The Blessings of Multiple Causes: A Causal Graphical View

Yixin Wang
Columbia University, New York City, USA

David M. Blei
Columbia University, New York City, USA

Abstract

Unmeasured confounding is a major hurdle for causal inference from observational data. Confounders—the variables that affect both the causes and the outcome—induce spurious non-causal correlations between the two. Wang and Blei (2018) lower this hurdle with "the blessings of *multiple* causes," where the correlation structure of multiple causes provides indirect evidence for unmeasured confounding. They leverage these blessings with an algorithm, called the deconfounder, that uses probabilistic factor models to correct for the confounders. In this paper, we take a causal graphical view of the deconfounder. In a graph that encodes shared confounding, we show how the multiplicity of causes can help identify intervention distributions. We then justify the deconfounder, showing that it makes valid inferences of the intervention. Finally, we expand the class of graphs, and its theory, to those that include other confounders and selection variables. Our results expand the theory in Wang and Blei (2018), justify the deconfounder for causal graphs, and extend the settings where it can be used.

Keywords: Causal inference, probabilistic models, structural causal models

Introduction

Unmeasured confounding is the major hurdle for causal inference from observational data. Confounders are variables that affect both the causes and the outcome. When measured, we can account for them with adjustments. But when unmeasured, they open back-door paths that bias the causal inference; 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. Thus when lifestyle is unmeasured, 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 unmeasured confounder.

To lower the hurdle of unmeasured confounding, Wang and Blei (2018) propose to dwell on *multiple* causes. They focus on settings where multiple causes affect a single outcome. They found that the correlation structure of the causes can reveal unmeasured *multi-cause* confounders, those that affect multiple causes and the outcome. They estimate those confounders with probabilistic factor models and then use them downstream in a causal inference. The correlation structure of the causes are "the blessings of multiple causes."

The example fits into this setting. Each type of food—broccoli, burgers, granola bars, pizza, and so on—is a potential cause of body fat. Further, each person's lifestyle affects multiple causes, i.e., their consumption of multiple types of food. People with a healthy lifestyle eat broccoli and granola more than burgers and pizza; people with an unhealthy lifestyle eat pizza and burgers more than they eat broccoli. Thus, using observed data, patterns of lifestyle might be revealed in the dependency structure of the foods. Then, by analyzing which foods each person eats, we can infer something about their lifestyle, something about the unmeasured confounder.

Leveraging this idea, Wang and Blei (2018) develop the "deconfounder" algorithm for causal inference. The deconfounder constructs a random variable—called the "substitute confounder"—that renders the causes conditionally independent; it then uses the substitute confounder to adjust for confounding bias. In the Rubin causal model, they prove the deconfounder leads to unbiased estimates of potential outcomes under the *single ignorability* assumption: each cause is conditionally independent of the potential outcome given the observed confounders. This assumption is weaker than the classical ignorability assumption prevalent in the potential outcomes literature.

Here we take a causal graphical view of the blessings of multiple causes. What causal quantities can be identified? How does single ignorability translate to assumptions on causal graphs? How does the multiplicity of the causes resolve causal identification? Does the deconfounder algorithm lead to valid causal estimates on causal graphs? These are the questions we study in this paper.

Consider multiple causal inference with *shared confounding*. This setting is in the causal graph of Figure 1b, where an unmeasured confounder U (lifestyle) affects multiple causes $\{A_1, \ldots, A_m\}$ (food choices) and an outcome Y (body fat). Further consider a subset of causes $\mathcal{C}$. We first prove that, under suitable conditions, the intervention distribution

$p(y \mid \text{do}(a_{\mathcal{C}}))$ is identifiable; it can be written in terms of the observational distribution. We then revisit the deconfounder. We show that it produces correct estimates of $p(y \mid \text{do}(a_{\mathcal{C}}))$; this result justifies the deconfounder on causal graphs.

Finally we generalize the result to the larger class of graphs in Figure 2b. This graph contains shared confounding, measured single-cause confounders (that only affect one cause), selection on the unobservables, and other structures. We prove identifiability in this larger class as well as the correctness of the deconfounder.

Taken together, these results expand the theory in Wang and Blei (2018), justify the deconfounder for causal graphs, and extend the settings where it can be used.

Related work. This work uses and extends causal identification with proxy variables (Kuroki and Pearl, 2014; Miao et al., 2018). While these works focus on a single cause and a single outcome, we analyze multiple causality. We leverage multiple causes to establish causal identification.

In more detail, proxy variables are the observable children of unmeasured confounders. Our key observation is that, in multiple causal inference, some causes can serve as proxies for causal identification of others. This observation helps identify the intervention distributions of subsets of causes; for example, the intervention distributions of each individual cause is identifiable. Further, unlike previous work in proxy variables, we do not need to find two independent proxies for the unobserved confounder; some causes themselves can serve as proxies for identifying the effect of the other causes.

Finally, this paper connects to the growing literature on multiple causal inference (Tran and Blei, 2017; Wang and Blei, 2018; Ranganath and Perotte, 2018; Heckerman, 2018). While most of these works focus on developing algorithms, we focus on theoretical aspects of the problem, expanding the ideas of Wang and Blei (2018) to identification and estimation in causal graphs. We note that D'Amour (2019) provides examples in multiple causal inference where the intervention distribution is not identifiable. The results in this work do not contradict those. Rather, we focus on the intervention distributions of *subsets* of the causes; D'Amour (2019) focuses on the intervention distributions of *all* the causes.

Multiple causes with shared confounding

Consider a causal inference problem where multiple causes of interest affect a single outcome; it is a *multiple causal inference*. Start with a simple causal graph (Figure 1b) where all the causes share the same unmeasured confounding. We will establish causal identification in this setting and prove the validity of causal estimation with the deconfounder algorithm. We will extend these results to more general causal graphs.

Multiple causal inference

Multiple causal inference focuses on a setting where multiple causes of interest affect a single outcome. The goal is to study the distribution of the outcome if we intervene on the causes. Multiple causal inference deviates from classical causal inference where a causal graph is built for a single cause and a single outcome.

Think of the meal/body-fat example in the introduction. It is a multiple causal inference: the three causes are the amount of pizza, burger, and broccoli a person eats everyday; the outcome is the person's body fat. Figure 1a illustrates the causal graph of this multiple causal inference. Figure 1b gives another example of multiple causal inference: we have m causes $A_1, \ldots, A_m$ that all affect the outcome Y.

The goal of multiple causal inference is to estimate the intervention distributions

$$P(y \mid \text{do}(a_{\mathcal{C}})) \triangleq P(y \mid \text{do}(A_{\mathcal{C}} = a_{\mathcal{C}})).$$

They describe the distribution of the outcome Y if we intervene on the set of causes $A_{\mathcal{C}} = \{A_i : i \in \mathcal{C}\}$; the set $\mathcal{C} \subseteq \{1, \ldots, m\}$ is the indices of the causes we intervene on.

Causal identification

We start with studying causal identification in multiple causal inference. We focus on the setting of shared unmeasured confounding: m causes $A_1, \ldots, A_m$ share the same unobserved confounder U as in Figure 1b.

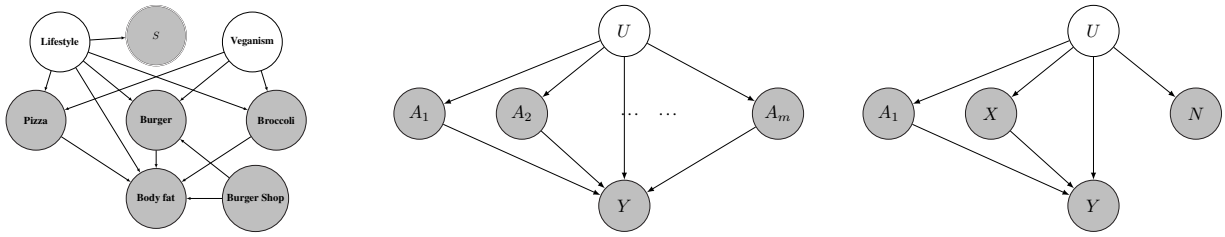
What intervention distributions can be identified with multiple causes under shared unmeasured confounding? In this section, we prove that the intervention distributions of *subsets of the causes* $P(y \mid \text{do}(a_{\mathcal{C}})), \mathcal{C} \subset \{1, \ldots, m\}$ are identifiable under suitable conditions.

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 1b) (Pearl, 2009). Identifiability ensures that an intervention distribution is *estimable* from the observed data.

The starting point of causal identification with multiple causes is the *proxy variable* approach; it focuses on causal identification with a *single* cause (Kuroki and Pearl, 2014; Miao et al., 2018). Consider the causal graph (Figure 1c) with a cause A_1, an outcome Y, and an unobserved confounder U. The goal is to estimate the intervention distribution $P(y \mid \text{do}(a_1))$. A *proxy* is an observable child of the unobserved confounder, e.g. X; a *null proxy* is a proxy that does not affect the outcome, e.g. N. The intervention distribution $P(y \mid \text{do}(a_1))$ 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, $P(y \mid \text{do}(a_1))$ is identifiable with the proxies N and X in Figure 1c.

Leveraging the proxy variable approach, we identify intervention distributions for *multiple causes* with shared unobserved confounding. The idea is to use some causes as proxies to identify the intervention distributions of the other causes. With multiple causes, we do not need to seek external proxy variables in the data collection process; the causes themselves can serve as proxies. Nor do we need to find a single variable as the null proxy that does not affect the outcome (like N in Figure 1c). We only need to find some function of the causes that does not affect the outcome. For example, the sum of two causes can be a null proxy if the sum does not affect the outcome. This idea liberates us from collecting external data about the proxy variables in observational studies. With multiple causes, we can work solely with the data about the causes and the outcome. This is the "blessings of multiple causes" from the causal graphical view.

Below we formally state the identification result for multiple causes. Assume the causal graph in Figure 1b with m



(a) (b) (c)

Figure 1: (a) Causal graph of the meal/body-fat example. (b) Multiple causes with shared confounding. (c) Proxy variables for an unobserved confounder. (Only the shaded nodes are observed.)

causes $A_1, \ldots, A_m$, an outcome Y, and a shared unobserved confounder U. The goal is to identify the intervention distribution of a *strict subset* of the causes $P(y \mid \mathrm{do}(a_\mathcal{C}))$ where $A_\mathcal{C} = \{A_i : i \in \mathcal{C}\}$ and $\mathcal{C} \subset \{1, \ldots, m\}$.

We partition the m causes $\{A_1, \ldots, A_m\}$ into three sets: $A_\mathcal{C}$ is the set of causes we intervene on; $A_\mathcal{X}$ is the set of causes we use as proxy; $A_\mathcal{N}$ is the set of causes such that $f(A_\mathcal{N})$ can serve as a null proxy. The latter two mimics the proxy X and the null proxy N in the proxy variable approach. The 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[1] 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 ensures that a set of causes $A_\mathcal{N}$ exists such that some function of it—$f(A_\mathcal{N})$—can serve as a null proxy (Equation (1)). Roughly, it requires $f(A_\mathcal{N})$ does not affect the outcome. Such an f often exists when the number of causes is large (Dasgupta and Gupta, 2003).

Assumption 1.2 and Assumption 1.3 are two completeness conditions on the true causal model; they are required by the proxy variable approach (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.

Many common statistical models satisfy the completeness condition. Examples include exponential families (Newey and Powell, 2003), location-scale families (Hu and Shiu, 2018), and nonparametric regression models (Darolles et al., 2011). Completeness is a common assumption posited in nonparametric causal identification (Miao et al., 2018; D'Haultfoeuille, 2011); it is often used to guarantee the existence and the uniqueness of solutions to integral equations. We refer the readers to Chen et al. (2014) for a detailed discussion of completeness.

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

Theorem 1. *(Causal identification under shared confounding) Assume the causal graph Figure 1b. Under Assumption 1, the intervention distribution of the causes $A_\mathcal{C}$ 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 Equation (3) *always exists under weak regularity conditions in the appendix.*

Proof sketch. The proof of Theorem 1 relies on the partition of the m causes: $A_\mathcal{C}$ as the causes, $A_\mathcal{X}$ 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 1c; all the m causes 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 the appendix. □

Theorem 1 identifies the intervention distributions of subsets of the causes $A_\mathcal{C}$; 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 causes $P(y \mid \mathrm{do}(a_i)), i = 1, \ldots, m$. By using the causes themselves as proxies, Theorem 1 exemplifies how the multiplicity of the causes enables causal identification under shared unmeasured confounding.

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,

[1]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 $a_\mathcal{C}$,

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

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

it involves solving an integral equation (Equation (3)). This integral equation is hard to solve except in the simplest linear Gaussian case (Carrasco et al., 2007). How can we estimate $P(y \mid \text{do}(a_{\mathcal{C}}))$ in practice?

In this section, we revisit the deconfounder algorithm developed in Wang and Blei (2018). We show that the deconfounder correctly estimates the intervention distribution $P(y \mid \text{do}(a_{\mathcal{C}}))$; it implicitly solves the integral equation in Equation (3) by modeling the data. These results justify the deconfounder from a causal graphical perspective.

The deconfounder algorithm. Given the causes $A_1, \ldots, A_m$ and the outcome Y, the deconfounder proceeds in three steps:

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

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

The random variable $\hat{Z}$ is called a *substitute confounder*; it does not necessarily coincide with the true confounder U. It can be constructed using probabilistic factor models (e.g. Mnih and Salakhutdinov (2008)).

2. **Fit an outcome model.** It obtains an estimate of how the outcome depends on the causes and the substitute confounder $\hat{Z}$,

$$\hat{P}(y \mid a_1, \ldots, a_m, \hat{z}).$$

The outcome model is fit to be consistent with the observed data:

$$P(y, a_1, \ldots, a_m) = \int \hat{P}(y \mid a_1, \ldots, a_m, \hat{z}) \hat{P}(a_1, \ldots, a_m, \hat{z}) \, \mathrm{d}\hat{z}. \tag{6}$$

Together with the first step, the deconfounder gives the joint distribution

$$\hat{P}(y, a_1, \ldots, a_m, \hat{z}).$$

We note that many possible $\hat{P}(\cdot)$'s satisfy the deconfounder requirement (Equations (4) to (6)). The deconfounder outputs one such $\hat{P}$. (We will show that any such $\hat{P}$ will lead to the correct causal estimate under suitable conditions.)

3. **Estimate the intervention distribution.** It estimates the intervention distribution $P(y \mid \text{do}(a_{\mathcal{C}}))$ by integrating out the causes that are not intervened on:

$$\hat{P}(y \mid \text{do}(a_{\mathcal{C}})) \triangleq \int \hat{P}(y \mid a_1, \ldots, a_m, \hat{z}) \times \hat{P}(a_{\{1,\ldots,m\} \setminus \mathcal{C}}, \hat{z}) \, \mathrm{d}\hat{z} \, \mathrm{d}a_{\{1,\ldots,m\} \setminus \mathcal{C}}. \tag{7}$$

The correctness of the deconfounder. We next show that the deconfounder estimate $\hat{P}(y \mid \text{do}(a_{\mathcal{C}}))$ is correct for $P(y \mid \text{do}(a_{\mathcal{C}}))$.

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

1. *It is consistent with Assumption 1.1:*

$$\hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}) = \hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, \hat{z}). \tag{8}$$

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

Both sides of Equation (8) *and $\hat{P}(\hat{z} \mid a_{\mathcal{C}}, a_{\mathcal{X}})$ are computed from the deconfounder output $\hat{P}(y, a_1, \ldots, a_m, \hat{z})$.*

Assumption 2.1—together with Assumption 1.1—roughly require that there exist some function f and a subset of the causes $A_{\mathcal{N}}$ such that $f(A_{\mathcal{N}})$ does not affect the outcome in both the deconfounder outcome model and the true causal model. We emphasize that the deconfounder does not require the specification of f and $A_{\mathcal{N}}$. However, we can still verify this assumption given the deconfounder estimate $\hat{P}(y, a_1, \ldots, a_m, \hat{z})$: we form candidates of $f(A_{\mathcal{N}})$ that satisfies Assumption 2.1; then we check if $f(A_{\mathcal{N}})$ satisfies Assumption 1.1. If one such $f(A_{\mathcal{N}})$ exists, both assumptions are satisfied.

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 result of the deconfounder.

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

$$\hat{P}(y \mid \text{do}(a_{\mathcal{C}})) = P(y \mid \text{do}(a_{\mathcal{C}})), \tag{9}$$

where the left hand side is computed from Equation (7)*.*

Proof sketch. The proof of Theorem 2 relies on a key observation: the deconfounder implicitly solves the integral equation (Equation (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

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}) \, \mathrm{d}a_{\mathcal{X}} \tag{10}$$

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

Theorem 2 justifies the deconfounder for multiple causal inference under shared confounding (Figure 1b). It proves that the deconfounder correctly estimates the intervention distributions when they are identifiable. This result complements Theorem 5.3 of Wang and Blei (2018); it focuses on estimating the intervention distributions of *subsets* of the causes. In contrast, Theorem 5.3 of Wang and Blei (2018) focuses on estimating the intervention distributions of *all* the causes. Their identification result relies on stronger assumptions.

Multiple causes on general causal graphs

We have discussed causal identification and estimation when multiple causes share the same unobserved confounder. In this section, we extend these discussions to more general causal graphs. We will describe these general causal graphs (Figure 2b). Under suitable conditions, we reduce them to one with shared confounding (Figure 2c). We then establish causal identification and estimate intervention distributions using results in the previous section.

General causal graphs

We focus on the class of general causal graphs[2] in Figure 2b. As in the shared confounding graph (Figure 1b), it has m causes $A_1, \ldots, A_m$ and an outcome Y; the goal is to estimate $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$, where $A_{\mathcal{C}} \subseteq \{A_1, \ldots, A_m\}$ is the set of the causes we intervene on. Apart from the causes and the outcome, the causal graph has a few other components. (Figure 2a contains a glossary of terms.)

Confounders U. Confounders are parents of *both* the causes and the outcome; they can be unobserved. For example, the U variables (U_i^{sng} and U_i^{mlt}) in Figure 2b are confounders; they have arrows into the outcome Y and at least one of the causes A_i. We differentiate between *single-cause* and *multi-cause* confounders. Single-cause confounders affect only one cause, e.g. U_i^{sng}; multi-cause confounders affect two or more causes, e.g. U_i^{mlt}.

Null confounders W. Null confounders are parents of the causes but *not* the outcome; they can be unobserved. In Figure 2b, the W variables are null confounders; they have arrows into at least one of the causes A_i but no arrow into the outcome Y. As with the confounders, we also differentiate between *single-cause* null confounders (e.g. W_i^{sng}) and *multi-cause* null confounder (e.g. W_i^{mlt}).

Covariates V. Covariates are parents of the outcome but *not* the causes, e.g., the V variables in Figure 2b. They do not affect any of the m causes; they can be unobserved.

Selection operator S. Following Bareinboim and Pearl (2012), we introduce a selection operator S into the causal graph, $S \in \{0, 1\}$. $S = 1$ indicates an individual being selected; otherwise, $S = 0$. We only observe the outcome of those individuals with $s = 1$. Figure 2b allows selection to occur on the confounders (e.g. $U_i^{\mathrm{sng}}, U_i^{\mathrm{mlt}}$).

The four sets of variables—confounder, null confounder, covariates, and the selection operator—compose the more general causal graphs with multiple causes. We study identification and estimation on these causal graphs.

Causal identification

We first extend the causal identification result with shared confounding (Theorem 1) to the more general causal graphs. The idea is to first reduce the general causal graph (Figure 2b) to the one with shared confounding (Figure 2c). After the reduction, we leverage Theorem 1 to establish causal identification on the reduced graph (Figure 2c). These results finally lead to causal identification on the original general causal graph (Figure 2b).

[2]There exist causal graphs that do not fall in this class (Figure 2b); we leave them for future work.

Reduction to shared confounding. We first reduce the general causal graph (Figure 2b) to one with shared confounding (Figure 2c). In graphs with *shared confounding* (e.g. Figure 1b and Figure 2c), all the causes share an unobserved confounder, which renders the causes conditionally independent. Assume the general graph; we will show that we can equivalently use the reduced graph to identify the intervention distributions $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

The key to this reduction is the following observation: a shared confounder Z must "capture"[3] all multi-cause confounders and multi-cause null confounders because Z renders all the causes conditionally independent.

Consider a random variable Z that renders all the causes $A_1, \ldots, A_m$ conditionally independent as in Figure 2c. We claim that Z must capture all the multi-cause confounders and null confounders $\{U^{\mathrm{mlt}}, W^{\mathrm{mlt}}\}$. We can prove this claim by contradiction. Imagine there exists some multi-cause confounder U_i^{mlt} that is not captured by Z. This multi-cause confounder U_i^{mlt} will induce dependence among the causes because U_i^{mlt} affects two or more causes by definition. Due to this dependence, the m causes could not have been conditionally independent given Z because Z does not capture U_i^{mlt}. It contradicts the fact that Z renders all the causes conditionally independent. This argument shares the same spirit with the substitute confounder argument in Wang and Blei (2018).

Following this discussion, we can reduce all the multi-cause confounders and null confounders $\{U^{\mathrm{mlt}}, W^{\mathrm{mlt}}\}$ into a shared confounder Z; $\{U^{\mathrm{mlt}}, W^{\mathrm{mlt}}\}$ constitute an admissible set. Therefore, assuming the general causal graph (Figure 2b), we can equivalently identify the intervention distributions $P(y \mid \mathrm{do}(a_{\mathcal{C}}))$ using a reduced causal graph (Figure 2c); it involves only the single-cause confounders U^{sng} and a shared confounder Z. This reduction allows us to generalize the identification result under shared confounding (Theorem 1) to the more general causal graphs (Figure 2b).

We now formally state the validity of this reduction step.

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

$$P(y \mid u^{\mathrm{sng}}, u^{\mathrm{mlt}}, w^{\mathrm{mlt}}, a_1, \ldots, a_m, s = 1) = P(y \mid u^{\mathrm{sng}}, z, a_1, \ldots, a_m, s = 1). \quad (11)$$

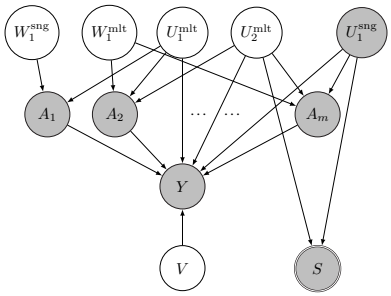
Proof sketch. The proof uses a measure-theoretic argument to characterize the information contained in the Z variable in Figure 2c. The full proof is in the appendix. □

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

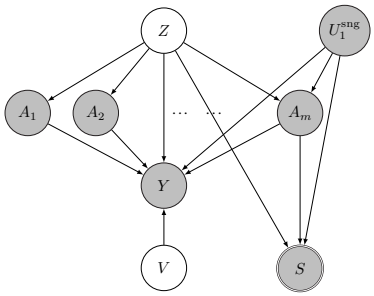
[3]The random variable A "captures" the random variable B if A contains all the information of B. Technically, it means the sigma algebra of the former is large than or equal to that of the latter: $\sigma(B) \subset \sigma(A)$.

Name	Definition	E.g.
Confounder	Parents of ≥ 1 cau. & ≥ 1 out.	U^{mlt} U^{sng}
Null Confounder	Parents of ≥ 1 cau. & 0 out.	W^{mlt} W^{sng}
Covariate	Parents of 0 cau. & ≥ 1 out.	V

(a)



(b)



(c)

Figure 2: (a) Glossary of terms. ("cau." = causes; "out." = outcome.) (b) The class of more general causal graphs. (S is the selection operator.) (c) The reduced causal graph with shared confounding. (The shaded nodes are observed.)

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 causes. This structure is similar to the unobserved shared confounding Figure 1b. In addition to the shared confounder Z, the reduced graph involves single-cause 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-cause confounders U_i^{sng}'s are observed.*
2. *The selection operator S satisfies*
$$S \perp (A, Y) \mid Z, U^{\text{sng}}. \tag{12}$$
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 causes must be observed. It allows us to adjust for confounding due to these single-cause 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 causes and single-cause-confounders. It aligns with common conditions required by recovery under selection bias (e.g., Theorem 2 of Bareinboim et al. (2014)).

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-cause 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 $f(A_{\mathcal{N}})$:*
$$f(A_{\mathcal{N}}) \perp Y \mid Z, A_{\mathcal{C}}, A_{\mathcal{X}}, U^{\text{sng}}, S = 1 \tag{13}$$
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-cause 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 causes $A_{\mathcal{C}}$ is identifiable:*

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

for any solution h to the integral equation

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) \tag{15}$$
$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \text{d}a_{\mathcal{X}}, \tag{16}$$

where $U_{\mathcal{C}}^{\text{sng}}$ is the single-cause confounders affecting $A_{\mathcal{C}}$. Moreover, the solution to Equation (15) *always exists under weak regularity conditions in the appendix.*

Proof sketch. The proof adopts a similar argument as in the proof of Theorem 1. We only need to take care of the additional (observed) single-cause confounders and the selection operator. In particular, Assumption 3.2 lets us shift from the selection biased distribution $P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1)$ to the non-selection-biased one $P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}})$.The full proof is in the appendix. □

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

Theorem 5. *Assume the causal graph Figure 2b. Assume a variant of Assumption 3 and Assumption 4 (detailed in the appendix), the intervention distribution of the causes $A_{\mathcal{C}}$ is identifiable using Equation* (14) *and Equation* (15)*.*

Proof sketch. This result is a direct consequence of Lemma 3 and Lemma 4. The full proof is in the appendix. □

Causal estimation with the deconfounder

We finally prove the correctness of the deconfounder algorithm on general causal graphs. We build on the identification result on general causal graphs (Theorem 5). We then show that the deconfounder provides correct causal estimates by implicitly solving the integral equation (Equation (15)). This argument is similar to the argument of Theorem 2.

The deconfounder algorithm for general causal graphs with selection bias extends the version described in the previous section (See the appendix for details). It outputs $\hat{P}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)$ and $\hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$.

We make a variant of Assumption 2 and state the correctness result for the deconfounder on general causal graphs.

Assumption 5. *The deconfounder estimates satisfy:*

1. *It is consistent with Assumption 3.1:*

$$\hat{P}(a_1, \ldots, a_m \mid \hat{z}, u^{\text{sng}}, s = 1) = \hat{P}(a_1, \ldots, a_m \mid \hat{z}, u^{\text{sng}}). \quad (17)$$

2. *It is consistent with Assumption 4.1:*

$$\hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}, u^{\text{sng}}, s = 1) = \hat{P}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, \hat{z}, u^{\text{sng}}, s = 1). \quad (18)$$

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 $a_{\mathcal{C}}$.*

$\hat{P}(\hat{z} \mid a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}, s = 1)$, Equation (17)*, and Equation* (18) *are computed from the deconfounder estimate.*

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

$$\hat{P}(y \mid \text{do}(a_{\mathcal{C}})) = P(y \mid \text{do}(a_{\mathcal{C}})). \quad (19)$$

Proof sketch. The proof of Theorem 6 follows a similar argument as in the proof of Theorem 2. We only need to additionally take care of the single-cause confounders and the selection operator. The full proof is in the appendix. □

Theorem 6 establishes the correctness of the deconfounder on general causal graphs under certain types of selection bias. It justifies the deconfounder on general causal graphs.

We illustrate Theorem 5 and Theorem 6 with a linear example in the appendix.

Discussion

We take a causal graphical view of "the blessings of multiple causes". Treating some causes as proxy variables of the shared confounder, we can identify the intervention distributions of the other causes. Thus, for a general class of causal graphs, we prove that the intervention distribution of subsets of causes is identifiable. Further, we show that the deconfounder algorithm makes valid inferences of these intervention distributions, a result that justifies the deconfounder on causal graphs. The results of this paper generalize the theory in Wang and Blei (2018) and extends the applicability of the deconfounder.

References

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

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

Carrasco, M., Florens, J.-P., and 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., and Newey, W. K. (2014). Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809.

D'Amour, A. (2019). On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. *arXiv preprint arXiv:1902.10286*.

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

Dasgupta, S. and Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65.

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

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

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

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

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

Mnih, A. and Salakhutdinov, R. R. (2008). Probabilistic matrix factorization. In *Advances in neural information processing systems*, pages 1257–1264.

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

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

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

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

Wang, Y. and Blei, D. M. (2018). The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*.

Appendix

Example: A linear causal model

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

Consider the meal/body-fat example. The causes are 10 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-cause confounder. Whether a person is vegan W^{mlt} is a multi-cause null confounder. Both U^{mlt} and W^{mlt} are unobserved. Whether one has easy access to good burger shops U^{sng} is a single-cause 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. Figure 1a shows the causal graph with a subset of the causes.

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 causes.) 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 and Blei (2018) also checks the model to ensure it fits the distribution of the assigned causes. (Let's assume that 2-D PPCA passes this check.)

PPCA leads to a linear estimate of the substitute confounder,

$$\hat{Z} = \left(\sum_{i=1}^{10} \gamma_{1i}A_i + \epsilon_{1\hat{Z}}, \sum_{i=1}^{10} \gamma_{2i}A_i + \epsilon_{2\hat{Z}}\right), \tag{20}$$

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

This substitute confounder $\hat{Z}$ satisfies Assumption 5. Plausibly, 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 Equation (22). Theorem 6 guarantees that this estimate is correct.

The deconfounder on general causal graphs

The deconfounder algorithm for general causal graphs with selection bias extends the version described in the previous section. Specifically, Assumption 2 allows the deconfounder 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 deconfounder algorithm outputs two estimates:

$$\hat{P}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z}) = \hat{P}(\hat{z})\hat{P}(u^{\text{sng}} \mid a_1, \ldots, a_m, \hat{z}) \prod_{i=1}^{n} \hat{P}(a_i \mid \hat{z}), \tag{21}$$

and

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

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

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

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

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

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}. Finally the deconfounder estimates

$$\begin{aligned} &\hat{P}(y \,|\, \mathrm{do}(a_{\mathcal{C}})) \\ &\triangleq \int \hat{P}(y \,|\, a_1, \ldots, a_m, \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s = 1) \\ &\quad \times \hat{P}(a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \hat{z}) P(u_{\mathcal{C}}^{\mathrm{sng}})\, \mathrm{d}u_{\mathcal{C}}^{\mathrm{sng}}\, \mathrm{d}\hat{z}\, \mathrm{d}a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \end{aligned} \tag{22}$$

where $U_{\mathcal{C}}^{\mathrm{sng}}$ are the single-cause confounders that affect the causes $A_{\mathcal{C}}$.

Proof of Theorem 1

Proof. The proof of Theorem 1 relies on two observations.

The first observation starts with the integral equation we solve:

$$P(y \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}})) \tag{23}$$

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

$$= \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \,|\, u) P(u \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}a_{\mathcal{X}}\, \mathrm{d}u. \tag{25}$$

The first equality is due to Equation (3). The second equality is due to the conditional independence implied by Figure 1b:

$$A_{\mathcal{X}} \perp A_{\mathcal{C}}, f(a_{\mathcal{N}}) \,|\, U.$$

The second observation relies on the null proxy:

$$P(y \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}})) \tag{26}$$

$$= \int P(y \,|\, u, a_{\mathcal{C}}, f(a_{\mathcal{N}})) P(u \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}u \tag{27}$$

$$= \int P(y \,|\, u, a_{\mathcal{C}}) P(u \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}u. \tag{28}$$

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 $Y \perp f(a_{\mathcal{N}}) \,|\, U, A_{\mathcal{C}}$. The reason is that

$$P(y \,|\, u, a_{\mathcal{C}}, f(a_{\mathcal{N}})) \tag{29}$$

$$= \int P(y \,|\, u, a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}})) P(a_{\mathcal{X}} \,|\, u, a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}a_{\mathcal{X}} \tag{30}$$

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

$$= P(y \,|\, u, a_{\mathcal{C}}). \tag{32}$$

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

Comparing Equation (25) and Equation (28) gives

$$\int \left[P(y \,|\, u, a_{\mathcal{C}}) - \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \,|\, u)\, \mathrm{d}a_{\mathcal{X}} \right] \times P(u \,|\, a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}u = 0, \tag{33}$$

which implies

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

This step is due to the completeness condition in Assumption 1.2.

Equation (34) leads to identification:

$$P(y \mid \mathrm{do}(a_{\mathcal{C}})) \tag{35}$$
$$= \int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u)\, \mathrm{d}a_{\mathcal{X}} P(u)\, \mathrm{d}u \tag{36}$$
$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}})\, \mathrm{d}a_{\mathcal{X}}. \tag{37}$$

Consider the special case of a single cause as in Figure 1c. 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)).

□

Proof of Theorem 2

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

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z})\, \mathrm{d}a_{\mathcal{X}} \tag{38}$$

under weak regularity conditions. (We will discuss the reason in later sections of the appendix.)

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

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}a_{\mathcal{X}}. \tag{39}$$

Given this claim (Equation (84)), we have

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

which proves the theorem. The first equality is due to Equation (7); the second is due to Equation (84); the third is due to Equation (5); the fourth is due to the above claim (Equation (84)) and Theorem 1.

We next prove the claim (Equation (84)). Start with the right side of the equality.

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

which establishes the claim. The first equality is due to Equations (4) and (5); the second is due to Equation (38); the third is due to Assumption 2.1, which implies

$$\hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}) = \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}). \tag{40}$$

Similar to Assumption 1.1, it is sufficient to assume Equation (40) directly. However, Assumption 2.1 is easier to check and more interpretable; it directly relates to the deconfounder outcome model.

□

Existence of solutions to the integral equations

Theorem 1 involves solving the integral equation

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}a_{\mathcal{X}}. \tag{41}$$

When does a solution exist for Equation (41)? 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*

$$P(y \mid z, x) = \int h(w, x, y) P(w \mid z, x)\, \mathrm{d}w \tag{42}$$

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 $K_x : L^2\{F(w \mid x)\} \to L^2\{F(z \mid x)\}, K_x(h) = \mathbb{E}\left[h(w) \mid z, x\right]$ for $h \in L^2\{F(w \mid x)\}$.

Leveraging Proposition 7, we can establish sufficient conditions for existence of a solution to Equation (41).

Corollary 8. *A solution exist for the integral equation Equation* (41) *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 Equation (10), Equation (15). The same argument also applies to the integral equation involved in Theorem 6:

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s = 1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\mathrm{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s = 1)\, \mathrm{d}a_{\mathcal{X}}. \tag{43}$$

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.

Proof of Lemma 3

The idea of the proof is to start with the structural equations of the general causal graph Figure 2b. Then posit the existence of a latent variable Z that renders all the causes 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-cause confounders U^{mlt}, multi-cause null confounders W^{mlt}, and some independent error. This result lets us establish

$$P(y \mid u^{\mathrm{sng}}, u^{\mathrm{mlt}}, w^{\mathrm{mlt}}, a_1, \ldots, a_m, s = 1) = P(y \mid u^{\mathrm{sng}}, z, a_1, \ldots, a_m, s = 1). \tag{44}$$

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

$$W_k = f_{W_k}(\epsilon_{W_k}), \qquad k = 1, \ldots, K, K \geq 0, \tag{45}$$
$$U_j = f_{U_j}(\epsilon_{U_j}), \qquad j = 1, \ldots, J, J \geq 0, \tag{46}$$
$$V_l = f_{V_l}(\epsilon_{V_l}), \qquad l = 1, \ldots, L, L \geq 0, \tag{47}$$
$$A_i = f_{A_i}(W_{S_{A_i}^W}, U_{S_{A_i}^U}, \epsilon_{A_i}), \qquad i = 1, \ldots, m, m \geq 2, \tag{48}$$
$$Y = f_y(A_1, \ldots, A_m, U_1, \ldots, U_K, V_1, \ldots V_L, \epsilon_Y), \tag{49}$$

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, $U_{S^U_{A_i}}$ indicates the set of confounders that have an arrow to both A_i and Y. $W_{S^W_{A_i}}$ indicates the set of null confounders that have an arrow to A_i; they do not have arrows to Y.

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

$$W^{\text{sng}} \triangleq \{W_1, \ldots, W_K\} / \bigcup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}). \tag{50}$$

To parse the notation above, recall that $W_{S^W_{A_i}}$ is the set of null confounders that affects A_i. $\bigcup_{i,j\in\{1,\ldots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}})$ 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-cause 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-cause confounders, all multi-cause null confounders and some independent error:*

$$\sigma(Z) = \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j\in\{1,\ldots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right), \tag{51}$$

$$= \sigma\left(\{\epsilon_Z\} \bigcup W^{\text{mlt}} \bigcup U^{\text{mlt}}\right). \tag{52}$$

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

We can parse the notation in Lemma 9 in the same way as in Equation (50): $\cup_{i,j\in\{1,\ldots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}})$ denotes the set of all multi-cause confounders; $\cup_{i,j\in\{1,\ldots,m\}:i\neq j}(U_{S^U_{A_i}} \cap U_{S^U_{A_j}})$ denotes the set of all multi-cause null confounders.

Proof. Without the loss of generality, we assume the compactness of representation in Equations (48) and (49). For any subset $\mathcal{S}$ of the random variables $\mathcal{S} \subset \{A_1, \ldots, A_m, Y\}$, 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 Equation (48). This assumption does not limit the class of causal graphs we study.

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

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

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

$$A_i \perp A_j \,|\, Z, \tag{54}$$

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

$$A_i \perp A_j \,|\, \sigma\left((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right), \tag{55}$$

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

$$\sigma\left((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{56}$$

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

$$\sigma\left(\bigcup_{i,j\in\{1,\ldots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{57}$$

Next Figure 2c implies

$$\sigma(A_1, \ldots, A_M) \not\subset \sigma(Z), \tag{58}$$

and

$$\sigma(Y) \not\subset \sigma(Z). \tag{59}$$

Therefore, we have

$$\sigma(Z) \subset \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right), \tag{60}$$

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 Equation (57) can be made equal to the right side of Equation (60). We have

$$\sigma(Z) = \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right) \tag{61}$$

$$= \sigma\left(\{\epsilon_Z\} \bigcup W^{\mathrm{mlt}} \bigcup U^{\mathrm{mlt}}\right). \tag{62}$$

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

$$P(y \mid u^{\mathrm{sng}}, u^{\mathrm{mlt}}, w^{\mathrm{mlt}}, a_1, \dots, a_m, s=1) = P(y \mid u^{\mathrm{sng}}, z, a_1, \dots, a_m, s=1), \tag{63}$$

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

Proof of Lemma 4

Proof. Denote $U^{\mathrm{sng}}_{\mathcal{C}}$ as the set of single-cause 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:

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1) \tag{64}$$

$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\mathrm{sng}}_{\mathcal{C}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}a_{\mathcal{X}} \tag{65}$$

$$= \int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\mathrm{sng}}_{\mathcal{C}}) P(a_{\mathcal{X}} \mid z) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}a_{\mathcal{X}}\, \mathrm{d}z \tag{66}$$

The first equality is due to Equation (15). The second equality is due to Assumption 3.2.

The second observation relies on the null proxy:

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1) \tag{67}$$

$$= \int P(y \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}z \tag{68}$$

$$= \int P(y \mid z, a_{\mathcal{C}}, u^{\mathrm{sng}}_{\mathcal{C}}, s=1) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}z \tag{69}$$

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 $Y \perp f(a_{\mathcal{N}}) \mid Z, U^{\mathrm{sng}}_{\mathcal{C}}, A_{\mathcal{C}}, S=1$. The reason is that

$$P(y \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1) \tag{70}$$

$$= \int P(y \mid z, a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1) P(a_{\mathcal{X}} \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}a_{\mathcal{X}} \tag{71}$$

$$= \int P(y \mid z, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\mathrm{sng}}_{\mathcal{C}}, s=1) P(a_{\mathcal{X}} \mid z, a_{\mathcal{C}}, u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}a_{\mathcal{X}} \tag{72}$$

$$= P(y \mid z, a_{\mathcal{C}}, u^{\mathrm{sng}}_{\mathcal{C}}, s=1). \tag{73}$$

The second equality is again due to Assumption 3.2.

Comparing Equation (66) and Equation (69) gives

$$\int \left[P(y \mid z, a_{\mathcal{C}}, u^{\mathrm{sng}}_{\mathcal{C}}, s=1) - \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\mathrm{sng}}_{\mathcal{C}}) P(a_{\mathcal{X}} \mid z)\, \mathrm{d}a_{\mathcal{X}}\right] \times P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\mathrm{sng}}_{\mathcal{C}}, s=1)\, \mathrm{d}z = 0, \tag{74}$$

which implies

$$P(y \mid z, a_{\mathcal{C}}, u^{\mathrm{sng}}_{\mathcal{C}}, s=1) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\mathrm{sng}}_{\mathcal{C}}) P(a_{\mathcal{X}} \mid z)\, \mathrm{d}a_{\mathcal{X}}. \tag{75}$$

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

Equation (75) leads to identification:

$$
\begin{aligned}
& P(y \mid \mathrm{do}(a_{\mathcal{C}})) && (76)\\
=& P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\mathrm{sng}}) P(z) P(u_{\mathcal{C}}^{\mathrm{sng}}) \,\mathrm{d}z \,\mathrm{d}u_{\mathcal{C}}^{\mathrm{sng}} && (77)\\
=& P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\mathrm{sng}}, s=1) P(z) P(u_{\mathcal{C}}^{\mathrm{sng}}) \,\mathrm{d}z \,\mathrm{d}u_{\mathcal{C}}^{\mathrm{sng}} && (78)\\
=& \int\int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\mathrm{sng}}) P(a_{\mathcal{X}} \mid z) \,\mathrm{d}a_{\mathcal{X}} P(z) P(u_{\mathcal{C}}^{\mathrm{sng}}) \,\mathrm{d}z \,\mathrm{d}u_{\mathcal{C}}^{\mathrm{sng}} && (79)\\
=& \int\int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\mathrm{sng}}) P(a_{\mathcal{X}}) P(u_{\mathcal{C}}^{\mathrm{sng}}) \,\mathrm{d}a_{\mathcal{X}} \,\mathrm{d}u_{\mathcal{C}}^{\mathrm{sng}}. && (80)
\end{aligned}
$$

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

□

Proof of Theorem 5

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

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

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

$$S \perp (A, Y) \mid U^{\mathrm{mlt}}, W^{\mathrm{mlt}}, U^{\mathrm{sng}}. \tag{81}$$

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

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

and the selection-biased distribution

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

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 $f(a_{\mathcal{N}})$:*

$$f(a_{\mathcal{N}}) \perp Y \mid A_{\mathcal{C}}, A_{\mathcal{X}}, U^{\mathrm{mlt}}, W^{\mathrm{mlt}}, U^{\mathrm{sng}}, S=1 \tag{82}$$

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

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

Proof of Theorem 6

We assume Assumption 6 and Assumption 7 as described in **??** .

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

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s=1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\mathrm{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s=1) \,\mathrm{d}a_{\mathcal{X}} \tag{83}$$

under weak regularity conditions. (We discuss the reason in **??** .)

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

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\mathrm{sng}}, s=1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\mathrm{sng}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\mathrm{sng}}, s=1) \,\mathrm{d}a_{\mathcal{X}}. \tag{84}$$

Given this claim (Equation (84)), we have

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

which proves the theorem. The first equality is due to Equation (22); the second is due to Equation (83); the third is due to Assumption 5 and $U_{\mathcal{C}}^{\mathrm{sng}}$ being the single-cause confounders for $A_{\mathcal{C}}$; the fourth is due to marginalizing out $\hat{Z}$; the fifth is due to the above claim (Equation (84)) and Theorem 5.

We next prove the claim (Equation (84)). Start with the right side of the equality.

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

which establishes the claim. The first equality is due to Equation (21); the second is due to Equation (83); the third equality is due to Assumption 5.2, which implies

$$\hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s=1) = \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s=1). \tag{85}$$

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

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

$$Y = \sum_{i=1}^{10} \alpha_{YA_i} A_i + \alpha_{YZ}\hat{Z} + \alpha_{YU'}U^{\mathrm{sng}} + \epsilon_Y. \tag{86}$$

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

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

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.

References

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

Wang, Y. and Blei, D. M. (2018). The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*.