
A Proxy Variable View of Shared Confounding

Yixin Wang¹ David M. Blei²

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 $\{A_1, \dots, A_m\}$ (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 $p(y \mid \text{do}(a_{\mathcal{C}}))$ 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

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

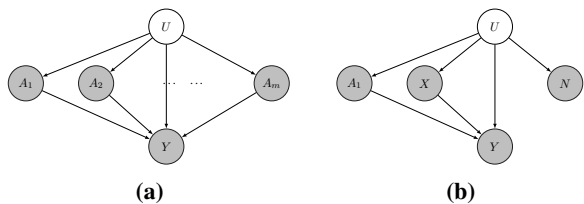


Figure 1. (a) Multiple treatments with shared confounding. (b) Proxy variables for an unobserved confounder (Miao et al., 2018). (Only the shaded nodes are observed.)

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; Čevič 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, \dots, 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, $P(Y | \text{do}(A_C = a_C))$. It is the distribution of the outcome Y if we intervene on $A_C \subset \{A_1, \dots, A_m\}$, 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., $P(y, a_1, \dots, a_m)$ 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 $P(y | \text{do}(a_C))$ 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 $P(y | \text{do}(a_1))$. 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 $P(y | \text{do}(a_1))$ 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, $P(y | \text{do}(a_1))$ 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 $P(y | \text{do}(a_C))$.

Partition the m treatments into three sets: A_C is the set of treatments on which we intervene; A_X is the set of treatments we use as a proxy; A_N is the set of treatments such that there exists a function $f(A_N)$ 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

¹We abbreviate $P(y | \text{do}(a_C)) \triangleq P(y | \text{do}(A_C = a_C))$.

variable strategy. Sets A_C , A_X and A_N must be non-empty.

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

1. *The outcome Y does not depend on $f(A_N)$:*

$$f(A_N) \perp Y | U, A_C, A_X, \quad (1)$$

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

2. *The conditional distribution $P(u | a_C, f(a_N))$ is complete² in $f(a_N)$ for almost all a_C .*
3. *The conditional distribution $P(f(a_N) | a_C, a_X)$ is complete in a_X for almost all a_C .*

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

²Definition of “complete”: The conditional distribution $P(u | a_C, f(a_N))$ is complete in $f(a_N)$ for almost all a_C means for any square-integrable function $g(\cdot)$ and almost all a_C ,

$$\int g(u, a_C) P(u | a_C, f(a_N)) du = 0 \text{ for almost all } f(a_N)$$

if and only if $g(u, a_C) = 0$ for almost all u .

heart attack through the ratios HDL/LDL and TG/HDL (Milán 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:

$$P(y | \text{do}(a_{\mathcal{C}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}}) da_{\mathcal{X}} \quad (2)$$

for any solution h to the integral equation

$$P(y | a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} | a_{\mathcal{C}}, f(a_{\mathcal{N}})) da_{\mathcal{X}}. \quad (3)$$

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](#). \square

[Theorem 1](#) identifies the intervention distributions of subsets of the treatments $A_{\mathcal{C}}$; it writes $P(y | \text{do}(a_{\mathcal{C}}))$ 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 $P(y | \text{do}(a_i))$, $i = 1, \dots, m$. 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 $P(y | \text{do}(a_{\mathcal{C}}))$ 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 $P(y | \text{do}(a_{\mathcal{C}}))$ in practice?

We revisit the deconfounder algorithm in [Wang & Blei \(2019a\)](#). We show that the deconfounder correctly estimates the intervention distribution $P(y | \text{do}(a_{\mathcal{C}}))$; 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, \dots, 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, \dots, A_m , it first constructs a random variable \hat{Z} such that all the treatments are conditionally independent:

$$\hat{P}(a_1, \dots, a_m, \hat{z}) = \hat{P}(\hat{z}) \prod_{j=1}^m \hat{P}(a_j | \hat{z}), \quad (4)$$

where $\hat{P}(\cdot)$ is consistent with the observed data $P(a_1, \dots, a_m) = \int \hat{P}(a_1, \dots, a_m, \hat{z}) d\hat{z}$. 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 $\hat{P}(y | a_1, \dots, a_m, \hat{z})$. This *outcome model* is fit to be consistent with the observed data:

$$P(y, a_1, \dots, a_m) = \int \hat{P}(y | a_1, \dots, a_m, \hat{z}) \hat{P}(a_1, \dots, a_m, \hat{z}) d\hat{z}. \quad (5)$$

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

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

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

$$\hat{P}(y | \text{do}(a_C)) \triangleq \int \hat{P}(y | a_1, \dots, a_m, \hat{z}) \times \hat{P}(a_{\{1, \dots, m\} \setminus C}, \hat{z}) d\hat{z} da_{\{1, \dots, m\} \setminus C}. \quad (6)$$

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 $P(y | \text{do}(a_C))$.

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

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

$$\hat{P}(y | \text{do}(a_C)) = P(y | \text{do}(a_C)), \quad (7)$$

where $\hat{P}(y | \text{do}(a_C))$ 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 $\hat{P}(y, a_1, \dots, a_m, \hat{z})$. Assumption 2.2 guarantees that the deconfounder estimate can be written as

$$\hat{P}(y | a_C, \hat{z}) = \int \hat{h}(y, a_C, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} | \hat{z}) da_{\mathcal{X}} \quad (8)$$

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

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 $P(y | \text{do}(a_C))$, where $A_C \subset \{A_1, \dots, A_m\}$ 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^{sing} 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^{sing} 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 $P(y | \text{do}(a_C))$ because $\{U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}\}$ constitute an admissible set.

We can equivalently identify the intervention distributions $P(y | \text{do}(a_C))$ 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:*

$$\begin{aligned} P(y | u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \dots, a_m, s = 1) \\ = P(y | u^{\text{sng}}, z, a_1, \dots, a_m, s = 1). \end{aligned} \quad (9)$$

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

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

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. \square

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 $P(y | \text{do}(a_C))$ 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

$$S \perp (A, Y) | Z, U^{\text{sng}}. \quad (11)$$

3. We observe the non-selection-biased distribution

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

and the selection-biased distribution

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

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)).

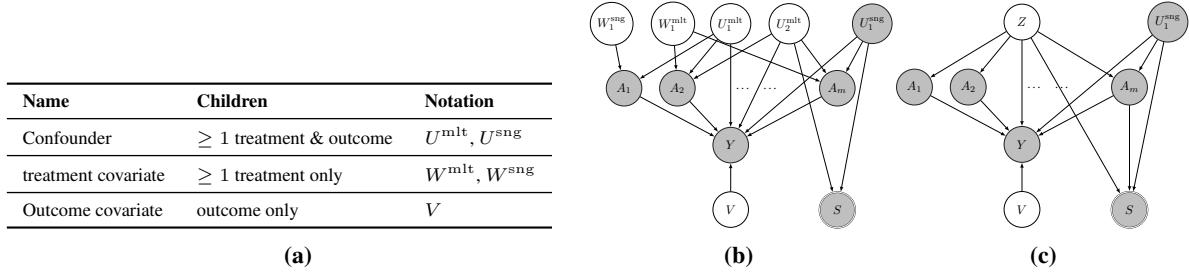


Figure 2. (a) Types of nodes (b) The expanded class of causal graphs. S is the selection operator. (c) The reduced causal graph with shared confounding.

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 $\emptyset \neq \mathcal{N} \subset \{1, \dots, m\} \setminus \mathcal{C}$ such that*

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

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

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

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:*

$$P(y \mid \text{do}(a_{\mathcal{C}})) \quad (13)$$

$$= \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}}) P(u_{\mathcal{C}}^{\text{sng}}) da_{\mathcal{X}} du_{\mathcal{C}}^{\text{sng}}$$

for any solution h to the integral equation

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

$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}})$$

$$\times P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) da_{\mathcal{X}}, \quad (14)$$

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 $P(a_1, \dots, a_m, u^{\text{sng}})$ and the selection-biased data $P(y, u^{\text{sng}}, a_1, \dots, a_m \mid s = 1)$. In this case, the algorithm outputs two estimates:

$$(1) \hat{P}(a_1, \dots, a_m, u^{\text{sng}}, \hat{z})$$

$$= \hat{P}(\hat{z}) \hat{P}(u^{\text{sng}} \mid a_1, \dots, a_m, \hat{z}) \prod_{i=1}^n \hat{P}(a_i \mid \hat{z}),$$

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

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

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

$$(2) \int \hat{P}(y, a_1, \dots, a_m, u^{\text{sg}}, \hat{z} | s = 1) d\hat{z} \\ = P(y, a_1, \dots, a_m, u^{\text{sg}} | s = 1).$$

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 $\hat{P}(a_1, \dots, a_m, u^{\text{sg}}, \hat{z})$ and $\hat{P}(y, a_1, \dots, a_m, u^{\text{sg}}, \hat{z} | s = 1)$ 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 $\hat{P}(y | \text{do}(a_C))$.

Finally the algorithm estimates

$$\hat{P}(y | \text{do}(a_C)) \\ \triangleq \int \hat{P}(y | a_1, \dots, a_m, \hat{z}, u_C^{\text{sg}}, s = 1) \\ \times \hat{P}(a_{\{1, \dots, m\} \setminus C}, \hat{z}) P(u_C^{\text{sg}}) du_C^{\text{sg}} d\hat{z} da_{\{1, \dots, m\} \setminus C}, \quad (15)$$

where U_C^{sg} are the single-treatment confounders that affect the treatments A_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 $\hat{P}(y, a_1, \dots, a_m, u^{\text{sg}}, \hat{z} | s = 1)$ and $\hat{P}(a_1, \dots, a_m, u^{\text{sg}}, \hat{z})$ that satisfy the following:*

1. *It is consistent with [Assumption 3.1](#):*

$$\hat{P}(a_1, \dots, a_m | \hat{z}, u^{\text{sg}}, s = 1) \\ = \hat{P}(a_1, \dots, a_m | \hat{z}, u^{\text{sg}}). \quad (16)$$

2. *It is consistent with [Assumption 4.1](#):*

$$\hat{P}(y | a_C, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}, u^{\text{sg}}, s = 1) \\ = \hat{P}(y | a_C, a_{\mathcal{X}}, \hat{z}, u^{\text{sg}}, s = 1). \quad (17)$$

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

The conditional $\hat{P}(\hat{z} | a_C, a_{\mathcal{X}}, u^{\text{sg}}, s = 1)$, [Eq. 16](#), and [Eq. 17](#) can be computed from $\hat{P}(a_1, \dots, a_m, u^{\text{sg}}, \hat{z})$ and $\hat{P}(y, a_1, \dots, a_m, u^{\text{sg}}, \hat{z} | s = 1)$.

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:*

$$\hat{P}(y | \text{do}(a_C)) = P(y | \text{do}(a_C)). \quad (18)$$

(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, \dots, 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^{sg} is a single-treatment confounder; it affects both burger consumption A_1 and body fat percentage Y ; U^{sg} 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,

$$U^{\text{mlt}} = \epsilon_{U^{\text{mlt}}}, U^{\text{sg}} = \epsilon_{U^{\text{sg}}}, W^{\text{mlt}} = \epsilon_{W^{\text{mlt}}}, \\ A_1 = \alpha_{A_1 U} U^{\text{mlt}} + \alpha_{A_1 W} W^{\text{mlt}} + \alpha_{A_1 U'} U^{\text{sg}} + \epsilon_{A_1}, \\ A_i = \alpha_{A_i U} U^{\text{mlt}} + \alpha_{A_i W} W^{\text{mlt}} + \epsilon_{A_i}, i = 2, \dots, 10, \\ Y = \sum_{i=1}^{10} \alpha_{Y A_i} A_i + \alpha_{Y U} U^{\text{mlt}} + \alpha_{Y U'} U^{\text{sg}} + \epsilon_Y.$$

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): $P(y | \text{do}(a_1, a_2))$. (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 $P(y | \text{do}(a_1, a_2))$ is identifiable.

Now consider a simple deconfounder. Fit a 2-D probabilistic principal component analysis (PPCA) to the data about food consumption $\{A_1, \dots, A_{10}\}$; 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,

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

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 $\propto \text{Unif}(U; 0, 0.5)$. 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

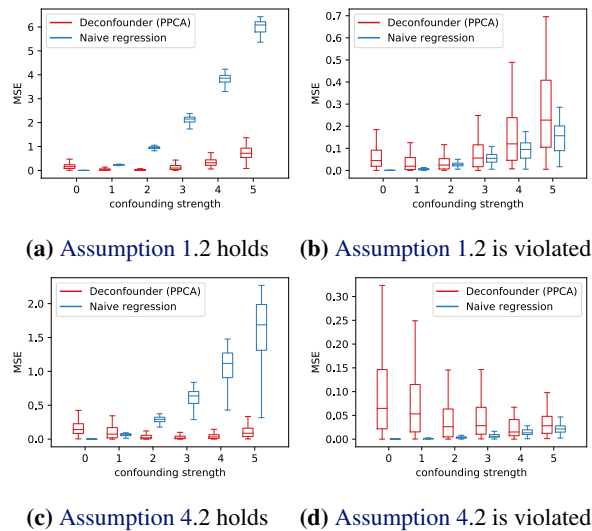


Figure 3. The deconfounder outperforms naive regression when the identification conditions are satisfied, but fails to otherwise.

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.

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 semi-parametric 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 confounding: 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). Naïve 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 inferring 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*.
- Ćevič, D., Bühlmann, P., & Meinshausen, N. (2018). Spectral deconfounding via perturbed sparse linear models.