Wang and Blei in December of 2019 and two simulations that were publicly posted tutorials on the `blei-lab` github page from September 20, 2018 (Wang, 2019) until its removal on March 22nd, 2020, after we downloaded and analyzed them. After reading our working paper and discussing our analyses with them, Wang, Zhang and Blei posted reference implementations on July 2nd, 2020 at `https://github.com/blei-lab/deconfounder_public` and `https://github.com/zhangly811/Medical_deconfounder_simulation`. Because the reference implementation was produced in response to our analyses, all references are with respect to the December 2019 code and the details provided in the published papers.

This appendix provides additional details on the simulations and attempts to explain why our results deviate from the previously published results. In this introduction to the appendix, we define a few terms that we will use repeatedly. We then detail common deviations between our simulations and the original designs and provide discussion of why our results differ. We then run through each simulation in turn.

Common Terms:

- Naïve Regression
  When not otherwise specified the naïve regression is an OLS regression of the outcome on all the treatments. When creating confidence intervals they are always 95% intervals calculated using the classical covariance matrix estimated under homoskedasticity.
- Oracle Regression
  This follows the same setup as the naïve regression but also controls for the unmeasured confounder.
- PCA
  When we compute a principal components analysis we always center and scale the treatments. We take the top k principal components where k is set to the true number of unobserved components. This is obviously not feasible in real world settings, but is as generous as possible to the deconfounder.
- pPCA
  To compute probabilistic Principal Components Analysis we follow the `rstan` code provided to us by Wang and Blei for replicating the smoking example in Wang and Blei (2019). We remove the computation of heldout samples for reasons that will be

explained below. We estimate the model using automatic differentiation variational inference using their convergence settings. When unspecified we use the posterior mean as our estimate of $\hat{Z}_i$.

- Deconfounder
  When not otherwise specified we are using the substitute confounder version of the deconfounder not the reconstructed causes.

C.1 Common Deviations in Simulation Setups

We have tried to remain as faithful as possible to the original simulation setups, but we have made changes where necessary to assess the deconfounder's performance. In each simulation and application, we detail all deviations in procedures—here we summarize the most common such deviations.

Stabilizing Estimation There are two areas where estimation issues become a concern in the deconfounder: the estimation of the factor model and the estimation of the outcome model.

Instability in the factor model arises for a number of reasons. In at least two of the simulations (Medical Deconfounder 2 and Smoking), the original design calls for simulating factor models which have more latent dimensions than observed data dimensions (m). In these settings the extra factors are just noise as n factors would be sufficient to exactly reconstruct the data under a frequentist model. In these settings, we report results for the models with $k > m$ but caution against overinterpreting the results. See the smoking simulation for an examination of factor model instability in this setting. In other settings (GWAS, Actors and Breast Cancer), the code uses a procedure that analyzes a subset of the data when fitting factor models and updates in batches. In Breast Cancer and Actors we replace their procedure with code that analyzes the entire dataset at once. In many of the applications we replace their PPCA with standard PCA where possible to avoid the noise from the posterior approximation.

Many of the original simulations estimate the naïve and oracle regressions with bayesian linear regression with normal priors fit with black-box varitional estimation in `Stan`. We always use the computationally cheaper and more stable OLS which explains why in several simulations (e.g. Medical Deconfounder) the original papers report oracle coverage rates that are below the nominal levels. Instability in the deconfounder outcome model is particularly high because of near-perfect collinearity between the treatments and the substitute confounder. This problem is in turn exacerbated by the black box variational estimation in `Stan`. This drives our development of the PCA+CV-Ridge deconfounder variant we deploy in the Medical Deconfounder 1 simulation.

Probing the Simulations We extend many of the simulations to more thoroughly probe the properties of the estimator. A number of the original simulations (Quadratic, Logistic, Medical Deconfounder 1) only report results from a single realization of the data generating process, while we repeat hundreds or thousands of times. We also extend our search over reasonable parameter spaces of the original data generating process by examining different numbers of treatments and levels of confounding (Quadratic), different levels of noise added to the substitute confounder (Logistic) and different coefficients in the outcome model

(Medical Deconfounder 1). In other cases, we use the same simulation designs, but explore the results in different ways (e.g. breaking out GWAS results by causal and non-causal coefficients).

Removing the Heldout Procedure Many of the original simulations describe holding out 20% of the data for predictive checks. We generally assume the correct latent specification and are interested in comparing all models that are presented in the original paper so we skip this step. See also our discussion in Supplement D for an explanation of why the heldout procedure implemented in the smoking example yields different results.

C.2 Medical Deconfounder Simulations

The simulations are both taken from Section 3 of Zhang et al. (2019) on the medical deconfounder. They are used in conjunction with two case studies to establish the performance of the medical deconfounder. We recreate these simulations from details provided in the paper (reusing `Stan` code provided to us by Wang and Blei provided for the smoking example).

C.2.1 Simulation Study 1: One Real Cause

Summary: We replicate and extend a simulation design designed to support the medical deconfounder from Zhang et al. (2019) which uses penalized regression to be estimable.

The true data generating process for each of 1000 patients indexed by i with true scalar confounder Z_i is as follows.

$$\begin{aligned} Z_i &\sim \text{Normal}(0, 1) \\ A_{1,i} &\sim \text{Normal}(0.3Z_i, 1) \\ A_{2,i} &\sim \text{Normal}(0.4Z_i, 1) \\ Y_i &\sim \text{Normal}(0.5Z_i + 0.3A_{2,i}, 1) \end{aligned}$$

In the actual paper, the notation writes the error term as ϵ_i for that is shared across A_1, A_2, Y, but we take this to be a typo.

The original simulation fits a probabilistic PCA model with $k = 1$ using black box variational inference (Ranganath et al., 2014) in Edward and fit the outcome model using ADVI (Kucukelbir et al., 2017). To keep the entire workflow in R, we fit the probabilistic PCA model with $k = 1$ in `rstan` using ADVI based on the probabilistic PCA model provided to us for replicating the smoking example. We modified the code to remove the heldout procedure designed for posterior predictive checks but otherwise kept the model with same, with the same priors. We use the default settings in `rstanarm` for the outcome regression. We increased the maximum number of iterations on both the factor model and the outcome model by a factor of 10 to try to stabilize the estimates.

We introduce the PCA + CV-Ridge estimator. We estimate PCA with $k = 1$ and then estimate a ridge regression using `glmnet` (Friedman et al., 2010) and choosing the penalty parameter by cross-validation. In the paper, Zhang et al. (2019) present one realization of this data-generating process and we repeat this process 1000 times to create a simulation. Table 2 of Zhang et al. (2019) contains their results. The results are not directly comparable in that they are showing a single realization of a set of estimators and are showing more systematic results.

Finally, we show an additional simulation where the true treatment effects are $(-.3, .3)$ instead of $(0, .3)$ to demonstrate the results are sensitive to the settings of the true coefficients, consistent with our results about the subset deconfounder.

Deviations

- we take more than one draw from the simulation and report aggregate quantities
- we correct a typo in the manuscript and do not share errors across treatments
- we use `rstan` instead of `Edward` for pPCA
- we increase the maximum iterations of the bayesian procedures
- we add the PCA + CV-Ridge Estimator.
- we assess in terms of bias, standard deviation and RMSE rather than with coverage or p-values.
- we don't holdout 20% of the data or do predictive checks
- we use OLS rather than bayesian linear regression for our naïve and oracle estimators

Their Results: They argue based on one draw from their simulated process that the deconfounder leads to the right conclusions and the naïve leads to the wrong conclusions.
Our Results: We demonstrate that even with the PCA + CV-Ridge Estimator which provides large improvements over the medical deconfounder, the performance of the deconfounder is in practice no better than the naïve regression.

C.2.2 Simulation Study 2: A multi-medication simulation

Summary: We replicate and extend a simulation design created to support the medical deconfounder from Zhang et al. (2019) which uses a nonlinear functional form to be estimable.

Zhang et al. (2019) simulate 50 total treatments of which only 10 have a non-zero effect on the outcome. They use the data generating process,

$$
\begin{aligned}
Z_{i,k} &\sim \text{Normal}(0, 1), && k = 1 \dots 10 \\
A_{i,j} &\sim \text{Bernoulli}\left(\sigma\left(\sum_{k=1}^{10} \lambda_{k,j} Z_{i,k}\right)\right), && j = 1 \dots 50 \\
Y_i &\sim \text{Normal}\left(\sum_{j=1}^{10} \beta_j A_{ij} + \sum_{k=1}^{10} \gamma_k Z_{i,k}, 1\right) \\
\lambda_{k,j} &\sim \text{Normal}(0, .5^2) \\
\gamma_k, \beta_j &\sim \text{Normal}(0, .25^2)
\end{aligned}
$$

where σ is the sigmoid function. Only the first 10 treatments have non-zero coefficients—these are the medications that work.

Zhang et al. (2019) report results for the oracle, the naïve estimator, a poisson matrix factorization (PMF) with $K = 450$ and a deep exponential family—the latter two implemented in `Edward`. We estimate the Poisson matrix factorization using the `R` package `poismf`

with L_2 regularization (`l2_reg=.01`) and $K = 450$ run for 100 iterations. All outcome regressions are computed using standard linear regression with homoskedastic standard errors. Coverage is computed with respect to 95% confidence intervals.

		Coverage Proportion		
	RMSE	All	Non-Zero	Zero
Oracle	0.03	0.95	0.95	0.95
Naïve	0.13	0.42	0.41	0.42
Deconfounder	0.13	0.43	0.43	0.44

Table 6: **Replication of Simulation Study 2 of Zhang et al. (2019).** The deconfounder is estimated using a 450-dimensional poisson matrix factorization. Coverage rates for nominal 95% intervals are reported separately for zero coefficients (no causal effect) and non-zero coefficients.

We calculate RMSE by calculating all the squared errors $(\hat{\beta}_j - \beta)^2$ and taking the square root of the mean over all coefficients in all simulations. We simulate the entire data generating process each time and conduct 1000 simulations.

We note that original results in Zhang et al. (2019) do not show the oracle achieving 95% coverage—in fact, they show only 50% coverage for the non-zero coefficients whereas we achieve the nominal rate. The results of the deconfounder and the naïve estimator are essentially indistinguishable in our setting, but we note that they aren't particularly different as reported in the original paper, either. Because the factor models are fit with $k = 450$ to a 50-dimensional the exact form of the prior specification will have a huge impact on the values of the substitute confounder. In an unpenalized setting, $k = 50$ should be sufficient to exactly reconstruct the observed data.

Deviations

- we report only the poisson matrix factorization (not the deep exponential family) and use `poismf` instead of `Edward`.
- we don't holdout 20% of the data or do predictive checks
- we use OLS rather than bayesian linear regression for all outcome regressions

Their Results: They argue that the deconfounder produces better results than the naïve.

Our Results: We cannot compare to the deep exponential family, but the poisson matrix factorization does not show the gains in improvement over naïve that they claim. We show better RMSE for all estimators as well as better coverage.

C.3 Tutorial Simulations

The simulations were IPython notebooks in a folder marked `toy simulations`. Each shows a simulated data generating process and then walks through a single draw and compares the naïve regression (labeled "noncausal estimation") with deconfounder estimates based on reconstructed causes and the substitute confounder. We understand that public tutorials need to be fast to run and thus may often be less nuanced than authors would prefer. That

said, we think these tutorials are important to replicate because they are they way many potential users would be exposed to the deconfounder and would come to understand its properties.

C.3.1 Logistic Simulation

Summary: This simulation adds noise to the substitute confounder to make the model estimable—we explore how variations on the amount of noise affect simulation results.

The clearest application of the noised deconfounder is in the logistic tutorial simulation (Wang, 2019). The simulation uses the following data generating process to create 10,000 observations:

$$\begin{aligned}(X_1, X_2, Z) &\sim \text{Multivariate Normal}\left((0,0,0), \begin{pmatrix} 1 & 0.4 & 0.4 \\ 0.4 & 1 & 0.4 \\ 0.4 & 0.4 & 1 \end{pmatrix}\right) \\ y &\sim \text{Bernoulli}\left(\frac{1}{1+\exp\left(-(.4+.2X_1+1X_2+.9Z)\right)}\right)\end{aligned}$$

The substitute confounder is found by using PCA of (X_1, X_2) and then adding random noise, such that

$$\hat{Z} = \text{Normal}(\underbrace{\eta_1 X_1 + \eta_2 X_2}_{\text{PCA}}, .1^2)$$

The simulation then estimates a logistic regression with a linear predictor $(X_1, X_2, \hat{Z})$. The random noise is introduced to break the perfect collinearity between the treatments and the substitute confounder. For a single draw of the data generating process, the tutorial simulation claims that with the naïve regression "none of the confidence intervals include the truth", but with the deconfounder "both of the confidence intervals (for $X1, X2$) include the truth." The implication is that the deconfounder improves upon the naïve regression estimates. We repeat the simulation 100,000 times to summarize properties more generally and vary the standard deviation of the noise added to the deconfounder from .1 to $(1, .1, .01, .001)$.

We extend the tutorial to assess bias, standard deviation, coverage and RMSE all at different values of the noise parameter with results shown in Table 7. At no point does the overall performance of the deconfounder exceed that of the naïve estimator. For large noise, the deconfounder approaches the performance of the naïve estimator; as the noise grows small and collinearity increases, estimator variance and RMSE get large very quickly.

To help explain the discrepancy in our results, we note that the single draw shown in the workbook is unusual in terms of its error. The argument made in the writeup is that the confidence intervals of the deconfounder contain the truth while the naïve estimator doesn't. This is a fairly common occurrence—approximately 75% of the simulations because the naïve estimator has coverage close to zero. However, in only 42% of the simulations did the deconfounder produce answers closer to the truth than the naïve (along both dimensions). The deconfounder is unusually close to (and the the naïve estimator unusually far from) the truth in terms of mean absolute error across the two coefficients. The deconfounder only

		Treatment 1		Treatment 2	
	Noise S.D.	Deconfounder	Naïve	Deconfounder	Naïve
Bias	10^{-3}	0.200	0.210	0.116	0.126
	10^{-2}	0.202		0.118	
	10^{-1}	0.210		0.127	
	10^{0}	0.210		0.126	
Std. Dev.	10^{-3}	16.566	0.026	16.567	0.030
	10^{-2}	1.659		1.659	
	10^{-1}	0.167		0.168	
	10^{0}	0.031		0.035	
Coverage	10^{-3}	0.949	0.000	0.949	0.012
	10^{-2}	0.948		0.949	
	10^{-1}	0.759		0.884	
	10^{0}	0.000		0.042	
RMSE	10^{-3}	16.568	0.211	16.567	0.130
	10^{-2}	1.671		1.663	
	10^{-1}	0.269		0.211	
	10^{0}	0.212		0.131	

Table 7: **Logistic Tutorial Simulation.** 100,000 simulations are summarized in terms of bias, standard deviation of the sampling distribution, coverage of 95% confidence intervals, and root mean squared-error for various levels of simulated noise. As the noise gets small, the standard deviation and the RMSE of the deconfounder explode (the estimator approaches perfect collinearity). As the noise increases, the deconfounder collapses on the performance of the naïve estimator.

performs as well as reported 8% of the time and naïve only performs as poorly as reported 16% of the time. Thus the reported draw is not a representative indicator of performance.

Deviations

- we repeat the process to create a simulation
- we examine only substitute confounder and not reconstructed causes
- we explore different noise levels

Their Results: The simulation shows an example where the confidence interval for the deconfounder covers the truth and the naïve estimator does not.

Our Results: We show that the coverage result is relatively typical but the one draw shown is abnormally accurate for the deconfounder. By evaluating across levels of noise added to the substitute confounder we demonstrate that the results are highly sensitive to the noise level and are at no level better than naïve on bias, variance or RMSE.

C.3.2 Quadratic Simulation

Summary: This simulation uses a transformation of PCA to make the model estimable—we explore how variations in simulation parameters affect results.

10,000 observations are simulated from the following data-generating process,

$$\begin{bmatrix} A_{i,1} \\ A_{i,2} \\ Z_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \right)$$

$$Y_i \sim \mathcal{N}(0.4 + 0.2A_{i,1}^2 + 1A_{i,2}^2 + 0.9Z_i^2,\ 1)$$

for $\rho = .4$. The substitute confounder, $\hat{Z}$, is based on a PCA of (A_1, A_2), $\hat{Z} = \eta_1 A_1 + \eta_2 A_2$. The outcome regression is $Y = \tau_0 + \tau_1 A_1^2 + \tau_2 A_2^2 + \gamma \hat{Z}^2 + \epsilon_i$.

We modify this simulation by adding noise to the outcome drawn from Normal(0,1). We also evaluate the performance of the parametric specification $\hat{Z}_{(\text{alt})} = A_1^2 + A_2^2 + 2A_1A_2$ and show that it has superior performance. We also demonstrate that the deconfounder breaks down for negative values of the correlation coefficient.

In a footnote of the main text, we noted that the proposed inference procedure uses PCA on A_1, A_2 but estimates treatment effects on their squares. This either presumes that we are interested in the treatment effect of the squared treatment or that we have access to the square root of our real treatments of interest. However, in that case, we lose the true sign of A_1, A_2. To approximate this scenario, we ran a version of the analysis where we estimate PCA on the absolute value of the variables $(|A_1|)$, $|A_2|$) with results shown in Table 8. This leads to an approximately five-fold increase in the RMSE for the deconfounder and roughly a doubling in RMSE for the parametric specification. The relatively worse performance of the deconfounder is due to centering before performing PCA, but the knowledge to not center can only be leveraged if we are sure that the underlying $\boldsymbol{A}$ was centered which requires more knowledge that we cannot have.

	Treatment Number	
model	1	2
Naïve	0.1248	0.1252
Oracle	0.0072	0.0074
Parametric	0.0425	0.0431
Deconfounder	0.1028	0.1033

Table 8: RMSE of Quadratic Simulation with Original Settings ($\rho = .4$ and $m = 2$) with PCA of Absolute Value of $\boldsymbol{A}_1, \boldsymbol{A}_2$.

Deviations

- we simulate the outcome with error
- we repeat the process to create a simulation
- we examine only substitute confounder and not reconstructed causes

Their Results: The original simulation shows for a single draw that the deconfounder is closer to the truth than the naïve and they claim that the confidence intervals contain the truth.

Our Results: We extend the simulation and while performance is in fact strong at the

given settings, changing the correlation between $\boldsymbol{A}$'s to moderately negative causes the deconfounder to perform much worse than the naïve. As our theory predicts, when the number of treatments is large, the difference between the deconfounder and the naïve regression disappears.

C.4 GWAS

Summary: Wang and Blei (2019) applies the deconfounder to study the effects of genes on traits. We use replication provided to us by Wang and Blei to perform a simulation under similar conditions and we find that the deconfounder and the naïve regression perform almost identically. On closer inspection this is expected. Figure 4 of Wang and Blei (2019) shows nearly identical performance between the deconfounder and the naïve regression.

C.4.1 Overview

As a motivating example of the deconfounder, Wang and Blei (2019) evoke the use of methods similar to the deconfounder in the genetics literature to explain the effect of genes on the expression of traits. The genetics database that was used to produce the simulations in the original paper could not be shared with us because of restrictions in data access. Instead, Wang and Blei shared a purely synthetic simulation procedure intended to replicate the characteristics of the simulations in the original paper as well as code for several factor models applied to this data set. In this section we describe our results using this synthetic example of the GWAS simulation. We find that the deconfounder offers essentially identical performance to the naïve regression. This is not surprising, because Figure 4 of Wang and Blei (2019) shows that RMSE from the deconfounder and the naïve regression are nearly identical.

C.4.2 Simulation Procedure

We follow the description of the data generating process in Wang and Blei (2019) for the high SNR setting, using a synthetic genetic simulation provided to us to generate data under the Balding-Nichols procedure. We generate data with 5000 individuals and 5000 genetic markers, with genetic and environmental variation set to 0.4, and with 10% of the genes assumed to have a causal effect on the outcome. Note, that because of this assumption, any method that shrinks coefficient estimates towards zero will obtain better performance on the non-causal genes, so we divide our results into causal and non-causal genes. We use the provided code to estimate substitute confounders for deep exponential families (with a 100-dimensional substitute confounder), pca (10-dimensional), poisson matrix factorization (10-dimensional), linear factor analysis (10-dimensional), and probabilistic principal components (5-dimensional). We avoid the use of the holdout procedure described in the code because it incorrectly sets all held out values to be zero. The posterior predictive checks, as implemented in the code, suggest that the factor models have unrealistic model fits.

We then generate a single set of effects on the causal genes $\boldsymbol{\beta} \sim \mathcal{N}(0, 1)$ and confounding variables $\boldsymbol{\lambda}$ using a slightly modified function from WB, where the modification enabled us to draw the coefficients only once. Following the original simulation design, we set all non-

causal coefficients to zero. Using the draws of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ we simulated the outcome vector, $\boldsymbol{Y}$, 100 times.

As in the original simulation, we use a ridge regression and nonlinear functional form to render the deconfounder estimable. In each simulation we estimate a ridge regression, cross validating to obtain the penalty parameter. For the naïve regression we condition on all 5,000 genes. For each of the factor models we also include the corresponding estimated substitute confounders. We write our own code to estimate the average root mean squared error, which we display in Table 9.

Table 9 shows that the naïve regression outperforms the deconfounder on this simulation on the genes that have a causal effect on the outcome. On the non-causal genes the other models perform slightly better, but all models offer a nearly identical improvement over the naïve regression of the outcomme on one gene at a time.

Table 9: Using the Synthetic Genetic Data Set, The Deconfounder Offers No Improvement Over Naïve Regression

	RMSE		
	Causal	Non-Causal	Overall
Naïve	0.737	0.127	0.263
Oracle	0.742	0.125	0.263
Deconfounder (DEF)	0.746	0.123	0.263
Deconfounder (PCA)	0.745	0.123	0.263
Deconfounder (PMF)	0.745	0.123	0.263
Deconfounder (LFA)	0.746	0.123	0.263
Bivariate Naive	1.576	1.607	1.604

Deviations

- We use a synthetic simulation created to approximate the simulation in the original paper
- We draw the genes and λ once.
- We evaluate bias and RMSE (see discussion in Supplement D.1.5.
- We use a ridge regression with a cross validation-selected penalty, using mean squard error as a cross validation statistic
- Following the genetics literature, we examine performance differences on causal and non-causal genes

Takeaways

- The deconfounder offers marginal improvements on the non-causal genes and performs worse than the naïve estimator on the causal genes.
- Wang and Blei (2019) uses this simulation as evidence that DEFs are useful. While DEFs do provide a better estimate of the non-causal genes, they are worse on the causal genes and offer—at best—marginal improvements in effect estimation.

C.5 Subset Deconfounder

We use a new simulation design to examine the finite sample properties of the subset deconfounder under the linear-linear data generating process. We create a single-dimensional confounder, $\boldsymbol{Z}$, and allow this confounder to satisfy strong infinite confounding.

Our simulation is designed to demonstrate how the average RMSE of the subset deconfounder depends on the underlying treatment effect sizes. In all of our settings we set each $\theta_m = 10$, ensuring strong infinite confounding is satisfied. We suppose $A_i \sim \theta_m Z_i + \nu_m$ where $\nu_m \sim \mathcal{N}(0, 0.01)$. We then generate outcome data using the linear outcome model using the following coefficient values:

1. $\beta_m = 10$
2. $\beta_m = 100$
3. $\beta_m \sim \mathcal{N}(1, 4)$
4. $\beta_m = \frac{1}{m}$

We suppose $\gamma = 10$ for all simulations and that $Y_i = A_i\boldsymbol{\beta} + Z_i\gamma + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 0.01)$.

The results from this simulation align exactly with the predictions from Proposition 4. Specifically, from Proposition 4 we predict for $\beta_m = 10$ a bias of magnitude 10, $\beta_m = 100$ a bias of magnitude 100, $\beta_m \sim \mathcal{N}(1, 4)$ a bias of magnitude 1, and for $\beta_m = \frac{1}{m}$ the bias at M treatments will bias$_M = \frac{\sum_{m=1}^{M} \frac{1}{M}}{M}$ or the average value of β_m.

Appendix D. Smoking Simulation

Summary: We replicate the first empirical case study in Wang and Blei (2019), a semi-synthetic dataset about the causes of smoking. We argue that the simulation design is not informative about the performance of the deconfounder because (1) the factor models often have $k \geq m$, and (2) the controls used to compare with a strategy of measuring confounders are themselves uninformative about the confounding. We first quickly review the details of the original design based on the paper and replication code provided by Wang and Blei in December 2019. We then briefly detail our argument along with some additional results.

D.1 Original Design

The smoking simulation is a semi-synthetic study that uses data from the 1987 National Medical Expenditures Survey (NMES) to generate a real joint distribution of three variables which are combined with a linear model to create a synthetic outcome.

D.1.1 Data Generating Process in Wang and Blei (2019)

The original simulation in Wang and Blei (2019) selects two observed treatments from the NMES: the individual's martial status, A_{mar}, and the exposure to smoking measured as the number of cigarettes per day divided by 20 times number of years smoked,A_{exp}. The last age at which the person smoked, A_{age}, is designated as the unobserved confounder.

All variables are centered and scaled. In equations 23 of Wang and Blei (2019), the data generating process for the synthetic outcome is laid out as,

$$Y_i = \text{Normal}\left(\beta_0 + \beta_{\text{mar}} A_{\text{mar},i} + \beta_{\text{exp}} A_{\text{exp},i} + \beta_{\text{age}} A_{\text{age},i}, 1\right)$$

where the intercept β_0 is included in the replication code provided to us but not the paper.

In Wang and Blei (2019) equation 24, the data generating process of the coefficients is described as,

$$\begin{aligned}\beta_{\text{mar}} &\sim \text{Normal}(0,1)\\ \beta_{\text{exp}} &\sim \text{Normal}(0,1)\\ \beta_{\text{age}} &\sim \text{Normal}(0,1)\end{aligned}$$

although in the provided replication code the coefficient for the last variable (which will be used as the unobserved confounder) is multiplied by 2.5, leading to,

$$\beta_{\text{age}} \sim \text{Normal}(0, 2.5^2)$$

One of the two treatments, A_{mar}, is a factor variable with 5 levels. While the factor variable is unlabeled, by examining earlier sources[17], we are confident that level 1 corresponds to married and level 5 corresponds to never married. We think levels 2-4 correspond to widowed, divorced and separated respectively. This treatment is treated as a numeric variable in `R` although factor levels 2-4 (22% of the data) aren't meaningfully ordered.

The original simulation treats the first two variables as observed and the final($A_{\text{age},i}$) as the unobserved confounder. The paper reports the results for 12 configurations of models: naïve regression, oracle, linear factor with one dimension (substitute confounder and reconstructed cause), quadratic factor with one, two and three dimensions (substitute confounder and reconstructed cause) and the one dimensional quadratic factor model with additional covariates.

D.1.2 Factor Model Inference in Wang and Blei (2019)

For each simulation in Wang and Blei (2019), factor models are fit using automatic variational bayes as implemented in `rstan`. The model for the quadratic factor analysis (the linear is analogous) as implemented in the provided replication code is,

$$\begin{aligned}\alpha &\sim \text{Gamma}(1,1)\\ \theta^{(0)} &\sim \text{Normal}(0, 1/\alpha)\\ \theta_k^{(1)} &\sim \text{Normal}(0, 1/\alpha)\\ Z_{i,k} &\sim \text{Normal}(0, 2^2)\\ \boldsymbol{A}_i &\sim \text{Normal}\left(\theta^{(0)} + \sum_{k=1}^{K} \theta_k^{(1)} Z_{i,k} + \sum_{k=1}^{K} \theta_k^{(2)} Z_{i,k}^2, .1^2\right)\end{aligned}$$

17. The data comes from Imai and Van Dyk (2004) which in turn gets it from Johnson et al. (2003) which obtains the data from the original source.

The Normal variances are held fixed for $\boldsymbol{Z}$ and $\boldsymbol{A}$ and in the equations above we have set them to the values given in the code. The model is fit with the default settings for ADVI (fully factorized gaussian approximation) except with a fixed step size of .25. Before beginning the variational approximation, the initial values are set by optimizing the joint posterior with LBFGS for a maximum of 1000 iterations.

For the substitute confounder the $\boldsymbol{Z}$ variables are used directly. For the reconstructed causes, the model outputs:

$$\hat{\boldsymbol{A}}_{\text{WB}} \sim \text{Normal}\left(\theta^{(0)} + \sum_{k=1}^{K} \theta_k^{(1)} Z_{i,k} + \sum_{k=1}^{K} \theta_k^{(2)} Z_{i,k}^2, .1^2\right)$$

This differs from Wang and Blei (2019) on page 1582 which defines the reconstructed causes as the posterior predictive mean.

D.1.3 A Note on the Holdout Procedure.

In order to calculate the posterior predictive checks, the code holds back approximately 5% of the individual cells of the matrix $\boldsymbol{A}$ sampled at random. The holdout percentage is approximate because the sampling procedure allows duplicates which are then removed. The heldout values are replaced by zero (which due to centering is also the mean of the data). These values are not resampled in the inference program and so they are effectively treated as mean single imputations of the missing values. This presumably has both an effect on the fit of the factor model and the posterior predictive checks themselves (which are now conducted exclusively on data that the model is trained believing are exact zeroes). This procedure was corrected in the reference implementation released in July 2020.

D.1.4 Evaluation Procedure in Wang and Blei (2019)

After estimating the factor model the code from Wang and Blei (2019) fits one of the following adjustment strategies:

1. Substitute Confounder
   control for $\hat{\boldsymbol{Z}}$
2. Reconstructed Causes
   replacing the treatment with $\boldsymbol{A} - \hat{\boldsymbol{A}}$. This is the version described in the paper
3. Reconstructed Causes 2
   the two-parameter version where they control for $\hat{A}$
4. Substitute Confounder with Controls
   controlling for $\hat{\boldsymbol{Z}}$ and five controls (see below)
5. Reconstructed Causes with Controls
   replacing treatment with $\boldsymbol{A} - \hat{\boldsymbol{A}}$ and five controls
6. Oracle
   controlling for the true confounder
7. Naïve
   regression of Y on all treatments only

The controls include the following variables: age started smoking, binary sex indicator, 3-level factor variable for race, 3-level factor variable for seatbelt use (rarely/sometimes/(always or almost always), 4-level factor variable for education (college graduate/some college/high school/other).[18] Unlike the treatments and confounders, these control variables are not standardized or centered. The factor variables are entered as scalars (rather than contrast coded factors). For some variables like education that are ordered this produces a linear approximation to the factor model but for the race value there is no guarantee it produces anything in particular.Each of these models is estimated with one of two different outcome regressions: bayesian linear regression estimated using ADVI in `rstanarm` and OLS.

D.1.5 Outcome Regression: Bayesian Linear Regression:

The ultimate goal for the simulations from Wang and Blei (2019) is to study properties of the joint posterior distribution $f(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$. Samples from this joint distribution are obtained by factorizing as $f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z})f(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$. Samples from $f(\boldsymbol{z}|\boldsymbol{A})$ are taken from the factor model's posterior—ignoring information from $\boldsymbol{Y}$—and used as an approximation to $f(\boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$. Then a Bayesian linear regression of $\boldsymbol{Y}$ on $\boldsymbol{A}$ and $\boldsymbol{z}$ is used to sample from the conditional posterior $f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z})$.

Let $\tilde{\beta}_{j,s,f,d}$ be a draw from this approximate posterior distribution for treatment j in simulation s where:

- $j \in 1 \dots 2$ indexes the two observed treatments
- $s \in 1 \dots S$ indexes the simulation (i.e. one dataset drawn from the semi-synthetic data generating process)
- $f \in 1 \dots F$ indexes samples from the factor model's posterior distribution $f(\boldsymbol{z}|\boldsymbol{A})$
- $d \in 1 \dots D$ indexes the sample from the outcome regression's posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z})$.

and we use $\tilde{\cdot}$ to emphasize that it is a sample from a posterior distribution. In the replication code $f(\boldsymbol{z}|\boldsymbol{A})$ (the factor model posterior) is approximated with five samples and so we will set $F = 5$. The code then approximates $f(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{A}, \boldsymbol{z})$ (outcome regression conditional posterior) with a single sample and so we will set $D = 1$. Let $\beta_{j,s}$ indicate the true treatment effect for treatment j in simulation s.

The Wang and Blei (2019) code computes three quantities which effectively treat the sample from the posterior as the estimator and compute properties of that estimator within a given simulation (treating the posterior draws as independent realizations of that estimator) and then average over simulations and treatments. We explore each below.

Bias Calculation. The following quantity is computed:

$$\text{Bias}^2_{\text{WB}} = \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{S}\sum_{s=1}^{S}\left(\left(\frac{1}{5}\sum_{f=1}^{5}\underbrace{\tilde{\beta}_{j,s,f,1}}_{\text{estimator}}\right) - \underbrace{\beta_{j,s}}_{\text{truth}}\right)^2\right)$$

The quantity marked "estimator" is a draw from the posterior distribution and the expectation for the bias is taken with respect to that posterior distribution. The metric can also

18. We pieced these together from Johnson et al. (2003) and Imai and Van Dyk (2004) but we can't be sure without definitions. These definitions due line up approximately with the summary statistics reported in Johnson et al. (2003).

be interpreted as the mean-squared error of the posterior mean estimator approximated with five samples from the posterior distribution.

Variance Calculation. Let the function $\widehat{\text{Var}}(\cdot)$ be the sample variance of its arguments. The code defines,

$$\text{Var}_{\text{WB}} = \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{S}\sum_{s=1}^{S}\underbrace{\widehat{\text{Var}}\left(\tilde{\beta}_{j,s,1,1}\dots\tilde{\beta}_{j,s,5,1}\right)}_{\text{posterior var}}\right)$$

Thus, this reports the per-simulation average posterior variance summed over treatments.

Mean Squared Error. Finally the code computes,

$$\text{MSE}_{\text{WB}} = \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{5}\sum_{f=1}^{5}\left(\tilde{\beta}_{j,s,f,1}-\beta_{j,s}\right)^2\right)\right)$$

This is the per-simulation, average squared posterior deviation summed over the treatments.

Naïve and Oracle Regressions. In both the oracle and naïve regression there are, of course, no samples from the factor model. The oracle simply averages over all 1000 samples from the outcome regression's posterior and the naïve regression averages over 5 samples from the outcome regression's posterior. This will make estimates for the naïve regression much noisier.

Priors. In the original replication code, the Bayesian outcome regression uses different priors depending on the specification. The substitute confounder uses `rstanarm`'s default Normal(0,1) prior. The reconstructed causes, oracle and naïve regressions use the default `hs_plus()` prior which is hierarchical shrinkage prior which is a Normal centered at 0 with a standard deviation that is the product of two independent half Cauchy parameters. The latter has much more mass at 0 and correspondingly fatter tails.

D.1.6 Outcome Regression: OLS

The original replication code employs a corresponding set of definitions when using ordinary least squares. Denote $\hat{\beta}_{j,s,f}$ to be the coefficient for treatment j fit on simulation s conditional on draw f from the factor model.

Reported Bias Calculation. The following quantities are computed:

$$\text{Bias}^2_{\text{WB}} = \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{S}\sum_{s=1}^{S}\left(\underbrace{\left(\frac{1}{5}\sum_{f=1}^{5}\hat{\beta}_{j,s,f}\right)}_{\text{estimator}}-\underbrace{\beta_{j,s}}_{\text{truth}}\right)^2\right)$$

This corresponds to the calculation in the Bayesian case but plugging in the coefficient estimates for the sample from the posterior.

Reported Variance Calculation. Let the function $\widehat{\text{Var}}(\cdot)$ be the same variance of its arguments and $\widehat{SE}(\cdot)$ be the estimated standard error of its argument. The code defines:

$$\text{Var}_{\text{WB}} = \frac{1}{2}\sum_{j=1}^{2}\left(\frac{1}{S}\sum_{s=1}^{S}\underbrace{\widehat{\text{Var}}\left(\hat{\beta}_{j,s,1}\dots\hat{\beta}_{j,s,5}\right)}_{\text{var of coefs}} + \underbrace{\frac{1}{5}\sum_{f=1}^{5}\widehat{SE}(\hat{\beta_{j,s,f}})^2}_{\text{avg of vars}}\right)$$

This uses the law of total variance to provide a more efficient estimator of the sum of the average variance of $f(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{Y}, \boldsymbol{A})$.

Reported Mean Squared Error. The code computes,

$$\text{MSE}_{\text{WB}} = \text{Bias}^2_{\text{WB}} + \text{Var}_{\text{WB}}$$

D.1.7 Reported Results

Table 3 of Wang and Blei (2019) presents the findings. The discussion highlights the improved performance of the one-dimensional and two-dimensional quadratic models over the naïve regression although no estimator is particularly close to the oracle. In the corresponding discussion (1585-1586), the results are used to emphasize three points: (1) the value of the posterior predictive check for signaling whether results are biased, (2) controlling for observed confounders increases variance but does not decrease bias, and (3) the deconfounder outperforms naïve regression.

D.2 New Results

In this section, we briefly present a conceptual argument about the design before providing some broader results to contextualize the findings.

D.2.1 Factor models where $k \geq m$

In four of the eight original factor model specifications including two of the three highlighted in Wang and Blei (2019) for performance, the dimensionality of the latent factors exceeds the dimensionality of the data. In a frequentist setting, these models would exactly reconstruct the observed treatments (they all nest 2-dimensional PCA as a special case). The models are fit with bayesian methods but using broad priors and so it would appear that they only reason they don't perfectly reconstruct the treatments is noise in the posterior approximation. This renders the models estimable, but uninformative about performance of the deconfounder.

D.2.2 Deviations in Our Procedure

In re-implementing the simulation we tried to strike a balance between remaining comparable to the original design and making changes that we felt were essential to being able to interpret the simulation. In total, we estimate three baseline specifications: naïve regression, the oracle model and a regression controlling for WB's controls as well as five

sets of deconfounder models based on specifications by WB (Linear Factor Model with 1 dimension, Quadratic Factor Model with 1-3 dimensions and Quadratic Factor Model with 1 dimension and controls). For each set of models we report three variants: the substitute confounder (controlling for $\hat{\boldsymbol{Z}}$), reconstructed causes as stated in the paper (replacing the treatment with $\boldsymbol{A} - \hat{\boldsymbol{A}}$) and reconstructed causes as implemented in their code (controlling for $\hat{\boldsymbol{A}}$). This adds the specification of the controls alone and includes both versions of the reconstructed causes. We outline the other changes we make here along with our rationale.

Deviation 1: OLS Outcome Regression. We use an OLS outcome regression and average results over 1000 draws of the factor model's posterior distribution (rather than 5). This ensures the computation of the approximate joint posterior mean is not too noisy. The 1000 regressions can be done computationally efficiently by noting that the design matrix stays fixed. Denoting the design matrix $\boldsymbol{X}$ we precompute $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ which means the full set of 1000 regressions can be computed with a single additional matrix multiply.

Deviation 2: Fixed Simulation Coefficients It is difficult to evaluate properties like bias when the coefficients are varying across runs. To see why this would be complicated, consider an estimator that was biased towards zero, and imagine that the calculation applied was $E[\hat{\beta}_i - \beta_i]$ where β_i was sampled from a normal distribution. The upward and downward biases would cancel each other out over draws, leading to an estimate of approximately 0. We avoid this problem by fixing the coefficients to arbitrarily chosen values (1,1,1) and then assessing the bias, variance and mean-squared-error of the posterior mean. The results are driven primarily by bias so we report only the root mean squared error.

Deviation 3: Removed Holdout Procedure We were concerned about the effects of the mean imputation so we removed the holdout procedure entirely. Once this was done, the treatments do not change across simulations.

Deviation 4: Change in the Reconstructed Causes. We replace the reconstructed causes with

$$\hat{\boldsymbol{A}}_i = \theta^{(0)} + \sum_{k=1}^{K} \theta_k^{(1)} Z_{i,k} + \sum_{k=1}^{K} \theta_k^{(2)} Z_{i,k}^2$$

in each sample from the posterior. Due to the sampling of $\boldsymbol{Z}$ and θ, the reconstructed cause is almost surely not collinear with $\boldsymbol{A}$.

Deviation 5: Remove the intercept. We remove the intercept from the true model to be consistent with the equations in Wang and Blei (2019).

Deviation 6: Make control variables factors We treat the control variables which are categorical as factors rather than numerics as done in the original simulation design.

Concerns We Do Not Address There were several issues in the simulation design that we did not address because we felt that to do so would too fundamentally alter the simulation.

- for the treatments, we continue treating the factor variables as continuous variables because otherwise the dimension of $\boldsymbol{A}$ changes

- A_{exp} is skewed (it is logged in Imai and Van Dyk (2004)) which affects the normality assumptions
- many of the models considered here involve many more dimensions than the two in the observed data. It is unclear why the model isn't fitting the data perfectly in these settings or why we would use latent variable modes with many more latent variables than observed dimensions.

D.2.3 New Results: Instability in Factor Models

We first demonstrate that a substantial amount of variation in the original results is due to instability in factor model estimation. Table 10 shows results across four different factor model fits, labeled F1–4. Because we removed the model checking using held-out data, the inputs across all four models are identical, so only the seed changes across the four iterations. The differences in the learned factor models induce substantial differences in the RMSE of the resulting estimates. For example, the two-dimensional quadratic model with substitute confounder (one of the preferred specifications in Wang and Blei (2019) and row 11 of our Table 10) ranges from 12% better to 30% worse than the naïve estimate for the effect of Treatment 1.

We note that different adjustment strategies (Sub., Rec., Rec. 2) used with the same factor model can yield substantially different results. Because the PPC is specific to the factor model and not the adjustment strategy, it cannot provide information about which would provide better performance.

As in Wang and Blei (2019) we do not observe substantial benefits from including covariates with the deconfounder. However, line three of the table makes clear that this is because the covariates alone are not sufficient to improve over the Naïve regression. In practice this is because they are essentially uncorrelated with the variable chosen to be the unobserved confounder. Thus, we should not draw conclusions from this study about the role of measured confounders.

We have only shown one set of simulated coefficients here. However, because the RMSE is driven almost entirely by the bias term, the results here are extremely well predicted by the standard omitted variable bias formula. Define $\boldsymbol{X} = (A_{\text{mar}}, A_{\text{exp}}, \hat{\boldsymbol{Z}})$ then,

$$\text{bias}(\beta_{\text{mar}}, \beta_{\text{exp}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Z}\beta_{\text{age}}$$

For a fixed factor model, and thus a fixed $\hat{\boldsymbol{Z}}$, we can calculate the bias for any setting of the true coefficient β_{age}.

D.2.4 New Results: Max ELBO of the Factor Models

The variational inference procedure comes with a natural mechanism for choosing among the factor models. We run each model twenty times and choose the one that maximizes the evidence lower bound. In `rstan` this has to be parsed from a log that collects the material printed to the screen as it is not included in the returned output. The results are in Table 11.

In practice we observe that all the linear factor model fits are very similar, but the quadratic factor models vary substantially. Those where $k \geq m$ vary the most. Thus while we maximized the ELBO over twenty different fits, we would expect that results would be

		Treatment 1				Treatment 2			
		F1	F2	F3	F4	F1	F2	F3	F4
Baseline	Naïve	1.00				1.00			
	Oracle	0.043				0.026			
	Controls	0.900				1.094			
Lin. (1 dim.)	Sub.	1.020	1.176	1.065	0.980	1.011	1.093	1.033	0.987
	Rec.	0.253	0.181	0.843	0.665	0.617	0.529	0.911	0.816
	Rec. 2	1.020	1.176	1.065	0.980	1.011	1.093	1.033	0.987
Quad. (1 dim.)	Sub.	0.657	1.054	0.759	1.097	0.861	0.989	1.022	0.908
	Rec.	0.533	1.298	1.135	0.375	1.733	1.278	2.243	0.481
	Rec. 2	1.299	1.503	1.026	0.136	2.186	1.503	1.702	0.545
Quad. (2 dim.)	Sub.	0.882	1.028	1.317	1.012	0.834	0.941	0.569	0.921
	Rec.	3.399	1.814	1.094	6.526	1.315	0.872	2.431	0.701
	Rec. 2	0.640	0.711	0.498	0.538	1.848	1.117	0.915	1.329
Quad. (3 dim.)	Sub.	1.638	1.856	1.003	1.398	1.060	0.964	1.003	0.440
	Rec.	6.870	5.692	1.441	4.737	0.113	3.327	3.864	2.089
	Rec. 2	0.502	1.777	0.850	0.445	1.859	1.420	1.124	1.116
Quad.(1 dim.) w/ Controls	Sub.	0.586	0.945	0.665	0.990	0.966	1.085	1.116	1.009
	Rec.	0.636	1.112	1.033	0.446	1.744	1.179	2.146	0.515
	Rec. 2	1.221	1.334	0.876	0.173	2.179	1.528	1.653	0.659

Table 10: **Smoking Simulation Results Vary Substantially By Factor Model Run**: This table shows the ratio of root mean squared-error to the naïve regression for 18 different specifications and four different runs of each factor model. Values above 1 indicate that the model is performing worse than the naïve regression and models below 1 indicate it is performing better. The left column provides the factor model and the second column provides the adjustment strategy. "Sub." is the substitute confounder; "Rec." is the reconstructed causes approach stated in the paper; and "Rec. 2" is the two-parameter reconstructed causes approach implemented in code. Models do not consistently perform better than naïve.

unstable under replication. We present these results simply to demonstrate that the ELBO does not provide a way to resolve the problem demonstrated in the previous subsection.

D.3 Conclusions on Smoking

The smoking simulation in Wang and Blei (2019) seeks to use a semisynthetic design to justify a number of conclusions about the deconfounder's performance. Unfortunately, as we have shown, these conclusions do not hold under reasonable adjustments and extensions to the simulation design.

Appendix E. Breast Cancer Tutorial

Summary: The github tutorial examines the effect of various tumor features on the diagnosis of breast cancer tumors. The tutorial uses approximate inference to fit a probabilistic

		Treatment 1	Treatment 2
Baseline	Naïve	1.000	1.000
	Oracle	0.044	0.025
	Controls	0.900	1.094
Lin. (1 dim.)	Sub.	1.062	1.033
	Rec.	1.325	1.151
	Rec. 2	1.061	1.033
Quad. (1 dim.)	Sub.	0.648	0.844
	Rec.	1.027	1.865
	Rec. 2	1.748	2.437
Quad. (2 dim.)	Sub.	2.648	0.308
	Rec.	4.073	1.214
	Rec. 2	0.718	1.353
Quad. (3 dim.)	Sub.	2.192	0.840
	Rec.	3.290	1.362
	Rec. 2	0.605	0.924
Quad.(1 dim.) w/ Controls	Sub.	0.566	0.947
	Rec.	1.166	1.889
	Rec. 2	1.670	2.421

Table 11: **Deconfounder Does Not Outperform the Naïve Regression In the Smoking Simulation**: This table shows the ratio of root mean squared-error to the naïve regression for 18 different specifications using the factor model which maximized the ELBO over twenty runs. Values above 1 indicate that the model is performing worse than the naïve regression and models below 1 indicate it is performing better. The left column provides the factor model and the second column provides the adjustment strategy. "Sub." is the substitute confounder; "Rec." is the reconstructed causes approach stated in the paper; and "Rec. 2" is the two-parameter reconstructed causes approach implemented in code. Models do not consistently perform better than naïve. See cautionary note in main text, results are very unstable.

principal components model to estimate the substitute confounder and then assert this provides valid causal estimates. This assertion is based on a non-standard assessments of whether a model is causal or not. We show that the full deconfounder is only estimable because the approximate inference leads to considerable noise in the estimated substitute confounder. When a more standard estimation procedure for the substitute confounder is deployed the full deconfounder is only estimable with a penalized regression. And the coefficient estimate that we obtain is entirely dependent on the amount of penalization. This demonstrates that the deconfounder is not particularly helpful for causal inference in this setting.

Using a breast cancer data set that is distributed with SciKit learn, the tutorial estimates a substitute confounder using black box variational inference. The tutorial argues that approximate inference is completely acceptable and can be ignored. We show this is not the case—approximate inference adds considerable noise to the estimated substitute

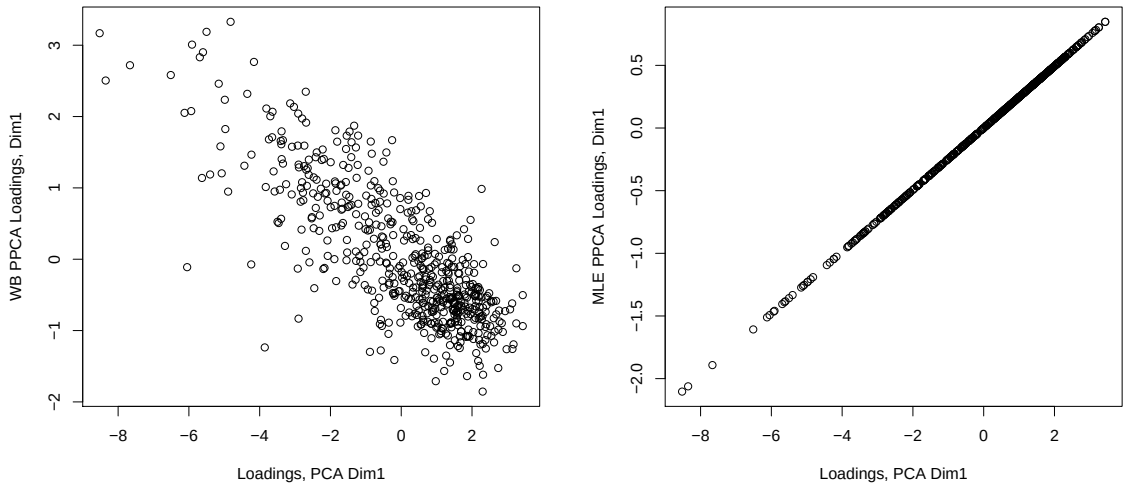

Figure 2: **The tutorial's procedure for estimating the substitute confounder adds unncessary noise.** The left-hand plots shows the relationship between the substitute confounder estimated using approximate inference and the substitute confounder estimated using traditional PCA. The approximate inference adds considerable noise. The right-hand plot shows that if we use well known MLE routines for estimating PPCA there is no disagreement between the loadings. Approximate inference, then, leads to considerable and unnecessary error.

confounder. Consider the left-hand facet of Figure 2 which compares the first dimension of the estimated substitute confounder from approximate inference (vertical axis) to the substitute confounder estimated using PCA. This shows that the estimated substitute confounder is a noisy version of PCA (we have not rotated the loadings, which explains the negative relationship). But the right-hand facet compares the estimate of the substitute confounder estimated from PPCA using maximum likelihood estimation against PCA. The estimated loadings using maximum likelihood to fit a PPCA model are effectively identical to the loadings from PCA, just scaled differently. In short, the approximate inference procedure leads to a poorly estimated model.

The result of this poorly estimated factor model is that including the substitute confounder has little effect on the actual coefficient estimates, yielding estimates that are nearly identical to a naïve regression. Column 1 of Table 12 provides the coefficient estimates of features on breast cancer diagnosis using the original estimates of the substitute confounder in a logistic model. Rather than make the non-standard step of subsetting to only 80% of the data, we fit the model to the entire data set. We want to emphasize that this model is estimable *solely* because the approximate inference procedure yields a poor estimate of the parameters of the PPCA model. As we might expect give this fact, the second column shows that a naïve logistic regression that simply ignores those confounders yields nearly identical results. In the third column we obtain coefficient estimates, but now we estimate the substitute confounder using standard PCA. Without the errors in the estimated substitute

confounder from approximate inference, the model is no longer estimable using a standard logistic regression. Instead we rendered the model estimable with a ridge regression, with the penalty selected using cross validation. This yields dramatically different results. Many of the coefficients have effect sizes that are orders of magnitude smaller. The final column shows the estimates from the one-at-a-time deconfounder, fit using a logistic regression. This reveals strikingly different results from those reported in the online tutorial: the sign flips on half of the coefficients.

	Full Deconfounder WB PCA	Naïve	Full Deconfounder True PCA	One-at-a-Time True PCA
mean radius	-3.10	-3.48	-0.81	-0.56
mean texture	-1.38	-1.64	-0.51	-0.07
mean smoothness	-0.87	-1.07	-0.38	-1.31
mean compactness	1.28	0.88	-0.29	2.59
mean concavity	-0.77	-1.19	-0.50	1.37
mean concave points	-1.78	-2.27	-0.75	-0.97
mean symmetry	-0.24	-0.50	-0.24	-0.34
mean fractal dimension	0.27	0.19	0.29	0.98

Table 12: **Correctly estimating PCA leads to dramatically different results** The first column replicates the exact procedure from the github tutorial, but estimate the coefficients on the entire sample. The second column merely drops the substitute confounder and yields very similar results. The third column estimates the full deconfounder using the true PCA estimates, using a penalized ridge regression to fit the model. This yields dramatically different results, consistent with our theoretical results. The fourth column provides the results from a one-at-a-time deconfounder. This shows even more deviations, with half of the coefficients changing signs, and many coefficients exhibiting orders of magnitude effect estimate changes.

Deviations

- We fit the outcome regression on the entire data set, rather than using an 80% held out data set.
- Given the severe errors in the approximate inference procedure, we use standard PCA estimation routines

Their Results: The tutorial claims this model provides robust causal effect estimates.
Our Results: We show that the model in WB's tutorial is estimable solely because of errors in the estimation of the PCA model. Once corrected, we obtain different coefficient results that vary substantially depending on how we apply the deconfounder.

Appendix F. Posterior Predictive Model Checking Does Not Reliably Identify When the Deconfounder Works

We have shown that it is impossible to know when the deconfounder improves over the naïve regression in practice. Throughout their papers, Wang and Blei (2019) and Zhang

et al. (2019) use a framework of posterior predictive checks (PPCs) to "greenlight" their use of the deconfounder in practice and adjudicate between alternative estimators. Wang and Blei (2019) explain,

> We consider predictive scores with predictive scores larger than 0.1 to be satisfactory; we do not have enough evidence to conclude significant mismatch of the assignment model. Note that the threshold of 0.1 is a subjective design choice. We find such assignment models that pass this threshold often yield satisfactory causal estimates in practice (Wang and Blei, 2019, p. 1581)

If PPCs could be used in this way, it would allow highly flexible density estimation models to be used, even when the true parametric form was unknown—as is always the case in practice. The proof of the subset deconfounder establishes that this is impossible in that setting because the performance of the subset deconfounder depends on untestable assumptions about the treatment effects. For the full deconfounder, the check in WB are not well-suited to evaluating the conditional independence of $\boldsymbol{A}$ given $\hat{\boldsymbol{Z}}$ which is perhaps the most relevant observable property (Imai and Jiang, 2019).

We evaluate the performance of PPCs on a quadratic-poisson factor model with $n = 10000$ observations and $m = 100$ treatments, where

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, 0.2) \\ A_{ij} &\sim \text{Poisson}\left(\exp\left(\theta_{j1} Z_i + \theta_{j2} Z_i^2\right)\right) \\ Y_i &\sim \mathcal{N}\left(\boldsymbol{A}_i \boldsymbol{\beta} + Z_i \gamma, 0.1\right) \end{aligned}$$

,

where $\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2} \sim \mathcal{N}(0, 1)$ and $\boldsymbol{\beta}$ is set equal to $(0.8, -0.6, 0.4, -1.2)$ repeated 25 times. We compare the naïve regression and the oracle to the two versions of the subset deconfounder: (i) using a Deep Exponential Family (DEF) (Ranganath et al., 2015) with (5,3,1) layers to estimate the substitute confounder, and (ii) using a two-dimensional PCA to estimate the substitute confounder.

The results in Figure 3 show the average RMSE for each adjustment strategy plotted at the corresponding PPC for the DEF. The average RMSE for the DEF-deconfounder is approximately equal, whether the model passes the PPC or not. Further, we see that there is considerable variation. We see that there can be extremely large RMSEs from estimation when the deconfounder passes the PPC and quite small when it fails the PPC. We also find that more complex models do not outperform simple alternatives. The average RMSE of the deconfounder using DEF is over three-times larger than the average RMSE when using PCA—even though the true underlying model that generated the treatments is nonlinear in the substitute confounder. But both implementations of the deconfounder perform considerably worse than the naive regression. The DEF deconfounder that pass the PPC has an average RMSE 5.8 times the average RMSE of the naive regression, while the PCA deconfounder has an average RMSE 1.8 times the average RMSE of the naive estimator. In every simulation, the average RMSE from the naive regression is better than either implementation of the deconfounder.

This suggests that PPCs cannot help us distinguish when the deconfounder will improve over alternatives. There is considerable noise in the RMSE of models that pass or do

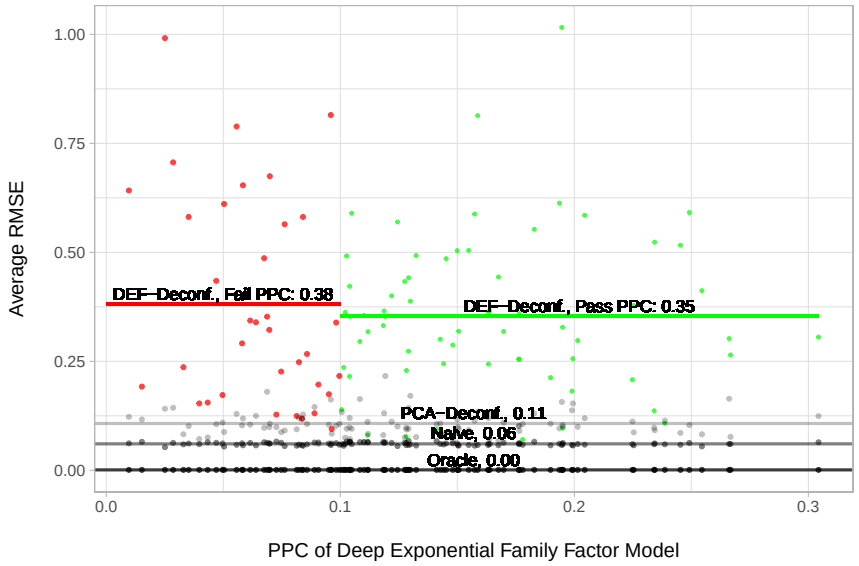

Figure 3: **Posterior Predictive Checks Do Not Reliably Identify When the Deconfounder Outperforms the Naïve Estimator.** The horizontal axis is the predictive score from the posterior predictive check of the DEF and the vertical axis is the average RMSE for the treatment effect estimates. The red points are the average RMSE from applications of the DEF that failed the predictive score check, while the green points are the average RMSE from applications of the DEF that passed the predictive score check.

not pass PPCs, so the safest conclusion is that there is no difference between the RMSE of applications of the deconfounder that do and do not pass the PPC. The simulation also demonstrates that if the functional form of the factor model is unknown, a flexible factor model can perform considerably worse than simpler models, even when the true data generating process is nonlinear and the flexible model passes model-fit checks. Most importantly, the deconfounder's estimates can be considerably worse than a naïve estimator which simply ignores confounding. In short, model checking and flexible nonlinear factor models cannot solve the deconfounder's problems.

Appendix G. Additional Results from Actor Application

We offer several variants of the deconfounder results below. "WB Deconfounder" is the cached results of the Poisson Matrix Factorization run in Wang and Blei (2019). The "Deconfounder, Full" is our re-estimated model using the same code. In Wang and Blei (2019), the analyses did not condition on any observed covariates. After we shared our draft with Wang and Blei in July 2020, they posted a reference implementation of the deconfounder that provided an illustration of conditioning on budget and runtime.

G.1 Top Actors by Multiplicative Effect $e^{\hat{\beta}_j}$

Naive, Full Stan Lee (×9.31), John Ratzenberger (×9.26), Sacha Baron Cohen (×7.09), Leonardo DiCaprio (×5.50), Josh Hutcherson (×5.19), Corey Burton (×4.80), Brian Doyle Murray (×4.79), Tom Hanks (×4.76), Julie Andrews (×4.64), Desmond Llewelyn (×4.55), Will Smith (×4.36), Ava Acres (×4.24), Crispin Glover (×4.12), Warwick Davis (×4.10), Andy Serkis (×4.02), Eddie Murphy (×3.88), Shia LaBeouf (×3.81), Tomas Arana (×3.79), Judi Dench (×3.75), Penélope Cruz (×3.66), Tom Cruise (×3.49), Blythe Danner (×3.46), Jay Baruchel (×3.46), Harrison Ford (×3.42), Michael Douglas (×3.37)

WB Deconfounder, Full Stan Lee (×9.29), John Ratzenberger (×8.29), Sacha Baron Cohen (×8.12), Josh Hutcherson (×5.02), Corey Burton (×4.91), Leonardo DiCaprio (×4.87), Tom Hanks (×4.71), Ava Acres (×4.55), Julie Andrews (×4.38), Will Smith (×4.37), Desmond Llewelyn (×4.19), Crispin Glover (×4.19), Eddie Murphy (×4.18), Andy Serkis (×4.14), Judi Dench (×3.99), Shia LaBeouf (×3.92), Brian Doyle Murray (×3.91), Tobin Bell (×3.63), Tomas Arana (×3.61), George Lopez (×3.56), Warwick Davis (×3.55), Blythe Danner (×3.52), Mary Elizabeth Winstead (×3.52), Tom Cruise (×3.46), Penélope Cruz (×3.45)

Deconfounder, Full Courteney Cox (×15.32), Tom Cruise (×14.74), John Ratzenberger (×11.13), Vera Farmiga (×9.86), Sacha Baron Cohen (×9.83), Kim Coates (×9.76), Charlize Theron (×8.74), James Badge Dale (×8.65), Patrick Wilson (×8.58), Penélope Cruz (×8.41), Rafe Spall (×8.09), Tom Hanks (×8.02), Jeffrey Tambor (×7.63), Albert Finney (×7.17), Crispin Glover (×6.94), Garrett Hedlund (×6.86), Octavia Spencer (×6.78), Emile Hirsch (×6.75), Leonardo DiCaprio (×6.61), Tim Robbins (×6.48), Chiwetel Ejiofor (×6.45), Ewen Bremner (×6.39), John Heard (×6.20), Johnny Knoxville (×6.01), George Clooney (×6.00)

Budget Adjustment, Full Sacha Baron Cohen (×6.93), Brian Doyle Murray (×4.08), Conrad Vernon (×4.04), Julie Andrews (×3.84), Tomas Arana (×3.83), Crispin Glover (×3.73), John Ratzenberger (×3.55), Tobin Bell (×3.46), Leonardo DiCaprio (×3.41), Keira Knightley (×3.29), Reese Witherspoon (×3.27), Josh Hutcherson (×3.10), Desmond Llewelyn (×3.09), Tom Hanks (×2.92), George Lopez (×2.87), James Rebhorn (×2.85), Clea DuVall (×2.85), Jason Sudeikis (×2.83), Allen Covert (×2.79), Alan Rickman (×2.76), Anna Kendrick (×2.75), Jodie Foster (×2.73), Andy Samberg (×2.69), John Heard (×2.64), Michael Kelly (×2.64)

Naive, One at a Time Jess Harnell (×12.28), Ava Acres (×10.16), Warwick Davis (×10.09), Stan Lee (×9.85), Orlando Bloom (×9.50), Hugo Weaving (×8.42), Chris Miller (×8.35), Conrad Vernon (×7.92), John Ratzenberger (×7.91), Mickie McGowan (×7.67), Julie Walters (×7.11), Christopher Lee (×7.03), Danny Mann (×6.98), Lasco Atkins (×6.50), Ian McKellen (×6.44), Tom Felton (×6.26), Daniel Radcliffe (×6.23), Andy Serkis (×6.19), Sacha Baron Cohen (×6.10), Timothy Spall (×6.03), Frank Oz (×6.01), Emma Watson (×5.82), Pat Kiernan (×5.69), Bonnie Hunt (×5.68), Denis Leary (×5.26)

WB Deconfounder, One at a Time Jess Harnell (×13.49), Ava Acres (×10.49), Chris Miller (×9.19), Orlando Bloom (×9.02), Stan Lee (×8.77), Conrad Vernon (×8.61), Hugo Weaving (×7.72), Warwick Davis (×7.39), John Ratzenberger (×7.35), Mickie McGowan

(×7.09), Christopher Lee (×6.86), Danny Mann (×6.50), Julie Walters (×6.27), Ian McKellen (×6.21), Lasco Atkins (×6.16), Denis Leary (×5.68), Sacha Baron Cohen (×5.67), Tom Felton (×5.57), John DiMaggio (×5.56), Chris Ellis (×5.43), Andy Serkis (×5.41), Julie Andrews (×5.37), Frank Oz (×5.22), Bonnie Hunt (×5.08), Timothy Spall (×5.06)

Deconfounder, One at a Time Lasco Atkins (×5.64), Sacha Baron Cohen (×4.24), John Ratzenberger (×4.01), Desmond Llewelyn (×3.91), Will Smith (×3.64), Tom Cruise (×3.63), Brad Garrett (×3.56), Hugo Weaving (×3.46), Ving Rhames (×3.38), Ian McKellen (×3.30), Naomie Harris (×3.27), Jeffrey Tambor (×3.13), Warwick Davis (×3.10), Ava Acres (×3.06), Orlando Bloom (×3.06), Jet Li (×2.94), Julie Walters (×2.92), Brent Spiner (×2.91), Stan Lee (×2.91), Lois Maxwell (×2.90), Corey Burton (×2.79), Christoph Waltz (×2.78), Jess Harnell (×2.77), Michelle Rodriguez (×2.72), Judi Dench (×2.66)

G.2 Top Actors by Appearance-weighted Log-scale Coefficients $n_j\hat{\beta}_j$

Naive, Full Stan Lee (58.01), John Ratzenberger (48.98), Tom Hanks (46.81), Tom Cruise (41.21), Harrison Ford (36.87), Arnold Schwarzenegger (36.84), Morgan Freeman (36.02), Will Smith (35.36), Eddie Murphy (32.55), Leonardo DiCaprio (32.41), Brad Pitt (31.35), Bruce Willis (30.70), Judi Dench (30.42), Robert De Niro (29.12), John Travolta (28.56), Denzel Washington (27.26), Jim Carrey (26.84), Jack Black (26.83), Robin Williams (26.06), Desmond Llewelyn (25.75), John Leguizamo (23.72), Scarlett Johansson (22.62), Octavia Spencer (21.71), Leslie Mann (21.57), Matt Damon (21.46)

WB Deconfounder, Full Stan Lee (57.96), John Ratzenberger (46.53), Tom Hanks (46.48), Tom Cruise (40.95), Harrison Ford (37.05), Arnold Schwarzenegger (36.59), Morgan Freeman (35.86), Will Smith (35.39), Eddie Murphy (34.33), Brad Pitt (32.93), Judi Dench (31.80), Leonardo DiCaprio (30.07), Bruce Willis (29.15), Robin Williams (28.20), Jack Black (28.01), Robert De Niro (27.60), Jim Carrey (27.26), John Travolta (26.68), John Leguizamo (25.68), Liam Neeson (25.24), Denzel Washington (24.91), Desmond Llewelyn (24.36), Matt Damon (24.22), Scarlett Johansson (24.21), Octavia Spencer (24.11)

Deconfounder, Full Tom Cruise (88.80), Tom Hanks (62.45), George Clooney (55.52), John Ratzenberger (53.01), Octavia Spencer (44.04), Charlize Theron (43.36), Mark Wahlberg (42.11), Bruce Willis (40.40), Morgan Freeman (38.59), Ryan Reynolds (38.07), Harrison Ford (37.17), Leonardo DiCaprio (35.88), Will Smith (35.71), Jim Broadbent (34.99), Jason Statham (34.81), Stellan Skarsgård (32.83), Jeffrey Tambor (32.52), Patrick Wilson (32.25), Jason Flemyng (31.98), Penélope Cruz (31.94), Zoe Saldana (30.97), Laurence Fishburne (30.86), Sylvester Stallone (30.23), James Remar (30.12), Mickey Rourke (29.07)

Budget Adjustment, Full Tom Hanks (32.14), John Ratzenberger (27.85), Harrison Ford (27.61), Morgan Freeman (23.66), Leonardo DiCaprio (23.30), Octavia Spencer (21.69), Reese Witherspoon (21.35), Robert De Niro (21.35), Tom Cruise (21.23), Denzel Washington (20.95), Scarlett Johansson (20.85), Eddie Murphy (20.59), Keira Knightley (20.27), Ralph Fiennes (20.20), Sacha Baron Cohen (19.36), Alan Rickman (19.27), Desmond Llewelyn (19.17), John Leguizamo (19.00), Jim Carrey (18.88), George Clooney (18.87), John Travolta (18.86), Robin Williams (18.82), Philip Seymour Hoffman (18.16), Patricia Clarkson (18.02), Brad Pitt (17.34)

Naive, One at a Time Stan Lee (59.49), John Ratzenberger (45.50), Tom Cruise (42.97), Morgan Freeman (42.31), Hugo Weaving (40.47), Tom Hanks (37.46), Samuel L Jackson (36.98), Frank Welker (36.69), Will Smith (35.78), Jess Harnell (35.11), Ian McKellen (33.52), Christopher Lee (31.19), Bill Hader (30.36), Liam Neeson (29.64), Bruce Willis (29.30), Orlando Bloom (29.26), Cameron Diaz (29.09), Brad Pitt (28.29), Judi Dench (27.80), Warwick Davis (27.74), Danny Mann (27.20), Cate Blanchett (27.14), Stellan Skarsgård (27.05), Alan Tudyk (26.92), Harrison Ford (26.72)

WB Deconfounder, One at a Time Stan Lee (56.45), John Ratzenberger (43.90), Tom Cruise (40.47), Morgan Freeman (39.43), Hugo Weaving (38.83), Tom Hanks (38.07), Jess Harnell (36.42), Frank Welker (35.07), Will Smith (34.04), Samuel L Jackson (32.97), Ian McKellen (32.86), Liam Neeson (32.22), Bill Hader (31.23), Christopher Lee (30.82), Cameron Diaz (30.40), Judi Dench (29.17), Bruce Willis (28.90), Cate Blanchett (28.68), Orlando Bloom (28.59), Brad Pitt (28.00), Jonah Hill (27.59), Stellan Skarsgård (27.30), Alan Tudyk (27.15), Danny Mann (26.21), Harrison Ford (26.00)

Deconfounder, One at a Time Tom Cruise (42.54), Will Smith (31.01), John Ratzenberger (30.58), Morgan Freeman (29.26), Harrison Ford (27.97), Stan Lee (27.76), Tom Hanks (25.50), Arnold Schwarzenegger (24.14), Frank Welker (23.91), Hugo Weaving (23.59), Desmond Llewelyn (23.18), Liam Neeson (22.72), Jim Carrey (22.59), Judi Dench (22.51), Ving Rhames (21.90), Ian McKellen (21.52), Bruce Willis (21.47), Robin Williams (19.48), Jeffrey Tambor (18.27), Patrick Stewart (18.06), Mark Wahlberg (17.77), Alan Tudyk (17.41), Lasco Atkins (17.31), Hugh Jackman (16.05), Lois Maxwell (15.98)