A Proxy Variable View of Shared Confounding

Yixin Wang [1] **David M. Blei** [2]

Abstract

Causal inference from observational data can be biased by unobserved confounders. Confounders—the variables that affect both the treatments and the outcome—induce spurious non-causal correlations between the two. Without additional conditions, unobserved confounders generally make causal quantities hard to identify. In this paper, we focus on the setting where there are many treatments with shared confounding, and we study under what conditions is causal identification possible. The key observation is that we can view subsets of treatments as proxies of the unobserved confounder and identify the intervention distributions of the rest. Moreover, while existing identification formulas for proxy variables involve solving integral equations, we show that one can circumvent the need for such solutions by directly modeling the data. Finally, we extend these results to an expanded class of causal graphs, those with other confounders and selection variables.

1. Introduction

Causal inference from observational data can be biased by unobserved confounders. Confounders are variables that affect both the treatments and the outcome. When measured, we can account for them with adjustments (Pearl, 2009). But when unobserved, they open back-door paths that bias the causal inference; back-door adjustments are not possible.

Consider the following causal problem. How does a person's diet affect her body fat percentage? One confounder is lifestyle: someone with a healthy lifestyle will eat healthy foods such as boiled broccoli; but she will also exercise frequently, which lowers her body fat. When lifestyle is unobserved, the composition of diet will be correlated with body fat, regardless of its true causal effect. Compounding the difficulty, accurate measurements of lifestyle (the confounder) are difficult to obtain, e.g., requiring expensive real-time tracking of activities. Lifestyle is necessarily an unobserved confounder.

Here we focus on the setting where multiple treatments share the same unobserved confounder. The example fits into this setting. Each type of food—broccoli, burgers, granola bars, pizza, and so on—is a potential "treatment" for body fat. Further, each person's lifestyle affects multiple treatments, i.e., their consumption of multiple types of food. People with a healthy lifestyle eat broccoli and granola; people with an unhealthy lifestyle eat pizza and burgers. Thus the different foods share the same unobserved confounder, i.e. each person's lifestyle.

When multiple treatments share the same unobserved confounding, which causal quantities can be identified? How can we estimate them? These are the questions we address.

Begin with the causal graph of Figure 1a, where an unobserved confounder U (lifestyle) affects multiple treatments $\{A_1, \ldots, A_m\}$ (food choices) and an outcome Y (body fat). Further consider a subset of treatments $\mathcal{C}$. We prove that, under suitable conditions, the intervention distribution $p(y \mid \mathrm{do}(a_{\mathcal{C}}))$ is identifiable.

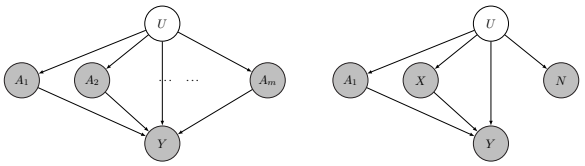
The key observation is that, under shared confounding, some treatments can serve as proxies of unobserved confounders (Miao et al., 2018; Kuroki & Pearl, 2014), enabling causal identification of other treatments. This observation helps identify the intervention distributions of subsets of treatments. Unlike prior work, we do not need to find two external proxies for the unobserved confounder; some treatments themselves can serve as proxies for other treatments.

We then turn to estimation. The identification formula we obtain requires solving an integral equation (Miao et al., 2018), which might be difficult. We show that the deconfounder algorithm of Wang & Blei (2019a) can help bypass this requirement, producing correct causal estimates by directly modeling the data. With a simulation study, we demonstrate that the identification conditions we require are crucial for the algorithm to produce correct causal inferences. We note that, while we use the same algorithm, the theoretical setting considered here is different from Wang & Blei (2019a).

We finally generalize the identification and estimation results to an expanded class of graphs in Figure 2b. This class

[1]University of California, Berkeley [2]Columbia University. Correspondence to: David M. Blei <david.blei@columbia.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.



(a) (b)

Figure 1. (a) Multiple treatments with shared confounding. (b) Proxy variables for an unobserved confounder (Miao et al., 2018). (Only the shaded nodes are observed.)

contains shared confounding, measured single-treatment confounders (that only affect one treatment), and selection on the unobservables. We establish identifiability as well as the applicability of the deconfounder in this larger class.

Contributions. The main contributions of this paper are identification and estimation results that target multiple treatments with shared confounding, allowing for certain types of selection bias. We derive conditions under which the intervention distributions of the treatments are identifiable and further conditions under which the deconfounder algorithm can produce correct causal inference. The key idea is to use some treatments as proxies of the unobserved confounder to identify the effect of other treatments. Rather than solving integral equations, the algorithm estimates the intervention distributions by directly modeling the data.

Related work. This work uses and extends causal identification with proxy variables (Kuroki & Pearl, 2014; Miao et al., 2018; Shi et al., 2020). While these works focus on a single treatment and a single outcome, we leverage the multiplicity of the treatments to establish causal identification. With multiple treatments, the recent work of Miao et al. (2020) proposes two approaches to identifying the intervention distributions: the auxiliary variable approach and the null treatments approach. These approaches utilize the shared confounding structure via assuming at least half of the confounded treatments do not causally affect the outcome, and the treatment-confounder distribution is identifiable from observational data. Our approach differs from this approach in how we leverage the shared confounding structure for causal identification. As the treatments share the same unobserved confounder, we view some treatments as proxies of the shared unobserved confounder for identifying the effects of the other treatments.

A second body of related work is on causal inference with multiple treatments (Ranganath & Perotte, 2018; Heckerman, 2018; Janzing & Schölkopf, 2018; D'Amour, 2019b; Frot et al., 2017; Ćevid et al., 2018; Wang et al., 2017; Tran & Blei, 2017; Wang & Blei, 2019a; Puli et al., 2020). While many of these works focus on developing algorithms, we focus on theoretical aspects of the problem. The deconfounder algorithm that we use was developed in Wang & Blei (2019a) and has been heavily debated and discussed (Ogburn et al., 2019; 2020; Imai & Jiang, 2019; Grimmer et al., 2020; D'Amour, 2019a;b; Wang & Blei, 2020; 2019b). Here we delineate settings and assumptions, different from those in Wang & Blei (2019a), where the algorithm provides correct causal inferences. We also demonstrate that the effectiveness of the algorithm in practice relies on these assumptions.

The identification results in this paper differ from those in Wang & Blei (2019a). First, that work assumes the unobserved confounder is a deterministic function of the treatments; in contrast, we allow the substitute confounder to be random given the treatments. Second, we establish identification by assuming the existence of a function of the treatments that does not affect the outcome; this assumption is not made in Wang & Blei (2019a). Finally, we extend the ideas to allow for selection bias (Bareinboim & Pearl, 2012), including selection driven by unobserved confounders.

We note that D'Amour (2019b) provides negative examples of causal identification where some intervention distributions are not identifiable; it also suggests collecting additional proxy variables to resolve non-identification. The results below do not contradict those of D'Amour (2019b). Rather, we focus on the intervention distributions of *subsets* of the treatments; D'Amour (2019b) focuses on the intervention distributions of *all* the treatments. Further, the way we use proxy variables differs in that we use existing causes as proxy variables, as opposed to collecting additional proxies.

2. Multiple treatments & shared confounders

Consider a causal inference problem where multiple treatments of interest affect a single outcome. It deviates from classical causal inference, where the main interest is a single treatment and a single outcome.

Figure 1a provides an example. There are m treatments $A_1, \ldots, A_m$ that all affect the outcome Y; and there is an unobserved confounder U that affects Y and the treatments. This graph exemplifies *shared unobserved confounding*, where U affects multiple treatments.

In this paper, the goal is to estimate the intervention distributions on subsets of treatments, $P(Y \mid \mathrm{do}(A_{\mathcal{C}} = a_{\mathcal{C}}))$. It is the distribution of the outcome Y if we intervene on $A_{\mathcal{C}} \subset \{A_1, \ldots, A_m\}$, which is a (strict) subset. (E.g., if we are interested in each treatment individually then each subset contains one treatment.) We will establish causal identification and then discuss an algorithm for estimation. Section 3 extends these results to an expanded class of graphs.

2.1. Causal identification

An intervention distribution is *identifiable* if it can be written as a function of the observed data distribution (e.g., $P(y, a_1, \ldots, a_m)$ in Figure 1a) (Pearl, 2009). In Figure 1a, which intervention distributions can be identified? In this section we prove that, under suitable conditions, the intervention distributions of subsets of the treatments $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$ are identifiable.[1]

The starting point for causal identification with multiple treatments is the *proxy variable* strategy, which focuses on causal identification with a single treatment (Kuroki & Pearl, 2014; Miao et al., 2018). Consider the causal graph in Figure 1b: it has a single treatment A_1, an outcome Y, and an unobserved confounder U. The goal is to estimate the intervention distribution $P(y \mid \mathrm{do}(a_1))$. There are some other variables in the graph too. A *proxy* X is an observable child of the unobserved confounder; a *null proxy* N is a proxy that does not affect the outcome. The theory around proxy variables says that the intervention distribution $P(y \mid \mathrm{do}(a_1))$ is identifiable if (1) we observe two proxies of the unobserved confounder U and (2) one of the proxies is a null proxy (Miao et al., 2018). In particular, since N and X are observed, $P(y \mid \mathrm{do}(a_1))$ is identifiable.

We leverage the idea of proxy variables to identify intervention distributions in Figure 1a, multiple treatments with shared unobserved confounding. The main idea is to use some treatments as proxies to identify the intervention distributions of other treatments. The benefit is that, with multiple treatments, we do not need to observe external proxy variables; rather the treatments themselves serve as proxies. Nor do we need to observe a null proxy, one that does not affect the outcome (like N in Figure 1b); we only need to assume that there is a function of the treatments that does not affect the outcome. (We do not need to know this function either, just that at least one such function exists.) In short, we can use the idea of the proxy but without collecting external data; we can work solely with the data about the treatments and the outcome.

We formally state the identification result. To repeat, assume the causal graph in Figure 1a with m treatments $A_{1:m}$, an outcome Y, and a shared unobserved confounder U. The goal is to identify the intervention distribution of a *strict subset* of the treatments $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

Partition the m treatments into three sets: $A_{\mathcal{C}}$ is the set of treatments on which we intervene; $A_{\mathcal{X}}$ is the set of treatments we use as a proxy; $A_{\mathcal{N}}$ is the set of treatments such that there exists a function $f(A_{\mathcal{N}})$ that can serve as a null proxy. (We discuss this assumption below.) The latter two sets mimic the proxy X and the null proxy N in the proxy variable strategy. Sets $A_{\mathcal{C}}$, $A_{\mathcal{X}}$ and $A_{\mathcal{N}}$ must be non-empty.

Assumption 1. *There exists some function f and a set $\emptyset \neq \mathcal{N} \subset \{1, \ldots, m\} \backslash \mathcal{C}$ such that*

1. *The outcome Y does not depend on $f(A_{\mathcal{N}})$:*

$$f(A_{\mathcal{N}}) \perp Y \mid U, A_{\mathcal{C}}, A_{\mathcal{X}}, \tag{1}$$

   where $\mathcal{X} = \{1, \ldots, m\} \backslash (\mathcal{C} \cup \mathcal{N}) \neq \emptyset$.

2. *The conditional distribution $P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}))$ is complete[2] in $f(a_{\mathcal{N}})$ for almost all $a_{\mathcal{C}}$.*

3. *The conditional distribution $P(f(a_{\mathcal{N}}) \mid a_{\mathcal{C}}, a_{\mathcal{X}})$ is complete in $a_{\mathcal{X}}$ for almost all $a_{\mathcal{C}}$.*

Assumption 1.1 posits that a set of treatments $A_{\mathcal{N}}$ exists such that some function of them $f(A_{\mathcal{N}})$ can serve as a null proxy (Eq. 1). Roughly, it requires $f(A_{\mathcal{N}})$ does not affect the outcome. It does not require that we know $\mathcal{N}$ or $f(A_{\mathcal{N}})$, just that they exist.

When might this assumption be satisfied? First, suppose some of the multiple treatments do not affect the outcome. Then Assumption 1.1 reduces to the null proxy assumption (Kuroki & Pearl, 2014; Miao et al., 2018; D'Amour, 2019b). This might be plausible, e.g., in a genetic study or other setting where there are many treatments. Again, we do not need to know *which* treatments are "null treatments." Indeed, as long as two treatments are null, the theory below implies that the intervention distributions of each individual treatment is identifiable.

But this assumption goes beyond a restatement of the null proxy assumption. Suppose two (or more) treatments only affect the outcome as a bundle. Then the bundle can form the set $\mathcal{N}$ and the function is one that is "orthogonal" to how they are combined. As a (silly) example, consider two of the treatments to be bread and butter. Suppose they must be served together to induce the joyfulness of food, but not individually. (If either is served alone, it has no effect on joyfulness one way or the other.) Then the function $f(A_{\mathcal{N}})$ is XOR of the bundle; the quantity (bread XOR butter) does not affect Y. Again, the function and set must exist; we do not need to know them.

As a more serious example, consider that HDL cholesterol, LDL cholesterol, and triglycerides (TG) affect the risk of a

[1] We abbreviate $P(y \mid \mathrm{do}(a_{\mathcal{C}})) \triangleq P(y \mid \mathrm{do}(A_{\mathcal{C}} = a_{\mathcal{C}}))$.

[2] Definition of "complete": The conditional distribution $P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}))$ is complete in $f(a_{\mathcal{N}})$ for almost all $a_{\mathcal{C}}$ means for any square-integrable function $g(\cdot)$ and almost all $a_{\mathcal{C}}$,

$$\int g(u, a_{\mathcal{C}}) P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}u = 0 \text{ for almost all } f(a_{\mathcal{N}})$$

if and only if $g(u, a_{\mathcal{C}}) = 0$ for almost all u.

heart attack through the ratios HDL/LDL and TG/HDL (Millán et al., 2009). Then HDL×LDL and TG×HDL are both examples of $f(A_{\mathcal{N}})$ that do not affect Y. The existence of one of them suffices for Assumption 1.1. (We discuss this assumption in more technical detail in Appendix B.)

Assumption 1.2 and Assumption 1.3 are two completeness conditions on the true causal model; they are required by the proxy variable strategy (e.g. Conditions 2 and 3 of Miao et al. (2018)). Roughly, they require that the distributions of U corresponding to different values of $f(A_{\mathcal{N}})$ are distinct; the distributions of $f(A_{\mathcal{N}})$ relative to different $A_{\mathcal{X}}$ values are also distinct.

The two assumptions are satisfied when we work with a causal model that satisfies the completeness condition. Many common models satisfy this condition. Examples include exponential families (Newey & Powell, 2003), location-scale families (Hu & Shiu, 2018), and nonparametric regression models (Darolles et al., 2011). Completeness is a common assumption posited in nonparametric causal identification (Miao et al., 2018; Yang et al., 2017; D'Haultfoeuille, 2011); it is often used to guarantee the existence and the uniqueness of solutions to integral equations. Chen et al. (2014) provides a discussion of completeness.

Under Assumption 1, we can identify the intervention distribution of the subset of the treatments $A_{\mathcal{C}}$.

Theorem 1. *(Causal identification under shared confounding) Assume the causal graph Figure 1a. (Note the data does not need to be "faithful" to the graph—some edges can be missing.) Under Assumption 1, the intervention distribution of the treatments $A_{\mathcal{C}}$ is identifiable:*

$$P(y \mid \mathrm{do}(a_{\mathcal{C}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}}) \, \mathrm{d}a_{\mathcal{X}} \tag{2}$$

for any solution h to the integral equation

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}}. \tag{3}$$

Moreover, the solution to Eq. 3 always exists under weak regularity conditions in Appendix D.

Proof sketch. The proof relies on the partition of the m treatments: $A_{\mathcal{C}}$ as the treatments, $A_{\mathcal{X}}$ as the proxies, and $A_{\mathcal{N}}$ such that $f(A_{\mathcal{N}})$ can be a null proxy. We then follow the proxy variable strategy to identify the intervention distributions of $A_{\mathcal{C}}$ using $A_{\mathcal{X}}$ as a proxy and $f(A_{\mathcal{N}})$ as a null proxy. We no longer have a null proxy like N as in Figure 1b; all the m treatments can affect the outcome. However, Assumption 1.1 allows $f(A_{\mathcal{N}})$ to play the role of a null proxy. The full proof is in Appendix A. □

Theorem 1 identifies the intervention distributions of subsets of the treatments $A_{\mathcal{C}}$; it writes $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$ as a function of the observed data distribution $P(y, a_{\mathcal{C}}, a_{\mathcal{X}}, a_{\mathcal{N}})$. In particular, it lets us identify the intervention distributions of individual treatments $P(y \mid \mathrm{do}(a_i)), i = 1, \ldots, m$. By using the treatments themselves as proxies, Theorem 1 exemplifies how the multiplicity of the treatments enables causal identification under shared unobserved confounding.

2.2. Causal estimation with the deconfounder

Theorem 1 guarantees that the intervention distribution $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$ is estimable from the observed data. However, it involves solving an integral equation (Eq. 3). This integral equation is hard to solve except in the simplest linear Gaussian case (Carrasco et al., 2007). How can we estimate $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$ in practice?

We revisit the deconfounder algorithm in Wang & Blei (2019a). We show that the deconfounder correctly estimates the intervention distribution $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$; it implicitly solves the integral equation in Eq. 3 by modeling the data. (This is an alternative justification of the algorithm from Wang & Blei (2019a).)

We first review the algorithm. Given the treatments $A_1, \ldots, A_m$ and the outcome Y, the deconfounder proceeds in three steps:

1. **Construct a substitute confounder.** Based *only* on the (observed) treatments $A_1, \ldots, A_m$, it first constructs a random variable $\hat{Z}$ such that all the treatments are conditionally independent:

$$\hat{P}(a_1, \ldots, a_m, \hat{z}) = \hat{P}(\hat{z}) \prod_{j=1}^{m} \hat{P}(a_j \mid \hat{z}), \tag{4}$$

   where $\hat{P}(\cdot)$ is consistent with the observed data $P(a_1, \ldots, a_m) = \int \hat{P}(a_1, \ldots, a_m, \hat{z}) \, \mathrm{d}\hat{z}$. The random variable $\hat{Z}$ is called a *substitute confounder*; it does not necessarily coincide with the unobserved confounder U. The substitute is constructed using probabilistic models with local and global variables (Bishop, 2006), such as probabilistic PCA (Tipping & Bishop, 1999).

2. **Fit an outcome model.** The next step is to estimate how the outcome depends on the treatments and the substitute confounder $\hat{P}(y \mid a_1, \ldots, a_m, \hat{z})$. This *outcome model* is fit to be consistent with the observed data:

$$P(y, a_1, \ldots, a_m) = \int \hat{P}(y \mid a_1, \ldots, a_m, \hat{z}) \hat{P}(a_1, \ldots, a_m, \hat{z}) \, \mathrm{d}\hat{z}. \tag{5}$$

   Along with the first step, the deconfounder gives the joint distribution $\hat{P}(y, a_1, \ldots, a_m, \hat{z})$.

3. **Estimate the intervention distribution.** The final step estimates the intervention distribution $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$ by

integrating out the non-intervened treatments and the substitute confounder,

$$\hat{P}(y \mid \mathrm{do}(a_{\mathcal{C}})) \triangleq \int \hat{P}(y \mid a_1, \ldots, a_m, \hat{z}) \times \hat{P}(a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \hat{z})\, \mathrm{d}\hat{z}\, \mathrm{d}a_{\{1,\ldots,m\}\setminus\mathcal{C}}. \quad (6)$$

This is the estimate.

The correctness of the deconfounder. Note that many possible $\hat{P}(\cdot)$'s satisfy the deconfounder requirements (Eqs. 4 and 5); the algorithm outputs one such $\hat{P}$. Under suitable conditions, we show that any such $\hat{P}$ provides the correct causal estimate $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

Assumption 2. *The deconfounder estimate $\hat{P}(y, a_1, \ldots, a_m, \hat{z})$ satisfies two conditions:*

1. *It is consistent with Assumption 1.1, $\hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}) = \hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, \hat{z})$.*
2. *The conditional distribution $\hat{P}(\hat{z} \mid a_{\mathcal{C}}, a_{\mathcal{X}})$ is complete in $a_{\mathcal{X}}$ for almost all $a_{\mathcal{C}}$.*

Assumption 2.1 roughly requires that there exists a function f and a subset of the treatments $A_{\mathcal{N}}$ such that $f(A_{\mathcal{N}})$ does not affect the outcome in the deconfounder outcome model. (When the number of treatments goes to infinity, Assumption 2.1 reduces to Assumption 1.1.) We emphasize that $f(A_{\mathcal{N}})$ is not involved in calculating the estimate (Eq. 6); it only appears in Assumption 2.1. Hence the correctness of the algorithm does not require specifying $f(\cdot)$ and $A_{\mathcal{N}}$, just that it exists.

Assumption 2.2 requires that the distributions of $\hat{Z}$ corresponding to different values of $A_{\mathcal{X}}$ are distinct. It is a similar completeness condition as in Assumption 1.

Now we state the correctness of the algorithm.

Theorem 2. *(Correctness of the deconfounder under shared confounding) Assume the causal graph Figure 1a. Under Assumption 1, Assumption 2 and weak regularity conditions, the deconfounder provides correct estimates of the intervention distribution:*

$$\hat{P}(y \mid \mathrm{do}(a_{\mathcal{C}})) = P(y \mid \mathrm{do}(a_{\mathcal{C}})), \quad (7)$$

where $\hat{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ is computed from Eq. 6.

Proof sketch. The proof of Theorem 2 relies on a key observation: the deconfounder implicitly solves the integral equation (Eq. 3) by modeling the observed data with $\hat{P}(y, a_1, \ldots, a_m, \hat{z})$. Assumption 2.2 guarantees that the deconfounder estimate can be written as

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z})\, \mathrm{d}a_{\mathcal{X}} \quad (8)$$

under weak regularity conditions; this function $\hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}})$ also solves the integral equation (Eq. 3). The deconfounder uses this solution to form an estimate of $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$; this estimate is correct because of Theorem 1. The full proof is in Appendix C. □

Theorem 2 justifies the deconfounder for multiple causal inference under shared confounding (Figure 1a). It proves that the deconfounder correctly estimates the intervention distributions when they are identifiable. This result complements Theorems 6–8 of Wang & Blei (2019a); it establishes identification and correctness by assuming there exists some function of the treatments that does not affect the outcome. In contrast, Theorems 6–8 of Wang & Blei (2019a) assume a "consistent substitute confounder," that the substitute confounder is a deterministic function of the treatments. Their assumption is stronger; conditional on the treatments, Theorems 1 and 2 allow the substitute confounder to be random.

Theorem 2 also shows that we can leverage the deconfounder algorithm to put the proxy variable strategy into practice. While existing identification formulas of proxy variables involves solving integral equations (Miao et al., 2018), Theorem 2 shows how to circumvent this need by directly modeling the data and applying the deconfounder; it implicitly solves the integral equations.

Section 4 illustrates these theorems with a linear example.

3. An expanded class of causal graphs

We discussed causal identification and estimation when multiple treatments share the same unobserved confounder. We now extend these results to an expanded class of causal graphs, those with several types of nodes and, in particular, those that include a selection variable (Bareinboim et al., 2014; Bareinboim & Pearl, 2012). Using the results in Section 2, we establish causal identification and estimate intervention distributions.

3.1. An expanded class of causal graphs

The expanded class of graphs is illustrated in Figure 2b. As above, there are m treatments $A_{1:m}$ and an outcome Y. The goal is to estimate $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$, where $A_{\mathcal{C}} \subset \{A_1, \ldots, A_m\}$ is a subset of treatments on which we intervene. Apart from treatments and outcome, the graph has other types of variables; Figure 2a contains a glossary.

Confounders. Confounders are parents of both the treatments and the outcome; they can be unobserved. In Figure 2b, for example, U_i^{sng} and U_i^{mlt} are confounders; they have arrows into the outcome Y and at least one of the treatments A_i. We differentiate between *single-treatment* and *multi-treatment* confounders. Single-treatment confounders like U_i^{sng} affect only one treatment; multi-treatment con-

founders like U_i^{mlt} affect two or more treatments.

Covariates. There are two types of covariates—treatment covariates and outcome covariates. treatment covariates are parents of the treatments, but not the outcome; they can be unobserved. As with confounders, we differentiate between *single-treatment* covariates W_i^{sng} and *multi-treatment* covariates W_i^{mlt}. Outcome covariates like V are parents of the outcome but not the treatments. They do not affect any of the m treatments; they can be unobserved.

Selection operator. Following Bareinboim & Pearl (2012), we introduce a selection operator $S \in \{0, 1\}$ into the causal graph. The value $S = 1$ indicates an individual being selected; otherwise, $S = 0$. We only observe the outcome of those individuals with $S = 1$, but we may observe the treatments on unselected individuals. (E.g., consider a genome-wide association study where we collect an expensive-to-measure trait on a subset of the population but have genome data on a much larger set.) Note that Figure 2b allows selection to occur on the confounders.

3.2. Causal identification

We extend the results around causal identification and estimation under shared confounding (Theorems 1 and 2) to the expanded class of graphs. We first reduce the graph of Figure 2b to one close to the shared confounding case; then we handle the complications of selection bias.

Reduction to shared confounding. To reduce the graph of Figure 2b, we bundle all the unobserved multi-treatment confounders and null confounders $\{U^{\text{mlt}}, W^{\text{mlt}}\}$ into a single unobserved confounder Z. This variable Z is shared by all the treatments as in Figure 1a and renders all the treatments conditionally independent. Moreover, it is sufficient to adjust for Z and single-treatment confounders U^{sng} to estimate $P(y \mid \text{do}(a_{\mathcal{C}}))$ because $\{U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}\}$ constitute an admissible set.

We can equivalently identify the intervention distributions $P(y \mid \text{do}(a_{\mathcal{C}}))$ in the graph of Figure 2b using a reduced graph of Figure 2c; it involves only the single-treatment confounders U^{sng} and a shared confounder Z. Below we formally state the validity of the reduction.

Lemma 3. *(Validity of reduction) Assume the causal graph in Figure 2b. Adjusting for the multi-treatment confounders and null confounders on the graph of Figure 2b is equivalent to adjusting for the shared confounder in Figure 2c:*

$$P(y \mid u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \ldots, a_m, s = 1) = P(y \mid u^{\text{sng}}, z, a_1, \ldots, a_m, s = 1). \quad (9)$$

Proof sketch. The proof uses a measure-theoretic argument to characterize the information contained in the Z variable in Figure 2c. Roughly, the information in Z is same as the information of all multi-treatment confounders, all null confounders, and some independent error:

$$\sigma(z) = \sigma(u^{\text{mlt}}, w^{\text{mlt}}, \epsilon_Z), \quad (10)$$

where $\sigma(\cdot)$ denotes the σ-algebra of a random variable. The independent error ϵ_Z satisfies

$$\epsilon_Z \perp Y, S, U^{\text{sng}}, U^{\text{mlt}}, W^{\text{mlt}}, A_1, \ldots, A_m.$$

Eq. 10 implies that conditioning on Z is equivalent to conditioning on $U^{\text{mlt}}, W^{\text{mlt}}, \epsilon_Z$; it leads to Eq. 9. The full proof is in Appendix E. □

Causal identification on the reduced causal graph (Figure 2c). We reduced the expanded class of graphs (Figure 2b) to one with shared confounding (Figure 2c). This reduction allows us to establish causal identification on the expanded class. We extend Theorem 1 from Figure 1a to Figure 2c. With the reduction step (Lemma 3), it leads to causal identification.

How can we identify the intervention distributions $P(y \mid \text{do}(a_{\mathcal{C}}))$ on the reduced graph (Figure 2c)? Figure 2c has a confounder Z that is shared across all treatments. This structure is similar to the unobserved shared confounding of Figure 1a. In addition to the shared confounder Z, the reduced graph involves single-treatment confounders U^{sng} and the selection operator S. We posit two assumptions on them to enable causal identification.

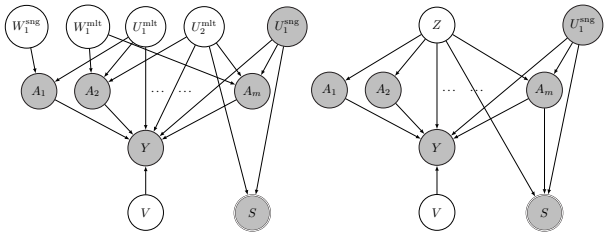
Assumption 3. *The causal graph Figure 2c satisfies the following conditions:*

1. *All single-treatment confounders U_i^{sng}'s are observed.*
2. *The selection operator S satisfies*
$$S \perp (A, Y) \mid Z, U^{\text{sng}}. \quad (11)$$
3. *We observe the non-selection-biased distribution*
$$P(a_1, \ldots, a_m, u^{\text{sng}})$$
and the selection-biased distribution
$$P(y, u^{\text{sng}}, a_1, \ldots, a_m \mid s = 1).$$

Assumption 3.1 requires that the confounders that affect the outcome and only one of the treatments must be observed. It allows us to adjust for confounding due to these single-treatment confounders. Assumption 3.2 roughly requires that selection can only occur on the confounders. Assumption 3.3 requires access to the non-selection-biased distribution of the treatments and single-treatment-confounders. It aligns with common conditions required by recovery under selection bias (e.g., Theorem 2 of Bareinboim et al. (2014)).

Name	Children	Notation
Confounder	≥ 1 treatment & outcome	$U^{\text{mlt}}, U^{\text{sng}}$
treatment covariate	≥ 1 treatment only	$W^{\text{mlt}}, W^{\text{sng}}$
Outcome covariate	outcome only	V

(a)



(b) **(c)**

Figure 2. (a) Types of nodes (b) The expanded class of causal graphs. S is the selection operator. (c) The reduced causal graph with shared confounding.

We next establish causal identification on the reduced causal graph Figure 2c. We additionally make Assumption 4; it is a variant of Assumption 1 but involves single-treatment confounders and the selection operator.

Assumption 4. *There exists some function f and a set $\emptyset \neq \mathcal{N} \subset \{1, \ldots, m\} \backslash \mathcal{C}$ such that*

1. *The outcome Y does not causally depend on $f(A_{\mathcal{N}})$:*

$$f(A_{\mathcal{N}}) \perp Y \mid Z, A_{\mathcal{C}}, A_{\mathcal{X}}, U^{\text{sng}}, S = 1 \tag{12}$$

   where $\mathcal{X} = \{1, \ldots, m\} \backslash (\mathcal{C} \cup \mathcal{N}) \neq \emptyset$.

2. *The conditional $P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1)$ is complete in $f(a_{\mathcal{N}})$ for almost all $a_{\mathcal{C}}$ and $u_{\mathcal{C}}^{\text{sng}}$, where $U_{\mathcal{C}}^{\text{sng}}$ is the single-treatment confounders affecting $A_{\mathcal{C}}$.*

3. *The conditional $P(f(a_{\mathcal{N}}) \mid a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}, s = 1)$ is complete in $a_{\mathcal{X}}$ for almost all $a_{\mathcal{C}}$ and $u_{\mathcal{C}}^{\text{sng}}$.*

Under Assumption 3 and Assumption 4, we can identify the intervention distributions $P(y \mid \text{do}(a_{\mathcal{C}}))$.

Lemma 4. *Assume the causal graph Figure 2c. Under Assumption 3 and Assumption 4, the intervention distribution of the treatments $A_{\mathcal{C}}$ is identifiable:*

$$P(y \mid \text{do}(a_{\mathcal{C}})) \tag{13}$$
$$= \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}}) P(u_{\mathcal{C}}^{\text{sng}}) \, \text{d}a_{\mathcal{X}} \, \text{d}u_{\mathcal{C}}^{\text{sng}}$$

for any solution h to the integral equation

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1)$$
$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \times P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \text{d}a_{\mathcal{X}}, \tag{14}$$

where $U_{\mathcal{C}}^{\text{sng}}$ is the single-treatment confounders affecting $A_{\mathcal{C}}$. Moreover, the solution to Eq. 14 always exists under weak regularity conditions in Appendix D.

(The proof is in Appendix F, similar to Theorem 1.)

Causal identification on the expanded class of causal graphs (Figure 2b). Based on the previous analysis on the reduced graph, we establish causal identification result on the expanded class of causal graphs.

Theorem 5. *Assume the causal graph Figure 2b. Assume a variant of Assumption 3 and Assumption 4 (detailed in Appendix G), the intervention distribution of the treatments $A_{\mathcal{C}}$ is identifiable using Eq. 13 and Eq. 14.*

(The proof is in Appendix G.)

3.3. Causal estimation with the deconfounder

We finally extend the deconfounder to the expanded class of causal graphs (Figure 2b) with selection bias and prove its correctness. We build on the identification result of Theorem 5. We then show that the deconfounder provides correct causal estimates by implicitly solving the integral equation (Eq. 14). This argument is similar to the argument of Theorem 2.

The algorithm for the expanded class of graphs with selection bias extends the version described in Section 2.2. Specifically, Assumption 2 allows the algorithm to have access to both the non-selection-biased data $P(a_1, \ldots, a_m, u^{\text{sng}})$ and the selection-biased data $P(y, u^{\text{sng}}, a_1, \ldots, a_m \mid s = 1)$. In this case, the algorithm outputs two estimates:

$$(1) \hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$$
$$= \hat{P}(\hat{z}) \hat{P}(u^{\text{sng}} \mid a_1, \ldots, a_m, \hat{z}) \prod_{i=1}^{n} \hat{P}(a_i \mid \hat{z}),$$
$$(2) \hat{P}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1).$$

We note that the former is constructed using only the treatments $A_1, \ldots, A_m$ and single-treatment confounders U^{sng}. Moreover, both estimates must be consistent with the observed data:

$$(1) \int \hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z}) \, \text{d}\hat{z} = P(a_1, \ldots, a_m, u^{\text{sng}}),$$

$$(2)\ \int \hat{P}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)\, \mathrm{d}\hat{z} = P(y, a_1, \ldots, a_m, u^{\text{sng}} \mid s = 1).$$

We note that the substitute confounder $\hat{Z}$ does not necessarily coincide with the true confounders U^{mlt} or the true null confounders W^{mlt}. Nor do $\hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$ and $\hat{P}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)$ need to be unique. We will show that any $\hat{Z}$ and $\hat{P}$ that the algorithm outputs will lead to a correct estimate of $\hat{P}(y \mid \text{do}(a_{\mathcal{C}}))$.

Finally the algorithm estimates

$$\hat{P}(y \mid \text{do}(a_{\mathcal{C}})) \triangleq \int \hat{P}(y \mid a_1, \ldots, a_m, \hat{z}, u^{\text{sng}}_{\mathcal{C}}, s = 1) \times \hat{P}(a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \hat{z}) P(u^{\text{sng}}_{\mathcal{C}})\, \mathrm{d}u^{\text{sng}}_{\mathcal{C}}\, \mathrm{d}\hat{z}\, \mathrm{d}a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \tag{15}$$

where $U^{\text{sng}}_{\mathcal{C}}$ are the single-treatment confounders that affect the treatments $A_{\mathcal{C}}$.

We now prove the correctness of the deconfounder on the expanded class of causal graphs. We make a variant of Assumption 2 and state the correctness result.

Assumption 5. *The deconfounder outputs the estimates $\hat{P}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)$ and $\hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$ that satisfy the following:*

1. *It is consistent with Assumption 3.1:*

$$\hat{P}(a_1, \ldots, a_m \mid \hat{z}, u^{\text{sng}}, s = 1) = \hat{P}(a_1, \ldots, a_m \mid \hat{z}, u^{\text{sng}}). \tag{16}$$

2. *It is consistent with Assumption 4.1:*

$$\hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}, u^{\text{sng}}, s = 1) = \hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, \hat{z}, u^{\text{sng}}, s = 1). \tag{17}$$

3. *The conditional $\hat{P}(\hat{z} \mid a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}, s = 1)$ is complete in $a_{\mathcal{X}}$ for almost all $a_{\mathcal{C}}$.*

The conditional $\hat{P}(\hat{z} \mid a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}, s = 1)$, Eq. 16, and Eq. 17 can be computed from $\hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$ and $\hat{P}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)$).

Under these assumptions, Theorem 6 establishes the correctness of the deconfounder on causal graphs under certain types of selection bias.

Theorem 6. *(Correctness of the deconfounder on the expanded class of causal graphs) Assume the causal graph Figure 2b. Assume a variant of Assumption 3 and Assumption 4 (detailed in Appendix H). Under Assumption 5 and weak regularity conditions, the deconfounder provides correct estimates of the intervention distribution:*

$$\hat{P}(y \mid \text{do}(a_{\mathcal{C}})) = P(y \mid \text{do}(a_{\mathcal{C}})). \tag{18}$$

(The proof is in Appendix H.)

4. Example: A linear causal model

We illustrate Theorems 5 and 6 in a linear causal model.

Consider the meal/body-fat example. The treatments are ten types of food $A_1, \ldots, A_{10}$; the outcome is a person's body fat Y. How does food consumption affect body fat?

In this example, the individual's lifestyle U^{mlt} is a multi-treatment confounder. Whether a person is vegan W^{mlt} is a multi-treatment null confounder. Both U^{mlt} and W^{mlt} are unobserved. Whether one has easy access to good burger shops U^{sng} is a single-treatment confounder; it affects both burger consumption A_1 and body fat percentage Y; U^{sng} is observed. Finally, the observational data comes from a survey with selection bias S; people with healthy lifestyle are more likely to complete the survey.

Every variable is associated with a disturbance term ϵ, which comes from a standard normal. Given these variables, suppose the real world is linear,

$$\begin{aligned} U^{\text{mlt}} &= \epsilon_{U^{\text{mlt}}}, U^{\text{sng}} = \epsilon_{U^{\text{sng}}}, W^{\text{mlt}} = \epsilon_{W^{\text{mlt}}}, \\ A_1 &= \alpha_{A_1U} U^{\text{mlt}} + \alpha_{A_1W} W^{\text{mlt}} + \alpha_{A_1U'} U^{\text{sng}} + \epsilon_{A_1}, \\ A_i &= \alpha_{A_iU} U^{\text{mlt}} + \alpha_{A_iW} W^{\text{mlt}} + \epsilon_{A_i}, i = 2, \ldots, 10, \\ Y &= \sum_{i=1}^{10} \alpha_{YA_i} A_i + \alpha_{YU} U^{\text{mlt}} + \alpha_{YU'} U^{\text{sng}} + \epsilon_Y. \end{aligned}$$

These equations describe the true causal model of the world. The confounders and null confounders $\{U^{\text{mlt}}, W^{\text{mlt}}\}$ are unobserved.

We are interested in the intervention distribution of the first two food categories, burger (A_1) and broccoli (A_2): $P(y \mid \text{do}(a_1, a_2))$. (We emphasize that we might be interested in any subsets of the treatments.) This world satisfies the assumptions of Theorem 5. Even though the confounders U^{mlt} are unobserved, the intervention distribution $P(y \mid \text{do}(a_1, a_2))$ is identifiable.

Now consider a simple deconfounder. Fit a 2-D probabilistic principal component analysis (PPCA) to the data about food consumption $\{A_1, \ldots, A_{10}\}$; we do not model the outcome Y. Wang & Blei (2019a) also checks the model to ensure it fits the distribution of the assigned treatments. (Let's assume that 2-D PPCA passes this check.)

PPCA leads to a linear estimate of the substitute confounder,

$$\hat{Z} = \left(\sum_{i=1}^{10} \gamma_{1i} A_i + \epsilon_{1\hat{Z}}, \sum_{i=1}^{10} \gamma_{2i} A_i + \epsilon_{2\hat{Z}} \right), \tag{19}$$

for parameters γ_{1i} and γ_{2i}, and Gaussian noise $\epsilon_{i,\hat{Z}}$.

This substitute confounder $\hat{Z}$ satisfies Assumption 5. Plausi-

bly, the real world satisfies the variant of Assumption 3 and Assumption 4. These assumptions greenlight us to calculate the intervention distribution. We fit an outcome model using the substitute confounder $\hat{Z}$ and calculate the intervention distribution using Eq. 15. Theorem 6 guarantees that this estimate is correct.

5. A Simulation Study

In this section, we see the identification results in action. We find that the identification conditions discussed in Sections 2 and 3 are crucial for producing correct causal estimates. The theoretical results and the conditions required by Theorems 1, 2, 5 and 6 are practically important.

Specifically, we consider a linear data generating process in Section 4 with a one-dimensional U and three treatments A_1, A_2, A_3. We explore two configurations of the unobserved confounder U.

In one configuration, U is normally distributed, and the resulting observational data satisfies the completeness condition in Assumption 1.2. Figure 3a shows the mean squared error (RMSE) of the deconfounder average treatment effect (ATE) estimate stays low even if the confounding strength is high while the RMSE of naive regression quickly blows up.

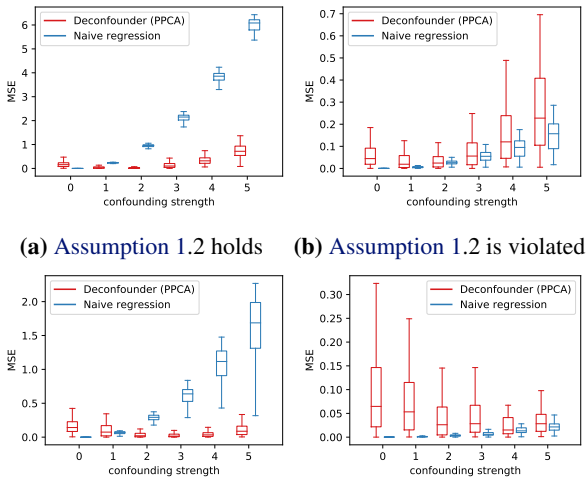
In a second configuration, U is uniformly distributed; it results in an observational data distribution that violates the completeness condition (Assumption 1.2). Figure 3b shows that the deconfounder can no longer control for confounding in this setting. It produces causal estimates that have consistently lower quality than naive regression.

Finally we extend the simulation to study selection bias. Given a normally or uniformly distributed U, we generate observational data from the same linear model. We then introduce selection bias by selecting samples with probability $\propto \mathcal{N}(U\,;\,0, 0.5^2)$ and $\propto \mathrm{Unif}(U\,;\,0, 0.5)$. We apply the deconfounder estimation algorithm.

Figures 3c and 3d exhibit the similar phenomenon as above. When the identification conditions hold, the deconfounder produces significant improvement in ATE estimation. When these conditions are violated, the deconfounder produces low quality ATE estimates. Notice that under selection bias, the variance of the estimate tends to go down as confounding strength goes up. We observe this phenomenon because stronger confounding strength makes it easier to infer the latent confounder U, reducing the variance of the estimate.

6. Discussion

We study causal identification and estimation when multiple treatments share the same unobserved confounder. By treating some treatments as proxies of the shared confounder, we can identify the intervention distributions of the other treatments. For an expanded class of causal graphs, we prove that the intervention distribution of subsets of treatments is identifiable. We further show that the deconfounder algorithm of Wang & Blei (2019a) makes valid inferences of these intervention distributions when causal identification holds. We demonstrate the practical relevance of these theoretical results in a simulation study, showing how violating the identification conditions can fail the deconfounder in practice.



(a) Assumption 1.2 holds **(b)** Assumption 1.2 is violated

(c) Assumption 4.2 holds **(d)** Assumption 4.2 is violated

Figure 3. The deconfounder outperforms naive regression when the identification conditions are satisfied, but fails to otherwise.

Acknowledgements

We thank Elias Bareinboim and Victor Veitch for their insightful comments on the manuscript. This work is supported by ONR N00014-17-1-2131, ONR N00014-15-1-2209, NIH 1U01MH115727-01, NSF CCF-1740833, DARPA SD2 FA8750-18-C-0130, Amazon, and Simons Foundation.

References

Bareinboim, E. & Pearl, J. (2012). Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics* (pp. 100–108).

Bareinboim, E., Tian, J., & Pearl, J. (2014). Recovering from selection bias in causal and statistical inference.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Carrasco, M., Florens, J.-P., & Renault, E. (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6, 5633–5751.

Chen, X., Chernozhukov, V., Lee, S., & Newey, W. K. (2014). Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2), 785–809.

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

D'Amour, A. (2019b). 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* (pp. 3478–3486).

Darolles, S., Fan, Y., Florens, J.-P., & Renault, E. (2011). Nonparametric instrumental regression. *Econometrica*, 79(5), 1541–1565.

D'Haultfoeuille, X. (2011). On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3), 460–471.

Frot, B., Nandy, P., & Maathuis, M. H. (2017). Learning directed acyclic graphs with hidden variables via latent gaussian graphical model selection. *arXiv preprint arXiv:1708.01151*.

Grimmer, J., Knox, D., & Stewart, B. M. (2020). Na\" ive regression requires weaker assumptions than factor models to adjust for multiple cause confounding. *arXiv preprint arXiv:2007.12702*.

Heckerman, D. (2018). Accounting for hidden common causes when infering cause and effect from observational data. *arXiv preprint arXiv:1801.00727*.

Hu, Y. & Shiu, J.-L. (2018). Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3), 659–693.

Imai, K. & Jiang, Z. (2019). Comment: The challenges of multiple causes. *Journal of the American Statistical Association*, 114(528), 1605–1610.

Janzing, D. & Schölkopf, B. (2018). Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1).

Kuroki, M. & Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2), 423–437.

Miao, W., Geng, Z., & Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4), 987–993.

Miao, W., Hu, W., Ogburn, E. L., & Zhou, X. (2020). Identifying effects of multiple treatments in the presence of unmeasured confounding. *arXiv preprint arXiv:2011.04504*.

Millán, J., Pintó, X., et al. (2009). Lipoprotein ratios: physiological significance and clinical usefulness in cardiovascular prevention. *Vascular health and risk management*, 5, 757.

Newey, W. K. & Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5), 1565–1578.

Ogburn, E. L., Shpitser, I., & Tchetgen, E. J. T. (2019). Comment on "blessings of multiple causes". *Journal of the American Statistical Association*, 114(528), 1611–1615.

Ogburn, E. L., Shpitser, I., & Tchetgen, E. J. T. (2020). Counterexamples to "the blessings of multiple causes" by wang and blei.

Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition.

Puli, A., Perotte, A., & Ranganath, R. (2020). Causal estimation with functional confounders. *Advances in Neural Information Processing Systems*, 33.

Ranganath, R. & Perotte, A. (2018). Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*.

Shi, X., Miao, W., Nelson, J. C., & Tchetgen Tchetgen, E. J. (2020). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2), 521–540.

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

Tran, D. & Blei, D. M. (2017). Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*.

Wang, J., Zhao, Q., Hastie, T., & Owen, A. B. (2017). Confounder adjustment in multiple hypothesis testing. *Annals of statistics*, 45(5), 1863.

Wang, Y. & Blei, D. M. (2019a). The blessings of multiple causes. *Journal of the American Statistical Association*, 114(528), 1574–1596.

Wang, Y. & Blei, D. M. (2019b). The blessings of multiple causes: Rejoinder. *Journal of the American Statistical Association*, 114(528), 1616–1619.

Wang, Y. & Blei, D. M. (2020). Towards clarifying the theory of the deconfounder. *arXiv preprint arXiv:2003.04948*.

Yang, S., Wang, L., & Ding, P. (2017). Identification and estimation of causal effects with confounders subject to instrumental missingness. *arXiv preprint arXiv:1702.03951*.

Ćevid, D., Bühlmann, P., & Meinshausen, N. (2018). Spectral deconfounding via perturbed sparse linear models.

Supplementary Material: A Proxy Variable View of Shared Confounding

A. Proof of Theorem 1

Proof. The proof of Theorem 1 relies on two observations. The first observation starts with the integral equation we solve:

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}} \tag{20}$$

$$= \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u) P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}} \, \mathrm{d}u. \tag{21}$$

The first equality is due to Eq. 3. The second equality is due to the conditional independence implied by Figure 1a: $A_{\mathcal{X}} \perp A_{\mathcal{C}}, f(a_{\mathcal{N}}) \mid U$.

The second observation relies on the null proxy:

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int P(y \mid u, a_{\mathcal{C}}, f(a_{\mathcal{N}})) P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}u \tag{22}$$

$$= \int P(y \mid u, a_{\mathcal{C}}) P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}u. \tag{23}$$

The first equality is due to the definition of conditional probability. The second equality is due to the second part of Assumption 1, which implies $Y \perp f(a_{\mathcal{N}}) \mid U, A_{\mathcal{C}}$. The reason is that

$$P(y \mid u, a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int P(y \mid u, a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}})) P(a_{\mathcal{X}} \mid u, a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}} \tag{24}$$

$$= \int P(y \mid u, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u, a_{\mathcal{C}}) \, \mathrm{d}a_{\mathcal{X}} \tag{25}$$

$$= P(y \mid u, a_{\mathcal{C}}). \tag{26}$$

In fact, it is sufficient to assume $Y \perp f(a_{\mathcal{N}}) \mid U, A_{\mathcal{C}}$ instead of $Y \perp f(a_{\mathcal{N}}) \mid U, A_{\mathcal{C}}, A_{\mathcal{X}}$ in Theorem 1. However, the latter is easier to check and interpret.

Comparing Eq. 21 and Eq. 23 gives

$$\int \left[P(y \mid u, a_{\mathcal{C}}) - \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u) \, \mathrm{d}a_{\mathcal{X}} \right] \times P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}u = 0, \tag{27}$$

which, by the completeness condition in Assumption 1.2, implies

$$P(y \mid u, a_{\mathcal{C}}) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u) \, \mathrm{d}a_{\mathcal{X}}. \tag{28}$$

Eq. 28 leads to identification:

$$P(y \mid \mathrm{do}(a_{\mathcal{C}})) = \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u) \, \mathrm{d}a_{\mathcal{X}} P(u) \, \mathrm{d}u \tag{29}$$

$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}}) \, \mathrm{d}a_{\mathcal{X}}. \tag{30}$$

Consider the special case of a single treatment as in Figure 1b. Let $a_{\mathcal{C}} = \{A_1\}$, $a_{\mathcal{X}} = \{X\}$, $a_{\mathcal{N}} = N$, and $f(a_{\mathcal{N}}) = N$. The above proof reduces to the identification proof for proxy variables (Theorem 1 of Miao et al. (2018)). □

B. Examples of Assumption 1

As an example, if the structural equation writes

$$Y = g(A_1 + A_2, A_3, \ldots, A_m, U, \epsilon),$$

where $\epsilon \perp U, A_1, \ldots, A_m$, then Assumption 1.1 is satisfied if A_1 and A_2 are identically Gaussian: $A_{\mathcal{N}} = (A_1, A_2)$ and $f(A_{\mathcal{N}}) = A_1 - A_2$ satisfies

$$A_1 - A_2 \perp Y \,|\, U, A_3, \ldots, A_m.$$

If A_1 and A_2 are both Gaussian but not identically distributed, then $f(A_{\mathcal{N}}) = \alpha_1 A_1 - \alpha_2 A_2$ would satisfy

$$\alpha_1 A_1 - \alpha_2 A_2 \perp Y \,|\, U, A_3, \ldots, A_m,$$

for some constant α_1 and α_2.

Similarly, if the structural equation writes

$$Y = g(A_1 \times A_2, A_3, \ldots, A_m, U, \epsilon),$$

where $\epsilon \perp U, A_1, \ldots, A_m$, then Assumption 1.1 is satisfied if A_1 and A_2 are identically log-normal: $A_{\mathcal{N}} = (A_1, A_2)$ and $f(A_{\mathcal{N}}) = A_1/A_2$ satisfies

$$A_1/A_2 \perp Y \,|\, U, A_3, \ldots, A_m.$$

As a final example, if the structural equation writes

$$Y = g(A_1 \&\& A_2, A_3, \ldots, A_m, U, \epsilon),$$

where $\epsilon \perp U, A_1, \ldots, A_m$ and A_1, A_2 are both binary, then Assumption 1.1 is satisfied: $A_{\mathcal{N}} = (A_1, A_2)$ and $f(A_{\mathcal{N}}) = A_1$ XOR A_2 satisfies

$$A_1 \text{ XOR } A_2 \perp Y \,|\, U, A_3, \ldots, A_m.$$

C. Proof of Theorem 2

Proof. Assumption 2.2 guarantees the existence of some function $\hat{h}$ such that

$$\hat{P}(y \,|\, a_{\mathcal{C}}, \hat{z}) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \,|\, \hat{z}) \, \mathrm{d}a_{\mathcal{X}} \tag{31}$$

under weak regularity conditions. (We will discuss the reason in Appendix D.)

We first claim that $\hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}})$ solves

$$P(y \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}}. \tag{32}$$

Given this claim (Eq. 77), we have

$$\begin{aligned}
&\hat{P}(y \,|\, \mathrm{do}(a_{\mathcal{C}})) \\
=&\int \hat{P}(y \,|\, \hat{z}, a_{\mathcal{C}}) \hat{P}(\hat{z}) \, \mathrm{d}\hat{z} \\
=&\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \,|\, \hat{z}) \, \mathrm{d}a_{\mathcal{X}} \hat{P}(\hat{z}) \, \mathrm{d}\hat{z} \\
=&\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}}) \, \mathrm{d}a_{\mathcal{X}} \\
=&P(y \,|\, \mathrm{do}(a_{\mathcal{C}})),
\end{aligned}$$

which proves the theorem. The first equality is due to Eq. 6; the second is due to Eq. 77; the third is due to the deconfounder estimate being consistent with the observed data distribution by construction; the fourth is due to the above claim (Eq. 77) and Theorem 1.

We next prove the claim (Eq. 77). Start with the right side of the equality.

$$\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}}$$

$$
\begin{aligned}
&= \int \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}) \hat{P}(\hat{z} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}} \, \mathrm{d}\hat{z} \\
&= \int \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}) \hat{P}(\hat{z} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}\hat{z} \\
&= P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})),
\end{aligned}
$$

which establishes the claim. The first equality is due to Eq. 4 and the deconfounder estimate being consistent with the observed data; the second is due to Eq. 31; the third is due to Assumption 2.1, which implies

$$
\hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}) = \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}). \tag{33}
$$

Similar to Assumption 1.1, it is sufficient to assume Eq. 33 directly. However, Assumption 2.1 is easier to check and more interpretable; it directly relates to the deconfounder outcome model.

□

D. Existence of solutions to the integral equations

Theorem 1 involves solving the integral equation

$$
P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}}. \tag{34}
$$

When does a solution exist for Eq. 34? We appeal to Proposition 1 of Miao et al. (2018).

Proposition 7. *(Proposition 1 of Miao et al. (2018)) Denote $L^2\{F(t)\}$ as the space of all square-integrable function of t with respect to a c.d.f. $F(t)$. A solution to integral equation*

$$
P(y \mid z, x) = \int h(w, x, y) P(w \mid z, x) \, \mathrm{d}w \tag{35}
$$

exists if

1. *the conditional distribution $P(z \mid w, x)$ is complete in w for all x,*
2. $\int \int P(w \mid z, x) P(z \mid w, x) \, \mathrm{d}w \, \mathrm{d}z < +\infty$,
3. $\int [P(y \mid z, x)]^2 P(z \mid x) \, \mathrm{d}z < +\infty$,
4. $\sum_{n=1}^{+\infty} | < P(y \mid z, x), \psi_{x,n} >|^2 < +\infty$,

where the inner product is $< g, h >= \int g(t) h(t) \, \mathrm{d}F(t)$, and $(\lambda_{x,n}, \phi_{x,n}, \psi_{x,n})_{n=1}^{\infty}$ is a singular value decomposition of the conditional expectation operator $K_x : L^2\{F(w \mid x)\} \to L^2\{F(z \mid x)\}, K_x(h) = \mathbb{E}[h(w) \mid z, x]$ for $h \in L^2\{F(w \mid x)\}$.

Leveraging Proposition 7, we can establish sufficient conditions for existence of a solution to Eq. 34.

Corollary 8. *A solution exist for the integral equation Eq. 34 if*

1. *the conditional distribution $P(f(a_{\mathcal{N}}) \mid a_{\mathcal{X}}, a_{\mathcal{C}})$ is complete in $a_{\mathcal{X}}$ for all $a_{\mathcal{C}}$,*
2. $\int \int P(a_{\mathcal{X}} \mid f(a_{\mathcal{N}}), a_{\mathcal{C}}) P(f(a_{\mathcal{N}}) \mid a_{\mathcal{X}}, a_{\mathcal{C}}) \, \mathrm{d}a_{\mathcal{X}} \, \mathrm{d}f(a_{\mathcal{N}}) < +\infty$,
3. $\int [P(y \mid f(a_{\mathcal{N}}), a_{\mathcal{C}})]^2 P(f(a_{\mathcal{N}}) \mid a_{\mathcal{C}}) \, \mathrm{d}f(a_{\mathcal{N}}) < +\infty$,
4. $\sum_{n=1}^{+\infty} | < P(y \mid f(a_{\mathcal{N}}), a_{\mathcal{C}}), \psi_{a_{\mathcal{C}},n} >|^2 < +\infty$,

where $\psi_{a_{\mathcal{C}},n}$ is similarly defined as a component of the singular value decomposition.

We remark that the first condition is precisely Theorem 1.3; others are weak regularity conditions.

By the same token, we can establish sufficient conditions for solution existence of Eq. 8, Eq. 14. The same argument also applies to the integral equation involved in Theorem 6:

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}}. \tag{36}$$

It is easy to show that the conditions described in the main text are sufficient to guarantee the existence of solutions under weak regularity conditions. We omit the details here.

E. Proof of Lemma 3

The idea of the proof is to start with the structural equations of the expanded class of causal graphs Figure 2b. Then posit the existence of a latent variable Z that renders all the treatments conditionally independent; Figure 2c features this conditional independence structure. We will quantify the information (i.e. the σ-algebra) of this latent variable Z; Z contains the information of the union of multi-treatment confounders U^{mlt}, multi-treatment null confounders W^{mlt}, and some independent error. This result lets us establish

$$P(y \mid u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \ldots, a_m, s=1) = P(y \mid u^{\text{sng}}, z, a_1, \ldots, a_m, s=1). \tag{37}$$

We start with a generic structural equation model for multiple treatments.

$$W_k = f_{W_k}(\epsilon_{W_k}), \qquad k = 1, \ldots, K, K \geq 0, \tag{38}$$
$$U_j = f_{U_j}(\epsilon_{U_j}), \qquad j = 1, \ldots, J, J \geq 0, \tag{39}$$
$$V_l = f_{V_l}(\epsilon_{V_l}), \qquad l = 1, \ldots, L, L \geq 0, \tag{40}$$
$$A_i = f_{A_i}(W_{S_{A_i}^W}, U_{S_{A_i}^U}, \epsilon_{A_i}), \qquad i = 1, \ldots, m, m \geq 2, \tag{41}$$
$$Y = f_y(A_1, \ldots, A_m, U_1, \ldots, U_K, V_1, \ldots V_L, \epsilon_Y), \tag{42}$$

where all the errors $\epsilon_{W_k}, \epsilon_{U_j}, \epsilon_{V_l}, \epsilon_{A_i}, \epsilon_Y$ are independent. Notation wise, we note that $S_{A_i}^W \subset \{1, \ldots, K\}$ is an index set; if $S_{A_1}^W = \{1, 3, 4\}$, then $W_{S_{A_i}^W} = (W_1, W_3, W_4)$. The same notion applies to $S_{A_i}^U \subset \{1, \ldots, J\}$.

The notation in this structural equation model is consistent with the set up in Figure 2b. W_k's are null confounders; U_j's are confounders; V_l's are covariates. Moreover, $U_{S_{A_i}^U}$ indicates the set of confounders that have an arrow to both A_i and Y. $W_{S_{A_i}^W}$ indicates the set of null confounders that have an arrow to A_i; they do not have arrows to Y.

Relating to the single-treatment and multi-treatment notion, we have single-treatment null confounders as

$$W^{\text{sng}} \triangleq \{W_1, \ldots, W_K\} / \bigcup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S_{A_i}^W} \cap W_{S_{A_j}^W}). \tag{43}$$

To parse the notation above, recall that $W_{S_{A_i}^W}$ is the set of null confounders that affects A_i. $\bigcup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S_{A_i}^W} \cap W_{S_{A_j}^W})$ describes the set of null confounders that affect at least two of the A_i's. Hence, W^{sng} denotes the set of null confounders that affect only one of the A_i's, a.k.a. single-treatment null confounders.

Before proving Lemma 3, we first prove the following lemma that quantifies the information in Z (in Figure 2c).

Lemma 9. *The random variable Z in Figure 2c "captures" all multi-treatment confounders, all multi-treatment null confounders and some independent error:*

$$\sigma(Z) = \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S_{A_i}^W} \cap W_{S_{A_j}^W}) \cup (U_{S_{A_i}^U} \cap U_{S_{A_j}^U}))\right), \tag{44}$$
$$= \sigma\left(\{\epsilon_Z\} \bigcup W^{\text{mlt}} \bigcup U^{\text{mlt}}\right). \tag{45}$$

where $\epsilon_Z \perp (\epsilon_Y, V_1, \ldots, V_L, \cup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S_{A_i}^W} \cap W_{S_{A_j}^W}) \cup (U_{S_{A_i}^U} \cap U_{S_{A_j}^U}), S)$.

We can parse the notation in Lemma 9 in the same way as in Eq. 43: $\cup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}})$ denotes the set of all multi-treatment confounders; $\cup_{i,j\in\{1,\dots,m\}:i\neq j}(U_{S^U_{A_i}} \cap U_{S^U_{A_j}})$ denotes the set of all multi-treatment null confounders.

Proof. Without the loss of generality, we assume the compactness of representation in Eqs. 41 and 42. For any subset $\mathcal{S}$ of the random variables $\mathcal{S} \subset \{A_1, \dots, A_m, Y\}$, we assume the σ-algebra $\sigma(\bigcap_\tau (S^W_{S_\tau}, S^U_{S_\tau}, S^V_{S_\tau}))$ is the *smallest* σ-algebra that makes all the random variables in $\mathcal{S}$ jointly independent. The assumption is made for technical convenience. We simply ensure the arrows from the W, U, V's to the A_i's do exist. In other words, all the W, U, V's "whole-heartedly" contribute to the A_i's when they appear in Eq. 41. This assumption does not limit the class of causal graphs we study.

First we show that all multi-treatment confounders and all multi-treatment null confounders are measurable with respect to the substitute confounder Z:

$$\sigma\left(\bigcup_{i,j\in\{1,\dots,m\}:i\neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{46}$$

Consider any pair of A_i and A_j. Figure 2c implies that

$$A_i \perp A_j \,|\, Z, \tag{47}$$

for $i \neq j$ and $i, j \in \{1, \dots, M\}$. On the other hand, we have

$$A_i \perp A_j \,|\, \sigma\left((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right), \tag{48}$$

by the independence of errors assumption. Therefore, by the compactness of representation assumption, $\sigma((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))$ is the smallest σ-algebra that renders A_i independent of A_j. This implies

$$\sigma\left((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{49}$$

The argument can be applied to any pair of $i \neq j, i, j \in \{1, \dots, M\}$, so we have

$$\sigma\left(\bigcup_{i,j\in\{1,\dots,m\}:i\neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{50}$$

Next Figure 2c implies

$$\sigma(A_1, \dots, A_M) \not\subset \sigma(Z), \tag{51}$$

and

$$\sigma(Y) \not\subset \sigma(Z). \tag{52}$$

Therefore, we have

$$\sigma(Z) \subset \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j\in\{1,\dots,m\}:i\neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right), \tag{53}$$

where ϵ_Z is independent of all the other errors in the structural model, including those of A and Y.

The error ϵ_Z can have an empty σ-algebra: for example, ϵ_Z is a constant. Therefore, the left side of Eq. 50 can be made equal to the right side of Eq. 53. We have

$$\sigma(Z) = \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j\in\{1,\dots,m\}:i\neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right) \tag{54}$$

$$= \sigma\left(\{\epsilon_Z\} \bigcup W^{\text{mlt}} \bigcup U^{\text{mlt}}\right). \tag{55}$$

for some random variable ϵ_Z that is independent of all other random errors ϵ's. □

As a direct consequence of Lemma 9, we have

$$P(y \mid u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \ldots, a_m, s=1) = P(y \mid u^{\text{sng}}, z, a_1, \ldots, a_m, s=1), \tag{56}$$

due to the definition of conditional probabilities and $\epsilon_Z \perp Y \mid S, U^{\text{sng}}, U^{\text{mlt}}, W^{\text{mlt}}, A_1, \ldots, A_m$. The latter is because ϵ_Z is independent of all other errors.

F. Proof of Lemma 4

Proof. Denote $U_{\mathcal{C}}^{\text{sng}}$ as the set of single-treatment confounders that affects $A_{\mathcal{C}}$.

The proof of Lemma 4 relies on two observations.

The first observation starts with the integral equation we solve:

$$\begin{aligned}
&P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1) && (57)\\
=&\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}} && (58)\\
=&\int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}}\, \mathrm{d}z && (59)
\end{aligned}$$

The first equality is due to Eq. 14. The second equality is due to Assumption 3.2.

The second observation relies on the null proxy:

$$\begin{aligned}
&P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1) && (60)\\
=&\int P(y \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}z && (61)\\
=&\int P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s=1) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}z && (62)
\end{aligned}$$

The first equality is due to the definition of conditional probability. The second equality is due to the second part of Assumption 4; it implies $Y \perp f(a_{\mathcal{N}}) \mid Z, U_{\mathcal{C}}^{\text{sng}}, A_{\mathcal{C}}, S=1$. The reason is that

$$\begin{aligned}
&P(y \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1) && (63)\\
=&\int P(y \mid z, a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1) P(a_{\mathcal{X}} \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}} && (64)\\
=&\int P(y \mid z, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}, s=1) P(a_{\mathcal{X}} \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}} && (65)\\
=&P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s=1). && (66)
\end{aligned}$$

The second equality is again due to Assumption 3.2.

Comparing Eq. 59 and Eq. 62 gives

$$\int \left[P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s=1) - \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z)\, \mathrm{d}a_{\mathcal{X}} \right] \times P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}z = 0, \tag{67}$$

which implies

$$P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s=1) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z)\, \mathrm{d}a_{\mathcal{X}}. \tag{68}$$

This step is due to the completeness condition in Assumption 4.2.

Eq. 68 leads to identification:

$$P(y \mid \mathrm{do}(a_{\mathcal{C}})) \tag{69}$$

$$= P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}) P(z) P(u_{\mathcal{C}}^{\text{sng}}) \,\mathrm{d}z \,\mathrm{d}u_{\mathcal{C}}^{\text{sng}} \tag{70}$$

$$= P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1) P(z) P(u_{\mathcal{C}}^{\text{sng}}) \,\mathrm{d}z \,\mathrm{d}u_{\mathcal{C}}^{\text{sng}} \tag{71}$$

$$= \int\int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z) \,\mathrm{d}a_{\mathcal{X}} P(z) P(u_{\mathcal{C}}^{\text{sng}}) \,\mathrm{d}z \,\mathrm{d}u_{\mathcal{C}}^{\text{sng}} \tag{72}$$

$$= \int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}}) P(u_{\mathcal{C}}^{\text{sng}}) \,\mathrm{d}a_{\mathcal{X}} \,\mathrm{d}u_{\mathcal{C}}^{\text{sng}}. \tag{73}$$

In particular, the second equality is due to Assumption 3.2.

□

G. Proof of Theorem 5

We first state the variant of Assumption 3 and Assumption 4 required by Theorem 5. We essentially replace Z with $(U^{\text{mlt}}, W^{\text{mlt}})$ in these assumptions.

Assumption 6. *(Assumption 3') The causal graph Figure 2b satisfies the following conditions:*

1. *All single-treatment confounders U_i^{sng}'s are observed.*
2. *The selection operator S satisfies*

$$S \perp (A, Y) \mid U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}. \tag{74}$$

3. *We observe the non-selection-biased distribution*

$$P(a_1, \ldots, a_m, u^{\text{sng}})$$

and the selection-biased distribution

$$P(y, u^{\text{sng}}, a_1, \ldots, a_m \mid s = 1).$$

Assumption 7. *(Assumption 4') There exists some function f and a set $\emptyset \neq \mathcal{N} \subset \{1, \ldots, m\} \backslash \mathcal{C}$ such that*

1. *The outcome Y does not causally depend on $f(a_{\mathcal{N}})$:*

$$f(a_{\mathcal{N}}) \perp Y \mid A_{\mathcal{C}}, A_{\mathcal{X}}, U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}, S = 1 \tag{75}$$

   where $\mathcal{X} = \{1, \ldots, m\} \backslash (\mathcal{C} \cup \mathcal{N}) \neq \emptyset$.

2. *The conditional $P(u^{\text{mlt}}, w^{\text{mlt}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1)$ is complete in $f(a_{\mathcal{N}})$ for almost all $a_{\mathcal{C}}$ and $u_{\mathcal{C}}^{\text{sng}}$, where $U_{\mathcal{C}}^{\text{sng}}$ is the single-treatment confounders affecting $A_{\mathcal{C}}$.*
3. *The conditional $P(f(a_{\mathcal{N}}) \mid a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}, s = 1)$ is complete in $a_{\mathcal{X}}$ for almost all $a_{\mathcal{C}}$ and $u_{\mathcal{C}}^{\text{sng}}$.*

Under these assumptions, Theorem 5 is a direct consequence of Lemma 3 and Lemma 4. The reason is that $U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}$ constitutes an admissible set to identify the intervention distributions $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

H. Proof of Theorem 6

We assume Assumption 6 and Assumption 7 as described in Appendix G.

Proof. Assumption 5.2 guarantees the existence of some function $\hat{h}$ such that

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s = 1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s = 1) \,\mathrm{d}a_{\mathcal{X}} \tag{76}$$

under weak regularity conditions. (We discuss the reason in Appendix D.)

We first claim that $\hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}})$ solves

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}}. \tag{77}$$

Given this claim (Eq. 77), we have

$$\begin{aligned}
&\hat{P}(y \mid \mathrm{do}(a_{\mathcal{C}})) \\
=&\int\int \hat{P}(y \mid \hat{z}, u_{\mathcal{C}}^{\text{sng}}, a_{\mathcal{C}}, s=1) \hat{P}(\hat{z}) P(u_{\mathcal{C}}^{\text{sng}})\, \mathrm{d}\hat{z}\, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} \\
=&\int\int\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}} \hat{P}(\hat{z}) P(u_{\mathcal{C}}^{\text{sng}})\, \mathrm{d}\hat{z}\, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} \\
=&\int\int\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z})\, \mathrm{d}a_{\mathcal{X}} \hat{P}(\hat{z}) P(u_{\mathcal{C}}^{\text{sng}})\, \mathrm{d}\hat{z}\, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} \\
=&\int\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}})\, \mathrm{d}a_{\mathcal{X}} P(u_{\mathcal{C}}^{\text{sng}})\, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} \\
=&P(y \mid \mathrm{do}(a_{\mathcal{C}})),
\end{aligned}$$

which proves the theorem. The first equality is due to Eq. 15; the second is due to Eq. 76; the third is due to Assumption 5 and $U_{\mathcal{C}}^{\text{sng}}$ being the single-treatment confounders for $A_{\mathcal{C}}$; the fourth is due to marginalizing out $\hat{Z}$; the fifth is due to the above claim (Eq. 77) and Theorem 5.

We next prove the claim (Eq. 77). Start with the right side of the equality.

$$\begin{aligned}
&\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}} \\
=&\int\int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\text{sng}}, a_{\mathcal{C}}, s=1) \hat{P}(\hat{z} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}a_{\mathcal{X}}\, \mathrm{d}\hat{z} \\
=&\int \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1) \hat{P}(\hat{z} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}\hat{z} \\
=&\int \hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1) \hat{P}(\hat{z} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1)\, \mathrm{d}\hat{z} \\
=&P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s=1),
\end{aligned}$$

which establishes the claim. The first equality is due to Eq. 15; the second is due to Eq. 76; the third equality is due to Assumption 5.2, which implies

$$\hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1) = \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s=1). \tag{78}$$

The fourth equality is due to marginalizing out $\hat{z}$. □

I. Constructing candidate $f(a_{\mathcal{N}})$'s from the deconfounder outcome model

We illustrate how to construct candidate $f(a_{\mathcal{N}})$'s in the deconfounder outcome model.

Consider a fitted linear outcome model

$$Y = \sum_{i=1}^{10} \alpha_{YA_i} A_i + \alpha_{YZ} \hat{Z} + \alpha_{YU'} U^{\text{sng}} + \epsilon_Y. \tag{79}$$

where all the random variables are Gaussian.

It implies that there exists $f_1(A_9, A_{10}) = A_9 + \alpha_{9,10} A_{10}$ that satisfies

$$f_1(A_9, A_{10}) \perp Y \mid \hat{Z}, U^{\text{sng}}, A_1, \ldots, A_8,$$

where

$$\alpha_{9,10} = -\frac{\alpha_9 \mathrm{Var}(A_9) + \alpha_{10}\mathrm{Cov}(A_9, A_{10})}{\alpha_9 \mathrm{Cov}(A_9, A_{10}) + \alpha_{10}\mathrm{Var}(A_{10})}.$$

The reason is that $f(A_9, A_{10}) \perp (\alpha_9 A_9 + \alpha_{10} A_{10})$. Hence $f(a_{\mathcal{N}}) = A_9 + \alpha_{9,10} A_{10}$ satisfies Assumption 5.2.

J. Details of the simulation study

Figure 3a. We simulate $n = 10,000$ data points from a linear Gaussian model and apply the deconfounder. For $\gamma_U = 0, 1, 2, 3, 4, 5$,

$$U_{n\times 1} \sim \mathcal{N}(0, I), \tag{80}$$
$$\theta_{1\times 3} \sim \mathrm{Unif}(0, I), \tag{81}$$
$$A_{n\times 3} \sim \mathcal{N}(U\theta, I), \tag{82}$$
$$\beta_{1\times 3} \sim \mathrm{Unif}(0, I), \tag{83}$$
$$\beta_0 \sim \mathrm{Unif}(0, 1), \tag{84}$$
$$Y \sim \mathcal{N}(\beta_0 + A\beta^\top + \gamma_U \cdot U, I). \tag{85}$$

To apply the deconfounder, we perform maximum likelihood estimation of PPCA on A and then fit a linear model of Y against both A and the PPCA factor.

As (1) the distributions of U, A, Y are all Gaussian, and (2) the Gaussianity of A leads to the existence of null proxy (as is discussed in Appendix I, the completeness conditions in Assumption 1 are satisfied.

Figure 3b. We perform the same simulation as above except that $U_{n\times 1} \sim \mathrm{Unif}(0, I)$. In this case, the distributions of A and Y no longer belong to the exponential family and violate the completeness conditions in Assumption 1.

Figures 3c and 3d. We perform the same pair of simulation as above except that we add an additional selection step to U. After generating U from $U_{n\times 1} \sim \mathcal{N}(0, I)$, we select U w.p. proportional to $\mathcal{N}(U\,;\, 0, 0.5^2)/\mathcal{N}(U\,;\, 0, I)$ and $\mathrm{Unif}(U\,;\, 0, 0.5)/\mathrm{Unif}(U\,;\, 0, I)$ respectively. The resulting U distribution is $\mathcal{N}(U\,;\, 0, 0.5^2)$ and $\mathrm{Unif}(U\,;\, 0, 0.5)$ respectively.

References

Miao, W., Geng, Z., & Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4), 987–993.