Multiple Causal Inference with Bayesian Factor Models

Yixin Wang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Multiple Causal Inference with Bayesian Factor Models

Yixin Wang

Causal inference from observational data is a vital problem, but it comes with strong assumptions. Most methods assume that we observe all confounders, variables that affect both the cause variables and the outcome variables. But whether we have observed all confounders is a famously untestable assumption. In this dissertation, we develop algorithms for causal inference from observational data, allowing for unobserved confounding. These algorithms focus on problems of multiple causal inference: scientific studies that involve many causes or many outcomes that are simultaneously of interest.

Begin with multiple causal inference with many causes. We develop the *deconfounder*, an algorithm that accommodates unobserved confounding by leveraging the multiplicity of the causes. How does the deconfounder work? The deconfounder uses the correlation among the multiple causes as evidence for unobserved confounders, combining Bayesian factor models and predictive model checking to perform causal inference.

We study the theoretical requirements for the deconfounder to provide unbiased causal estimates, along with its limitations and trade-offs. We also show how the deconfounder connects to the proxy-variable strategy for causal identification (Miao et al., 2018) by treating subsets of causes as proxies of the unobserved confounder. We demonstrate the deconfounder in simulation studies and real-world data. As an application, we develop the *deconfounded recommender*, a variant of the deconfounder tailored to causal inference on recommender systems.

Finally, we consider multiple causal inference with many outcomes. We develop the *control-outcome deconfounder*, an algorithm that corrects for unobserved confounders using multiple negative control outcomes. Negative control outcomes are outcome variables for which the cause is *a priori* known to have no effect. The control-outcome deconfounder uses the correlation among these outcomes as evidence for unobserved confounders. We discuss the theoretical and empirical properties of the control-outcome deconfounder. We also show how the control-outcome deconfounder generalizes the method of synthetic controls (Abadie et al., 2010, 2015; Abadie and Gardeazabal, 2003), expanding its scope to nonlinear settings and non-panel data.

Table of Contents

Acknowledgments

This dissertation would not have been possible without the support of many people.

My adviser David Blei has been an exceptional mentor, collaborator, and role model. Dave's vision for machine learning led to this dissertation about causal inference, and his support sustained me through our adventure in causality. I will always look up to Dave for his ability to articulate ideas, his writing as "clear and simple as the truth," and his pursuit of both intuitive and rigorous research.

I thank the members of my dissertation committee: John Cunningham, Andrew Gelman, David Madigan, and Liam Paninski. John's advice and encouragement helped me persist through the challenging times of graduate school. Andrew's statistical communication class was an enlightening episode in my first year. David has been an amazing source of mentorship and inspiration. Liam's computational neuroscience class helped me appreciate serious machine learning endeavors for science.

I am also fortunate to have worked closely with many fascinating collaborators, including Alp Kucukelbir, Scott Linderman, Andrew Miller, Aaron Schein, Dhanya Sridhar, Wesley Tansey, Victor Veitch, Linying Zhang, and Jose Zubizarreta. All of them have influenced my research in significant ways, from teaching me to code to offering illuminating perspectives on research. I particularly thank Jose Zubizarreta for introducing me to causal inference and getting me started in causality research. I also wish to thank all members of the Blei lab and the Department of Statistics at Columbia. The stimulating reading group meetings and the fun tenth-floor lounge chats will always be part of my fond memories.

Finally, many thanks go to all my friends and family for their support. I especially thank my parents, who have given me the very best of everything they can offer. They have been extremely supportive of all my academic endeavors and have always pushed me to be the best I can be. I am forever grateful for their unbounded love and support.

Dedicated to my parents

1

Introduction

Here is a frivolous, but perhaps lucrative problem. Table 1.1 contains data about movies. For each movie, the table shows its cast of actors and how much money the movie made. Consider a movie producer deciding which actors to hire to maximize the revenue of her new movie. Making this decision requires her to contemplate "what if" questions: what would the movie revenue be if she hires Oprah Winfrey? What would the revenue be if she does not hire Oprah Winfrey?

These "what if" questions are key subjects of interest in causal inference. They characterize potential outcomes ("what *would* ... *be*"), also known as counterfactual outcomes. Differences between the potential outcomes quantify the causal effects; for example, how much would the movie revenue increase (or decrease) if she hires Oprah Winfrey the movie? To solve the producer's problem, we need to perform causal inference with the observational data from Table 1.1.

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, . . . }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, . . . }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, . . . }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, . . . }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, . . . }	$1506M
⋮	⋮	⋮

Table 1.1: Top earning movies in the TMDB dataset

Traditionally, causal inference from observational data is a difficult enterprise and requires strong assumptions. The challenge is that the observational dataset is limited; it contains the revenue of each movie, but only at its assigned cast. However, the producer's problem is not a traditional causal inference. While causal inference usually considers a single possible cause, such as whether a subject receives a drug or a placebo, our producer is considering a *multiple causal inference*, where each actor might causally contribute to the revenue.

In this dissertation, we study multiple causal inference, i.e. causal inference problems with many causes or many outcomes. We show how a multiple causal inference can be easier than traditional causal inference. Thanks to the multiplicity structure of the problem, the producer can make causal inferences under weaker untestable assumptions than traditional ones.

The contribution of the dissertation is a family of causal inference algorithms that exploit the multiplicity of the causes (or the outcomes) using Bayesian factor models. We study the theoretical requirements for these algorithms to provide unbiased causal estimates, along with their limitations and trade-offs. We finally examine their empirical properties through simulation studies and real applications.

Outline of the dissertation

- Chapter 2 reviews (multiple) causal inference with the potential outcomes notation. We introduce the task of causal identification and discuss a common identification strategy via a so-called unconfoundedness assumption. We finally describe the basic machinery of causal directed acyclic graphs (DAGs) and illustrate how it can help assess the unconfoundedness assumption.

- Chapter 3 develops the *deconfounder*, an algorithm that addresses unobserved confounding—a key challenge of causal inference from observational data—by leveraging the multiplicity of the causes. This algorithm is central to this dissertation, illustrating how Bayesian factor models and predictive model checking can help with multiple causal inference.
- Chapter 4 discusses the theoretical requirements of the deconfounder. We further develop theory around the deconfounder and establish causal identification. Specifically, we show that the deconfounder can provide unbiased estimates of three different causal quantities under suitable assumptions.
- Chapter 5 presents a causal graphical view of the deconfounder, delineating a different set of assumptions under which the deconfounder can correctly identify causal effects. Specifically, we treat subsets of causes as proxies of the unobserved confounder. This result also illustrates how the deconfounder can help put into practice the proxy-variable strategy for identification (Miao et al., 2018).
- Chapter 6 studies the deconfounder in three empirical studies. Two studies involve simulations of realistic scenarios; these help assess how well the deconfounder performs relative to the ground truth. The third study is a real-world analysis for the movie producer above. The studies demonstrate the benefits of the deconfounder, showing how the deconfounder can provide closer-to-truth causal estimates.
- Chapter 7 adapts the deconfounder to perform causal inference on recommender systems. The goal of recommendation is to show users items that they will like. Though usually framed as a prediction, the spirit of recommendation is to answer a causal question—for each user and movie, what would the rating be if we "forced" the user to watch the movie? To this end, we develop the *deconfounded recommender*, a causal inference approach to recommendation, one where watching a movie is a "treatment" and a user's rating is an "outcome."
- Chapter 8 extends the deconfounder to multiple causal inference with many outcomes. We develop the *control-outcome deconfounder*, an algorithm that corrects for unobserved confounding using multiple negative control outcomes. Negative control outcomes are outcome variables for which the cause is *a priori* known to have no effect. We show that the negative control outcomes can serve as evi-

dence for some unobserved confounders if they are affected by the same confounders as the outcome of interest. We also show that the control-outcome deconfounder generalizes the synthetic control method (Abadie et al., 2010, 2015; Abadie and Gardeazabal, 2003), expanding its scope to nonlinear settings and non-panel data.

- Chapter 9 discusses the debate around the deconfounder (D'Amour, 2018, 2019a,b,c; Ogburn et al., 2020, 2019a,b) and how it informs the theory and practice of the deconfounder.
- Chapter 10 summarizes the dissertation and discusses directions of future work in multiple causal inference.

2

(Multiple) causal inference

In this chapter, we set up (multiple) causal inference with the potential outcome notation, discuss a common causal inference strategy via a so-called unconfoundedness assumption, and illustrate how causal directed acyclic graphs (DAGs) can help assess this assumption.*

2.1 The potential outcome notation for (multiple) causal inference

Let's first return to the question of the movie producer: What would the movie revenue be if she hires Oprah Winfrey? This is a question about the potential movie revenue, a potential outcome. So, to answer this question, the producer can use the potential outcomes approach to causal inference (Imbens and Rubin, 2015a; Rubin, 1974, 2005).

*This chapter is in part based on Wang and Blei (2019a).

Begin with the potential outcomes notation. In the data, there are m *possible causes*, encoded in a vector $\boldsymbol{a} = (a_1, \ldots, a_m) \in \mathcal{A}^m$. We can consider a variety of types: real-valued causes, binary causes, integer causes, and so on. In the example of movie revenue, the causes are binary: a_j encodes whether actor j is in the movie.

For each individual i (movie) there is a *potential outcome function* that maps configurations of causes to the outcome(s). The outcome can be a scalar (e.g. the total movie revenue) or a vector (e.g. the movie revenue in l different countries). For the ith movie, the potential outcome function maps each possible cast to the movie's revenue had it had that cast, $\boldsymbol{y}_i(\boldsymbol{a}) : \mathcal{A}^m \to \mathcal{Y}^l$. The potential outcome function encodes, for example, how much money *Star Wars* would have made if Robert Redford replaced Harrison Ford as Han Solo.

When doing causal inference, the producer's goal is to estimate something about the population distribution of $\boldsymbol{Y}_i(\boldsymbol{a})$. For example, she might consider a particular cast $\boldsymbol{a}$ and characterize the sampling distribution of the potential outcomes $\boldsymbol{Y}_i(\boldsymbol{a})$ for each configuration of the causes $\boldsymbol{a}$. This distribution provides causal inferences, such as the expected outcome for a particular array of causes (a particular cast of actors) $\mu(\boldsymbol{a}) = \mathbb{E}[\boldsymbol{Y}_i(\boldsymbol{a})]$ or the average effect of individual causes (how much a particular actor contributes to revenue). Such a causal inference is a *multiple causal inference* if it involves a multi-dimensional cause vector ($m > 1$) and/or a multi-dimensional outcome ($l > 1$).

2.2 Observational data and causal identification

To make causal inferences, we draw data from the sampling distribution of assigned causes $\boldsymbol{a}_i$ (the cast of movie i) and realized outcomes $\boldsymbol{y}_i(\boldsymbol{a}_i)$ (its revenue).[†] The data is $\mathcal{D} = \{(\boldsymbol{a}_i, \boldsymbol{y}_i(\boldsymbol{a}_i)\}_{i=1}^n$. Note we only observe the outcome for the assigned causes $\boldsymbol{y}_i(\boldsymbol{a}_i)$, which is just one of the values of the potential outcome function. But we want to use such data to characterize the full distribution of $\boldsymbol{Y}_i(\boldsymbol{a})$ for any $\boldsymbol{a}$; this is the "fundamental problem of causal inference" (Holland, 1986).

[†]We use the term *assigned causes* for the vector of what some might call the "assigned treatments." Because some variables may not exhibit a causal effect, a more precise term would be "assigned potential causes" (but it is too cumbersome).

This disconnect between the observed data $\mathcal{D}$ and the potential outcome function $\boldsymbol{Y}_i(\boldsymbol{a})$ leads to a core question in causal inference: Is it even *possible* to estimate the expected potential movie revenue from the observed movie casts and revenues? Or more generally, are causal quantities like $\mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{a}))$ and $\mathbb{E}[\boldsymbol{Y}_i(\boldsymbol{a})]$ estimable from the observed data? Under what conditions are they estimable? These questions constitute the task of *causal identification*, whose goal is to find suitable conditions under which a causal quantity of interest is estimable, a.k.a. identifiable, from observational data.

More precisely, a causal quantity is *identifiable* if it can be written as a deterministic function of the observed data distribution. For example, suppose we can calculate the expected potential revenue of a new cast by fitting a linear regression to the observed data and predicting for the new cast. Then the mapping from the observed data to the prediction is the deterministic function that identifies the potential revenue, which is also known as the identification strategy. It tells us how to calculate the causal quantity of interest if we are given an infinite amount of data and hence can perfectly recover the observed data distribution.

If a causal quantity is non-identifiable, then no estimators or algorithms can estimate this quantity accurately, even with infinite observational data. However, this is not the end of the world. This causal quantity may still be partially identifiable, namely, the causal quantity must reside in an interval, whose endpoints are deterministic functions of the observed data distribution. For example, the observed movie revenues may imply that the expected potential revenue must be between \$3M and \$1300M for a cast of Brad Pitt and Oprah Winfrey. In such cases, we seek interval estimates of the causal quantity that cover the true value with high probability. These estimates are sometimes called *partial identification intervals* (Manski, 2003, 2010). These intervals are generally different from confidence intervals; confidence intervals often shrink to zero with infinite data but partial identification intervals do not.

2.3 Confounders and the unconfoundedness assumption

How can the producer calculate (or, more technically, identify) the expected potential revenue $\mathbb{E}[\boldsymbol{Y}_i(\boldsymbol{a})]$ from the observed data? Is it possible? What assumptions are required? We discuss these questions in this section.

Naively, the producer subsets the data in Table 1.1 to those with cast equal to $\boldsymbol{a}$ and then computes a Monte Carlo estimate of the revenue. This procedure is unbiased when $\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right] = \mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}\right]$. But there is a problem. The data in Table 1.1 hide *confounders*, variables that affect both the causes and the effect. For example, every movie has a genre, such as comedy, action, or romance. This genre has an effect on both who is in the cast and the revenue. (E.g., action movies cast a certain set of actors and tend to make more money than comedies.) When left unadjusted, the genre of the movie produces a statistical dependence between whether an actor is cast and the revenue; this dependence biases the causal estimates, $\mathbb{E}\left[\boldsymbol{Y}_i \mid \boldsymbol{A}_i = \boldsymbol{a}\right] \neq \mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]$.

Thus the main activities of traditional causal inference are to identify, measure, and control for confounders. Suppose the producer measures confounders for each movie W_i. Then the inference is simple: Use the data (now with confounders) to take Monte Carlo estimates of $\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a}) \mid W_i, \boldsymbol{A}_i = \boldsymbol{a}\right]\right]$; this iterated expectation "controls" for the confounders. But the problem is that whether the estimate is equal to $\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]$ rests on an uncheckable assumption: There are no other confounders. For many applied causal inference problems, this assumption is a leap of faith.

More precisely, suppose we measure covariates $\boldsymbol{X}_i$ and append them to each data point, $\mathcal{D} = \{(\boldsymbol{a}_i, \boldsymbol{x}_i, \boldsymbol{y}_i(\boldsymbol{a}_i)\}_{i=1}^n$. If these covariates contain all confounders then

$$\mathbb{E}\left[\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{A}_i) \mid \boldsymbol{X}_i, \boldsymbol{A}_i = \boldsymbol{a}\right]\right] = \mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]. \tag{2.1}$$

With augmented data, estimate the left side with Monte Carlo; thus estimate $\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]$. Equation (2.1) presents a deterministic mapping from the observed data distribution $\mathrm{P}(\boldsymbol{A}_i, \boldsymbol{Y}_i(\boldsymbol{A}_i), \boldsymbol{X}_i)$ to the causal quantity of interest $\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]$. In other words, Equation (2.1) identifies the expected potential outcome $\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]$.

Equation (2.1) is only true when $\boldsymbol{X}_i$ capture all confounders. More precisely, it is true under the assumption of *unconfoundedness*[‡] (Imai and Van Dyk, 2004; Rosenbaum and Rubin, 1983).

[‡]Here we describe the weak version of the unconfoundedness assumption, which requires individual potential outcomes $\boldsymbol{Y}_i(\boldsymbol{a})$ be marginally independent of the causes $\boldsymbol{A}_i$, i.e. $\boldsymbol{A}_i \perp\!\!\!\perp \boldsymbol{Y}_i(\boldsymbol{a}) \mid \boldsymbol{X}_i$ for all $\boldsymbol{a}$. Imbens (2000) and Hirano and Imbens (2004) call this assumption *weak unconfoundedness*. In contrast, the strong version of unconfoundedness says $\boldsymbol{A}_i \perp\!\!\!\perp (\boldsymbol{Y}_i(\boldsymbol{a}))_{\boldsymbol{a}\in\mathcal{A}} \mid \boldsymbol{X}_i$, which requires all possible potential outcomes $(\boldsymbol{Y}_i(\boldsymbol{a}))_{\boldsymbol{a}\in\mathcal{A}}$ be jointly independent of the causes $\boldsymbol{A}_i$.

Assumption 2.3.1: Unconfoundedness. The potential outcomes $\boldsymbol{Y}_i(\boldsymbol{a})$ are unconfounded if, conditional on observed covariates $\boldsymbol{X}_i$, the assigned causes are independent of the potential outcomes,

$$\boldsymbol{A}_i \perp\!\!\!\perp \boldsymbol{Y}_i(\boldsymbol{a}) \mid \boldsymbol{X}_i \qquad \forall \boldsymbol{a}. \tag{2.2}$$

The nuance is that Equation (2.2) needs to hold for all possible $\boldsymbol{a}$'s, not only for the value of $\boldsymbol{Y}_i(\boldsymbol{a})$ at the assigned causes. Unconfoundedness implies no unobserved confounders.

Beyond unconfoundedness, the identification of the potential outcome function further requires two other assumptions: *stable unit treatment value assumption (SUTVA)* (Rubin, 1980, 1990) and *overlap* (Imai and Van Dyk, 2004).

Assumption 2.3.2: Stable unit treatment value assumption (SUTVA) (Rubin, 1980, 1990). The potential outcomes satisfy the stable unit treatment value assumption (SUTVA) if

1. The treatment applied to one unit does not affect the outcome of another unit.
2. There is only one single version of each treatment level.

Loosely, SUTVA ensures that the potential outcome $\boldsymbol{Y}_i(\boldsymbol{a})$ is well-defined. In more details, the first part of SUTVA requires that there is no interference among the units; the second part requires that the same value of the assigned cause will always lead to the same distribution of potential outcomes. Both requirements guarantee that the potential outcome $\boldsymbol{Y}_i(\boldsymbol{a})$ is indeed only a function of $\boldsymbol{a}$, the assigned causes of unit i.

Assumption 2.3.3: Overlap (Imai and Van Dyk, 2004). The causes $\boldsymbol{A}_i$ and the covariates $\boldsymbol{X}_i$ satisfy the overlap assumption if

$$\mathrm{P}(\boldsymbol{A}_i \in \bar{\mathcal{A}} \mid \boldsymbol{X}_i) > 0 \tag{2.3}$$

for all sets $\bar{\mathcal{A}}$ with positive measure.

The overlap assumption roughly requires that for any values of the covariates, all values of the assigned causes must have positive probability. For example, suppose the cause is the presence of an actor and the covariate is the genre. Then, for each genre, it must be possible for the actor to be present or not present. The overlap assumption is violated if the actor can never appear in a particular genre, e.g. action movies.

Assumptions 2.3.1 to 2.3.3 identify the distribution of the potential outcomes:

$$\begin{aligned} \mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{a})) &= \mathbb{E}_X\left[\mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{a}) \mid \boldsymbol{X}_i)\right] && (2.4)\\ &= \mathbb{E}_X\left[\mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{X}_i)\right] && (2.5)\\ &= \mathbb{E}_X\left[\mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{A}_i) \mid \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{X}_i)\right] && (2.6)\\ &= \mathbb{E}_X\left[\mathrm{P}(\boldsymbol{Y}_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{X}_i)\right] && (2.7) \end{aligned}$$

Equation (2.4) is due to the tower property. Equation (2.5) is due to Assumption 2.3.1. Equation (2.6) is due to $\boldsymbol{Y}_i(\boldsymbol{A}_i) = \boldsymbol{Y}_i(\boldsymbol{a})$ when $\boldsymbol{A}_i = \boldsymbol{a}$. Equation (2.7) is due $\boldsymbol{Y}_i(\boldsymbol{A}_i) = \boldsymbol{Y}_i$. Finally, Equation (2.7) is estimable from observational data because $\mathrm{P}(\boldsymbol{A}_i = \boldsymbol{a} \mid \boldsymbol{X}_i) > 0$ due to Assumption 2.3.3.

Equation (2.7) underlies the common practice of causal inference: find and measure the confounders, estimate conditional expectations, and average. (It is also known as the *g*-formula (Hernán and Robins, 2020) or the backdoor adjustment (Pearl, 1988).) For example, the genre of the movie is a confounder of the causal inference of movie revenues. The genre affects both which cast is selected and the potential earnings of the film. But the assumption that there are no unobserved confounders is significant. One of the central challenges around causal inference from observational data is that unconfoundedness is untestable—it fundamentally depends on the entire potential outcome function, of which we only observe one value (Holland, 1986).

2.4 Assessing unconfoundedness with causal directed acyclic graphs (DAGs)

To assess the unconfoundedness assumption (Assumption 2.3.1), we have to rely on domain experts: Which measured covariates satisfy unconfoundedness? A useful tool for this question is causal directed acyclic graphs (DAGs).

A causal DAG is a directed acyclic graph that encodes the causal mechanisms between variables. Figure 2.1a is one example. Each node is a variable involved in the causal problem. Each arrow indicates a possible causal relationship from one variable to the other; nevertheless, the causal effect can be zero. The absence of an arrow indicates no causal relationship from one to the other. For example, an arrow exists from movie genre to movie revenue. It indicates that the movie genre can affect movie revenue but not the other way around.

Causal DAGs can involve both observed and unobserved (i.e. latent) variables. To distinguish the two, only the shaded nodes in the graphs are observed. For example, the movie genre is an unobserved variable in Figure 2.1a and all other variables are observed. Further, causal DAGs also encode confounding relationships: The common causes of the two variables of interest are confounders. For example, the movie genre confounds the effect of hiring Oprah Winfrey on the movie revenue.

As another example of the causal DAG, we consider Figure 2.1b for assessing how a person's diet affects her body fat percentage. Each type of food—broccoli, burgers, granola bars, pizza, and so on—is a potential cause of body fat. One confounder in Figure 2.1b is lifestyle; it affects a person's consumption of multiple types of food; it also affects the outcome, body fat. 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. Moreover, someone with a healthy lifestyle exercises frequently, which lowers her body fat. The variable veganism also has a similar relationship with food and body fat. Burger shop is yet another confounder that may affect both the consumption of burger and the level of body fat. In Figure 2.1b, both lifestyle and veganism are unobserved.

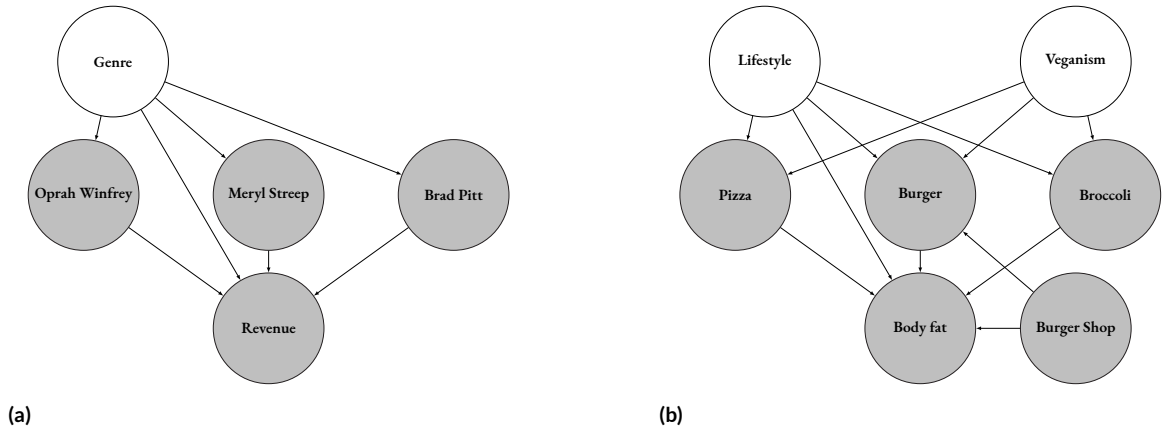


Figure 2.1: (a) A causal dag for movie actors and revenues. (b) A causal dag for food and body fat. (Only the shaded nodes are observed.)

Causal DAGs are useful for assessing the unconfoundedness assumption. Begin with designating the cause variable(s) and the outcome variable(s) of interest in the causal DAG. We can apply the back-door criterion to find out which variables will compose a covariate set that satisfies the unconfoundedness assumption (Pearl, 2009).

Definition 2.4.1: Backdoor criterion. A set of variables $\boldsymbol{X}$ satisfy the backdoor criterion to identify the effect of $\boldsymbol{A}$ on $\boldsymbol{Y}$ if

1. All directed paths between $\boldsymbol{A}$ and $\boldsymbol{Y}$ are unperturbed; there are no descendants of $\boldsymbol{A}$ in $\boldsymbol{X}$;
2. $\boldsymbol{X}$ blocks every backdoor paths from $\boldsymbol{A}$ to $\boldsymbol{Y}$, where a backdoor path is a path that remains if we were to remove any arrows pointing out of $\boldsymbol{A}$. Backdoor paths are non-causal paths from $\boldsymbol{A}$ to $\boldsymbol{Y}$.

Intuitively, the backdoor criterion ensures that all non-causal paths are blocked by the set of variables $\boldsymbol{X}$. Therefore, covariates that satisfy the backdoor criterion let us estimate the causal effects without being confounded by non-causal relationships.

If a set of covariates $\boldsymbol{X}$ satisfy the backdoor criterion to identify the effect of $\boldsymbol{A}$ on $\boldsymbol{Y}$, then it must satisfy the unconfoundedness assumption: $\boldsymbol{A}_i \perp\!\!\!\perp \boldsymbol{Y}_i(\boldsymbol{a}) \mid \boldsymbol{X}_i, \forall \boldsymbol{a}$. In causal DAG notations, it is

$$\mathrm{P}(\boldsymbol{Y}_i \mid \mathrm{do}(\boldsymbol{A}_i = \boldsymbol{a}), X_i) = \mathrm{P}(\boldsymbol{Y}_i \mid \boldsymbol{A}_i = \boldsymbol{a}, X_i), \tag{2.8}$$

where $\mathrm{P}(\boldsymbol{Y}_i \mid \mathrm{do}(\boldsymbol{A}_i = \boldsymbol{a}), X_i)$ is an the intervention distribution; it represents the distribution of Y_i conditional on X_i if we intervene on the variable A_i by setting it to be $\boldsymbol{A}_i = \boldsymbol{a}$. It is equivalent to $\mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{a}) \mid X_i)$ in the potential outcomes notation. (In general, we have the following notational equivalence between the causal DAGs and the potential outcomes: $\mathrm{P}(\boldsymbol{Y}_i \mid \mathrm{do}(\boldsymbol{A}_i = \boldsymbol{a})) = \mathrm{P}(\boldsymbol{Y}_i(\boldsymbol{a}))$.)

This connection between the backdoor criterion and the unconfoundedness assumption suggests a workflow for causal inference from observational data:

1. Set up the causal question. Designate the causes and outcomes of interest;
2. Consult a domain expert. Ask her to draw a causal DAG that involves all the relevant variables;
3. Identify a set of covariates that satisfy the backdoor criterion, usually labeled as the *sufficient set*;
4. Collect a dataset that measures the causes, the outcomes, and the sufficient set of covariates;
5. Calculate the causal effect by conditioning on the sufficient set (Equation (2.7)).

2.5 Discussion

In this chapter, we reviewed the potential outcome notation for causal inference, discussed the task of causal identification from observational data, presented a common set of identification assumptions that involve unconfoundedness, and showed how the machinery of causal DAGs can help us assess unconfoundedness.

This classical approach to causal identification relies on unconfoundedness, roughly requiring "no unobserved confounders." In practice, we may not be able to observe or measure all confounders. For example, applying the backdoor criterion reveals that lifestyle is a variable that must be measured in the food and body fat example. But accurate measurements of lifestyle are difficult to obtain, e.g., requiring expensive real-time tracking of activities. In such cases, lifestyle is necessarily an unmeasured confounder.

How can we identify causal quantities in the presence of unobserved confounders? Common approaches appeal to further domain knowledge. For example, the front door adjustment (Pearl, 2009) requires a set of measured variables that intercepts all the causal paths between the causes and outcomes of interest. The instrumental variable strategy (Angrist and Imbens, 1995; Angrist et al., 1996) uses the

so-called instruments, variables that affect the outcomes only through the causes. The proxy variable strategy (Kuroki and Pearl, 2014; Miao et al., 2018) leverages multiple proxies of unobserved confounders to assist causal identification. Finally, we can perform sensitivity analysis (Robins et al., 2000b) to assess how strong the unobserved confounding has to be to overturn causal conclusions. In the coming chapters, we will focus on this setting of causal inference with unobserved confounding.

3

Multiple causes and the deconfounder

In this chapter, we explore multiple causal inference with multiple causes and a single outcome.* We develop the deconfounder algorithm, an algorithm for the movie producer who worries about missing a confounder. First, the producer finds and fits a good latent-variable model to capture the dependence among actors. It should be a Bayesian factor model, one that contains a per-movie latent variable that renders the assigned cast conditionally independent. (Probabilistic principal component analysis (Tipping and Bishop, 1999) is a simple example, but there are many others.) Given the model, she then estimates the per-movie variable for each cast in the dataset; this estimated variable is a substitute for unobserved confounders. Finally, she controls for the substitute confounder and obtains valid causal inferences.

All methods for causal inference rely on assumptions. The deconfounder makes two major ones. First, it assumes that the fitted latent-variable model is a good model of the assigned causes. This assumption is testable, and we will use predictive checks to assess how well the fitted model captures the data. Second, it assumes that there are no unobserved single-cause confounders, variables that affect one cause

*This chapter is in part based on Wang and Blei (2019a).

(e.g., actor) and the potential outcome function (e.g., revenue). While this assumption is not testable, it is weaker than the usual assumption of unconfoundedness, which requires no unobserved confounders. Depending on the nature of the causal quantities of interest, the deconfounder requires further assumptions, which we will discuss in Chapter 4.

Subject to the assumptions, the deconfounder provides valid causal inferences because it capitalizes on the dependency structure of the observed casts to uncover unobserved confounders. It uses patterns of how actors tend to appear together in movies as indirect evidence for confounders in the data.

Beyond making movies, many causal inference problems, especially from observational data, also classify as multiple causal inference problems. Such problems arise in many fields.

- **Genome-wide association studies (GWAS).** In GWAS, biologists want to know how genes causally connect to traits (Stephens and Balding, 2009; Visscher et al., 2017). The assigned causes are alleles on the genome, often encoded as either being common ("major") or uncommon ("minor"), and the effect is the trait under study. Confounders, such as shared ancestry among the population, bias naive estimates of the effect of genes. We study GWAS problems in Section 6.2.

- **Computational neuroscience.** Neuroscientists want to know how the electrical activity of neurons produces observed behavior, such as limb movement (Churchland et al., 2012). The possible causes are multiple measurements about the brain's activity, e.g., one per neuron, and the effect is a measured behavior. Confounders, particularly through dependencies among neural activity, bias the estimated connections between brain activity and behavior.

- **Social science.** Sociologists and policy-makers want to know how social programs affect social outcomes, such as poverty levels and upward mobility (Morgan and Winship, 2015). However, individuals may enroll in several such programs, blurring information about their possible effects. In social science, controlled experiments are difficult to engineer; using observational data for causal inference is typically the only option.

- **Medicine.** Doctors want to know how medical treatments affect the progression of the disease. The multiple causes are medications and procedures; the outcome is a measurement of a disease (e.g., a lab test). There are many confounders—such as when and where a patient is treated or the treatment preferences of the attending doctor—and these variables bias the estimates of effects. While gold-standard data from clinical trials are expensive to obtain, the abundance of electronic health records could inform medical practices.
- **Recommender systems.** Technology companies want to know whether recommending different items to a user will increase revenue. The multiple causes are the recommendation of each item; the outcome is the total revenue of the company. However, the past purchase history of the users affects both which items are recommended and which items they buy, i.e., the revenue. Users' past purchase history thus confounds the observed effect of recommendation.

All of these problems of causal inference can use the deconfounder. Fit a good factor model of the assigned causes, infer substitute confounders, and use the substitutes in causal inference.

The rest of the chapter is organized as follows. Section 3.1 presents the deconfounder. Section 3.2 describes the identification strategy of the deconfounder and its main assumptions. Section 3.3 discusses the practical details of the deconfounder and presents the full algorithm. Section 3.4 answers some questions a reader might have.

3.1 Multiple causal inference with the deconfounder

In this section, we discuss the problem of multiple causal inference and develop the deconfounder.

3.1.1 The deconfounder

We develop the *deconfounder*, an algorithm that uses the multiplicity of causes to infer unobserved confounders. There are three steps. First, find a good latent variable model of the assignment mechanism. (A good model is one that accurately captures the joint distribution of the causes.) Second, use the model to infer the latent variable for each individual. Finally, use the inferred variable as a substitute for unobserved confounders and form causal inferences.

We explain the method and discuss why and when it provides unbiased causal inferences.

In the first step of the deconfounder, define and fit a *Bayesian factor model* (a.k.a. a probabilistic factor model) to capture the joint distribution of causes $p(a_1, \ldots, a_m)$. A Bayesian factor model posits per-individual latent variables Z_i, which we call local factors, and uses them to model the assigned causes. The model is

$$\begin{aligned} Z_i &\sim p(\cdot \mid \alpha) \quad i = 1, \ldots, n, \\ A_{ij} \mid Z_i &\sim p(\cdot \mid z_i, \theta_j) \quad j = 1, \ldots, m, \end{aligned} \tag{3.1}$$

where α parameterizes the distribution of Z_i and θ_j parameterizes the per-cause distribution of A_{ij}. Notice that Z_i can be multi-dimensional. Factor models encompass many methods from Bayesian statistics and probabilistic machine learning. Examples include probabilistic PCA (Tipping and Bishop, 1999), mixture models (McLachlan and Basford, 1988), mixed-membership models (Airoldi et al., 2008; Blei et al., 2003; Erosheva, 2003; Pritchard et al., 2000b), and deep generative models (Kingma and Welling, 2013; Mohamed and Lakshminarayanan, 2016; Neal, 1990; Ranganath et al., 2015, 2016; Rezende and Mohamed, 2015; Tran et al., 2017).

We can fit using any appropriate method, such as Bayesian inference. And exact fitting is not required; we can use approximate methods like the EM algorithm, Markov chain Monte Carlo, or variational inference. What the deconfounder requires is that the fitted factor model provides an accurate approximation of the population distribution of $p(\boldsymbol{a})$.

In the next step, use the fitted factor model to calculate the conditional expectation of each individual's local factor weights $\hat{z}_i = \mathbb{E}_M [Z_i \mid \boldsymbol{A}_i = \boldsymbol{a}_i]$. We emphasize that this expectation is from the fitted model M (not the population distribution). We can use approximate expectations.

In the final step, condition on $\hat{z}_i$ as a substitute confounder and proceed with causal inference. For example, estimate $\mathbb{E}[Y_i(\boldsymbol{a})] = \mathbb{E}\left[\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid \hat{Z}_i, \boldsymbol{A}_i = \boldsymbol{a}\right]\right]$.

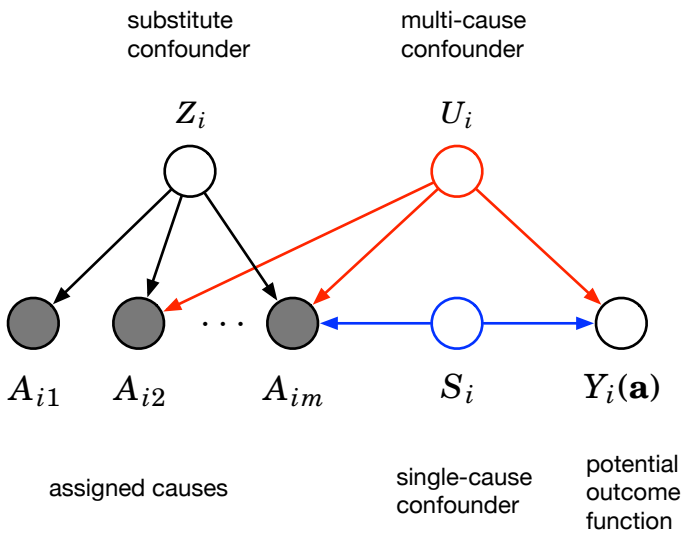


Figure 3.1: A graphical model argument for the deconfounder. The punchline is that if Z_i renders the A_{ij}'s conditionally independent then there cannot be a multi-cause confounder. The proof is by contradiction. Assume conditional independence holds, $p(a_{i1}, \ldots, a_{im} \mid z_i) = \prod_j p(a_{ij} \mid z_i)$; if there exists a multi-cause confounder U_i (red) then, by d-separation, conditional independence cannot hold (Pearl, 1988). Note we cannot rule out the single-cause confounder S_i (blue).

Why is this strategy sensible? Assume the fitted factor model captures the (unconditional) distribution of assigned causes $p(a_{i1}, \ldots, a_{im})$. This means that all causes are conditionally independent given the local latent factors,

$$p(a_{i1}, \ldots, a_{im} \mid z_i) = \prod_{j=1}^{m} p(a_{ij} \mid z_i). \tag{3.2}$$

Now make an additional assumption: there are no *single-cause confounders*, variables that affect just one of the assigned causes and the potential outcome function. (More precisely, we need to have observed all the single-cause confounders.) With this assumption, the independence statement of Equation (3.2) implies unconfoundedness, $\boldsymbol{A}_i \perp\!\!\!\perp Y_i(\boldsymbol{a}) \mid Z_i$, and unconfoundedness justifies causal inference. In summary, if the Bayesian factor model captures the distribution of assigned causes—a testable proposition—then we can use $\hat{z}_i$ as a variable that contains the (multi-cause) confounders.

The graphical model in Figure 3.1 justifies the deconfounder and reveals its assumptions.[†] Suppose we observe a Z_i such that the conditional independence in Equation (3.2) holds. Further suppose there exists an unobserved multi-cause confounder U_i (illustrated in red), which connects to multiple assigned causes and the outcome. If such a U_i exists then the causes would be dependent, even conditional on Z_i. (This fact comes from d-separation.) But such dependence leads to a contradiction, specifically that Equation (3.2) does not hold. Thus U_i cannot exist.

There is a nuance. The conditional independence in Equation (3.2) cannot rule out the existence of an unobserved single-cause confounder, denoted S_i in Figure 3.1. Even if such a confounder exists, conditional independence still holds.

Here is the punchline. If we find a Bayesian factor model that accurately represents the distribution of causes then that model can provide a variable that captures the unobserved multi-cause confounders. The reason is that the multi-cause confounders induce dependence among the causes; a good factor model provides a variable that renders the causes conditionally independent; thus that variable captures the confounders. This is the blessing of multiple causes.

3.2 The identification strategy of the deconfounder

How does the deconfounder identify potential outcomes? The classical strategy for causal identification is that unconfoundedness, together with SUTVA and overlap, identifies the potential outcomes (Hirano and Imbens, 2004; Imai and Van Dyk, 2004; Imbens, 2000). The deconfounder continues to assume SUTVA and overlap, but it weakens the unconfoundedness assumption.

Roughly, unconfoundedness requires that there are no unobserved confounders. To weaken this assumption, the deconfounder constructs a substitute confounder that captures multi-cause confounders. (The proof is in Section 4.1.) Uncovering multi-cause confounders from data weakens the unconfoundedness assumption to one of no unobserved *single-cause* confounders.

Thus the deconfounder relies on three main assumptions: (1) SUTVA (Rubin, 1980, 1990); (2) no unobserved single-cause confounders; (3) overlap (Imai and Van Dyk, 2004).

[†]Figure 3.1 uses a graphical model to represent and reason about conditional dependencies in the population distribution. It is not a causal graphical model or a structural equation model.

3.2.1 Stable unit treatment value assumption (SUTVA)

The SUTVA requires that the potential outcomes of one individual are independent of the assigned causes of another individual. It assumes that there is no interference between individuals and there is only a single version of each assigned cause. See Rubin (1980, 1990) and Imbens and Rubin (2015a) for discussion of this assumption.

3.2.2 No unobserved single-cause confounders

"No unobserved single-cause confounders" requires that we observe any confounders that affect only one of the causes; see Figure 3.1. (The precise technical definition is in Assumption 4.1.1 of Section 4.1.)

This assumption is weaker than the classical assumption of unconfoundedness, which requires "no unobserved confounders." That said, whether the assumption is plausible depends on the particulars of the problem. Note that "no unobserved single-cause confounders" reduces to the "no unobserved confounders" when there is only one cause; all confounders are single-cause in this case.

When might "no unobserved single-cause confounders" be plausible? Consider the movie-actor example. One possible confounder is the reputation of the director. Famous directors have access to a circle of capable actors; they also tend to make good movies with large revenues. If the dataset contains many actors, several are likely in the circle of capable actors; the director's reputation is a multi-cause confounder. (If only one actor in the dataset is capable then the director's reputation is a single-cause confounder.)

Or consider the GWAS problem. If a confounder affects single-nucleotide polymorphisms (SNPs)—and we observe 100,000 SNPs per individual—then the confounder may be unlikely to affect only one. The same reasoning can apply to other settings—medications in medical informatics data, neurons in neuroscience recordings, and vocabulary terms in text data.

By the same token, "no unobserved single-cause confounders" may not be satisfied when there are very few assigned causes. Consider the neuroscience problem of inferring the relationship between brain activity and animal behavior, but where the scientist only records the activity of a small number of neurons. While unlikely that a confounder affects only one neuron in the brain, it may be more possible that a confounder affects only one of the observed neurons. This would violate "no unobserved single-cause confounders."

In domains where "no unobserved single-cause confounders" is likely not satisfied, we suggest performing sensitivity analysis (Cinelli et al., 2019; Franks et al., 2019; Gilbert et al., 2003; Imai and Van Dyk, 2004; Robins et al., 2000b) on the deconfounder estimates. It assesses the robustness of the estimate against unobserved single-cause confounding. In the context of GWAS, Section 6.2 will illustrate the effect of violating "no unobserved single-cause confounders."

3.2.3 Overlap

The final main assumption of the deconfounder is that the substitute confounder Z_i satisfies the overlap condition,

$$p(A_{ij} \in \mathcal{A} \mid Z_i) > 0 \text{ for all sets } \mathcal{A} \text{ with positive measure, i.e. } p(\mathcal{A}) > 0. \tag{3.3}$$

Overlap asserts that, given the substitute confounder, the conditional probability of any vector of assigned causes is positive. This assumption is sometimes stated as the second half of ignorability (Imai and Van Dyk, 2004).[‡]

The potential outcome $Y_i(\boldsymbol{a})$ is not identifiable if the substitute confounder does not satisfy overlap. When the overlap is limited, i.e. $p(A_{ij} \in \mathcal{A} \mid Z_i)$ is small for all values of Z_i, then the deconfounder estimates of the potential outcome $Y_i(\boldsymbol{a})$ will have a high variance.

[‡]We also require the observed covariates X_i satisfy the overlap condition if they are single-cause confounders, i.e. $p(A_{ij} \in \mathcal{A} \mid X_i) > 0$ for all sets $\mathcal{A}$ with positive measure, i.e. $p(\mathcal{A}) > 0$.

For probabilistic factor models, the overlap condition is usually satisfied. For example, probabilistic PCA assumes $A_{ij} \mid Z_i \sim \mathcal{N}(Z_i^\top \theta_j, \sigma^2)$. The normal distribution has a support of the whole real line, which ensures $P(A_{ij} \in \mathcal{A} \mid Z_i) > 0$ for all $\mathcal{A}$ with positive measure. That said, as the dimensionality of Z_i increases, overlap often becomes increasingly limited (D'Amour et al., 2017). For example, probabilistic PCA returns increasingly small σ^2, which signals $P(A_{ij} \in \mathcal{A} \mid Z_i)$ is small.

We can enforce overlap by constraining the allowable family of factor models. With continuous causes, we restrict to models with continuous densities on $\mathbb{R}$. (We assume the causes are full-rank, i.e., that no two causes are measurable with each other; if such a pair exists, merge them into a single cause.) With discrete causes, we restrict to factor models with support on the whole $\mathcal{A}$ and a Z_i lower-dimensional than the causes.

Alternatively, we can merge highly correlated causes as a preprocessing step. For example, consider two causes that are always assigned the same value, e.g. two actors who either both appear in a movie or both not. We can merge them into one cause. This merging step prevents the deconfounder from extrapolating for the assigned causes which the data carries little evidence. We can also resort to classical strategies of causal inference under limited overlap, for example subsampling the population (Crump et al., 2009).

How can we assess the overlap with respect to the substitute confounder? With a fitted factor model, we can analyze the conditional distribution of the assigned causes given the substitute confounder $P(A_{ij} \mid Z_i)$ for all individual i's. A conditional with low variance or low entropy signals limited overlap and the possibility of high-variance causal estimates.

3.2.4 Other assumptions

We have described the main assumptions of the deconfounder. Note that point identification of causal quantities requires further assumptions; see Chapter 4 for a discussion of these additional assumptions. One of the key additional assumptions is pinpointability (Assumption 4.1.2). It requires that we can estimate the substitute confounder Z_i from the causes $\boldsymbol{A}_i$ with full certainty. Notice that an unobserved latent variable must affect an infinite number of causes to be pinpointable. It implies the unobserved multi-cause confounder needs to non-trivially affect an infinite number of causes to be picked up intact by the substitute confounder. We discuss this assumption further in Chapter 4.

3.2.5 The deconfounder is unbiased

With SUTVA, overlap, no unobserved single-cause confounders, and the additional technical assumptions, we use the deconfounder to estimate causal quantities.

The deconfounder (an informal version of Theorem 4.2.1). *Assume SUTVA and no unobserved single-cause confounders.Then the deconfounder provides an unbiased estimate of the average causal effect:*

$$\begin{aligned}
&\mathbb{E}_Y\left[Y_i(a_1, \ldots, a_m)\right] - \mathbb{E}_Y\left[Y_i(a'_1, \ldots, a'_m)\right] \\
&\quad = \mathbb{E}_{X,Z}\left[\mathbb{E}_Y\left[Y_i \mid A_{i1} = a_1, \ldots, A_{im} = a_m, X_i, Z_i\right]\right] \\
&\qquad - \mathbb{E}_{X,Z}\left[\mathbb{E}_Y\left[Y_i \mid A_{i1} = a'_1, \ldots, A_{im} = a'_m, X_i, Z_i\right]\right],
\end{aligned} \tag{3.4}$$

where Z_i denotes the substitute confounder constructed from the factor model.

The theorem relies on two properties of the substitute confounder: (1) it captures all multi-cause confounders; (2) it does not capture mediators. By its construction from probabilistic factor models, the substitute confounder captures all multi-cause confounders; again, see the graphical model argument in Figure 3.1. Moreover, the substitute confounder is constructed with only the observed causes; no outcome information is used, so it may not pick up any mediators. (We prove this fact in Lemma 9.2.1.) Thus, along with no unobserved single-cause confounders, the substitute confounder provides uncon-

foundedness. With unconfoundedness in hand, we treat the substitute confounder as if it were observed covariates. While this theorem does not require overlap, identifying other causal quantities with the deconfounder requires overlap. We discuss the identification of different causal quantities in Section 4.1 and lay out the assumptions required for each.

3.3 Practical details of the deconfounder

We next attend to some of the practical details of the deconfounder. The ingredients of the deconfounder are (1) a factor model of assigned causes, (2) a way to check that the factor model captures their population distribution, and (3) a way to estimate the conditional expectation $\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid \hat{Z}_i, \boldsymbol{A}_i = \boldsymbol{a}\right]$ for performing causal inference. We discuss each ingredient below (Section 3.3.1 and Section 3.3.2) and then describe the full deconfounder algorithm (Section 3.3.3).

3.3.1 Using the assignment model to infer a substitute confounder

The first ingredient is a factor model of the assigned causes, as defined in Equation (3.1), which we call the assignment model. Many models fall into this category, such as mixture models, mixed-membership models, and deep generative models. Each of these models can be written as Equation (3.1); they each involve a per-datapoint latent variable Z_i and a per-cause parameter θ_j. Fitting the factor model gives an estimate of the parameters θ_j, $j = 1, \ldots, m$. When the fitted factor model captures the population distribution of the assigned causes then inferences about Z_i can be used as substitute confounders in a downstream causal inference.

Example factor models The deconfounder requires that the investigator find an adequate factor model of the assigned causes and then use the factor model to estimate the posterior $p(z_i \mid \boldsymbol{a}_i)$. In the simulations and studies of Chapter 6, we will explore several classes of factor models; we describe some of them here.

One of the most common factor models is principal component analysis (PCA). PCA is appropriate when the assigned causes are real-valued. In its probabilistic form (Tipping and Bishop, 1999), both z_i and the per-cause parameters θ_j are real-valued K-vectors. The model is

$$Z_{ik} \sim \mathcal{N}(0, \lambda^2), \quad k = 1, \ldots, K, \tag{3.5}$$

$$A_{ij} \mid Z_i \sim \mathcal{N}\left(z_i^\top \theta_j, \sigma^2\right), \quad j = 1, \ldots, m. \tag{3.6}$$

We can fit probabilistic PCA with Bayesian methods and use standard conditional probability to calculate $p(z_i \mid \boldsymbol{a}_i)$. Exponential family extensions of PCA are also factor models (Collins et al., 2002; Mohamed et al., 2009) as are some deep generative models (Tran et al., 2017), which can be interpreted as a nonlinear probabilistic PCA.

When the causes are counts, Poisson factorization (PF) is an appropriate factor model (Cemgil, 2009; Gopalan et al., 2015; Schmidt et al., 2009). PF is a probabilistic form of nonnegative matrix factorization (Lee and Seung, 1999, 2001), where z_i and θ_j are positive K-vectors,

$$Z_{ik} \sim \text{Gamma}(\alpha_0, \alpha_1), \quad k = 1, \ldots, K, \tag{3.7}$$

$$A_{ij} \mid Z_i \sim \text{Poisson}(z_i^\top \theta_j), \quad j = 1, \ldots, m. \tag{3.8}$$

PF can be fit to large datasets with variational methods (Gopalan et al., 2015).

A final example of a factor model is the deep exponential family (DEF) (Ranganath et al., 2015). A DEF is a probabilistic deep neural network. It uses exponential families to generalize classical models like the sigmoid belief network (Neal, 1990) and deep Gaussian models (Rezende et al., 2014). For example, a two-layer DEF models each observation as

$$Z_{2,il} \sim \text{Exp-Fam}_2(\alpha), \quad l = 1, \ldots, L, \tag{3.9}$$

$$Z_{1,ik} \mid Z_{2,i} \sim \text{Exp-Fam}_1(g_1(z_{2,i}^\top \theta_{1,k})), \quad k = 1, \ldots, K, \tag{3.10}$$

$$A_{ij} \mid Z_{1,i} \sim \text{Exp-Fam}_0(g_0(z_{1,i}^\top \theta_{0,j})), \quad j = 1, \ldots, m. \tag{3.11}$$

Exp-Fam is an exponential family distribution, θ_* are parameters, and $g_*(\cdot)$ are link functions. Each layer of the DEF has the same functional form as a generalized linear model (McCullagh, 2018; McCullagh and Nelder, 1989). The DEF inherits the flexibility of deep neural networks, but uses exponential families to capture different types of layered representations and data. For example, if the assigned causes are counts then Expfam_0 can be Poisson; if they are reals then it can be Gaussian. Approximate inference in DEF can be performed with variational methods (Ranganath et al., 2014).

Predictive checks for the assignment model The deconfounder requires that its factor model captures the population distribution of the assigned causes. To assess the fidelity of the chosen model, we use predictive checks. A predictive check compares the observed assignments with assignments drawn from the model's predictive distribution. If the model is good, then there is little difference.

First hold out a subset of assigned causes for each individual $a_{i\ell}$, where ℓ indexes some held-out causes. The heldout assignments are written $\boldsymbol{a}_{i,\text{held}}$, and note we hold out randomly selected causes for each individual. The observed assignments are written $\boldsymbol{a}_{i,\text{obs}}$.

Next fit the factor model to the remaining assignment data $\mathcal{D} = \{\boldsymbol{a}_{i,\text{obs}}\}_{i=1}^n$. This results in a fitted assignment model $p(z, \theta \mid \boldsymbol{a})$. For each individual i, calculate the local posterior distribution of $p(z_i \mid \boldsymbol{a}_{i,\text{obs}})$.

Here is the predictive check. First sample held-out causes from their predictive distribution,

$$p(\boldsymbol{a}_{i,\text{held}}^{\text{rep}} \mid \boldsymbol{a}_{i,\text{obs}}) = \int p(\boldsymbol{a}_{i,\text{held}} \mid z_i)\, p(z_i \mid \boldsymbol{a}_{i,\text{obs}})\, \mathrm{d}z_i. \tag{3.12}$$

This distribution integrates out the local posterior $p(z_i \mid \boldsymbol{a}_{i,\text{obs}})$. (An approximate posterior also suffices; we discuss why in Section 3.4.)

Then compare replicated data to held-out data. We compare with expected log probability

$$t(\boldsymbol{a}_{i,\text{held}}) = \mathbb{E}_Z \left[\log p(\boldsymbol{a}_{i,\text{held}} \mid Z) \mid \boldsymbol{a}_{i,\text{obs}}\right], \tag{3.13}$$

which relates to the marginal log likelihood. In the nomenclature of posterior predictive checks, this is the "discrepancy function" that we use; one can use others.

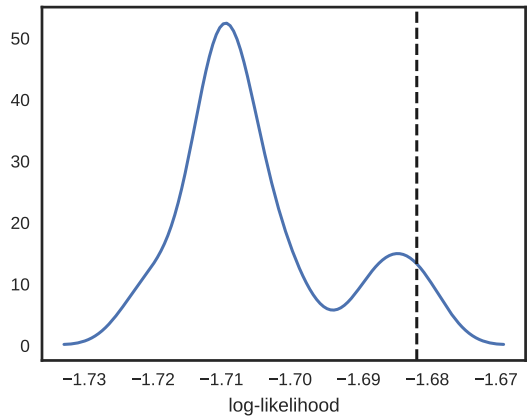


Figure 3.2: Predictive checks for the assignment model. The vertical dashed line shows $t(\boldsymbol{a}_{i,\text{held}})$. The blue curve shows the kernel density estimate (kde) of $t(\boldsymbol{a}^{\text{rep}}_{i,\text{held}})$. The predictive score is the area under the blue curve to the left of the vertical dashed line. The predictive score of this assignment model is larger than 0.1; we consider it satisfactory.

Finally calculate the predictive score,

$$\text{predictive score} = p\left(t(\boldsymbol{a}^{\text{rep}}_{i,\text{held}}) < t(\boldsymbol{a}_{i,\text{held}})\right). \tag{3.14}$$

Here the randomness stems from $\boldsymbol{a}^{\text{rep}}_{i,\text{held}}$ coming from the predictive distribution in Equation (3.12), and we approximate the predictive score with Monte Carlo.

How to interpret the predictive score? A good model will produce values of the held-out causes that give similar log-likelihoods to their real values—the predictive score will not be extreme. A mismatched model will produce an extremely small predictive score, often where the replicated data has much higher log-likelihood than the real data. An ideal predictive score is around 0.5. We consider predictive scores with predictive scores larger than 0.1 to be satisfactory; we do not have enough evidence to conclude a significant mismatch of the assignment model. Note that the threshold of 0.1 is a subjective design choice. We find such assignment models that pass this threshold often yield satisfactory causal estimates in practice. Figure 3.2 illustrates a predictive check of a good assignment model. Chapter 6 shows predictive checks in action.

These predictive checks blend ideas around posterior predictive checks (PPCs) (Rubin, 1984), PPCs with realized discrepancies (Gelman et al., 1996), PPCs with held-out data (Gelfand et al., 1992; Ranganath and Blei, 2019), and stage-wise checking of hierarchical models (Bayarri and Castellanos, 2007; Dey et al., 1998). They also relate to Bayesian causal model criticism (Tran et al., 2016b). More broadly, the process of iterative model building—-cycling between finding, fitting, and checking a model of the assignments—relates to the applied practice of Bayesian data analysis (Blei, 2014; Gelman et al., 2013).

3.3.2 The outcome model

We described how to fit and check a factor model of multiple assigned causes. We now fold in the observed outcomes and use the fitted factor model to correct for unobserved confounders.

Suppose $p(z_i \mid \boldsymbol{a}_i, \mathcal{D})$ concentrates around a point $\hat{z}_i$. Then we can use $\hat{z}_i$ as a confounder. Follow Section 2.3 to calculate the iterated expectation on the left side of Equation (2.1). However, replace the observed confounders with the substitute confounder; the goal is to calculate $\mathbb{E}\left[\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i\right]\right]$. First, approximate the outside expectation with Monte Carlo,

$$\mathbb{E}\left[\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i\right]\right] \approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_Y\left[Y_i(\boldsymbol{A}_i) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = \hat{z}_i\right]. \tag{3.15}$$

This approximation uses the substitute confounder $\hat{z}_i$, integrating over its population distribution. It uses the model to infer the substitute confounder from each data point and then integrates the distribution of that inferred variable induced by the population distribution of data.

Turn now to the inner expectation of Equation (3.15). We fit a function to estimate this quantity,

$$\mathbb{E}\left[Y_i(\boldsymbol{A}_i) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = z\right] = f(\boldsymbol{a}, z). \tag{3.16}$$

The function $f(\boldsymbol{a}, z)$ is called the *outcome model* and can be fit from the augmented observed data $\{\boldsymbol{a}_i, \hat{z}_i, y_i(\boldsymbol{a}_i)\}$. For example, we can minimize their discrepancy via some loss function ℓ:

$$\hat{f} = \arg\min_f \sum_{i=1}^{n} \ell(y_i(\boldsymbol{a}_i) - f(\boldsymbol{a}_i, \hat{z}_i)).$$

Like the factor model, we can check the outcome model—it is fit to observed data and should be predictive of held-out observed data (Tran et al., 2016b).

One outcome model we consider is a simple linear function,

$$f(\boldsymbol{a}, z) = \beta^\top \boldsymbol{a} + \gamma^\top z + \beta_0. \tag{3.17}$$

Another outcome model we consider is where $f(\cdot)$ is linear in the assigned causes $\boldsymbol{a}$ and the "reconstructed assigned causes" $\hat{\boldsymbol{a}}(z) = \mathbb{E}_M[\boldsymbol{A} \mid z]$, an expectation from the fitted factor model. This class of functions is

$$f(\boldsymbol{a}, z) = \beta^\top (\boldsymbol{a} - \hat{\boldsymbol{a}}(z)) + \beta_0. \tag{3.18}$$

This model relates to the generalized propensity score (Hirano and Imbens, 2004; Imbens, 2000), where it uses $\hat{a}(z)$ as a proxy for the propensity score. Note this substitution is used in Bayesian statistics (Geisser et al., 1990; Laird and Louis, 1982; Tierney and Kadane, 1986), and is justified when higher moments of the assignment are similar across individuals. In both Equations (3.17) and (3.18), the coefficient β represents the average causal effect of raising each cause by one unit.

If only a subset of the causes is of interest, then we fit the outcome model to the subset only. For example, suppose we are only interested in the first $k < m$ causes. Then we estimate

$$\mathbb{E}[Y_i(\boldsymbol{A}_{i,1:k}) \mid \boldsymbol{A}_{i,1:k} = \boldsymbol{a}_{1:k}, Z_i = z] = f(\boldsymbol{a}_{1:k}, z), \tag{3.19}$$

and consider outcome models like

$$f(\boldsymbol{a}_{1:k}, z) = \beta_{1:k}^\top \boldsymbol{a}_{1:k} + \gamma^\top z + \beta_0, \tag{3.20}$$

and

$$f(\boldsymbol{a}_{1:k}, z) = \beta_{1:k}^\top (\boldsymbol{a}_{1:k} - \hat{\boldsymbol{a}}(z)_{1:k}) + \beta_0. \tag{3.21}$$

Note we are not restricted to linear models. Other outcome models like random forests (Wager and Athey, 2017) and Bayesian additive regression trees (Hill, 2011) all apply here. Moreover, devising an outcome model is just one approach to approximating the inner expectation of Equation (3.15). Another approach is again to use Monte Carlo. There are several possibilities. In one, group the confounder $\hat{z}_i$ into bins and approximate the expectation within each bin. In another, bin by the propensity score $p(a_i \mid \hat{z}_i)$ and approximate the inner expectation within each propensity-score bin (Lunceford and Davidian, 2004; Rosenbaum and Rubin, 1983). A third possibility—if the assigned causes are discrete and the number of causes is small—is to use the propensity score with inverse propensity weighting (Dehejia and Wahba, 2002; Heckman et al., 1998; Horvitz and Thompson, 1952; Rosenbaum and Rubin, 1983).

3.3.3 The full algorithm, and an example

We described each component of the deconfounder. Algorithm 8.1 gives the full algorithm, a procedure for estimating Equation (3.15). The steps are: (1) find, fit, and check a factor model to the dataset of assigned causes; (2) estimate $\hat{z}_i$ for each data point; (3) find and fit an outcome model; (4) use the outcome model and estimated $\hat{z}_i$ to do causal inference.

Example Consider a causal inference problem in genome-wide association studies (GWAS) (Stephens and Balding, 2009; Visscher et al., 2017): how do human genes causally affect height? Here we give a brief account of how to use the deconfounder, omitting many of the details. We analyze GWAS problems extensively in Section 6.2. We discuss the connections of the deconfounder to existing GWAS methods in Section 3.5.

Consider a dataset of $n = 5,000$ individuals; for each individual, we measure height and genotype, specifically the alleles at $m = 100,000$ locations, called the single-nucleotide polymorphisms (SNPs). Each SNP is represented by a count of 0, 1, or 2; it encodes how many of the individual's two nucleotides differ from the most common pair of nucleotides at the location. Table 3.1 illustrates a snippet of the data (10 individuals).

Algorithm 3.1: The deconfounder

Input: a dataset of assigned causes and outcomes $\{(\boldsymbol{a}_i, y_i)\},\ i = 1, \ldots, n$
Output: the average potential outcome $\mathbb{E}\left[Y(\boldsymbol{a})\right]$ for any causes $\boldsymbol{a}$
repeat
  choose an assignment model from the class in Equation (3.1)
  fit the model to the assigned causes $\{\boldsymbol{a}_i\},\ i = 1, \ldots, n$
  check the fitted model $\hat{M}$
until *the assignment check is satisfactory*
foreach *datapoint* i **do**
  calculate $\hat{z}_i = \mathbb{E}_{\hat{M}}\left[Z_i \mid \boldsymbol{a}_i\right]$.
end
repeat
  choose an outcome model from Equation (3.16)
  fit the outcome model to the augmented dataset $\{(\boldsymbol{a}_i, y_i, \hat{z}_i)\},\ i = 1, \ldots, n$
  check the fitted outcome model
until *the outcome check is satisfactory*
estimate the average causal effect $\mathbb{E}\left[Y(\boldsymbol{a})\right] - \mathbb{E}\left[Y(\boldsymbol{a}')\right]$ by Equation (3.15)

We simulate such a dataset of genotypes and height. We generate each individual's genotypes by simulating heterogeneous mixing of populations (Pritchard et al., 2000b). We then generate the height from a linear model of the SNPs (i.e. the assigned causes) and some simulated confounders. (The confounders are only used to simulate data; when running the deconfounder, the confounders are unobserved.) In this simulated data, the coefficients of the SNPs are the true causal effects; we denote them $\beta^* = (\beta_1^*, \ldots, \beta_m^*)$. See Section 6.2 for more details of the simulation.

ID (i)	SNP1 ($a_{i,1}$)	SNP2 ($a_{i,2}$)	SNP3 ($a_{i,3}$)	SNP4 ($a_{i,4}$)	SNP5 ($a_{i,5}$)	SNP6 ($a_{i,6}$)	SNP7 ($a_{i,7}$)	SNP8 ($a_{i,8}$)	SNP9 ($a_{i,9}$)	$\cdots$	SNP100K ($a_{i,100K}$)	Height (feet) (y_i)
1	1	0	0	1	0	0	1	2	0	$\cdots$	0	5.73
2	1	2	2	1	2	1	1	0	1	$\cdots$	2	5.26
3	2	0	1	1	0	1	0	1	1	$\cdots$	2	6.24
4	0	0	0	1	1	0	1	2	0	$\cdots$	0	5.78
5	1	2	1	1	1	0	1	0	0	$\cdots$	1	5.09
$\vdots$						$\vdots$						$\vdots$

Table 3.1: How do SNPs causally affect height? This table shows a portion of a dataset: simulated SNPs as the multiple causes and height as the outcome.

The goal is to infer how the SNPs causally affect human height, even in the presence of unobserved confounders. The m-dimensional SNP vector $\boldsymbol{a}_i = (a_{i1}, a_{i2}, ..., a_{im})$ is the vector of assigned causes for individual i; the height y_i is the outcome. We want to estimate the potential outcome: what would the (average) height be if we set a person's SNP to be $\boldsymbol{a} = (a_1, a_2, ..., a_m)$? Mathematically, this is the average potential outcome function: $\mathbb{E}[Y_i(\boldsymbol{a})]$, where the vector of assigned causes $\boldsymbol{a}$ takes values in $\{0, 1, 2\}^m$.

We apply the deconfounder: model the assigned causes, infer a substitute confounder, and perform causal inference. To infer a substitute confounder, we fit a factor model of the assigned causes. Here we fit a 50-factor PF model, as in Equation (3.7). This fit results in estimates of non-negative factors $\hat{\theta}_j$ for each assigned cause and non-negative weights $\hat{z}_i$ for each individual (both K-vectors).

If the predictive check greenlights this fit, then we take the posterior predictive mean of the assigned causes as the reconstructed assignments, $\hat{a}_j(z_i) = \hat{z}_i^\top \hat{\theta}_j$. For brevity, we do not report the predictive check here. (The model passes.) We demonstrate predictive checks for GWAS in the empirical studies of Section 6.2.

Using the reconstructed assigned causes, we estimate the average potential outcome function. Here we fit a linear outcome model to the height y_i against both of the assigned causes $\boldsymbol{a}_i$ and reconstructed assignment $\hat{\boldsymbol{a}}(z_i)$,

$$y_i \sim \mathcal{N}\left(\beta_0 + \beta^\top(\boldsymbol{a}_i - \hat{\boldsymbol{a}}(z_i)), \sigma^2\right). \tag{3.22}$$

This regression is high dimensional ($m > n$); for regularization, we use an L_2-penalty on β (equivalently, normal priors). Fitting the outcome model gives an estimate of regression coefficients $\{\hat{\beta}_0, \hat{\beta}\}$. Because we use a linear outcome model, the regression coefficients $\hat{\beta}$ estimate the true causal effect β^*.

We evaluate the causal estimates obtained with and without the deconfounder. We focus on the root mean squred error (RMSE) of $\hat{\beta}$ to β^*. ("Causal estimation without the deconfounder" means fitting a linear model of the height y_i against the assigned causes $\boldsymbol{a}_i$.) TThe RMSE is 49.6×10^{-2} without the deconfounder and 41.6×10^{-2} with the deconfounder. The deconfounder produces closer-to-truth causal estimates.

3.4 A conversation with the reader

In this section, we answer some questions a reader might have.

Why do I need multiple causes? The deconfounder uses latent variables to capture dependence among the assigned causes. The theory in Section 4.1 says that a latent variable that captures this dependence will contain all valid multi-cause confounders. But estimating this latent variable requires evidence for the dependence, and evidence for dependence cannot exist with just one assigned cause. Thus the deconfounder requires multiple causes.

Is the deconfounder free lunch? The deconfounder is not free lunch—it trades confounding bias for estimation variance. To see this, take an information point of view: the deconfounder uses a portion of information in the data to estimate a substitute confounder; then it uses the rest to estimate causal effects. By contrast, classical causal inference uses all the information to estimate causal effects, but it must assume unconfoundedness.

Suppose unconfoundedness is satisfied, i.e. no unobserved confounders. Then both classical causal inference and the deconfounder provide unbiased causal estimates, though the deconfounder will be less confident; it has higher variance. Now suppose only "no unobserved single-cause confounders" is satisfied. The deconfounder still provides unbiased causal estimates, but the classical causal inference is biased.

Why does the deconfounder have two stages? Algorithm 8.1 first fits a factor model to the assigned causes and then fits the potential outcome function. This is a two-stage procedure. Why? Can we fit these two models jointly?

One reason is convenience. Good models of assigned causes may be known in the research literature, such as for genetic studies. Moreover, separately fitting the assignment model allows the investigator to fit models to any available data of assigned causes, including datasets where the outcome is not measured.

Another reason for the two stages is to ensure that Z_i does not contain mediators, variables along the causal path between the assigned causes and the outcome. Intuitively, excluding the outcome ensures that the substitute confounders are "pre-treatment" variables; we cannot identify a mediator by looking only at the assigned causes. More formally, excluding the outcome ensures that the model satisfies $p(z_i \mid \boldsymbol{a}_i, y_i(\boldsymbol{a}_i)) = p(z_i \mid \boldsymbol{a}_i)$; this equality cannot hold if Z_i contains a mediator.

How does the deconfounder relate to the generalized propensity score? What about instrumental variables? The deconfounder relates to both. The deconfounder can be interpreted as a generalized propensity score approach, except where the propensity score model involves latent variables. If we treat the substitute confounder Z_i as observed covariates, then the factor model $P(A_i \mid Z_i)$ is precisely the propensity score of the causes A_i. With this view, the innovation of the deconfounder is in Z_i being latent. Moreover, it is the multiplicity of the causes $A_{i1}, \ldots, A_{im}$ that makes a latent Z_i feasible; we can construct Z_i by finding a random variable that renders all the causes conditionally independent.

The deconfounder can also be interpreted as a way of constructing instruments using latent factor models. Think of a factor model of the causes with linearly separable noises: $A_{ij} \overset{a.s.}{=} f(Z_i) + \epsilon_{ij}$. Given the substitute confounder, consider the residual of the causes ϵ_{ij}. For example, with probabilistic PCA the residual is $\epsilon_{ij} = A_{ij} - Z_i^\top \theta_j \sim \mathcal{N}(0, \sigma^2)$.

Assuming no unobserved single-cause confounders, the variable ϵ_{ij} is an instrumental variable for the jth cause A_{ij}: (1) The residual ϵ_{ij} correlates with the cause A_{ij}. (2) The residual ϵ_{ij} affects the outcome only through the cause A_{ij}; this fact is true because the substitute confounder Z_i is constructed without using any outcome information. (3) The residual ϵ_{ij} cannot be correlated with a confounder; this is true because $Z_i \perp \epsilon_{ij}$ by construction from the factor model, where $P(Z_i)$ and $P(A_{ij} \mid Z_i)$ are specified separately.

However, the deconfounder differs from the classical instrumental variables approach because it uses latent variable models to construct instruments, rather than requiring that instruments be observed. The latent variable construction is feasible because the multiplicity of the causes allows us to construct Z_i and ϵ_{ij} from the conditional independence requirement.

Does the factor model of the assigned causes need to be the true assignment model? Which factor model should I choose if multiple factor models return good predictive scores? Finding a good factor model is not the same as finding the "true" model of the assigned causes. We do not assume the inferred variable Z_i reflects a real-world unobserved variable.

Rather, the deconfounder requires the factor model to capture the population distribution of the assigned causes and, more particularly, their dependence structure. This requirement is why predictive checking is important. If the deconfounder captures the population distribution—if the predictive check returns high predictive scores—then we can use the inferred local variables Z_i as substitute confounders.

Moreover, the deconfounder can rely on approximate inference methods to infer the substitute confounder. The predictive check evaluates whether Z_i provides a good predictive distribution, regardless of how it was inferred. Given the assumptions of the deconfounder, as long as the model and (approximate) inference method together give a good predictive distribution—one close to the population distribution of the assigned causes—then the downstream causal inference is valid. We use approximate inference for most of the factor models we study in Chapter 6.

Suppose multiple factor models give similarly good predictive scores in the predictive check. In this case, we recommend choosing the factor model with the lowest capacity. Factor models with similar predictive scores often result in causal estimates with similarly little bias. But the variance of these estimates can differ. Factor models with high capacity can compromise overlap and lead to high-variance estimates; factor models with low capacities tend to produce lower variance causal estimates. The empirical study in Section 6.1 demonstrates this phenomenon.

Should I condition on known confounders and covariates? Suppose we also observe known confounders and other covariates X_i. The deconfounder maintains its theoretical properties when we condition on observed covariates X_i as well as infer a substitute confounder Z_i. In particular, if X_i is "pre-treatment"—it does not include any mediators—then the causal estimate will be unbiased (Imai and Van Dyk, 2004) (also see Theorem 4.2.1 below). Moreover, to satisfy no unobserved single-cause confounders (Section 3.2.2), we must condition on single-cause confounders. In practice, if we observe covariates that we are confident are confounders, we should condition on them.

How can I assess the uncertainty of the deconfounder? The uncertainty in the deconfounder comes from two sources, the factor model and the outcome model. The deconfounder first fits (and checks) the factor model; it gives a substitute confounder $Z_i \sim p(z_i \mid \boldsymbol{a}_i)$. It then uses the mean of the substitute confounder $\hat{z}_i = \mathbb{E}_{\hat{M}}[Z_i \mid \boldsymbol{a}_i]$ to fit an outcome model $p(y_i \mid \boldsymbol{a}_i, \hat{z}_i)$ and compute the potential outcome estimate $\mathbb{E}[Y_i(\boldsymbol{a})]$.

To assess the uncertainty of the deconfounder, we consider the uncertainty from both sources. We first draw s samples $\{z_i^{(1)}, \ldots, z_i^{(s)}\}$ of the substitute confounder: $z_i^{(\ell)} \overset{iid}{\sim} p(z_i \mid \boldsymbol{a}_i), \ell = 1, \ldots, s$. For each sample $z_i^{(\ell)}$, we fit an outcome model and compute a point estimate of the potential outcome. (If the outcome model is probabilistic, we compute the posterior distribution of its parameters; this leads to a posterior of the potential outcome.) We aggregate the estimates of the potential outcome (or its distributions) from the s samples $\{z_i^{(1)}, \ldots, z_i^{(s)}\}$; the aggregated estimate is a collection of point estimates of the potential outcome (or a mixture of its posterior distributions).

The variance of this aggregated estimate describes the uncertainty of the deconfounder; it reflects how the finite data informs the estimation of the potential outcome. Section 2.6 of Gustafson (2015) shows that this procedure produces credible intervals that achieve a frequentist coverage averaging out to the nominal level; the averaging is over the prior distribution of the parameter space. This procedure is applicable no matter whether the causal quantities of interest are identifiable or only partially identifiable. In a two-cause smoking study, Section 6.1 illustrates this strategy for calculating the uncertainty of the deconfounder.

Can the causes be causally dependent among themselves? The short answer is no for the average (total) causal effects of the causes. With additional assumptions, the deconfounder may still be able to identify the direct causal effect of each cause. Finally, a valid substitute confounder may not exist in this setting. For example, consider a cause A_1 that causally affects A_2 according to $A_1 \sim \mathcal{N}(0, 1), A_2 = A_1 + \epsilon, \epsilon \sim \mathcal{N}(0, 1)$. In this case, a substitute confounder Z must satisfy $Z \overset{a.s.}{=} A_1$ or $Z \overset{a.s.}{=} A_2$, because it needs to render the two causes conditionally independent. But such a Z does not satisfy overlap.

3.5 Connections to genome-wide association studies (GWAS)

Many methods from the research literature, especially around genome-wide association studies (GWAS), can be reinterpreted as instances of the deconfounder algorithm. Each can be seen as positing a factor model of assigned causes (Section 3.3.1) and a conditional outcome model (Section 3.3.2).

The deconfounder justifies each of these methods as forms of multiple causal inference and, though predictive checks, points to how a researcher can usefully compare and assess them. Most of these methods were motivated by imagining true unobserved confounding structure. However, the theory around the deconfounder shows that a well-fitted factor model will capture confounders independent of a researcher imagining what they may be; see the question in Section 3.4.

Below we describe many methods from the GWAS literature and show how they can be viewed as deconfounder algorithms. The GWAS problem is described in Section 3.3.3.

Linear mixed models The linear mixed model (LMM) is one the most popular classes of methods for analyzing GWAS (Darnell et al., 2017; Kang et al., 2008; Lippert et al., 2011; Loh et al., 2015; Yang et al., 2014; Yu et al., 2006). Seen through the lens of the deconfounder, an LMM posits a linear outcome model that depends on both the SNPs and a scalar latent factor Z_i.

In the LMM literature, Z_i is not explicitly drawn from a factor model; rather, $Z_{1:n}$ are from a multivariate Gaussian whose covariance matrix, called the "kinship matrix," is calculated from the observed SNPs $\boldsymbol{a}_{1:n}$. However, this is mathematically equivalent to posterior latent factors from a one-dimensional PCA model. Subject to its capturing the distribution of SNPs, the LMM is performing a multiple causal inference with the deconfounder.

Principal component analysis A related approach is to first perform (multi-dimensional) PCA on the SNP matrix and then to estimate an outcome model from the corresponding residuals (Price et al., 2006). This too is an instance of the deconfounder. As a factor model, PCA is described in Equation (3.6). Fitting an outcome model to its residuals is equivalent to conditioning on the reconstructed assignments, Equation (3.18).

Logistic factor analysis Closely related to PCA is logistic factor analysis (LFA) (Hao et al., 2015; Song et al., 2015). LFA can be seen as the following factor model,

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, I) \\ \pi_{ij} \mid Z_i &\sim \mathcal{N}(z_i^\top \theta_j, \sigma^2), \quad j = 1, \ldots, m, \\ A_{ij} \mid \pi_{ij} &\sim \text{Binomial}(2, \text{logit}^{-1}(\pi_{ij})), \quad j = 1, \ldots, m. \end{aligned}$$

If it captures the SNP matrix well, then Z_i can be viewed as a substitute confounder.

With LFA in hand, Song et al. (2015) use inverse regression to perform association tests. Their approach is equivalent to assuming an outcome model conditional on the reconstructed assignments $a(\hat{z}_i)$, again Equation (3.18), and subsequently testing for non-zero coefficients.

In a variant of LFA, Tran and Blei (2017) use a neural-network-based model of the unobserved confounder, connecting this model to causal inference with a nonparametric structural equation model (Pearl, 2009). They take an explicitly causal view of the testing problem.

Mixed-membership models Finally, many statistical geneticists use mixed-membership models (Airoldi et al., 2014) to capture the latent population structure of SNPs, and then condition on that structure in downstream analyses (Falush et al., 2003, 2007; Pritchard et al., 2000a,b). In genetics, a mixed-membership model is a factor model that captures latent ancestral populations. The latent variable Z_i is on the $K - 1$ simplex; it represents how much individual i reflects each ancestral population. The observed SNP A_{ij} comes from a mixture of Binomials, where Z_i determines its mixture proportions.

Using these models, researchers use a linear outcome model conditional on z_i and devise tests for significant associations (Pritchard et al., 2000b; Song et al., 2015; Tran and Blei, 2017). The deconfounder justifies this practice from a causal perspective, and underlines the importance of finding a model of population structure that captures the per-individual distribution of SNPs.

3.6 Related work

The deconfounder relates to several threads of research in causal inference.

Probabilistic modeling for causal inference. Several lines of work use probabilistic modeling to aid causal inference. Mooij et al. (2010) use Gaussian processes to depict causal mechanisms; Zhang and Hyvärinen (2009) study post-nonlinear causal models and their identifiability; Mckeigue et al. (2010) build on sparse methods to infer causal structures; Moghaddass et al. (2016) use Bayesian factor models to generalize the self-controlled case series method to multiple causes and multiple outcomes. Louizos et al. (2017) use variational autoencoders to infer unobserved confounders from proxy variables, Shah and Meinshausen (2018) develop projection-based techniques for high-dimensional covariance estimation under latent confounding, Frot et al. (2019) use linear factor models for robust causal structure learning with hidden variables, and Kaltenpoth and Vreeken (2019) leverage information theory principles to differentiate causal and confounded connections.

With a related goal, Tran and Blei (2017) build implicit causal models. Like the GWAS example in Section 6.2, they take an explicit causal view of GWAS, treating the SNPs as the multiple causes. They connect implicit probabilistic models and nonparametric structural equation models for causal inference (Pearl, 2009), and develop inference algorithms for capturing shared confounding. Heckerman (2018) studies the same scenario with linear regression, where observing many causes makes it possible to account for shared confounders. Multiple causal inference and latent confounding was also formalized by Ranganath and Perotte (2018), who take an information-theoretic approach.

Most of these papers use Pearl's framework (Pearl, 2009); they hypothesize a causal graph with confounders, causes, and outcomes. This chapter complements these works. We develop the deconfounder in the potential outcomes framework (Imbens and Rubin, 2015a; Rubin, 1974, 2005).

Analyzing genome-wide association studies (GWAS). In GWAS, the latent population structure is an important unobserved confounder. Pritchard et al. (2000b) propose a probabilistic admixture model for unsupervised ancestry inference. Price et al. (2006) and Astle et al. (2009) estimate the unobserved population structure using the principal components of the genotype matrix. Yu et al. (2006) and Kang et al. (2010) estimate the population structure via the "kinship matrix" on the genotypes. Song et al. (2015) and Hao et al. (2015) rely on factor analysis and admixture models to estimate the population structure.

GTEx Consortium et al. (2017) adopt a similar idea to study the effect of genetic variations on gene expression levels. These methods can be seen as variants of the deconfounder (see Section 3.5). The deconfounder gives them a rigorous causal justification, provides principled ways to compare them, and suggests an array of new approaches. We study GWAS data in Section 6.2.

Assessing the unconfoundedness assumption. Rosenbaum and Rubin (1983) demonstrate that unconfoundedness, together with a good propensity score model, is sufficient to perform causal inference with observational data. Many subsequent efforts assess the plausibility of unconfoundedness. For example, Gilbert et al. (2003); Imai and Van Dyk (2004); Robins et al. (2000b) develop sensitivity analysis in various contexts, though focusing on data with a single cause. In contrast, this work uses predictive model checks to assess unconfoundedness with multiple causes. More recently, Sharma et al. (2016) leverage auxiliary outcome data to test for confounding; Janzing and Schölkopf (2018a,b); Liu and Chan (2018) develop tests for non-confounding in multivariate linear regression; Cinelli et al. (2019) develop sensitivity analysis for linear causal models; Franks et al. (2019) design flexible sensitivity analysis for causal inference with one binary treatment. Here we work without auxiliary data, focus on causal estimation, as opposed to testing, and move beyond linear models and one treatment.

The (generalized) propensity score. Lee et al. (2010); McCaffrey et al. (2004); Schneeweiss et al. (2009) and many others develop and evaluate different models for assigned causes. In particular, Chernozhukov et al. (2017a) introduce a semiparametric assignment model; they propose a principled way of correcting for the bias that arises when regularizing or overfitting the assignment model. The work in this chapter introduces latent variables into the assignment model. The multiplicity of causes enables us to infer these latent variables and then use them as substitutes for unobserved confounders.

Traditional causal inference with multiple treatments. Feng et al. (2012); Lechner (2001); Lopez et al. (2017); McCaffrey et al. (2013); Rassen et al. (2011); Zanutto et al. (2005) extend matching, subclassification, and weighting to multiple treatments, always assuming no unobserved confounders. This work relaxes that assumption to no unobserved single-cause confounders.

4

Causal identification of the deconfounder

In this chapter, we develop theoretical results around the deconfounder.* We first justify the use of factor models by connecting them to the unconfoundedness assumption. We show that factor models, together with no unobserved single-cause confounders and a so-called pinpointability assumption, imply unconfoundedness. We then discuss identification results around the deconfounder. Under SUTVA, no unobserved single-cause confounders, and pinpointability, we prove that the deconfounder identifies the average causal effects and the conditional potential outcomes under different conditions.

4.1 The unconfoundedness of the substitute confounder

Begin by defining a multi-cause separator, a type of random variable.

*This chapter is in part based on Wang and Blei (2019a, 2020).

Definition 4.1.1: Multi-cause separator. Consider all the causes $\boldsymbol{A}_i = \{A_{i1}, \ldots, A_{im}\}$. A *multi-cause separator* U_i is a smallest σ-algebra that renders all the causes conditionally independent,

$$\mathrm{P}(A_{i1}, \ldots, A_{im} \mid U_i) = \prod_{j=1}^{m} \mathrm{P}(A_{ij} \mid U_i), \tag{4.1}$$

and where none of the conditionals $\mathrm{P}(A_j \mid U_i)$ is a point mass.

The concept of the smallest σ-algebra defines the sense in which the variable U_i is "multi-cause." If U_i contains information about a single cause then it is not the smallest separating σ-algebra. We explain the reasons below.

Suppose the variable U_i contains a single cause component and a multi-cause component, $U_i = (U_i^s, U_i^*)$. Without loss of generality, suppose the single-cause component only depends on the first cause A_1. Then Equation (8.4) implies

$$\mathrm{P}(A_{i1}, \ldots, A_{im} \mid U_i) = \prod_{j=1}^{m} \mathrm{P}(A_{ij} \mid U_i) \tag{4.2}$$

$$= \mathrm{P}(A_{i1} \mid U_i^s, U_i^*) \cdot \prod_{j=2}^{m} \mathrm{P}(A_{ij} \mid U_i^*). \tag{4.3}$$

This implies that

$$\mathrm{P}(A_{i1}, \ldots, A_{im} \mid U_i^*) = \int \mathrm{P}(A_{i1}, \ldots, A_{im} \mid U) \cdot \mathrm{P}(U_i^s \mid U_i^*) \, \mathrm{d}U_i^s \tag{4.4}$$

$$= \prod_{j=2}^{m} \mathrm{P}(A_{ij} \mid U_i^*) \cdot \int \mathrm{P}(A_{i1} \mid U_i^*, U_i^s) \cdot \mathrm{P}(U_i^s \mid U_i^*) \, \mathrm{d}U_i^s \tag{4.5}$$

$$= \mathrm{P}(A_{i1} \mid U_i^*) \cdot \prod_{j=2}^{m} \mathrm{P}(A_{ij} \mid U_i^*), \tag{4.6}$$

which means that $U = (U_i^s, U_i^*)$ is not the smallest σ-algebra that renders the causes independent.

Next, we state the "no unobserved single-cause confounders" assumption.

Assumption 4.1.1: No unobserved single-cause confounders. There exists a random variable U_i that satisfies the following two requirements:

1. It is a multi-cause separator.

2. Together with the observed covariates X_i, it satisfies weak unconfoundedness,[†]

$$A_{i1}, \ldots, A_{im} \perp\!\!\!\perp Y_i(\boldsymbol{a}) \mid U_i, X_i \qquad \forall \mathrm{a} \in \mathcal{A}. \tag{4.7}$$

The first part ensures that U_i only involves multiple causes. The second part ensures that the variable U_i contains all multi-cause confounders. (It can contain other multi-cause variables as well.)

Why is Assumption 4.1.1 called "no unobserved single-cause confounders"? The variable U_i is a multi-cause separator: it cannot capture single-cause variables; it must capture ancestors of multiple causes, and it cannot capture descendants of multiple causes. For U_i and the observed covariates X_i to satisfy weak unconfoundedness, the observed covariates must include all single-cause confounders.

The next assumption is that every multi-cause separator is pinpointed by a single function of the observed causes. It is called a substitute confounder, though only after the theorem below will it deserve this name.

Assumption 4.1.2: Pinpointability. All multi-cause separators Z_i are pinpointed[‡] by a single deterministic function of the causes,

$$\mathrm{P}(Z \mid A_{i1}, \ldots, A_{im}) = \delta_{f(A_{i1},\ldots,A_{im})}, \tag{4.8}$$

where $\delta_{f(\cdot)}$ denotes a point mass at $f(\cdot)$.

Pinpointability requires that we can estimate the substitute confounder Z_i from the causes $\boldsymbol{A}_i$ with certainty; it is a deterministic function of the causes. Together with Proposition 4.1.1, pinpointability implies that the true unobserved multi-cause confounders are also deterministic functions of the causes. Nevertheless, the substitute confounder need not coincide with the true data-generating Z_i; nor does

[†]This assumption was suggested by Imai and Jiang at JSM 2019 (Wang and Blei, 2019c). Here X_i is a set of observed covariates and $Y_i(\boldsymbol{a})$ are potential outcomes. Definition 4 of Wang and Blei (2019a) says there exists a smallest-σ-algebra variable that renders the causes independent (Equation (40) of Wang and Blei (2019a)), and it satisfies weak unconfoundedness. For weak unconfoundedness, fix Equation (39) of Wang and Blei (2019a) by adding $A_{i,-j}$ to its conditioning.

[‡]This assumption is the same as the "consistency of substitute confounder" assumption in Wang and Blei (2019a).

it need to coincide with the true unobserved confounder. We only need to estimate the substitute confounder Z_i up to some deterministic bijective transformations (e.g. scaling and linear transformations). That said, we note that for the true unobserved multi-cause confounders to be pinpointability, it usually has to non-trivially affect an infinite number of causes.

Many factor models satisfy pinpointability when the number of causes is large. For example, probabilistic PCA and Poisson factorization lead to pinpointable Z_i as $(n+m)\cdot\log(nm)/(nm) \to 0$, where n is the number of individuals and m is the number of causes (Chen et al., 2019). Many studies also involve many causes, e.g. the GWAS study in Section 6.2 and the movie-actor study in Section 6.3.

Proposition 4.1.1: Weak unconfoundedness of the substitute confounder. Suppose Assumptions 4.1.1 and 4.1.2 hold. Consider a multi-cause separator Z_i. It satisfies weak unconfoundedness,

$$A_{i1}, \ldots, A_{im} \perp\!\!\!\perp Y_i(\boldsymbol{a}) \mid Z_i, X_i \qquad \forall \mathrm{a} \in \mathcal{A}. \tag{4.9}$$

Proof. Equation (4.8) implies the following conditional distribution of the substitute confounder given all of the other variables (including the potential outcomes),

$$\mathrm{P}(Z_i \mid A_{i1}, \ldots, A_{im}, \{Y(\mathrm{a})\}_{\mathrm{a}\in\mathcal{A}}, X_i) = \delta_{f(A_{i1},\ldots,A_{im})}. \tag{4.10}$$

Assumption 4.1.2 only concerns the probability space of the observed causes $\mathcal{A}$. But Equation (4.10) holds because $\mathrm{P}(Z_i \mid A_{i1}, \ldots, A_{im})$ is a point mass $\delta_{f(A_{i1},\ldots,A_{im})}$, which satisfies $\delta_{f(A_{i1},\ldots,A_{im})} \perp\!\!\!\perp \{Y_i(\mathrm{a})\}_{\mathrm{a}\in\mathcal{A}}, X_i$.

Equation (4.10) implies that the substitute confounder Z_i is unique. The reason is that if two variables Z_{i1} and Z_{i2} satisfy Equation (4.10) then they must be equal in the whole probability space, $Z_{i1} = Z_{i2} = f(A_{i1}, \ldots, A_{im})$ on $\mathcal{A} \times \mathcal{X} \times \{\mathcal{Y}(\mathrm{a})\}_{a\in\mathcal{A}}$.

The uniqueness of Z_i, together with Assumption 4.1.1, implies that the substitute confounder satisfies weak unconfoundedness,

$$A_{i1}, \ldots, A_{im} \perp\!\!\!\perp Y_i(\mathrm{a}) \mid Z_i, X_i \qquad \forall \boldsymbol{a} \in \mathcal{A}. \tag{4.11}$$

Why? Assumption 4.1.1 asserts the *existence* of a random variable that is (1) the smallest σ-algebra that renders the causes conditionally independent and (2) satisfies weak unconfoundedness. The argument above establishes the *uniqueness* of the random variable that satisfies (1). Thus the variable satisfying (1) must also satisfy (2). This unique random variable is the substitute confounder Z_i. [§]

□

Assumption 4.1.1 posits the existence of a multi-cause separator that also satisfies unconfoundedness. Assumption 4.1.2 implies there is only one multi-cause separator Z_i, it is unique. (The next paragraph discusses why.) These two assumptions together imply that the multi-cause separator Z_i also satisfies weak unconfoundedness.

Assumption 4.1.2 pinpoints the separator as a function of the causes. Why does this assumption imply the uniqueness of the multi-cause separator, particularly across the probability space expanded with all potential outcomes? Equation (4.8) implies

$$\mathrm{P}(Z_i \mid A_{i1}, \ldots, A_{im}, \{Y_i(\boldsymbol{a})\}_{\mathrm{a}\in\mathcal{A}}, X_i) = \delta_{f(A_{i1},\ldots,A_{im})}. \tag{4.12}$$

If two variables Z_{i1} and Z_{i2} both satisfy Equation (4.12) then they must be equal. The reason is that $Z_{i1} = Z_{i2} = f(\boldsymbol{A}_i)$ in the full probability space $\mathcal{A} \times \mathcal{X} \times \{\mathcal{Y}(\mathrm{a})\}_{a\in\mathcal{A}}$. Thus Assumption 4.1.2 implies the uniqueness of the multi-cause separator.

The deconfounder algorithm operationalizes this theory. *If* the investigator finds a deterministic function of the causes that renders them conditionally independent *then* the output of that function can be used as a substitute confounder in a downstream causal inference. Assumption 4.1.1 is that the observed covariates X_i and multi-cause separator U_i provide weak unconfoundedness: there are no unobserved single-cause confounders. Assumption 4.1.2 is that there is a single $f(\boldsymbol{A}_i)$ that provides the substitute confounder.

[§] This proof simplifies the proof of Lemma 2 in Wang and Blei (2019a).

The algorithm uses a probabilistic factor model and posterior predictive check to find $f(\boldsymbol{A}_i)$. Suppose a factor model describes well the distribution of the causes. Then its local latent variable renders the causes conditionally independent. When the number of causes is large and the local latent variable is low-dimensional, this inference approaches a deterministic function, satisfying Equation (4.8). The deconfounder infers the local latent variables and calls them substitute confounders.

Why are the inferred confounders multi-cause separators? Why do they form the smallest σ-algebra that renders the causes conditionally independent? The reason has two parts. (1) The factor model implies that its latent variable renders the causes conditionally independent. (2) The σ-algebra of a pinpointed separator cannot pick up single-cause variables; Lemma 9.2.1 provides a proof.

Assumption 4.1.2 requires a single deterministic function that provides the separator; the algorithm uses a factor model to find it. Kruskal (1989) and Allman et al. (2009) provide conditions that guarantee the uniqueness of the factor model that captures the distribution of the causes. Bai and Li (2012) and Chen et al. (2019) study conditions under which the latent variables of factor models are identifiable. With many causes and a low-dimensional factor model, the inference of its local variable approaches a deterministic function.

Finally, the deconfounder uses posterior predictive checks (Gelman et al., 1996; Guttman, 1967; Rubin, 1984) to assess the fidelity of the distribution of causes that is provided by the factor model. Specifically, the check evaluates the predictive distribution on sets of held-out causes. This strategy uses ideas from Bayarri and Berger (2000); Ranganath and Blei (2019); Robins et al. (2000a) to provide a better-calibrated check.

4.2 Causal identification of the deconfounder

Building on Proposition 4.1.1, we discuss a collection of causal identification results around the deconfounder. We prove that the deconfounder can identify three causal quantities under suitable conditions.[¶] These causal quantities include the average causal effect of all the causes, the average causal effect of subsets of the causes, and the conditional potential outcome.

[¶] Here "identify" means the causal quantity can be written as a function of the observed data. Moreover, the deconfounder can unbiasedly estimate it.

We first study the average causal effect of all the causes.

Theorem 4.2.1: Identification of the average causal effect of all the causes. Assume SUTVA, "no unobserved single-cause confounders," and pinpointability. Then, under conditions described below, the deconfounder non-parametrically identifies the average causal effect of all the causes. The average causal effect of changing the causes from $\boldsymbol{a} = (a_1, \ldots, a_m)$ to $\boldsymbol{a}' = (a'_1, \ldots, a'_m)$ is

$$\begin{aligned}&\mathbb{E}_Y\left[Y_i(\boldsymbol{a})\right] - \mathbb{E}_Y\left[Y_i(\boldsymbol{a}')\right] \\ &\quad = \mathbb{E}_{Z_i, X_i}\left[\mathbb{E}_Y\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i, X_i\right]\right] - \mathbb{E}_{Z_i, X_i}\left[\mathbb{E}_Y\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}', Z_i, X_i\right]\right]. \qquad (4.13)\end{aligned}$$

This holds with the following two conditions:[‖] (1) The substitute confounder is a piece-wise constant function of the (continuous) causes: $\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) = 0$ up to a set of Lebesgue measure zero; (2) The outcome is separable,

$$\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid Z_i = z, X_i = x\right] = f_1(\boldsymbol{a}, x) + f_2(z), \qquad (4.14)$$

$$\mathbb{E}\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = z, X_i = x\right] = f_3(\boldsymbol{a}, x) + f_4(z), \qquad (4.15)$$

for all $(\boldsymbol{a}, x, z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$ and some continuously differentiable[**] functions f_1, f_2, f_3, and f_4.[††]

Proof sketch. Theorem 4.2.1 relies on Proposition 4.1.1: "No unobserved single-cause confounders" and pinpointability ensure unconfoundedness given the substitute confounder Z_i and the observed covariates X_i. They greenlight us for causal inference.

Theorem 4.2.1 then leverages two additional conditions to identify average causal effects. It assumes $\nabla_{\boldsymbol{a}} f(a_1, \ldots, a_m) = 0$ almost everywhere and a separable outcome model. These two assumptions let us identify the average causal effect without assuming overlap.

[‖] We assume the two conditions—"piece-wise constant" and "separable"—for the substitute confounder. However, it suffices to assume the same two conditions for the unobserved multi-cause confounders. The former is easier to check; it also implies the latter because of Proposition 4.1.1.

[**] For binary causes, we can analogously assume that there exists $\boldsymbol{a}_{\text{new}}$ and $\boldsymbol{a}'_{\text{new}}$ such that $\boldsymbol{a}_{\text{new}} - \boldsymbol{a}'_{\text{new}} = \boldsymbol{a} - \boldsymbol{a}'$ and they lead to the same substitute confounder estimate $f(\boldsymbol{a}_{\text{new}}) = f(\boldsymbol{a}'_{\text{new}})$. Further, the outcome model is separable: $\mathbb{E}\left[Y_i(\boldsymbol{a}) - Y_i(\boldsymbol{a}') \mid Z_i = z, X_i = x\right] = f_1(\boldsymbol{a} - \boldsymbol{a}', x) + f_2(z)$.

[††] The expectation over Z_i and X_i is taken over $P(Z_i, X_i)$ in Equation (4.13): $\mathbb{E}_{Z_i, X_i}\left[\mathbb{E}_Y\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i, X_i\right]\right] = \int \mathbb{E}_Y\left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i, X_i\right] P(Z_i, X_i)\, \mathrm{d}Z_i\, \mathrm{d}X_i$.

More specifically, $\nabla_{\boldsymbol{a}} f(a_1, \ldots, a_m) = 0$ roughly requires that the substitute confounder is a step function of all causes. In other words, we can partition all possible values of $(a_1, \ldots, a_m)$ into countably many regions. In each region, the value of the substitute confounder must be a constant. But the substitute confounder can take different values in different regions. This condition ensures that the average causal effect $\mathbb{E}_Y[Y_i(\boldsymbol{a})] - \mathbb{E}_Y[Y_i(\boldsymbol{a}')]$ is identifiable if $\boldsymbol{a}$ and $\boldsymbol{a}'$ belong to the same region.

Further, we assume the outcome model be separable in the substitute confounder and the causes. It roughly requires that there is no interaction between the substitute confounder and the causes. This separability condition lets us identify the average causal effect for all values of $\boldsymbol{a}$ and $\boldsymbol{a}'$. The full proof is in Appendix A.1. □

Theorem 4.2.1 shows that the deconfounder can unbiasedly estimate the average causal effect of all the causes. It requires two conditions beyond "no unobserved single-cause confounders," SUTVA, and pinpointability. The first condition requires that the substitute confounder be a piece-wise constant function of the causes; it is satisfied when the substitute confounder is discrete and the causes are continuous. We remark that this piece-wise constant condition does not assume away all confounding. For example, it is satisfied when the substitute confounder (and hence the unobserved confounder) is a discretization of the causes. In this case, the substitute confounder still correlates with the causes while satisfying the piece-wise constant condition.

The second condition of Theorem 4.2.1 requires that the potential outcome be separable in the substitute confounder and the causes; the observed data also respects this separability. This condition is satisfied when the substitute confounder does not interact with the causes. For example, this condition is often satisfied in GWAS studies: the effect of SNPs on an individual's height does not depend on his/her ancestry (Veturi et al., 2019). A reader might ask: how can the outcome be separable in the substitute confounder Z_i and the causes A_i when $Z_i = f(A_i)$, which is required by the pinpointability assumption? The reason is that f is a non-differentiable piece-wise constant function by condition (1), while f_1, f_2, f_3, f_4 are differentiable required by condition (2). In this way, the conditional expectation $\mathbb{E}[Y_i(\boldsymbol{a}) \mid Z_i = z, X_i = x]$ is can be separated into two components, one differentiable $f_1(\boldsymbol{a}, x)$ and one non-differentiable $f_2(z)$. A similar argument also holds for $\mathbb{E}[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = z, X_i = x]$. It is this incongruence between X_i and Z_i in differentiability that leads to identification.

When the separability condition of Theorem 4.2.1 does not hold, we can still use the deconfounder to handle the unobserved multi-cause confounders that do not interact with the causes. As long as the observed covariates include those that do interact with the causes, the deconfounder produces unbiased estimates of the average causal effect.

We next discuss the identification of the average causal effect for subsets of the causes.

Theorem 4.2.2: . (Identification of the average causal effect of subsets of the causes) Assume SUTVA, "no unobserved single-cause confounders," and pinpointability. Then, under the condition described below, the deconfounder non-parametrically identifies the average causal effect of subsets of causes. The average causal effect of changing the first k ($k < m$) causes from $a_{1:k} = (a_1, \ldots, a_k)$ to $a'_{1:k} = (a'_1, \ldots, a'_k)$ is

$$\begin{aligned}&\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] - \mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a'_{1:k}, A_{i,(k+1):m})\right]\right]\\ &\quad = \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i,1:k} = a_{1:k}\right]\right] - \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i,1:k} = a'_{1:k}\right]\right]. \quad (4.16)\end{aligned}$$

This holds with the following condition: The first k causes $A_{i1}, \ldots, A_{ik}$ satisfy overlap, $P((A_{i1}, \ldots, A_{ik}) \in \mathcal{A} \mid Z_i, X_i) > 0$ for any set $\mathcal{A}$ such that $P(\mathcal{A}) > 0$.[‡‡]

Proof sketch. Similar to Theorem 4.2.1, Theorem 4.2.2 uses Proposition 4.1.1 to greenlight the use of a substitute confounder. It then relies on overlap to identify the average causal effect; we follow the classical argument that identifies the average treatment effect (Imbens and Rubin, 2015a). The full proof is in Appendix A.2. □

Theorem 4.2.2 shows that the deconfounder can unbiasedly estimate the average causal effect of subsets of the causes. It lets us answer "how would the movie revenue change, on average, if we place Meryl Streep and Sean Connery into a movie?" Beyond SUTVA, "no unobserved single-cause confounders," and pinpointability, Theorem 4.2.2 requires overlap. Overlap ensures that $\mathbb{E}_Y[Y_i \mid Z_i, X_i, A_{i,1:k} = a_{1:k}]$ is estimable from the observed data for all possible values of $(Z_i, X_i, A_{i,1:k})$. The overlap assumption about the causes in Theorem 4.2.2 replaces the separability assumption about the outcome model required by Theorem 4.2.1.

[‡‡]In full notation, $\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] = \mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1, \ldots, a_k, A_{ik+1}, \ldots, A_{im})\right]\right]$.

We note that the overlap condition and the pinpointability assumption are compatible. Though pinpointability requires $P(Z_i \mid \boldsymbol{A}_i) = \delta_{f(\boldsymbol{A}_i)}$, it is still possible for subsets of the causes to satisfy overlap; the pinpointability assumption only prevents the complete set of m causes from satisfying overlap. For example, consider a pinpointed substitute confounder that is one-dimensional, $Z_i = \sum_{j=1}^{m} \alpha_j A_{ij}$. Any $k \leq m - 1$ causes satisfy overlap, but the complete set of m causes do not.

Finally, we discuss the identification of the conditional mean potential outcome.

Theorem 4.2.3: . (Identification of the conditional mean potential outcome) Assume SUTVA, "no unobserved single-cause confounders," and pinpointability. Then, under the condition described below, the deconfounder non-parametrically identifies the mean potential outcome of an individual given its current assigned causes. If an individual is assigned with $\boldsymbol{a} = (a_1, \ldots, a_m)$, then its potential outcome under a different assignment $\boldsymbol{a}' = (a'_1, \ldots, a'_m)$ is

$$\mathbb{E}_Y \left[Y_i(\boldsymbol{a}') \mid \boldsymbol{A}_i = \boldsymbol{a} \right] = \mathbb{E}_{Z,X} \left[\mathbb{E}_Y \left[Y_i \mid Z_i, X_i, \boldsymbol{A}_i = \boldsymbol{a}' \right] \right].$$

This holds with the following condition: The cause assignment of interest $\boldsymbol{a}'$ leads to the same substitute confounder estimate as the observed assigned causes: $P(Z_i \mid \boldsymbol{A}_i = \boldsymbol{a}) = P(Z_i \mid \boldsymbol{A}_i = \boldsymbol{a}')$.

Proof sketch. As with Theorem 4.2.1 and Theorem 4.2.2, Theorem 4.2.3 relies on the unconfoundedness given the substitute confounders Z_i and the observed covariates X_i due to Proposition 4.1.1. It then identifies the potential outcome by focusing on the data points with the same substitute confounder estimate. We note that this identification result does not require overlap. The full proof is in Appendix A.3. □

Theorem 4.2.3 nonparametrically identifies the mean potential outcome of an individual $Y_i(\boldsymbol{a}')$ given its current assigned causes $\boldsymbol{A}_i = \boldsymbol{a}$. The only requirement is about the configurations of cause assignments we can query, $\boldsymbol{a}'$; these configurations should lead to the same substitute confounder estimate as the current assigned causes.

We illustrate this condition with actors causing movie revenue. For simplicity, assume the substitute confounder captures the genre of each movie. Start with one of the James Bond movies; it is a spy film. We can ask what its revenue would be if we make its cast to be that of "The Bourne Trilogy" (also a spy film). Alternatively, we can query what if we make its cast to include some actors from "The Bourne Trilogy" and other actors from "North By Northwest"; both are spy films. However, we can not query what if we make its cast to be that of "The Shawshank Redemption" (which is not a spy film).

Theorems 4.2.1 to 4.2.3 confirm the validity of the deconfounder by providing three sets of identification results. When the assumptions in Theorems 4.2.1 to 4.2.3 do not hold, we recommend evaluating the uncertainty of the deconfounder estimate. Section 3.4 discusses how; Section 6.1 gives an example. The posterior distribution of the deconfounder estimate reflects how the (finite) observed data informs causal quantities of interest (Gustafson, 2010, 2015). When the causal quantity is non-identifiable, the posterior distribution of the deconfounder estimate will reflect this non-identifiability. For example, if the causal quantity is non-identifiable over $\mathcal{R}$, the posterior distribution of the deconfounder estimate shall also be uniform over $\mathcal{R}$ (with non-informative priors).

5

A causal graphical view of the deconfounder

In this chapter, we take a causal graphical view of the blessings of multiple causes and provide an additional set of identification results via a connection to the proxy variable strategy (Miao et al., 2018).*

Consider a multiple causal inference with *shared confounding*. This setting is in the causal graph of Figure 5.1a, where an unobserved confounder U affects multiple causes $\{A_1, \ldots, A_m\}$ and an outcome Y. 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. We then revisit the deconfounder algorithm and show that it produces correct estimates of $p(y \mid \text{do}(a_{\mathcal{C}}))$; this result justifies the deconfounder on causal graphs.

*This chapter is in part based on Wang and Blei (2019d).

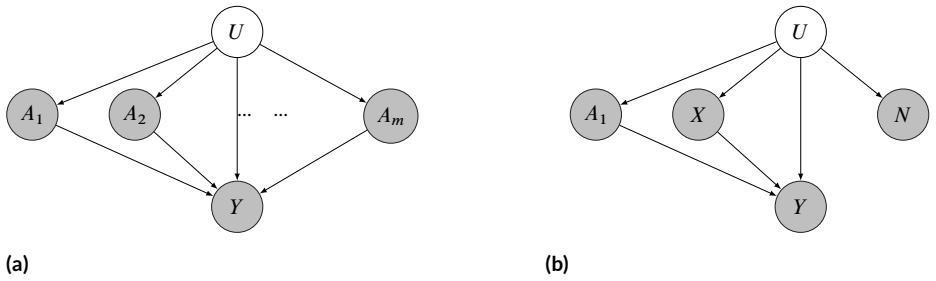


Figure 5.1: (a) Multiple causes with shared confounding. (b) Proxy variables for an unobserved confounder (Miao et al., 2018). (Only the shaded nodes are observed.)

The key observation here is that, under shared confounding, some causes can serve as proxies (Kuroki and Pearl, 2014; Miao et al., 2018) of unobserved confounders for causal identification of others. This observation helps identify the intervention distributions of subsets of causes. Further, unlike previous works in proxy variables, we do not need to find two external proxies for the unobserved confounder; some causes themselves can serve as proxies for other causes.

These results also illustrate how the deconfounder can help put into practice the proxy-variable strategy for identification (Miao et al., 2018). While existing identification formulas for proxy variables involve solving integral equations, the results here show how to circumvent the need for such solutions by directly modeling the data.

We finally generalize these results to the larger class of graphs in Figure 5.2a. 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.

5.1 Multiple causes with shared confounding

Consider the multiple causal inference problem in Figure 5.1a. There are m causes $A_1, \ldots, A_m$ that all affect the outcome Y; and there is an unobserved confounder U that affects Y and the causes. This graph exemplifies shared unobserved confounding, where U affects multiple causes.

In this chapter, the goal of multiple causal inference is to estimate the intervention distributions on subsets of causes, $\mathrm{P}(Y \mid \mathrm{do}(A_{\mathcal{C}} = a_{\mathcal{C}}))$. It is the distribution of the outcome Y if we intervene on $A_{\mathcal{C}} \subset \{A_1, \ldots, A_m\}$, which is a (strict) subset. (For example, we might be interested in each cause individually; then each subset contains one of the causes.) Relating this goal to Chapter 4, the intervention distribution on the first k causes $\mathrm{P}(Y \mid \mathrm{do}(A_{1:k} = a_{1:k}))$ is equivalent to the distribution of the potential outcomes $\mathrm{P}(Y_i(a_{1:k}))$ in Theorem 4.2.2. We will establish causal identification in this setting and prove the validity of causal estimation with the deconfounder algorithm. We extend these results to more general graphs Section 5.2.

5.1.1 Causal identification

An intervention distribution is *identifiable* if it can be written as a function of the observed data distribution (e.g., $\mathrm{P}(y, a_1, \ldots, a_m)$ in Figure 5.1a) (Pearl, 2009). Identifiability ensures that an intervention distribution is *estimable* from the observed data. In Figure 5.1a, which intervention distributions can be identified? In this section we prove that, under suitable conditions, the intervention distributions of subsets of the causes $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ are identifiable.[†]

The starting point for causal identification with multiple causes is the *proxy variable* strategy, which focuses on causal identification with a single cause (Kuroki and Pearl, 2014; Miao et al., 2018). Consider the causal graph in Figure 5.1b: it has a single cause A_1, an outcome Y, and an unobserved confounder U. The goal is to estimate the intervention distribution $\mathrm{P}(y \mid \mathrm{do}(a_1))$. There are some other variables in the graph too. A *proxy* X is an observable child of the unobserved confounder; a *null proxy* N is a proxy that does not affect the outcome. The theory around proxy variables says that the intervention distribution $\mathrm{P}(y \mid \mathrm{do}(a_1))$ is identifiable if (1) we observe two proxies of the unobserved confounder U and (2) one of the proxies is a null proxy (Miao et al., 2018). In particular, since N and X are observed, $\mathrm{P}(y \mid \mathrm{do}(a_1))$ is identifiable.

[†]We use the abbreviation $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}})) \triangleq \mathrm{P}(y \mid \mathrm{do}(A_{\mathcal{C}} = a_{\mathcal{C}}))$.

We leverage the idea of proxy variables to identify intervention distributions in Figure 5.1a, multiple causes with shared unobserved confounding. The main idea is to use some causes as proxies to identify the intervention distributions of other causes. The benefit is that, with multiple causes, we do not need to observe external proxy variables; rather the causes themselves serve as proxies. Nor do we need to observe a null proxy, one that does not affect the outcome (like N in Figure 5.1b); we only need to assume that there is a function of the causes 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 causes and the outcome. This is the "blessing of multiple causes" from the causal graphical view.

We formally state the identification result. To repeat, assume the causal graph in Figure 5.1a with m causes $A_{1:m}$, an outcome Y, and a shared unobserved confounder U. The goal is to identify the intervention distribution of a *strict subset* of the causes $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

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

Assumption 5.1.1: . 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 $f(A_{\mathcal{N}})$:

$$f(A_{\mathcal{N}}) \perp Y \mid U, A_{\mathcal{C}}, A_{\mathcal{X}}, \tag{5.1}$$

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

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

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

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

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

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

Assumption 5.1.1.1 posits that a set of causes $A_{\mathcal{N}}$ exists such that some function of them $f(A_{\mathcal{N}})$ can serve as a null proxy (Equation (5.1)). Roughly, it requires $f(A_{\mathcal{N}})$ does not affect the outcome. Note, it does not require that we know $f(A_{\mathcal{N}})$, just that it exists.

When might this assumption be satisfied? First, suppose some of the multiple causes do not affect the outcome. Then Assumption 5.1.1.1 reduces to the null proxy assumption (D'Amour, 2019b; Kuroki and Pearl, 2014; Miao et al., 2018). This might be plausible, e.g., in a genetic study or other settings where there are many causes. Again, we do not need to know *which* causes are "null causes," only that they exist. Indeed, as long as two causes are null, the theory below implies that the intervention distribution of each individual cause is identifiable.

But this assumption goes beyond a restatement of the null proxy assumption. Suppose two (or more) causes 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 causes 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 does not affect joyfulness one way or the other.) Then the function $f(A_{\mathcal{N}})$ is XOR of the bundle; the quantity (bread XOR butter) does not affect Y. Again, the function and set must exist; we do not need to know them.

As a more serious example, consider that HDL cholesterol, LDL cholesterol, and triglycerides (TG) affect the risk of a heart attack through the ratios HDL/LDL and TG/HDL (Millán et al., 2009). Then HDL×LDL and TG×HDL are both examples of $f(A_{\mathcal{N}})$ that do not affect Y. The existence of one of them suffices for Assumption 5.1.1.1. (We discuss this assumption in more technical detail in Appendix B.2.)

Assumption 5.1.1.2 and Assumption 5.1.1.3 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 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 (D'Haultfoeuille, 2011; Miao et al., 2018; Yang et al., 2017); it is often used to guarantee the existence and the uniqueness of solutions to integral equations. See Chen et al. (2014) for a detailed discussion of completeness.

Under Assumption 5.1.1, we identify the intervention distribution of the subset of the causes $A_{\mathcal{C}}$.

Theorem 5.1.1: . (Causal identification under shared confounding) Assume the causal graph Figure 5.1a. (Note the data does not need to be "faithful" to the graph—some edges can be missing.) Under Assumption 5.1.1, the intervention distribution of the causes $A_{\mathcal{C}}$ is identifiable:

$$\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \mathrm{P}(a_{\mathcal{X}}) \, \mathrm{d}a_{\mathcal{X}} \tag{5.2}$$

for any solution h to the integral equation

$$\mathrm{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \mathrm{P}(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}}. \tag{5.3}$$

Moreover, the solution to Equation (5.3) always exists under weak regularity conditions in Appendix B.5.

Proof sketch. The proof 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 5.1b; all the m causes can affect the outcome. However, Assumption 5.1.1.1 allows $f(A_{\mathcal{N}})$ to play the role of a null proxy. The full proof is in Appendix B.1. □

Theorem 5.1.1 identifies the intervention distributions of subsets of the causes $A_{\mathcal{C}}$; it writes $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ as a function of the observed data distribution $\mathrm{P}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, a_{\mathcal{N}})$. In particular, it lets us identify the intervention distributions of individual causes $\mathrm{P}(y \mid \mathrm{do}(a_i)), i = 1, \ldots, m$. By using the causes themselves as proxies, Theorem 5.1.1 exemplifies how the multiplicity of the causes enables causal identification under shared unobserved confounding.

5.1.2 Causal estimation with the deconfounder

Theorem 5.1.1 guarantees that the intervention distribution $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ is estimable from the observed data. However, it involves solving an integral equation (Equation (5.3)). This integral equation is hard to solve except in the simplest linear Gaussian case (Carrasco et al., 2007). How can we estimate $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ in practice?

We then rewrite the deconfounder algorithm in the causal DAG notation. We show that the deconfounder correctly estimates the intervention distribution $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$; it implicitly solves the integral equation in Equation (5.3) by modeling the data. This result justifies the deconfounder from a causal graphical perspective.

The deconfounder algorithm We first re-state the algorithm in causal DAG notations. 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:

$$\hat{\mathrm{P}}(a_1, \ldots, a_m, \hat{z}) = \hat{\mathrm{P}}(\hat{z}) \prod_{j=1}^{m} \hat{\mathrm{P}}(a_j \mid \hat{z}), \tag{5.4}$$

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

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

$$\mathrm{P}(y, a_1, \ldots, a_m) = \int \hat{\mathrm{P}}(y \mid a_1, \ldots, a_m, \hat{z}) \hat{\mathrm{P}}(a_1, \ldots, a_m, \hat{z})\, \mathrm{d}\hat{z}. \tag{5.5}$$

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

3. **Estimate the intervention distribution.** The final step estimates the intervention distribution $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ by integrating out the non-intervened causes and the substitute confounder,

$$\hat{\mathrm{P}}(y \mid \mathrm{do}(a_{\mathcal{C}})) \triangleq \int \hat{\mathrm{P}}(y \mid a_1, \ldots, a_m, \hat{z}) \times \hat{\mathrm{P}}(a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \hat{z})\, \mathrm{d}\hat{z}\, \mathrm{d}a_{\{1,\ldots,m\}\setminus\mathcal{C}}. \tag{5.6}$$

This is the estimate of the deconfounder.

The correctness of the deconfounder. Note that many possible $\hat{\mathrm{P}}(\cdot)$'s satisfy the deconfounder requirements (Equations (5.4) and (5.5)); the algorithm outputs one such $\hat{\mathrm{P}}$. Under suitable conditions, we show that any such $\hat{\mathrm{P}}$ provides the correct causal estimate $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

Assumption 5.1.2: . The deconfounder estimate $\hat{\mathrm{P}}(y, a_1, \ldots, a_m, \hat{z})$ satisfies two conditions:

1. It is consistent with Assumption 5.1.1.1,

$$\hat{\mathrm{P}}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}) = \hat{\mathrm{P}}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, \hat{z}). \tag{5.7}$$

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

Assumption 5.1.2.1 roughly requires that there exists a function f and a subset of the causes $A_{\mathcal{N}}$ such that $f(A_{\mathcal{N}})$ does not affect the outcome in the deconfounder outcome model. (When the number of causes goes to infinity, Assumption 5.1.2.1 reduces to Assumption 5.1.1.1.) We emphasize that $f(A_{\mathcal{N}})$ is not involved in calculating the deconfounder estimate (Equation (5.6)); it only appears in Assumption 5.1.2.1. Hence the correctness of the deconfounder does not require specifying $f(\cdot)$ and $A_{\mathcal{N}}$, just that it exists.

Assumption 5.1.2.2 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 5.1.1.

Now we state the correctness result of the deconfounder.

Theorem 5.1.2: . (Correctness of the deconfounder under shared confounding) Assume the causal graph Figure 5.1a. Under Assumption 5.1.1, Assumption 5.1.2 and weak regularity conditions, the deconfounder provides correct estimates of the intervention distribution:

$$\hat{\mathrm{P}}(y \mid \mathrm{do}(a_{\mathcal{C}})) = \mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}})), \tag{5.8}$$

where $\hat{\mathrm{P}}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ is computed from Equation (5.6).

Proof sketch. The proof of Theorem 5.1.2 relies on a key observation: the deconfounder implicitly solves the integral equation (Equation (5.3)) by modeling the observed data with $\hat{\mathrm{P}}(y, a_1, \ldots, a_m, \hat{z})$. Assumption 5.1.2.2 guarantees that the deconfounder estimate can be written as

$$\hat{\mathrm{P}}(y \mid a_{\mathcal{C}}, \hat{z}) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{\mathrm{P}}(a_{\mathcal{X}} \mid \hat{z}) \, \mathrm{d}a_{\mathcal{X}} \tag{5.9}$$

under weak regularity conditions; this function $\hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}})$ also solves the integral equation (Equation (5.3)). The deconfounder uses this solution to form an estimate of $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$; this estimate is correct because of Theorem 5.1.1. The full proof is in Appendix B.4. □

Theorem 5.1.2 justifies the deconfounder for multiple causal inference under shared confounding (Figure 5.1a). It proves that the deconfounder correctly estimates the intervention distributions when they are identifiable. To illustrate Theorems 5.1.1 and 5.1.2, Appendix B.3 gives a linear example.

Theorem 5.1.2 complements Theorems 4.2.1 to 4.2.3; it establishes identification and correctness by assuming there exists some function of the causes that does not affect the outcome. In contrast, Theorems 4.2.1 to 4.2.3 assume pinpointability, namely the substitute confounder is a deterministic function of the causes. The pinpointability assumption is stronger; conditional on the causes, Theorems 5.1.1 and 5.1.2 allow the substitute confounder to be random.

Theorem 5.1.2 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 5.1.2 shows how to circumvent this need by directly modeling the data and applying the deconfounder; it implicitly solves the integral equations with the modeled data.

Finally, how do we assess Assumption 5.1.1.1 and Assumption 5.1.2.1 in practice? We operate in two stages. In the first stage, we verify Assumption 5.1.2.1 (Equation (5.7)): we apply the deconfounder algorithm and find which pairs of $(f, A_{\mathcal{N}})$ satisfy Equation (5.7) in the fitted outcome model. We can often find many such pairs. In the second stage, we verify Assumption 5.1.1.1 (Equation (5.1)). That is, we ask the domain experts, among the $(f, A_{\mathcal{N}})$ pairs from the first stage, if there exists one pair that satisfies Equation (5.1). If so, both Assumption 5.1.1.1 and Assumption 5.1.2.1 are satisfied.

5.2 Multiple causes on general causal graphs

We discussed causal identification and estimation when multiple causes share the same unobserved confounder. We now extend these results to more general causal graphs, those with several types of nodes and, in particular, that include a selection variable (Bareinboim et al., 2014). Using the results in Section 5.1, we establish causal identification and estimate intervention distributions on this class of general causal graphs.

5.2.1 General causal graphs

We focus on the class of general causal graphs in Figure 5.2a.[§] As above, it has m causes $A_{1:m}$ and an outcome Y. The goal is to estimate $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$, where $A_{\mathcal{C}} \subset \{A_1, \ldots, A_m\}$ is a subset of causes on which we intervene. Apart from the causes and the outcome, the causal graph has a few other types of variables. (Table 5.1 contains a glossary of terms.)

Confounders. Confounders are parents of *both* the causes and the outcome; they can be unobserved. In Figure 5.2a, for example, U_i^{sng} and U_i^{mlt} 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 like U_i^{sng} affect only one cause; multi-cause confounders like U_i^{mlt} affect two or more causes.

[§]There exist causal graphs that do not fall in this class; we leave them for future work.

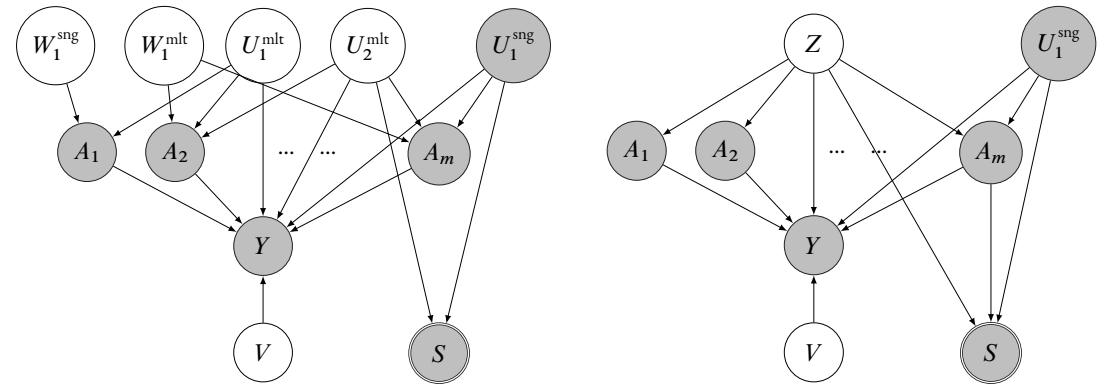


Figure 5.2: (a) The class of more general causal graphs. S is the selection operator. (b) The reduced causal graph with shared confounding.

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

Selection operator. Following Bareinboim and 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 causes 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 5.2a allows selection to occur on the confounders.

These variables—confounders, covariates, and the selection operator—compose the more general causal graphs with multiple causes. We study identification and estimation on these causal graphs.

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

Table 5.1: Types of nodes

5.2.2 Causal identification

We extend the results around causal identification under shared confounding (Theorems 5.1.1 and 5.1.2) to more general graphs. We first reduce the general graph (Figure 5.2a) to one close to the shared confounding case; then we handle the complications of selection bias. These steps lead to causal identification and the correctness of the deconfounder on general causal graphs.

Reduction to shared confounding. To reduce the general graph of Figure 5.2a, we bundle all the unobserved multi-cause confounders and null confounders $\{U^{\text{mlt}}, W^{\text{mlt}}\}$ into a single unobserved confounder Z. This variable Z is shared by all the causes as in Figure 5.1a and renders all the causes conditionally independent. Moreover, it is sufficient to adjust for Z and single-cause confounders U^{sng} to estimate $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ because $\{U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}\}$ constitute an admissible set.

Following this discussion, we can equivalently identify the intervention distributions $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ in the general causal graph (Figure 5.2a) using a reduced causal graph (Figure 5.2b); it involves only the single-cause confounders U^{sng} and a shared confounder Z. Below we formally state the validity of the reduction from general causal graphs to ones with shared confounding.

Lemma 5.2.1: . (Validity of reduction) Assume the causal graph in Figure 5.2a. Adjusting for the multi-cause confounders and null confounders on the general causal graph Figure 5.2a is equivalent to adjusting for the shared confounder in Figure 5.2b:

$$\mathrm{P}(y \mid u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \ldots, a_m, s = 1) = \mathrm{P}(y \mid u^{\text{sng}}, z, a_1, \ldots, a_m, s = 1). \tag{5.10}$$

Proof sketch. The proof uses a measure-theoretic argument to characterize the information contained in the Z variable in Figure 5.2b. Roughly, the information in Z is same as the information of all multi-cause confounders, all null confounders, and some independent error:

$$\sigma(z) = \sigma(u^{\text{mlt}}, w^{\text{mlt}}, \epsilon_Z), \tag{5.11}$$

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

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

Equation (5.11) implies that conditioning on Z is equivalent to conditioning on U^{mlt}, W^{mlt}, ϵ_Z; it leads to Equation (5.10). The full proof is in Appendix B.6. □

Causal identification on the reduced causal graph (Figure 5.2b). We have just reduced general causal graphs (Figure 5.2a) to one with shared confounding (Figure 5.2b). This reduction allows us to establish causal identification on general causal graphs. We extend Theorem 5.1.1 from Figure 5.1a to Figure 5.2b. With the reduction step (Lemma 5.2.1), it leads to causal identification on general causal graphs.

How can we identify the intervention distributions $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$ on the reduced graph (Figure 5.2b)? Figure 5.2b has a confounder Z that is shared across all causes. This structure is similar to the unobserved shared confounding Figure 5.1a. 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 5.2.1: . The causal graph Figure 5.2b satisfies the following conditions:

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

$$S \perp (A, Y) \mid Z, U^{\text{sng}}. \tag{5.12}$$

3. We observe the non-selection-biased distribution

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

and the selection-biased distribution

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

Assumption 5.2.1.1 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 5.2.1.2 roughly requires that selection can only occur on the confounders. Assumption 5.2.1.3 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 5.2b. We additionally make Assumption 5.2.2; it is a variant of Assumption 5.1.1 but involves single-cause confounders and the selection operator.

Assumption 5.2.2: . 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 $f(A_{\mathcal{N}})$:

$$f(A_{\mathcal{N}}) \perp Y \mid Z, A_{\mathcal{C}}, A_{\mathcal{X}}, U^{\text{sng}}, S = 1 \tag{5.13}$$

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

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

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

Under Assumption 5.2.1 and Assumption 5.2.2, we can identify the intervention distributions $\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}}))$.

Lemma 5.2.2: . Assume the causal graph Figure 5.2b. Under Assumption 5.2.1 and Assumption 5.2.2, the intervention distribution of the causes $A_{\mathcal{C}}$ is identifiable:

$$\mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}})) = \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}_{\mathcal{C}}) \mathrm{P}(a_{\mathcal{X}}) \mathrm{P}(u^{\text{sng}}_{\mathcal{C}}) \, \mathrm{d}a_{\mathcal{X}} \, \mathrm{d}u^{\text{sng}}_{\mathcal{C}} \tag{5.14}$$

for any solution h to the integral equation

$$\mathrm{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}_{\mathcal{C}}) \times \mathrm{P}(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1) \, \mathrm{d}a_{\mathcal{X}}, \tag{5.15}$$

where $U_{\mathcal{C}}^{\text{sng}}$ is the single-cause confounders affecting $A_{\mathcal{C}}$. Moreover, the solution to Equation (5.15) always exists under weak regularity conditions in Appendix B.5.

Proof sketch. The proof adopts a similar argument as in the proof of Theorem 5.1.1. We only need to take care of the additional (observed) single-cause confounders and the selection operator. In particular, Assumption 5.2.1.2 lets us shift from the selection biased distribution $\mathrm{P}(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1)$ to the non-selection-biased one $\mathrm{P}(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}})$.The full proof is in Appendix B.7. □

Causal identification on general causal graphs (Figure 5.2a). Based on the previous analysis of the reduced graph, we establish causal identification result on general causal graphs.

Theorem 5.2.1: . Assume the causal graph Figure 5.2a. Assume a variant of Assumption 5.2.1 and Assumption 5.2.2 (detailed in Appendix B.8), the intervention distribution of the causes $A_{\mathcal{C}}$ is identifiable using Equation (5.14) and Equation (5.15). (Proof in Appendix B.8.)

5.2.3 Causal estimation with the deconfounder

We finally extend the deconfounder to the more general causal graphs (Figure 5.2a) with selection bias and prove its correctness. We build on the identification result on general causal graphs (Theorem 5.2.1). We then show that the deconfounder provides correct causal estimates by implicitly solving the integral equation (Equation (5.15)). This argument is similar to the argument of Theorem 5.1.2.

The deconfounder algorithm for general causal graphs with selection bias extends the version described in Section 5.1.2. Specifically, Assumption 5.1.2 allows the deconfounder algorithm to have access to both the non-selection-biased data $\mathrm{P}(a_1, \ldots, a_m, u^{\text{sng}})$ and the selection-biased data $\mathrm{P}(y, u^{\text{sng}}, a_1, \ldots, a_m \mid s = 1)$. In this case, the deconfounder algorithm outputs two estimates:

$$\hat{\mathrm{P}}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z}) = \hat{\mathrm{P}}(\hat{z})\hat{\mathrm{P}}(u^{\text{sng}} \mid a_1, \ldots, a_m, \hat{z}) \prod_{i=1}^{n} \hat{\mathrm{P}}(a_i \mid \hat{z}), \tag{5.16}$$

and

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

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:

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

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

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

Finally the deconfounder estimates

$$\hat{\mathrm{P}}(y \mid \mathrm{do}(a_{\mathcal{C}})) \triangleq \int \hat{\mathrm{P}}(y \mid a_1, \ldots, a_m, \hat{z}, u^{\text{sng}}_{\mathcal{C}}, s = 1) \times \hat{\mathrm{P}}(a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \hat{z}) \mathrm{P}(u^{\text{sng}}_{\mathcal{C}})\, \mathrm{d}u^{\text{sng}}_{\mathcal{C}}\, \mathrm{d}\hat{z}\, \mathrm{d}a_{\{1,\ldots,m\}\setminus\mathcal{C}}, \tag{5.17}$$

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

We now prove the correctness of the deconfounder on general causal graphs. We make a variant of Assumption 5.1.2 and state the correctness result.

Assumption 5.2.3: . The deconfounder outputs the estimates $\hat{\mathrm{P}}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)$ and $\hat{\mathrm{P}}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$ that satisfy the following conditions:

1. It is consistent with Assumption 5.2.1.1:

$$\hat{\mathrm{P}}(a_1, \ldots, a_m \mid \hat{z}, u^{\text{sng}}, s = 1) = \hat{\mathrm{P}}(a_1, \ldots, a_m \mid \hat{z}, u^{\text{sng}}). \tag{5.18}$$

2. It is consistent with Assumption 5.2.2.1:

$$\hat{\mathrm{P}}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), \hat{z}, u^{\text{sng}}, s = 1) \tag{5.19}$$

$$= \hat{\mathrm{P}}(y \mid a_{\mathcal{C}}, a_{\mathcal{X}}, \hat{z}, u^{\text{sng}}, s = 1). \tag{5.20}$$

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

The conditional $\hat{\mathrm{P}}(\hat{z} \mid a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}, s = 1)$, Equation (5.18), and Equation (5.19) can be computed from the deconfounder estimate $\hat{\mathrm{P}}(a_1, \ldots, a_m, u^{\text{sng}}, \hat{z})$ and $\hat{\mathrm{P}}(y, a_1, \ldots, a_m, u^{\text{sng}}, \hat{z} \mid s = 1)$.

Theorem 5.2.2: . (Correctness of the deconfounder on general causal graphs) Assume the causal graph Figure 5.2a. Assume a variant of Assumption 5.2.1 and Assumption 5.2.2 (detailed in Appendix B.9). Under Assumption 5.2.3 and weak regularity conditions, the deconfounder provides correct estimates of the intervention distribution:

$$\hat{\mathrm{P}}(y \mid \mathrm{do}(a_{\mathcal{C}})) = \mathrm{P}(y \mid \mathrm{do}(a_{\mathcal{C}})). \tag{5.21}$$

Proof sketch. The proof of Theorem 5.2.2 follows a similar argument as in the proof of Theorem 5.1.2. We only need to additionally take care of the single-cause confounders and the selection operator. The full proof is in Appendix B.9. □

Theorem 5.2.2 establishes the correctness of the deconfounder on general causal graphs under certain types of selection bias. It justifies the deconfounder on general causal graphs.

5.3 Related work

This chapter 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 leverage the multiplicity of the causes to establish causal identification.

A second body of related work is on causal structural learning with latent variables (Anandkumar et al., 2014; Frot et al., 2019; Mckeigue et al., 2010; Silva et al., 2006). These works focus on learning whether an arrow exists between two variables, and its direction. We also use latent variables, but to a different end. We assume the direction of arrows is known and the set of causes of are always ancestors of the outcome. Unlike the work on structure learning, this chapter focuses on estimating the intervention distributions of the causes.

5.4 Discussion

We take a causal graphical view of the deconfounder algorithm. By treating some causes as proxies of the shared confounder, we can identify the intervention distributions of the other causes. For a general class of causal graphs, we prove that the intervention distribution of subsets of causes is identifiable. We further 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 chapter generalize the theory in Chapter 4 and extend the applicability of the deconfounder.

6

Empirical studies of the deconfounder

In this chapter, we study the deconfounder in three empirical studies.* Two studies involve simulations of realistic scenarios; these help assess how well the deconfounder performs relative to the ground truth. The third study is a real-world analysis. All three studies demonstrate the benefits of the deconfounder. They show how predictive checks reveal potential issues with downstream causal inference and how the deconfounder can provide closer-to-truth causal estimates.

In more detail, in Section 6.1 we study semi-synthetic data about smoking; the causes are a real dataset about smoking and the effect (medical expenses) is simulated. In Section 6.2 we study semi-synthetic data about genetics. Finally, in Section 6.3 we study real data about actors and movie revenue; there is no simulation.

*This chapter is in part based on Wang and Blei (2019a).

Each stage of the deconfounder requires computation: to fit the factor model, to check the factor model, to calculate the substitute deconfounder, and to fit the outcome model. In all these stages, we use black box variational inference (BBVI) (Kucukelbir et al., 2017; Ranganath et al., 2014). We use its RStan implementation (Carpenter et al., 2017) in Section 6.1 and its Edward implementation (Tran et al., 2017, 2016a) in Section 6.2 and 6.3. (This was a choice; other methods of computation can also be used.)

6.1 Two causes: How smoking affects medical expenses

We first study the deconfounder with semi-synthetic data about smoking. The 1987 National Medical Expenditures Survey (NMES) collected data about smoking habits and medical expenses in a representative sample of the U.S. population (Imai and Van Dyk, 2004; US Department of Health and Human Services Public Health service, 1987). The dataset contains 9,708 people and 8 variables about each. For each person, we focus on the current marital status (a_{mar}), the cumulative exposure to smoking (a_{exp}), and the last age of smoking (a_{age}). We standardize all variables.

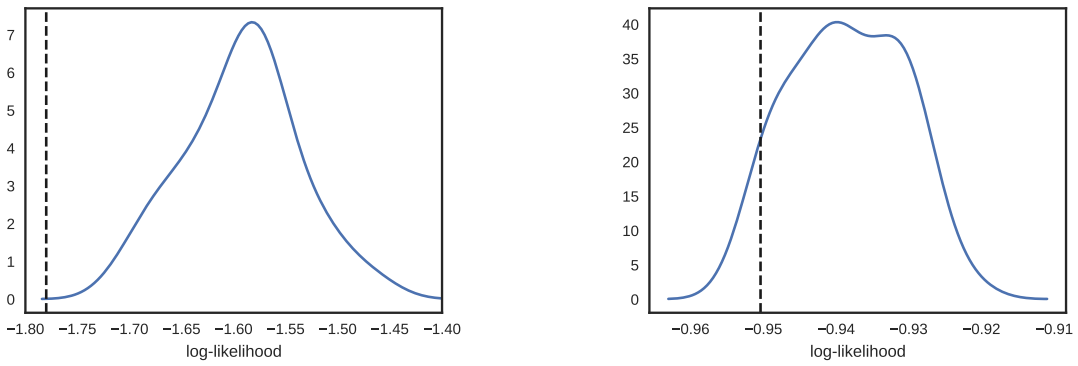
A true outcome model and causal inference problem We use the assigned causes from the survey to simulate a dataset of medical expenses, which we will consider as the outcome variable. In this simulation, the true model is linear,

$$y_i = \beta_{\text{mar}}\, a_{\text{mar},i} + \beta_{\text{exp}}\, a_{\text{exp},i} + \beta_{\text{age}}\, a_{\text{age},i} + \varepsilon_i, \tag{6.1}$$

where $\varepsilon_i \sim \mathcal{N}(0, 1)$. We generate the true causal coefficients from

$$\beta_{\text{mar}} \sim \mathcal{N}(0, 1), \quad \beta_{\text{exp}} \sim \mathcal{N}(0, 1), \quad \beta_{\text{age}} \sim \mathcal{N}(0, 1). \tag{6.2}$$

and from these coefficients we generate the outcome for each individual. The result is a dataset of 9,708 tuples $(a_{\text{mar},i}, a_{\text{exp},i}, a_{\text{age},i}, y_i)$. It is semi-synthetic: the assigned causes are from the real world, but we know the true outcome model. Note that the last smoking age is a multi-cause confounder—it affects both marital status and exposure and is one of the causes of the expenses.



(a) Linear Model

(b) Quadratic Model

Figure 6.1: Predictive checks for the substitute confounder z obtained from a linear factor model (a) and a quadratic factor model (b). The blue line is the kernel density estimate (kde) of the test-statistic based on the predictive distribution. The dashed vertical line shows the value of the test-statistic on the observed dataset. The figure shows that the linear model mismatches the data—the observed statistic falls in a low probability region of the kde. The quadratic factor model is a better fit to the data.

We are interested in estimating the causal effects of marital status and smoking exposure on medical expenses. But suppose we do not observe age; it is an unobserved confounder. We can use the deconfounder to solve the problem.

Modeling the assigned causes We begin by finding a good factor model of the assigned causes $(a_{\text{mar},i}, a_{\text{exp},i})$. Because there are two observed assigned causes, we consider models with a single scalar latent variable for overlap considerations. (See Section 3.2.) We consider two factor models.

The first is a linear factor model,

$$z_{\text{line},i} \sim \mathcal{N}(0, \sigma^2) \tag{6.3}$$

$$a_{\text{mar},i} = \eta^{(1)}_{\text{mar}}\, z_{\text{line},i} + \eta^{(0)}_{\text{mar}} + \epsilon_{i,\text{mar}} \tag{6.4}$$

$$a_{\text{exp},i} = \eta^{(1)}_{\text{exp}}\, z_{\text{line},i} + \eta^{(0)}_{\text{exp}} + \varepsilon_{i,\text{exp}}, \tag{6.5}$$

where all errors are standard normal. We use variational inference to approximate posterior estimates of the substitute confounders $z_{\text{line},i}$. Then we use the predictive check to evaluate it: following Section 3.3.1, we hold out a subset of the assigned causes and using the expected log probability as the test statistic. The resulting predictive score is 0.03, which signals a model mismatch. See Figure 6.1 (a).

We next consider a quadratic factor model,

$$z_{\text{quad},i} \sim \mathcal{N}(0, \sigma^2) \tag{6.6}$$

$$a_{\text{mar},i} = \eta^{(1)}_{\text{mar}}\, z_{\text{quad},i} + \eta^{(2)}_{\text{mar}}\, z^2_{\text{quad},i} + \eta^{(0)}_{\text{mar}} + \varepsilon_{i,\text{mar}} \tag{6.7}$$

$$a_{\text{exp},i} = \eta^{(1)}_{\text{exp}}\, z_{\text{quad},i} + \eta^{(2)}_{\text{exp}}\, z^2_{\text{quad},i} + \eta^{(0)}_{\text{exp}} + \varepsilon_{i,\text{exp}}, \tag{6.8}$$

where all errors are standard normal. (In Appendix C.1, we prove that the average causal effect is identifiable with this quadratic factor model and a linear outcome model. In Appendix C.3, We revisit the causal non-identifiability of a linear factor model and a linear outcome model. We also show that the causal effects become identifiable in this linear-linear setting if we have multiple independent outcomes.) We again use variational inference and a predictive check. The resulting predictive score is 0.12, Figure 6.1 (b). This value gives the green light. We use the model's posterior estimates $\hat{z}_i \sim p_{\text{quad}}(z \mid a_i)$ to form a substitute confounder in a causal inference.

Deconfounded causal inference Using a factor model to estimate substitute confounders, we proceed with causal inference. We set the outcome model of $\mathbb{E}\left[Y(A_{\text{mar}}, A_{\text{exp}}) \mid A, Z\right]$ to be linear in a_{mar} and a_{exp}. In one form, the linear model conditions on $\hat{z}$ directly. In another it conditions on the reconstructed causes, e.g. for the quadratic model and for age,

$$a_{\text{mar},i}(\hat{z}_i) = \mathbb{E}_{\text{quad}}\left[A_{\text{mar}} \mid Z = \hat{z}_i\right]. \tag{6.9}$$

See Equation (3.18).

We use predictive checks to evaluate the outcome models. Conditioning on $\hat{z}$ gives a predictive score of 0.05; conditioning on $a(\hat{z})$ gives a predictive score of 0.18. The model with reconstructed causes is better.

If the outcome model is good and if the substitute confounder captures the true confounders then the estimated coefficients for age and exposure will be close to the true β_{mar} and β_{exp} of Equation (6.1). We emphasize that Equation (6.1) is the true mechanism of the simulated world, which the deconfounder does not have access to. The linear model we posit for $\mathbb{E}\left[Y(A_{\text{mar}}, A_{\text{exp}}) \mid A, Z\right]$ is a functional form for the expectation we are trying to estimate.

Performance We compare all combinations of factor model (linear, quadratic) and outcome-expectation model (conditional on $\hat{z}_i$ or $a(\hat{z}_i)$). Table 6.1 gives the results, reporting the total bias and variance of the estimated causal coefficients β_{mar} and β_{exp}. We compute the variance by drawing posterior samples of the substitute confounder and the resulting posterior samples of the causal coefficients.

Table 6.1 also reports the estimates if we had observed the age confounder (oracle), and the estimates if we neglect causal inference altogether and fit a regression to the confounded data. Neglecting causal inference gives biased causal estimates; observing the confounder corrects the problem.

How does the deconfounder fare? Using the deconfounder with a linear factor model yields biased causal estimates, but we predicted this peril with a predictive check. Using the deconfounder with the quadratic assignment model, which passed its predictive check, produces less biased causal estimates. (The estimate with one-dimensional z_{quad} was still biased, but the outcome check revealed this issue.)

We also use this simulation study to illustrate a few questions discussed in Section 3.4:

- **What if multiple factor models pass the check?** We fit one-dimensional, two-dimensional, and three-dimensional quadratic factor models to the causes. All three models pass the check. Table 6.1 shows that they yield estimates with similar bias. However, factor models with higher capacity in general lead to higher variance. The one-dimensional factor model is the smallest factor model that passes the check, and it achieves the best mean squared error.
- **Should we additionally condition on the observed covariates?** Table 6.1 shows that using the deconfounder, along with covariates, preserves the unbiasedness of the causal estimates but inflates the variance. (The covariates include gender, race, seat belt usage, education level, and the age of starting to smoke.) This inflation of variance is likely because the covariates are not confounders and are therefore irrelevant for the causal inference here (Grimmer et al., 2020). This demonstrates how including non-confounding covariates can affect the efficiency of causal estimation.

This study provides two takeaway messages: (1) It is crucial to check both the assignment model and the outcome model; (2) Unless a single-cause confounder believably exists, we do not need to accompany the deconfounder with other observed covariates; (3) Use the deconfounder.

	Check	Bias^2 $\times 10^{-2}$	Variance $\times 10^{-2}$	MSE $\times 10^{-2}$
No control	–	24.19	0.28	24.48
Control for age (oracle)	–	5.06	0.07	5.14
Deconfounder				
Control for 1-dim z_{line}	✗	21.51	4.48	25.99
Control for 1-dim $a(z_{\text{line}})$	✗	20.02	4.77	24.80
Control for 1-dim z_{quad}	✓	17.77	5.59	23.36
Control for 1-dim $a(z_{\text{quad}})$	✓	**15.29**	5.26	**20.51**
Control for 2-dim z_{quad}	✓	**15.08**	7.49	22.58
Control for 2-dim $a(z_{\text{quad}})$	✓	15.45	6.26	**21.71**
Control for 3-dim z_{quad}	✓	16.24	7.74	23.99
Control for 3-dim $a(z_{\text{quad}})$	✓	15.62	9.15	24.77
Deconfounder with covariates				
Control for 1-dim z_{quad}, x	✓	16.15	6.22	22.38
Control for 1-dim $a(z_{\text{quad}}), x$	✓	**15.17**	7.13	**22.30**

Table 6.1: Total bias and variance of the estimated causal coefficients β_{exp} and β_{mar}. The bias and variance are posterior bias and posterior variance as defined in Chapter 2.5 of Gustafson (2010). ("Control for xxx" means we include xxx as a covariate in the linear outcome model. The ✓ symbol indicates the factor model gives a predictive score larger than 0.1; the ✗ symbol indicates otherwise.) Both not controlling for confounders and using the deconfounder with a poor Z-model that fails the predictive check bias the causal estimates. The deconfounder with a good Z-model and a good outcome model significantly reduces the bias in causal estimates; controlling for the "reconstructed causes" $\hat{a}$ in general yields less biased estimates than the substitute confounder Z in this study. Finally, the variance of causal estimates can increase if we increase the capacity of factor models or include additional covariates.

6.2 Many causes: Genome-wide association studies

Analyzing genome-wide association studies (GWAS) is an important problem in modern genetics (Stephens and Balding, 2009; Visscher et al., 2017). The GWAS problem involves large datasets of human genotypes and a trait of interest; the goal is to determine how genetic variation is causally connected to the trait. GWAS is a problem of multiple causal inference: for each individual, the data contains a trait and hundreds of thousands of single-nucleotide polymorphisms (SNPs), measurements on various locations on the genome.

One benefit of GWAS is that biology guarantees that genes are (typically) cast in advance; they are potential causes of the trait and not the other way around. However, there are many confounders. In particular, any correlation between the SNPs could induce confounding. Suppose the value of SNP i is correlated with the value of SNP j, and SNP j is causal for the outcome. Then a naive analysis will find a connection between gene i and the outcome.

There can be many sources of correlation; common sources include population structure, i.e., how the genetic codes of individuals exhibit their ancestral populations, and lifestyle variables. We study how to use the deconfounder to analyze GWAS data. (Many existing methods to analyze GWAS data can be seen as versions of the deconfounder; see Section 3.5.)

Simulated GWAS data and the causal inference problem We put the GWAS problem into our notation. The data are tuples $(\boldsymbol{a}_i, y_i)$, where y_i is a real-valued trait and $a_{ij} \in \{0, 1, 2\}$ is the value of SNP j in individual i. (The coding denotes "unphased data," where a_{ij} codes the number of minor alleles—deviations from the norm—at location j of the genome.) As usual, our goal is to estimate aspects of the distribution of $y_i(\boldsymbol{a})$, the trait of interest as a function of a specific genotype.

We generate synthetic GWAS data. Following Song et al. (2015), we simulate genotypes $\boldsymbol{a}_{1:n}$ from an array of realistic models. These include models generated from real-world fits, models that simulate heterogeneous mixing of populations, and models that simulate a smooth spatial mixing of populations. For each model, we produce multiple datasets of genotypes.

With the individuals in hand, we next generate their traits. Still following Song et al. (2015), we generate the outcome (i.e., the trait) from a linear model,

$$y_i = \sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i. \tag{6.10}$$

To introduce further confounding effects, we group the individuals by their SNPs; the ith individual is in group c_i. (Appendix C.6 describes how individuals are grouped.) Each group is associated with a per-group intercept term λ_c and a per-group error variance σ_c, where the noise $\varepsilon_i \sim \mathcal{N}(0, \sigma_c^2)$. In this study, the group indicator of each individual is an unobserved confounder.

In Equation (6.11), SNP j is associated with a true causal coefficient β_j. We draw this coefficient from $\mathcal{N}(0, 0.5^2)$ and truncate so that majority of the coefficients are set to zero (i.e., no causal effect). Such truncation mimics the sparse causal effects that are found in the real world. Further, we study both low and high SNR settings. In low SNR settings, the SNPs contribute only a small portion (e.g. 10%) of the variance, and vice versa. Appendix C.6 details the full configurations of the simulation.

In a separate set of studies, we generate binary outcomes. They come from a generalized linear model,

$$y_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp(\sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i)}\right). \tag{6.11}$$

We will study the deconfounder for both binary or real-valued outcomes.

For each true assignment model of $\boldsymbol{a}_i$, we simulate 100 datasets of genotypes $\boldsymbol{a}_i$, causal coefficients β_j, and outcomes y_i (real and binary). For each, the causal inference problem is to infer the causal coefficients β_j from tuples $(\boldsymbol{a}_i, y_i)$. The unobserved confounding lies in the correlation structure of the SNPs and the unobserved groups. We correct it with the deconfounder.

Deconfounding GWAS We apply the deconfounder with five assignment models discussed in Section 3.1.1: probabilistic principal component analysis (PPCA), Poisson factorization (PF), Gaussian mixture models (GMMs), the three-layer deep exponential family (DEF), and logistic factor analysis (LFA); none of these models is the true assignment model. (We use 50 latent dimensions so that most pass the predictive check; for the DEF we use the structure [100, 30, 15].) We fit each model to the observed SNPs and check them with the per-individual predictive checks from Section 3.3.1.

With the fitted assignment model, we estimate the causal effects of the SNPs. For real-valued traits, we use a linear model conditional on the SNPs and the reconstructed causes $a(\hat{z})$; see Equation (3.18). Each assignment model gives a different form of $a(\hat{z})$. For the binary traits, we use logistic regression, again conditional on the SNPs and reconstructed causes.

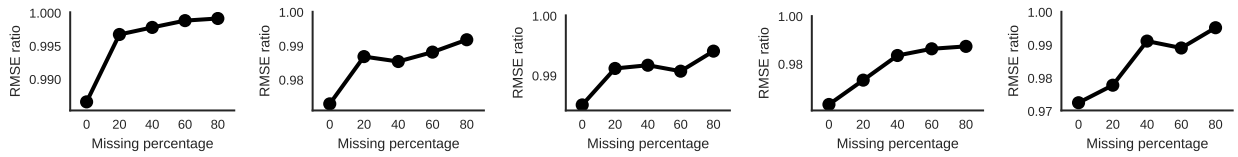
Performance We study the deconfounder for GWAS. Tables C.1 to C.10 present the full results across the 11 different configurations and both high and low signal-to-noise ratio (SNR) settings. Each table is attached to a true assignment model and reports results across different factor models of the SNPs. For each factor model, the tables report the results of the predictive check and the root mean squred error (RMSE) of the estimated causal coefficients (for real-valued and binary-valued outcomes). Tables C.1 to C.10 also report the error if we had observed the confounder and if we neglect causal inference by fitting a regression to the confounded data.

On both real and binary outcomes, the deconfounder gives good causal estimates with PPCA, PF, LFA, linear mixed models (LMMs), and DEFs: they produce lower RMSEs than blindly fitting regressions to the confounded data. (The linear mixed model does not explicitly posit an assignment model so we omit the predictive check. It can be interpreted as the deconfounder though; see Section 3.5.) Notably, the deconfounder often outperforms the regression where we include the (unobserved) confounder as a covariate under the low SNR setting; see Tables C.6 to C.9.

In general, predictive checks of the factor models reveal downstream issues with causal inference: better factor models of the assigned causes, as checked with the predictive checks, give closer-to-truth causal estimates. For example, the GMM does not perform well as a factor model of the assignments; it struggles with fitting high-dimensional data and can amplify the causal effects (see e.g. Table C.10). But checking the GMM signals this issue beforehand; the GMM constantly yields close-to-zero predictive scores in predictive checks.

Among the assignment models, the three-layer DEF almost always produces the best causal estimates. Inspired by deep neural networks, the DEF has layered latent variables; see Section 3.3.1. The DEF model of SNPs uses Gamma distributions on the latent variables (to induce sparsity) and a bank of Poisson distributions to model the observations.

The deconfounder is most challenged when the assigned SNPs are generated from a spatial model; see Tables C.5 and C.10. The spatial model produces spatially-correlated individuals; its parameter τ controls the spatial dispersion. (Consider each individual to sit in a unit square; as $\tau \to 0$, the individuals are placed closer to the corners of the unit square while when $\tau = 1$ they are distributed uniformly.) The



(a) Balding-Nichols **(b)** TGP **(c)** HDGP **(d)** PSD ($\alpha = 0.01$) **(e)** Spatial ($\tau = 0.1$)

Figure 6.2: The rmse ratio between the deconfounder with def and "No control" across simulations when only a subset of causes are unobserved. (Lower ratios means more correction.) As the percentage of observed causes decreases, the "no unobserved single-cause confounders" assumption is compromised; the deconfounder can no longer correct for all latent confounders.

five factor models—PPCA, PF, LFA, GMM, LMM, and DEF—all produce closer-to-truth causal estimates than when ignoring confounding effects. But they are farther from the truth than the estimates that use the (unobserved) confounder. Again, the predictive check hints at this issue. When the true distribution of SNPs is a spatial model, the predictive scores are generally more extreme (i.e., closer to zero).

Partially observed causes Finally, we study the situation where some assigned causes are unobserved, that is, where some of the SNPs are not measured. Recall that the deconfounder assumes that all single-cause confounders are observed. This assumption may be plausible when we measure all assigned causes but it may well be compromised when we only observe a subset—if a confounder affects multiple causes but only one of those causes is observed then the confounder becomes a single-cause confounder.

Using the simulated GWAS data, we randomly mask a percentage of the causes. We then use the deconfounder to estimate the causal effects of the remaining causes. To simplify the presentation, we focus on the DEF factor model. Figure 6.2 shows the ratio of the RMSE between the deconfounder and "no control"; a ratio closer to one indicates a more biased causal estimate. Across simulations, the RMSE ratio increases toward one as the percentage of observed causes decreases. With fewer observed causes, it becomes more likely for "no unobserved single-cause confounders" to be compromised.

Summary These studies provide three takeaway messages: (1) The deconfounder can produce closer-to-truth causal estimates, especially when we observe many assigned causes; (2) Predictive checks reveal downstream issues with causal inference, and better factor models give better causal estimates; (3) DEFs can be a handy class of factor models in the deconfounder.

6.3 Case study: How do actors boost movie earnings?

We now return to the example from Chapters 1 to 3: How much does an actor boost (or hurt) a movie's revenue? We study the deconfounder with the TMDB 5000 Movie Dataset.[†] It contains 901 actors (who appeared in at least five movies) and the revenue for the 2,828 movies they appeared in. The movies span 18 genres and 58 languages. (More than 60% of the movies are in English.) We focus on the cast and the log of the revenue. Note that this is a real-world observational data set. We no longer have ground truth of causal estimates.

The idea here is that actors are potential causes of movie earnings: some actors result in greater revenue. But confounders abound. Consider the genre of a movie; it will affect both who is in the cast and its revenue. For example, an action movie tends to cast action actors, and action movies tend to earn more than family movies. And the genre is just one possible confounder: movies in a series, directors, writers, language, and release season are all possible confounders.

We are interested in estimating the causal effects of individual actors on the revenue. The data are tuples of $(\boldsymbol{a}_i, y_i)$, where $a_{ij} \in \{0, 1\}$ is an indicator of whether actor j in movie i, and y_i is the revenue. Table 1.1 shows a snippet of the highest-earning movies in this dataset. The goal is to estimate the distribution of $Y_i(\boldsymbol{a})$, the (potential) revenue as a function of a movie cast.

Deconfounded causal inference We apply the deconfounder. We explore four assignment models: probabilistic principal component analysis (PPCA), Poisson factorization (PF), Gaussian mixture models (GMMs), and deep exponential familys (DEFs). (Each has 50 latent dimensions; the DEF has structure [50, 20, 5].) We fit each model to the observed movie casts and check the models with a predictive check on held-out data; see Section 3.3.1.

The GMM fails its check, yielding a predictive score < 0.01. The other models adequately capture patterns of actors: the checks return predictive scores of 0.12 (PPCA), 0.14 (PF), and 0.15 (DEF). These numbers give a green light to estimate how each actor affects movie earnings.

[†]https://www.kaggle.com/tmdb

With a fitted and checked assignment model, we estimate the causal effects of individual actors with a log-normal regression, conditional on the observed casts and "reconstructed casts," Equation (3.18).

Results: Predicting the revenue of uncommon movies We consider test sets of uncommon movies, where we simulate an "intervention" on the types of movies that are made. This changes the distribution of casts to be different from those in the training set.

For such data, a good causal model will provide better predictions than a purely predictive model. The reason is that predictions from a causal model will work equally well under interventions as for observational data. In contrast, a non-causal model can produce incorrect predictions if we intervene on the causes (Peters et al., 2016). This idea of invariance has also been discussed in Aldrich (1989); Dawid et al. (2010); Haavelmo (1944); Lanes (1988); Pearl (2009); Schölkopf et al. (2012) under the terms "autonomy," "modularity," and "stability."

In one test set, we hold out 10% of non-English-language movies. (Most of the movies are in English.) Table C.12 compares different models in terms of the average predictive log-likelihood. The deconfounder predicts better than both the purely predictive approach (no control) and a classical approach, where we condition on the observed (pre-treatment) covariates.

In another test set, we hold out 10% of movies from uncommon genres, i.e., those that are not comedies, action, or dramas. Table C.13 shows similar patterns of performance. The deconfounder predicts better than purely predictive models and than those that control for available confounders.

For comparison, we finally analyze a typical test set, one drawn randomly from the data. Here we expect a purely predictive method to perform well; this is the type of prediction it is designed for. Table C.11 shows the average predictive log-likelihood of the deconfounder and the purely predictive method. The deconfounder predicts slightly worse than the purely predictive method.

Exploratory analysis of actors and movies We show how to use the deconfounder to explore the data, understanding the causal value of actors and movies.[‡]

First, we examine how the coefficients of individual actors differ between a non-causal model and a deconfounded model. (In this section, we study the deconfounder with PF as the assignment model.) We explore actors with $n_j \beta_j$, their estimated coefficients scaled by the number of movies they appeared in. This quantity represents how much of the total log revenue is "explained" by actor j.

Consider the top 25 actors in both the corrected and uncorrected models. In the uncorrected model, the top actors are movie stars such as Tom Cruise, Tom Hanks, and Will Smith. Some actors, like Arnold Schwartzenegger, Robert De Niro, and Brad Pitt, appear in the top-25 uncorrected coefficients but not in the top-25 corrected coefficients. In their place, the top 25 causal actors include actors that do not appear in as many blockbusters, such as Owen Wilson, Nick Cage, Cate Blanchett, and Antonio Banderes.

Also, consider the actors whose estimated contribution improves the most from the non-causal to the causal model. The top five "most improved" actors are Stanley Tucci, Willem Dafoe, Susan Sarandon, Ben Affleck, and Christopher Walken. These (excellent) actors often appear in smaller movies.

Next, we look at how the deconfounder changes the causal estimates of movie casts. We can calculate the movie casts whose causal estimates are decreased most by the deconfounder. The "causal estimate of a cast" is the predicted revenue *without* including the term that involves the confounder; this is the portion of the predicted log revenue that is attributed to the cast.

At the top of this list are blockbuster series. Among the top 25 include all of the *X-Men* movies, all of the *Avengers* movies, and all of the *Ocean's* movies. Though unmeasured in the data, being part of a series is a confounder. It affects both the casting and the revenue of the movie: sequels must contain recurring characters and they are only made when the producers expect to profit. In capturing the correlations among casts, the deconfounder corrects for this phenomenon.

[‡]This section illustrates how to use the deconfounder to explore data. It is about these methods and the particular dataset that we studied, not a comment about the ground-truth quality of the actors involved. The author of this dissertation is a statistician, not a film critic.

7

The deconfounder on recommender systems

In this chapter, we adapt the deconfounder algorithm to perform causal inference on recommender systems.* We formulate recommendation as a multiple causal inference and develop the deconfounded recommender.

The goal of a recommender is to show its users items that they will like. Given a dataset of users' ratings, a recommender system learns the preferences of the users, predicts the users' ratings on those items they did not rate, and finally makes suggestions based on those predictions. In this chapter, we develop a causal inference approach to recommendation.

Why is recommendation a causal inference? Concretely, suppose the items are movies and the users rate movies they have seen. In prediction, the recommender system is trying to answer "How would the user rate this movie if she saw it?" However, recommending all the movies that users will like may not be the most cost-efficient strategy. Many recommendations, though costing money, will not make a difference in user behaviors. For example, users like certain movies so much that they will go see them no matter whether there is a recommendation. Therefore, we only want to recommend the movies that

*This chapter is in part based on Wang et al. (2020).

(1) if *made exposed*, the user will go see them and (2) if not, the user will not go see them. But this is a question about an *intervention*: what would the rating be if we *make* the user exposed (or not exposed) to the movie? One tenet of causal inference is that predictions under intervention are different from usual predictions.

Framing recommendation as a causal problem differs from the traditional approach. The traditional approach builds a model from observed rating data, often a matrix factorization, and then uses that model to predict unseen ratings. But this strategy only provides valid causal inferences—in the intervention sense above—if users randomly watched movies. (This is akin to a randomized clinical trial, where the treatment is exposure to a movie and the response is a rating.)

Users do not (usually) watch movies at random and, consequently, answering the causal question from observed rating data is challenging. The issue is that there may be *confounders*, variables that affect both the treatment assignments (which movies the users watch) and the outcomes (how they rate them). For example, because a user watches many movies by a particular director, they may often be recommended movies by this director and also tend to watch and like those movies. The director is a confounder that biases our inferences; it affects both which movies the user was recommended and whether they watch and like them. Compounding this issue, the confounders might be difficult (or impossible) to measure in recommender systems. Further, the theory around causal inferences says that these inferences are valid only if we have accounted for all confounders (Rosenbaum and Rubin, 1983). And, alas, whether we have indeed measured all confounders is uncheckable (Holland et al., 1985).

How can we overcome these obstacles? In this chapter, we develop the *deconfounded recommender*, extending the deconfounder to recommender systems. The deconfounded recommender builds on the two sources of information in recommendation data: which movies each user decided to watch and the user's rating for each of those movies. It posits that the two types of information come from different models—the *exposure* data comes from a model by which users discover movies to watch; the *ratings* data comes from a model by which users decide which movies they like. The ratings data entangles both types of information—users only rate movies that they see—and so classical matrix factorization is biased by the exposure model, i.e., that users are not randomly exposed to movies.

The deconfounded recommender tries to correct this bias. It first uses the exposure data to estimate a model of which movies each user is likely to consider. (In recommender systems, the exposure data is a form of "implicit" data.) It then uses this exposure model to estimate a substitute for the unobserved confounders. Second, it fits a rating model (e.g., matrix factorization) that accounts for the substitute confounders. The justification for this approach comes from Chapters 4 and 5; correlations among the considered movies provide indirect evidence for confounders. The deconfounded recommender focuses on how the exposure of each movie (i.e. one of the many causes) causally affects its observed rating (Equation (7.6)); we rely on Theorem 4.2.2 for the identification of causal parameters.

Consider a film enthusiast who mostly watches western action movies but who has also enjoyed two Korean dramas, even though non-English movies are not easily accessible in her area. A traditional recommender will infer preferences that center around westerns; the dramas carry comparatively little weight. The deconfounded recommender will also detect the preference for westerns, but it will further up-weight the preference for Korean dramas. The reason is that the history of the user indicates that she is unlikely to have been exposed to many non-English movies, and she liked the two Korean dramas that she did see. Compared to westerns, Korean dramas are likely movies that if recommended she might like, and if not recommended she might see. Consequently, when recommending from among the unwatched movies, the deconfounded recommender promotes other Korean dramas along with westerns.

Below we develop the deconfounded recommender. We empirically study it on both simulated data, where we control the amount of confounding, and real data, about shopping and movies. Compared to existing approaches, its performance is more robust to unobserved confounding; it predicts the ratings better and consistently improves recommendation.

7.1 The deconfounded recommender

We frame recommendation as a causal inference and develop the deconfounded recommender.

Matrix factorization and potential outcomes We first set up notation. Denote a_{ui} as the indicator of whether user u rated movie i. Let $y_{ui}(1)$ be the rating that user u would give movie i if she watches it. This rating is only observed if the user u watched and rated the movie i; otherwise it is unobserved. Similarly define $y_{ui}(0)$ to be the rating of user u on movie i if she does not see the movie. (We often "observe" $y_{ui}(0) = 0$ in recommendation data; unrated movie entries are filled with zeros.) The pair $(y_{ui}(0), y_{ui}(1))$ is the *potential outcomes* notation in the Rubin causal model (Imbens and Rubin, 2015b; Rubin, 1974, 2005), where watching a movie is a "treatment" and a user's rating of the movie is an "outcome."

A recommender system observes users' ratings of movies. We can think of these observations as two datasets. One dataset contains (binary) exposures, $\{a_{ui}, u = 1, \ldots, U, i = 1, \ldots, I\}$. It indicates who watched what. The other dataset contains the ratings for movies that users watched, $\{y_{ui}(a_{ui})$ for (u, i) such that $a_{ui} = 1\}$.

The goal of the recommender is to suggest movies its users will like. It first estimates $y_{ui}(1)$ for user-movie pairs with $a_{ui} = 0$; that is, it predicts each user's ratings for their unseen movies. It then uses these estimates to suggest movies to users. Note $y_{ui}(1)$ is a prediction under intervention: "What would the rating be if user u was made to watch movie i?"

To form the prediction of $y_{ui}(1)$, we recast matrix factorization in potential outcomes. First set up an *outcome model*,

$$y_{ui}(a) = \theta_u^\top \beta_i \cdot a + \epsilon_{ui}, \quad \epsilon_{ui} \sim \mathcal{N}(0, \sigma^2). \tag{7.1}$$

When $a = 1$ (i.e., user u watches movie i), this model says that the rating comes from a Gaussian distribution whose mean combines user preferences θ_u and item attributes β_i. When $a = 0$, the "rating" is a zero-mean Gaussian.

Fitting Equation (7.1) to the observed data recovers classical probabilistic matrix factorization (Mnih and Salakhutdinov, 2008). Its log-likelihood only involves observed ratings; it ignores the unexposed items. The fitted model can then predict $\mathbb{E}[y_{ui}(1)] = \theta_u^\top \beta_i$ for every (unwatched) user-movie pair. These predictions suggest movies that users would like.

Classical causal inference and adjusting for confounders in recommendation But matrix factorization does not provide an unbiased causal inference of $y_{ui}(1)$. The theory around potential outcomes says we can only estimate $y_{ui}(1)$ if we assume *unconfoundedness*. For all users u, unconfoundedness requires $\{\boldsymbol{y}_u(0), \boldsymbol{y}_u(1)\} \perp\!\!\!\perp \boldsymbol{a}_u$, where $\boldsymbol{y}_u(a) = (y_{u1}(a), \ldots, y_{uI}(a))$ and $\boldsymbol{a}_u = (a_{u1}, \ldots, a_{uI})$. In words, the vector of movies a user watches $\boldsymbol{a}_u$ is independent of how she would rate them if she watched them all $\boldsymbol{y}_u(1)$ (and if she watched none $\boldsymbol{y}_u(0)$).

Unconfoundedness does not hold for $\boldsymbol{y}_u(1)$—the process by which users find movies is not independent of how they rate them. Practically, this violation biases the estimates of user preferences θ_u: movies that u is not likely to see are down-weighted and vice versa. Again consider the American user who enjoyed two Korean dramas and rated them highly. Because she has only two high ratings of Korean dramas in the data, her preference for Korean dramas carries less weight than her other ratings; it is biased downward. Biased estimates of preferences lead to biased predictions of ratings.

When unconfoundedness does not hold, classical causal inference asks us to measure and control for confounders (Pearl, 2009; Rubin, 2005). These are variables that affect both the exposure and the ratings. Consider the location of a user as an example. It affects both which movies they are exposed to and (perhaps) what kinds of movies they tend to like.

Suppose we measured these per-user confounders w_u; they satisfy $\{\boldsymbol{y}_u(0), \boldsymbol{y}_u(1)\} \perp\!\!\!\perp \boldsymbol{a}_u \mid w_u$. Classical causal inference controls for them in the outcome model, $y_{ui}(a) = \theta_u^\top \beta_i \cdot a + \eta^\top w_u + \epsilon_{ui}, \epsilon_{ui} \sim \mathcal{N}(0, \sigma^2)$. However, this solution requires we measure *all* confounders. This assumption is known as *weak unconfoundedness*, as is defined by Assumption 2.3.1 in Chapter 2. In causal graphical models, this requirement is equivalent to "no open backdoor paths" (Pearl, 2009). Unfortunately, it is untestable (Holland et al., 1985).

The deconfounded recommender We now develop the deconfounded recommender. It leverages the dependencies among the exposure ("which movies the users watch") as indirect evidence for unobserved confounders. It uses a model of the exposure to construct a substitute confounder; it then conditions on the substitute when modeling the ratings.

The key idea is that causal inference for recommendation systems is a *multiple causal inference* problem: there are multiple treatments. Each user's binary exposure to each movie a_{ui} is a treatment; thus there are I treatments for each user. The vector of ratings $\boldsymbol{y}_u(1)$ is the outcome; this is an I-vector, which is partially observed. The multiplicity of treatments enables causal inference with unobserved confounders (Wang and Blei, 2019a).

The first step is to fit a model to the exposure data. We use Poisson factorization (PF) model (Gopalan et al., 2015). PF assumes the data come from the following process,

$$a_{ui} \mid \pi_u, \lambda_i \sim \text{Poisson}(\pi_u^\top \lambda_i), \qquad \forall u, i, \tag{7.2}$$

where both $\pi_u \overset{iid}{\sim} \text{Gamma}(c_1, c_2)$ and $\lambda_i \overset{iid}{\sim} \text{Gamma}(c_3, c_4)$ are nonnegative K-vectors. The user factor π_u captures user preferences (in picking what movies to watch) and the item vector λ_i captures item attributes. PF is a scalable variant of nonnegative factorization and is especially suited to binary data (Gopalan et al., 2015). It is fit with coordinate ascent variational inference.[†]

With a fitted PF model, the deconfounded recommender computes a substitute for unobserved confounders. It reconstructs the exposure matrix $\hat{a}$ from the PF fit,

$$\hat{a}_{ui} = \mathbb{E}_{\text{PF}}[\pi_u^\top \lambda_i \mid \boldsymbol{a}], \tag{7.3}$$

where $\boldsymbol{a}$ is the observed exposure for all users, and the expectation is taken over the posteriors computed from the PF model. This is the posterior predictive mean of $\pi_u^\top \lambda_i$, which serves as a substitute confounder (Wang and Blei, 2019a).

Finally, the deconfounded recommender posits an outcome model conditional on the substitute confounders $\hat{a}$,

$$y_{ui}(a) = \theta_u^\top \beta_i \cdot a + \gamma_u \cdot \hat{a}_{ui} + \epsilon_{ui}, \quad \epsilon_{ui} \sim \mathcal{N}(0, \sigma^2), \tag{7.4}$$

[†]The Bernoulli distribution is more natural to model binary exposure, but PF is more computationally efficient and several precedents use a Poisson to model binary data (Gopalan et al., 2015, 2014). PF scales linearly with the number of *nonzero* entries in the exposure matrix $\{a_{ui}\}_{U \times I}$ while Bernoulli scales with the number of *all* entries. Further, the Poisson distribution closely approximates the Bernoulli when the exposure matrix $\{a_{ui}\}_{U \times I}$ is sparse (DeGroot and Schervish, 2012). Finally, PF can also model non-binary *count* exposures: e.g., PF can model exposures that count how many times a user has been exposed to an item.

where γ_u is a user-specific coefficient that describes how much the substitute confounder $\hat{a}$ contributes to the ratings. The deconfounded recommender fits this outcome model to the observed data; it infers $\theta_u, \beta_i, \gamma_u$, via MAP estimation. The coefficients θ_u, β_i in Equation (7.4) are fit only with the observed user ratings (i.e., $a_{ui} = 1$) because $a_{ui} = 0$ zeroes out the term that involves them; in contrast, the coefficient γ_u is fit to all movies ($a_{ui} = 0$ and $a_{ui} = 1$) since $\hat{a}_{ui}$ is always non-zero.

To form recommendations, the deconfounded recommender calculates all the potential ratings $y_{ui}(1)$ with the fitted $\hat{\theta}_u, \hat{\beta}_i, \hat{\gamma}_u$. It then orders the potential ratings of the unseen movies. These are causal recommendations. Algorithm 8.1 provides the algorithm for forming recommendations with the deconfounded recommender.

Why does it work? Poisson factorization (PF) learns a per-user latent variable π_u from the exposure matrix a_{ui}, and we take π_u as a substitute confounder. What justifies this approach is that PF admits a special conditional independence structure: conditional on π_u, the treatments a_{ui} are independent (Equation (7.2)). If the exposure model PF fits the data well, then the per-user latent variable π_u (or functions of it, like $\hat{a}_{ui}$) captures multi-treatment confounders, i.e., variables that correlate with multiple exposures and the ratings vector (Lemma 3 of (Wang and Blei, 2019a)). We note that the true confounding mechanism does not need to coincide with PF and nor does the real confounder need to coincide with π_u. Rather, PF produces a substitute confounder that is sufficient to debias confounding.

Beyond probabilistic matrix factorization The deconfounder involves two models, one for exposure and one for outcome. We have introduced PF as the exposure model and probabilistic matrix factorization (Mnih and Salakhutdinov, 2008) as the outcome model.Focusing on PF as the exposure model, we extend the deconfounded recommender to general outcome models.

We start with a general form of matrix factorization,

$$y_{ui}(a) \sim p(\cdot \mid m(\theta_u^\top \beta_i, a), v(\theta_u^\top \beta_i, a)), \tag{7.5}$$

where $m(\theta_u^\top \beta_i, a)$ characterizes the mean and $v(\theta_u^\top \beta_i, a)$ the variance of the ratings $y_{ui}(a)$. This form encompasses many factorization models. Probabilistic matrix factorization (Mnih and Salakhutdinov, 2008) is

$$m(\theta_u^\top \beta_i, a) = a \cdot \theta_u^\top \beta_i, \quad v(\theta_u^\top \beta_i, a) = \sigma^2,$$

and $p(\cdot)$ is the Gaussian distribution. Weighted matrix factorization (Hu et al., 2008) also involves a Gaussian p, but its variance changes based on whether a user has seen the movie:

$$m(\theta_u^\top \beta_i, a) = \theta_u^\top \beta_i, \quad v(\theta_u^\top \beta_i, a) = \sigma_a^2,$$

where $\sigma_0^2 = \alpha\sigma_1^2$. This model leads us to downweight the zeros; we are less confident about the zero ratings. Poisson matrix factorization as an outcome model (Gopalan et al., 2015) only takes in a mean parameter

$$m(\theta_u^\top \beta_i, a) = \theta_u^\top \beta_i$$

and set $p(\cdot)$ to be the Poisson distribution.

With the general matrix factorization of Equation (7.5), the deconfounded recommender fits an augmented outcome model M_Y. This outcome model M_Y includes the substitute confounder as a covariate,

$$y_{ui}(a) \sim p(\cdot \mid m(\theta_u^\top \beta_i, a) + \gamma_u \hat{a}_{ui} + \beta_0, v(\theta_u^\top \beta_i, a)). \tag{7.6}$$

Notice the parameter γ_u is a user-specific coefficient; for each user, it characterizes how much the substitute confounder $\hat{a}$ contributes to the ratings. Note the deconfounded recommender also includes an intercept β_0. These deconfounded outcome model can be fit by *maximum a posteriori* estimation.

Algorithm 7.1: The Deconfounded Recommender

Input: a dataset of exposures and ratings $\{(a_{ui}, y_{ui}(a_{ui}))\}_{u,i},\ i = 1, \ldots, I, u = 1, \ldots, U$

Output: the potential outcome given treatment $\hat{y}_{ui}(1)$

1. Fit PF to the exposures $\{a_{ui}\}_{u,i}$ from Equation (7.2)
2. Compute substitute confounders $\{\hat{a}_{ui}\}_{u,i}$ from Equation (7.3)
3. Fit the outcome model $\{(a_{ui}, y_{ui}(a_{ui}))\}_{u,i}$ from Equation (7.6)
4. Estimate all potential ratings $y_{ui}(1)$ with the fitted outcome model (Equation (7.6))

It solves

$$\begin{aligned}
&\hat{\theta}_u, \hat{\beta}_i, \hat{\gamma}_u, \hat{\beta}_0 \\
&= \arg\max \sum_{u=1}^{U} \sum_{i=1}^{I} \log p(y_{ui}; m(\theta_u^\top \beta_i, a_{ui}) + \gamma_u \hat{a}_{ui} + \beta_0, \\
&\quad v(\theta_u^\top \beta_i, a_{ui})) + \textstyle\sum_{u=1}^{U} \log p(\theta_u) + \sum_{i=1}^{I} \log p(\beta_i) \\
&\quad + \sum_{u=1}^{U} \log p(\gamma_u) + \log p(\beta_0),
\end{aligned}$$

where $p(\theta_u)$, $p(\beta_i)$, $p(\gamma_u)$, and $p(\beta_0)$ are priors of the latent variables.

To form recommendations, the deconfounded recommender predicts all of the potential ratings, $y_{ui}(1)$. For an existing user u, it computes the potential ratings from the fitted outcome model,

$$\hat{y}_{ui}(1) = m(\hat{\theta}_u^\top \hat{\beta}_i, 1) + \hat{\gamma}_u \cdot \hat{a}_{ui} + \hat{\beta}_0. \tag{7.7}$$

For a new user u' with only a few ratings, it fixes the item vectors λ_i and β_i, and compute user vectors for the new user: it fits $\pi_{u'}$, $\theta_{u'}$, and $\gamma_{u'}$ from the exposure and the ratings of this new user u'. It finally computes the prediction,

$$\hat{y}_{u'i}(1) = m(\hat{\theta}_{u'}^\top \hat{\beta}_i, 1) + \hat{\gamma}_{u'} \cdot \hat{\pi}_{u'}^\top \lambda_i + \hat{\beta}_0. \tag{7.8}$$

The deconfounded recommender ranks all the items for each user based on $\hat{y}_{ui}(1), i = 1, \ldots, I$, and recommends highly ranked items.

7.2 Empirical Studies

We study the deconfounded recommender on simulated and real datasets. We examine its recommendation performance and compare it with existing recommendation algorithms. We find that the deconfounded recommender is more robust to unobserved confounding than existing approaches; it predicts the ratings better and consistently improves recommendation.

7.2.1 Evaluation of causal recommendation models

We first describe how we evaluate the recommender. Recommender systems are trained on a set of user ratings and tested on heldout ratings (or exposures). The goal is to evaluate the recommender with the test sets. Traditionally, we evaluate the accuracy (e.g. mean squared error) of the predicted ratings. Or we compute ranking metrics: were the items with high ratings also ranked high in our predictions?

However, causal recommendation models pose unique challenges for evaluation. In causal inference, we need to evaluate how a model performs across all potential outcomes,

$$\text{err}_{\text{cau}} = \frac{1}{U} \sum_{u=1}^{U} \ell(\{\hat{y}_{ui}\}_{i \in \{1,\dots,I\}}, \{y_{ui}(1)\}_{i \in \{1,\dots,I\}}), \tag{7.9}$$

where ℓ is a loss function, such as mean squared error (MSE) or normalized discounted cumulative gain (NDCG). The challenge is that we don't observe all potential outcomes $y_{ui}(1)$.

Which test sets can we use for evaluating causal recommendation models? One option is to generate a test set by randomly splitting the data; we call this a "regular test set." However, evaluation on the regular test set gives a biased estimate of err_{cau}; it emphasizes popular items and active users.

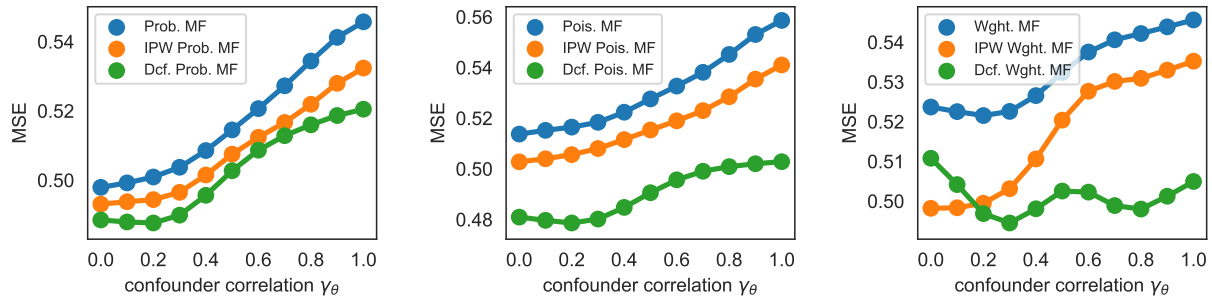
An (expensive) solution is to measure a randomized test set. Randomly select a subset $\mathcal{I}_u$ from all items and ask the users to interact and rate all of them. Then compute the average loss across users,

$$\text{err}_{\text{rand}} = \frac{1}{U} \sum_{u=1}^{U} \ell(\{\hat{y}_{ui}\}_{i \in \mathcal{I}_u}, \{y_{ui}(1))\}_{i \in \mathcal{I}_u}). \tag{7.10}$$

Equation (7.10) is an unbiased estimate of the average across all items in Equation (7.9); it tests the recommender's ability to answer the causal question. Two available datasets that include such random test sets are the Yahoo! R3 dataset (Marlin and Zemel, 2009) and the coat shopping dataset (Schnabel et al., 2016). We also create random test sets in simulation studies.

However, a randomized test set is often difficult to obtain. In this case, our solution is to evaluate the average per-item predictive accuracy on a "regular test set." For each item, we compute the MSE of all the ratings on this movie; we then average the MSEs of all items. While popular items receive more ratings, this average per-item predictive accuracy treats all items equally, popular or unpopular. We use this metric to evaluate the deconfounded recommender on Movielens 100k and Movielens 1M datasets.

7.2.2 Simulation studies



(a) Probabilistic MF **(b)** Poisson MF **(c)** Weighted MF

Figure 7.1: Varying confounder correlation γ_θ from 0.0 to 1.0 ($\gamma_y = 3.0$). The rating predictions of the deconfounded recommender (green) is more robust to unobserved confounding than its classical counterpart (blue) and the existing causal approach, IPW MF (Schnabel et al., 2016) (orange). (Lower is better.)

We study the deconfounded recommender on simulated datasets. We simulate movie ratings for $U = 5,000$ users and $I = 5,000$ items, where effect of preferences on the rating is confounded.

Simulation setup We simulate a K-vector confounder for each user $c_u \sim \text{Gamma}_K(0.3, 0.3)$ and a K-vector of attributes for each item $\beta_i \sim \text{Gamma}_K(0.3, 0.3)$. We then simulate the user preference K-vectors θ_u conditional on the confounders,

$$\theta_u \sim \gamma_\theta \cdot c_u + (1 - \gamma_\theta) \cdot \text{Gamma}_K(0.3, 0.3).$$

The constant $\gamma_\theta \in [0, 1]$ controls the exposure-confounder correlation; higher values imply stronger confounding.

We next simulate the binary exposures $a_{ui} \in \{0, 1\}$, the ratings for all users watching all movies $y_{ui}(1) \in \{1, 2, 3, 4, 5\}$, and calculate the observed ratings y_{ui}. The exposures and ratings are both simulated from truncated Poisson distributions, and the observed ratings mask the ratings by the exposure,

$$\begin{aligned} a_{ui} &\sim \min(\text{Poisson}(c_u^\top \beta_i), 1), \\ y_{ui}(1) &\sim \min(1 + \text{Poisson}((\theta_u + \gamma_y \cdot c_u)^\top \beta_i), 5), \\ y_{ui} &= a_{ui} \cdot y_{ui}(1). \end{aligned}$$

The constant $\gamma_y \geq 0$ controls how much the confounder c_u affects the outcome; higher values imply stronger confounding.

Competing methods We compare the deconfounded recommender to baseline methods. One set of baselines are the classical counterparts of the deconfounded recommender. We explore probabilistic matrix factorization (Mnih and Salakhutdinov, 2008), Poisson matrix factorization (Gopalan et al., 2015), and weighted matrix factorization (Hu et al., 2008); see Section 7.1 for details of these baseline models. We additionally compare to inverse propensity weighting (IPW) matrix factorization (Schnabel et al., 2016), which also handles selection bias in observational recommendation data.

Results Figure 7.1 shows how an unobserved confounder high correlated with exposures can affect rating predictions. (Its effects on ranking quality is in Appendix D.1.) Although the performances of all algorithms degrade as the unobserved confounding increases, the deconfounded recommender is more robust. It leads to smaller MSEs in rating prediction and higher NDCGs in recommendation than its classical counterparts (Probabilistic/Poisson/Weighted Matrix Factorization) and the existing causal approach (Inverse Propensity Weighted Probabilistic/Poisson/Weighted Matrix Factorization) (Schnabel et al., 2016).

7.2.3 Case studies I: The deconfounded recommender on random test sets

We next study the deconfounded recommender on two real datasets: Yahoo! R3 (Marlin and Zemel, 2009) and coat shopping (Schnabel et al., 2016). Both datasets are comprised of an observational training set and a random test set. The training set comes from users rating user-selected items; the random test set comes from the recommender system asking its users to rate randomly selected items. The latter enables us to evaluate how different recommendation models predict *potential outcomes*: what would the rating be if we *make* a user watch and rate a movie?

Datasets Yahoo! R3 (Marlin and Zemel, 2009) contains user-song ratings. The training set contains over 300K user-selected ratings from 15400 users on 1000 items. Its random test set contains 5400 users who were asked to rate 10 randomly chosen songs. The coat shopping dataset (Schnabel et al., 2016) contains user-coat ratings. The training set contains 290 users. Each user supplies 24 user-selected ratings among 300 items. Its random test contains ratings for 16 randomly selected coats per user.

Evaluation metrics We use the recommenders for two types of prediction: weak generalization and strong generalization (Marlin, 2004). Weak generalization predicts the preferences of existing users in the training set on their unseen movies. Strong generalization predicts preferences of new users—users not in the training set—on their unseen movies. Based on the predictions, we rank the items with nonzero ratings. To evaluate recommendation performance, we report three standard measures: NDCG, recall, and MSE. See Appendix D.2 for formal definitions.

Experimental setup For each dataset, we randomly split 80/20 the training set into training/validation sets. We leave the random test set intact. Across all experiments, we use the validation set to select the best hyperparameters for the recommendation models. We choose the hyperparameters that yield the best validation log NDCG. The latent dimension is chosen from $\{1, 2, 5, 10, 20, 50, 100\}$.

The deconfounded recommender has two components: the treatment assignment model and the outcome model. We always use PF as the treatment assignment model. The hyperparameters of both models are chosen together based on the same validation set as above. The best hyperparameters are those that yield the best validation NDCG.

	Yahoo! R3			Coat		
	NDCG	Recall@5	MSE	NDCG	Recall@5	MSE
Probabilistic MF [192]	0.772	0.540	1.963	0.728	0.552	1.422
IPW Probabilistic MF [241]	0.791	0.603	1.893	0.732	0.547	1.358
Deconfounded Probabilistic MF	**0.819**	**0.640**	**1.768**	**0.743**	**0.569**	**1.341**
Poisson MF [103]	0.789	0.564	1.539	0.713	0.451	1.657
IPW Poisson MF [241]	0.792	0.564	1.513	0.693	0.458	**1.631**
Deconfounded Poisson MF	**0.802**	**0.610**	**1.447**	**0.743**	**0.521**	1.657
Weighted MF [129]	0.820	0.639	2.047	0.738	0.560	1.658
IPW Weighted MF [241]	0.804	0.592	1.845	0.681	0.554	1.612
Deconfounded Weighted MF	**0.823**	**0.645**	**1.658**	**0.744**	**0.581**	**1.569**

Table 7.1: Recommendation on random test sets for existing users (weak generalization). The deconfounded recommender improves recommendation over classical approaches and the existing causal approach (Schnabel et al., 2016). (Higher is better for NDCG and Recall@5; lower is better for MSE.)

	Yahoo! R3			Coat		
	NDCG	Recall@5	MSE	NDCG	Recall@5	MSE
Probabilistic MF [192]	0.802	0.792	2.307	0.833	0.737	**1.490**
IPW Probabilistic MF [241]	0.818	0.827	2.625	0.811	0.724	1.587
Deconfounded Probabilistic MF	**0.824**	**0.829**	**2.244**	**0.847**	**0.808**	1.540
Poisson MF [103]	0.765	0.752	1.913	0.764	0.671	1.889
IPW Poisson MF [241]	0.769	0.761	**1.876**	0.769	**0.678**	1.817
Deconfounded Poisson MF	**0.774**	**0.769**	1.881	**0.772**	**0.678**	**1.811**
Weighted MF [129]	0.809	0.793	1.904	0.850	0.814	2.859
IPW Weighted MF [241]	0.786	0.788	1.883	0.837	0.800	2.477
Deconfounded Weighted MF	**0.820**	**0.818**	**1.644**	**0.854**	**0.829**	**2.421**

Table 7.2: Recommendation on random test sets for new users (strong generalization). The deconfounded recommender improves recommendation over classical approaches and the existing causal approach (Schnabel et al., 2016). (Higher is better for NDCG and Recall@5; lower is better for MSE.)

Results Tables 7.1 and 7.2 show the recommendation performance of the deconfounded recommender and its competitors. Across the three metrics and the two datasets, the deconfounded recommender outperforms its classical counterpart for both weak and strong generalization: it produces better item rankings and improves retrieval quality; its predicted ratings are also more accurate. The deconfounded recommender also outperforms the IPW matrix factorization (Schnabel et al., 2016), which is the main existing approach that targets selection bias in recommendation systems. These results show that the deconfounded recommender produces more accurate predictions of user preferences by accommodating unobserved confounders in item exposures.[‡]

[‡]The MSEs of IPW MF in Table 7.1 differ from those reported in Schnabel et al. (2016), because Schnabel et al. (2016) relies on a random validation set in training, which is not available here.

	Movielens 100K		Movielens 1M	
	MSE	MAE	MSE	MAE
Probabilistic MF [192]	2.926	1.425	2.774	1.321
IPW Probabilistic MF [241]	2.609	1.275	2.714	1.303
Deconfounded Probabilistic MF	**2.554**	**1.260**	**2.699**	**1.299**
Poisson MF [103]	3.374	1.475	2.357	1.305
IPW Poisson MF [241]	3.439	1.480	**2.196**	**1.220**
Deconfounded Poisson MF	**3.268**	**1.454**	2.325	1.295
Weighted MF [129]	2.359	1.219	3.558	1.516
IPW Weighted MF [241]	2.344	1.198	2.872	1.363
Deconfounded Weighted MF	**2.101**	**1.138**	**2.864**	**1.360**

Table 7.3: Average per-item predictive accuracy on heldout ratings. The deconfounded recommender leads to lower MSEs and MAEs on predictions over all items; it outperforms classical approaches and the existing causal approach (Schnabel et al., 2016). (Lower is better.)

7.2.4 Case studies II: The deconfounded recommender on regular test sets

We study the deconfounded recommender on two MovieLens datasets: ML100k and ML1M.[§] These datasets only involve observational data; they do not contain randomized test sets. We focus on how well we predict on *all* items, popular or unpopular.

Datasets The Movielens 100k dataset contains 100,000 ratings from 1,000 users on 1,700 movies. The Movielens 1M dataset contains 1 million ratings from 6,000 users on 4,000 movies.

Experimental setup and performance measures We employ the same experimental protocols for hyperparameter selection as before. For each dataset, we randomly split the training set into training/validation/test sets with 60/20/20 proportions. We measure the recommendation performance by the average per-item predictive accuracy, which equally treats popular and unpopular items; see Section 7.2.1 for details.

[§]http://grouplens.org/datasets/movielens/

Results Table 7.3 presents the recommendation performance of the deconfounded recommender and their classical counterpart on average per-item MSEs and MAEs. Across the two metrics and two datasets, the deconfounded recommender leads to lower MSEs and MAEs on predictions over all items than classical approaches and the existing causal approach, IPW MF (Schnabel et al., 2016). Instead of focusing on popular items, the deconfounded recommender targets accurate predictions on *all* items. Hence it improves the prediction quality across all items.

7.3 Related work

This chapter draws on several threads of previous research in recommendation algorithms.

The first is on evaluating recommendation algorithms via biased data. It is mostly explored in the multi-armed bandit literature (Li et al., 2015, 2010; Vanchinathan et al., 2014; Zhao et al., 2013). These works focus on online learning and rely on importance sampling. Here we consider an orthogonal problem. We reason about user preferences, rather than recommendation algorithms, and we use offline learning and parametric models.

The second thread is around the missing-not-completely-at-random assumption in recommendation algorithms. Marlin and Zemel (2009) studied the effect of violating this assumption in ratings. Similar to our exposure model, they posit an explicit missingness model that leads to improvements in predicting ratings. Later, other researchers proposed different rating models to accommodate this violated assumption (Bonner and Vasile, 2018; Hernández-Lobato et al., 2014; Liang et al., 2016; Ling et al., 2012; Smith and Elkan, 2004; Srebro and Salakhutdinov, 2010). In this chapter, we take an explicitly causal view of the problem. While violating the missing-not-completely-at-random assumption is one form of confounding bias (Ding et al., 2018), the explicit causal view opens up the door to other recent debiasing tools, such as the deconfounder. It also articulates the rationale of such adjustments: By modeling which movies users tend to watch, we avoid recommending the movies that users will watch anyway even without a recommendation. Rather, we only want to recommend movies that if recommender the user will watch and otherwise not.

Finally, the recent work of Schnabel et al. (2016) also adapted causal inference—IPW, in particular—to address missingness. Their propensity models rely on either observed ratings of a randomized trial or externally observed user and item covariates. In contrast, our work relies solely on the observed ratings: we do not require ratings from a gold-standard randomized exposure nor do we use external covariates. In Section 7.2, we show that the deconfounded recommender provides better recommendations than Schnabel et al. (2016).

7.4 Discussion

We develop the *deconfounded recommender*, a strategy to use classical recommendation models for causal predictions: how would a user rate a recommended movie? The deconfounded recommender uses Poisson factorization to infer confounders in treatment assignments; it then augments common recommendation models to correct for confounding bias. The deconfounded recommender improves recommendation performance and rating predictions; it is also more robust to unobserved confounding in user exposures.

8

Multiple control outcomes and the control-outcome deconfounder

In this chapter, we consider multiple causal inference with a single cause and multiple outcomes. In particular, many of the outcomes are negative control outcomes that cannot be affected by the cause. We then develop the *control-outcome deconfounder*, an algorithm that corrects for unobserved confounders using multiple negative control outcomes. We describe the causal identification assumptions required for this algorithm to produce unbiased causal estimates. We illustrate how it generalizes the method of synthetic controls (Abadie et al., 2010, 2015; Abadie and Gardeazabal, 2003), expanding its scope to nonlinear settings and non-panel data.

Suppose we want to evaluate the effectiveness of a movie recommendation system. How does the recommendation of a movie affect how long a user spends watching it? To answer this question, we collect data about n users. For each one, the data contains whether the user was recommended the movie *Frozen*, $A_i \in \{0, 1\}$, and how long she spent watching it, Y_i.

One metric of the system's quality is how much more time do users spend watching *Frozen* if the system decides to recommend it to them. But estimating this metric is a causal inference. Using potential outcomes notation (Imbens and Rubin, 2015a; Rubin, 1974, 2005), let $Y_i(1)$ be user i's time with *Frozen* if she is recommended it and $Y_i(0)$ be her time with *Frozen* if she is not recommended it. The metric is $\mathbb{E}\left[Y_i(1) - Y_i(0) \mid A_i = 1\right]$, the difference a recommendation makes given that the system decided to make a recommendation. This quantity is called the average treatment effect on the treated (ATT). (A full picture of the recommendation system estimates the ATT for each movie and averages the results.)

The ATT decomposes as $\mathbb{E}\left[Y_i(1) \mid A_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid A_i = 1\right]$. The first term is easy to estimate from the data, as we always observe $Y_i(1)$ when $A_i = 1$. But the second term is a counterfactual: it is the expected time a user would have (counterfactually) watched *Frozen* even if she was not recommended it, given that the user was (in fact) recommended it. A naive estimate of $\mathbb{E}\left[Y_i(0) \mid A_i = 1\right]$ is the average time spent on *Frozen* for users who did not receive the recommendation. But this estimate is only unbiased if the treatment is randomized, which is usually not the case. (And especially not for recommendation systems—we don't want a randomized recommendation system.)

How can we estimate the counterfactual? A classical solution estimates it by correcting for confounders, variables that affect the treatment and the outcome. One confounder is the user's preferences for movie genres, such as drama, children's movies, and action movies. Each user's preferences U_i affects both the output of the recommendation system and also whether the user likes *Frozen*. (If the recommendation system is good, then it will be designed to capture the users' preferences.) Suppose the preferences U_i suffice to estimate causal effects, as when there are no other confounders. Then $\mathbb{E}\left[Y_i(0) \mid A_i = 1\right] = \mathbb{E}_{U_i \mid A_i = 1}\left[\mathbb{E}\left[Y_i(0) \mid U_i, A_i = 0\right]\right]$, and everything on the right-hand side is estimable if the preferences U_i are also observed. But this estimate relies on observing sufficient confounders. When some are unobserved, the estimate is biased.

This chapter shows how to use external data to control for some of the unobserved confounders. Suppose we also observe how much time users spend watching other movies, such as *Star Wars*, *Brazil*, and *High School Musical*; and assume these variables are not affected by receiving a recommendation for *Frozen*. These are *negative control outcomes*, variables for which the treatment is known to have no effect

(Lipsitch et al., 2010; Rosenbaum, 1989). The method in this chapter uses negative control outcomes—more briefly, control outcomes—to form a substitute for unobserved confounders. The intuition is that the control outcomes contain indirect evidence for some of the unobserved confounders. Intuitively, information about user behavior on other movies can provide evidence for a user's preferences.

The algorithm is called the *control-outcome deconfounder*. It constructs a substitute confounder by fitting a Bayesian factor model (e.g. probabilistic principal component analysis (Tipping and Bishop, 1999) or a mixture model (McLachlan and Basford, 1988)) to all the outcomes—both control and non-control—and then uses these substitute confounders in a downstream causal inference. We prove that the substitute confounder can correct for *multi-outcome confounders*, those that affect the treatment, the study outcome, and the control outcomes. The control-outcome deconfounder mirrors the deconfounder but operates in a setting with a single treatment, a study outcome, and many control outcomes. And we show how the seminal method of synthetic controls (Abadie et al., 2010, 2015; Abadie and Gardeazabal, 2003) can be viewed as a control-outcome deconfounder, expanding the settings where synthetic controls can be used.

The control-outcome deconfounder requires several assumptions to provide unbiased estimates, and these are discussed at length below. The main requirement is a causal inference problem with many control outcomes. Each problem involves units of study, a treatment of interest, an outcome of interest, many control outcomes, and the possibility of an unobserved multi-outcome confounder. Here are some examples.

- **Recommendation systems.** This is the example we discussed. The units are users of a recommendation system; the treatment is whether a user is recommended a movie; the outcome is how much time she spent watching that movie. The goal is to estimate $\mathbb{E}\left[Y_i(1) - Y_i(0) \mid A_i = 1\right]$. It measures how effective is the recommendation system.

  The control outcomes are users' time watching other movies; they are not affected by the recommendation of the movie under study. These control outcomes can help correct for multi-outcome confounders such as user preferences. User preferences are multi-outcome because they affect whether the movie was recommended, how much time the user spent on it, and how much time the user spent on other movies.

- **Panel data for evaluating social policy.** This problem is inspired by the original synthetic controls literature (Abadie et al., 2015). The units are states; the treatment is whether a state bans smoking in 2020; the outcome is the state's increase in the popularity of vaping. The goal is to estimate $\mathbb{E}\left[Y_i(1) - Y_i(0) \mid A_i = 1\right]$: for states that ban smoking, does it increase the popularity of vaping?

  Control outcomes are each state's change in vaping popularity for other years, from 2019 and back; these changes are not affected by the 2020 policy to ban vaping. The control outcomes can help correct for confounders such as each state's budget for anti-smoking education. This budget affects the decision to ban smoking (through state-wide public sentiment), the 2020 change in vaping popularity, and the previous changes in vaping popularity.

  Using previous time points as control outcomes, we need to assume that there is no "anticipatory effect" (Malani and Reif, 2010, 2015), that the coming ban in 2020 does not affect vaping popularity in 2019. It also assumes that the changes in popularity from year to year are conditionally independent, given the unobserved confounders.

- **Supermarket promotions.** The units are shoppers on shopping trips; the treatment is whether a shopper is given a coupon for gift wrap; the outcome is how much gift wrap he purchased. The goal is to estimate $\mathbb{E}\left[Y_i(1) - Y_i(0) \mid A_i = 1\right]$: how effective is the coupon delivery program?

  Control outcomes are the shopper's purchases for other items; these purchases are not affected by the coupon for gift wrap. The control outcomes can help correct for confounders such as the season. The winter holiday season affects the decision to give a coupon, the amount of gift wrap purchases, and the number of other items purchased.

  One nuance is that the control outcomes must be items that are not complements (Mankiw, 2020) to gift wrap (such as gift cards). Complementary items do not meet the criterion for a control outcome because, through the purchase of gift wrap, they are affected by the delivery of the coupon. The control-outcome deconfounder requires there are no causal connections among the outcomes.

In the coming sections, we characterize the class of unobserved confounders that can be handled with the control-outcome deconfounder and describe the theoretical assumptions required for it to produce unbiased causal estimates. We discuss how the control-outcome deconfounder and its surrounding theory generalizes the method of synthetic controls, focusing on settings with many treated units and a long

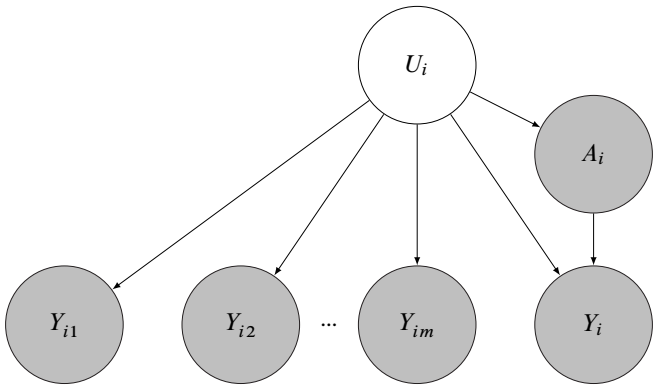


Figure 8.1: Multiple control outcomes with shared unobserved confounding. (Shaded nodes are observed; unshaded ones are latent.)

pre-treatment period. Specifically, the control-outcome deconfounder can produce synthetic controls if it is coupled with inverse probability weighting in the downstream causal inference; thus the results here justify synthetic controls beyond the current theory, to nonlinear settings and non-panel data. Finally, we present a semi-synthetic simulation study of the control-outcome deconfounder on recommendation systems. We show that both the control-outcome deconfounder and the synthetic control method can produce accurate causal estimates in the presence of unobserved confounding.

The rest of the chapter is organized as follows. Section 8.1 sets up the causal inference problem with many control outcomes and develops the control-outcome deconfounder algorithm for causal identification. Section 8.2 discusses the connection between the control-outcome deconfounder and the synthetic control method. Section 8.3 presents an empirical study of the control-outcome deconfounder. Section 8.5 concludes the chapter with a brief discussion.

8.1 Causal inference with multiple control outcomes

8.1.1 Unobserved confounders and a valid substitute confounder

Consider a dataset of n independent and identical units. Each unit i receives a treatment $A_i \in \mathcal{A}$ and elicits an outcome $Y_i \in \mathcal{Y}$; the treatment and outcome are both scalars. Each unit is also accompanied with a potential outcome function $\{Y_i(a)\}_{a\in\mathcal{A}}$, where $Y_i(a)$ is the potential outcome of unit i if it were treated with value a. As a consequence of this notation, $Y_i = Y_i(A_i)$.

One goal of causal inference in this setting is to estimate the population expectation of a potential outcome at treatment a, $\mathbb{E}[Y_i(a)]$. Another is to estimate a conditional expectation of a potential outcome at treatment a given the assigned treatment a', $\mathbb{E}[Y_i(a) \mid A_i = a']$. For now consider the simpler one, the expected potential outcome. As with previous chapters, we worry that some confounders may be unobserved.

Suppose, along with covariates, the data include *control outcomes*—outcome variables that are *a priori* known to be unaffected by the treatment. We will show that observing many control outcomes can help alleviate some of the unobserved confounding.

For each unit i, we observe m control outcomes $Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$. Denote the jth potential control outcome as $Y_{ij}^{\emptyset}(a)$; it is the value of the jth control outcome if unit i were treated with value a. By the definition of the control outcome, the potential control outcome does not depend on the value of the treatment,

$$(Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) \stackrel{d}{=} (Y_{i1}^{\emptyset}(a), \ldots, Y_{im}^{\emptyset}(a)) \qquad \forall a \in \mathcal{A}. \tag{8.1}$$

Observing many control outcomes can help with some types of unobserved confounding. Consider a *multi-outcome confounder*, one where the same unobserved confounder U_i affects the treatment assignment, the control outcomes, and the study outcome. Then the observed control outcomes can serve as evidence of the unobserved confounder U_i.

How? Define a *valid substitute confounder* as having two properties. First, it is a random variable Z_i that renders all the outcomes, both the study outcome and control outcomes, conditionally independent. This implies that there is a model θ that describes the population distribution of the outcomes as a marginal over Z_i,

$$\mathrm{P}(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) = \int \mathrm{P}_\theta(Y_i \mid Z_i) \cdot \left(\prod_{j=1}^{m} \mathrm{P}_\theta(Y_{ij}^{\emptyset} \mid Z_i) \right) \mathrm{P}(Z_i) \, \mathrm{d}Z_i. \tag{8.2}$$

This representation of the joint is seen in Figure 8.2a. Second, we can pinpoint Z_i from the observed outcomes,

$$\mathrm{P}_\theta(Z_i \mid Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) = \delta_{f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})}, \tag{8.3}$$

where $\delta_{f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})}$ represents a point mass located at $f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$ and f is some deterministic function. For many models, this property holds when there are infinitely many control outcomes, $m \to \infty$.

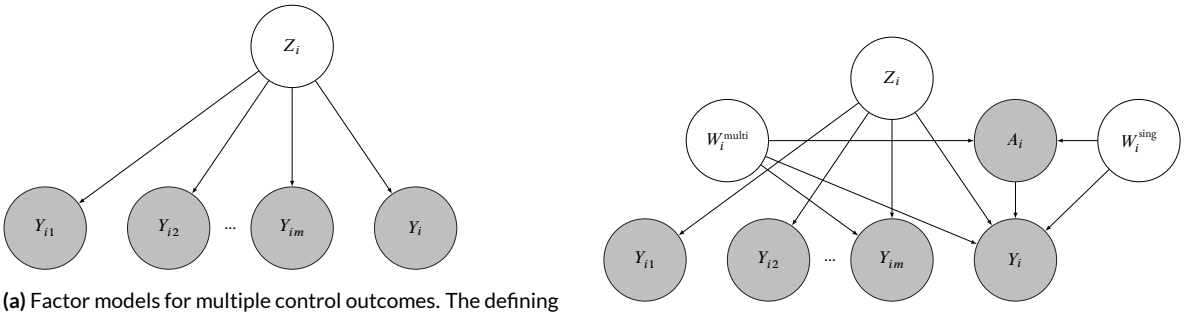
We claim that a valid substitute confounder, as defined above, must capture* all multi-outcome confounders. This is true by proof of contradiction. By construction, Z_i satisfies the conditional independence structure in the likelihood in Equation (8.2), also seen in Figure 8.2a. But suppose it misses† some multi-outcome confounders W_i^{multi}. Then conditional independence can not hold because W_i^{multi} induces dependence between two or more outcomes (see Figure 8.2b) and so Z_i cannot have rendered all the outcomes conditionally independent.

How about single-outcome confounders? The same argument cannot rule out single-outcome confounders W_i^{sing}. Missing out a single-outcome confounder (Figure 8.2b) does not prevent Z_i from satisfying the conditional independence (Figure 8.2a). Note this reasoning follows a similar argument as the one in Chapter 3, though they use multiple treatments rather than multiple outcomes.

In fact, the substitute confounder Z_i cannot pick up‡ any single-outcome variables, because a substitute confounder cannot be both single-outcome and pinpointed. (We prove this fact in Appendix E.1.) This means that Z_i also cannot pick up any post-treatment variables. Post-treatment variables must be single-outcome because there is only one outcome Y_i that can be affected by the treatment.

In summary, observing many control outcomes provides the possibility of constructing a valid substitute confounder. It is a variable that renders the outcomes conditionally independent and can be pinpointed from the observations. A valid substitute confounder captures the information about multi-outcome confounders but does not capture information about single-outcome confounders.

*Technically, "capture" means all multi-outcome confounders are measurable with respect to the substitute confounder.
†"miss" means some multi-outcome confounders are not measurable with respect to the substitute confounder.
‡"not pick up" means no single-outcome variables are measurable with respect to the substitute confounder.



(a) Factor models for multiple control outcomes. The defining feature of factor models is conditional independence: Z_i renders all outcomes $Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$ conditionally independent.

(b) Substitute confounders must pick up all multi-outcome confounders but no single-outcome confounders.

Figure 8.2: Factor models and substitute confounders.

Assuming no unobserved single-outcome confounders, this means we can correct for unobserved multi-outcome confounding by using the substitute Z_i in place of the unobserved confounder U_i. The next section makes this statement more precise, articulating the assumptions needed to identify causal effects in the presence of unobserved multi-outcome confounding. We then operationalize these ideas into the control-outcome deconfounder, an algorithm that uses the identification results to estimate causal effects.

8.1.2 Causal identification with multiple control outcomes

Why do multiple control outcomes help identify causal effects, even in the face of unobserved multi-outcome confounders?

The key idea is to use the observed control outcomes to pinpoint substitute multi-outcome confounders. This strategy requires several assumptions: (1) the stable unit treatment value assumption (SUTVA); (2) a valid substitute confounder, as described above; (3) no unobserved single-outcome confounders; (4) overlap, also known as positivity.

This approach inherits from the classical strategy of measuring sufficient confounders, but with the difference that some confounders—the multi-outcome confounders—may not be observed. To address this, we use the observations of multiple control outcomes to pinpoint a substitute multi-outcome confounder. In this way, we assume "no unobserved single-outcome confounders" instead of the classically assumed "no unobserved confounders." But note we additionally assume (3)—(4) to handle multi-outcome confounders.

We discuss each assumption in detail and then state the theorem.

Assumption 8.1.1: Stable unit treatment value assumption (SUTVA). There is no interference between the units; there is only one version of the treatment for each value $a \in \mathcal{A}$.

SUTVA is the classical assumption we make to perform causal inference on independent units (Rubin, 1980, 1990). It ensures that the potential outcome for each unit $Y_i(a)$ is well-defined.

Assumption 8.1.2: Valid substitute confounder. There exists a valid substitute confounder Z_i. It has the following two properties.

1. It renders the outcomes conditionally independent,

$$\mathrm{P}(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) = \int \mathrm{P}_\theta(Y_i \mid Z_i) \cdot \left(\prod_{j=1}^{m} \mathrm{P}_\theta(Y_{ij}^{\emptyset} \mid Z_i) \right) \mathrm{P}(Z_i) \, \mathrm{d}Z_i. \tag{8.4}$$

2. It can be pinpointed from the outcomes,

$$\mathrm{P}_\theta(Z_i \mid Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) = \delta_{f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})}, \tag{8.5}$$

   for some deterministic function f.

Assumption 8.1.2 describes the two requirements of a valid substitute confounder Z_i. The first part ensures that Z_i renders all the outcomes conditionally independent. This conditional independence ensures that it picks up all multi-outcome confounders. (See the discussion in Section 8.1.1.) The second part implies that we can infer the substitute confounder from the data with full certainty. It

is satisfied when (1) the distribution of all outcomes can be described by a factor model and (2) the number of control outcomes increases to infinity, i.e. $m \to \infty$. For example, if the distribution of all outcomes can be described by structured latent factor models, e.g. probabilistic principal component analysis (Tipping and Bishop, 1999) or Poisson factorization (Gopalan et al., 2013, 2015, 2014), then the substitute confounder can be pinpointed when $(n + m) \cdot \log(nm)/(nm) \to 0$ (Chen et al., 2019).

Assumption 8.1.3: No unobserved single-outcome confounders. Denote U_i as the smallest σ-algebra that renders all the outcomes conditionally independent. Moreover, neither $\mathrm{P}(Y_{ij}^{\emptyset} \mid U_i)$, $j = 1, \ldots, m$ nor $\mathrm{P}(Y_i \mid U_i)$ is a point mass. Then "no unobserved single-outcome confounders" requires that U_i, together with the observed covariates X_i, satisfy weak unconfoundedness,

$$Y_{i1}^{\emptyset}(a), \ldots, Y_{im}^{\emptyset}(a), Y_i(a) \perp A_i \mid Z_i, X_i, \quad \forall a \in \mathcal{A}. \tag{8.6}$$

Assumption 8.1.3 roughly requires that we observe any confounders that affect only one of the outcomes, whether control outcome or study outcome. This assumption is expressed mathematically by assuming the pinpointed substitute confounder Z_i, along with the observed covariates X_i, satisfies weak unconfoundedness (Equation (8.6)) (Hirano and Imbens, 2004; Imbens, 2000). We note that this assumption differs from a common identification assumption used in synthetic control (Angrist and Pischke, 2008; Kinn, 2018; O'Neill et al., 2016),

$$Y_i(a) \perp A_i \mid X_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}, \quad \forall a \in \mathcal{A}. \tag{8.7}$$

Assumption 8.1.3 posits that the substitute confounder Z_i and the covariates X_i satisfy weak unconfoundedness, while Equation (8.7) assumes that all control outcomes $Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$ and the covariates X_i satisfy weak unconfoundedness.

Assumption 8.1.4: Overlap, also known as positivity. For all sets $\mathcal{A}_S$ with positive measure, $\mathrm{P}(A_i \in \mathcal{A}_S \mid X_i = x, Z_i = z) > 0$ for all values of (x, z), where $Z_i \overset{a.s.}{=} f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$ as is required by Assumption 8.1.2.

Overlap is a classical assumption required in causal inference, which often requires that all pairs of treatment and confounder value can occur with positive probability. The overlap assumption here is similar but concerns both the substitute confounder Z_i and the covariates X_i. We emphasize that while the substitute confounder is a deterministic function of the outcomes, this does not prevent the overlap assumption from being satisfied.

Under these assumptions, controlling for a valid substitute confounder leads to unbiased estimates.

Theorem 8.1.1: The control-outcome deconfounder. Under Assumptions 8.1.1 to 8.1.4, controlling for a valid substitute confounder yields unbiased causal estimates,

$$\mathbb{E}\left[Y_i(a)\right] = \mathbb{E}_{Z_i, X_i}\left[\mathbb{E}\left[Y_i \mid A_i = a, Z_i, X_i\right]\right] \tag{8.8}$$

$$\mathbb{E}\left[Y_i(a) \mid A_i = a'\right] = \mathbb{E}_{Z_i, X_i \mid A_i = a'}\left[\mathbb{E}\left[Y_i \mid A_i = a, Z_i, X_i\right]\right]. \tag{8.9}$$

A sketch of the proof is as follows. The substitute confounder Z_i and the covariate X_i together satisfy weak unconfoundedness. Hence adjusting for them with g-formula leads to unbiased estimates of the potential outcomes. The full proof of Theorem 8.1.1 is in Appendix E.2.

At a higher level, Assumptions 8.1.1 to 8.1.4 delineates a setting where the class of multi-outcome confounders are factually unobserved, but effectively observed—they can be pinpointed by observing many control outcomes. This fact is why the control-outcome deconfounder can identify causal quantities from observational data. We note that Assumptions 8.1.2 and 8.1.3 of the control-outcome deconfounder mirror Assumptions 4.1.1 and 4.1.2 of the deconfounder. Moreover, Theorem 8.1.1 mirros Theorem 4.2.2.

8.1.3 The control-outcome deconfounder

We now develop the control-outcome deconfounder, an algorithm that infers valid substitute confounders from multiple control outcomes. We first construct the valid substitute confounder that renders all outcomes conditionally independent. Then we use that substitute in a downstream causal inference.

Constructing a substitute confounder Z_i The key requirement of the substitute confounder is that it renders all the outcomes $(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$ conditionally independent,

$$\mathrm{P}(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset} \mid Z_i) = \mathrm{P}(Y_i \mid Z_i) \cdot \prod_{j=1}^{m} \mathrm{P}(Y_{ij}^{\emptyset} \mid Z_i). \tag{8.10}$$

We enforce this conditional independence by finding factor models that captures the distribution of the outcomes $\mathrm{P}(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$. A factor model posits latent variables Z_i such that

$$Z_i \sim \mathrm{P}(Z_i), \qquad i = 1, \ldots, n, \tag{8.11}$$

$$Y_i \mid Z_i \sim \mathrm{P}(Y_i \mid Z_i, \theta) \tag{8.12}$$

$$Y_{ij}^{\emptyset} \mid Z_i \sim \mathrm{P}(Y_{ij}^{\emptyset} \mid Z_i, \theta_j^{\emptyset}), \qquad j = 1, \ldots, m, \tag{8.13}$$

where $\theta, \theta_1^{\emptyset}, \ldots, \theta_m^{\emptyset}$ are parameters of the factor model. In particular, the latent variable Z_i renders the outcomes conditionally independent.

We fit the factor model by optimizing the likelihood of the factor model over the parameters

$$\hat{\theta}_1^{\emptyset}, \ldots, \hat{\theta}_m^{\emptyset}, \hat{\theta} = \underset{\theta_1^{\emptyset}, \ldots, \theta_m^{\emptyset}, \theta}{\arg\max} \int \mathrm{P}(Y_i \mid Z_i, \theta) \cdot \prod_{j=1}^{m} \mathrm{P}(Y_{ij}^{\emptyset} \mid Z_i, \theta_j^{\emptyset}) \mathrm{P}(Z_i) \, \mathrm{d}Z_i. \tag{8.14}$$

Finally, we obtain a point estimate of the substitute confounders Z_i:

$$\hat{z}_i = \mathbb{E}\left[Z_i \mid Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}, \hat{\theta}_1^{\emptyset}, \ldots, \hat{\theta}_m^{\emptyset}, \hat{\theta} \right]. \tag{8.15}$$

When $m \to \infty$, i.e. the number of control outcomes grows to infinity, the conditional distribution $\mathrm{P}(Z_i \mid Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) \xrightarrow{d} \delta_{f_\theta(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})}$ for some function $f_\theta(\cdot)$. In this case, the point estimate also converge to the same point, i.e. $\hat{z}_i \xrightarrow{a.s.} f_\theta(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$.

Merely fitting a factor model to the outcomes does not guarantee that its latent variable Z_i satisfies the conditional independence requirement (Equation (8.10)). We need to further check that the fitted factor model does capture the distribution of the outcomes well. Many existing algorithms can perform this check. Here we employ a predictive check (Ranganath and Blei, 2019). We randomly hold out some outcomes for each user and fit the factor model. Then we generate the heldout entries using the fitted factor model. If the generated heldout entries are indistinguishable (in terms of log-likelihood values) from the observed values in these heldout entries, we claim the factor model passes the predictive check and can capture the distribution of the outcomes.

Note we infer the substitute confounder by only looking at the outcomes, both the control outcomes and the study outcome. It does not involve the treatments.

Deconfounding with the substitute confounder After obtaining a substitute confounder $\{\hat{z}_i\}_{i=1}^n$, we use them in a downstream causal inference, as though they were observed confounders. For example, we can compute the conditional expected potential outcome $\mathbb{E}\left[Y_i(a) \mid A_i = a'\right]$ by fitting an outcome model

$$\mathbb{E}\left[Y_i \mid A_i = a, Z_i, X_i\right] = \beta_0 + \beta_A \cdot a + \beta_Z \cdot Z_i + \beta_X \cdot X_i \tag{8.16}$$

and apply the adjustment

$$\begin{aligned}\mathbb{E}_{Z_i, X_i \mid A_i = a'}&\left[\mathbb{E}\left[Y_i \mid A_i = a, Z_i, X_i\right]\right] \\ &\approx \beta_0 + \beta_A \cdot a + \beta_Z \cdot \frac{\sum_{A_i = a'} Z_i}{\sum_{i=1}^n \mathbb{1}\{A_i = a'\}} + \beta_X \cdot \frac{\sum_{A_i = a'} X_i}{\sum_{i=1}^n \mathbb{1}\{A_i = a'\}}.\end{aligned} \tag{8.17}$$

Alternatively, we may use the inverse propensity score weighting to calculate $\mathbb{E}\left[Y_i(a) \mid A_i = a'\right]$

$$\sum_{A_i = a} Y_i \cdot \frac{1}{\sum_{i=1}^n \mathbb{1}\{A_i = a'\}} \cdot \frac{\mathrm{P}(A_i = a' \mid Z_i, X_i)}{\mathrm{P}(A_i = a \mid Z_i, X_i)}. \tag{8.18}$$

Both the adjustment formula and inverse propensity score weighting are standard confounder adjustment methods; we may use other adjustment methods and estimate other quantities too. The only twist is that we join the covariates X_i with the substitute confounder Z_i, treating both as confounders we adjust for.

Algorithm 8.1: The control-outcome deconfounder

Input: a dataset of treatment, study outcome, many outcomes, and covariates
$\{(a_i, y_i, y_{i1}^{\emptyset}, \ldots, y_{im}^{\emptyset}, x_i)\},\ i = 1, \ldots, n$
Output: the conditional average potential outcome $\mathbb{E}\left[Y_i(a) \mid A_i = a'\right] \quad \forall a \in \mathcal{A}$
repeat
  choose a factor model from the class in Equation (8.2)
  fit the model to the outcomes $\{(y_i, y_{i1}^{\emptyset}, \ldots, y_{im}^{\emptyset})\},\ i = 1, \ldots, n$
  check the fitted model $\hat{M}$
until *the factor model check is satisfactory*
foreach *datapoint* i **do**
  calculate $\hat{z}_i = \mathbb{E}_{\hat{M}}\left[Z_i \mid Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}\right]$.
end
repeat
  choose an outcome model, e.g. Equation (8.16)
  fit the outcome model to the augmented dataset $\{(a_i, y_i, \hat{z}_i, x_i)\},\ i = 1, \ldots, n$
  check the fitted outcome model
until *the outcome check is satisfactory*
estimate the conditional average potential outcome $\mathbb{E}\left[Y_i(a) \mid A_i = a'\right]$ by Equation (8.17), Equation (8.18), or other adjustment methods

Algorithm 8.1 presents the full algorithm. In summary, this strategy helps deconfound unobserved multi-outcome confounders U_i, those shared between both the study outcome Y_i and the control outcomes $Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$. Thanks to the identification theorem of the Theorem 8.1.1, and subject to the assumptions, this algorithm provides unbiased causal inferences.

8.2 Synthetic control as the control-outcome deconfounder

8.2.1 Panel data and the synthetic control method

The control-outcome deconfounder can be applied to many settings where there exist many control outcomes. One such setting is panel data, where we observe many units and measure their outcomes for m time steps. We are interested in how a policy executed on a subset of units at time $m + 1$ can causally affect their outcome at time $m + 1$.

A popular approach to panel data is the synthetic control method (Abadie et al., 2010, 2015; Abadie and Gardeazabal, 2003), where the researcher weighs the control units so that the treated and the (weighted) control match in the pre-treatment outcomes. In this section, we will establish a connection between the synthetic control method and the control-outcome deconfounder. That is, the synthetic control method can be seen as one instance of the control outcome deconfounder coupled with inverse propensity score weighting. This connection will weaken the linearity assumption that commonly justifies synthetic control. It allows us to apply the synthetic control method to nonlinear settings with many treated units and long pre-treatment periods.

To think concretely about the panel data setting, we consider a dataset of stocks. For each stock i, we measure its covariates X_i and its prices from time 1 to time m, denoted as $(Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$. At time $m+1$, we execute a treatment for some stocks, i.e. the trading commission is doubled. We use $A_i = 1$ to indicate that the commission of stock i is doubled and $A_i = 0$ otherwise. In particular, a stock is treated if its trading volume U_i is high, though the trading volume is unobserved. Finally we measure the stock prices of all stocks at time $m+1$; stock i has price Y_i. The goal is to estimate the causal effect of the treatment (i.e. the commission raise) on the treated units, i.e. $\mathbb{E}\left[Y_i(1) - Y_i(0) \mid A_i = 1\right]$. We observe $\mathbb{E}\left[Y_i(1) \mid A_i = 1\right] = \mathbb{E}\left[Y_i \mid A_i = 1\right]$; the key challenge is to estimate $\mathbb{E}\left[Y_i(0) \mid A_i = 1\right]$.

Relating this stock example to the control outcomes setting (Figure 8.1), we can view the pre-treatment stock prices $(Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})$ as the control outcomes; they are not affected by the commission raise. The variable pair (A_i, Y_i) is the treatment and the outcome of interest; they are confounded by the unobserved trading volume U_i. Finally, the trading volume U_i also affects the pre-treatment outcomes.

To apply the synthetic control method, we need to assume a linear structural model of the panel data:

$$Y_{ij}^{\emptyset} = \delta_j + \theta_j U_i + \lambda_j X_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{8.19}$$

$$Y_i = \delta_{m+1} + \theta_{m+1} U_i + \lambda_{m+1} X_i + \alpha A_i + \epsilon_{i(m+1)}, \tag{8.20}$$

where $\{\delta_j, \theta_j, \lambda_j\}_{j=1}^{m+1}$ are parameters of the structural model, Z_i denotes some unobserved covariates, and $\{\epsilon_{ij}\}_{j=1}^{m+1}$ are independent and identically distributed zero-mean random variables that satisfy $A_i, U_i, X_i \perp \{\epsilon_{ij}\}_{j=1}^{m+1}$. Moreover, the linear structural model allows (U_i, X_i, A_i) to be dependent.

The synthetic control method finds weights for each control unit such that their weighted pre-treatment outcomes match that of the average pre-treatment outcomes treated units; the weighted covariates also match the average of treated units. In mathematical notation, synthetic control finds weights $\{w_i^{\text{SC}}\}_{i=1}^{m}$ that satisfy

$$\sum_{A_i=0} w_i{}^{\text{SC}} Y_{ij}^{\emptyset} = \frac{1}{\sum_{i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} Y_{ij}^{\emptyset}, j = 1, \ldots, m, \tag{8.21}$$

$$\sum_{A_i=0} w_i{}^{\text{SC}} X_i = \frac{1}{\sum_{i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} X_i, \tag{8.22}$$

$$\sum_{A_i=0} w_i^{\text{SC}} = 1, \text{ and } w_i^{\text{SC}} \geq 0, i = 1, \ldots, n. \tag{8.23}$$

Finally we estimate the expected potential outcome $\mathbb{E}\left[Y_i(0) \mid A_i = 1\right]$ with

$$\sum_{A_i=0} w_i^{\text{SC}} \cdot Y_i. \tag{8.24}$$

8.2.2 The synthetic control method and the control-outcome deconfounder

In this section, we show that the synthetic control method can be viewed as an instance of the control-outcome deconfounder. More precisely, the inverse probability weights implied by the substitute confounder will satisfy the synthetic control constraints (Equations (8.21) and (8.22)) in expectation, given both the number of units and the number of pre-treatment periods go to infinity. In other words, these weights can balance all the pre-treatment outcomes. Under additional assumptions, the synthetic control will be equivalent to the inverse probability weights.

Why might the control-outcome deconfounder relate to synthetic control? Suppose the dataset $\{A_i, Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}\}_{i=1}^{n}$ is generated by the linear structural model (Equations (8.19) and (8.20)). This connection relies on a few observations:

1. The linear structural model implies that U_i, X_i are the multi-outcome confounders. They affect all outcomes, control outcomes or the study outcome; they may also affect the treatment A_i. Moreover, U_i, X_i render the outcomes weakly unconfounded:

$$Y_i(a), Y_{i1}^{\emptyset}(a), \ldots, Y_{im}^{\emptyset}(a) \perp A_i \mid X_i, U_i, \quad \forall a \in \mathcal{A}.$$

2. Applying the control-outcome deconfounder to the dataset, we obtain a substitute confounder $Z_i \overset{a.s}{=} (U_i, X_i)$ if the number of units and the number of pre-treatment outcomes both go to infinity. The reason is that Z_i can render all the outcomes $Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$ conditionally independent in Equations (8.19) and (8.20). (See also Section 8.1.) Hence, Assumption 8.1.2 and Assumption 8.1.3 hold for the control-outcome deconfounder.

3. To estimate $\mathbb{E}[Y_i(0) \mid A_i = 1]$, we calculate the inverse probability weighting estimator (Horvitz and Thompson, 1952)

$$\sum_{A_i=0} Y_i \cdot w_i^{\text{IPW}}, \tag{8.25}$$

where the weights are

$$w_i^{\text{IPW}} = \frac{1}{n\mathrm{P}(A_i = 1)} \cdot \frac{\mathrm{P}(A_i = 1 \mid Z_i, X_i)}{\mathrm{P}(A_i = 0 \mid Z_i, X_i)}. \tag{8.26}$$

These weights can balance the pre-treatment outcomes because of Assumption 8.1.3, i.e. these outcomes are weakly unconfounded given Z_i, X_i (Rosenbaum and Rubin, 1983), i.e.

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} Y_{ij}^{\emptyset}\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} Y_{ij}^{\emptyset}\right], j = 1, \ldots, m. \tag{8.27}$$

Roughly, these observations imply that the control-outcome deconfounder finds the unobserved confounder U_i by fitting Bayesian factor models. The resulting inverse probability weights hence satisfy the synthetic control constraints because U_i, X_i render the outcomes weakly unconfounded.

The following theorem summarizes the connection between the control outcome deconfounder and the synthetic control method in general settings.

Theorem 8.2.1: Synthetic control as the control-outcome deconfounder. Suppose Assumptions 8.1.1 to 8.1.4 hold.

1. (Control-outcome deconfounder → Synthetic control) The inverse probability weights implied by the control outcome deconfounder satisfy

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} f_y(f(Y_{ij}^{\emptyset}))\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} f_y(f(Y_{ij}^{\emptyset}))\right] \tag{8.28}$$

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} f_x(X_i)\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} f_x(X_i)\right], \tag{8.29}$$

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} f_z(Z_i)\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} f_z(Z_i)\right], \tag{8.30}$$

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} f_{y_0}(Y_i(0))\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} f_{y_0}(Y_i(0))\right], \tag{8.31}$$

for any functions f_y, f_x, f_z, f_{y_0}. Hence they satisfy the synthetic control constraints (Equations (8.21) to (8.23)) in expectation.

2. (Synthetic control → Control-outcome deconfounder) Suppose the substitute confounder is linear in the outcomes, i.e. $\mathrm{P}(Z_i \mid Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) = \delta_{f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})}$, where $f(\cdot)$ is linear. Further assume that the propensity score function is logistic-linear in Z_i, X_i. The synthetic control weights with the maximum entropy is equivalent to inverse probability weighting with the control-outcome

deconfounder. That is, if

$$\{w_i^{\text{SC}}\}_{i=1}^n = \underset{\{w_i\}_{i=1}^n}{\arg\max} \quad \mathbb{E}\left[\sum_{i=1}^n w_i \log w_i\right] \tag{8.32}$$

$$\text{s.t. } \mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} Y_{ij}^{\emptyset}\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^n \mathbb{1}\{i=1\}} \sum_{i=1} Y_{ij}^{\emptyset}\right] \tag{8.33}$$

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}} X_i\right] = \mathbb{E}\left[\frac{1}{\sum_{i=1}^n \mathbb{1}\{A_i=1\}} \sum_{A_i=1} X_i\right], \tag{8.34}$$

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\text{IPW}}\right] = 1 \text{ and } w_i^{\text{IPW}} \geq 0, i = 1, \ldots, n. \tag{8.35}$$

then

$$w_i^{\text{SC}} = w_i^{\text{IPW}}. \tag{8.36}$$

Theorem 8.2.1 describes the general connection between the control-outcome deconfounder and the synthetic control method. The proof of Theorem 8.2.1 is in Appendix E.3.

The first part of Theorem 8.2.1 shows that the inverse probability weights implied by the control-outcome deconfounder satisfies the weight constraints of the synthetic control method; they balance the pre-treatment outcomes and the observed covariates. Additionally, they can balance any functions of the pre-treatment outcomes, observed covariates, substitute confounders, and the potential study outcomes. Loosely, the control-outcome deconfounder satisfies the synthetic control constraints and more.

The main intuition behind the first part of Theorem 8.2.1 is that inverse probability weights, in general, can balance covariates and potential outcomes under weak unconfoundedness (Imbens and Rubin, 2015a). In this setting, the outcomes satisfy weak unconfoundedness given both the pinpointed substitute confounder and the observed covariates. Hence the implied inverse probability weights balance them both, along with all the outcomes, including both the control outcomes and the study outcomes.

The second part of Theorem 8.2.1 delineates a setting where the synthetic control weights are equivalent to the control-outcome deconfounder. The equivalence holds with a linear substitute confounder and a logistic-linear propensity score model. We note that requiring a linear substitute confounder is weaker than requiring a linear structural model of Equations (8.19) and (8.20) with additive errors. A

substitute confounder can be linear when the outcomes come from nonlinear factor models. For example, this linearity holds when the factor model is any generalized latent factor analysis with canonical exponential family distributions (Chen et al., 2019; Skrondal and Rabe-Hesketh, 2004), including the Poisson factorization model (Gopalan et al., 2013, 2014).

The second part of Theorem 8.2.1 is due to the following observation: exponential family distributions are the maximum entropy solution subject to moment matching constraints. Under first order moment constraints like Equation (8.27), the maximum entropy solution coincides with inverse probability weights with a logistic-linear propensity score model of the covariates and the substitute confounder. Similar results have been developed in the context of covariate-balancing methods (Athey et al., 2018b; Ben-Michael et al., 2019; Chan et al., 2016; Deville and Särndal, 1992; Fan et al., 2016; Hainmueller, 2012; Hellerstein and Imbens, 1999; Hirshberg and Wager, 2018; Imai and Ratkovic, 2014; Li et al., 2018; Wang and Zubizarreta, 2019, 2020; Wong and Chan, 2018; Yiu and Su, 2018; Zhao, 2018; Zhao and Percival, 2017; Zubizarreta, 2015). These results often operate under weak unconfoundedness (Hirano and Imbens, 2004; Imbens, 2000). In contrast, Theorem 8.2.1.2 works with a different set of identifications assumptions, i.e. Assumptions 8.1.1 to 8.1.4.

Together with Theorem 8.1.1, Theorem 8.2.1 also justifies the synthetic controls beyond linear structural models. Under assumptions of Theorem 8.2.1, the synthetic control methods can produce unbiased causal estimates, as with the control outcome deconfounder.

This connection between synthetic control and the control outcome deconfounder also generalizes to other variants of synthetic control methods. For example, the same connection also holds with robust synthetic controls and its multi-dimensional variant (Amjad et al., 2019, 2018). The exact same argument applies if we consider all outcomes Y_i, $Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$ being $L-$dimensional vectors ($L > 1$).

Finally, we remark that the control-outcome deconfounder relates closely to Xu (2017) and Gobillon and Magnac (2016) as two-stage algorithms for panel data. The difference lies in that the control-outcome deconfounder explicitly uncover potential unobserved confounders by fitting Bayesian factor models, while the others optimize weights, which implicitly handle unobserved confounders. The setting of many control outcome ($m \to \infty$) also closely relates to Ferman (2019), which studies the limiting properties of the synthetic control method but with infinitely many pre-treatment periods. However, their results focus on linear structural models while we study more general settings in this chapter.

8.3 Empirical study: How effective are recommendations?

We study the control-outcome deconfounder in a semi-synthetic study about recommendation systems. We simulate movie recommendations and the resulting user behaviors, where the causal effect of recommendations is confounded by the (unobserved) user preferences. The goal is to estimate this causal effect of recommendation.

Across different strengths of unobserved confounding, we demonstrate that the control-outcome deconfounder can correct for confounding bias in causal estimates. Moreover, the control-outcome deconfounder produces causal estimates with smaller mean squared error than the synthetic control method.

Simulation setup Start with the MovieLens dataset ML1M, [§] which contains one million ratings (0-5 stars) from $6,000$ users on $4,000$ movies; denote the rating of user i on movie j as R_{ij}. For each user i, ($i = 1, \ldots, 6000$), we generate the confounder U_i by calibrating user preferences on drama, i.e. what percentage of movies to which user i gave a nonzero rating is drama. Next randomly choose a movie, e.g. "Othello"; for notation convenience, we relabel the movies so that "Othello" is the last movie ($j = 4000$). "Othello" will be the movie whose recommendation and its resulting user behaviors are of interest. We simulate the treatment A_i, i.e. whether user i was recommended with the movie "Othello":

$$A_i \sim \text{Bernoulli}(\text{sigmoid}(\gamma_0 + \gamma_U U_i)), \tag{8.37}$$

where γ_U is the correlation between confounder and the log-odds of the treatment. Finally we generate the outcomes, i.e. how many 20-minute blocks user i spent on each movie. The time spent on movies other than "Othello" are the control outcomes $Y^N_{i1}, \ldots, Y^N_{i,3999}$; they are made equal to the ratings which take values in $\{0, 1, 2, 3, 4, 5\}$:

$$Y^N_{ij} = R_{ij}, \quad i = 1, \ldots, 6000, j = 1, \ldots, 3999. \tag{8.38}$$

[§] http://grouplens.org/datasets/movielens/

Finally, the time user i spent on "Othello" is the outcome of interest Y_i; it is affected by both the treatment A_i and the confounder U_i:

$$Y_i \sim \text{Poisson}(R_{i,4000} + \beta_A A_i + \beta_U U_i), \tag{8.39}$$

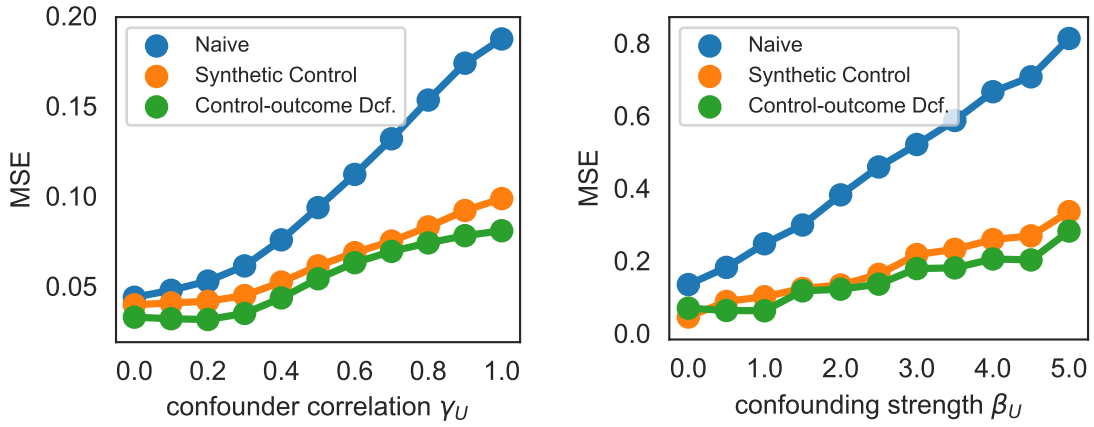
where the coefficients $\beta_U > 0$ is the confounder strength and $\beta_A > 0$ is the causal effect of the treatment. The goal is to estimate average treatment effect on the treated (ATT), whose true value is β_U.

Competing methods We apply the control-outcome deconfounder to the semi-synthetic dataset. We use Poisson factorization (Gopalan et al., 2013, 2014) as the factor model and decide its dimensionality by choosing the smallest dimension that passes the predictive check. (The chosen latent dimensionality is 35.) Further, we choose the outcome model as Poisson regression of the outcome Y_i against the treatment A_i and the substitute confounder Z_i. We use the g-formula for confounder correction.

We compare the control-outcome deconfounder with two baseline methods: One is a naive ATT estimator $\mathbb{E}[Y_i \mid A_i = 1] - \mathbb{E}[Y_i \mid A_i = 0]$, where the expectations are computed using Poisson regression; the other is a variant of synthetic control estimator, implemented as in Doudchenko and Imbens (2016).

Results Figure 8.3a demonstrates how the confounder correlation γ_U affects the ATT estimates obtained from the naive estimator, the synthetic control method, and the control-outcome deconfounder. Figure 8.3b illustrates the effect of the confounder strength β_U. The naive ATT estimator is susceptible to confounding bias. However, both the synthetic control method and the control-outcome deconfounder are robust to unobserved confounding; they are both able to correct for the bias due to U_i.

These results corroborate Theorem 8.1.1, which establishes the validity of the control-outcome deconfounder in the presence of unobserved multi-outcome confounding. They also echo Theorem 8.2.1, which describes the close connection between the synthetic control method and the control-outcome deconfounder with logistic treatment model and linear substitute confounder. Finally, we observe that



(a) Varying confounder correlation γ_U (β_U = 3.0)

(b) Varying confounder strength β_U (γ_U = 0.5)

Figure 8.3: Mse of att estimates under varying strengths of confounding. Each dot is the average MSE of 30 replications. The naive att estimate suffers from unobserved confounding. In contrast, both the control-outcome deconfounder and the synthetic control method are able to correct for unobserved confounding. The control-outcome deconfounder produces att estimates with slightly smaller MSEs.

the control-outcome deconfounder produces ATT estimates with slightly smaller MSEs. This difference in MSEs is likely due to the difference in efficiency between the two estimators; the synthetic control estimate is a weighting estimate, which may bear a higher variance than the correction with parametric outcome models.

8.4 Related work

This chapter relates to multiple threads of research in causal inference.

Causal inference on panel data. Causal inference on panel data has been studied from many perspectives. Brodersen et al. (2015) approach it with Bayesian structural time-series models. Bertrand et al. (2004); Card (1990) study it with differences-in-difference estimates, along with more recent works of Abadie (2005); Abraham and Sun (2018); Athey and Imbens (2006); Bonhomme and Sauder (2011); Botosaru and Gutierrez (2018); Callaway et al. (2018); Callaway and Sant'Anna (2019); Cengiz et al. (2019); De Chaisemartin and D'HaultfŒuille (2017); Goodman-Bacon (2018); Heckman et al. (1998, 1997); Imai and Kim (2019); Qin et al. (2008).

Abadie et al. (2010, 2015); Abadie and Gardeazabal (2003) approach causal estimation on panel data with the synthetic control method, along with many variants, extensions, and discussions (Abadie and L'Hour, 2017; Amjad et al., 2019, 2018; Ben-Michael et al., 2019; Chernozhukov et al., 2017b, 2018; Doudchenko and Imbens, 2016; Dube and Zipperer, 2015; Dusetzina et al., 2015; Ferman, 2019; Gobillon and Magnac, 2016; Hsiao et al., 2012; Kinn, 2018; Li, 2019; O'Neill et al., 2016; Robbins et al., 2017; Samartsidis et al., 2019; Viviano and Bradic, 2019; Xu, 2017). Taking the matrix completion perspective, Athey et al. (2018a) unify the synthetic control approaches with the unconfoundedness approaches (Imbens and Rubin, 2015a; Rosenbaum and Rubin, 1983). Arkhangelsky et al. (2019) develop synthetic difference-in-differences, which generalizes the synthetic control method and the difference-in-difference estimator.

This chapter differs from all these works in that we study the more general setting of multiple control outcomes than the setting of panel data. The identification strategy we develop do not rely on linearity of the structural models as in Abadie et al. (2010); Xu (2017); it also differs from other existing identification conditions like Kinn (2018); O'Neill et al. (2016) in relying on the asymptotics of latent variable models (Chen et al., 2019).

Negative controls. Negative controls, also known as proxy variables, are often used to detect the presence of unobserved confounding (Lipsitch et al., 2010; Rosenbaum, 1989). More recently, researchers have developed strategies that use negative controls to identify causal effects under unobserved confounding; they posit different assumptions under which certain observed or counterfactual variables suffice to account for unobserved confounding (Arnold et al., 2016; Dusetzina et al., 2015; Egami, 2018; Flanders et al., 2017; Kasza et al., 2017; Miao and Tchetgen, 2018; Miao and Tchetgen Tchetgen, 2017; Richardson et al., 2015; Sanderson et al., 2017; Shi et al., 2018; Sofer et al., 2016; Tchetgen Tchetgen, 2013). More recently, Kuroki and Pearl (2014); Miao et al. (2018) leverage completeness conditions on the distributions of variables to identify causal effects with negative controls. Madigan et al. (2014); Schuemie et al. (2016, 2018) use negative controls to calibrate p-values of average treatment effects in observational studies. This chapter differs from them in that causal identification is achieved via the consistent estimability of latent factor models and letting the number of outcomes increase to infinity.

Unconfoundedness and sensitivity analysis. Hirano and Imbens (2004); Imbens (2000); Rosenbaum and Rubin (1983) show that weak unconfoundedness is a key assumption that helps identify causal effects. Many works attempt to assess the plausibility of these assumptions with sensitivity analysis, including Franks et al. (2019); Gilbert et al. (2003); Imai and Van Dyk (2004); Imbens (2003); Robins et al. (2000b). More recently, researchers have developed a test for confounding leveraging external data (Janzing and Schölkopf, 2018a,b; Liu and Chan, 2018; Sharma et al., 2016). This chapter differs from them in that we leverage external data of multiple control outcomes to adjust for unobserved confounding.

8.5 Discussion

In this chapter, we study how the external data of many negative control outcomes can help control for unobserved confounding in causal inference. We develop the control-outcome deconfounder, an algorithm that handles unobserved multi-outcome confounders with Bayesian factor models. We describe the theoretical assumptions required for the algorithm to produce unbiased causal estimates. We also show how the control-outcome deconfounder generalizes the synthetic control method in settings where the number of units and the number of pre-treatment periods go to infinity. This connection justifies the synthetic control method beyond the current theory, to non-linear settings and non-panel data. We demonstrate in a semi-synthetic study that both the control-outcome deconfounder and the synthetic control method are robust to certain unobserved confounding.

9

The debate around the deconfounder

In this chapter, we summarize the debate around the deconfounder (D'Amour, 2018, 2019a,b,c; Ogburn et al., 2020, 2019a,b) and discuss how they inform the theory and practice of the deconfounder.*

9.1 D'Amour (2018, 2019a,b,c)

D'Amour (2018, 2019b,c) explore nonparametric non-identification of the potential outcome distribution $P(Y_i(\boldsymbol{a}))$ in multiple causal inference with many causes. He shows that $P(Y_i(\boldsymbol{a}))$ cannot be nonparametrically identifiable without any assumptions on the causal relationship involved.

The causal identification results of the deconfounder in Chapters 4 and 5 (Theorems 4.2.1 to 4.2.3, 5.1.2 and 5.2.2) do not contradict these negative results from D'Amour (2018, 2019b,c). These results establish causal identification for different causal quantities and make different assumptions than D'Amour (2018, 2019b,c). More specifically, under pinpointability and other suitable conditions, The-

*This chapter is in part based on Wang and Blei (2019a,b,c, 2020).

Causal quantity	Result	Condition	Source
$P(Y(\boldsymbol{a}))$	Non-ID	No conditions	D'Amour (2018, 2019b,c)
$\mathbb{E}[Y(\boldsymbol{a})] - \mathbb{E}[Y(\boldsymbol{a}')]$	ID	Pinpointed substitute confounder; Categorical substitute confounder; No confounder/cause interaction; Differentiable relationships	Theorem 4.2.1
$\mathbb{E}_A\left[\mathbb{E}_Y\left[Y(a_{1:k}, A_{(k+1):m})\right]\right]$	ID	Pinpointed substitute confounder; $A_{1:k}$ satisfy overlap	Theorem 4.2.2
$\mathbb{E}[Y(\boldsymbol{a}') \mid \boldsymbol{A} = \boldsymbol{a}]$	ID	Pinpointed substitute confounder; $\boldsymbol{a}'$ and $\boldsymbol{a}$ map to same substitute	Theorem 4.2.3
$\mathbb{E}[Y(\boldsymbol{a})]$	ID	$\mathbb{E}[U \mid \mathrm{A}]$ nonlinear; $\mathbb{E}[Y \mid \mathrm{A}, U]$ linear	Sec 2.1 (IJ)
$\mathbb{E}[Y(\boldsymbol{a})]$	ID	Measure instrument W; Instrument W satisfies overlap	Sec 2.2 (IJ)
$\int Y(\boldsymbol{a}) q_1(\boldsymbol{a})\, \mathrm{d}\boldsymbol{a}$ $- \int Y(\boldsymbol{a}) q_2(\boldsymbol{a})\, \mathrm{d}\boldsymbol{a}$	ID	$p(\boldsymbol{a} \mid z) > 0$ when $q_1(\boldsymbol{a}), q_2(\boldsymbol{a}) > 0$	Sec 2.3 (IJ)

Table 9.1: Identification in multiple causal inference

orem 4.2.1 shows that the average causal effect of all the causes $\mathbb{E}[Y_i(\boldsymbol{a})] - \mathbb{E}[Y_i(\boldsymbol{a}')]$ is nonparametrically identifiable; Theorem 4.2.2 shows that the average causal effect of subsets of the causes $\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] - \mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a'_{1:k}, A_{i,(k+1):m})\right]\right]$ is nonparametrically identifiable; Theorem 4.2.3 shows that the conditional mean potential outcome $\mathbb{E}[Y_i(\boldsymbol{a}') \mid \boldsymbol{A}_i = \boldsymbol{a}]$ is nonparametrically identifiable. Table 9.1 summarizes these results, together with those from D'Amour (2018, 2019b,c); Imai and Jiang (2019), and describes the current identification landscape of multiple causal inference with many causes.

When a causal quantity of interest is not known to be identified, we must evaluate the uncertainty of the deconfounder estimate. The posterior distribution of the deconfounder estimate will reflect this non-identifiability. It calibrates how the (finite) observed data informs causal quantities of interest (Gustafson, 2010, 2015).

D'Amour (2019a,c) also articulates the fundamental tension between using the causes to infer unobserved confounding and using them to estimate causal effects. In other words, the deconfounder does not provide free lunch: the more information is baked into estimating the substitute confounder, the less information is available for estimating causal quantities. Moreover, the assumption that we can pinpoint the substitute confounder is at odds with "all-cause" overlap, i.e., that $P(\boldsymbol{A} \mid Z) > 0$. Both cannot be simultaneously satisfied.

Theorems 4.2.1 to 4.2.3 live at one extreme of this tension. They assume a pinpointed substitute confounder and forgo overlap on all the causes. (Note it is still possible for subsets of the causes to satisfy overlap, as in Theorem 4.2.2.) To achieve identification without overlap, Theorem 4.2.1 extrapolates the observed data by appealing to a non-differentiable substitute confounder and differentiable causal relationships. Moreover, the pinpointed substitute confounder is achievable thanks to the multiplicity of the causes and the consistent estimability of factor models.

9.2 Ogburn et al. (2020, 2019a,b)

Ogburn et al. (2019a,b) provide a technical meditation on some of the theoretical aspects of the deconfounder. Among their remarks, they claim that there are "foundational errors" with the work and that the "premise is incorrect." These claims are not substantiated. There are no foundational errors; the premise is correct. Further, in a continuation of their commentary, Ogburn et al. (2020) propose counterexamples to the theory of the deconfounder, claiming that the substitute confounder cannot satisfy weak unconfoundedness. Using Proposition 4.1.1, we show how the proposed counterexamples do not satisfy the required assumptions. Ogburn et al. (2020) also claim that, even if weak unconfoundedness holds, Theorems 4.2.1 and 4.2.2 do not hold because their causal estimands are not well-defined under pinpointability. We show how the other assumptions of Theorems 4.2.1 and 4.2.2 enable the causal estimands in question to be well-defined.

9.2.1 Ogburn et al. (2019a,b)

The identification results in Theorems 4.2.1 to 4.2.3 capitalize on two requirements: (1) the distribution of the causes $p(\boldsymbol{a})$ can be described by a Bayesian factor model and (2) the factor model pinpoints the substitute confounder Z, i.e. $Z \overset{a.s.}{=} f_\theta(\mathrm{A})$ for some f_θ. The first requirement relies on the successful execution of the deconfounder, i.e., finding a factor model that captures $p(\boldsymbol{a})$. The conditional independence structure of factor models guarantees that the substitute confounder Z pick up all multi-cause confounders and no multi-cause mediators or colliders. The second requirement is pinpointability. It is satisfied when the number of causes goes to infinity and Z remains finite-dimensional. From Lemma 4, it guarantees that Z cannot pick up single-cause confounders, mediators, or colliders.

Ogburn et al. (2019a,b)'s main concern revolves around mediators; they worry that the substitute confounder picks up mediators and bias the causal estimate. Below we show that the substitute confounder can not pick up any information about multi-cause mediators, single-cause mediators, or any of the other graphs that Ogburn et al. (2019a,b) put forward.[†]

Lemma 9.2.1: Substitute confounders can not pick up mediators. No post-treatment variable can be measurable with respect to a pinpointed substitute confounder.

Proof. First, the substitute cannot pick up any multi-cause post-treatment variables. Otherwise, the substitute can not render all the causes conditionally independent.

The substitute also cannot pick up any single-cause variables. These variables include pre-treatment variables, such as single-cause confounders, and single-cause post-treatment variables, such as single-cause mediators or colliders.

The key idea behind the proof is the following. We assume the causes pinpoint the substitute confounder $z = f(\boldsymbol{a}\,;\,\theta)$, as is the case where there are many causes. The deconfounder further requires that the converse is not true, i.e., that the substitute does not pinpoint the causes. This fact holds in a probabilistic model of the causes, such as when the dimension of the substitute stays fixed as the number

[†]This result is stated as Lemma 4 in Wang and Blei (2019a). This lemma is correct, as is the proof in the paper. But Lemma 4 might also seem surprising. Here we present an alternative proof.

of causes increases. Further, the deconfounder requires that the factor model can not have one component of the substitute *a priori* be a deterministic function of another component; this fact also holds in probabilistic factor models. The proof then follows by contradiction: if the substitute picks up single-cause variables then the factor model must be "degenerate," i.e., non-probabilistic.

Here are the details. Suppose the substitute Z does pick up a single-cause variable. Then separate Z into a single-cause component and a multi-cause one, $Z = (Z_s, Z_m)$. Without loss of generality, assume the single-cause component only depends on the first cause. The assumption of a pinpointed substitute confounder says

$$p(z \mid \boldsymbol{a}, \theta) = p(z_s, z_m \mid \boldsymbol{a}, \theta) = \delta_{(f_s(\boldsymbol{a}\,;\,\theta), f_m(\boldsymbol{a}\,;\,\theta))}, \tag{9.1}$$

where $\boldsymbol{a} = (a_1, \ldots, a_m)$ are the m causes and $f(\cdot)$ are the deterministic functions that map causes to substitute confounders.

Now calculate the conditional distribution of the single-cause component given the causes,

$$\begin{aligned}
&p(z_s \mid \boldsymbol{a}) \\
&= p(z_s \mid \boldsymbol{a}, z_m = f_m(\boldsymbol{a}\,;\,\theta))) && (9.2)\\
&= p(z_s \mid a_1, z_m = f_m(\boldsymbol{a}\,;\,\theta))) && (9.3)\\
&= \frac{p(z_s \mid z_m = f_m(\boldsymbol{a}\,;\,\theta)) \cdot p(a_1 \mid z_s, z_m = f_m(\boldsymbol{a}\,;\,\theta))}{p(a_1 \mid z_m = f_m(\boldsymbol{a}\,;\,\theta))}. && (9.4)
\end{aligned}$$

Equation (E.2) is due to the pinpointability of substitute confounder. Equation (E.3) is due to $Z_s \perp A_2, \ldots, A_m \mid A_1, Z_m$. Equation (9.4) is due to the definition of conditional probability.

Equation (9.4) and Equation (9.1) imply that at least one of $p(z_s \mid z_m = f_m(\boldsymbol{a}\,;\,\theta))$ and $p(a_1 \mid z_s, z_m = f_m(\boldsymbol{a}\,;\,\theta))$ is a point mass. But this is a contradiction: either term being a point mass implies that the factor model is degenerate. The former is a point mass when one component Z_s of the substitute is a deterministic function of another component Z_m. The latter is a point mass when the first cause is a deterministic function of the latent Z.

Note the same argument would not reach a contradiction for multi-cause variables Z_{m}. The reason is that

$$
\begin{aligned}
& p(z_{\mathrm{m}} \mid \boldsymbol{a}) \\
&= p(z_{\mathrm{m}} \mid \boldsymbol{a}, z_{\mathrm{s}} = f_{\mathrm{s}}(\boldsymbol{a}\,;\,\theta))) && (9.5) \\
&= \frac{p(a_1, z_{\mathrm{m}} \mid z_{\mathrm{s}} = f_{\mathrm{s}}(\boldsymbol{a}\,;\,\theta))) \cdot \prod_{j=2}^{m} p(a_j \mid z_{\mathrm{m}})}{p(\boldsymbol{a})}, && (9.6)
\end{aligned}
$$

where $\prod_{j=2}^{m} p(a_j \mid z_{\mathrm{m}})$ can converge to a point mass with non-degenerate factor models and $m \to \infty$.

□

9.2.2 Ogburn et al. (2020)

Ogburn et al. (2020)'s proposed counterexamples violate the assumptions Ogburn et al. (2020) propose counterexamples to the theory of the deconfounder. Using Proposition 4.1.1, the proposed counterexamples do not satisfy the required assumptions.

Example 1 There are two independent causes A_1 and A_2 and a substitute confounder $Z \sim \text{Bernoulli}(0.5)$ that is independent of all other variables $(A_1, A_2, Y(\text{a}))$. Here the substitute confounder Z does not satisfy Assumption 4.1.2; its conditional distribution $\text{P}(Z \mid A_1, A_2) = \text{Bernoulli}(0.5)$ is not a point mass.

Example 2 There are two causes A_1 and A_2, and another variable U. Assume that $A_j \perp\!\!\!\perp Y(\boldsymbol{a}) \mid U, j = 1, 2$ and U is the smallest σ-algebra that renders A_1 and A_2 conditionally independent. Set the substitute confounder $Z = U$. Again $\text{P}(Z \mid A_1, A_2)$ is not a point mass, violating Assumption 4.1.2. Further, U does not satisfy weak unconfoundedness because $(A_1, A_2) \not\perp\!\!\!\perp Y(\boldsymbol{a}) \mid U$, violating Assumption 4.1.1.[‡]

[‡]This example does satisfy the "no unobserved single-cause confounder" assumption as stated in Definition 4 of Wang and Blei (2019a). But it violates the refined assumption in Assumption 4.1.1, due to Imai and Jiang at JSM and also published in Wang and Blei (2019c).

Example 3 A variant of the second example involves a third cause A_3 and sets the substitute confounder $Z = A_3$. This example violates Assumption 4.1.2. A pinpointed substitute cannot be a function of only one cause (Wang and Blei, 2019c); see also Lemma 9.2.1.

The importance of conditional independence Ogburn et al. (2020) claim that the conditional independence requirement of factor models does not "drive the success" of deconfounder-like methods. But the conditional independence requirement, along with pinpointability, plays an important role in confirming the assumptions required by the algorithm.

1. Requiring conditional independence outlines the class of confounders that the deconfounder targets; they must be multi-cause confounders. This requirement is why, with the assumption of no unobserved single-cause confounder, the deconfounder handles all confounders.

2. Requiring conditional independence prevents the substitute confounder from capturing multi-cause colliders or mediators; capturing such variables violates the conditional independence requirement.

3. As for single-cause post-treatment variables, Wang and Blei (2019c) shows that substitute confounders that satisfy Assumption 4.1.2 cannot capture any single-cause variables. Thus, along with point #2 above, the substitute confounder does not capture any post-treatment variables.

The causal estimand is well-defined Ogburn et al. (2020) also claim that Theorems 4.2.1 and 4.2.2 do not hold even if the substitute confounder satisfies weak unconfoundedness. They write that the conditional distribution $P(Y \mid A_1 = a_1, \ldots, A_m = a_m, Z = z, X = x)$ appears in Theorems 4.2.1 and 4.2.2, but this conditional probability is not always well defined under the pinpointability assumption (Assumption 4.1.2) which requires $P(Z \mid A_1, \ldots, A_m) = \delta_{f(A_1,\ldots,A_m)}$. In particular, the conditioning event $\{A_1 = a_1, \ldots, A_m = a_m, Z = z\}$ is an impossible event when $z \neq f(a_1, \ldots, a_m)$.

However, Theorems 4.2.1 and 4.2.2 do hold.

Theorem 4.2.2 does not involve the conditional $P(Y \mid A_1 = a_1, \ldots, A_m = a_m, Z = z, X = x)$ that Ogburn et al. (2020) question. It identifies the expected potential outcome when intervening on subsets of the causes, so it only involves the conditional given the first k ($k < m$) causes, i.e. $\mathbb{E}[Y \mid Z, X, A_1 = a_1, \ldots, A_k = a_k]$; see Equation (4.16) in Theorem 4.2.2 and Equations (A.28) to (A.29) in Appendix A.2.

Further, this conditional expectation $\mathbb{E}[Y \mid Z, X, A_1 = a_1, \ldots, A_k = a_k]$ is well-defined because the conditioning event $\{Z, X, A_1 = a_1, \ldots, A_k = a_k\}$ is assumed to be possible. The reason is that Theorem 4.2.2 requires the overlap condition, i.e. $P((A_1, \ldots, A_k) \in \mathcal{A} \mid Z, X) > 0$ for any set $\mathcal{A}$ with $P(\mathcal{A}) > 0$, which implies that the conditioning event $\{Z, X, A_1 = a_1, \ldots, A_k = a_k\}$ must be possible. Further, assuming overlap is not at odds with pinpointability (Assumption 4.1.2); the two assumptions can hold simultaneously. It is because the conditioning event only involve the first k ($k < m$) causes. In contrast, pinpointability is a constraint between Z and *all* the m causes.

To understand the two assumptions intuitively, consider Z as the genre of the movie and $A_1, \ldots, A_m$ as the m actors. Pinpointablity roughly means that each cast $(a_1, \ldots, a_m)$ deterministically maps to a genre z. For example, an all-action-movie-cast can only map to the genre action movie. However, this deterministic mapping does not exclude the possibility that a subset of the actors (e.g. two of the action movie actors, Jason Statham and Sylvester Stallone) appear in both action movies and drama movies.

Theorem 4.2.1 does involve the full conditional $\mathbb{E}[Y \mid A_1 = a_1, \ldots, A_m = a_m, Z = z, X = x]$ in question. However, it is well-defined due to the structural assumption on the potential outcome (Equation (4.14)):

$$\mathbb{E}[Y(\boldsymbol{a}) \mid Z = z, X = x] = f_1(\boldsymbol{a}, x) + f_2(z) \quad \forall (\boldsymbol{a}, z, x) \in \mathcal{A} \times \mathcal{Z} \times \mathcal{X}. \tag{9.7}$$

This structural assumption implies the form of the full conditional

$$\mathbb{E}[Y \mid Z = z, X = x, \boldsymbol{A} = \boldsymbol{a}] = f_1(\boldsymbol{a}, x) + f_2(z) \quad \forall (\boldsymbol{a}, z, x) \in \mathcal{A} \times \mathcal{Z} \times \mathcal{X}. \tag{9.8}$$

This step is due to the weak unconfoundedness of the substitute confounder (Proposition 4.1.1), which implies

$$\mathbb{E}\left[Y(\boldsymbol{a}) \mid Z = z, X = x\right] = \mathbb{E}\left[Y(\boldsymbol{a}) \mid Z = z, X = x, \boldsymbol{A} = \boldsymbol{a}\right] \tag{9.9}$$

$$= \mathbb{E}\left[Y \mid Z = z, X = x, \boldsymbol{A} = \boldsymbol{a}\right]. \tag{9.10}$$

Equation (9.8) holds for all $(\boldsymbol{a}, z, x) \in \mathcal{A} \times \mathcal{Z} \times \mathcal{X}$. They include the tuples where $z \neq f(\boldsymbol{a})$ does not satisfy the deterministic constraint in the pinpointability assumption. At a higher level, it is this structural assumption on the potential outcome (Equation (9.7)) that makes the full conditional in question $\mathbb{E}\left[Y \mid A_1 = a_1, \ldots, A_m = a_m, Z = z, X = x\right]$ well-defined; it allows Equation (9.8) to extrapolate to these $(\boldsymbol{a}, z, x)$ tuples.

One might ask: how does it make sense to ask about the expected outcome Y conditional on an impossible event $\{Z = z, X = x, \boldsymbol{A} = \boldsymbol{a}\}$ where $z \neq f(\boldsymbol{a})$? The reason is that this event is only deemed impossible in the current observational dataset: the pinpointability assumption only applies to the current observational dataset. Yet, the same causal relationship (Equation (9.7)) persists across many datasets with different relationships between $\boldsymbol{A}$ and Z, i.e. $P(\boldsymbol{A} \mid Z, X)$. Hence the pinpointability assumption does not preclude the existence of another observational dataset violating this constraint while obeying the causal relationship (Equation (9.7)). The full conditional $\mathbb{E}\left[Y \mid A_1 = a_1, \ldots, A_m = a_m, Z = z, X = x\right]$ is thus a valid conditional in these datasets.

To understand this argument intuitively, we return to the movie actor example. Suppose Z represents the movie genre and $\boldsymbol{A}$ is the vector of actors. Suppose, in the dataset at hand, all directors are rational and an all-action-movie-cast can only appear in an action movie. That said, another dataset can exist where an all-action-movie-cast appears in a drama movie made by an inexperienced director. The full conditional, i.e. the expected revenue Y of such a movie, is a valid quantity to compute.

As another example, consider Z as one's birthplace and $\boldsymbol{A}$ as his or her genetic markers. It is possible to have one dataset where one configuration of the genetic markers maps deterministically to being born in Europe. However, it doesn't preclude the possibility of another dataset with this same configuration probabilistically mapping to both Europe or America. The query about the expected height Y given this configuration and being born in America is legitimate; hence the full conditional $\mathbb{E}\left[Y \mid A = a, Z = z, X = x\right]$ is well-defined.

9.2.3 Discussion

To reiterate Wang and Blei (2019b,c, 2020), the deconfounder is not a turnkey solution to causal inference. It does not relieve the researcher from trying to measure confounders. As for all causal inference with observational data, it comes with uncheckable assumptions. In particular, Assumption 4.1.1 is that there are no unobserved single-cause confounders.

Both Ogburn et al. (2019a,b) and Ogburn et al. (2020) question the correctness of the theory behind the deconfounder. The objections are incorrect, discussed above and also in Wang and Blei (2019b,c, 2020). But more broadly, Ogburn et al.'s resistance to the idea stems from an important misunderstanding. Both commentaries reiterate the fact that no information about unobserved confounders can be inferred from observational data. The deconfounder does not challenge this fact. Rather, the theory finds confounders that are *effectively observed*, even if not explicitly so, and embedded in the multiplicity of the causes. The deconfounder extracts this information for causal inference.

10

Conclusions

Classical causal inference studies how a univariate cause affects a univariate outcome. In this dissertation, we studied *multiple causal inference*, where multiple causes or multiple outcomes are simultaneously of interest. Multiple causes or multiple outcomes might at first appear to be a complication, but we showed that it can help reveal unobserved confounding.

We developed the family of deconfounder algorithms: first, fit a good factor model of the multiple causes (or multiple outcomes); then, use the factor model to infer a substitute confounder; finally, perform causal inference. We showed how a substitute confounder from a good factor model must capture all multi-cause (or multi-outcome) confounders in theory. We also demonstrated that the deconfounder improve causal estimates in empirical studies.

Directions of future work

Identification landscape of multiple causal inference We provided three identification results for the deconfounder when the causal question involves an infinite number of causes (or outcomes). (Table 9.1 summarizes all the known results.) However, the general identification landscape of multiple causal inference is unknown. Which causal quantities are identifiable and under what conditions? Painting the full identification landscape will help draw the boundaries for the causal problems to which the deconfounder may be applied.

On the parametric identification front, one strategy to develop identification results is to use the incongruence between the factor model and the outcome model. For example, Appendices C.1 and C.2 employ moment generating functions of the observed data to establish identification for a quadratic factor model and a linear outcome model. Imai and Jiang (2019) show that causal parameters are identified with nonlinear factor models and linear outcome models. Kong et al. (2019) establish causal identification for Gaussian causes and a binary outcome.

This incongruence approach to causal identification relates to Theorem 4.2.1, which dwells on the incongruence between the non-differentiability of the unobserved confounder and the differentiability of all other relationships. It has also been used to learn causal graphs and to see the arrow of time (Bauer et al., 2016; Moneta et al., 2013; Peters et al., 2009, 2013, 2017; Shimizu, 2014), which relies on the incongruence between the non-Gaussian variables and the Gaussian noise.

On the non-parametric identification front, any causal identification results that allow for flexible outcome models will be valuable. Known identification results of the deconfounder can only apply to causal problems with infinite causes (Theorems 4.2.1 to 4.2.3). Moreover, these results do not apply to important causal quantities like the expected potential outcome $\mathbb{E}\left[\boldsymbol{Y}_i(\boldsymbol{a})\right]$. To prove identification for flexible outcome models, we can achieve identification by assuming additive errors (Hoyer et al., 2009), composing invertible nonlinear mappings with additive errors (Zhang and Hyvärinen, 2010; Zhang et al., 2016), appealing to the mutual independence of the factor model error terms and the out-

come model error terms (Carroll et al., 2010; Hu, 2008; Hu and Ridder, 2012; Hu and Sasaki, 2015; Hu and Schennach, 2008; Hu and Shum, 2012; Schennach, 2004; Schennach and Hu, 2013), or relying on the incongruence between smooth causal relationships and non-smooth observed variable distributions (Delaigle et al., 2016).

Understanding the practical implications of the deconfounder assumptions Most theoretical results around the deconfounder relies on the pinpointability assumption (Assumption 4.1.2), which roughly requires that the unobserved confounder non-trivially affect an infinite number of causes (or outcomes). This assumption can be too strong to be satisfied in practice. More generally, all of the deconfounder assumptions may be violated in real applications (Grimmer et al., 2020). Studying the implications of these assumptions across applications will be a worthwhile exercise. Developing sensitivity analysis for these assumptions will also be a crucial addition to the deconfounder workflow. Moreover, we shall seek to weaken these assumptions for particular settings: What assumptions can we remove if we *a priori* know the causal effects are sparse? What if we only need to test whether the causal effects are nonzero as opposed to estimating them?

Partial identification with the deconfounder In multiple causal inferences with only a few causes, the pinpointability assumption of the deconfounder cannot be satisfied. In such cases, the deconfounder will not be able to identify the causal quantities of interest in general. Instead, it can only partially identify them, namely, the causal quantities cannot be accurately estimated but only bounded with an interval. Section 3.4 suggests that the deconfounder should handle this setting by properly assessing the uncertainty of the deconfounder estimates. Specifically, it should propagate the uncertainty in estimating the substitute confounder and compute posterior credible intervals of the causal quantities. This procedure for uncertainty quantification is justified by Gustafson (2010, 2015), but theoretical guarantees for its coverage probability remain absent.

More broadly, the behavior of the deconfounder transitioning from infinite causes (identified) to finite causes (partially identified) is barely explored. Unlike the usual dichotomy between non-identification and identification, the deconfounder smoothly transition from the non-identified setting to the identified. Loosely, as the number of causes increases, the identification uncertainty of the substitute confounder decreases in the deconfounder, and the causal quantities of interest become "more identified." A rigorous study of this transition can help bridge existing causal identification strategies to partially-identified settings.

Causal estimation with the deconfounder The theory of the deconfounder has focused on causal identification, characterizing the behavior of the deconfounder given an infinite amount of data. However, its asymptotic properties and finite-sample properties have not been studied in general. Along this line, Grimmer et al. (2020) establish the consistency of the deconfounder with linear factor models and linear outcomes models. Moreover, they show that naive regression also achieves consistency under the same assumptions. General asymptotic results of the deconfounder will add to our understanding of the similarities and differences between the deconfounder and traditional methods.

Another interesting direction is to improve the estimation efficiency of the deconfounder. The deconfounder operates in a two-stage fashion and completely ignores the outcome in the first stage. While this practice is designed to promote the objectivity of the study, the deconfounder will suffer in its estimation efficiency. The reason is that the substitute confounder may capture multi-cause variables that do not affect the outcome; conditioning on these variables in the second stage inflates the estimation variance. Therefore, techniques like sufficient dimensionality reduction (Adragni and Cook, 2009; Cook, 2018; Globerson and Tishby, 2003) shall be used in the second stage to improve estimation efficiency.

Finally, the deconfounder can be viewed as a blend of Bayesian hierarchical modeling with causal inference. It considers multiple parallel causal inference problems simultaneously, allowing them to share strength with each other. Following this thinking, classical empirical Bayes ideas like Stein's estimation (Efron, 2012) may also help improve estimation efficiency.

Extensions of the deconfounder to new settings The deconfounder algorithm cannot handle causally-dependent causes or outcomes, because the deconfounder requires that no causal connections exist among the multiple causes (or outcomes). However, many applied causal inference problems violate this requirement. For example, causally dependent causes, longitudinal settings, sequential decision making, and networked data all have causal connections between its causes. Extending the deconfounder to these settings would broaden the applications of the deconfounder. Along this line, Bica et al. (2019) develop the time series deconfounder for longitudinal settings, and Guo et al. (2020) develop the network deconfounder for networked observational data.

Empirical applications of the deconfounder Though targeting multiple causal inference in general, deconfounder-type algorithms have mostly been used only in genome-wide association studies (GWAS) (Astle et al., 2009; GTEx Consortium et al., 2017; Hao et al., 2015; Kang et al., 2010; Price et al., 2006; Pritchard et al., 2000b; Song et al., 2015; Tran and Blei, 2017; Yu et al., 2006), econometrics (Bai and Ng, 2006; Cheng and Hansen, 2015; Gonçalves and Perron, 2014; Stock and Watson, 2016), and multiple testing (Fan et al., 2019; Friguet et al., 2009). Some researchers have recently explored the suitability of the deconfounder in other applications. For example, Athey et al. (2019) consider the deconfounder in the context of supply and demand models. Zhang et al. (2019) explore the deconfounder in medical applications. Saini et al. (2019) adapt the deconfounder to evaluate the next best set of marketing actions in digital marketing. Pölsterl and Wachinger (2020) apply the deconfounder to control for unknown confounders in neuro-imaging data. Verma et al. (2020a,b) use the deconfounder to estimate how stylistic attributes of a textual piece can causally affect reader consumption.

Structured causal inference The core idea of the deconfounder is to use the shared confounding structure of the observational data to deconfound; we fit Bayesian factor models to extract information about the confounders shared among multiple causes. At a higher level, the deconfounder capitalizes on the structure of the problem to help with causal inference. The same idea can be applied to causal inference problems with other structures, including hierarchical confounding structures, multiple independent decision makers, and networked structures. Along this line, Witty et al. (2020) develop a hierarchical Bayesian model for latent confounders shared across observations.

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132.

Abadie, A. and L'Hour, J. (2017). A penalized synthetic control estimator for disaggregated data. *Work. Pap., Mass. Inst. Technol., Cambridge, MA*.

Abraham, S. and Sun, L. (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Available at SSRN 3158747*.

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405.

Airoldi, E., Blei, D. M., Erosheva, E., and Fienberg, S., editors (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41(1):15–34.

Allman, E., Matias, C., and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37:3099–3132.

Amjad, M., Misra, V., Shah, D., and Shen, D. (2019). mrsc: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):37.

Amjad, M., Shah, D., and Shen, D. (2018). Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

Angrist, J. and Imbens, G. (1995). Identification and estimation of local average treatment effects.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2019). Synthetic difference in differences. Technical report, National Bureau of Economic Research.

Arnold, B. F., Ercumen, A., Benjamin-Chung, J., and Colford Jr, J. M. (2016). Brief report: negative controls to detect selection bias and measurement bias in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 27(5):637.

Astle, W., Balding, D. J., et al. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2018a). Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Athey, S., Imbens, G. W., and Pollmann, M. (2019). Comment on: "The blessings of multiple causes" by Yixin Wang and David M. Blei. *Journal of the American Statistical Association*, 114(528):1602–1604.

Athey, S., Imbens, G. W., and Wager, S. (2018b). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B*.

Bai, J. and Li, K. (2012). Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40:436–465.

Bai, J. and Ng, S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, 74(4):1133–1150.

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.

Bauer, S., Schölkopf, B., and Peters, J. (2016). The arrow of time in multivariate time series. In *International Conference on Machine Learning*, pages 2043–2051.

Bayarri, M. and Berger, J. O. (2000). P values for composite null models. *Journal of the American Statistical Association*, 95(452):1127–1142.

Bayarri, M. and Castellanos, M. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22:322–343.

Ben-Michael, E., Feller, A., and Rothstein, J. (2019). Synthetic controls and weighted event studies with staggered adoption. *arXiv preprint arXiv:1912.03290*.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Bica, I., Alaa, A. M., and van der Schaar, M. (2019). Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders. *arXiv preprint arXiv:1902.00450*.

Blei, D. M. (2014). Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent (d)irichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Bonhomme, S. and Sauder, U. (2011). Recovering distributions in difference-in-differences models: A comparison of selective and comprehensive schooling. *Review of Economics and Statistics*, 93(2):479–494.

Bonner, S. and Vasile, F. (2018). Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 104–112. ACM.

Botosaru, I. and Gutierrez, F. H. (2018). Difference-in-differences when the treatment status is observed in only one period. *Journal of Applied Econometrics*, 33(1):73–90.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L., et al. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1):247–274.

Callaway, B., Li, T., and Oka, T. (2018). Quantile treatment effects in difference in differences models under dependence restrictions and with only two time periods. *Journal of Econometrics*, 206(2):395–413.

Callaway, B. and Sant'Anna, P. H. (2019). Difference-in-differences with multiple time periods. *Available at SSRN 3148250*.

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review*, 43(2):245–257.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

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.

Carroll, R. J., Chen, X., and Hu, Y. (2010). Identification and estimation of nonlinear models using two samples with nonclassical measurement errors. *Journal of nonparametric statistics*, 22(4):379–399.

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorization models. *Computational Intelligence and Neuroscience*, 2009.

Cengiz, D., Dube, A., Lindner, A., and Zipperer, B. (2019). The effect of minimum wages on low-wage jobs: Evidence from the united states using a bunching estimator. Technical report, National Bureau of Economic Research.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B*, 78(3):673–700.

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

Chen, Y., Li, X., and Zhang, S. (2019). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*.

Cheng, X. and Hansen, B. E. (2015). Forecasting with factor-augmented regression: A frequentist model averaging approach. *Journal of Econometrics*, 186(2):280–293.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017a). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2017b). An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*.

Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018). Practical and robust t-test based inference for synthetic control and related methods.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405):51.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E. (2019). Sensitivity analysis of linear structural causal models. In *International Conference on Machine Learning*, pages 1252–1261.

Collins, M., Dasgupta, S., and Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, pages 617–624.

Cook, R. D. (2018). Principal components, sufficient dimension reduction, and envelopes.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

D'Amour, A. (2018). (Non-)identification in latent confounder models. http://www.alexdamour.com/blog/public/2018/05/18/non-identification-in-latent-confounder-models/. Accessed: 2018-05-29.

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

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

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

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2017). Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

Darnell, G., Georgiev, S., Mukherjee, S., and Engelhardt, B. E. (2017). Adaptive randomized dimension reduction on massive data. *Journal of Machine Learning Research*, 18(140):1–30.

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

Dawid, A. P., Didelez, V., et al. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231.

De Chaisemartin, C. and D'HaultfŒuille, X. (2017). Fuzzy differences-in-differences. *The Review of Economic Studies*, 85(2):999–1028.

DeGroot, M. and Schervish, M. (2012). *Probability and Statistics*. Addison-Wesley.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.

Delaigle, A., Hall, P., et al. (2016). Methodology for non-parametric deconvolution when the error distribution is unknown. *JR Stat. Soc. Ser. B. Stat. Methodol*, 78(1):231–252.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Dey, D., Gelfand, A., Swartz, T., and Vlachos, P. (1998). Simulation based model checking for hierarchical models. *Test*.

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

Ding, P., Li, F., et al. (2018). Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237.

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Dube, A. and Zipperer, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies.

Dusetzina, S. B., Brookhart, M. A., and Maciejewski, M. L. (2015). Control outcomes and exposures for improving internal validity of nonrandomized studies. *Health services research*, 50(5):1432–1451.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Egami, N. (2018). Identification of causal diffusion effects using stationary causal directed acyclic graphs. *arXiv preprint arXiv:1810.07858*.

Erosheva, E. A. (2003). Bayesian estimation of the grade of membership model. *Bayesian Statistics*, 7:501–510.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Resources*, 7(4):574–578.

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach.

Fan, J., Ke, Y., Sun, Q., and Zhou, W.-X. (2019). Farmtest: Factor-adjusted robust multiple testing with approximate false discovery control. *Journal of the American Statistical Association*, 114(528):1880–1893.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697.

Ferman, B. (2019). On the properties of the synthetic control estimator with many periods and many controls.

Flanders, W. D., Strickland, M. J., and Klein, M. (2017). A new method for partial correction of residual confounding in time-series and other observational studies. *American Journal of Epidemiology*, 185(10):941–949.

Franks, A., D'Amour, A., and Feller, A. (2019). Flexible sensitivity analysis for observational studies without observable implications. *Journal of the American Statistical Association*, (just-accepted):1–38.

Friguet, C., Kloareg, M., and Causeur, D. (2009). A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406–1415.

Frot, B., Nandy, P., and Maathuis, M. H. (2019). Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):459–487.

Geisser, S., Hodges, J., Press, S., and ZeUner, A. (1990). The validity of posterior expansions based on laplace method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 7:473.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541.

Globerson, A. and Tishby, N. (2003). Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3(Mar):1307–1331.

Gobillon, L. and Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. *Review of Economics and Statistics*, 98(3):535–551.

Gonçalves, S. and Perron, B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, 182(1):156–173.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2013). Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*.

Gopalan, P. K., Charlin, L., and Blei, D. M. (2014). Content-based recommendations with Poisson factorization. In *Advances in Neural Information Processing Systems*, pages 3176–3184.

Grimmer, J., Knox, D., and Stewart, B. (2020). Naïve regression requires weaker assumptions than factor models to adjust for multiple cause confounding. http://bit.ly/gks_paper.

GTEx Consortium, Battle*, A., Brown*, C. D., Engelhardt*, B. E., and Montgomery*, S. M. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550:204–213.

Guo, R., Li, J., and Liu, H. (2020). Learning individual causal effects from networked observational data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 232–240.

Gustafson, P. (2010). Bayesian inference for partially identified models. *The International Journal of Biostatistics*, 6(2).

Gustafson, P. (2015). *Bayesian inference for partially identified models: Exploring the limits of limited data*, volume 140. CRC Press.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 83–100.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Hainmueller, J. (2012). Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hao, W., Song, M., and Storey, J. D. (2015). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721.

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

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. Technical report, National bureau of economic research.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, 81(1):1–14.

Hernán, M. A. and Robins, J. M. (2020). Causal inference: what if. *Boca Raton: Chapman & Hill/CRC*, 2020.

Hernández-Lobato, J. M., Houlsby, N., and Ghahramani, Z. (2014). Probabilistic matrix factorization with non-random missing data. In *International Conference on Machine Learning*, pages 1512–1520.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, 226164:73–84.

Hirshberg, D. A. and Wager, S. (2018). Augmented minimax linear estimation. *arXiv:1712.00038*.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Holland, P. W., Glymour, C., and Granger, C. (1985). Statistics and causal inference. *ETS Research Report Series*, 1985(2).

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696.

Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, 27(5):705–740.

Hu, Y. (2008). Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. IEEE.

Hu, Y. and Ridder, G. (2012). Estimation of nonlinear models with mismeasured regressors using marginal information. *Journal of Applied Econometrics*, 27(3):347–385.

Hu, Y. and Sasaki, Y. (2015). Closed-form estimation of nonparametric models with non-classical measurement errors. *Journal of Econometrics*, 185(2):392–408.

Hu, Y. and Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216.

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

Hu, Y. and Shum, M. (2012). Nonparametric identification of dynamic models with unobserved state variables. *Journal of Econometrics*, 171(1):32–44.

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

Imai, K. and Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? *American Journal of Political Science*, 63(2):467–490.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B*, 76(1):243–263.

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

Imbens, G. and Rubin, D. (2015a). *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Imbens, G. W. and Rubin, D. B. (2015b). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

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

Janzing, D. and Schölkopf, B. (2018b). Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810*.

Kaltenpoth, D. and Vreeken, J. (2019). We are not your real parents: Telling causal from confounded using mdl. *arXiv preprint arXiv:1901.06950*.

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Kasza, J., Wolfe, R., and Schuster, T. (2017). Assessing the impact of unmeasured confounding for binary outcomes using confounding functions. *International Journal of Epidemiology*, 46(4):1303–1311.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kinn, D. (2018). Synthetic control methods and big data. *arXiv preprint arXiv:1803.00096*.

Kong, D., Yang, S., and Wang, L. (2019). Multi-cause causal inference with unmeasured confounding and binary outcome. *arXiv preprint arXiv:1907.13323*.

Kruskal, J. (1989). Rank, decomposition, and uniqueness for three-way and n-way arrays. In *Multiway Data Analysis*. Elsevier Science Publishers B.V.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

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

Laird, N. M. and Louis, T. A. (1982). Approximate posterior distributions for incomplete data problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–200.

Lanes, S. F. (1988). The logic of causal inference. *Causal Inference. ERI, Boston*, pages 59–75.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labor Market Policies*, pages 43–58. Springer.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Li, K. T. (2019). Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, pages 1–16.

Li, L., Chen, S., Kleban, J., and Gupta, A. (2015). Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, pages 929–934. ACM.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM.

Liang, D., Charlin, L., McInerney, J., and Blei, D. M. (2016). Modeling user exposure in recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 951–961. International World Wide Web Conferences Steering Committee.

Ling, G., Yang, H., Lyu, M. R., and King, I. (2012). Response aware model-based collaborative filtering. *arXiv preprint arXiv:1210.4869*.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, 8(10):833.

Lipsitch, M., Tchetgen, E. T., and Cohen, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)*, 21(3):383.

Liu, F. and Chan, L. (2018). Confounder detection in high dimensional linear models using first moments of spectral measures. *arXiv preprint arXiv:1803.06852*.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284.

Lopez, M. J., Gutman, R., et al. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6449–6459.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

Madigan, D., Stang, P. E., Berlin, J. A., Schuemie, M., Overhage, J. M., Suchard, M. A., Dumouchel, B., Hartzema, A. G., and Ryan, P. B. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1:11–39.

Malani, A. and Reif, J. (2010). Accounting for anticipation effects: an application to medical malpractice tort reform. Technical report, National Bureau of Economic Research.

Malani, A. and Reif, J. (2015). Interpreting pre-trends as anticipation: Impact on estimated treatment effects from tort reform. *Journal of Public Economics*, 124:1–17.

Mankiw, N. G. (2020). *Principles of economics*. Cengage Learning.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. (2010). Partial identification in econometrics. In *Microeconometrics*, pages 178–188. Springer.

Marlin, B. M. (2004). Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 627–634.

Marlin, B. M. and Zemel, R. S. (2009). Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, pages 5–12. ACM.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.

McCullagh, P. (2018). *Generalized linear models*. Routledge.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, vol. 37 of monographs on statistics and applied probability.

Mckeigue, P., Krohn, J., Storkey, A. J., and Agakov, F. V. (2010). Sparse instrumental variables (spiv) for genome-wide studies. In *Advances in Neural Information Processing Systems*, pages 28–36.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker.

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.

Miao, W. and Tchetgen, E. T. (2018). A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*.

Miao, W. and Tchetgen Tchetgen, E. (2017). Invited commentary: bias attenuation and identification of causal effects with multiple negative controls. *American Journal of Epidemiology*, 185(10):950–953.

Millán, J., Pintó, X., Muñoz, A., Zúñiga, M., Rubiés-Prat, J., Pallardo, L. F., Masana, L., Mangas, A., Hernández-Mijares, A., González-Santos, P., et al. (2009). Lipoprotein ratios: physiological significance and clinical usefulness in cardiovascular prevention. *Vascular health and risk management*, 5:757.

Mnih, A. and Salakhutdinov, R. R. (2008). Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1257–1264.

Moghaddass, R., Rudin, C., and Madigan, D. (2016). The factorized self-controlled case series method: An approach for estimating the effects of many drugs on many outcomes. *Journal of Machine Learning Research*, 17(185):1–24.

Mohamed, S., Ghahramani, Z., and Heller, K. A. (2009). Bayesian exponential family pca. In *Advances in Neural Information Processing Systems*, pages 1089–1096.

Mohamed, S. and Lakshminarayanan, B. (2016). Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*.

Moneta, A., Entner, D., Hoyer, P. O., and Coad, A. (2013). Causal inference by independent component analysis: Theory and applications. *Oxford Bulletin of Economics and Statistics*, 75(5):705–730.

Mooij, J. M., Stegle, O., Janzing, D., Zhang, K., and Schölkopf, B. (2010). Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695.

Morgan, S. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition.

Neal, R. M. (1990). Learning stochastic feedforward networks. *Department of Computer Science, University of Toronto*, 64(9).

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

Ogburn, E., Shpitser, I., and Tchetgen Tchetgen, E. (2020). Counterexamples to "The blessings of multiple causes" by Wang and Blei. *arXiv:2020.001*.

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

Ogburn, E. L., Shpitser, I., and Tchetgen, E. J. T. (2019b). Comment on" blessings of multiple causes". *arXiv preprint arXiv:1910.05438*.

O'Neill, S., Kreif, N., Grieve, R., Sutton, M., and Sekhon, J. S. (2016). Estimating causal effects: considering three alternatives to difference-in-differences estimation. *Health Services and Outcomes Research Methodology*, 16(1-2):1–21.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems.

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

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Peters, J., Janzing, D., Gretton, A., and Schölkopf, B. (2009). Detecting the direction of causal time series. In *Proceedings of the 26th annual international conference on machine learning*, pages 801–808.

Peters, J., Janzing, D., and Schölkopf, B. (2013). Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

Pölsterl, S. and Wachinger, C. (2020). Controlling for unknown confounders in neuroimaging. *arXiv preprint arXiv:2006.13135*.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000a). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000b). Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181.

Qin, J., Zhang, and Biao (2008). Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):329–349.

Ranganath, R. and Blei, D. M. (2019). Population predictive checks. *arXiv:1908.00882*.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

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

Ranganath, R., Tang, L., Charlin, L., and Blei, D. M. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771.

Ranganath, R., Tran, D., and Blei, D. M. (2016). Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333.

Rassen, J. A., Solomon, D. H., Glynn, R. J., and Schneeweiss, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety*, 20(7):675–683.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and variational inference in deep latent gaussian models. In *International Conference on Machine Learning*, volume 2.

Richardson, D. B., Keil, A., Tchetgen, T. E., and Cooper, G. (2015). Negative control outcomes and the analysis of standardized mortality ratios. *Epidemiology (Cambridge, Mass.)*, 26(5):727.

Robbins, M. W., Saunders, J., and Kilmer, B. (2017). A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *Journal of the American Statistical Association*, 112(517):109–126.

Robins, J., van der Vaart, A., and Ventura, V. (2000a). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, 95(452):1143–1156.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000b). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94. Springer.

Rosenbaum, P. R. (1989). The role of known effects in observational studies. *Biometrics*, 45(2):557–569.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Saini, S. K., Dhamnani, S., Ibrahim, A. A., and Chavan, P. (2019). Multiple treatment effect estimation using deep generative model with task embedding. In *The World Wide Web Conference*, pages 1601–1611.

Samartsidis, P., Seaman, S. R., Presanis, A. M., Hickman, M., De Angelis, D., et al. (2019). Assessing the causal effect of binary interventions from observational panel data with few treated units. *Statistical Science*, 34(3):486–503.

Sanderson, E., Macdonald-Wallis, C., and Davey Smith, G. (2017). Negative control exposure studies in the presence of measurement error: implications for attempted effect estimate calibration. *International Journal of Epidemiology*, 47(2):587–596.

Schennach, S. M. (2004). Nonparametric regression in the presence of measurement error. *Econometric Theory*, 20(6):1046–1093.

Schennach, S. M. and Hu, Y. (2013). Nonparametric identification and semiparametric estimation of classical measurement error models without side information. *Journal of the American Statistical Association*, 108(501):177–186.

Schmidt, M. N., Winther, O., and Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pages 540–547. Springer.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352*.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., and Suchard, M. A. (2016). Robust empirical calibration of p-values using observational data. *Statistics in medicine*, 35(22):3883.

Schuemie, M. J., Hripcsak, G., Ryan, P. B., Madigan, D., and Suchard, M. A. (2018). Empirical confidence interval calibration for population-level effect estimation studies in observational healthcare data. *Proceedings of the National Academy of Sciences*, 115(11):2571–2577.

Shah, R. D. and Meinshausen, N. (2018). Rsvp-graphs: Fast high-dimensional covariance matrix estimation under latent confounding. *arXiv preprint arXiv:1811.01076*.

Sharma, A., Hofman, J. M., and Watts, D. J. (2016). Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv preprint arXiv:1611.09414*.

Shi, X., Miao, W., and Tchetgen, E. T. (2018). Multiply robust causal inference with double negative control adjustment for unmeasured confounding. *arXiv preprint arXiv:1808.04906*.

Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98.

Silva, R., Scheine, R., Glymour, C., and Spirtes, P. (2006). Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(Feb):191–246.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC.

Smith, A. and Elkan, C. (2004). A Bayesian network framework for reject inference. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 286–295. ACM.

Sofer, T., Richardson, D. B., Colicino, E., Schwartz, J., and Tchetgen, E. J. T. (2016). On negative outcome control of unobserved confounding as a generalization of difference-in-differences. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 31(3):348.

Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554.

Srebro, N. and Salakhutdinov, R. R. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems*, pages 2056–2064.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of Macroeconomics*, volume 2, pages 415–525. Elsevier.

Tchetgen Tchetgen, E. (2013). The control outcome calibration approach for causal inference with unobserved confounding. *American Journal of Epidemiology*, 179(5):633–640.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

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

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

Tran, D., Hoffman, M. D., Saurous, R. A., Brevdo, E., Murphy, K., and Blei, D. M. (2017). Deep probabilistic programming. In *International Conference on Learning Representations*.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016a). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

Tran, D., Ruiz, F. J., Athey, S., and Blei, D. M. (2016b). Model criticism for Bayesian causal inference. *arXiv preprint arXiv:1610.09037*.

US Department of Health and Human Services Public Health service (1987). National medical expenditure survey series (nmes).

Vanchinathan, H. P., Nikolic, I., De Bona, F., and Krause, A. (2014). Explore-exploit in top-n recommender systems via gaussian processes. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 225–232. ACM.

Verma, G., BV, S., Sharma, S., and Srinivasan, B. V. (2020a). Generating need-adapted multimodal fragments. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 335–346.

Verma, G., Srinivasan, B. V., Saini, S. K., and Chhaya, N. (2020b). Modeling causal impact of textual style on a targeted goal. In *Companion Proceedings of the Web Conference 2020*, pages 59–60.

Veturi, Y., de los Campos, G., Yi, N., Huang, W., Vazquez, A. I., and Kühnel, B. (2019). Modeling heterogeneity in the genetic architecture of ethnically diverse groups using random effect interaction models. *Genetics*, pages genetics–301909.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Viviano, D. and Bradic, J. (2019). Synthetic learner: model-free inference on treatments over time. *arXiv preprint arXiv:1904.01490*.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

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

Wang, Y. and Blei, D. M. (2019b). The blessings of multiple causes: A reply to Ogburn et al.(2019). *arXiv preprint arXiv:1910.07320*.

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

Wang, Y. and Blei, D. M. (2019d). Multiple causes: A causal graphical view. *arXiv preprint arXiv:1905.12793*.

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

Wang, Y., Liang, D., Charlin, L., and Blei, D. M. (2020). Causal inference on recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*.

Wang, Y. and Zubizarreta, J. R. (2019). Large sample properties of matching for balance. *arXiv preprint arXiv:1905.11386*.

Wang, Y. and Zubizarreta, J. R. (2020). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*, 107(1):93–105.

Witty, S., Takatsu, K., Jensen, D., and Mansinghka, V. (2020). Causal inference using gaussian processes with structured latent confounders.

Wong, R. K. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M., and Price, A. L. (2014). Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics*, 46(2):100.

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

Yiu, S. and Su, L. (2018). Covariate association eliminating weights: a unified weighting framework for causal effect estimation. *Biometrika*.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203.

Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press.

Zhang, K. and Hyvärinen, A. (2010). Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*, pages 157–164.

Zhang, K., Zhang, J., Huang, B., Schölkopf, B., and Glymour, C. (2016). On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *UAI*.

Zhang, L., Wang, Y., Ostropolets, A., Mulgrave, J. J., Blei, D. M., and Hripcsak, G. (2019). The medical deconfounder: Assessing treatment effects with electronic health records. In *Machine Learning for Healthcare Conference*, pages 490–512.

Zhao, Q. (2018). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*, page in press.

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1).

Zhao, X., Zhang, W., and Wang, J. (2013). Interactive collaborative filtering. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 1411–1420. ACM.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

A

Proofs for Chapter 4

A.1 Proof of Theorem 4.2.1

Proof. For notational simplicity, denote $\boldsymbol{a} = (a_1, \ldots, a_m)$; $\boldsymbol{a}' = (a'_1, \ldots, a'_m)$; $\boldsymbol{A}_i = (A_{i1}, \ldots, A_{im})$. We also write $f_\theta(\cdot) = f(\cdot)$.

We start with rewriting $\mathbb{E}_Y[Y_i(\boldsymbol{a})] - \mathbb{E}_Y[Y_i(\boldsymbol{a}')]$ using the unconfoundedness assumption and the separability assumption.

First notice that

$$\begin{aligned}\mathbb{E}_Y[Y_i(\boldsymbol{a})] =& \mathbb{E}_{Z,X}[\mathbb{E}_Y[Y_i(\boldsymbol{a}) \mid X_i, Z_i]] && \text{(A.1)}\\ =& \mathbb{E}_X[f_1(\boldsymbol{a}, X_i)] + \mathbb{E}_Z[f_2(Z_i)]. && \text{(A.2)}\end{aligned}$$

The first equality is due to the tower property. The second equality is due to the separability assumption. The third equality is due to the linearity of expectations.

Hence we have

$$\mathbb{E}_Y\left[Y_i(\boldsymbol{a})\right] - \mathbb{E}_Y\left[Y_i(\boldsymbol{a}')\right] = \mathbb{E}_X\left[f_1(\boldsymbol{a}, X_i)\right] - \mathbb{E}_X\left[f_1(\boldsymbol{a}', X_i)\right] \tag{A.3}$$

$$= \int_{C(\boldsymbol{a},\boldsymbol{a}')} \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_1(\boldsymbol{a}, X_i)\right] \mathrm{d}\boldsymbol{a}, \tag{A.4}$$

where $C(\boldsymbol{a}, \boldsymbol{a}')$ is a line where $\boldsymbol{a}$ and $\boldsymbol{a}'$ are the end points. The second equality is due to the fundamental theorem of calculus.

Next, we see how the gradient of the potential outcome function $\nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_1(\boldsymbol{a}, X_i)\right]$ relates to the gradient of the outcome model we fit. The key idea here is that the two gradients are equal in regions $\{\boldsymbol{a} : f(\boldsymbol{a}) = c\}$ for each c.

We will rely on the pinpointability assumption. Notice that, for almost all $\boldsymbol{a}$, we have

$$\nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_1(\boldsymbol{a})\right] = \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_3(\boldsymbol{a})\right] \tag{A.5}$$

It is due to two observations. The first observation is that

$$\nabla_{\boldsymbol{a}} \mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i \mid Z_i = f(\boldsymbol{a}), A_i = \boldsymbol{a}, X_i\right]\right] \tag{A.6}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i(\boldsymbol{a}) \mid Z_i = f(\boldsymbol{a}), A_i = \boldsymbol{a}, X_i\right]\right] \tag{A.7}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i(\boldsymbol{a}) \mid Z_i = f(\boldsymbol{a}), X_i\right]\right] \tag{A.8}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_1(\boldsymbol{a}, X_i)\right] + \nabla_{\boldsymbol{a}} f_2(f(\boldsymbol{a})) \tag{A.9}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_1(\boldsymbol{a}, X_i)\right] + \nabla_{f(\boldsymbol{a})} f_2 \cdot \nabla_{\boldsymbol{a}} f(\boldsymbol{a}) \tag{A.10}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X\left[f_1(\boldsymbol{a}, X_i)\right] \tag{A.11}$$

The first equality is due to stable unit treatment value assumption (SUTVA). The second equality is due to is due to Proposition 4.1.1: $Y_i(\boldsymbol{a}) \perp \boldsymbol{A}_i \mid X_i, Z_i$. The third equality is due to the separability condition. The fourth equality is due to the chain rule. The fifth equality is due to $\nabla_{\boldsymbol{a}} f(\boldsymbol{a}) = 0$ up to a set of Lebesgue measure zero.

The second observation is that

$$\nabla_{\boldsymbol{a}} \mathbb{E}_X \left[\mathbb{E}_Y \left[Y_i \mid Z_i = f(\boldsymbol{a}), \boldsymbol{A}_i = \boldsymbol{a}, X_i \right] \right] \tag{A.12}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X \left[f_3(\boldsymbol{a}, X_i) \right] + \nabla_{\boldsymbol{a}} f_4(f(\boldsymbol{a})) \tag{A.13}$$

$$= \nabla_{\boldsymbol{a}} \mathbb{E}_X \left[f_3(\boldsymbol{a}, X_i) \right] \tag{A.14}$$

Hence Equation (A.5) is true because f_1 and f_3 are continuously differentiable.

Therefore, we have

$$\mathbb{E}_Y \left[Y_i(\boldsymbol{a}) \right] - \mathbb{E}_Y \left[Y_i(\boldsymbol{a}') \right] \tag{A.15}$$

$$= \int_{C(\boldsymbol{a}, \boldsymbol{a}')} \nabla_{\boldsymbol{a}} \mathbb{E}_X \left[f_1(\boldsymbol{a}, X_i) \right] \mathrm{d}\boldsymbol{a} \tag{A.16}$$

$$= \int_{C(\boldsymbol{a}, \boldsymbol{a}')} \nabla_{\boldsymbol{a}} \mathbb{E}_X \left[f_3(\boldsymbol{a}, X_i) \right] \mathrm{d}\boldsymbol{a} \tag{A.17}$$

$$= \mathbb{E}_X \left[f_3(\boldsymbol{a}, X_i) \right] - \mathbb{E}_X \left[f_3(\boldsymbol{a}', X_i) \right] \tag{A.18}$$

$$= \left(\mathbb{E}_X \left[f_3(\boldsymbol{a}, X_i) \right] + \mathbb{E} \left[f_4(Z_i) \right] \right) - \left(\mathbb{E}_X \left[f_3(\boldsymbol{a}', X_i) \right] + \mathbb{E} \left[f_4(Z_i) \right] \right) \tag{A.19}$$

$$\begin{aligned} &= \int \mathbb{E}_Y \left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}', X_i, Z_i \right] P(Z_i, X_i) \, \mathrm{d}Z_i \, \mathrm{d}X_i \\ &\quad - \int \mathbb{E}_Y \left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, X_i, Z_i \right] P(Z_i, X_i) \, \mathrm{d}Z_i \, \mathrm{d}X_i \end{aligned} \tag{A.20}$$

$$= \mathbb{E}_{Z,X} \left[\mathbb{E}_Y \left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i, X_i \right] \right] - \mathbb{E}_{Z,X} \left[\mathbb{E}_Y \left[Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}', Z_i, X_i \right] \right]. \tag{A.21}$$

The first equality is due to Equation (A.4). The second equality is due to Equation (A.5). The third equality is due to the fundamental theorem of calculus. The fourth equality is due to simple algebra. The fifth equality is due to the separability condition.

□

A.2 Proof of Theorem 4.2.2

Proof. Proposition 4.1.1 ensures that the substitute confounder Z_i and the observed covariate X_i satisfies

$$(A_{i1}, \ldots, A_{im}) \perp\!\!\!\perp Y_i(a_{i1}, \ldots, a_{im}) \mid Z_i, X_i. \tag{A.22}$$

Therefore, we have

$$\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] \tag{A.23}$$

$$= \mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1, \ldots, a_k, A_{i,k+1}, \ldots, A_{im})\right]\right] \tag{A.24}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1, \ldots, a_k, A_{i,k+1}, \ldots, A_{im}) \mid Z_i, X_i\right]\right]\right] \tag{A.25}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_1, \ldots, a_k, A_{i,k+1}, \ldots, A_{im}) \mid Z_i, X_i, A_{i1} = a_1, \ldots, A_{ik} = a_k\right]\right]\right] \tag{A.26}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(A_{i1}, \ldots, A_{ik}, A_{i,k+1}, \ldots, A_{im}) \mid Z_i, X_i, A_{i1} = a_1, \ldots, A_{ik} = a_k\right]\right]\right] \tag{A.27}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i1} = a_1, \ldots, A_{ik} = a_k\right]\right]\right] \tag{A.28}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i1} = a_1, \ldots, A_{ik} = a_k\right]\right] \tag{A.29}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i,1:k} = a_{1:k}\right]\right] \tag{A.30}$$

The first equality is an expansion of the notations. The second equality is due to the tower property. The third equality is due to Equation (A.22). The fourth equality is due to $A_{i1} = a_1, \ldots, A_{ik} = a_k$. The fifth equality is due to SUTVA. The sixth equality is because the inner expectation does not depend on $A_{(k+1):m}$.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a_{1:k}, A_{i,(k+1):m})\right]\right] - \mathbb{E}_{A_{(k+1):m}}\left[\mathbb{E}_Y\left[Y_i(a'_{1:k}, A_{i,(k+1):m})\right]\right] \\
=&\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i,1:k} = a_{1:k}\right]\right] - \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i,1:k} = a'_{1:k}\right]\right]
\end{aligned}$$

by the linearity of expectation.

Finally, $\mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid Z_i, X_i, A_{i,1:k} = a_{1:k}\right]\right]$ can be estimated from the observed data because (1) $A_{i,1:k}$ satisfy overlap with respect to (Z_i, X_i) and (2) the substitute confounder Z can be pinpointed. □

A.3 Proof of Theorem 4.2.3

Proof. As with Theorem 4.2.1 and Theorem 4.2.2, Theorem 4.2.3 relies on the unconfoundedness given the substitute confounders Z_i and the observed covariates X_i due to Proposition 4.1.1.

Given this unconfoundedness, Theorem 4.2.3 identifies the mean potential outcome of an individual given its current cause assignment $A_i = (a_1, \ldots, a_m)$; it only requires that the new cause assignment of interest $(a'_1, \ldots, a'_m)$ lead to the same substitute confounder estimate: $f(a_1, \ldots, a_m) = f(a'_1, \ldots, a'_m)$.

To prove identification, we rewrite this conditional mean potential outcome

$$\mathbb{E}_Y\left[Y_i(a'_1, \ldots, a'_m) \mid A_{i1} = a_1, \ldots, A_{im} = a_m\right] \tag{A.31}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i(a'_1, \ldots, a'_m) \mid A_{i1} = a_1, \ldots, A_{im} = a_m, Z_i, X_i\right]\right] \tag{A.32}$$

$$= \mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i(a'_1, \ldots, a'_m) \mid A_{i1} = a_1, \ldots, A_{im} = a_m, Z_i = f(a_1, \ldots, a_m), X_i\right]\right] \tag{A.33}$$

$$= \mathbb{E}_X\left[\mathbb{E}_Y\left[Y_i(a'_1, \ldots, a'_m) \mid A_{i1} = a'_1, \ldots, A_{im} = a'_m, Z_i = f(a_1, \ldots, a_m), X_i\right]\right] \tag{A.34}$$

$$= \mathbb{E}_{Z,X}\left[\mathbb{E}_Y\left[Y_i \mid A_{i1} = a'_1, \ldots, A_{im} = a'_m, Z_i, X_i\right]\right] \tag{A.35}$$

The first equality is due to the tower property. The second equality is due to the pinpointability of the substitute confounder $P(Z_i \mid \boldsymbol{A}_i) = \delta_{f(\boldsymbol{A}_i)}$. The third equality is due to unconfoundedness given Z_i, X_i. The fourth equality is estimable from the data because $f(a_1, \ldots, a_m) = f(a'_1, \ldots, a'_m)$. Hence the identification of $\mathbb{E}_Y\left[Y_i(a'_1, \ldots, a'_m) \mid A_{i1} = a_1, \ldots, A_{im} = a_m\right]$ is established. We note that this identification result does not require overlap. □

B

Proofs for Chapter 5

B.1 Proof of Theorem 5.1.1

Proof. The proof of Theorem 5.1.1 relies on two observations. The first observation starts with the integral equation we solve:

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}} \tag{B.1}$$

$$= \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid u) P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}} \, \mathrm{d}u. \tag{B.2}$$

The first equality is due to Equation (5.3). The second equality is due to the conditional independence implied by Figure 5.1a: $A_{\mathcal{X}} \perp A_{\mathcal{C}}, f(a_{\mathcal{N}}) \mid U$.

The second observation relies on the null proxy:

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

$$= \int P(y \mid u, a_{\mathcal{C}}) P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}u. \tag{B.4}$$

The first equality is due to the definition of conditional probability. The second equality is due to the second part of Assumption 5.1.1, which implies $Y \perp f(a_{\mathcal{N}}) \mid U, A_{\mathcal{C}}$. The reason is that

$$P(y \mid u, a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int P(y \mid u, a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}))P(a_{\mathcal{X}} \mid u, a_{\mathcal{C}}, f(a_{\mathcal{N}}))\,\mathrm{d}a_{\mathcal{X}} \tag{B.5}$$

$$= \int P(y \mid u, a_{\mathcal{C}}, a_{\mathcal{X}})P(a_{\mathcal{X}} \mid u, a_{\mathcal{C}})\,\mathrm{d}a_{\mathcal{X}} \tag{B.6}$$

$$= P(y \mid u, a_{\mathcal{C}}). \tag{B.7}$$

In fact, it is sufficient to assume $Y \perp f(a_{\mathcal{N}}) \mid U, A_{\mathcal{C}}$ instead of $Y \perp f(a_{\mathcal{N}}) \mid U, A_{\mathcal{C}}, A_{\mathcal{X}}$ in Theorem 5.1.1. However, the latter is easier to check and interpret.

Comparing Equation (B.2) and Equation (B.4) gives

$$\int \left[P(y \mid u, a_{\mathcal{C}}) - \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}})P(a_{\mathcal{X}} \mid u)\,\mathrm{d}a_{\mathcal{X}} \right] \times P(u \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}))\,\mathrm{d}u = 0, \tag{B.8}$$

which, by the completeness condition in Assumption 5.1.1.2, implies

$$P(y \mid u, a_{\mathcal{C}}) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}})P(a_{\mathcal{X}} \mid u)\,\mathrm{d}a_{\mathcal{X}}. \tag{B.9}$$

Equation (B.9) leads to identification:

$$P(y \mid \mathrm{do}(a_{\mathcal{C}})) = \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}})P(a_{\mathcal{X}} \mid u)\,\mathrm{d}a_{\mathcal{X}}P(u)\,\mathrm{d}u \tag{B.10}$$

$$= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}})P(a_{\mathcal{X}})\,\mathrm{d}a_{\mathcal{X}}. \tag{B.11}$$

Consider the special case of a single cause as in Figure 5.1b. Let $a_{\mathcal{C}} = \{A_1\}, a_{\mathcal{X}} = \{X\}, a_{\mathcal{N}} = N$, and $f(a_{\mathcal{N}}) = N$. The above proof reduces to the identification proof for proxy variables (Theorem 1 of Miao et al. (2018)). □

B.2 Examples of Assumption 5.1.1

As an example, if the structural equation writes

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

where $\epsilon \perp U, A_1, \ldots, A_m$, then Assumption 5.1.1.1 is satisfied if A_1 and A_2 are identically Gaussian: $A_{\mathcal{N}} = (A_1, A_2)$ and $f(A_{\mathcal{N}}) = A_1 - A_2$ satisfies

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

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

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

for some constant α_1 and α_2.

Similarly, if the structural equation writes

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

where $\epsilon \perp U, A_1, \ldots, A_m$, then Assumption 5.1.1.1 is satisfied if A_1 and A_2 are identically log-normal: $A_{\mathcal{N}} = (A_1, A_2)$ and $f(A_{\mathcal{N}}) = A_1/A_2$ satisfies

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

As a final example, if the structural equation writes

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

where $\epsilon \perp U, A_1, \ldots, A_m$ and A_1, A_2 are both binary, then Assumption 5.1.1.1 is satisfied: $A_{\mathcal{N}} = (A_1, A_2)$ and $f(A_{\mathcal{N}}) = A_1 \text{ XOR } A_2$ satisfies

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

B.3 Example: A linear causal model

We illustrate Theorems 5.2.1 and 5.2.2 in a linear causal model.

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

In this example, the individual's lifestyle U^{mlt} is a multi-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 a healthy lifestyle are more likely to complete the survey.

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

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

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 5.2.1. 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 PPCA to the data about food consumption $A_{1:10}$; we do not model the outcome Y. Wang and Blei (2019a) 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,

$$
\hat{Z} = \left(\sum_{i=1}^{10} \gamma_{1i} A_i + \epsilon_{1\hat{Z}}, \sum_{i=1}^{10} \gamma_{2i} A_i + \epsilon_{2\hat{Z}} \right), \tag{B.12}
$$

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

This substitute confounder $\hat{Z}$ satisfies Assumption 5.2.3. Plausibly, the real world satisfies the variant of Assumption 5.2.1 and Assumption 5.2.2. 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 (5.17). Theorem 5.2.2 guarantees that this estimate is correct.

B.4 Proof of Theorem 5.1.2

Proof. Assumption 5.1.2.2 guarantees the existence of some function $\hat{h}$ such that

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}) \, \mathrm{d}a_{\mathcal{X}} \tag{B.13}$$

under weak regularity conditions. (We will discuss the reason in Appendix B.5.)

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

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) \, \mathrm{d}a_{\mathcal{X}}. \tag{B.14}$$

Given this claim (Equation (B.57)), 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 (5.6); the second is due to Equation (B.57); the third is due to the deconfounder estimate being consistent with the observed data distribution by construction; the fourth is due to the above claim (Equation (B.57)) and Theorem 5.1.1.

We next prove the claim (Equation (B.57)). 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 Equation (5.4) and the deconfounder estimate being consistent with the observed data; the second is due to Equation (B.13); the third is due to Assumption 5.1.2.1, which implies

$$\hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}) = \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}). \tag{B.15}$$

Similar to Assumption 5.1.1.1, it is sufficient to assume Equation (B.15) directly. However, Assumption 5.1.2.1 is easier to check and more interpretable; it directly relates to the deconfounder outcome model.

□

B.5 Existence of solutions to the integral equations

Theorem 5.1.1 involves solving the integral equation

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}})) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}))\, \mathrm{d}a_{\mathcal{X}}. \tag{B.16}$$

When does a solution exist for Equation (B.16)? We appeal to Proposition 1 of Miao et al. (2018).

Proposition B.5.1: . (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

$$P(y \mid z, x) = \int h(w, x, y) P(w \mid z, x)\, \mathrm{d}w \tag{B.17}$$

exists if

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

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

Leveraging Proposition B.5.1, we can establish sufficient conditions for existence of a solution to Equation (B.16).

Corollary B.5.1: . A solution exist for the integral equation Equation (B.16) 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 5.1.1.3; others are weak regularity conditions.

By the same token, we can establish sufficient conditions for solution existence of Equation (5.9), Equation (5.15). The same argument also applies to the integral equation involved in Theorem 5.2.2:

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s = 1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\mathrm{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\mathrm{sng}}, s = 1) \, \mathrm{d}a_{\mathcal{X}}. \tag{B.18}$$

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.

B.6 Proof of Lemma 5.2.1

The idea of the proof is to start with the structural equations of the general causal graph Figure 5.2a. Then posit the existence of a latent variable Z that renders all the causes conditionally independent; Figure 5.2b 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

$$P(y \mid u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \ldots, a_m, s = 1) = P(y \mid u^{\text{sng}}, z, a_1, \ldots, a_m, s = 1). \tag{B.19}$$

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

$$W_k = f_{W_k}(\epsilon_{W_k}), \qquad k = 1, \ldots, K, K \geq 0, \tag{B.20}$$

$$U_j = f_{U_j}(\epsilon_{U_j}), \qquad j = 1, \ldots, J, J \geq 0, \tag{B.21}$$

$$V_l = f_{V_l}(\epsilon_{V_l}), \qquad l = 1, \ldots, L, L \geq 0, \tag{B.22}$$

$$A_i = f_{A_i}(W_{S^W_{A_i}}, U_{S^U_{A_i}}, \epsilon_{A_i}), \qquad i = 1, \ldots, m, m \geq 2, \tag{B.23}$$

$$Y = f_y(A_1, \ldots, A_m, U_1, \ldots, U_K, V_1, \ldots V_L, \epsilon_Y), \tag{B.24}$$

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 $S^W_{A_i} \subset \{1, \ldots, K\}$ is an index set; if $S^W_{A_1} = \{1, 3, 4\}$, then $W_{S^W_{A_i}} = (W_1, W_3, W_4)$. The same notion applies to $S^U_{A_i} \subset \{1, \ldots, J\}$.

The notation in this structural equation model is consistent with the set up in Figure 5.2a. 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

$$W^{\text{sng}} \triangleq \{W_1, \ldots, W_K\} / \bigcup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}). \tag{B.25}$$

To parse the notation above, recall that $W_{S^W_{A_i}}$ is the set of null confounders that affects A_i. The set of variables $\bigcup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}}\cap W_{S^W_{A_j}})$ 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 5.2.1, we first prove the following lemma that quantifies the information in Z (in Figure 5.2b).

Lemma B.6.1: . The random variable Z in Figure 5.2b "captures" all multi-cause confounders, all multi-cause null confounders and some independent error:

$$\sigma(Z)=\sigma\left(\{\epsilon_Z\}\bigcup(\cup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}}\cap W_{S^W_{A_j}})\cup(U_{S^U_{A_i}}\cap U_{S^U_{A_j}}))\right), \tag{B.26}$$

$$=\sigma\left(\{\epsilon_Z\}\bigcup W^{\text{mlt}}\bigcup U^{\text{mlt}}\right). \tag{B.27}$$

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

We can parse the notation in Lemma B.6.1 in the same way as in Equation (B.25): The set of variables $\cup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}}\cap W_{S^W_{A_j}})$ denotes the set of all multi-cause confounders; $\cup_{i,j\in\{1,\dots,m\}:i\neq j}(U_{S^U_{A_i}}\cap U_{S^U_{A_j}})$ denotes the set of all multi-cause null confounders.

Proof. Without the loss of generality, we assume the compactness of representation in Equations (B.23) and (C.58). For any subset $\mathcal{S}$ of the random variables $\mathcal{S}\subset\{A_1,\dots,A_m,Y\}$, we assume the σ-algebra $\sigma(\bigcap_\tau(S^W_{S_\tau},S^U_{S_\tau},S^V_{S_\tau}))$ is the *smallest* σ-algebra that makes all the random variables in $\mathcal{S}$ jointly independent. The assumption is made for technical convenience. We simply ensure the arrows from the W,U,V's to the A_i's do exist. In other words, all the W,U,V's "whole-heartedly" contribute to the A_i's when they appear in Equation (B.23). 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:

$$\sigma\left(\bigcup_{i,j\in\{1,\dots,m\}:i\neq j}(W_{S^W_{A_i}}\cap W_{S^W_{A_j}})\cup(U_{S^U_{A_i}}\cap U_{S^U_{A_j}})\right)\subset\sigma(Z). \tag{B.28}$$

Consider any pair of A_i and A_j. Figure 5.2b implies that

$$A_i \perp A_j \mid Z, \tag{B.29}$$

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

$$A_i \perp A_j \mid \sigma\left((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right), \tag{B.30}$$

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

$$\sigma\left((W_{S^W_{A_i}} \cap W_{S^W_{A_j}}), (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{B.31}$$

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

$$\sigma\left(\bigcup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}})\right) \subset \sigma(Z). \tag{B.32}$$

Next Figure 5.2b implies $\sigma(A_1, \ldots, A_M) \not\subset \sigma(Z)$ and $\sigma(Y) \not\subset \sigma(Z)$.

Therefore, we have

$$\sigma(Z) \subset \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right), \tag{B.33}$$

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 (B.32) can be made equal to the right side of Equation (B.33). We have

$$\sigma(Z) = \sigma\left(\{\epsilon_Z\} \bigcup (\cup_{i,j \in \{1,\ldots,m\}: i \neq j} (W_{S^W_{A_i}} \cap W_{S^W_{A_j}}) \cup (U_{S^U_{A_i}} \cap U_{S^U_{A_j}}))\right) \tag{B.34}$$

$$= \sigma\left(\{\epsilon_Z\} \bigcup W^{\text{mlt}} \bigcup U^{\text{mlt}}\right). \tag{B.35}$$

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

As a direct consequence of Lemma B.6.1, we have

$$P(y \mid u^{\text{sng}}, u^{\text{mlt}}, w^{\text{mlt}}, a_1, \ldots, a_m, s = 1) = P(y \mid u^{\text{sng}}, z, a_1, \ldots, a_m, s = 1), \tag{B.36}$$

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

B.7 Proof of Lemma 5.2.2

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

The proof of Lemma 5.2.2 relies on two observations.

The first observation starts with the integral equation we solve:

$$\begin{aligned}
&P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1) && \text{(B.37)}\\
&= \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}_{\mathcal{C}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1)\, \mathrm{d}a_{\mathcal{X}} && \text{(B.38)}\\
&= \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u^{\text{sng}}_{\mathcal{C}}) P(a_{\mathcal{X}} \mid z) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1)\, \mathrm{d}a_{\mathcal{X}}\, \mathrm{d}z && \text{(B.39)}
\end{aligned}$$

The first equality is due to Equation (5.15). The second equality is due to Assumption 5.2.1.2.

The second observation relies on the null proxy:

$$\begin{aligned}
&P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1) && \text{(B.40)}\\
&= \int P(y \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1)\, \mathrm{d}z && \text{(B.41)}\\
&= \int P(y \mid z, a_{\mathcal{C}}, u^{\text{sng}}_{\mathcal{C}}, s = 1) P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u^{\text{sng}}_{\mathcal{C}}, s = 1)\, \mathrm{d}z && \text{(B.42)}
\end{aligned}$$

The first equality is due to the definition of conditional probability. The second equality is due to the second part of Assumption 5.2.2; it implies $Y \perp f(a_{\mathcal{N}}) \mid Z, U_{\mathcal{C}}^{\text{sng}}, A_{\mathcal{C}}, S = 1$. The reason is that

$$
\begin{aligned}
& P(y \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) && \text{(B.43)}\\
=& \int P(y \mid z, a_{\mathcal{C}}, a_{\mathcal{X}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) P(a_{\mathcal{X}} \mid z, a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \mathrm{d}a_{\mathcal{X}} && \text{(B.44)}\\
=& \int P(y \mid z, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}, s = 1) P(a_{\mathcal{X}} \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \mathrm{d}a_{\mathcal{X}} && \text{(B.45)}\\
=& P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1). && \text{(B.46)}
\end{aligned}
$$

The second equality is again due to Assumption 5.2.1.2.

Comparing Equation (B.39) and Equation (B.42) gives

$$
\begin{aligned}
\int \left[P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1) - \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z) \, \mathrm{d}a_{\mathcal{X}} \right] & \\
\times P(z \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \mathrm{d}z = 0, & \quad \text{(B.47)}
\end{aligned}
$$

which implies

$$
P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1) = \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z) \, \mathrm{d}a_{\mathcal{X}}. \tag{B.48}
$$

This step is due to the completeness condition in Assumption 5.2.2.2.

Equation (B.48) leads to identification:

$$
\begin{aligned}
& P(y \mid \mathrm{do}(a_{\mathcal{C}})) && \text{(B.49)}\\
=& P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}) P(z) P(u_{\mathcal{C}}^{\text{sng}}) \, \mathrm{d}z \, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} && \text{(B.50)}\\
=& P(y \mid z, a_{\mathcal{C}}, u_{\mathcal{C}}^{\text{sng}}, s = 1) P(z) P(u_{\mathcal{C}}^{\text{sng}}) \, \mathrm{d}z \, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} && \text{(B.51)}\\
=& \int \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid z) \, \mathrm{d}a_{\mathcal{X}} P(z) P(u_{\mathcal{C}}^{\text{sng}}) \, \mathrm{d}z \, \mathrm{d}u_{\mathcal{C}}^{\text{sng}} && \text{(B.52)}\\
=& \int \int h(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}}) P(u_{\mathcal{C}}^{\text{sng}}) \, \mathrm{d}a_{\mathcal{X}} \, \mathrm{d}u_{\mathcal{C}}^{\text{sng}}. && \text{(B.53)}
\end{aligned}
$$

In particular, the second equality is due to Assumption 5.2.1.2.

□

B.8 Proof of Theorem 5.2.1

We first state the variant of Assumption 5.2.1 and Assumption 5.2.2 required by Theorem 5.2.1. We essentially replace Z with $(U^{\text{mlt}}, W^{\text{mlt}})$ in these assumptions.

Assumption B.8.1: . (Assumption 5.2.1') The causal graph Figure 5.2a satisfies the following conditions:

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

$$S \perp (A, Y) \mid U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}. \tag{B.54}$$

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 B.8.2: . (Assumption 5.2.2') There exists some function f and an nonempty set $\mathcal{N} \subset \{1, \ldots, m\} \backslash \mathcal{C}$ such that

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

$$f(a_{\mathcal{N}}) \perp Y \mid A_{\mathcal{C}}, A_{\mathcal{X}}, U^{\text{mlt}}, W^{\text{mlt}}, U^{\text{sng}}, S = 1, \tag{B.55}$$

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

2. The conditional $P(u^{\text{mlt}}, w^{\text{mlt}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1)$ is complete in $f(a_{\mathcal{N}})$ for almost all $a_{\mathcal{C}}$ and $u_{\mathcal{C}}^{\text{sng}}$, where $U_{\mathcal{C}}^{\text{sng}}$ is the single-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 these assumptions, Theorem 5.2.1 is a direct consequence of Lemma 5.2.1 and Lemma 5.2.2. The reason is that U^{mlt}, W^{mlt}, U^{sng} constitutes an admissible set to identify the intervention distributions $P(y \mid \text{do}(a_{\mathcal{C}}))$.

B.9 Proof of Theorem 5.2.2

We assume Assumption B.8.1 and Assumption B.8.2 as described in Appendix B.8.

Proof. Assumption 5.2.3.2 guarantees the existence of some function $\hat{h}$ such that

$$\hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s = 1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) \hat{P}(a_{\mathcal{X}} \mid \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \mathrm{d}a_{\mathcal{X}} \tag{B.56}$$

under weak regularity conditions. (We discuss the reason in Appendix B.5.)

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

$$P(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) = \int \hat{h}(y, a_{\mathcal{C}}, a_{\mathcal{X}}, u_{\mathcal{C}}^{\text{sng}}) P(a_{\mathcal{X}} \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), u_{\mathcal{C}}^{\text{sng}}, s = 1) \, \mathrm{d}a_{\mathcal{X}}. \tag{B.57}$$

Given this claim (Equation (B.57)), we have

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

which proves the theorem. The first equality is due to Equation (5.17); the second is due to Equation (B.56); the third is due to Assumption 5.2.3 and $U_{\mathcal{C}}^{\text{sng}}$ 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 (B.57)) and Theorem 5.2.1.

We next prove the claim (Equation (B.57)). Start with the right side of the equality.

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

which establishes the claim. The first equality is due to Equation (5.16); the second is due to Equation (B.56); the third equality is due to Assumption 5.2.3.2, which implies

$$
\hat{P}(y \mid a_{\mathcal{C}}, f(a_{\mathcal{N}}), \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s = 1) = \hat{P}(y \mid a_{\mathcal{C}}, \hat{z}, u_{\mathcal{C}}^{\text{sng}}, s = 1). \tag{B.58}
$$

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

B.10 Constructing candidate $f(a_{\mathcal{N}})$'s from the deconfounder outcome model

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

Consider a fitted linear outcome model

$$
Y = \sum_{i=1}^{10} \alpha_{YA_i} A_i + \alpha_{YZ} \hat{Z} + \alpha_{YU'} U^{\text{sng}} + \epsilon_Y. \tag{B.59}
$$

where all the random variables are Gaussian.

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

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

where

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

The reason is that $f(A_9, A_{10}) \perp (\alpha_9 A_9 + \alpha_{10} A_{10})$. Hence $f(a_{\mathcal{N}}) = A_9 + \alpha_{9,10} A_{10}$ satisfies Assumption 5.2.3.2.

C

Additional results for Chapter 6

C.1 Causal identification with a quadratic factor model and a linear outcome model

We establish causal identification when the true causal model is composed of a quadratic factor model and a linear outcome model.

We first write down the causal model:

$$Z = \epsilon_Z, \tag{C.1}$$

$$A_1 = \alpha_{10} + \alpha_{11} Z + \alpha_{12} Z^2 + \epsilon_{A1}, \tag{C.2}$$

$$A_2 = \alpha_{20} + \alpha_{21} Z + \alpha_{22} Z^2 + \epsilon_{A2}, \tag{C.3}$$

$$Y = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \gamma Z + \epsilon_Y, \tag{C.4}$$

where all the errors $\epsilon_Z, \epsilon_{A1}, \epsilon_{A2}, \epsilon_Y$ are independent zero-mean Gaussian with a fixed but unknown variance.

We note that all variables Z, A_1, A_2, Y are scalars in this example; only A_1, A_2, Y are observable; Z is unobserved.

To prove identification, we show that the causal parameters β_1 and β_2 are both functions of the moment generating function of (A_1, A_2, Y).

We first rewrite Y:

$$\begin{aligned} Y =&(\beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20}) + (\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma) \cdot Z + (\beta_1\alpha_{12} + \beta_2\alpha_{22}) \cdot Z^2 + \beta_1\epsilon_{A1} \\ &+ \beta_2\epsilon_{A2} + \epsilon_Y, \\ =&(\beta_1\alpha_{12} + \beta_2\alpha_{22}) \cdot \left(Z + \frac{\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma}{2 \cdot (\beta_1\alpha_{12} + \beta_2\alpha_{22})}\right)^2 + \beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y \\ &+ \left(\beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20} - \left(\frac{\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma}{2 \cdot (\beta_1\alpha_{12} + \beta_2\alpha_{22})}\right)^2\right) \end{aligned}$$

In other words, the observed random variable Y is a sum of a constant, a non-central χ^2 random variable and a zero mean Gaussian random variable $\beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y$.

For notation simplicity, we denote the constants with separate symbols:

$$B_0 \triangleq \beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20}, \tag{C.5}$$

$$B_1 \triangleq \beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma, \tag{C.6}$$

$$B_2 \triangleq \beta_1\alpha_{12} + \beta_2\alpha_{22}. \tag{C.7}$$

Therefore, we have

$$Y = B_0 + B_1 \cdot Z + B_2 \cdot Z^2 + \epsilon_Y, \tag{C.8}$$

where $(\frac{Z}{\sigma_Z} + \frac{B_1}{2B_2\sigma_Z})^2$ is a non-central χ^2 random variable with the non-centrality parameter $\lambda = \left(\frac{B_1}{2B_2\sigma_Z}\right)^2$ and degree of freedom $k = 1$. (σ_Z^2 is the variance of Z.)

We leverage this property to identify the distribution ϵ_Y. Notice the moment generating function of A_1, A_2, Y is

$$M_{A_1,A_2,Y}(t_1,t_2,t_3) \tag{C.9}$$

$$=\mathbb{E}\left[\exp(t_1A_1+t_2A_2+t_3Y)\right] \tag{C.10}$$

$$=\exp(B_0t_3+\alpha_{10}t_1+\alpha_{20}t_2) \tag{C.11}$$

$$\cdot\,\mathbb{E}\left[\exp\left((\alpha_{11}t_1+\alpha_{21}t_2+B_1t_3)\cdot Z+(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\cdot Z^2\right)\right] \tag{C.12}$$

$$\cdot\,\mathbb{E}\left[t_1\epsilon_{A1}+t_2\epsilon_{A2}+t_3(\beta_1\epsilon_{A1}+\beta_2\epsilon_{A2}+\epsilon_Y)\right] \tag{C.13}$$

$$=\exp\left(B_0t_3+\alpha_{10}t_1+\alpha_{20}t_2-(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\cdot\left(\frac{\alpha_{11}t_1+\alpha_{21}t_2+B_1t_3)}{2(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)}\right)^2\right) \tag{C.14}$$

$$\cdot\,\mathbb{E}\left[\exp\left((\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\sigma_Z^2\cdot\left(\frac{\alpha_{11}t_1+\alpha_{21}t_2+B_1t_3}{2(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\sigma_Z}+\frac{Z}{\sigma_Z}\right)^2\right)\right] \tag{C.15}$$

$$\cdot\,\mathbb{E}\left[t_1\epsilon_{A1}+t_2\epsilon_{A2}+t_3(\beta_1\epsilon_{A1}+\beta_2\epsilon_{A2}+\epsilon_Y)\right] \tag{C.16}$$

$$=\exp\left(B_0t_3+\alpha_{10}t_1+\alpha_{20}t_2-(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\cdot\left(\frac{\alpha_{11}t_1+\alpha_{21}t_2+B_1t_3}{2(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)}\right)^2\right) \tag{C.17}$$

$$\cdot\,\frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{1/2}} \tag{C.18}$$

$$\cdot\exp(\frac{1}{2}(t_1+t_3\beta_1)^2\sigma_{A_1}^2)\exp(\frac{1}{2}(t_2+t_3\beta_2)^2\sigma_{A_2}^2)\exp(\frac{1}{2}t_3\sigma_Y^2) \tag{C.19}$$

$$=\exp\left(B_0t_3+\alpha_{10}t_1+\alpha_{20}t_2-(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\cdot\left(\frac{\alpha_{11}t_1+\alpha_{21}t_2+B_1t_3}{2(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)}\right)^2\right) \tag{C.20}$$

$$\cdot\,\frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{1/2}} \tag{C.21}$$

$$\cdot\exp(\frac{1}{2}(t_1\sigma_{A_1}^2+\beta_1^2\sigma_{A_1}^2t_3^2+2\beta_1\sigma_{A_1}^2t_1t_3+t_2\sigma_{A_2}^2+\beta_2^2\sigma_{A_2}^2t_3^2+2\beta_2\sigma_{A_2}^2t_2t_3+\sigma_Y^2t_3), \tag{C.22}$$

where $t=(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\sigma_Z^2$ and $\lambda=\left(\frac{\alpha_{11}t_1+\alpha_{21}t_2+B_1t_3}{2(\alpha_{12}t_1+\alpha_{22}t_2+B_2t_3)\sigma_Z}\right)^2$.

Notice that the ratio of the coefficients in front of t_3^2 and $t_1 t_3$ is β_1. Hence we can identify β_1 from the moment generating function of the unobserved random variables A_1, A_2, Y. The reason is the incongruence between exponential functions, polynomial functions, and square root functions, i.e. exponential functions can not be written as polynomials and others. The other components of the moment generating functions Equations (C.20) and (C.21) do not contain the terms t_3^2 and $t_1 t_3$.

The high-level intuition behind the above calculation is the incongruence between the nonlinear (quadratic) factor model and the linear outcome model. More specifically, the variance due to ϵ_Y in the linear outcome model cannot be attributed wrongfully to the causes and the confounder $\beta_1 A_1 + \beta_2 A_2 + \gamma Z$; the former is Gaussian while the latter is non-Gaussian except when $\alpha_{12} = \alpha_{22} = 0$. (This incongruence does not hold for the linear factor model and the linear outcome model.)

For the same reason, we can identify the other causal parameter β_2.

This result can be extended to other nonlinear factor models and linear outcome models.

C.2 Causal identification with a Bernoulli factor model and a linear outcome model

Consider the model

$$Z_i \sim \text{Bernoulli}(p), \tag{C.23}$$

$$A_i \sim \text{Bernoulli}(a + bZ_i), \tag{C.24}$$

$$Y_i = \beta A_i + \gamma Z_i + \epsilon, \tag{C.25}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2). \tag{C.26}$$

We show that given $P(A_i, Y_i)$, the parameter β is identifiable in this model.

To prove identifiability of β, we will show that β can be written as deterministic function of $P(A_i, Y_i)$. We establish this argument through showing how β can be written as a deterministic function of the moment generating functions of $P(A_i, Y_i)$.

We first try to understand how β relates to the moment generating functions of $P(A_i, Y_i)$.

First consider Y_i. Notice that the conditional distribution $Y_i \mid Z_i$ is Bernoulli-Gaussian composition

$$P(Y_i \mid Z_i = 0) = \mathcal{N}(\beta \cdot \text{Benroulli}(a), \sigma^2), \tag{C.27}$$

$$P(Y_i \mid Z_i = 1) = \mathcal{N}(\beta \cdot \text{Bernoulli}(a+b) + \gamma, \sigma^2). \tag{C.28}$$

Hence, the distribution of Y_i is a mixture of the two,

$$P(Y_i) = p \cdot \mathcal{N}(\beta \cdot \text{Bernoulli}(a+b) + \gamma, \sigma^2) + (1-p)\mathcal{N}(\beta \cdot \text{Benroulli}(a), \sigma^2) \tag{C.29}$$

It implies that the moment generating function of Y_i must be

$$\begin{aligned}
\mathbb{E}\left[e^{tY_i}\right] =& p \cdot e^{\frac{1}{2}t^2\sigma^2}\mathbb{E}\left[(e^{\beta\cdot\text{Bernoulli}(a+b)+\gamma)\cdot t}\right] + (1-p)\cdot e^{\frac{1}{2}t^2\sigma^2}\mathbb{E}\left[e^{\beta\cdot\text{Benroulli}(a)\cdot t}\right] \\
=& p \cdot e^{\frac{1}{2}t^2\sigma^2}((1-a-b)e^{\gamma t} + (a+b)e^{(\beta+\gamma)t}) + (1-p)\cdot e^{\frac{1}{2}t^2\sigma^2}((1-a) + ae^{\beta t}) \\
=& p\cdot(1-a-b)e^{\gamma t + \frac{1}{2}t^2\sigma^2} + p\cdot(a+b)e^{(\beta+\gamma)t+\frac{1}{2}t^2\sigma^2} \\
&+ (1-p)ae^{\beta t+\frac{1}{2}t^2\sigma^2} + (1-p)\cdot(1-a)e^{\frac{1}{2}t^2\sigma^2}
\end{aligned}$$

Assume the coefficients in the moment generating functions are nonzero. That is, $p \neq 0, a+b \neq 1$, $a+b \neq 0$, $p \neq 1$, and $a \neq 1$.

This calculation implies that we can recover the parameter σ^2 from $P(Y_i)$. In more detail, suppose the moment generating function of the $P(Y_i)$ is

$$\mathbb{E}\left[e^{tY_i}\right] = \tilde{c}_1 e^{\frac{1}{2}t^2c_0} + \tilde{c}_2 e^{\frac{1}{2}t^2c_0+c_2t} + \tilde{c}_3 e^{\frac{1}{2}t^2c_0+c_3t} + \tilde{c}_4 e^{\frac{1}{2}t^2c_0+c_4t}, \tag{C.30}$$

where the constants $\tilde{c}_1, \tilde{c}_2, \tilde{c}_3, \tilde{c}_4, c_0, c_2, c_3, c_5$ are deterministic function of $P(Y_i)$, which itself is a deterministic function of $P(Y_i, A_i)$.

Suppose $\beta\cdot\gamma \neq 0$. Then we can write the parameter σ^2 as a deterministic functions of $P(Y_i, A_i)$:

$$\sigma^2 = c_0. \tag{C.31}$$

Next, to find the other parameters β, γ, $\beta+\gamma$, we need to match c_2, c_3, c_5 with the values β, γ, $\beta+\gamma$. We find which two of $\{c_2, c_3, c_4\}$ add up to the third. For example, if (without loss of generality) $c_2 + c_3 = c_4$, then we know

$$\beta + \gamma = c_4 \tag{C.32}$$

$$p(a+b) = \tilde{c}_4. \tag{C.33}$$

The reason is that only $(\beta) + (\gamma) = (\beta + \gamma)$ will satisfy the constraint of two of $\{\beta, \gamma, \beta + \gamma\}$ adding up being equal to the third. If $(\beta) + (\beta + \gamma) = (\gamma)$ or $(\gamma) + (\beta + \gamma) = (\beta)$, then $\gamma = 0$ or $\beta = 0$, which contradicts with the assumption that $\beta \cdot \gamma \neq 0$.

Next, consider A_i. The conditional distribution of $A_i \mid Z_i$ is

$$P(A_i \mid Z_i = 0) = \text{Benroulli}(a), \tag{C.34}$$

$$P(A_i \mid Z_i = 1) = \text{Bernoulli}(a+b) + \gamma. \tag{C.35}$$

Hence the distribution of A_i is

$$P(A_i) = p\text{Bernoulli}(a+b) + (1-p)\text{Benroulli}(a). \tag{C.36}$$

Again compute the moment generating function of A_i:

$$\mathbb{E}\left[e^{tA_i}\right] = p(1-a-b+(a+b)e^t) + (1-p)(1-a+ae^t) \tag{C.37}$$

$$= [p(1-a-b) + (1-p)(1-a)] + [p(a+b) + (1-p)a]e^t. \tag{C.38}$$

If the moment generating function of the $P(A_i)$ is

$$\mathbb{E}\left[e^{tA_i}\right] = \tilde{d}_1 + \tilde{d}_2 e^{d_2 t}, \tag{C.39}$$

where $\tilde{d}_1, \tilde{d}_2, d_2$ are deterministic functions of the $P(A_i)$, which itself is a deterministic function of $P(A_i, Y_i)$. Then, we can have

$$p(a+b) + (1-p)a = \tilde{d}_2. \tag{C.40}$$

Equations (C.33) and (C.40) imply that

$$(1-p)a = \tilde{d}_2 - \tilde{c}_4. \tag{C.41}$$

Return to Equation (C.30), we notice that the coefficient in front of $e^{\beta t + \frac{1}{2}t^2\sigma^2}$ is $(1-p)a$. Therefore, we match $(1-p)a = \tilde{d}_2 - \tilde{c}_4$ with the coefficients in front of the exponential terms $\tilde{c}_2 e^{\frac{1}{2}t^2 c_0 + c_2 t}$ and $\tilde{c}_3 e^{\frac{1}{2}t^2 c_0 + c_3 t}$; the one that matches is the term $(1-p)a e^{\beta t + \frac{1}{2}t^2\sigma^2}$.

$$\beta = \begin{cases} c_2 & \text{if } \tilde{c}_2 = \tilde{d}_2 - \tilde{c}_4 \\ c_3 & \text{if } \tilde{c}_3 = \tilde{d}_2 - \tilde{c}_4. \end{cases} \tag{C.42}$$

We have written β as a deterministic function of the observed data distribution $P(A_i, Y_i)$. Therefore, it is an identifiable parameter as long as $\beta \cdot \gamma \neq 0$.

Finally, consider the edge cases $\beta \cdot \gamma = 0$. First, when $\mathbb{E}\left[e^{tY_i}\right] = \tilde{c}_1 e^{\frac{1}{2}t^2 c_0}$, we conclude $\beta = \gamma = 0$; β is identifiable. Next, suppose only one of γ and β is zero. Then we have

$$\mathbb{E}\left[e^{tY_i}\right] = \tilde{c}_1 e^{\frac{1}{2}t^2 c_0} + \tilde{c}_2 e^{\frac{1}{2}t^2 c_0 + c_2 t} \tag{C.43}$$

We conclude $\gamma = 0$ and $\beta = c_2$ if $\tilde{c}_2 = \tilde{d}_2 = p(a+b) + (1-p)a$ and $\beta = 0$ if otherwise. Again, β is identifiable when only only one of γ and β is zero.

We thus have proved that β is identifiable given $P(Y_i, A_i)$, i.e. β can be written as a deterministic function of $P(Y_i, A_i)$ (if $p \neq 0, a+b \neq 1, a+b \neq 0, p \neq 1$, and $a \neq 1$).

C.3 Causal identification with a linear factor model and a linear outcome model

We revisit the non-identification of the deconfounder in linear Gaussian factor models (D'Amour, 2018, 2019a,b). Without pinpointed substitute confounder Z_i, the joint causal effect of all m causes might not be identifiable. We also show how this non-identification can be resolved by considering multiple independent outcomes, hence "the blessings of multiple outcomes."

Importantly, this discussion of non-identification does not contradict the identification result in Theorem 4.2.1; it does not make the same assumptions.

We start with restating the linear Gaussian non-identification example from D'Amour (2018).

Denote the confounder as Z, the causes A, and the (observed) outcome Y:

$$Z_{n\times k} = \epsilon_Z, \tag{C.44}$$

$$A_{n\times m} = Z_{n\times k}\alpha_{k\times m} + \epsilon_A, \tag{C.45}$$

$$Y_{n\times 1} = A_{n\times m}\beta_{m\times 1} + Z_{n\times k}\gamma_{k\times 1} + \epsilon_Y, \tag{C.46}$$

where all random variables are zero mean Gaussian. In particular, the errors $\epsilon_A, \epsilon_Y, \epsilon_Z$ have diagonal covariance matrices. The above equations describe the true data generating process.

We observe the causes and the outcome; the confounder is unobserved. The goal is to estimate the coefficient β; it describes the causal effect of A on Y.

We observe the following covariance matrices:

$$\Sigma_{AA} = \alpha^\top \Sigma_{ZZ}\alpha + \sigma_A^2 \cdot I, \tag{C.47}$$

$$\Sigma_{AY} = \Sigma_{AA}\beta + \alpha^\top \Sigma_{ZZ}\gamma, \tag{C.48}$$

$$\Sigma_{YY} = \beta^\top \Sigma_{AA}\beta + \beta^\top\alpha^\top \Sigma_{ZZ}\gamma + \gamma^\top \Sigma_{ZZ}\gamma + \sigma_Y^2 \cdot I. \tag{C.49}$$

We can obtain two equally good solutions of $(\alpha, \beta, \gamma, Z, \sigma_A, \sigma_Y, \sigma_Z)$ by fixing γ and rescaling the latent variable Z (D'Amour, 2018).

The crux of this non-identifiability issue lies in Equation (C.48), where the $p \times 1$ matrix Σ_{AY} is mapped to a rank p matrix Σ_{AA} and a rank k matrix Σ_{ZZ}. This mapping is non-identifiable.

We can resolve this non-identifiability issue by assuming σ_Y is known. That is, we assume the measurement error in the outcome is $\mathcal{N}(0, \sigma_Y^2)$. This requires external information about the measurement error of the outcome. Such information could be obtained from running independent experiments on the outcome. We emphasize that this is an assumption on the data generating process.

We now prove causal identification assuming under this additional assumption on measurement error. We will write the causal parameter β in terms of observed quantities Σ_{AA}, Σ_{AY}, Σ_{YY}.

We first write the reconstructed causes $\hat{\boldsymbol{a}}(z)$ as $W = Z\alpha$. Then we re-write the model Equation (C.46) as

$$Z_{n\times k} = \epsilon_Z, \tag{C.50}$$

$$W_{n\times m} = Z_{n\times k}\alpha_{k\times m} \tag{C.51}$$

$$A_{n\times m} = W_{n\times m} + \epsilon_A, \tag{C.52}$$

$$Y_{n\times 1} = A_{n\times m}\beta_{m\times 1} + W_{n\times m}\delta_{m\times 1} + \epsilon_Y, \tag{C.53}$$

where $\alpha\delta = \gamma$.

It leads to the following covariances:

$$\Sigma_{AA} = \Sigma_{WW} + \sigma_A^2 \cdot I \tag{C.54}$$

$$\Sigma_{WA} = \Sigma_{WW}, \tag{C.55}$$

$$\Sigma_{WY} = \Sigma_{WA}\beta + \Sigma_{WW}\delta, \tag{C.56}$$

$$\Sigma_{AY} = \Sigma_{AA}\beta + \Sigma_{AW}\delta. \tag{C.57}$$

$$\Sigma_{YY} = \beta^\top \Sigma_{AY} + \delta^\top \Sigma_{WW}\delta + \sigma_Y^2 \cdot I. \tag{C.58}$$

In particular, we observe Σ_{AY}, Σ_{AA}, Σ_{YY}. First of all, we can write Σ_{WW} and σ_A^2 in terms of Σ_{AA}:

$$\sigma_A^2 = \Sigma_{AA,11} - \Sigma_{AA,21} \cdot \frac{\Sigma_{AA,13}}{\Sigma_{AA,23}}, \tag{C.59}$$

$$\Sigma_{WW} = \Sigma_{AA} - \sigma_A^2 \cdot I. \tag{C.60}$$

Second, we re-arrange Equation (C.57):

$$\delta = \Sigma_{AW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) \tag{C.61}$$

Finally, we substitute Equation (C.61) into Equation (C.58)

$$\begin{aligned}\Sigma_{YY} =& \beta^\top \Sigma_{AY} + (\Sigma_{AW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta))^\top \Sigma_{WW} \Sigma_{AW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) + \sigma_Y^2 \cdot I && \text{(C.62)}\\ =& \beta^\top \Sigma_{AY} + (\Sigma_{AY} - \Sigma_{AA}\beta)^\top \Sigma_{WW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) + \sigma_Y^2 \cdot I && \text{(C.63)}\\ =& \beta^\top \Sigma_{AY} + (\Sigma_{AY} - \Sigma_{AA}\beta)^\top (\Sigma_{AA} - \Sigma_{AA,11} - \Sigma_{AA,21} \cdot \frac{\Sigma_{AA,13}}{\Sigma_{AA,23}} \cdot I)^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta)\\ & + \sigma_Y^2 \cdot I. && \text{(C.64)}\end{aligned}$$

When σ_Y^2 is assumed known, then we can solve for the causal parameter β.

This calculation shows that we can identify the causal effect if the variance of the measurement error σ_Y^2 is known. Concretely, using outcome models like linear regression with a known measurement error or logistic regression will not result in the non-identification pathology. This aligns with our practices in the empirical studies in Chapter 6.

In the same vein of this calculation, the causal effect will be identifiable if we observe two or more independent outcomes for each unit. We can use the off-diagonal terms of the covariance matrix Σ_{YY} to estimate β with a similar equation to Equation (C.64).

More generally, if (1) the distribution of measurement error is known, that is

$$P(Y \mid A, \hat{A}) = p_\epsilon(Y \mid f(A, \hat{A}))$$

with p_ϵ known and (2) f can be identifiable from $P(f(A, \hat{A}))$ and $P(A, \hat{A})$, then the mean potential outcome function can be identifiable.

C.4 Detailed results of the GWAS Study

In this section, we present tables of results from the GWAS study in Section 6.2.

Tables C.1 to C.5 contain the result under the high SNR setting.

Tables C.6 to C.10 contain the result under the low SNR setting.

	Pred. check	Real-valued outcome RMSE$\times 10^{-2}$	Binary outcome RMSE$\times 10^{-2}$
No control	—	49.66	39.39
Control for confounders*	—	40.27	31.09
(G)LMM	—	46.22	37.81
PPCA	0.13	46.05	36.01
PF	0.15	44.58	36.30
LFA	0.14	43.02	36.65
GMM	0.01	47.33	40.24
DEF	0.18	**41.05**	**33.88**

Table C.1: GWAS high-snr simulation I: Balding-Nichols Model. ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

	Pred. check	Real-valued outcome RMSE$\times 10^{-2}$	Binary outcome RMSE$\times 10^{-2}$
No control	—	68.78	38.16
Control for confounders*	—	60.29	32.76
(G)LMM	—	65.25	35.41
PPCA	0.15	65.98	36.11
PF	0.17	64.25	34.79
LFA	0.17	64.00	37.08
GMM	0.02	67.23	35.40
DEF	0.20	**63.73**	**33.71**

Table C.2: GWAS high-snr simulation II: 1000 Genomes Project (TGP). ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

C.5 Detailed results of the Movie Study

In this section, we present tables of results from the movies study in Section 6.3.

C.6 Details of Section 6.2

We follow Song et al. (2015) in simulating the allele frequencies. We present the full details here.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	77.35	45.93
Control for confounders*	—	67.53	39.43
(G)LMM	—	74.38	42.79
PPCA	0.14	74.45	43.27
PF	0.14	71.40	42.75
LFA	0.13	72.11	42.34
GMM	0.03	76.27	46.88
DEF	0.16	**69.86**	**41.61**

Table C.3: GWAS high-snr simulation III: Human Genome Diversity Project (HGDP). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

We simulate the $n \times m$ matrix of genotypes A from $A_{ij} \sim \text{Binomial}(2, F_{ij})$, where F is the $n \times m$ matrix of allele frequencies. Let $F = \Gamma S$, where Γ is $n \times d$ and S is $d \times m$ with $d \leq m$. The $d \times m$ matrix S encodes the genetic population structure. The $n \times d$ matrix Γ maps how the structure affects the allele frequencies of each SNP. Table C.14 details how we generate Γ and S for each simulation setup.

With low SNR, we simulate datasets of $n = 5000$ individuals and $m = 100,000$ SNPs where the SNPs $\sum_j \beta_j a_{ij}$ contributes 10% of the variance, the per-group intercept λ_{c_i} contributes 20%, and the error ε_i contributes 70%. With high SNR, we simulate datasets of $n = 100,000$ individuals and $m = 5,000$ SNPs where the SNPs signal contributes 40%, the per-group intercept contributes 40% and the error contributes 20%.

For each simulation scenario, we generate 100 independent studies. We then simulate a trait; we consider two types: one continuous and one binary. For each trait, three components contributing to the trait: causal signals $\sum_{j=1}^{m} \beta_j a_{ij}$, confounder λ_i, and random effects ϵ_i.

First, without loss of generality, we set the first 1% of the m SNPs to be the true causal SNPs ($\beta_j \neq 0, \beta_j \overset{iid}{\sim} \mathcal{N}(0, 0.5^2)$. We set $\beta_j = 0$ for the rest of the SNPs.

Notice that the SNPs are affected by some latent population structure. We simulate the confounder λ_i and the random effects ϵ_i so that they depend on the latent population structure as well.

For the confounder λ_i, we first perform K-means clustering on the columns of S with $K = 3$ using Euclidean distance. This assigns each individual i to one of three mutually exclusive cluster sets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$, where $\mathcal{S}_k \subset \{1, 2, \ldots, n\}$. Set $\lambda_j = k$ if $j \in \mathcal{S}_k, k = 1, 2, 3$.

We then simulate the random effects ϵ_i. Let $\tau_1^2, \tau_2^2, \tau_3^2 \overset{iid}{\sim} \text{InvGamma}(3,1)$, and set $\sigma_i^2 = \tau_k^2$ for all $j \in \mathcal{S}_i, k = 1, 2, 3$. Draw $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

We control the SNR to mimic the highly noisy nature of GWAS data sets: in the low SNR setting, we let the causal signals $\sum_{j=1}^m \beta_j a_{ij}$ contribute to $\nu_{\text{gene}} = 0.1$ of the variance, the confounder λ_i contribute $\nu_{\text{conf}} = 0.2$, and the random effects ϵ_i contribute $\nu_{\text{noise}} = 0.7$. In the high SNR setting, we have $\nu_{\text{gene}} = 0.4$, $\nu_{\text{conf}} = 0.4$, and $\nu_{\text{noise}} = 0.2$.

We set

$$\lambda_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^m \beta_j a_{ij}\}_{i=1}^n}{\sqrt{\nu_{\text{gene}}}}\right]\left[\frac{\sqrt{\nu_{\text{conf}}}}{s.d.\{\lambda_i\}_{i=1}^n}\right]\lambda_i, \tag{C.65}$$

$$\epsilon_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^m \beta_j a_{ij}\}_{i=1}^n}{\sqrt{\nu_{\text{gene}}}}\right]\left[\frac{\sqrt{\nu_{\text{noise}}}}{s.d.\{\epsilon_i\}_{i=1}^n}\right]\epsilon_i. \tag{C.66}$$

We finally generate a real-valued outcome from a linear model and a binary outcome from a logistic model:

$$y_{i,\text{real}} = \sum_{j=1}^m \beta_j a_{ij} + \lambda_i + \epsilon_i, \tag{C.67}$$

$$y_{i,\text{binary}} \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(\sum_{j=1}^m \beta_j a_{ij} + \lambda_i + \epsilon_i)}\right). \tag{C.68}$$

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\alpha = 0.01$	No control	—	40.68	30.37
	Control for confounders*	—	34.35	28.21
	(G)LMM	—	39.09	28.36
	PPCA	0.15	38.14	28.97
	PF	0.16	**34.77**	28.67
	LFA	0.16	35.87	28.33
	GMM	0.02	38.15	29.69
	DEF	0.18	34.84	**28.04**
$\alpha = 0.1$	No control	—	43.87	36.77
	Control for confounders*	—	37.62	33.89
	(G)LMM	—	39.97	35.76
	PPCA	0.21	39.60	35.61
	PF	0.19	38.95	**34.28**
	LFA	0.18	39.28	34.73
	GMM	0.00	44.38	36.44
	DEF	0.20	**38.75**	34.85
$\alpha = 0.5$	No control	—	47.38	41.84
	Control for confounders*	—	43.63	39.86
	(G)LMM	—	47.28	42.91
	PPCA	0.14	46.90	41.41
	PF	0.16	43.29	40.69
	LFA	0.17	43.60	40.77
	GMM	0.02	46.95	42.47
	DEF	0.18	**43.09**	**40.03**
$\alpha = 1.0$	No control	—	53.94	49.32
	Control for confounders*	—	47.12	45.96
	(G)LMM	—	49.21	48.96
	PPCA	0.21	50.57	47.58
	PF	0.19	48.07	46.16
	LFA	0.17	49.27	46.16
	GMM	0.02	52.28	50.31
	DEF	0.23	**47.82**	**45.62**

Table C.4: GWAS high-snr simulation IV: Pritchard-Stephens-Donnelly (PSD). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms (G)LMM; DEF often performs the best among the five factor models. Predictive checking offers a good indication of when the deconfounder fails.

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\tau = 0.1$	No control	—	47.47	45.16
	Control for confounders*	—	44.22	43.85
	(G)LMM	—	47.35	44.15
	PPCA	0.08	47.61	44.36
	PF	0.09	47.13	43.69
	LFA	0.09	47.16	43.87
	GMM	0.01	47.55	45.95
	DEF	0.10	**46.95**	**43.62**
$\tau = 0.25$	No control	—	44.68	41.10
	Control for confounders*	—	41.23	39.65
	(G)LMM	—	43.42	**40.67**
	PPCA	0.11	**43.26**	41.28
	PF	0.12	43.30	41.10
	LFA	0.13	43.62	41.65
	GMM	0.01	44.81	41.02
	DEF	0.13	43.35	40.97
$\tau = 0.5$	No control	—	45.18	40.92
	Control for confounders*	—	41.33	37.35
	(G)LMM	—	44.83	40.59
	PPCA	0.10	43.78	**39.99**
	PF	0.09	43.65	40.23
	LFA	0.10	43.88	40.04
	GMM	0.01	46.08	40.76
	DEF	0.12	**43.57**	40.02
$\tau = 1.0$	No control	—	56.57	57.70
	Control for confounders*	—	52.98	55.46
	(G)LMM	—	56.44	56.33
	PPCA	0.14	55.18	57.36
	PF	0.12	55.29	56.31
	LFA	0.13	**54.75**	56.66
	GMM	0.01	57.15	57.55
	DEF	0.12	55.07	**56.22**

Table C.5: GWAS high-snr simulation V: Spatial model. ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder often outperforms (G)LMM. Predictive checking offers a good indication of when the deconfounder fails: GMM poorly captures the SNPs; it can amplify the error in causal estimates.

	Pred. check	Real-valued outcome RMSE$\times 10^{-2}$	Binary outcome RMSE$\times 10^{-2}$
No control	—	6.55	5.75
Control for confounders*	—	6.54	5.75
(G)LMM	—	6.54	**5.74**
PPCA	0.14	6.52	**5.74**
PF	0.16	6.53	**5.74**
LFA	0.14	6.54	**5.74**
GMM	0.01	6.54	**5.74**
DEF	0.19	**6.47**	**5.74**

Table C.6: GWAS low-snr simulation I: Balding-Nichols Model. ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

	Pred. check	Real-valued outcome RMSE$\times 10^{-2}$	Binary outcome RMSE$\times 10^{-2}$
No control	—	8.31	4.85
Control for confounders*	—	8.28	4.85
(G)LMM	—	8.29	4.85
PPCA	0.14	8.29	4.85
PF	0.15	8.29	4.85
LFA	0.17	8.26	4.85
GMM	0.02	8.30	4.85
DEF	0.20	**8.11**	**4.84**

Table C.7: GWAS low-snr simulation II: 1000 Genomes Project (TGP). ("Control for all confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

	Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
No control	—	9.59	5.84
Control for confounders*	—	9.52	5.84
(G)LMM	—	9.57	5.84
PPCA	0.14	9.55	5.84
PF	0.13	9.56	5.84
LFA	0.14	9.54	5.84
GMM	0.03	9.59	5.84
DEF	0.16	**9.47**	**5.83**

Table C.8: GWAS low-snr simulation III: Human Genome Diversity Project (HGDP). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\alpha = 0.01$	No control	—	3.73	3.23
	Control for confounders*	—	3.71	3.23
	(G)LMM	—	3.71	3.23
	PPCA	0.13	3.64	3.23
	PF	0.16	3.67	3.23
	LFA	0.16	3.66	3.23
	GMM	0.02	3.72	3.23
	DEF	0.18	**3.59**	**3.22**
$\alpha = 0.1$	No control	—	4.09	3.84
	Control for confounders*	—	4.09	3.84
	(G)LMM	—	4.09	3.84
	PPCA	0.20	4.08	3.84
	PF	0.18	4.08	3.84
	LFA	0.18	4.07	3.84
	GMM	0.00	4.09	3.84
	DEF	0.20	**4.05**	**3.83**
$\alpha = 0.5$	No control	—	4.82	4.14
	Control for confounders*	—	4.81	4.14
	(G)LMM	—	4.82	4.14
	PPCA	0.14	4.81	**4.13**
	PF	0.17	**4.80**	**4.13**
	LFA	0.16	4.81	4.14
	GMM	0.03	4.82	4.14
	DEF	0.19	**4.80**	**4.13**
$\alpha = 1.0$	No control	—	5.43	4.58
	Control for confounders*	—	5.38	4.57
	(G)LMM	—	5.40	4.58
	PPCA	0.21	5.38	**4.57**
	PF	0.16	5.41	**4.57**
	LFA	0.19	5.40	**4.57**
	GMM	0.02	5.43	4.58
	DEF	0.24	**5.37**	**4.57**

Table C.9: GWAS low-snr simulation IV: Pritchard-Stephens-Donnelly (PSD). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms LMM; DEF performs the best among the five factor models; it also outperforms using the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

		Pred. check	**Real-valued outcome** RMSE$\times 10^{-2}$	**Binary outcome** RMSE$\times 10^{-2}$
$\tau = 0.1$	No control	—	4.66	4.74
	Control for confounders*	—	4.63	4.73
	(G)LMM	—	4.57	**4.73**
	PPCA	0.09	4.62	4.74
	PF	0.08	4.58	4.74
	LFA	0.09	4.54	**4.73**
	GMM	0.02	4.70	4.74
	DEF	0.10	**4.53**	**4.73**
$\tau = 0.25$	No control	—	4.30	3.81
	Control for confounders*	—	3.81	3.79
	(G)LMM	—	4.28	**3.80**
	PPCA	0.10	4.26	**3.80**
	PF	0.12	4.26	**3.80**
	LFA	0.12	4.27	**3.80**
	GMM	0.01	4.30	3.81
	DEF	0.13	**4.25**	**3.80**
$\tau = 0.5$	No control	—	4.30	3.85
	Control for confounders*	—	3.82	3.83
	(G)LMM	—	4.28	**3.83**
	PPCA	0.11	4.27	**3.83**
	PF	0.09	4.28	3.84
	LFA	0.11	4.27	3.84
	GMM	0.01	4.29	3.84
	DEF	0.13	**4.25**	3.84
$\tau = 1.0$	No control	—	6.71	5.52
	Control for confounders*	—	5.43	5.51
	(G)LMM	—	6.70	5.52
	PPCA	0.14	6.70	5.52
	PF	0.12	6.70	5.52
	LFA	0.12	6.69	5.52
	GMM	0.01	6.72	5.53
	DEF	0.13	**6.62**	**5.51**

Table C.10: GWAS low-snr simulation V: Spatial model. ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder often outperforms LMM; DEF often performs the best among the five factor models. Yet, the deconfounder does not outperform using the (unobserved) confounder information. Spatially-induced SNPs challenge many latent variable models to capture its patterns and fully deconfound causal inference. Predictive checking offers a good indication of when the deconfounder fails: GMM poorly captures the SNPs; it can amplify the error in causal estimates.

Control	Average predictive log-likelihood
No Control	**-1.1**
Control for X	**-1.1**
Control for $\hat{a}_{\text{PPCA}}$	-1.2
Control for $\hat{a}_{\text{PF}}$	-1.2
Control for $\hat{a}_{\text{DEF}}$	-1.2
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.3
Control for $(\hat{a}_{\text{PF}}, X)$	-1.2
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.2

Table C.11: Average predictive log-likelihood on a holdout set of all movies. (X represents the observed covariates.) Causal models (the deconfounder) predicts slightly worse than prediction models.

Control	Average predictive log-likelihood
No Control	-2.5
Control for X	-2.1
Control for $\hat{a}_{\text{PPCA}}$	-1.6
Control for $\hat{a}_{\text{PF}}$	**-1.5**
Control for $\hat{a}_{\text{DEF}}$	**-1.5**
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.7
Control for $(\hat{a}_{\text{PF}}, X)$	**-1.5**
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.6

Table C.12: Average predictive log-likelihood on the holdout set of non-English movies. (X represents the observed covariates.) On a test set of uncommon movies, causal models with the deconfounder predict better than prediction models.

Control	Average predictive log-likelihood
No Control	-2.1
Control for X	-1.9
Control for $\hat{a}_{\text{PPCA}}$	-1.4
Control for $\hat{a}_{\text{PF}}$	**-1.2**
Control for $\hat{a}_{\text{DEF}}$	-1.3
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.4
Control for $(\hat{a}_{\text{PF}}, X)$	-1.3
Control for $(\hat{a}_{\text{DEF}}, X)$	**-1.2**

Table C.13: Average predictive log-likelihood on the holdout set of non-drama/comedy/action movies. (X represents the observed covariates.) On a test set of uncommon movies, causal models with the deconfounder predict better than prediction models.

Model	Simulation details
Balding-Nichols Model (Balding-Nichols)	Each row i of Γ has i.i.d. three independent and identically distributed draws from the Balding- Nichols model: $\gamma_{ik} \overset{iid}{\sim} \text{BN}(p_i, F_i)$, where $k \in \{1, 2, 3\}$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HapMap data set, calculating its observed allele frequency and estimating its F_{ST} value using the Weir & Cockerham estimator (Weir and Cockerham, 1984). The columns of S were Multinomial$(60/210, 60/210, 90/210)$, which reflect the subpopulation proportions in the HapMap data set.
1000 Genomes Project (TGP)	The matrix Γ was generated by sampling $\gamma_{ik} \overset{iid}{\sim} 0.9\times\text{Uniform}(0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. To generate S, we compute the first two principal components of the TGP genotype matrix after mean centering each SNP. We then transformed each principal component to be between $(0, 1)$ and set the first two rows of S to be the transformed principal components. The third row of S was set to 1, i.e. an intercept.
Human Genome Diversity Project (HGDP)	Same as TGP but generating S with the HGDP genotype matrix.
Pritchard-Stephens-Donnelly (PSD)	Each row i of Γ has i.i.d. three independent and identically distributed draws from the Balding- Nichols model: $\gamma_{ik} \overset{iid}{\sim} \text{BN}(p_i, F_i)$, where $k \in \{1, 2, 3\}$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HGPD data set, calculating its observed allele frequency and estimating its F_{ST} value using the Weir & Cockerham estimator (Weir and Cockerham, 1984). The estimator requires each individual to be assigned to a subpopulation, which were made according to the $K = 5$ subpopulations from the analysis in Rosenberg et al. (2002). The columns of S were sampled $(s_{1j}, s_{2j}, s_{3j} \overset{iid}{\sim} \text{Dirichlet}(\alpha, \alpha, \alpha)$ for $j = 1, \ldots, m, \alpha = 0.01, 0.1, 0.5, 1$.
Spatial	The matrix Γ was generated by sampling $\gamma_{ik} \overset{iid}{\sim} 0.9\times\text{Uniform}(0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. The first two rows of S correspond to coordinates for each individual on the unit square and were set to be independent and identically distributed samples from $\text{Beta}(\tau, \tau)$, $\tau = 0.1, 0.25, 0.5, 1$, while the third row of S was set to be 1, i.e. an intercept. As $\tau \to 0$, the individuals are placed closer to the corners of the unit square, while when $\tau = 1$, the individuals are distributed uniformly.

Table C.14: Simulating allele frequencies.

D

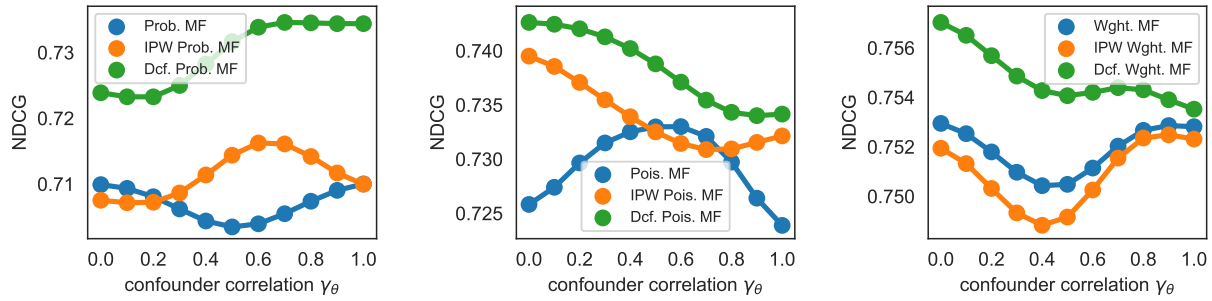
Additional results for Chapter 7

D.1 Detailed simulation results

Figures 7.1 and D.1 explore how an unobserved confounder high correlated with exposures can affect the rating predictions and ranking quality of recommendation algorithms.

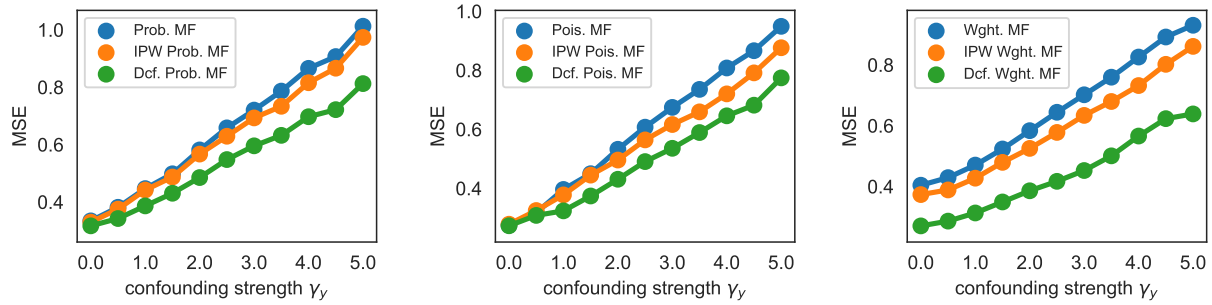
Figures D.2 and D.3 explore how an unobserved confounder high correlated with the rating outcome can affect the rating predictions and ranking quality of recommendation algorithms.

In both simulations, we find that the deconfounded recommender is more robust to unobserved confounding; it leads to smaller MSEs in rating prediction and higher NDCGs in recommendation than its classical counterparts (Probabilistic/Poisson/Weighted MF) and the existing causal approach (IPW MF (Schnabel et al., 2016)).



(a) Probabilistic MF **(b)** Poisson MF **(c)** Weighted MF

Figure D.1: Varying confounding correlation γ_θ from 0.0 to 1.0 ($\gamma_y = 3.0$). The ranking quality of the deconfounded recommender (green) is more robust to unobserved confounding than its classical counterpart (blue) and the existing causal approach (orange) (Schnabel et al., 2016). Higher is better. (Results averaged over 10 runs.)



(a) Probabilistic MF **(b)** Poisson MF **(c)** Weighted MF

Figure D.2: Varying confounding strength γ_y from 0.0 to 5.0 ($\gamma_\theta = 0.6$). The rating predictions of the deconfounded recommender (green) is more robust to unobserved confounding than its classical counterpart (blue) and the existing causal approach (orange) (Schnabel et al., 2016). Lower is better. (Results averaged over 10 runs.)

D.2 Performance measures

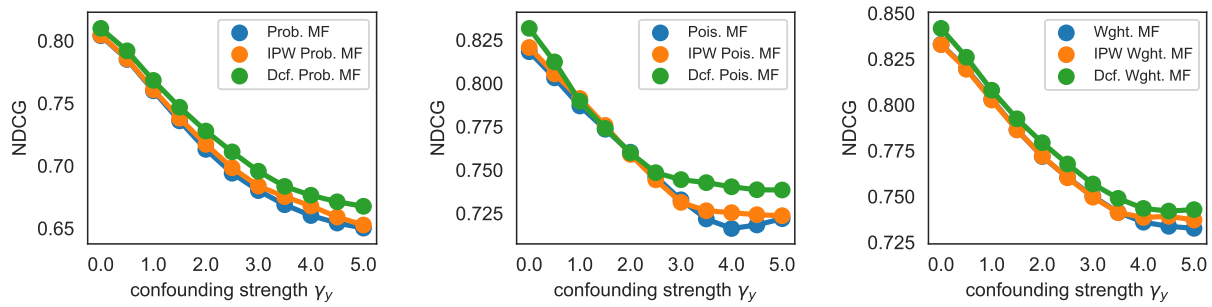
Denote $rank(u, i)$ as the rank of item i in user u's predicted list; let y_u^{test} as the set of relevant items in the test set of user u.*

- NDCG measures ranking quality:

$$DCG = \frac{1}{U}\sum_{u=1}^{U}\sum_{i=1}^{L}\frac{2^{\text{rel}_{ui}}-1}{\log_2(i+1)}, \quad NDCG = \frac{DCG}{IDCG},$$

  where L is the number of items in the test set of user u, the relevance rel_{ui} is set as the rating of user u on item i, and $IDCG$ is a normalizer that ensures $NDCG$ sits between 0 and 1.

*We consider items with a rating greater than or equal to three as relevant.



(a) Probabilistic MF **(b)** Poisson MF **(c)** Weighted MF

Figure D.3: Varying confounding strength γ_y from 0.0 to 5.0 ($\gamma_\theta = 0.6$). The ranking quality of the deconfounded recommender (green) is more robust to unobserved confounding than its classical counterpart (blue) and the existing causal approach (orange) (Schnabel et al., 2016). Higher is better. (Results averaged over 10 runs.)

- Recall@k. Recall@k evaluates how many relevant items are selected:

$$\text{Recall@k} = \frac{1}{U}\sum_{u=1}^{U}\sum_{i\in y_u^{\text{test}}}\frac{\mathbb{1}\{rank(u,i)\leq k\}}{\min(k,|y_u^{\text{test}}|)}.$$

- MSE. MSE evaluates how far the predicted ratings are away from the true rating:

$$\text{MSE} = \frac{1}{|(u,i)\in y^{\text{test}}|}\sum_{(u,i)\in y^{\text{test}}}(\hat{y}_{ui}-y_{ui})^2.$$

D.3 Experimental Details

For weighted matrix factorization, we set weights of the observation by $c_{ui} = 1 + \alpha y_{ui}$ where $\alpha = 40$ following (Hu et al., 2008).

For Gaussian latent variables in probabilistic matrix factorization, we use priors $\mathcal{N}(0,1)$ or $\mathcal{N}(0,0.1^2)$.

For Gamma latent variables in PF, we use prior Gamma(0.3, 0.3).

For inverse propensity weighting (IPW) matrix factorization (Schnabel et al., 2016), we following Section 6.2 of Schnabel et al. (2016). We set the propensity p_{ui} to be $p_{ui} = k$ for ratings $y_{ui} \geq 4$ and $p_{ui} = k\alpha^{4-y_{ui}}$ for ratings $y_{ui} < 4$, where $\alpha = 0.25$ and k is set so that $\frac{1}{U\cdot I}\sum_{u=1}^{U}\sum_{i=1}^{I} p_{ui} = 0.05$. We then perform inverse propensity weighting (IPW) matrix factorization with the estimated propensities.

E

Proofs for Chapter 8

E.1 Substitute confounder and single-outcome variables

In this section, we show that a substitute confounder cannot be both single-outcome and pinpointed.

Lemma E.1.1: . No single-outcome variable can be measurable with respect to a valid substitute confounder.

Proof. This proof repeats the proof of Proposition 4.1.1 (or equivalently Lemma 4 in Wang and Blei (2019a)).

The key idea behind the proof is the following. We assume the outcomes pinpoint the substitute confounder (Assumption 8.1.2). Moreover, we require that the converse is not true, i.e., that the substitute does not pinpoint the outcomes, or equivalently, the outcomes are not deterministic functions of the substitute confounder. Further, we require that the factor model can not have one component of the substitute *a priori* be a deterministic function of another component. Probabilistic factor models also satisfy all these requirements.

The proof then follows by contradiction: if the substitute picks up single-outcome variables then the factor model must be "degenerate," i.e., non-probabilistic.

Suppose the substitute confounder Z does pick up a single-outcome variable. Then separate Z into a single-outcome component and a multi-outcome one, $Z = (Z_{\mathrm{s}}, Z_{\mathrm{m}})$. Without loss of generality, assume the single-outcome component only depends on the study outcome. A valid substitute confounder must satisfy

$$p(z \mid y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}, \theta) = p(z_{\mathrm{s}}, z_{\mathrm{m}} \mid y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta) = \delta_{(f_{\mathrm{s}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta),\, f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta))}, \quad \text{(E.1)}$$

where $f(\cdot)$ are the deterministic functions that map outcomes to substitute confounders.

Now calculate the conditional distribution of the single-outcome component given the outcomes,

$$\begin{aligned}
&p(z_{\mathrm{s}} \mid y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}) \\
&\quad = p(z_{\mathrm{s}} \mid y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}, z_{\mathrm{m}} = f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta))) && \text{(E.2)} \\
&\quad = p(z_{\mathrm{s}} \mid y, z_{\mathrm{m}} = f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta))) && \text{(E.3)} \\
&\quad = \frac{p(z_{\mathrm{s}} \mid z_{\mathrm{m}} = f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta)) \cdot p(y \mid z_{\mathrm{s}}, z_{\mathrm{m}} = f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta))}{p(y \mid z_{\mathrm{m}} = f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}\,;\,\theta))}. && \text{(E.4)}
\end{aligned}$$

Equation (E.2) is due to the pinpointability assumption. Equation (E.3) is due to $Z_{\mathrm{s}} \perp Y_{1:m}^{\emptyset} \mid Y, Z_{\mathrm{m}}$. Equation (E.4) is due to the definition of conditional probability.

Equations (E.1) and (E.4) imply that at least one of $p(z_{\mathrm{s}} \mid z_{\mathrm{m}} = f_{\mathrm{m}}(\boldsymbol{a}\,;\,\theta))$ and $p(y \mid z_{\mathrm{s}}, z_{\mathrm{m}} = f_{\mathrm{m}}(y, y_1^{\emptyset}, \ldots, y_m^{\emptyset}, \,;\,\theta))$ is a point mass. But this is a contradiction: either term being a point mass implies that the factor model is degenerate. The former is a point mass when one component Z_{s} of the substitute is a deterministic function of another component Z_{m}. The latter is a point mass when the study outcome is a deterministic function of the latent Z. □

E.2 Proof of Theorem 8.1.1

Proof. Assumptions 8.1.2 and 8.1.3 guarantee that the substitute confounder Z_i and the covariates X_i satisfy weak unconfoundedness. Moreover, Z_i can be pinpointed from the observed data. Further, Assumption 8.1.4 ensures that overlap holds with Z_i, X_i. Hence, we can treat Z_i as if it were observed confounders as X_i and proceed with causal inference:

$$\mathbb{E}\left[Y_i(a)\right] = \mathbb{E}_{Z_i,X_i}\left[\mathbb{E}\left[Y_i(a) \mid Z_i, X_i\right]\right] \tag{E.5}$$

$$= \mathbb{E}_{Z_i,X_i}\left[\mathbb{E}\left[Y_i(a) \mid A_i = a, Z_i, X_i\right]\right] \tag{E.6}$$

$$= \mathbb{E}_{Z_i,X_i}\left[\mathbb{E}\left[Y_i \mid A_i = a, Z_i, X_i\right]\right], \tag{E.7}$$

where the first equality is due to the tower property, the second equality is due to Assumption 8.1.3, and the third equality is due to SUTVA.

Similarly, we have

$$\mathbb{E}\left[Y_i(a) \mid A_i = a'\right] = \mathbb{E}_{Z_i,X_i \mid A_i=a'}\left[\mathbb{E}\left[Y_i(a) \mid A_i = a', Z_i, X_i\right]\right] \tag{E.8}$$

$$= \mathbb{E}_{Z_i,X_i \mid A_i=a'}\left[\mathbb{E}\left[Y_i(a) \mid A_i = a, Z_i, X_i\right]\right] \tag{E.9}$$

$$= \mathbb{E}_{Z_i,X_i \mid A_i=a'}\left[\mathbb{E}\left[Y_i \mid A_i = a, Z_i, X_i\right]\right]. \tag{E.10}$$

□

E.3 Proof of Theorem 8.2.1

We first prove the first part.

$$\mathbb{E}[\sum_{A_i=0} w_i^{\text{IPW}} f_y(Y_{ij}^{\emptyset})] \tag{E.11}$$

$$= \mathbb{E}\left[\frac{1}{nP(A_i=1)}\sum_{i=1}^{n} \cdot \mathbb{1}\{A_i=0\}\frac{P(A_i=1\,|\,Z_i,X_i)}{P(A_i=0\,|\,Z_i,X_i)} f_y(Y_{ij}^{\emptyset})\right] \tag{E.12}$$

$$= \frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n} \frac{P(A_i=1\,|\,Z_i,X_i)}{P(A_i=0\,|\,Z_i,X_i)}\mathbb{E}\left[\mathbb{1}\{A_i=0\} f_y(Y_{ij}^{\emptyset})\,|\,Z_i,X_i\right]\right] \tag{E.13}$$

$$= \frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n} \frac{P(A_i=1\,|\,Z_i,X_i)}{P(A_i=0\,|\,Z_i,X_i)}\mathbb{E}\left[\mathbb{1}\{A_i=0\}\,|\,Z_i,X_i\right]\mathbb{E}\left[f_y(Y_{ij}^{\emptyset})\,|\,Z_i,X_i\right]\right] \tag{E.14}$$

$$= \frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\{A_i=1\}\,|\,Z_i,X_i)\right]\mathbb{E}\left[f_y(Y_{ij}^{\emptyset})\,|\,Z_i,X_i\right]\right] \tag{E.15}$$

$$= \frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}\{A_i=1\} f_y(Y_{ij}^{\emptyset})\,|\,Z_i,X_i)\right]\right] \tag{E.16}$$

$$= \frac{1}{nP(A_i=1)}\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\{A_i=1\} f_y(Y_{ij}^{\emptyset})\right] \tag{E.17}$$

$$= \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n}\mathbb{1}\{A_i=1\}}\sum_{A_i=1} f_y(Y_{ij}^{\emptyset})\right] \tag{E.18}$$

The first equality is due to the definition of the inverse probability weights. The second equality is due to the tower property. The third equality is due to $f_y(Y_{ij}^{\emptyset}) \perp A_i \,|\, Z_i, X_i$. The fourth equality is due to canceling out $P(A_i = 0\,|\,Z_i, X_i)$. The fifth equality is again due to $f_y(Y_{ij}^{\emptyset}) \perp A_i \,|\, Z_i, X_i$.. The sixth equality is due to the tower property.

We can prove the other equalities of the first part with the exact same argument.

Next, prove the second part.

First, when $m \to \infty$ (due to Assumption 8.1.2),

$$P(Z_i \,|\, Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}) = \delta_{f(Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})} = \delta_{f_1(Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset})}, \tag{E.19}$$

for some function f_1. The reason is that when $m \to \infty$, any infinite subset of $Y_i, Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$ would pinpoint Z_i (Chen et al., 2019). Hence $Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$ would be sufficient to pinpoint Z_i, i.e. $Z_i \perp Y_i \mid Y_{i1}^{\emptyset}, \ldots, Y_{im}^{\emptyset}$. Therefore, Equation (E.19) holds and f_1 must be linear as is f, which implies

$$\mathbb{E}\left[\sum_{A_i=0} w_i^{\mathrm{IPW}} Z_i\right] = \mathbb{E}\left[\frac{1}{\sum_{A_i=1}^{n} \mathbb{1}\{A_i = 1\}} \sum_{A_i=1} Z_i\right] \tag{E.20}$$

Next, using the Lagrangian multipliers, one can show the solution to the optimization problem with constraints Equation (8.35), Equation (E.20) and Equation (8.34) is

$$w_i^{\mathrm{SC}} = \exp(\alpha + \theta_Z Z_i + \theta_X \cdot X_i)), \tag{E.21}$$

where $\alpha = \log \frac{P(A_i=1)}{P(A_i=0)}$ (Cover and Thomas, 2012; Zhao and Percival, 2017). This solution of w_i^{SC} coincides with w_i^{IPW} when the propensity score model is log-linear in Z_i, X_i. Moreover, w_i^{SC} satisfies Equation (8.33) because $Y_{ij}^{\emptyset}$ is weak unconfounded given Z_i, X_i (Assumption 8.1.3).

F

Other research work during Ph.D. studies

This chapter summarizes my other research work during Ph.D. studies.

Theory of variational inference The empirical success of many machine learning algorithms is exciting. However, a fundamental theoretical understanding of why they work well is necessary for practitioners to apply them with trust. One such algorithm is variational Bayes (VB), which is a modern alternative to the classical Markov chain Monte Carlo for latent variable inference. Despite its wide empirical success in massive datasets, it has been a long-standing open problem to develop general theoretical guarantees for variational Bayes.

Along this line, Wang and Blei (2019b) establish frequentist consistency and asymptotic normality of VB methods. Specifically, we connect VB methods to point estimates based on variational approximations, called frequentist variational approximations, and we use the connection to prove a variational Bernstein–von Mises theorem. The theorem leverages the theoretical characterizations of frequentist variational approximations to understand asymptotic properties of VB. In summary, we prove that (1) the VB posterior converges to the Kullback-Leibler (KL) minimizer of a normal distribution, centered

at the truth and (2) the corresponding variational expectation of the parameter is consistent and asymptotically normal. As applications of the theorem, we derive asymptotic properties of VB posteriors in Bayesian mixture models, Bayesian generalized linear mixed models, and Bayesian stochastic block models. We conduct a simulation study to illustrate these theoretical results.

Further, Wang and Blei (2019c) study VB under model misspecification. We prove the VB posterior is asymptotically normal and centers at the value that minimizes the KL divergence to the true data-generating distribution. Moreover, the VB posterior mean centers at the same value and is also asymptotically normal. These results generalize the variational Bernstein–von Mises theorem (Wang and Blei, 2019b) to misspecified models. As a consequence of these results, we find that the model misspecification error dominates the variational approximation error in VB posterior predictive distributions. It explains the widely observed phenomenon that VB achieves comparable predictive accuracy with Markov chain Monte Carlo (MCMC) even though VB uses an approximating family. As illustrations, we study VB under three forms of model misspecification, ranging from model over-/under-dispersion to latent dimensionality misspecification. We conduct two simulation studies that demonstrate the theoretical results.

Finally, Wang et al. (2019a) establish a connection between variational autoencoder (VAE) and empirical Bayes. We show that a VAE approximates one form of empirical Bayes inference: in Efron's language (Efron et al., 2019), it performs g-modeling with a particular parametric form of g.

Model misspecification and robust machine learning Many machine learning algorithms require positing a correct model of the data generating process. However, in practice, all models are wrong; they only approximate the reality. The challenge of model misspecification calls for new ways of thinking in designing robust machine learning algorithms.

In this direction, Wang et al. (2017) propose a way to systematically detect and mitigate mismatch of a large class of probabilistic models. The idea is to raise the likelihood of each observation to a weight and then to infer both the latent variables and the weights from data. Inferring the weights allows a model to identify observations that match its assumptions and down-weight others. This enables robust inference and improves predictive accuracy. We study four different forms of mismatch with reality, ranging from missing latent groups to structure misspecification. A Poisson factorization analysis of the Movielens 1M dataset shows the benefits of this approach in a practical scenario.

Wang and Zubizarreta (2019) study causal inference algorithms that are robust to model misspecification. We focus on *minimal dispersion approximately balancing weights*, an algorithm that can consistently estimate average treatment effects without positing a treatment assignment model. This model-free approach liberates us from biases due to model misspecification. In more detail, we observe a connection between approximate covariate balance and shrinkage estimation of the propensity score. This connection leads to both theoretical and practical developments. From a theoretical standpoint, we characterize the asymptotic properties of minimal weights and show that, under standard smoothness conditions on the propensity score function, minimal weights are consistent estimates of the true inverse probability weights. Also, we show that the resulting weighting estimator is consistent, asymptotically normal, and semiparametrically efficient. From a practical standpoint, we present a finite sample oracle inequality that bounds the loss incurred by balancing more functions of the covariates than strictly needed. This inequality shows that minimal weights implicitly bound the number of active covariate balance constraints. We finally provide a tuning algorithm for choosing the degree of approximate balance in minimal weights. We conclude with four empirical studies that suggest approximate balance is preferable to exact balance, especially when there is limited overlap in covariate distributions. In these studies, we show that the root mean squared error of the weighting estimator can be reduced by as much as a half with approximate balance.

Fair machine learning In additional to theoretical correctness, machine learning algorithms are often required to be fair in order to be deployed in practice. To this end, Wang et al. (2019b) design algorithms that automate decision-making while reliably repairing historical discriminations. In more detail, we propose two algorithms that adjust fitted machine learning (ML) predictors to make them fair. We focus on two legal notions of fairness: (a) providing equal opportunities (EO) to individuals regardless

of sensitive attributes and (b) repairing historical disadvantages through affirmative action (AA). More technically, we produce fair EO and AA predictors by positing a causal model and considering counterfactual decisions. We prove that the resulting predictors are theoretically optimal in predictive performance while satisfying fairness. We evaluate the algorithms, and the trade-offs between accuracy and fairness, on datasets about admissions, income, credit and recidivism.

Bibliography

Chen, Y., Li, X., and Zhang, S. (2019). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

D'Amour, A. (2018). (Non-)identification in latent confounder models. http://www.alexdamour.com/blog/public/2018/05/18/non-identification-in-latent-confounder-models/. Accessed: 2018-05-29.

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

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

Efron, B. et al. (2019). Bayes, oracle Bayes and empirical Bayes. *Statistical Science*, 34(2):177–201.

Hu, Y., Koren, Y., and Volinsky, C. (2008). Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. IEEE.

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.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298(5602):2381–2385.

Schnabel, T., Swaminathan, A., Singh, A., Chandak, N., and Joachims, T. (2016). Recommendations as treatments: Debiasing learning and evaluation. *arXiv preprint arXiv:1602.05352*.

Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554.

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

Wang, Y. and Blei, D. M. (2019b). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Wang, Y. and Blei, D. M. (2019c). Variational Bayes under model misspecification. In *Advances in Neural Information Processing Systems*, pages 13357–13367.

Wang, Y., Kucukelbir, A., and Blei, D. M. (2017). Robust probabilistic modeling with Bayesian data reweighting. In *International Conference on Machine Learning*, pages 3646–3655.

Wang, Y., Miller, A. C., Blei, D. M., et al. (2019a). Comment: Variational autoencoders as empirical Bayes. *Statistical Science*, 34(2):229–233.

Wang, Y., Sridhar, D., and Blei, D. M. (2019b). Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*.

Wang, Y. and Zubizarreta, J. R. (2019). Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika*. asz050.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

Zhao, Q. and Percival, D. (2017). Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1).