



The Blessings of Multiple Causes

Yixin Wang & David M. Blei

To cite this article: Yixin Wang & David M. Blei (2019) The Blessings of Multiple Causes, Journal of the American Statistical Association, 114:528, 1574-1596, DOI: [10.1080/01621459.2019.1686987](https://doi.org/10.1080/01621459.2019.1686987)

To link to this article: <https://doi.org/10.1080/01621459.2019.1686987>



View supplementary material [↗](#)



Accepted author version posted online: 31 Oct 2019.
Published online: 03 Jan 2020.



Submit your article to this journal [↗](#)



Article views: 613



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



The Blessings of Multiple Causes

Yixin Wang^a and David M. Blei^{a,b}

^aDepartment of Statistics, Columbia University, New York, NY; ^bDepartment of Computer Science, Columbia University, New York, NY

ABSTRACT

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 causal variables and the outcome variables. This assumption is standard but it is also untestable. In this article, we develop the deconfounder, a way to do causal inference with weaker assumptions than the traditional methods require. The deconfounder is designed for problems of multiple causal inference: scientific studies that involve multiple causes whose effects are simultaneously of interest. Specifically, the deconfounder combines unsupervised machine learning and predictive model checking to use the dependencies among multiple causes as indirect evidence for some of the unobserved confounders. We develop the deconfounder algorithm, prove that it is unbiased, and show that it requires weaker assumptions than traditional causal inference. We analyze its performance in three types of studies: semi-simulated data around smoking and lung cancer, semi-simulated data around genome-wide association studies, and a real dataset about actors and movie revenue. The deconfounder is an effective approach to estimating causal effects in problems of multiple causal inference. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received June 2018
Accepted September 2019

KEYWORDS

Causal inference; Machine learning; Probabilistic models; Unconfoundedness; Unobserved confounding.

1. Introduction

Here is a frivolous, but perhaps lucrative, causal inference problem. Table 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 interested in the causal effect of each actor; for example, how much does revenue increase (or decrease) if Oprah Winfrey is in the movie?

To solve this problem, the producer wants to use the potential outcomes approach to causal inference (Rubin 1974, 2005; Imbens and Rubin 2015). Following the methodology, she associates each movie to a *potential outcome function*, $y_i(\mathbf{a})$. This function maps each possible cast \mathbf{a} to its revenue if the movie i had that cast. (The cast \mathbf{a} is a binary vector with one element per actor; each element encodes whether the actor is in the movie.) 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 $Y_i(\mathbf{a})$. For example, she might consider a particular cast \mathbf{a} and estimate the expected revenue of a movie with that cast, $\mathbb{E}[Y_i(\mathbf{a})]$.



Traditionally, causal inference from observational data is a difficult enterprise and requires strong assumptions. The challenge is that the 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. This article shows how multiple causal inference can be easier than traditional causal inference. Thanks to the multiplicity of causes, the producer can make causal inferences under weaker assumptions than the traditional approaches require.


Let's discuss the producer's inference in more detail: how can she calculate $\mathbb{E}[Y_i(\mathbf{a})]$? Naively, she subsets the data in Table 1 to those with cast equal to \mathbf{a} , and then computes a Monte Carlo estimate of the revenue. This procedure is unbiased when $\mathbb{E}[Y_i(\mathbf{a})] = \mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}]$.

But there is a problem. The data in Table 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 unobserved, the genre of the movie produces a statistical dependence between whether an actor is cast and the revenue; this dependence biases the causal estimates, $\mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}] \neq \mathbb{E}[Y_i(\mathbf{a})]$.

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 inference is simple: use the data (now with confounders) to take Monte Carlo estimates of $\mathbb{E}[\mathbb{E}[Y_i(\mathbf{a}) | W_i, \mathbf{A}_i = \mathbf{a}]]$; this iterated expectation “controls” for the confounders. But the problem is that whether the estimate is equal to $\mathbb{E}[Y_i(\mathbf{a})]$ rests on an uncheckable assumption: there are no other confounders. For many applied causal inference problems, this assumption is a leap of faith.

CONTACT Yixin Wang  yixin.wang@columbia.edu  Department of Statistics, Columbia University, Room 1005 SSW, MC 4690, 1255 Amsterdam Avenue, New York, NY 10027.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2019 American Statistical Association

Table 1. The TMDB dataset of movie earnings.

Title	Cast	Revenue
<i>Avatar</i>	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, ...}	\$2788M
<i>Titanic</i>	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, ...}	\$1845M
<i>The Avengers</i>	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, ...}	\$1520M
<i>Jurassic World</i>	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, ...}	\$1514M
<i>Furious 7</i>	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, ...}	\$1506M
⋮	⋮	⋮
<i>The Divide</i>	{Lauren German, Michael Biehn, Milo Ventimiglia, Courtney B. Vance, ...}	\$22,000

We develop *the deconfounder*, an alternative method for the 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 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. 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 (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.

Subject to the assumptions, the deconfounder provides valid causal inferences because it capitalizes on the dependency structure of the observed casts. 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. 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, for example, 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 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 affect both which items are recommended and which items they buy, that is, 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.

1.1. Related Work

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

1.1.1. 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) builds on sparse methods to infer causal structures; Moghaddass, Rudin, and Madigan (2016) use 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 covari-

ance estimation under latent confounding, Frot, Nandy, and Maathuis (2019) use linear factor models for robust causal structure learning with hidden variables, and Kaltenpoth and Vreeken (2019) leverages 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 **genome-wide association studies (GWAS)**, treating the **single-nucleotide polymorphisms (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 articles use Pearl's framework (Pearl 2009); they hypothesize a causal graph with confounders, causes, and outcomes. This article complements these works. We develop the deconfounder in the potential outcomes framework (Rubin 1974, 2005; Imbens and Rubin 2015).

1.1.2. Analyzing GWAS

In **GWAS**, latent population structure is an important unobserved confounder. Pritchard et al. (2000) propose a probabilistic admixture model for unsupervised ancestry inference. Price et al. (2006) and Astle and Balding (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, Hao, and Storey (2015) and Hao, Song, and Storey (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 Appendix A in the supplementary materials). 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.

1.1.3. Assessing the Unconfoundedness Assumption

Rosenbaum and Rubin (1983) demonstrate that unconfoundedness and a good propensity score model are sufficient to perform causal inference with observational data. Many subsequent efforts assess the plausibility of unconfoundedness. For example, Robins, Rotnitzky, and Scharfstein (2000), Gilbert, Bosch, and Hudgens (2003), and Imai and Van Dyk (2004) 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, Hofman, and Watts (2016) leveraged auxiliary outcome data to test for confounding; Janzing and Schölkopf (2018a; 2018b), and Liu and Chan (2018) developed tests for non-confounding in multivariate linear regression; Cinelli et al. (2019) developed sensitivity analysis for linear causal models;

Franks, D'Amour, and Feller (2019) designed 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.

1.1.4. The (Generalized) Propensity Score

Schneeweiss et al. (2009), McCaffrey, Ridgeway, and Morral (2004), Lee, Lessler, and Stuart (2010), and many others develop and evaluate different models for assigned causes. In particular, Chernozhukov et al. (2017) 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 article 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.

1.1.5. Traditional Causal Inference With Multiple Treatments

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

1.2. This Article

Section 2 reviews traditional causal inference, sets up multiple causal inference, presents the deconfounder. Section 3 describes the identification strategy of the deconfounder and its main assumptions. Section 4 discusses the practical details of the deconfounder and presents the full algorithm. Section 5 answers some questions a reader might have. Section 6 presents three empirical studies, two semi-synthetic and one real. Section 7 further develops the theory around the deconfounder and establishes causal identification. Section 8 concludes the article.

2. Multiple Causal Inference With the Deconfounder

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

2.1. Multiple Causal Inference

We first describe multiple causal inference. In the data, there are m possible causes, encoded in a vector $\mathbf{a} = (a_1, \dots, a_m)$. 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 (revenue). We focus on real-valued outcomes. For the i th movie, the potential outcome function maps each possible cast to the log of the movie's revenue had it had that cast, $y_i(\mathbf{a}) : \{0, 1\}^m \rightarrow \mathbb{R}$.

The goal of causal inference is to characterize the sampling distribution of the potential outcomes $Y_i(\mathbf{a})$ for each configura-

tion of the causes \mathbf{a} . This distribution provides causal inferences, such as the expected outcome for a particular array of causes (a particular cast of actors) $\mu(\mathbf{a}) = \mathbb{E}[Y_i(\mathbf{a})]$ or the average effect of individual causes (how much a particular actor contributes to revenue).

To help make causal inferences, we draw data from the sampling distribution of assigned causes \mathbf{a}_i (the cast of movie i) and realized outcomes $y_i(\mathbf{a}_i)$ (its revenue).¹ The data is $\mathcal{D} = \{(\mathbf{a}_i, y_i(\mathbf{a}_i))_{i=1}^n\}$. Note we only observe the outcome for the assigned causes $y_i(\mathbf{a}_i)$, 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 $Y_i(\mathbf{a})$ for any \mathbf{a} ; this is the “fundamental problem of causal inference” (Holland 1986).

To estimate $\mu(\mathbf{a})$, consider using the data to calculate conditional Monte Carlo approximations of $\mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}]$. These estimates are simply averages of the outcomes for each configuration of the causes. But this approach may not be accurate. There might be *unobserved confounders*—hidden variables that affect both the assigned causes \mathbf{A}_i and the potential outcome function $Y_i(\mathbf{a})$. When there are unobserved confounders, the assigned causes are correlated with the observed outcome. Consequently, Monte Carlo estimates of $\mu(\mathbf{a})$ are biased,

$$\mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}] \neq \mathbb{E}[Y_i(\mathbf{a})]. \quad (1)$$

We can estimate $\mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}]$ with the dataset; but the goal is to estimate $\mathbb{E}[Y_i(\mathbf{a})]$.²

Suppose we measure covariates x_i and append to each data point, $\mathcal{D} = \{(\mathbf{a}_i, x_i, y_i(\mathbf{a}_i))_{i=1}^n\}$. If these covariates contain all confounders then

$$\mathbb{E}[\mathbb{E}[Y_i(\mathbf{a}) | X_i, \mathbf{A}_i = \mathbf{a}]] = \mathbb{E}[Y_i(\mathbf{a})]. \quad (2)$$

With augmented data, estimate the left side with Monte Carlo; thus, estimate $\mathbb{E}[Y_i(\mathbf{a})]$.

Equation (2) is true when X capture all confounders. More precisely, it is true under the assumption of *unconfoundedness*³ (Rosenbaum and Rubin 1983; Imai and Van Dyk 2004): conditional on observed X , the assigned causes are independent of the potential outcomes,

$$\mathbf{A}_i \perp\!\!\!\perp Y_i(\mathbf{a}) | X_i \quad \forall \mathbf{a}. \quad (3)$$

¹We 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).

²Here is the notation. Capital letters denote a random variable. For example, the random variable \mathbf{A}_i is a randomly chosen vector of assigned causes from the population. The random variable $Y_i(\mathbf{A}_i)$ is a randomly chosen potential outcome from the population, evaluated at its assigned causes. A lowercase letter is a realization. For example, \mathbf{a}_i is in the dataset—it is the vector of assigned causes of individual i . The left side of Equation (1) is an expectation with respect to the random variables; it conditions on the random vector of assigned causes to be equal to a certain realization $\mathbf{A}_i = \mathbf{a}$. The right side is an expectation over the same population of the potential outcome functions, but always evaluated at the realization \mathbf{a} .

³Here we describe the weak version of the unconfoundedness assumption, which requires individual potential outcomes $Y_i(\mathbf{a})$ be marginally independent of the causes \mathbf{A}_i , that is, $\mathbf{A}_i \perp\!\!\!\perp Y_i(\mathbf{a}) | X_i$ for all \mathbf{a} . Imbens (2000) and Hirano and Imbens (2004) call this assumption *weak unconfoundedness*. In contrast, the strong version of unconfoundedness says $\mathbf{A}_i \perp\!\!\!\perp (Y_i(\mathbf{a}))_{\mathbf{a} \in \mathcal{A}} | X_i$, which requires all possible potential outcomes $(Y_i(\mathbf{a}))_{\mathbf{a} \in \mathcal{A}}$ be jointly independent of the causes \mathbf{A}_i .

The nuance is that Equation (3) needs to hold for all possible \mathbf{a} ’s, not only for the value of $Y_i(\mathbf{a})$ at the assigned causes. Unconfoundedness implies no unobserved confounders.⁴

Equation (2) underlies the practice of causal inference: find and measure the confounders, estimate conditional expectations, and average. In the introduction, for example, we pointed out that the genre of the movie is a confounder to 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).

2.2. 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 *probabilistic factor model* to capture the joint distribution of causes $p(\mathbf{a}_1, \dots, \mathbf{a}_m)$. A 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

$$\begin{aligned} Z_i &\sim p(\cdot | \alpha) \quad i = 1, \dots, n, \\ \mathbf{A}_{ij} | Z_i &\sim p(\cdot | z_i, \theta_j) \quad j = 1, \dots, m, \end{aligned} \quad (4)$$

where α parameterizes the distribution of Z_i and θ_j parameterizes the per-cause distribution of \mathbf{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 (Pritchard et al. 2000; Blei, Ng, and Jordan 2003; Erosheva 2003; Airoldi et al. 2008), and deep generative models (Neal 1990; Kingma and Welling 2013; Rezende and Mohamed 2015; Mohamed and Lakshminarayanan 2016; Ranganath et al. 2015; Ranganath, Tran, and Blei 2016; Tran et al. 2017).

We can fit using any appropriate method, such as maximum likelihood estimation or 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 $p(\mathbf{a})$.

⁴We also assume *stable unit treatment value assumption* (SUTVA) (Rubin 1980, 1990) and *overlap* (Imai and Van Dyk 2004), roughly that any vector of assigned causes has positive probability. These three assumptions together identify the potential outcome function (Imbens 2000; Hirano and Imbens 2004; Imai and Van Dyk 2004).

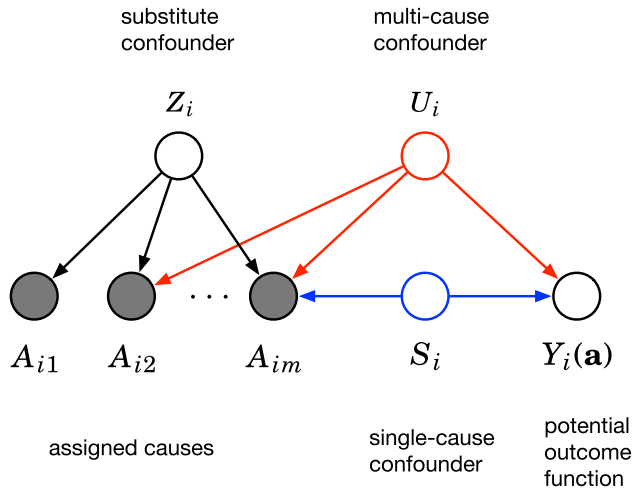


Figure 1. A graphical model argument for the deconfounder. The punchline is that if Z_i renders the A_{ij} 's conditionally independent then there cannot be a multi-cause confounder. The proof is by contradiction. Assume conditional independence holds, $p(a_{i1}, \dots, a_{im} | z_i) = \prod_j p(a_{ij} | z_i)$; 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).

In the next step, use the fitted factor model to calculate the conditional expectation of each individual's local factor weights $\hat{z}_i = \mathbb{E}_M[Z_i | A_i = \mathbf{a}_i]$. 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 $\mathbb{E}[Y_i(\mathbf{a})] = \mathbb{E}\left[\mathbb{E}\left[Y_i(\mathbf{a}) | \hat{Z}_i, A_i = \mathbf{a}\right]\right]$.

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

$$p(a_{i1}, \dots, a_{im} | z_i) = \prod_{j=1}^m p(a_{ij} | z_i). \quad (5)$$

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 (5) implies unconfoundedness, $A_i \perp\!\!\!\perp Y_i(\mathbf{a}) | Z_i$, and unconfoundedness justifies causal inference. In summary, if the 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 1 justifies the deconfounder and reveals its assumptions.⁵ Suppose we observe a Z_i such that the conditional independence in Equation (5) 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 (5) does not hold. Thus, U_i cannot exist.

There is a nuance. The conditional independence in Equation (5) cannot rule out the existence of an unobserved single-cause confounder, denoted S_i in Figure 1. Even if such a confounder exists, the conditional independence still holds.

Here is the punchline. If we find a factor model that accurately represents the distribution of causes then that model can provide a variable that captures the unobserved multiple-cause confounders. The reason is that the multiple-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. The Identification Strategy of the Deconfounder

How does the deconfounder identify potential outcomes? The classical strategy for causal identification is that unconfoundedness, together with **stable unit treatment value assumption (SUTVA)** and overlap, identifies the potential outcomes (Imbens 2000; Hirano and Imbens 2004; Imai and Van Dyk 2004). 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 multiple-cause confounders. (The proof is in Section 7.) 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).

3.1. Stable Unit Treatment Value Assumption (SUTVA)

The **stable unit treatment value assumption (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 (2015) for discussion of this assumption.

3.2. 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 1. (The precise technical definition is in Definition 4 of Section 7.)

This assumption is weaker than 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.

⁵Figure 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.

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, it is likely that several are 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 SNPs—and we observe 100,000 SNPs per individual—then the confounder may be unlikely to have an effect on 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 (Robins, Rotnitzky, and Scharfstein 2000; Gilbert, Bosch, and Hudgens 2003; Imai and Van Dyk 2004; Cinelli et al. 2019; Franks, D’Amour, and Feller 2019) 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.”

3.3. Overlap

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

$$p(A_{ij} \in \mathcal{A} | Z_i) > 0 \text{ for all sets } \mathcal{A} \text{ with positive measure,} \\ \text{i.e., } p(\mathcal{A}) > 0. \quad (6)$$

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).⁶

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

For probabilistic factor models, the overlap condition is usually satisfied. For example, probabilistic PCA assumes $A_{ij} | Z_i \sim \mathcal{N}(Z_i^\top \theta_j, \sigma^2)$. The normal distribution has support over the real line, which ensures $p(A_{ij} \in \mathcal{A} | Z_i) > 0$ 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 $p(A_{ij} \in \mathcal{A} | Z_i)$ 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, that is, 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, for example, 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 $p(A_{ij} | Z_i)$ for all individual i ’s. A conditional with low variance or low entropy signals limited overlap and the possibility of high-variance causal estimates.

3.4. The Deconfounder Is Unbiased

We have described the main assumptions of the deconfounder. With **SUTVA**, overlap, and no unobserved single-cause confounders, we use the deconfounder to estimate causal quantities. Note that point identification of causal quantities requires further assumptions; see Section 7 for a discussion of these additional assumptions.

The deconfounder (informal version of Theorem 6). Assume **SUTVA** and no unobserved single-cause confounders. Then the deconfounder provides an unbiased estimate of the average causal effect:

$$\mathbb{E}_Y[Y_i(a_1, \dots, a_m)] - \mathbb{E}_Y[Y_i(a'_1, \dots, a'_m)] \\ = \mathbb{E}_{X,Z}[\mathbb{E}_Y[Y_i | A_{i1} = a_1, \dots, A_{im} = a_m, X_i, Z_i]] \\ - \mathbb{E}_{X,Z}[\mathbb{E}_Y[Y_i | A_{i1} = a'_1, \dots, A_{im} = a'_m, X_i, Z_i]], \quad (7)$$

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 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 4.) Thus, along with no unobserved single-cause confounders, the substitute confounder provides unconfoundedness. 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 identification of

⁶We also require the observed covariates X_i satisfy the overlap condition if they are single-cause confounders, that is, $p(A_{ij} \in \mathcal{A} | X_i) > 0$ for all sets \mathcal{A} with positive measure, i.e. $p(\mathcal{A}) > 0$.

different causal quantities in Section 7 and lay out the assumptions required for each.

4. 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 $\mathbb{E}[Y_i(\mathbf{a}) | \hat{Z}_i, \mathbf{A}_i = \mathbf{a}]$ for performing causal inference. We discuss each ingredient below (Sections 4.1 and 4.2) and then describe the full deconfounder algorithm (Section 4.3).

4.1. Using the Assignment Model to Infer a Substitute Confounder

The first ingredient is a factor model of the assigned causes, as defined in Equation (4), 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 (4); 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 $\theta_j, j = 1, \dots, m$. 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.

4.1.1. 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 $p(z_i | \mathbf{a}_i)$. In the simulations and studies of Section 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

$$Z_{ik} \sim \mathcal{N}(0, \lambda^2), \quad k = 1, \dots, K, \quad (8)$$

$$A_{ij} | Z_i \sim \mathcal{N}(z_i^\top \theta_j, \sigma^2), \quad j = 1, \dots, m. \quad (9)$$

We can fit probabilistic PCA with maximum likelihood (or Bayesian methods) and use standard conditional probability to calculate $p(z_i | \mathbf{a}_i)$. Exponential family extensions of PCA are also factor models (Collins, Dasgupta, and Schapire 2002; Mohamed, Ghahramani, and Heller 2009) as are some deep generative models (Tran et al. 2017), which can be interpreted as a non-linear probabilistic PCA.

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

$$Z_{ik} \sim \text{Gamma}(\alpha_0, \alpha_1), \quad k = 1, \dots, K, \quad (10)$$

$$A_{ij} | Z_i \sim \text{Poisson}(z_i^\top \theta_j), \quad j = 1, \dots, m. \quad (11)$$

PF can be fit to large datasets with variational methods (Gopalan, Hofman, and Blei 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, Mohamed, and Wierstra 2014). For example, a two-layer DEF models each observation as

$$Z_{2,l} \sim \text{Exp-Fam}_2(\alpha), \quad l = 1, \dots, L, \quad (12)$$

$$Z_{1,k} | Z_{2,i} \sim \text{Exp-Fam}_1(g_1(z_{2,i}^\top \theta_{1,k})), \quad k = 1, \dots, K, \quad (13)$$

$$A_{ij} | Z_{1,i} \sim \text{Exp-Fam}_0(g_0(z_{1,i}^\top \theta_{0,j})), \quad j = 1, \dots, m. \quad (14)$$

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 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, Gerrish, and Blei 2014).

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

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

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

$$p(\mathbf{a}_{i,\text{held}}^{\text{rep}} | \mathbf{a}_{i,\text{obs}}) = \int p(\mathbf{a}_{i,\text{held}} | z_i) p(z_i | \mathbf{a}_{i,\text{obs}}) dz_i. \quad (15)$$

This distribution integrates out the local posterior $p(z_i | \mathbf{a}_{i,\text{obs}})$. (An approximate posterior also suffices; we discuss why in Section 5.)

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

$$t(\mathbf{a}_{i,\text{held}}) = \mathbb{E}_Z [\log p(\mathbf{a}_{i,\text{held}} | Z) | \mathbf{a}_{i,\text{obs}}], \quad (16)$$

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.

Finally calculate the predictive score,

$$\text{predictive score} = p\left(t(\mathbf{a}_{i,\text{held}}^{\text{rep}}) < t(\mathbf{a}_{i,\text{held}})\right). \quad (17)$$

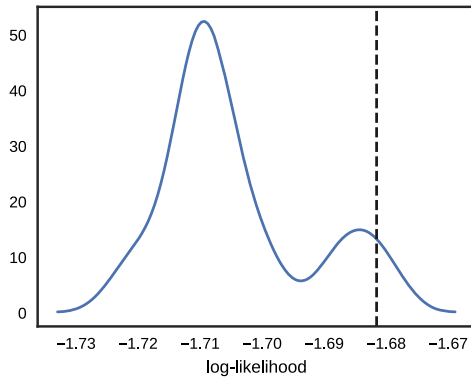


Figure 2. An example of a predictive check for an assignment model. The vertical dashed line shows $t(\mathbf{a}_i, \text{held})$. The blue curve shows the kde of $t(\mathbf{a}_i^{\text{rep}}, \text{held})$. The predictive score is the area under the blue curve to the left of the vertical dashed line. The predictive score of this assignment model is larger than 0.1; we consider it satisfactory.

Here the randomness stems from $\mathbf{a}_i^{\text{rep}}$ coming from the predictive distribution in Equation (15), 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 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 2 illustrates a predictive check of a good assignment model. Section 6 shows predictive checks in action.

These predictive checks blend ideas around **posterior predictive checks (PPCs)** (Rubin 1984), **PPCs** with realized discrepancies (Gelman, Meng, and Stern 1996), **PPCs** with held-out data (Gelfand, Dey, and Chang 1992; Ranganath and Blei 2019), and stage-wise checking of hierarchical models (Dey et al. 1998; Bayarri and Castellanos 2007). 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 (Gelman et al. 2013; Blei 2014).

4.2. 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 $p(z_i | \mathbf{a}_i, \mathcal{D})$ concentrates around a point \hat{z}_i . Then we can use \hat{z}_i as a confounder. Follow Section 2.1 to calculate the iterated expectation on the left side of Equation (2). However, replace the observed confounders with the substitute confounder; the goal is to calculate $\mathbb{E}[\mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}, Z_i]]$.

First, approximate the outside expectation with Monte Carlo,

$$\begin{aligned} \mathbb{E}[\mathbb{E}[Y_i(\mathbf{a}) | \mathbf{A}_i = \mathbf{a}, Z_i]] \\ \approx \frac{1}{n} \sum_{i=1}^n \mathbb{E}_Y[Y_i(\mathbf{A}_i) | \mathbf{A}_i = \mathbf{a}, Z_i = \hat{z}_i]. \end{aligned} \quad (18)$$

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 (18). We fit a function to estimate this quantity,

$$\mathbb{E}[Y_i(\mathbf{A}_i) | \mathbf{A}_i = \mathbf{a}, Z_i = z] = f(\mathbf{a}, z). \quad (19)$$

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

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

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,

$$f(\mathbf{a}, z) = \beta^\top \mathbf{a} + \gamma^\top z + \beta_0. \quad (20)$$

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

$$f(\mathbf{a}, z) = \beta^\top (\mathbf{a} - \hat{\mathbf{a}}(z)) + \beta_0. \quad (21)$$

This model relates to the generalized propensity score (Imbens 2000; Hirano and Imbens 2004), where it uses $\hat{\mathbf{a}}(z)$ as a proxy for the propensity score. Note this substitution is used in Bayesian statistics (Laird and Louis 1982; Tierney and Kadane 1986; Geisser et al. 1990), and is justified when higher moments of the assignment are similar across individuals. In both Equations (20) and (21), the coefficient β represents the average causal effect of raising each cause by one unit.

Note we are not restricted to linear models. Other outcome models like random forests (Wager and Athey 2018) 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 (18). 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 $p(\mathbf{a}_i | \hat{z}_i)$ and approximate the inner expectation within each propensity-score bin (Rosenbaum and Rubin 1983; Lunceford and Davidian 2004). 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 (Horvitz and Thompson 1952; Rosenbaum and Rubin 1983; Heckman et al. 1998; Dehejia and Wahba 2002).

Algorithm 1: The deconfounder

Input: a dataset of assigned causes and outcomes $\{(\mathbf{a}_i, y_i)\}, i = 1, \dots, n$

Output: the average potential outcome $\mathbb{E}[Y(\mathbf{a})]$ for any causes \mathbf{a}

repeat

- choose an assignment model from the class in Equation (4)
- fit the model to the assigned causes $\{\mathbf{a}_i\}, i = 1, \dots, n$
- check the fitted model \hat{M}

until the assignment check is satisfactory

foreach datapoint i **do**

- calculate $\hat{z}_i = \mathbb{E}_{\hat{M}}[Z_i | \mathbf{a}_i]$.

end

repeat

- choose an outcome model from Equation (19)
- fit the outcome model to the augmented dataset $\{(\mathbf{a}_i, y_i, \hat{z}_i)\}, i = 1, \dots, n$
- check the fitted outcome model

until the outcome check is satisfactory

estimate the average causal effect $\mathbb{E}[Y(\mathbf{a})] - \mathbb{E}[Y(\mathbf{a}')] \text{ by Equation (18)}$

4.3. The Full Algorithm and an Example

We described each component of the deconfounder. Algorithm 1 gives the full algorithm, a procedure for estimating Equation (18). The steps are: (1) find, fit, and check a factor model to the dataset of assigned causes; (2) estimate \hat{z}_i for each datapoint; (3) find and fit a outcome model; (4) use the outcome model and estimated \hat{z}_i to do causal inference.

4.3.1. 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 Appendix A in the supplementary materials.

Consider a dataset of $n = 5000$ 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 2 illustrates a snippet of the data (10 individuals).

We simulate such a dataset of genotypes and height. We generate each individual's genotypes by simulating heterogeneous mixing of populations (Pritchard et al. 2000). 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 $\beta^* = (\beta_1^*, \dots, \beta_m^*)$. See Section 6.2 for more details of the simulation.

The goal is to infer how the SNPs causally affect human height, even in the presence of unobserved confounders. The m -dimensional SNP vector $\mathbf{a}_i = (a_{i1}, a_{i2}, \dots, a_{im})$ 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 $\mathbf{a} = (a_1, a_2, \dots, a_m)$? Mathematically, this is the average potential outcome function: $\mathbb{E}[Y_i(\mathbf{a})]$, where the vector of assigned causes \mathbf{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 (10). This fit results in estimates of nonnegative factors $\hat{\theta}_j$ for each assigned cause and nonnegative 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 \mathbf{a}_i and reconstructed assignment $\hat{\mathbf{a}}(z_i)$,

$$y_i \sim \mathcal{N}(\beta_0 + \beta^\top (\mathbf{a}_i - \hat{\mathbf{a}}(z_i)), \sigma^2). \quad (22)$$

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^* .

Table 2. How do SNPs causally affect height?

ID (i)	SNP_1 ($a_{i,1}$)	SNP_2 ($a_{i,2}$)	SNP_3 ($a_{i,3}$)	SNP_4 ($a_{i,4}$)	SNP_5 ($a_{i,5}$)	SNP_6 ($a_{i,6}$)	SNP_7 ($a_{i,7}$)	SNP_8 ($a_{i,8}$)	SNP_9 ($a_{i,9}$)	...	SNP_100K ($a_{i,100K}$)	Height (feet) (y_i)
1	1	0	0	1	0	0	1	2	0	...	0	5.73
2	1	2	2	1	2	1	1	0	1	...	2	5.26
3	2	0	1	1	0	1	0	1	1	...	2	6.24
4	0	0	0	1	1	0	1	2	0	...	0	5.78
5	1	2	1	1	1	0	1	0	0	...	1	5.09
...
10000	1	1	0	0	0	2	0	0	1	...	2	5.45

NOTE: This table shows a portion of a dataset: simulated SNPs as the multiple causes and height as the outcome.

We evaluate the causal estimates obtained with and without the deconfounder. We focus on the root mean squared 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 \mathbf{a}_i .) The rmse is 49.6×10^{-2} without the deconfounder and 41.2×10^{-2} with the deconfounder. The deconfounder produces closer-to-truth causal estimates.

5. 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 7](#) says that a latent variable which 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 a free lunch? The deconfounder is not a 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, that is, 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 classical causal inference is biased.

Why does the deconfounder have two stages? [Algorithm 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 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 “pretreatment” variables; we cannot identify a mediator by looking only at the assigned causes. More formally, excluding the outcome ensures that the model satisfies $p(z_i | \mathbf{a}_i, y_i(\mathbf{a}_i)) = p(z_i | \mathbf{a}_i)$; 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 $P(A_i | Z_i)$ 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}, \dots, 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: $A_{ij} \stackrel{\text{a.s.}}{=} f(Z_i) + \epsilon_{ij}$. 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 j th 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 $P(A_{ij} | Z_i)$ are specified separately.

However, the deconfounder differs from classical instrumental variables approaches 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 [Section 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 “pretreatment”—it does not include any mediators—then the causal estimate will be unbiased (Imai and Van Dyk 2004) (also see Theorem 6). Moreover, to satisfy no unobserved single-cause confounders (Section 3.2), we must condition on single-cause confounders.

That said, we do not need to condition on observed confounders that affect more than one of the causes; it suffices to condition only on the substitute confounder Z_i . And there is a tradeoff. Conditioning on covariates X_i maintains unbiasedness but it hurts efficiency. If the true causal effect size is small then large confidence or credible intervals will conclude these small effects as insignificant—inefficient causal estimates can bury the real causal effects. The empirical study in Section 6.1 explores this phenomenon.

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 $Z_i \sim p(z_i | \mathbf{a}_i)$. It then uses the mean of the substitute confounder $\hat{z}_i = \mathbb{E}_{\hat{M}}[Z_i | \mathbf{a}_i]$ to fit an outcome model $p(y_i | \mathbf{a}_i, \hat{z}_i)$ and compute the potential outcome estimate $\mathbb{E}[Y_i(\mathbf{a})]$.

To assess the uncertainty of the deconfounder, we consider the uncertainty from both sources. We first draw s samples $\{z_i^{(1)}, \dots, z_i^{(s)}\}$ of the substitute confounder: $z_i^{(\ell)} \stackrel{\text{iid}}{\sim} p(z_i | \mathbf{a}_i)$, $\ell = 1, \dots, s$. 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 $\{z_i^{(1)}, \dots, z_i^{(s)}\}$; 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. In a two-cause smoking study, Section 6.1 illustrates this strategy for calculating the uncertainty of the deconfounder.

6. Empirical Studies

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 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 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.

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)** (Ranganath, Gerrish, and Blei 2014; Kucukelbir et al. 2017). We use its RStan implementation (Carpenter et al. 2017) in Section 6.1 and its Edward implementation (Tran et al. 2016a, 2017) in Sections 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 (US Department of Health and Human Services Public Health Service 1987; Imai and Van Dyk 2004). The dataset contains 9708 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.

6.1.1. 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,

$$y_i = \beta_{\text{mar}} a_{\text{mar},i} + \beta_{\text{exp}} a_{\text{exp},i} + \beta_{\text{age}} a_{\text{age},i} + \varepsilon_i, \quad (23)$$

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

$$\beta_{\text{mar}} \sim \mathcal{N}(0, 1) \quad \beta_{\text{exp}} \sim \mathcal{N}(0, 1) \quad \beta_{\text{age}} \sim \mathcal{N}(0, 1). \quad (24)$$

and from these coefficients we generate the outcome for each individual. The result is a dataset of 9708 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.

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.

6.1.2. 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.) We consider two factor models.

The first is a linear factor model,

$$z_{\text{line},i} \sim \mathcal{N}(0, \sigma^2) \quad (25)$$

$$a_{\text{mar},i} = \eta_{\text{mar}}^{(1)} z_{\text{line},i} + \eta_{\text{mar}}^{(0)} + \varepsilon_{i,\text{mar}} \quad (26)$$

$$a_{\text{exp},i} = \eta_{\text{exp}}^{(1)} z_{\text{line},i} + \eta_{\text{exp}}^{(0)} + \varepsilon_{i,\text{exp}}, \quad (27)$$

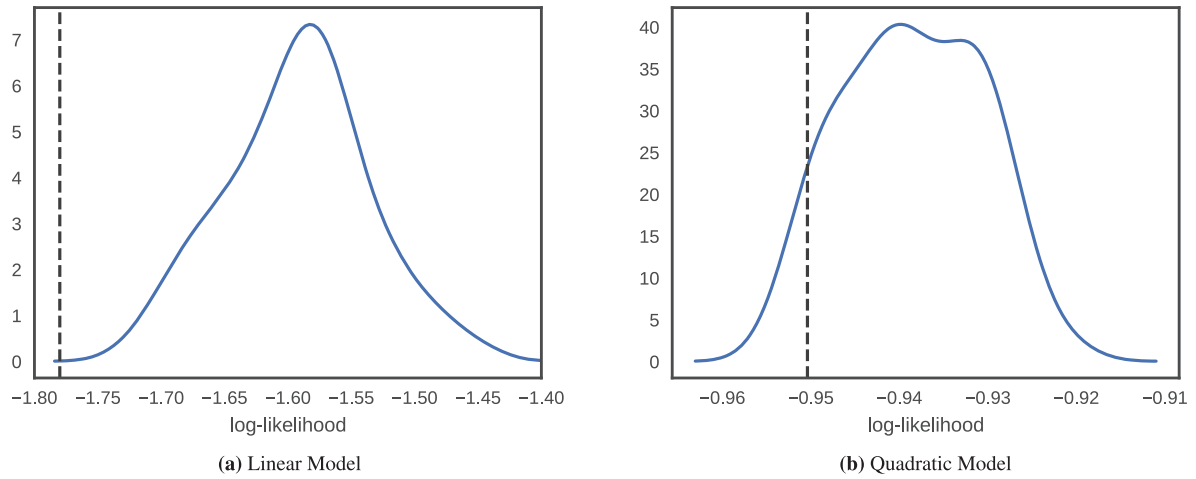


Figure 3. Predictive checks for the substitute confounder z obtained from a linear factor model (a) and a quadratic factor model (b). The blue line is the *kde* of the test-statistic based on the predictive distribution. The dashed vertical line shows the value of the test-statistic on the observed dataset. 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.

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 4.1, 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 3(a).

We next consider a quadratic factor model,

$$z_{\text{quad},i} \sim \mathcal{N}(0, \sigma^2) \quad (28)$$

$$a_{\text{mar},i} = \eta_{\text{mar}}^{(1)} z_{\text{quad},i} + \eta_{\text{mar}}^{(2)} z_{\text{quad},i}^2 + \eta_{\text{mar}}^{(0)} + \varepsilon_{i,\text{mar}} \quad (29)$$

$$a_{\text{exp},i} = \eta_{\text{exp}}^{(1)} z_{\text{quad},i} + \eta_{\text{exp}}^{(2)} z_{\text{quad},i}^2 + \eta_{\text{exp}}^{(0)} + \varepsilon_{i,\text{exp}}, \quad (30)$$

where all errors are standard normal. (In Appendix C in the supplementary materials, we prove that the average causal effect is identifiable with this quadratic factor model and a linear outcome model.) We again use variational inference and a predictive check. The resulting predictive score is 0.12, Figure 3(b). This value gives the green light. We use the model's posterior estimates $\hat{z}_i \sim p_{\text{quad}}(z | a_i)$ to form a substitute confounder in a causal inference.

6.1.3. Deconfounded Causal Inference

Using a factor model to estimate substitute confounders, we proceed with causal inference. We set the outcome model of $\mathbb{E}[Y(A_{\text{mar}}, A_{\text{exp}}) | A, Z]$ 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, for example, for the quadratic model and for age,

$$a_{\text{mar},i}(\hat{z}_i) = \mathbb{E}_{\text{quad}}[A_{\text{mar}} | Z = \hat{z}_i]. \quad (31)$$

See Equation (21).

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 (23). We emphasize that Equation (23) is the true mechanism of the simulated world, which the deconfounder does not have access to. The linear model we posit for $\mathbb{E}[Y(A_{\text{mar}}, A_{\text{exp}}) | A, Z]$ is a functional form for the expectation we are trying to estimate.

6.1.4. 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 3 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 3 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 5:

- *What if multiple factor models pass the check?* We fit to the causes one-dimensional, two-dimensional, and three-dimensional quadratic factor models. All three models pass the check. Table 3 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.

Table 3. Total bias and variance of the estimated causal coefficients β_{exp} and β_{mar} .

	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
<i>Deconfounder</i>				
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
<i>Deconfounder with covariates</i>				
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

NOTE: (“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. The bold values are the smallest values in each block.

- *Should we additionally condition on the observed covariates?* Table 3 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 demonstrates how including covariates trades variance for the risk of missing a confounder.

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.

6.2. Many Causes: Genome-Wide Association Studies

Analyzing gene-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, that is, how the genetic codes of an individuals exhibits 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 Appendix A in the supplementary materials.)

6.2.1. Simulated GWAS Data and the Causal Inference Problem

We put the GWAS problem into our notation. The data are tuples (\mathbf{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 $y_i(\mathbf{a})$, the trait of interest as a function of a specific genotype.

We generate synthetic GWAS data. Following Song, Hao, and Storey (2015), we simulate genotypes $\mathbf{a}_{1:n}$ 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, Hao, and Storey (2015), we generate the outcome (i.e., the trait) from a linear model,

$$y_i = \sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i. \quad (32)$$

To introduce further confounding effects, we group the individuals by their SNPs; the i th individual is in group c_i . (Appendix N in the supplementary materials 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 (33), 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 N in the supplementary materials details the full configurations of the simulation.

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

$$y_i \sim \text{Bernoulli} \left(\frac{1}{1 + \exp(\sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i)} \right). \quad (33)$$

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

For each true assignment model of \mathbf{a}_i , we simulate 100 datasets of genotypes \mathbf{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 (\mathbf{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.

6.2.2. Deconfounding GWAS

We apply the deconfounder with five assignment models discussed in Section 2.2: 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 4.1.

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 $\mathbf{a}(\hat{\mathbf{z}})$; see Equation (21). Each assignment model gives a different form of $\mathbf{a}(\hat{\mathbf{z}})$. For the binary traits, we use a logistic regression, again conditional on the SNPs and reconstructed causes.

6.2.3. Performance

We study the deconfounder for GWAS. Tables 6–15 in the supplementary materials 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 squared error (RMSE) of the estimated causal coefficients (for real-valued and binary-valued outcomes). Tables 6–15 in the supplementary materials 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 Appendix A in the supplementary mate-

rials.) Notably, the deconfounder often outperforms the regression where we include the (unobserved) confounder as a covariate under the low SNR setting; see Tables 11–14 in the supplementary materials.

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 15 in the supplementary materials). 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 4.1. 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 10 and 15 in the supplementary materials. The spatial model produces spatially correlated individuals; its parameter τ controls the spatial dispersion. (Consider each individual to sit in a unit square; as $\tau \rightarrow 0$, the individuals are placed closer to the corners of the unit square while when $\tau = 1$ they are distributed uniformly.) The 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).

6.2.4. 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 4 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.

6.2.5. Summary

These studies provide three take-away messages: (1) the deconfounder can produce closer-to-truth causal estimates, especially

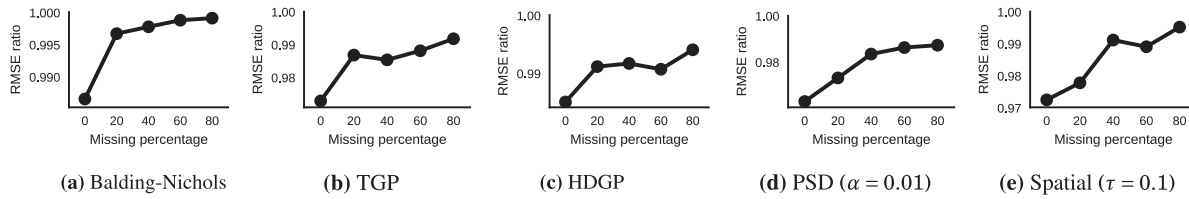


Figure 4. The *rmse* ratio between the deconfounder with *def* and “No control” across simulations when only a subset of causes are unobserved. (Lower ratios means more correction.) As the percentage of observed causes decreases, the “no unobserved single-cause confounders” assumption is compromised; the deconfounder can no longer correct for all latent confounders.

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 Section 1: How much does an actor boost (or hurt) a movie’s revenue? We study the deconfounder with the TMDB 5000 Movie Dataset.⁷ It contains 901 actors (who appeared in at least five movies) and the revenue for the 2828 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 dataset. 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 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 (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 shows a snippet of the highest-earning movies in this dataset. The goal is to estimate the distribution of $Y_i(a)$, the (potential) revenue as a function of a movie cast.

6.3.1. 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 families (*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 4.1.

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.

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 (21).

6.3.2. 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 noncausal model can produce incorrect predictions if we intervene on the causes (Peters, Bühlmann, and Meinshausen 2016). This idea of invariance has also been discussed in Haavelmo (1944), Aldrich (1989), Lanes (1988), Pearl (2009), Schölkopf et al. (2012), and Dawid and Didelez (2010) 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 17 in the supplementary materials 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 (pretreatment) covariates.

In another test set, we hold out 10% of movies from uncommon genres, that is, those that are not comedies, action, or dramas. Table 18 in the supplementary materials 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 16 in the supplementary materials shows the average predictive log-likelihood of the deconfounder and the purely predictive method. The deconfounder predicts slightly worse than the purely predictive method.

6.3.3. 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.⁸

First we examine how the coefficients of individual actors differ between a noncausal model and a deconfounded model.

⁸This 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 authors of this article are statisticians, not film critics.

⁷<https://www.kaggle.com/tmdb>.

(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 Schwarzenegger, 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 Banderas.

Also consider the actors whose estimated contribution improves the most from the noncausal 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.

7. Theory

We develop theoretical results around the deconfounder. (All proofs are in the Appendix.)

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,” imply unconfoundedness. We next establish theoretical properties of the substitute confounder: it captures all multiple-cause confounders and it does not capture any mediators. These results imply that if the factor model captures the distribution of the assigned causes then the substitute confounder renders the assignment ignorable. Moreover, such a factor model always exists.

We then discuss identification results around the deconfounder. Under **stable unit treatment value assumption (SUTVA)** and “no unobserved single-cause confounders,” we prove that the deconfounder identifies the average causal effects and the conditional potential outcomes under different conditions.

7.1. Factor Models and the Substitute Confounder

To study the deconfounder, we first connect unconfoundedness to factor models. Recall the definitions of unconfoundedness and factor model.

Unconfoundedness assumes that the assigned causes are conditionally independent of the potential outcomes (Rosenbaum and Rubin 1983; Imbens 2000):

Definition 1 (Weak unconfoundedness (Imbens 2000)). The assigned causes are weakly unconfounded given Z_i if

$$(A_{i1}, \dots, A_{im}) \perp\!\!\!\perp Y_i(\mathbf{a}) \mid Z_i \quad (34)$$

for all $(a_1, \dots, a_m) \in \mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_m$, and $i = 1, \dots, n$.

Roughly, the assigned causes are weakly unconfounded given Z_i if all confounders are captured by Z_i . More technically, the assigned causes are weakly unconfounded if all confounders are measurable with respect to the σ -algebra generated by Z_i .

A factor model of assigned causes describes each assigned cause of a individual with a latent variable specific to this individual and another specific to this cause:

Definition 2 (Factor model of assigned causes). Consider the assigned causes $\mathbf{A}_{1:n}$, a set of latent variables $\mathbf{Z}_{1:n}$ and a set of parameters $\theta_{1:m}$. A factor model of the assigned causes is a latent-variable model,

$$p(\mathbf{z}_{1:n}, \mathbf{a}_{1:n}; \theta_{1:m}) = p(\mathbf{z}_{1:n}) \prod_{i=1}^n \prod_{j=1}^m p(a_{ij} \mid z_i, \theta_j). \quad (35)$$

The distribution of assigned causes is the corresponding marginal,

$$p(\mathbf{a}_{1:n}) = \int p(\mathbf{z}_{1:n}, \mathbf{a}_{1:n}; \theta_{1:m}) d\mathbf{z}_{1:n}. \quad (36)$$

In a factor model, each latent variable Z_i of individual i renders its assigned causes $A_{ij}, j = 1, \dots, m$, conditionally independent. Each cause is accompanied with an unknown parameter θ_j . As we mentioned in Section 4.1, many common models from Bayesian statistics and machine learning can be written as factor models. In the deconfounder, we fit factor models to construct substitute confounders, where we infer $\mathbf{Z}_{1:n}$ as a function of $\mathbf{a}_{1:n}$ and check its fidelity against the distribution of the causes $p(\mathbf{a}_{1:n})$ using a predictive check. When a fitted factor model passes the check, it captures $p(\mathbf{a}_{1:n})$ well. In other words, factor models in the deconfounder satisfy Equations (35) and (36) with $p(\mathbf{z}_{1:n}) = \delta_{f_\theta(\mathbf{a}_{1:n})}$ for some function $f_\theta(\cdot)$.

To connect unconfoundedness to factor models, consider an intermediate construct, the “Kallenberg construction.” The Kallenberg construction is inspired by the idea of randomization variables, Uniform[0,1] variables from which we can construct a random variable with an arbitrary distribution (Kallenberg 1997). The Kallenberg construction of assigned causes will bridge the conditional independence statement in Equation (34) with the factor models of the deconfounder.

Definition 3 (Kallenberg construction of assigned causes). Consider a random variable Z_i taking values in \mathcal{Z} . The distribution of assigned causes (A_{i1}, \dots, A_{im}) admits a Kallenberg construction if there exists (deterministic) measurable functions, $f_j: \mathcal{Z} \times [0, 1] \rightarrow \mathcal{A}_j$ and random variables $U_{ij} \in [0, 1]$ ($j = 1, \dots, m$) such that

$$A_{ij} \stackrel{a.s.}{=} f_j(Z_i, U_{ij}); \quad (37)$$

the variables U_{ij} must marginally follow $\text{Uniform}[0,1]$ and jointly satisfy

$$(U_{i1}, \dots, U_{im}) \perp\!\!\!\perp (Z_i, Y_i(a_1, \dots, a_m)) \quad (38)$$

for all $(a_1, \dots, a_m) \in \mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_m$.

Using these definitions, the first lemma relates unconfoundedness to the Kallenberg construction.

Lemma 1 (Kallenberg construction \Leftrightarrow weak unconfoundedness). The assigned causes are weakly unconfounded given a random variable Z_i if and only if the distribution of the assigned causes (A_{i1}, \dots, A_{im}) admits a Kallenberg construction from Z_i .

What Lemma 1 says is that if the distribution of the assigned causes has a Kallenberg construction from a random variable Z_i then Z_i is a valid substitute confounder: it renders the causes unconfounded. Moreover, a valid substitute confounder must always come from a Kallenberg construction.

We next relate the Kallenberg construction to factor models. We show that factor models admit a Kallenberg construction. This fact suggests the deconfounder: if we fit a factor model to capture the distribution of assigned causes then we can use the fitted factor model to construct a substitute confounder. This step relies on a key assumption of the deconfounder, “no unobserved single-cause confounders.”

Definition 4 (No unobserved single-cause confounders). Denote X_i as the observed covariates. There are no unobserved single-cause confounders for the assigned causes A_{i1}, \dots, A_{im} if, for $j = 1, \dots, m$,

1. There exist some random variable V_{ij} such that

$$A_{ij} \perp\!\!\!\perp Y_i(\mathbf{a}) \mid X_i, V_{ij}, \quad (39)$$

$$A_{ij} \perp\!\!\!\perp A_{i,-j} \mid V_{ij}, \quad (40)$$

where $A_{i,-j} = \{A_{i1}, \dots, A_{im}\} \setminus A_{ij}$ is the complete set of m causes excluding the j th cause;

2. There exists no proper subset of the sigma algebra $\sigma(V_{ij})$ satisfies Equation (40).

At a higher level, V_{ij} refers to the multiple-cause confounders that affect the j th cause A_{ij} . Equation (39) then ensures that the observed covariates X_i and the multiple-cause confounders V_{ij} satisfy unconfoundedness. In other words, X_i must contain all single-cause confounders. Equation (40) ensures that V_{ij} indeed induces a dependence between A_{ij} and $A_{i,-j}$. It guarantees that V_{ij} can be recovered by constructing a random variable Z_i that renders all the causes conditionally independent.

This “no unobserved single-cause confounders” assumption differs from the classical weak unconfoundedness assumption (Definition 1) by only requiring marginal independence between individual causes A_{ij} and the potential outcome $Y_i(\mathbf{a})$. In contrast, weak unconfoundedness requires $(A_{i1}, \dots, A_{im}) \perp\!\!\!\perp Y_i(\mathbf{a}) \mid X_i$, that is, the joint independence between the causes (A_{i1}, \dots, A_{im}) and the potential outcome function $Y_i(\mathbf{a})$. Moreover, it involves multiple-cause confounders V_{ij} . We remark that “no unobserved single-cause confounders” reduces to weak unconfoundedness when there is only one cause; both require $\mathbf{A}_i \perp\!\!\!\perp Y_i(\mathbf{a}) \mid X_i$, where \mathbf{A}_i and \mathbf{a} are one-dimensional.

Now we state the connection between the Kallenberg construction and factor models.

Lemma 2 (Factor models \Rightarrow Kallenberg construction). Under weak regularity conditions and “no unobserved single-cause confounders,” every factor model of the assigned causes $p(z_{1:n}, \mathbf{a}_{1:n}; \theta_{1:m})$ admits a Kallenberg construction from Z_i .

Lemmas 1 and 2 connect unconfoundedness to Kallenberg constructions and then Kallenberg constructions to factor models. The two lemmas together connect factor models to unconfoundedness. These connections enable the deconfounder: they explain how the distribution of assigned causes relates to the substitute confounder Z in a Kallenberg construction. They justify why we can take a set of assigned causes and do inference on Z via factor models.

Next we establish two properties of the substitute confounder. We assume the substitute confounder comes from a factor model that captures the population distribution of the causes.

The first property is that the substitute confounder must capture all multiple-cause confounders. It implies that the inferred substitute confounder, together with all single-cause confounders (if there is any), deconfounds causal inference.

Lemma 3. Any multiple-cause confounder C_i must be measurable with respect to the σ -algebra generated by the substitute confounder Z_i .

A multiple-cause confounder is a confounder that confounds two or more causes. (Its technical definition stems from Definition 4 of VanderWeele and Shpitser (2013); see Appendix H in the supplementary materials.) Figure 1 gives the intuition with a graphical model and Appendix H in the supplementary materials gives a detailed proof.

Lemma 3 shows that the deconfounder captures unobserved confounders. But might the inferred substitute confounder pick up a mediator? If the substitute confounder also picks up a mediator then conditioning on it will yield conservative causal estimates (Baron and Kenny 1986; Imai, Keele, and Yamamoto 2010). The next proposition alleviates this concern.

Lemma 4. Any mediator is almost surely not measurable with respect to the σ -algebra generated by the substitute confounder Z_i and the pretreatment observed covariates X_i .

Lemma 4 implies that the substitute confounder does not pick up mediators, variables along the path between causes and effects. This property greenlights us for treating the inferred substitute confounder as a pretreatment covariate.

Lemmas 3 and 4 qualify the substitute confounder for mimicking confounders. We condition on the substitute confounder and proceed with causal inference.

These lemmas lead to justifications of the deconfounder algorithm. We first describe their implications on the substitute confounders and factor models.

Proposition 5 (Substitute confounders and factor models). Under weak regularity conditions,

1. Under “no unobserved single-cause confounders,” the assigned causes are weakly unconfounded given the substitute confounder Z_i and the pretreatment covariates X_i if the true distribution $p(\mathbf{a}_{1:m})$ can be written as a factor model that uses the substitute confounder, $p(z_{1:m}, \mathbf{a}_{1:m}; \theta_{1:m})$.
2. There always exists a factor model that captures the distribution of assigned causes.

Proof sketch. The first part follows from [Lemmas 1 and 2](#). The second part follows from the Reichenbach’s common cause principle ([Reichenbach 1956](#); [Sober 1976](#); [Peters, Janzing, and Schölkopf 2017](#)) and Sklar’s theorem ([Sklar 1959](#)): any multivariate joint distribution can be factorized into the product of univariate marginal distributions and a copula which describes the dependence structure between the variables. The full proof is in Appendix G in the supplementary materials. \square

Proposition 5 justifies the use of factor models in the deconfounder. The first part of [Proposition 5](#) suggests how to find a valid substitute confounder, one that renders the causes weakly unconfounded. Two conditions suffice: (1) the substitute confounder comes from a factor model; (2) the factor model captures the population distribution of the assigned causes. The assignment model in the deconfounder stems from this result: fit a factor model to the assigned causes, check that it captures their population distribution, and finally use the fitted factor model to infer a substitute confounder. The first part of the theorem says that the deconfounder does deconfound. The second part ensures that there is hope to find a deconfounding factor model. There always exists a factor model that captures the population distribution of the assigned causes.

7.2. Causal Identification of the Deconfounder

Building on the characterizations of the substitute confounder ([Lemmas 1–4](#)), we discuss a collection of causal identification results around the deconfounder. We prove that the deconfounder can identify three causal quantities under suitable conditions.⁹ 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.

Before stating the identification results, we first describe the notion of a *consistent* substitute confounder; we will rely on this notion for identification.

Definition 5 (Consistency of substitute confounders). The factor model $p(\theta, z, \mathbf{a})$ admits consistent estimates of the substitute confounder Z_i if, for some function f_θ ,

$$p(z_i | \mathbf{a}_i, \theta) = \delta_{f_\theta(\mathbf{a}_i)}. \quad (41)$$

Consistency of substitute confounders requires that we can estimate the substitute confounder Z_i from the causes \mathbf{A}_i with certainty; it is a deterministic function of the causes.¹⁰ Nevertheless, the substitute confounder need not coincide with the

true data-generating Z_i ; nor does 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).

Many factor models admit consistent substitute confounder estimates when the number of causes is large. For example, probabilistic PCA and Poisson factorization lead to consistent Z_i as $(n + m) \cdot \log(nm)/(nm) \rightarrow 0$, where n is the number of individuals and m is the number of causes ([Chen, Li, and Zhang 2017](#)). Many studies also involve many causes, for example, the [genome-wide association studies \(GWAS\)](#) study in [Section 6.2](#) and the movie-actor study in [Section 6.3](#).

We now describe three identification results under the [SUTVA](#), “no unobserved single-cause confounders,” and consistency of substitute confounders. We first study the average causal effect of all the causes.

Theorem 6 (Identification of the average causal effect of all the causes). Assume [SUTVA](#), “no unobserved single-cause confounders,” and consistency of substitute confounders. Then, under conditions described below, the deconfounder nonparametrically identifies the average causal effect of all the causes. The average causal effect of changing the causes from $\mathbf{a} = (a_1, \dots, a_m)$ to $\mathbf{a}' = (a'_1, \dots, a'_m)$ is

$$\begin{aligned} \mathbb{E}_Y [Y_i(\mathbf{a})] - \mathbb{E}_Y [Y_i(\mathbf{a}')] &= \mathbb{E}_{Z, X} [\mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}, Z_i, X_i]] \\ &\quad - \mathbb{E}_{Z, X} [\mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}', Z_i, X_i]]. \end{aligned} \quad (42)$$

This holds with the following two conditions¹¹: (1) the substitute confounder is a piece-wise constant function of the (continuous) causes: $\nabla_{\mathbf{a}} f_\theta(\mathbf{a}) = 0$ up to a set of Lebesgue measure zero; (2) the outcome is separable,

$$\begin{aligned} \mathbb{E} [Y_i(\mathbf{a}) | Z_i = z, X_i = x] &= f_1(\mathbf{a}, x) + f_2(z), \\ \mathbb{E} [Y_i | \mathbf{A}_i = \mathbf{a}, Z_i = z, X_i = x] &= f_3(\mathbf{a}, x) + f_4(z), \end{aligned}$$

for all $(\mathbf{a}, x, z) \in \mathcal{A} \times \mathcal{X} \times \mathcal{Z}$ and some continuously differentiable¹² functions f_1, f_2, f_3 , and f_4 .¹³

Proof sketch. [Theorem 6](#) relies on two results: (1) “No unobserved single-cause confounders” and [Lemma 3](#) ensure (Z_i, X_i) capture all confounders; (2) The pretreatment nature of X_i and [Lemma 4](#) ensure (Z_i, X_i) capture no mediators. These results assert unconfoundedness given the substitute confounder Z_i and the observed covariates X_i . They greenlight us for causal inference given consistency of substitute confounder estimates. [Theorem 6](#) then leverages two additional conditions to identify average causal effects without assuming overlap. The full proof is in Appendix K in the supplementary materials. \square

¹¹We 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 multiple-cause confounders. The former is easier to check; it also implies the latter because of [Lemma 3](#).

¹²For binary causes, we can analogously assume that there exists \mathbf{a}_{new} and \mathbf{a}'_{new} such that $\mathbf{a}_{\text{new}} - \mathbf{a}'_{\text{new}} = \mathbf{a} - \mathbf{a}'$ and they lead to the same substitute confounder estimate $f(\mathbf{a}_{\text{new}}) = f(\mathbf{a}'_{\text{new}})$. Further, the outcome model is separable: $\mathbb{E} [Y_i(\mathbf{a}) - Y_i(\mathbf{a}') | Z_i = z, X_i = x] = f_1(\mathbf{a} - \mathbf{a}', x) + f_2(z)$.

¹³The expectation over Z_i and X_i is taken over $P(Z_i, X_i)$ in Equation (42): $\mathbb{E}_{Z_i, X_i} [\mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}, Z_i, X_i]] = \int \mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}, Z_i, X_i] P(Z_i, X_i) dZ_i dX_i$.

⁹Here “identify” means the causal quantity can be written as a function of the observed data. Moreover, the deconfounder can unbiasedly estimate it.

¹⁰Together with [Lemma 3](#), consistency of substitute confounders implies that the true unobserved multiple-cause confounders are also deterministic functions of the causes.

Theorem 6 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 consistency of substitute confounders. 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 6** 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 $Z_i = f_\theta(A_i)$, which is required by the consistency of substitute confounders? The reason is that f_θ 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 $\mathbb{E}[Y_i(\mathbf{a}) | Z_i = z, X_i = x]$ can be separated into two components, one differentiable $f_1(\mathbf{a}, x)$ and one non-differentiable $f_2(z)$. A similar argument also holds for $\mathbb{E}[Y_i | A_i = \mathbf{a}, Z_i = z, X_i = x]$. It is this incongruence between X_i and Z_i in differentiability that leads to identification.

When the separability condition of **Theorem 6** does not hold, we can still use the deconfounder to handle the unobserved multiple-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 7 (Identification of the average causal effect of subsets of the causes). Assume **SUTVA**, “no unobserved single-cause confounders,” and consistency of substitute confounders. Then, under the condition described below, the deconfounder nonparametrically identifies the average causal effect of subsets of causes. The average causal effect of changing the first k ($k < m$) causes from $a_{1:k} = (a_1, \dots, a_k)$ to $a'_{1:k} = (a'_1, \dots, a'_k)$ is

$$\begin{aligned} & \mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_{1:k}, A_{i,(k+1):m})]] \\ & - \mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a'_{1:k}, A_{i,(k+1):m})]] \\ & = \mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i,1:k} = a_{1:k}]] \\ & - \mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i,1:k} = a'_{1:k}]]. \end{aligned} \quad (43)$$

This holds with the following condition: The first k causes A_{i1}, \dots, A_{ik} satisfy overlap, $P((A_{i1}, \dots, A_{ik}) \in \mathcal{A} | Z_i, X_i) > 0$ for any set \mathcal{A} such that $P(\mathcal{A}) > 0$.¹⁴

Proof sketch. Similar to **Theorem 6**, **Theorem 7** uses Lemmas 3 and 4 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 2015). The full proof is in Appendix L in the supplementary materials. \square

Theorem 7 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 “no unobserved single-cause confounders,” **SUTVA**, and consistency of substitute confounders, **Theorem 7** requires overlap. Overlap ensures that $\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i,1:k} = a_{1:k}]$ 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 7** replaces the separability assumption about the outcome model required by **Theorem 6**.

We note that the overlap condition and the consistency of substitute confounders are compatible. Though consistency requires $P(Z_i | A_i) = \delta_{f_\theta(A_i)}$, it is still possible for subsets of the causes to satisfy overlap; the consistency condition only prevents the complete set of m causes from satisfying overlap. For example, consider a consistent estimate of the 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 8 (Identification of the conditional mean potential outcome). Assume **SUTVA**, “no unobserved single-cause confounders,” and consistency of substitute confounders. Then, under the condition described below, the deconfounder nonparametrically identifies the mean potential outcome of an individual given its current assigned causes. If an individual is assigned with $\mathbf{a} = (a_1, \dots, a_m)$, then its potential outcome under a different assignment $\mathbf{a}' = (a'_1, \dots, a'_m)$ is

$$E_Y[Y_i(\mathbf{a}) | A_i = \mathbf{a}] = E_{Z,X}[E_Y[Y_i | Z_i, X_i, A_i = \mathbf{a}']].$$

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

Proof sketch. As with Theorems 6 and 7, **Theorem 8** relies on the unconfoundedness given the substitute confounders Z_i and the observed covariates X_i due to Lemmas 3 and 4. 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 M in the supplementary materials. \square

¹⁴In full notation, $\mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_{1:k}, A_{i,(k+1):m})]] = \mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_1, \dots, a_k, A_{ik+1}, \dots, A_{im})]]$.

Given consistency of substitute confounders, [Theorem 8](#) nonparametrically identifies the mean potential outcome of an individual $Y_i(\mathbf{a}')$ given its current assigned causes $\mathbf{A}_i = \mathbf{a}$. The only requirement is about the configurations of cause assignments we can query, \mathbf{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 movie; 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 cannot query what if we make its cast to be that of “The Shawshank Redemption” (which is not a spy film).

Theorems 6–8 confirm the validity of the deconfounder by providing three sets of nonparametric identification results. When the assumptions in Theorems 6–8 may not hold, we recommend evaluating the uncertainty of the deconfounder estimate. [Section 5](#) 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. 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 will be uniform over \mathcal{R} (with noninformative priors).

We finally remark that the identification results in Theorems 6–8 do not contradict the negative results of D’Amour (2019). D’Amour (2019) explore nonparametric non-identification of a particular multi-causal quantity, the mean potential outcome $\mathbb{E}[Y_i(\mathbf{a})]$. In this article, Theorems 6–8 establish the nonparametric identification of different causal quantities. D’Amour (2019) do not make the same assumptions as in Theorems 6–8. More specifically, under consistency of substitute confounders and other suitable conditions, [Theorem 6](#) shows that the average causal effect of all the causes $\mathbb{E}[Y_i(\mathbf{a})] - \mathbb{E}[Y_i(\mathbf{a}')] is nonparametrically identifiable; [Theorem 7](#) shows that the average causal effect of subsets of the causes $\mathbb{E}_{A_{(k+1):m}}[\mathbb{E}_Y[Y_i(a_{1:k}, A_{i,(k+1):m})]] - \mathbb{E}_{A_{(k+1):m}}[\mathbb{E}_Y[Y_i(a'_{1:k}, A_{i,(k+1):m})]]$ is nonparametrically identifiable; [Theorem 8](#) shows that the conditional mean potential outcome $\mathbb{E}[Y_i(\mathbf{a}') | \mathbf{A}_i = \mathbf{a}]$ is nonparametrically identifiable.$

8. Discussion

Classical causal inference studies how a univariate cause affects an outcome. Here we studied *multiple causal inference*, where there are multiple causes that contribute to the effect. Multiple causes might at first appear to be a curse, but we showed that it can be a blessing.

We developed the *deconfounder*: first fit a good factor model of assigned causes; 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 confounders, and we demonstrated that whether a factor model is satisfactory is a checkable proposition.

There are many directions for future work.

We estimated the potential outcomes under configurations of the causes. Which potential outcomes can be reliably estimated? Can we trade off confounding bias and estimation variance?

We checked factor models for downstream causal unbiasedness. But model checking is an imprecise science. Can we develop rigorous model checking algorithms for causal inference?

We focused on estimation. Can we develop a testing counterpart? How can we identify significant causes while still preserving family-wise error rate or false discovery rate?

We analyzed univariate outcomes. Can we work with both multiple causes and multiple outcomes. Can dependence among outcomes further help causal inference?

Supplementary Materials

The supplementary materials contain further discussions of the deconfounder algorithm, detailed results of the empirical studies, and proofs of the theoretical results.

Acknowledgments

We have had many useful discussions about the previous versions of this article. We thank Edo Airoldi, Elias Barenboim, Léon Bottou, Alexander D’Amour, Barbara Engelhart, Andrew Gelman, David Heckerman, Jennifer Hill, Ferenc Huszár, George Hripcsak, Daniel Hsu, Guido Imbens, Thorsten Joachims, Fan Li, Lydia Liu, Jackson Loper, David Madigan, Joris Mooij, Suresh Naidu, Xinkun Nie, Elizabeth Ogburn, Georgia Papadogeorgou, Judea Pearl, Alex Peysakhovich, Rajesh Ranganath, Jason Roy, Cosma Shalizi, Dylan Small, Hal Stern, Amos Storkey, Wesley Tansey, Eric Tchetgen Tchetgen, Dustin Tran, Victor Veitch, Linbo Wang, Stefan Wager, Kilian Weinberger, Jeannette Wing, Linying Zhang, Qingyuan Zhao, and José Zubizarreta.

Funding

This work is supported by ONR N00014-17-1-2131, ONR N00014-15-1-2209, NIH 1U01MH115727-01, NSF CCF-1740833, DARPA SD2 FA8750-18-C-0130, IBM, 2Sigma, Amazon, NVIDIA, and Simons Foundation.

References

- Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), “Mixed Membership Stochastic Blockmodels,” *Journal of Machine Learning Research*, 9, 1981–2014. [1577]
- Aldrich, J. (1989), “Autonomy,” *Oxford Economic Papers*, 41, 15–34. [1588]
- Astle, W., and Balding, D. J. (2009), “Population Structure and Cryptic Relatedness in Genetic Association Studies,” *Statistical Science*, 24, 451–471. [1576]
- Baron, R. M., and Kenny, D. A. (1986), “The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations,” *Journal of Personality and Social Psychology*, 51, 1173. [1590]
- Bayarri, M., and Castellanos, M. (2007), “Bayesian Checking of the Second Levels of Hierarchical Models,” *Statistical Science*, 22, 322–343. [1581]

- Blei, D. M. (2014), "Build, Compute, Critique, Repeat: Data Analysis With Latent Variable Models," *Annual Review of Statistics and Its Application*, 1, 203–232. [1581]
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent (D)irichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1577]
- 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, 1–32. [1584]
- Cemgil, A. T. (2009), "Bayesian Inference for Nonnegative Matrix Factorization Models," *Computational Intelligence and Neuroscience*, 2009, 785152. [1580]
- Chen, Y., Li, X., and Zhang, S. (2017), "Structured Latent Factor Analysis for Large-Scale Data: Identifiability, Estimability, and Their Implications," arXiv no. 1712.08966. [1591]
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [1576]
- 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, 51. [1575]
- 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*, pp. 1252–1261. [1576,1579]
- 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*, pp. 617–624. [1580]
- 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, 187–199. [1579]
- D'Amour, A. (2019), "On Multi-Cause Causal Inference With Unobserved Confounding: Counterexamples, Impossibility, and Alternatives," arXiv no. 1902.10286. [1593]
- D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2017), "Overlap in Observational Studies With High-Dimensional Covariates," arXiv no. 1711.02582. [1579]
- Dawid, A. P., and Didelez, V. (2010), "Identifying the Consequences of Dynamic Treatment Strategies: A Decision-Theoretic Overview," *Statistics Surveys*, 4, 184–231. [1588]
- Dehejia, R. H., and Wahba, S. (2002), "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics*, 84, 151–161. [1581]
- Dey, D., Gelfand, A., Swartz, T., and Vlachos, P. (1998), "Simulation Based Model Checking for Hierarchical Models," *Test*, 7, 325–346. [1581]
- Erosheva, E. A. (2003), "Bayesian Estimation of the Grade of Membership Model," *Bayesian Statistics*, 7, 501–510. [1577]
- 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, 681–697. [1576]
- Franks, A., D'Amour, A., and Feller, A. (2019), "Flexible Sensitivity Analysis for Observational Studies Without Observable Implications," *Journal of the American Statistical Association*, 1–33. [1576,1579]
- 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*, 81, 459–487. [1576]
- 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. [1581]
- 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. [1581]
- Gelman, A., Meng, X., and Stern, H. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," *Statistica Sinica*, 6, 733–807. [1581]
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Boca Raton, FL: Chapman and Hall/CRC. [1581]
- 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, 531–541. [1576,1579]
- Gopalan, P., Hofman, J. M., and Blei, D. M. (2015), "Scalable Recommendation With Hierarchical Poisson Factorization," in *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, AUA Press, pp. 326–335. [1580]
- 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. [1576]
- Haavelmo, T. (1944), "The Probability Approach in Econometrics," *Econometrica: Journal of the Econometric Society*, 12, iii–115. [1588]
- Hao, W., Song, M., and Storey, J. D. (2015), "Probabilistic Models of Genetic Variation in Structured Populations Applied to Global Human Studies," *Bioinformatics*, 32, 713–721. [1576]
- Heckerman, D. (2018), "Accounting for Hidden Common Causes When Inferring Cause and Effect From Observational Data," arXiv no. 1801.00727. [1576]
- Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998), "Characterizing Selection Bias Using Experimental Data," Technical Report, National Bureau of Economic Research. [1581]
- Hill, J. L. (2011), "Bayesian Nonparametric Modeling for Causal Inference," *Journal of Computational and Graphical Statistics*, 20, 217–240. [1581]
- Hirano, K., and Imbens, G. W. (2004), "The Propensity Score With Continuous Treatments," in *Applied Bayesian Modeling and Causal Inference From Incomplete-Data Perspectives*, eds. A. Gelman and X.-L. Meng, New York: Wiley, pp. 73–84. [1577,1578,1581]
- Holland, P. (1986), "Statistics and Causal Inference," *Journal of the American Statistical Association*, 81, 945–960. [1577]
- 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, 663–685. [1581]
- Imai, K., Keele, L., and Yamamoto, T. (2010), "Identification, Inference and Sensitivity Analysis for Causal Mediation Effects," *Statistical Science*, 25, 51–71. [1590]
- 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, 854–866. [1576,1577,1578,1579,1584]
- Imbens, G. W. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710. [1577,1578,1581,1589]
- Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*, New York: Cambridge University Press. [1574,1576,1578,1592]
- Janzing, D., and Schölkopf, B. (2018a), "Detecting Confounding in Multivariate Linear Models via Spectral Analysis," *Journal of Causal Inference*, 6, 1–27. [1576]
- (2018b), "Detecting Non-Causal Artifacts in Multivariate Linear Regression Models," arXiv no. 1803.00810. [1576]
- Kallenberg, O. (1997), *Foundations of Modern Probability*, Collection: Probability and Its Applications, New York: Springer. [1589]
- Kaltenpoth, D., and Vreeken, J. (2019), "We Are Not Your Real Parents: Telling Causal From Confounded Using MDL," arXiv no. 1901.06950. [1576]
- Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., and Eskin, E. (2010), "Variance Component Model to Account for Sample Structure in Genome-Wide Association Studies," *Nature Genetics*, 42, 348. [1576]
- Kingma, D. P., and Welling, M. (2013), "Auto-Encoding Variational Bayes," arXiv no. 1312.6114. [1577]
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017), "Automatic Differentiation Variational Inference," *The Journal of Machine Learning Research*, 18, 430–474. [1584]
- Laird, N. M., and Louis, T. A. (1982), "Approximate Posterior Distributions for Incomplete Data Problems," *Journal of the Royal Statistical Society, Series B*, 44, 190–200. [1581]
- Lanes, S. F. (1988), "The Logic of Causal Inference," in *Causal Inference*, ed. K. J. Rothman, Boston: ERI, pp. 59–75. [1588]
- Lechner, M. (2001), "Identification and Estimation of Causal Effects of Multiple Treatments Under the Conditional Independence Assumption," in

- Econometric Evaluation of Labor Market Policies*, eds. M. Lechner and F. Pfeiffer, Heidelberg: Springer, pp. 43–58. [1576]
- Lee, B. K., Lessler, J., and Stuart, E. A. (2010), “Improving Propensity Score Weighting Using Machine Learning,” *Statistics in Medicine*, 29, 337–346. [1576]
- Lee, D. D., and Seung, H. S. (1999), “Learning the Parts of Objects by Non-Negative Matrix Factorization,” *Nature*, 401, 788. [1580]
- (2001), “Algorithms for Non-Negative Matrix Factorization,” in *Advances in Neural Information Processing Systems*, pp. 556–562. [1580]
- Liu, F., and Chan, L. (2018), “Confounder Detection in High Dimensional Linear Models Using First Moments of Spectral Measures,” arXiv no. 1803.06852. [1576]
- Lopez, M. J., and Gutman, R. (2017), “Estimation of Causal Effects With Multiple Treatments: A Review and New Ideas,” *Statistical Science*, 32, 432–454. [1576]
- 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*, pp. 6449–6459. [1575]
- 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, 2937–2960. [1581]
- 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, 3388–3414. [1576]
- 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, 403. [1576]
- McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, 37. [1580]
- 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*, pp. 28–36. [1575]
- McLachlan, G. J., and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering* (Vol. 84), New York: Marcel Dekker. [1577]
- 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, 1–24. [1575]
- Mohamed, S., Ghahramani, Z., and Heller, K. A. (2009), “Bayesian Exponential Family PCA,” in *Advances in Neural Information Processing Systems*, pp. 1089–1096. [1580]
- Mohamed, S., and Lakshminarayanan, B. (2016), “Learning in Implicit Generative Models,” arXiv no. 1610.03483. [1577]
- 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*, pp. 1687–1695. [1575]
- Morgan, S., and Winship, C. (2015), *Counterfactuals and Causal Inference* (2nd ed.), New York: Cambridge University Press. [1575]
- Neal, R. M. (1990), “Learning Stochastic Feedforward Networks,” *Department of Computer Science, University of Toronto*, 64, 1283. [1577,1580]
- Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, San Francisco, CA: Morgan Kaufmann Publishers Inc. [1578]
- (2009), *Causality* (2nd ed.), New York: Cambridge University Press. [1576,1588]
- 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*, 78, 947–1012. [1588]
- Peters, J., Janzing, D., and Schölkopf, B. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*, Cambridge, MA: MIT Press. [1591]
- 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, 904. [1576]
- Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000), “Association Mapping in Structured Populations,” *The American Journal of Human Genetics*, 67, 170–181. [1576,1577,1582]
- Ranganath, R., and Blei, D. M. (2019), “Population Predictive Checks,” arXiv no. 1908.00882. [1581]
- Ranganath, R., Gerrish, S., and Blei, D. (2014), “Black Box Variational Inference,” in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, eds. Samuel Kaski and Jukka Corander, Reykjavik, Iceland: PMLR, 33, pp. 814–822. [1580,1584]
- Ranganath, R., and Perotte, A. (2018), “Multiple Causal Inference With Latent Confounding,” arXiv no. 1805.08273. [1576]
- Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015), “Deep Exponential Families,” in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, eds. Lebanon, G. and S. V. N. Vishwanathan, San Diego, California, USA: PMLR, 38, pp. 762–771. [1577,1580]
- Ranganath, R., Tran, D., and Blei, D. (2016), “Hierarchical Variational Models,” in *International Conference on Machine Learning*, pp. 324–333. [1577]
- 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, 675–683. [1576]
- Reichenbach, H. (1956), *The Direction of Time*, ed. M. Reichenbach, Berkeley, CA: University of California Press. [1591]
- Rezende, D. J., and Mohamed, S. (2015), “Variational Inference With Normalizing Flows,” arXiv no. 1505.05770. [1577]
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014), “Stochastic Back-propagation and Variational Inference in Deep Latent Gaussian Models,” in *International Conference on Machine Learning* (Vol. 2). [1580]
- Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), “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*, eds. M. E. Halloran and D. Berry, New York: Springer, pp. 1–94. [1576,1579]
- Rosenbaum, P. R., and Rubin, D. B. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects,” *Biometrika*, 70, 41–55. [1576,1577,1581,1589]
- Rubin, D. B. (1974), “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies,” *Journal of Educational Psychology*, 66, 688. [1574,1576]
- (1980), “Randomization Analysis of Experimental Data: The Fisher Randomization Test Comment,” *Journal of the American Statistical Association*, 75, 591–593. [1577,1578]
- (1984), “Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician,” *The Annals of Statistics*, 12, 1151–1172. [1581]
- (1990), “Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies,” *Statistical Science*, 5, 472–480. [1577,1578]
- (2005), “Causal Inference Using Potential Outcomes: Design, Modeling, Decisions,” *Journal of the American Statistical Association*, 100, 322–331. [1574,1576]
- 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*, Springer, pp. 540–547. [1580]
- 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, 512. [1576]
- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012), “On Causal and Anticausal Learning,” arXiv no. 1206.6471. [1588]
- Shah, R. D., and Meinshausen, N. (2018), “Rsvp-Graphs: Fast High-Dimensional Covariance Matrix Estimation Under Latent Confounding,” arXiv no. 1811.01076. [1575]

- Sharma, A., Hofman, J. M., and Watts, D. J. (2016), “Split-Door Criterion for Causal Identification: Automatic Search for Natural Experiments,” arXiv no. 1611.09414. [1576]
- Sklar, M. (1959), “Fonctions de Repartition an Dimensions et Leurs Marges,” *Publications de l’Institut de Statistique de l’Université de Paris*, 8, 229–231. [1591]
- Sober, E. (1976), *Simplicity*. [1591]
- Song, M., Hao, W., and Storey, J. D. (2015), “Testing for Genetic Associations in Arbitrarily Structured Populations,” *Nature Genetics*, 47, 550–554. [1576,1586]
- Stephens, M., and Balding, D. J. (2009), “Bayesian Statistical Methods for Genetic Association Studies,” *Nature Reviews Genetics*, 10, 681. [1575,1582,1586]
- Tierney, L., and Kadane, J. B. (1986), “Accurate Approximations for Posterior Moments and Marginal Densities,” *Journal of the American Statistical Association*, 81, 82–86. [1581]
- Tipping, M. E., and Bishop, C. M. (1999), “Probabilistic Principal Component Analysis,” *Journal of the Royal Statistical Society, Series B*, 61, 611–622. [1575,1577,1580]
- Tran, D., and Blei, D. M. (2017), “Implicit Causal Models for Genome-Wide Association Studies,” arXiv no. 1710.10742. [1576]
- 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*. [1577,1580,1584]
- 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 no. 1610.09787. [1584]
- Tran, D., Ruiz, F. J., Athey, S., and Blei, D. M. (2016b), “Model Criticism for Bayesian Causal Inference,” arXiv no. 1610.09037. [1581]
- US Department of Health and Human Services Public Health Service (1987). National Medical Expenditure Survey Series (NMES). Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2006-03-30. <https://doi.org/10.3886/ICPSR06371.v1> [1584]
- VanderWeele, T. J., and Shpitser, I. (2013), “On the Definition of a Confounder,” *The Annals of Statistics*, 41, 196–220. [1590]
- 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*, 211, 1395–1407. [1592]
- 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, 5–22. [1575,1582,1586]
- Wager, S., and Athey, S. (2018), “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests,” *Journal of the American Statistical Association*, 113, 1228–1242. [1581]
- Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebly, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., and Kresovich, S. (2006), “A Unified Mixed-Model Method for Association Mapping That Accounts for Multiple Levels of Relatedness,” *Nature Genetics*, 38, 203. [1576]
- 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, 59–73. [1576]
- 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*, AUAI Press, pp. 647–655. [1575]



Comment: Reflections on the Deconfounder

Alexander D'Amour

To cite this article: Alexander D'Amour (2019) Comment: Reflections on the Deconfounder, Journal of the American Statistical Association, 114:528, 1597-1601, DOI: [10.1080/01621459.2019.1689138](https://doi.org/10.1080/01621459.2019.1689138)

To link to this article: <https://doi.org/10.1080/01621459.2019.1689138>



Published online: 03 Jan 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Comment: Reflections on the Deconfounder

Alexander D'Amour

Google Research, Cambridge, MA

I would like to congratulate the authors on their illuminating article, and thank the editors for the opportunity to discuss the article. The deconfounder method that this article presents is appealing: a number of important scientific investigations and high-stakes decisions fit into its template. Indeed, as the authors note, instances of the deconfounder have already been deployed without explicit causal language in a number of applied settings. By bringing to light the implicit causal argument that underlies this approach, the authors have sparked an important conversation with potentially far-reaching consequences. It is thus important to carefully outline when we expect the deconfounder method to succeed in characterizing causal relationships and when we expect it to fail.

I have personally been in conversation with the authors over the past two years about this work, and this discussion has yielded some interesting insights, some of which have been published (D'Amour 2019), and some of which now appear in the current version of the article and in follow-up work (Wang and Blei 2019). The aim of this note is to draw out some conclusions from this conversation about the role that the deconfounder can play in practical causal inference. In particular, I will make three points here. First, in my role as the critic in this conversation, I will summarize some arguments about the lack of causal identification in the bulk of settings where the “informal” message of the article suggests that deconfounder could be used. This is a point that is discussed at length in D'Amour (2019), which motivated the results concerning causal identification in Theorems 6–8. Second, I will argue that adding parametric assumptions to the working model to obtain identification of causal parameters (a strategy followed in Theorem 6 and in the experimental examples) is a risky strategy, and should only be done when extremely strong prior information is available. Finally, I will consider the implications of the nonparametric identification results provided for a narrow, but nontrivial, set of causal estimands in Theorems 7 and 8. I will highlight that these results may be even more interesting from the perspective of *detecting* causal identification from observed data, under relatively weak assumptions about confounders.

Throughout this note, I will draw connections to sensitivity analysis methods that probe the implications of unobserved confounding. This is a natural lens through which to study the deconfounder because many sensitivity analysis methods posit a similar latent variable model to the one that the deconfounder deploys as a working model (see, e.g., Rosenbaum and Rubin 1983). Well-designed sensitivity analyses can reveal how specific assumptions restrict the range of causal conclusions

that are compatible with the observed data, and are thus useful for understanding what is lost when assumptions like “no unobserved confounders” are relaxed to “no unobserved single-cause confounders.” Thus, I believe, as the authors suggest, that sensitivity analysis should be a core part of any workflow that deploys the deconfounder, and discuss at various places how sensitivity analysis could be used effectively in this setting.

Preliminaries. Following the article, I will denote causes as $A := (A^{(1)}, \dots, A^{(m)})$ taking specific values $a = (a^{(1)}, \dots, a^{(m)})$, potential outcomes as $Y(a)$. To avoid measure-theoretic considerations when writing conditioning statements, I will consider the treatments $A^{(k)}$ to be discrete. I will write observed outcomes as Y^{obs} , where, under the stable unit treatment value assumption (SUTVA), $Y^{\text{obs}} = Y(A)$. Finally, I will denote by Z any latent confounders.

Throughout, I will consider models of the joint distribution $P(A, Y^{\text{obs}}, Z)$, which I will refer to as latent variable models. I will assume that unconfoundedness is satisfied conditional on Z :

$$Y(a) \perp\!\!\!\perp A \mid Z \quad Z\text{-a.e., } \forall a.$$

Thus, if the latent variable model is fully specified, the potential outcome distributions $P(Y(a))$ are also specified by the following adjustment formula, which “adjusts” for the confounder Z

$$P(Y(a)) = E[P(Y^{\text{obs}} \mid Z, A = a)] \quad \forall a. \quad (1)$$

I will refer to the integrand in (1) $P(Y^{\text{obs}} \mid Z, A = a)$ as the outcome model. If the confounder Z is observed, and the overlap condition is satisfied, then $P(Y(a))$ is identified from observed data. The question at hand is whether $P(Y(a))$ can be identified when Z is unobserved.

1. Fundamental Limitations of the Deconfounder Approach

I will begin by summarizing the argument in D'Amour (2019) critiquing the “informal” message about the deconfounder approach (stated most explicitly in the informal statement of Theorem 6 and Section 3.4). Specifically, this message asserts that, under the “no unobserved single-cause confounders” assumption, any well-fitting latent variable model $P(Y^{\text{obs}}, A, Z)$ will yield the correct potential outcome distribution $P(Y(a))$ via the adjustment formula (1). This informal story is motivated by strong intuition. Lemmas 1–3 establish that multi-cause

confounding leaves an observable “imprint” of dependence between the causes A . Thus, it seems natural that we might be able to gain some information, and even adjust for, an unobserved multi-cause confounder Z by modeling the dependence between the causes A .

Unfortunately, this intuition can only be carried so far: while a factor model for the causes A can recover some information about multi-cause confounders from observed data, the potential outcome distributions $P(Y(a))$ are not nonparametrically identified, except in cases where all confounding is observed. Thus, without additional unverifiable assumptions, no method can recover the distributions $P(Y(a))$ when there is unobserved confounding. In this section, I briefly demonstrate why this is the case. For a more in-depth argument about lack of identification in this setting with concrete examples, see D'Amour (2019).

As I show formally below, the key difficulty is that the causes A cannot be used simultaneously as measurements of the unobserved confounder Z , and as treatments whose effects are being estimated. If the event $A = a$ provides only a noisy measurement of Z , there is ambiguity in how the outcome model $P(Y^{\text{obs}} | Z, A = a)$ should align the variability in the residual distributions $P(Y^{\text{obs}} | A = a)$ and $P(Z | A = a)$; there are many specifications of the residual dependence between Y^{obs} and Z that are compatible with the observed data. This is a classic problem that arises when confounders are measured with error (see, e.g., Ogburn and Vanderweele 2012). On the other hand, if the event $A = a$ provides a perfect measurement of Z , such that there is some function $\hat{z}(A)$ such that $\hat{z}(a) = Z$, then the overlap condition fails. In this case, $P(Y^{\text{obs}} | Z, A = a)$ is only identified when $Z = \hat{z}(a)$ because the event $Z \neq \hat{z}(a)$ has zero probability in the observed data.

Let us now make this argument formal. To do this, we will account for how the two deconfounder assumptions of (a) good model fit, and (b) “no unobserved single-cause confounders” constrain the factor model and its implications about the potential outcomes $P(Y(a))$. This accounting is convenient if we rewrite the joint distribution using copula densities $c(V, W) = \frac{P(V, W)}{P(V)P(W)}$, which characterize the dependence between random variables independently of their marginal distributions.

$$P(Y^{\text{obs}}, A, Z) = \underbrace{P(A, Y^{\text{obs}})}_{\text{Observed}} \cdot \underbrace{P(Z|A)c(Z, A)}_{\text{Factor Model}} \cdot \underbrace{c(Y^{\text{obs}}, Z | A)}_{\text{Outcome Copula}}. \quad (2)$$

Each factor in this composition corresponds to a different assumption. The requirement for good model fit constrains only the first term, which specifies the distribution of observable quantities, while the “no unobserved single-cause confounders” assumption constrains the second term by constraining the causes to be conditionally independent given Z (Lemma 2).¹ This leaves the outcome-confounder copula density $c(Y^{\text{obs}}, Z | A) = \frac{P(Y^{\text{obs}}, Z | A)}{P(Y | A)P(Z | A)}$ unconstrained. This copula specifies the residual dependence between Y^{obs} and Z after conditioning on the causes A , and plays a key role in specifying the outcome model $P(Y^{\text{obs}} | A, Z)$.

¹The “no unobserved single-cause confounders” assumption does not uniquely identify the factor model by itself. Some structure also needs to be put on the latent variable, and even then, the factor model may not be identified. See D'Amour (2019) for an example where the factor model $P(A, Z)$ is itself not identified.

To complete the argument, note that the potential outcome distributions $P(Y(a))$ implied by the latent variable model are sensitive to the specification of this copula. Specifically, the estimand in (1) can be written as

$$P(Y(a)) = \int_Z P(Y^{\text{obs}} | A = a) c(Y^{\text{obs}}, Z | A = a) dP(Z).$$

Plugging in different specifications of the copula here yields different conclusions about $P(Y(a))$. Whenever $P(Y(a)) \neq P(Y^{\text{obs}} | A = a)$, there are multiple specifications of the copula that yield different conclusions about the potential outcomes.² Thus, $P(Y(a))$ is not identified unless there is no confounding and $P(Y(a)) = P(Y^{\text{obs}} | A = a)$.

We can now revisit the tension between the roles of causes A as measurements of Z , and as treatments. In cases where Z can only be inferred inexactly (i.e., $P(Z | A = a)$ is nondegenerate), the marginals $P(Y^{\text{obs}} | A = a)$ and $P(Z | A = a)$ put some constraints on the outcome model $P(Y^{\text{obs}} | Z, A = a)$, but the ambiguity in the copula implies that this model is not identified for any value of Z . In cases where Z can be reconstructed deterministically from the causes by some function $\hat{z}(a)$, (i.e., $P(Z | A = a)$ is degenerate), the outcome model $P(Y^{\text{obs}} | Z, A = a)$ is identified when $Z = \hat{z}(a)$, but the copula is undefined whenever $Z \neq \hat{z}(a)$ because this event has zero probability.

The upshot of this argument is that neither the deconfounder nor any other estimation method can adjust for unobserved confounding when estimating $P(Y(a))$ under the “no unobserved single-cause confounders” assumption alone. This conclusion holds no matter how much information we can glean about an unobserved confounder Z from the causes A . Although the single-cause confounding assumption does put some non-trivial structure on the latent variable model, it is not enough for causal estimation.

This lack of identification leaves practitioners looking to apply the deconfounder with two options: either make additional assumptions about the latent variable model $P(Y^{\text{obs}}, A, Z)$ so that $P(Y(a))$ is identified, or seek out causal comparisons where all of the confounding is effectively observed. In the Theory section of the article, the authors consider both of these paths. I will discuss each of these options in turn.

2. Parametric Identification, If You Must

I now turn to the subject of parametric identification of causal parameters, and offer some cautions about employing this strategy. Parametric identification is a natural strategy to employ when the causal parameters of interest are not nonparametrically identified. One obtains parametric identification by adding parametric assumptions to the working model that constrain the implied potential outcome distributions $P(Y(a))$ to be unique. The authors employ this parametric identification strategy in the experimental demonstrations of the deconfounder, as well as the formal result in Theorem 6. In Theorem 6, the copula $c(Y^{\text{obs}}, Z | A)$ is restricted by assuming that there is no interaction between

²To see this, note that the independence copula $c(Y^{\text{obs}}, Z | A = a) = 1$ implies that $P(Y(a)) = P(Y^{\text{obs}} | A = a)$. Thus, because $P(Y(a)) \neq P(Y^{\text{obs}} | A = a)$, this copula and the true copula yield different conclusions about $P(Y(a))$.

the causes A and the latent variable Z in the outcome model (i.e., that they combine linearly), and assuming that the confounder is piecewise constant in A . In the article's experiments, the authors assume a parametric factor model (e.g., a quadratic factor model for the genome-wide association study simulation), and a true linear outcome model. In the cases of Theorem 6 and the GWAS simulation study, the authors prove that these parametric assumptions are sufficient for identification.

Parametric identification can be a risky strategy to employ in practice. Specifically, the fact that the parametric assumptions are necessary to identify causal parameters implies that some aspects of these assumptions are not testable in the observed data. The decomposition in (2) makes this clear: given that the observed data are insufficient to identify the causal parameters, the parametric assumptions must restrict some of the unidentified portions of the latent variable model. Thus, to have confidence in this approach, one needs to have confidence in the parametric model used to identify causal effects as a *true model of the world*, not merely as an acceptable description of the observed data. This is because the identifying parametric assumptions specify not only a descriptive model of the observed data, but also a structural model for unobserved counterfactual outcomes. Relying on parametric identification may be feasible in cases where one has strong prior knowledge—for example, about the quantity represented by the unmeasured confounder, or the specific distributions of measurement errors—but such knowledge is often unavailable.

In addition, uncertainty estimates that are based directly on the parametric specification, for example, Bayesian credible sets, do not capture the full extent of uncertainty about causal effects according to the data. Specifically, these uncertainty estimates only quantify uncertainty *within* the specified model, and do not include the fundamental uncertainty associated with the lack of nonparametric identification of the potential outcome distributions $P(Y(a))$. As a result, unless the prior information used to specify the parametric assumptions is very strong, these uncertainty estimates will understate the degree of uncertainty about a causal parameter estimate. This is a standard critique of parametric uncertainty quantification, but carries extra weight in the context where conclusions depend on untestable aspects of the parametric model. For example, for the parametrically identified latent variable model in the GWAS example, as the sample size grows, the posterior for the causal parameter will concentrate around a single value, even though there exists a range of outcome models that correspond to different copulas $c(Y^{\text{obs}}, Z \mid A = a)$ that are equivalently compatible with the observed data, but would concentrate on different causal parameters. In fact, even small, seemingly benign parametric choices can mask alternative causal explanations. Lessons from latent variable models in the missing data and causal inference literatures can be instructive here. For example, analyses of the widely used Heckman selection model (Heckman 1979) have noted that the tail thickness of priors on latent variables can induce starkly different conclusions that are hidden by using the Gaussian default (Little and Rubin 2015; Ding 2014). See also discussions in Robins, Rotnitzky, and Scharfstein (2000) and Linero and Daniels (2018) for other examples.

Here, sensitivity analysis can be a useful tool to account for the fundamental uncertainty due to nonidentification of the

causal estimand. When performed with parametric models, sensitivity analyses perturb the parametric assumptions made with the estimating model to understand what other causal conclusions could be obtained under different parametric specifications. Performing sensitivity analyses on deconfounder estimates is straightforward: a number of sensitivity analysis approaches employ a working model with the same latent variable structure (e.g., Rosenbaum and Rubin 1983; Imbens 2003; Dorie et al. 2016; Cinelli and Hazlett 2018). However, sensitivity analyses can also fall victim to spurious parametric identification if the perturbations are not appropriately parameterized (Gustafson and McCandless 2018). To avoid this issue, it can be useful to employ sensitivity analysis strategies that cleanly separate the portions of the model that are identified by the observed data from those that are identified by parametric assumptions (Robins, Rotnitzky, and Scharfstein 2000; Linero and Daniels 2018; Franks, D'Amour, and Feller 2019). In the context of the deconfounder, the decomposition in (2) is a promising place to start, and is the subject of current work.

3. Toward a More Selective Deconfounder Workflow

A more cautious alternative to pursuing parametric identification is to seek out causal questions that have definitive answers under the “no unobserved single-cause confounders” assumption. The authors take this path in Theorems 7 and 8, in a setting where the latent confounder Z can be deterministically reconstructed as a function of the causes $\hat{z}(A)$. Here, however, the factor model seems less interesting as a tool for calculating causal effects, and more interesting as a tool for establishing empirically when no unobserved confounding is present. In my opinion, this latter framing seems more promising.

To review, in Theorem 7 the authors consider partitioning the causes into a set of focal causes $A_{1:k}$ whose effects will be estimated, and a set of auxiliary causes $A_{k+1:m}$ that will serve as measurements of the latent confounder. The theorem then states that if the latent confounder Z can be written as a function of the auxiliary causes $Z = \hat{z}(A_{k+1:m})$ alone,³ then the distributions of potential outcomes defined with respect to the subset of focal causes $P(Y(a_{1:k}))$ are identifiable subject to an overlap condition. Meanwhile, Theorem 8 states that certain counterfactual potential outcome distributions of the form $P(Y(a) \mid A = a')$ are identifiable as long as the causes a and a' map to the same value of the latent confounder, that is, $\hat{z}(a) = \hat{z}(a')$.

In these results, the authors focus on the role of the factor model in the identification of causal estimands under the “no unobserved single-cause confounders” assumption. However, the factor model is not essential for this point. Note that Theorems 7 and 8 both imply that the causal parameters can be identified in terms of the causes A alone, because it is assumed that the confounder Z can be written as a function of A . Written with slightly more generality, the identification result in Theorem 7 implies

$$P(Y(a_{1:k})) = E[P(Y^{\text{obs}} \mid A_{1:k} = a_{1:k}, A_{k+1:m})], \quad (3)$$

³This is not how the theorem is stated, but this function restriction is implied by the subsequent overlap condition.

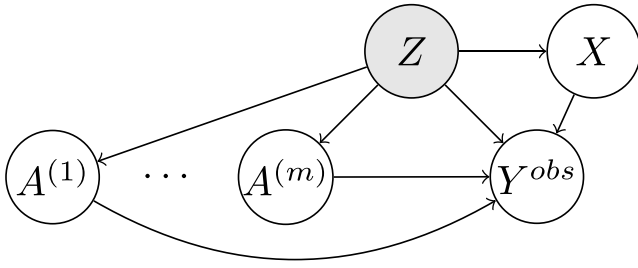


Figure 1. DAG assumed in Proposition 1, representing the relationship between causes A , latent confounder Z , covariates X , and observed outcome Y^{obs} .

while the identification result in Theorem 8 implies

$$P(Y(a') \mid A = a) = P(Y^{\text{obs}} \mid A = a') \quad \forall (a, a') \text{ s.t. } \hat{z}(a) = \hat{z}(a'). \quad (4)$$

To me, the more interesting point is that the factor model can be used in some cases to determine empirically whether some of the assumptions of the theorems are met. For example, the setting of Theorem 7 can be framed as a problem where the unobserved confounder Z is measured with proxies $A_{k+1:m}$. It is well-understood that in the limit where Z is perfectly recovered by the proxies, the potential outcome distribution $P(Y(a_{1:k}))$ is identified (Ogburn and Vanderweele 2012); however, in single-cause problems, one cannot determine whether this condition has been met. Similarly, Theorem 8 can be framed as a setting where one is imputing a set of counterfactual outcomes within a subpopulation where there is no confounding because, within this subpopulation, the confounder is fixed. Here, too, in single-cause problems, one cannot definitively identify such subpopulations from observed data. Interestingly, the theory of multi-cause confounding presented in the article suggests that these assumptions can be empirically validated under some restrictions on the causal DAG relating A to Y^{obs} and the “no unobserved single-cause confounders” assumption. For example, this theory supports the following proposition.

Proposition 1. Suppose there are no single-cause confounders, and the structural relationships between causes A , latent confounder Z , and observed outcomes Y^{obs} can be represented in the DAG in Figure 1. Suppose that in addition to causes A , we also have auxiliary covariates X , which are conditionally independent of the causes A conditional on the multi-cause confounder Z . Then for any function $\hat{z}(A, X)$ such that the causes A are mutually independent conditional on $\hat{z}(A, X)$, the conditional independence $A \perp\!\!\!\perp Y(a) \mid \hat{z}(A, X)$ also holds for each a .

Theorems 7 and 8 can be written as consequences of this proposition. This proposition is potentially useful because it shows that absence of certain confounding structures has observable implications. This insight is closely related to the literature on negative controls (see, e.g., Lipsitch, Tchetgen Tchetgen, and Cohen 2010).

This result suggests that one can use a similar workflow to the deconfounder to determine, at least in principle, whether identification statements like (3) or (4) are valid in a given setting. Specifically, one can obtain a function $\hat{z}(A, X)$ (perhaps by fitting a factor model), then test whether the causes A appear

to be mutually independent conditional on $\hat{z}(A, X)$. If one is satisfied that this is true, (3) or (4) can be applied. Importantly, this procedure is truly agnostic to the parametric specification of the model used to obtain $\hat{z}(A, X)$: all of the conditions are only functions of observables.

While the workflow in this procedure is similar to the deconfounder, it has a different use case. Instead of enabling causal inference in a wide range of cases, this procedure would be used to determine whether one can proceed with unconfounded inference at all, and can potentially give “no” as an answer. Still, this sort of procedure can prove useful in complex data contexts, where it can be valuable to surface causal questions that can be adequately answered with the available data. In a specific example of this approach, Sharma, Hofman, and Watts (2018) propose a similar testing procedure to uncover unconfounded comparisons, and use it to evaluate the causal effect of a recommender system on purchasing rates for certain products.

In outlining this procedure, I have belabored the point that it is a workflow “in principle” because it could prove tricky to implement. The observable implication that needs to be tested is a complex conditional independence statement, and these are notoriously difficult to test in practice (Shah and Peters 2018). In particular, one would receive the “green light” to estimate a causal parameter by failing to reject the null of conditional independence, which can only be reliably depended upon if the test has acceptably high power, but designing such tests is difficult, and in some settings, impossible.

Here, it can again be helpful to turn back to sensitivity analysis. Instead of attempting to rule out all possible forms of dependence between the causes A conditional on $\hat{z}(A, X)$, a sensitivity analysis approach could explore a number of candidate models for the residual dependence between the causes A and relate these models to the confounding induced by the unobserved confounder Z . For example, one could examine the range of causal effects that would be compatible with the assumption that, conditional on $\hat{z}(A, X)$, the causes A are no more predictive of a potential outcome $Y(a)$ than any leave-one-out set of the causes A_{-k} is able to predict a held-out cause $A^{(k)}$. This sort of calibration argument is common in more standard sensitivity analyses (Imbens 2003; Dorie et al. 2016; Cinelli and Hazlett 2018; Franks, D’Amour, and Feller 2019). In cases where dependence between the causes can be ruled out conclusively, this approach would yield a sensitivity region that collapses to a point; however, in the more likely case where many dependences cannot be ruled out, this approach would represent this uncertainty with a wider sensitivity region. It should be noted that constructing a plausible sensitivity analysis of this type would require deep domain knowledge to justify the analogy between different dependences between variables. Negative control methods and related identification strategies (Lipsitch, Tchetgen Tchetgen, and Cohen 2010; Miao, Geng, and Tchetgen Tchetgen 2018) could be framed as particularly successful executions of this type of argument.

4. Conclusion

In writing this article, the authors have drawn attention to a problem that is simultaneously scientifically impor-

tant, methodologically interesting, and conceptually subtle. Although I have taken on the role of critic in our conversations, I believe their contribution here is important. I remain skeptical about the deconfounder as a method for causal point estimation, but believe that the authors' characterization of multi-cause confounding could yield fruitful developments in sensitivity analysis, and in potentially obtaining identification results in more complex settings. This work has certainly inspired me to pay more attention to this problem, and to consider how new methods and tools can be developed to help practitioners draw principled causal conclusions in this setting.

Acknowledgments

I would like to thank Avi Feller, Alex Franks, Elizabeth Ogburn, James Atwood, and D. Sculley for helpful comments. Thanks to Yixin Wang and David Blei for open discussions about their work.

References

- Cinelli, C., and Hazlett C. (2018), "Making Sense of Sensitivity: Extending Omitted Variable Bias," Technical Report, Working Paper. [1599,1600]
- D'Amour, A. (2019), "On Multi-Cause Causal Inference With Unobserved Confounding: Counterexamples, Impossibility, and Alternatives," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3478–3486. [1597,1598]
- Ding, P. (2014), "Bayesian Robust Inference of Sample Selection Using Selection-t Models," *Journal of Multivariate Analysis*, 124, 451–464. [1599]
- Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016), "A Flexible, Interpretable Framework for Assessing Sensitivity to Unmeasured Confounding," *Statistics in Medicine*, 35, 3453–3470. [1599,1600]
- 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, DOI: 10.1080/01621459.2019.1604369. [1599,1600]
- Gustafson, P., and McCandless, L. C. (2018), "When Is a Sensitivity Parameter Exactly That?," *Statistical Science*, 33, 86–95. [1599]
- Heckman, J. J. (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. [1599]
- Imbens, G. W. (2003), "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132. [1599,1600]
- Linero, A. R., and Daniels, M. J. (2018), "Bayesian Approaches for Missing Not at Random Outcome Data: The Role of Identifying Restrictions," *Statistical Science*, 33, 198–213. [1599]
- Lipsitch, M., Tchetgen Tchetgen, E., and Cohen, T. (2010), "Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies," *Epidemiology*, 21, 383. [1600]
- Little, R. J. A., and Rubin, D. B. (2015), *Statistical Analysis With Missing Data*, New York: Wiley. [1599]
- Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018), "Identifying Causal Effects With Proxy Variables of an Unmeasured Confounder," *Biometrika*, 105, 987–993. [1600]
- Ogburn, E. L., and Vanderweele, T. J. (2012), "Bias Attenuation Results for Nondifferentially Mismeasured Ordinal and Coarsened Confounders," *Biometrika*, 100, 241–248. [1598,1600]
- Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000), "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*, eds. M. E. Halloran and D. Berry, New York: Springer, pp. 1–94. [1599]
- Rosenbaum, P. R., and Rubin, D. B. (1983), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society, Series B*, 45, 212–218. [1597,1599]
- Shah, R. D., and Peters, J. (2018), "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure," arXiv no. 1804.07203. [1600]
- Sharma, A., Hofman, J. M., and Watts, D. J. (2018), "Split-Door Criterion: Identification of Causal Effects Through Auxiliary Outcomes," *The Annals of Applied Statistics*, 12, 2699–2733. [1600]
- Wang, Y., and Blei, D. M. (2019), "Multiple Causes: A Causal Graphical View," arXiv no. 1905.12793. [1597]



Comment on: “The Blessings of Multiple Causes” by Yixin Wang and David M. Blei

Susan Athey, Guido W. Imbens & Michael Pollmann

To cite this article: Susan Athey, Guido W. Imbens & Michael Pollmann (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, DOI: [10.1080/01621459.2019.1691008](https://doi.org/10.1080/01621459.2019.1691008)

To link to this article: <https://doi.org/10.1080/01621459.2019.1691008>



Published online: 03 Jan 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Comment on: “The Blessings of Multiple Causes” by Yixin Wang and David M. Blei

Susan Athey^{a,b,c}, Guido W. Imbens^{a,b,c,d}, and Michael Pollmann^d

^aGraduate School of Business, Stanford University, Stanford, CA; ^bSIEPR, Stanford, CA; ^cNBER, Cambridge, MA; ^dDepartment of Economics, Stanford University, Stanford, CA

We congratulate the authors of Wang and Blei (2018) on a thought-provoking article on causal inference in settings with unobserved confounders. We expect that their ideas will lead to further developments in this important area. In this comment, we offer some thoughts on one such direction. Specifically, we explore the relevance of the Wang and Blei (2018) multiple causes ideas for a canonical problem in economics, namely the identification of demand functions in simultaneous equations (supply and demand) models. This is an old problem in economics, going back to the origins of the field of econometrics in the 1920s. Classic references include Tinbergen (1930), Haavelmo (1943), Wright (1928), and many econometric textbooks. See Angrist, Graddy, and Imbens (2000) for an interpretation in the modern causal inference literature. We show that the Wang and Blei (2018) multiple causes ideas bring new insights to this setting, but that they will not be a panacea.

First let us introduce a version of the canonical set up for demand function estimation to demonstrate how the presence of unobserved confounders naturally emerges. Consider initially a setting with a single product. For a number of markets, indexed by t , we observe the price for this product, P_t , and the quantity sold, Q_t . These markets may correspond to geographically separated markets or to the same location at different points in time. The interest is in the demand function $Q_t^D : \mathbb{R} \mapsto \mathbb{R}$ that describes how much consumers are willing to buy at different prices. Causal effects correspond to comparisons of the demand function at different prices, possibly scaled by the price differences, for example, $(Q_t^D(p) - Q_t^D(p'))/(p - p')$. For illustrative purposes, we assume the demand function is linear:

$$Q_t^D(p) = \alpha + \beta \times p + \varepsilon_t,$$

with β negative. The problem is that the prices we see are not randomly assigned. Instead they are determined by sellers who set prices to maximize profits (price times quantity minus cost). Suppose that the cost in market t is also linear with unobserved market-specific component η_t :

$$C_t(q) = (c + \eta_t) \times q,$$

so that profits are

$$\begin{aligned} \Pi_t(p) &= p \times Q_t^D(p) - C_t(Q_t^D(p)) = (\alpha + \beta p + \varepsilon_t) \\ &\quad \times p - (c + \eta_t) \times (\alpha + \beta p + \varepsilon_t). \end{aligned}$$

The profit maximizing price is

$$P_t = \arg \max_p \Pi_t(p) = \frac{c}{2} - \frac{\alpha}{2\beta} - \frac{\varepsilon_t}{2\beta} + \frac{\eta_t}{2}.$$

The realized price depends on ε_t , so the potential demand $Q_t^D(p)$ at a given price is correlated with the realized price P_t , and we do not have weak unconfoundedness (Rosenbaum and Rubin 1983; Imbens 2000). In econometric terminology, P_t is said to be endogenous. For the set up in directed acyclical graph (DAG) from Pearl (1995) and Pearl (2000), see Figure 1.

Traditionally, the econometric literature deals with the endogeneity of prices by using instruments. For example, in the analysis of the demand for fish in Angrist, Graddy, and Imbens (2000) the authors use weather conditions at sea as an instrument for the price, see Figure 2. These weather conditions are correlated with the unobserved cost shocks η_t , but assumed to be independent of the unobserved demand shocks, ε_t . Alternatively researchers exploit variation over time by using fixed effect methods and regression discontinuity designs (Angrist and Pischke 2008; Imbens and Lemieux 2008). Prices may change discontinuously, which can validate comparisons of quantities just before and after price changes.

The Wang and Blei (2018) article takes a different approach that we explore in the context of this demand function example. A critical component in their approach is the presence of additional information in the form of multiple causes. Suppose we have multiple products, for ease of exposition two, with prices P_{t1} and P_{t2} , and quantities Q_{t1} and Q_{t2} . The demand function for product i depends only on its own price, and is assumed to be linear, and the cost function also has the same structure as before, so that for $i = 1, 2$:

$$Q_{ti}^D(p) = \alpha_i + \beta_i \times p + \varepsilon_{ti}, \quad C_{ti}(q) = (c_i + \eta_{ti}) \times q,$$

so that the equilibrium price for product i is

$$P_{ti} = \frac{c_i}{2} - \frac{\alpha_i}{2\beta_i} - \frac{\varepsilon_{ti}}{2\beta_i} + \frac{\eta_{ti}}{2}.$$

Note that P_{t2} is not strictly speaking a second cause for Q_{t1} in the Wang and Blei (2018) sense: the direct effect of the second price on the quantity of the first product is absent in our set up. However, it does share with the Wang and Blei (2018) approach the key feature that this variable partly depends on what is

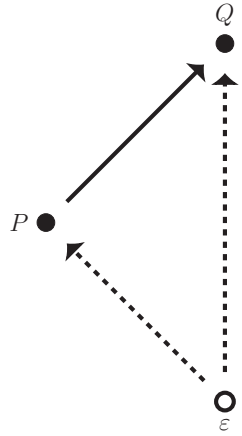


Figure 1. Unobserved confounder.

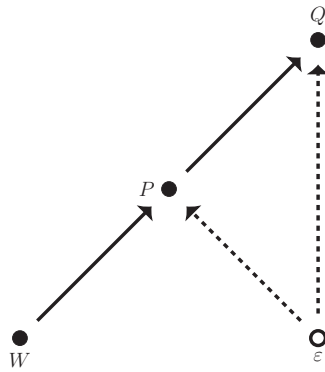


Figure 2. Instrumental variable.

potentially the same unobserved confounder as the first price P_{t1} , through the possible correlation of the residuals ε_{t1} and ε_{t2} .

Let the correlation between the demand shocks ε_{t1} and ε_{t2} be ρ_ε , and let the correlation between the supply shocks η_{t1} and η_{t2} be ρ_η , and assume that the demand and supply shocks are uncorrelated for both products. For ease of exposition we also assume that demand and supply shocks have the same variances for both products, σ_ε^2 and σ_η^2 , respectively. See Figure 3. We focus on estimation of the price effect in the demand function for the first product, β_1 .

Given this set up we consider various standard estimators for β_1 that deal with the unmeasured confounders in different ways. We calculate for each of these estimators the bias, and investigate under what conditions on the correlation structure of the demand and cost shocks these biases vanish.

The first estimator does not actually attempt to address the presence of the unobserved confounder, and is intended to set the stage and provide a baseline comparison. Consider least squares regression of the quantity for product 1, Q_{t1} , on the own price, P_{t1} :

$$Q_{t1} = \mu + \gamma \times P_{t1} + v_{t1}. \quad (1)$$

This leads to a biased estimate of β_1 because of the endogeneity of the price, or the presence of the unobserved confounder ε_{t1} .

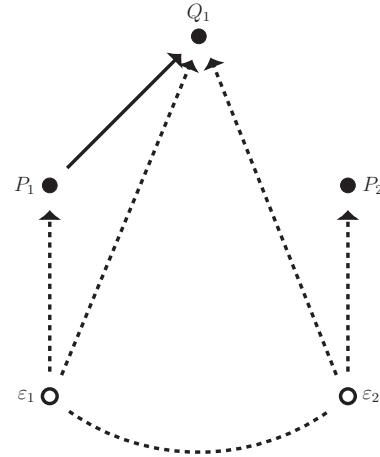


Figure 3. Two product example in general case.

The probability limit of the least squares estimator is

$$\gamma = \beta_1 - 2\beta_1 \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \beta_1^2 \sigma_\eta^2} > \beta_1$$

(bias because of unobserved confounder).

Note that β_1 is negative so there is an upward bias. This is a standard result in economics that one cannot in general estimate a demand function by regressing quantities on prices because of the endogeneity of the price.

The question now is what we can do with the additional information in the form of the price for the second good, given the structure of the model, including the correlation structure between the demand shocks and costs shocks. Here, we investigate the implications of four simple regression strategies, including some conventional ones and some that are in the spirit of the Wang and Blei (2018) multiple causes ideas. For each of these strategies, we explore when they remove or at least improve the biases relative to the true causal effect β_1 that we saw in the simple regression of Q_{t1} on P_{t1} .

1. First, consider simply controlling for the second price in the regression by adding P_{t2} to the specification in (1). This does not help in general. Regressing Q_{t1} on both prices, P_{t1} and P_{t2} , leads to

$$\gamma = \beta_1 - 2\beta_1 \frac{(\sigma_\varepsilon^2 + \beta_1^2 \sigma_\eta^2) \sigma_\varepsilon^2 - (\rho_\varepsilon \sigma_\varepsilon^2 + \beta_1^2 \rho_\eta \sigma_\eta^2) \rho_\varepsilon \sigma_\varepsilon^2}{(\sigma_\varepsilon^2 + \beta_1^2 \sigma_\eta^2)^2 - (\rho_\varepsilon \sigma_\varepsilon^2 + \beta_1^2 \rho_\eta \sigma_\eta^2)^2}.$$

Here, the asymptotic bias only vanishes in unusual settings.

2. Second, consider using P_{t2} as an instrument for P_{t1} in (1). This leads to

$$\gamma = \beta_1 - 2\beta_1 \frac{\rho_\varepsilon \sigma_\varepsilon^2}{\rho_\varepsilon \sigma_\varepsilon^2 + \beta_1^2 \rho_\eta \sigma_\eta^2}.$$

Here γ is equal to β_1 if two conditions are satisfied. First, the demand shocks must be uncorrelated ($\rho_\varepsilon = 0$), and second there must be some correlation in the cost shocks ($\rho_\eta \neq 0$). This type of identification strategy is sometimes used in Industrial Organization in analyses of markets for differentiated products (Hausman and Leonard 1996). This strategy is in fact the opposite of the common unobserved confounder

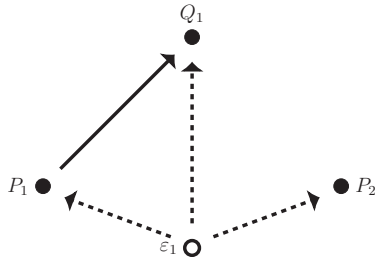


Figure 4. Proxy for unobserved confounder.

case that Wang and Blei (2018) focus on. In this case, the unobserved confounders are single cause confounders, and there is no common unobserved confounder at all.

3. Wang and Blei (2018) propose to use the second price in a different way. Their idea amounts in the current simple linear example to regressing Q_{t1} on P_{t1} controlling, not for the second price P_{t2} , but for the average price $\bar{P}_t = (P_{t1} + P_{t2})/2$. This leads to

$$\gamma = \beta_1 - 2\beta_1 \frac{(1 - \rho_\varepsilon)\sigma_\varepsilon^2}{(1 - \rho_\varepsilon)\sigma_\varepsilon^2 + \beta_1^2(1 - \rho_\eta)\sigma_\eta^2}.$$

The Wang and Blei (2018) insight in this setting corresponds to the fact that this is equal to the effect of interest β_1 if the demand shocks are the same for the two products, or $\rho_\varepsilon = 1$ (as long as the cost shocks are not perfectly correlated, $\rho_\eta \neq 1$). In the case with perfectly correlated demand shocks the unobserved confounder is common to both prices and there is no single-cause unobserved confounder, as in Figure 4.

4. The second Wang and Blei (2018) procedure leads to the same result. Regressing Q_{t1} on P_{t1} using the residual $Z_t = P_{t1} - \bar{P}_t$ as an instrument for P_{t1} also estimates

$$\beta_1 - 2\beta_1 \frac{(1 - \rho_\varepsilon)\sigma_\varepsilon^2}{(1 - \rho_\varepsilon)\sigma_\varepsilon^2 + \beta_1^2(1 - \rho_\eta)\sigma_\eta^2}.$$

Again this is equal to β_1 if the demand shocks are perfectly correlated, that is, if $\rho_\varepsilon = 1$.

Discussion

If we are interested in estimating demand functions, regressing quantity on price typically does not work because we expect that price is high when unobserved components of demand are high, so there is an upward bias in simple regression estimates. The Wang and Blei (2018) multiple-causes idea implies that if we have multiple prices that all depend on the same unobserved components, we can try to exploit this additional information by predicting and proxying for the unobserved component. Wang and Blei (2018) suggest doing so in two ways. One is by controlling for the principal components of the multiple prices, and one is by using part of the own price that cannot be predicted by these principal components as an instrument.

Can this work? Yes. If in this demand function example the demand shocks are highly correlated, and the supply shocks are not, then this will eliminate or at least reduce the biases. As such it is a welcome new addition to the identification strategies economists use in such settings. Does this work in general? No. It relies on assumptions about the structure of the demand and cost shocks. In a world where the demand shocks are only very weakly correlated relative to cost shocks (or not correlated at all), this does not work, and the Hausman and Leonard (1996) approach will work better. On the other hand, in a world where the demand shocks are highly correlated relative to cost shocks, or in particular if they are perfectly correlated, this will work well. Which case one is in depends on the relative magnitude of the demand and cost shock correlations, and will require assessing those in the specific context.

Wang and Blei (2018) bring new ideas to the widespread problem of accounting for the presence of unobserved confounders. Such problems are pervasive in economics and other social sciences, where researchers have developed many specific methods to deal with particular cases. The Wang and Blei (2018) multiple-cause proposals add to the methods that can be used in those settings. We expect they will find multiple applications in the social sciences.

Funding

We are grateful for support from the Office of Naval Research under grant N00014-17-1-2131.

References

- Angrist, J. D., Graddy, K., and Imbens, G. W. (2000), "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models With an Application to the Demand for Fish," *The Review of Economic Studies*, 67, 499–527. [1602]
- Angrist, J. D., and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press. [1602]
- Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1–12. [1602]
- Hausman, J. A., and Leonard, G. K. (1996), "Economic Analysis of Differentiated Products Mergers Using Real World Data," *George Mason Law Review*, 5, 321. [1603,1604]
- Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710. [1602]
- Imbens, G., and Lemieux, T. (2008), "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615–635. [1602]
- Pearl, J. (1995), "Causal Diagrams for Empirical Research," *Biometrika*, 82, 669–688. [1602]
- (2000), *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press. [1602]
- Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [1602]
- Tinbergen, J. (1930), "Determination and Interpretation of Supply Curves: An Example," *Zeitschrift für Nationalökonomie*, 1, 669–679. [1602]
- Wang, Y., and Blei, D. M. (2018), "The Blessings of Multiple Causes," arXiv no. 1805.06826. [1602,1603,1604]
- Wright, P. G. (1928), *Tariff on Animal and Vegetable Oils*, New York: Macmillan Company. [1602]



Comment: The Challenges of Multiple Causes

Kosuke Imai & Zhichao Jiang

To cite this article: Kosuke Imai & Zhichao Jiang (2019) Comment: The Challenges of Multiple Causes, Journal of the American Statistical Association, 114:528, 1605-1610, DOI: [10.1080/01621459.2019.1689137](https://doi.org/10.1080/01621459.2019.1689137)

To link to this article: <https://doi.org/10.1080/01621459.2019.1689137>



Published online: 03 Jan 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Comment: The Challenges of Multiple Causes

Kosuke Imai^a and Zhichao Jiang^b

^aDepartment of Government and Department of Statistics, Harvard University, Cambridge, MA; ^bDepartment of Biostatistics and Epidemiology, University of Massachusetts, Amherst, MA

We begin by congratulating Yixin Wang and David Blei for their thought-provoking article that opens up a new research frontier in the field of causal inference. The authors directly tackle the challenging question of how to infer causal effects of many treatments in the presence of unmeasured confounding. We expect their article to have a major impact by further advancing our understanding of this important methodological problem. This commentary has two goals. We first critically review the deconfounder method and point out its advantages and limitations. We then briefly consider three possible ways to address some of the limitations of the deconfounder method.

1. The Advantages and Limitations of the Deconfounder Method

We first discuss several advantages offered by the deconfounder method. We then examine the assumptions required by the method and discuss its limitations.

1.1. The Deconfounder Method

Suppose that we have a simple random sample of n units from a population. We have a total of m treatments, represented by the m -dimensional vector, $\mathbf{A}_i = (A_{i1}, A_{i2}, \dots, A_{im})^\top$, for unit i . For the sake of simplicity, we ignore the possible existence of observed confounders \mathbf{X}_i . But, all the arguments of this commentary are applicable, conditional on \mathbf{X}_i . The deconfounder method consists of the following two simple steps. The first step fits the following factor model to the observed treatments,

$$p(A_{i1}, A_{i2}, \dots, A_{im}) = \int p(\mathbf{Z}_i) \prod_{j=1}^m p(A_{ij} | \mathbf{Z}_i) d\mathbf{Z}_i, \quad (1)$$

where $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{ik})^\top$ represents the k -dimensional vector of latent factors.

Once the estimates of the factors $\hat{\mathbf{Z}}_i$, which Wang and Blei call the *substitute confounders*, are obtained, the second step estimates the average causal effects of multiple treatments by adjusting for these substitute confounders as follows,

$$\begin{aligned} \tau(\mathbf{a}, \mathbf{a}') &= \mathbb{E}\{Y_i(\mathbf{a}) - Y_i(\mathbf{a}')\} \\ &= \mathbb{E}\{\mathbb{E}(Y_i | \mathbf{A}_i = \mathbf{a}, \hat{\mathbf{Z}}_i) - \mathbb{E}(Y_i | \mathbf{A}_i = \mathbf{a}', \hat{\mathbf{Z}}_i)\}, \end{aligned} \quad (2)$$

where $\mathbf{a} \in \mathcal{A}$ and $\mathbf{a}' \in \mathcal{A}$ are the vectors of selected treatment values with $\mathbf{a} \neq \mathbf{a}'$ and \mathcal{A} represents the support of \mathbf{A}_i . In practice, a regression model may be used to adjust for the substitute confounders as demonstrated by Wang and Blei in their empirical application.

The deconfounder method is attractive to applied researchers for several reasons. First, it is a simple procedure based on two classes of familiar statistical models—factor models and regression models. Second, the method offers diagnostics in observational studies with unmeasured confounding. Specifically, researchers can check the conditional independence among the observed treatments given the estimated factors,

$$A_{ij} \perp\!\!\!\perp \mathbf{A}_{i,-j} | \hat{\mathbf{Z}}_i \quad (3)$$

for any $j = 1, \dots, m$ and $\mathbf{A}_{i,-j}$ represents all the treatments except A_{ij} . If this conditional independence does not hold, then there may exist unobserved confounders that affect both A_{ij} and some of $\mathbf{A}_{i,-j}$, yielding a biased causal estimate. As discussed below, however, the lack of conditional independence may also be due to the misspecification of factor model, which, for example, would be present if there are causal relationships among treatments.

In sum, the deconfounder method proposes a simple solution to a long-standing problem of inferring causal effects of multiple treatments in observational studies. Many analysts of observational studies rely upon the assumption that the treatments are unconfounded conditional on a set of observed pre-treatment covariates. And yet, it is often difficult to rule out the possible existence of unobserved confounders. The deconfounder method not only offers a new identification strategy in the presence of unobserved confounding, but also shows how to check the validity of the resulting estimates under certain assumptions.

1.2. Assumptions

What assumptions does the deconfounder method require? Wang and Blei use a graphical model to represent the conditional dependencies required by the deconfounder method. Here, we reproduce the graphical model using the directed acyclic graph (DAG) in Figure 1. In addition to the SUTVA (Rubin 1990), this DAG implies several key assumptions. First, the unobserved confounders \mathbf{Z} should represent all

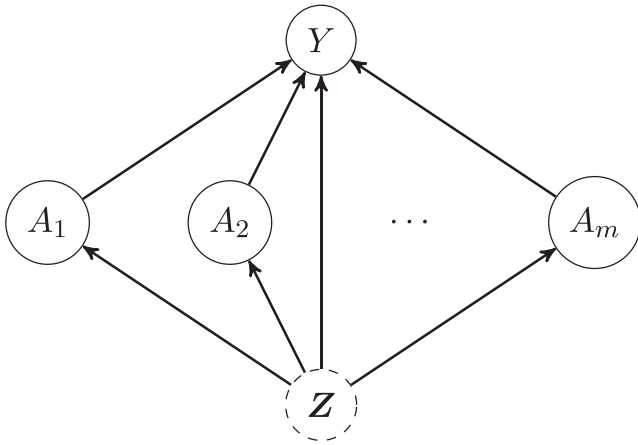


Figure 1. Directed acyclic graph for the deconfounder method.

confounding variables such that the treatments are ignorable given Z ,

$$Y_i(\mathbf{a}) \perp\!\!\!\perp A_i \mid Z_i \quad (4)$$

for any $\mathbf{a} \in \mathcal{A}$. The assumption implies that the multi-cause confounder Z_i suffices to adjust for the treatment-outcome confounding.

Second, the DAG also implies the following conditional independence assumption,

$$A_{ij} \perp\!\!\!\perp A_{i,-j} \mid Z_i \quad (5)$$

for any $j = 1, 2, \dots, m$. The assumption justifies the factor model in Equation (1). This assumption is violated if, for example, there exists a causal relationship among treatments. In the movie revenue application considered in the original article, the assumption is violated if the choice of actor for the main role (e.g., Sean Connery in a James Bond movie) influences the selection of actor for another role (e.g., Bernard Lee as the character of M). This is an important limitation of the deconfounder method as the problem may be common in applied research with multiple treatments.

In addition, according to Wang and Blei, the deconfounder method also requires the following overlap assumption that is

not explicitly represented in the DAG,

$$p(A_i \in \mathcal{A}^* \mid Z_i) > 0 \quad (6)$$

for all sets $\mathcal{A}^* \subset \mathcal{A}$ with $p(A_i \in \mathcal{A}^*) > 0$. The assumption implies that the choice of treatment values \mathbf{a} may be constrained when estimating $\mathbb{E}\{Y_i(\mathbf{a})\}$. If the selected value of \mathbf{a} does not belong to \mathcal{A}^* , then the resulting causal inference will be based on extrapolation.

Finally, the key identification condition of the deconfounder method is the assumption of “no unobserved single-cause confounder.” Wang and Blei formalize this assumption as the following set of conditional independence assumptions (see Definition 4 of the original article),

$$Y_i(\mathbf{a}) \perp\!\!\!\perp A_{ij} \mid \mathbf{V}_{ij}, \quad (7)$$

$$A_{ij} \perp\!\!\!\perp A_{i,-j} \mid \mathbf{V}_{ij} \quad (8)$$

for any $j = 1, 2, \dots, m$, $\mathbf{a} \in \mathcal{A}$, and some random variable \mathbf{V}_{ij} . In addition, the authors require that these conditional independence relations do not hold when conditioning on any proper subset of the sigma algebra of \mathbf{V}_{ij} .

Unfortunately, these conditional independence assumptions are not sufficient to eliminate the possible existence of unobserved single-cause confounders. Figure 2 presents two examples, in which single-cause confounders exist, but Equations (7) and (8) still hold. In addition, both cases can be reduced to the DAG in Figure 1 where no single-cause unobserved confounder exists by defining the unobserved multi-cause confounder as $Z = (Z_1, Z_2, Z_3)$. The examples demonstrate that a single multi-cause confounder can be decomposed into multiple single-cause confounders, and that several single-cause confounders can be combined into a single multi-cause confounder. Therefore, it is difficult to distinguish between single-cause and multiple-cause confounders without the knowledge of causal relationships among the variables.

We believe that it is important to develop the precise formal statement of the no unobserved single-cause confounder assumption. Such formalization allows us to understand how this assumption enables the identification of causal effects. In addition, our discussion implies that assessing the credibility of the assumption requires the scientific knowledge about the underlying causal structure involving unobserved confounders.

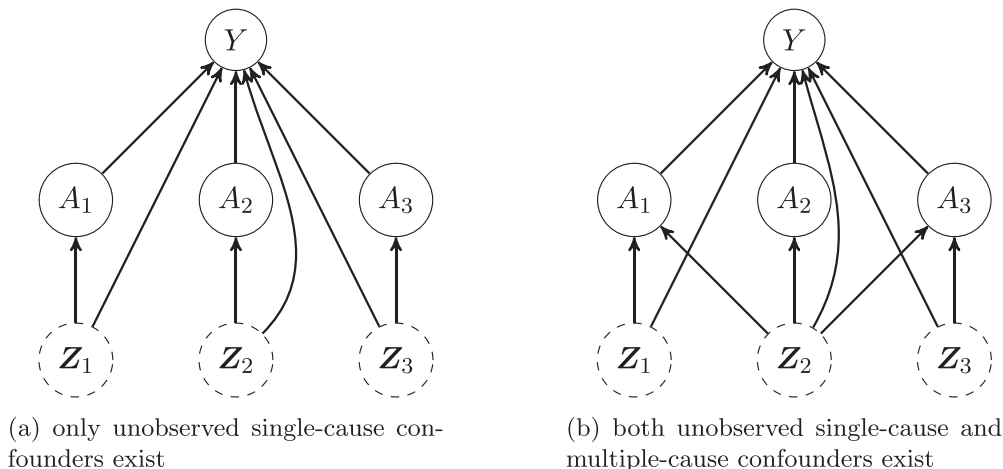


Figure 2. Examples of unobserved single-cause confounders.

1.3. Nonparametric Identification

Wang and Blei establish the nonparametric identification of the average treatment effect given in Equation (2) under the aforementioned assumptions in two steps. First, they show that a factor model of the observed treatments can be used to consistently estimate the substitute confounder. Second, they show that given the substitute confounder, the average treatment effects can be nonparametrically identified using Equation (2).

In an insightful paper, D'Amour (2019) demonstrates that this two-step proof strategy leads to two problems for the deconfounder method. First, there may be more than one factor model that is compatible with the distribution of the observed treatments. He provides an example where different factor models that are compatible with the distribution of the observed treatments under the structure of Figure 1 yield different causal estimates. Second, D'Amour shows that even if a factor model is uniquely identified, the nonparametric identification is in general impossible.

Moving beyond the counterexamples, we consider the identification assumption for the factor model, discuss the role of the substitute confounder, and assess the overlap assumption required by the deconfounder method.

With respect to the identifiability of factor models, Kruskal (1977) and Allman, Matias, and Rhodes (2009) give the general identification assumptions when observed variables are discrete. In this case, a crucial assumption is that the latent factor is correlated with the observed variables. In our context, this means that \mathbf{Z} must causally affect each treatment A_j . In the causal inference literature, this assumption is known as faithfulness (Spirtes et al. 2000), which states that there exists conditional independence among variables in the population distribution if and only if it is entailed in the corresponding DAG. Thus, although Wang and Blei only discuss a set of conditional independence assumptions, the deconfounder method requires the faithfulness assumption to ensure the identifiability of factor model.

Next, we discuss the role of the substitute confounder. In the proof of the deconfounder method, Wang and Blei not only assume that the true unobserved confounder \mathbf{Z}_i can be consistently estimated, but also treat the estimated substitute confounder $\widehat{\mathbf{Z}}_i$ as its true counterpart. This proof strategy ignores the crucial fact that the (estimated) substitute confounder is a function of observed treatments $\widehat{\mathbf{Z}}_i = \widehat{h}_M(\mathbf{A}_i) = \mathbb{E}_M(\mathbf{Z}_i | \mathbf{A}_i)$, where \widehat{h}_M indicates the fact that the substitute confounder is estimated from the data and depends on the choice of factor model and \mathbb{E}_M represents the expectation with respect to the fitted factor model. We emphasize that the substitute confounder $\widehat{\mathbf{Z}}_i$ does not converge in probability to the true confounder \mathbf{Z}_i , which in itself is a random variable. Rather, the substitute confounder converges to a function of observed treatments. Yet, this consistency result is required for the key results of the paper (i.e., Theorems 6–8).

We also closely examine the identification formula given in Equation (2) by explicitly writing out the conditional expectation,

$$\mathbb{E}\{\mathbb{E}(Y_i | \mathbf{A}_i = \mathbf{a}, \widehat{\mathbf{Z}}_i)\} = \int \mathbb{E}(Y_i | \mathbf{A}_i = \mathbf{a}, \widehat{\mathbf{Z}}_i) p(\widehat{\mathbf{Z}}_i) d\widehat{\mathbf{Z}}_i. \quad (9)$$

Notice that Equation (9) does not follow unless the support of $p(\widehat{\mathbf{Z}}_i | \mathbf{A}_i = \mathbf{a})$ is identical to the support of $p(\widehat{\mathbf{Z}}_i)$ for any

given $\mathbf{a} \in \mathcal{A}$. Unfortunately, since the substitute confounder is a function of the observed treatments, $p(\widehat{\mathbf{Z}}_i | \mathbf{A}_i = \mathbf{a})$ is in general degenerate. The overlap assumption given in Equation (6) is not applicable because the assumption is about the (true) unobserved confounders \mathbf{Z}_i rather than the (estimated) substitute confounders, $\widehat{\mathbf{Z}}_i$. This means that we can only identify $\mathbb{E}(Y_i | \mathbf{A}_i = \mathbf{a}, \widehat{\mathbf{Z}}_i = \mathbf{z}) = \mathbb{E}(Y_i | \mathbf{A}_i = \mathbf{a})$ for the values of \mathbf{z} with $\mathbf{z} = \widehat{h}_M(\mathbf{a})$, implying that only a certain set of causal effects are identifiable.

In Theorem 6 of the original paper, Wang and Blei address this problem by imposing two additional restrictions. First, it is assumed that the outcome is separable in the following sense,

$$\mathbb{E}\{Y_i(\mathbf{a}) | \widehat{\mathbf{Z}}_i\} = f_1(\mathbf{a}) + f_2(\widehat{\mathbf{Z}}_i), \quad (10)$$

$$\mathbb{E}(Y_i | \mathbf{A}_i, \widehat{\mathbf{Z}}_i) = f_3(\mathbf{A}_i) + f_4(\widehat{\mathbf{Z}}_i), \quad (11)$$

where we use $\widehat{\mathbf{Z}}_i$ instead of \mathbf{Z}_i to emphasize the fact that the substitute confounder is estimated. Although Equation (10) allows us to write the average treatment effect as a function of treatment values alone, that is, $\mathbb{E}\{Y_i(\mathbf{a}) - Y_i(\mathbf{a}')\} = f_1(\mathbf{a}) - f_1(\mathbf{a}')$, this assumption is not particularly helpful for identification since conditioning on $\widehat{\mathbf{Z}}_i$ is still required to identify the mean potential outcomes. In addition, Equation (11) can be rewritten as $\mathbb{E}(Y_i | \mathbf{A}_i) = f_3(\mathbf{A}_i) + f_4(\widehat{h}_M(\mathbf{A}_i))$ because $\widehat{\mathbf{Z}}_i$ is a deterministic function of \mathbf{A}_i . This suggests that the validity of this restriction about the outcome model critically depends on the choice of factor model.

The second restriction is that when the treatments are continuous, the substitute confounder is a piecewise constant function, that is, $\nabla_{\mathbf{a}} f_{\theta}(\mathbf{a}) = 0$ where a parametric model is assumed for $p(\widehat{\mathbf{Z}}_i | \mathbf{A}_i = \mathbf{a}, \theta) = \delta_{f_{\theta}(\mathbf{a})}$ with a vector of parameters θ . A similar restriction is proposed for the case of discrete treatments. Since $p(\widehat{\mathbf{Z}}_i | \mathbf{A}_i = \mathbf{a}, \theta) = \delta_{\widehat{h}_M(\mathbf{a})}$ automatically holds, the assumption is valid if $\widehat{h}_M(\mathbf{a})$ is a piece-wise constant function. Thus, this second restriction also suggests that the choice of factor model is critical for the validity of the deconfounder method.

In sum, we conclude that the nonparametric identification is generally difficult to obtain under the deconfounder method. Because the substitute confounder is a function of observed treatments, it leads to the violation of the overlap assumption. Wang and Blei introduce two additional restrictions to address this problem. However, these assumptions impose severe constraints on the choice of factor model as well as that of outcome model. As a consequence, they may significantly limit the practical applicability of the deconfounder method. Even when researchers carefully choose a factor model that satisfies these restrictions, they may obtain causal effects only for a restricted range of treatment values.

2. Alternative Approaches

We next consider three alternative approaches to the important question of identifying the causal effects of multiple treatments in the presence of unobserved confounders. The approaches in this section will be based on Equation (4). Unlike the deconfounder method, however, we will directly consider the identification of the probability distributions involving the (true) unobserved confounder $p(\mathbf{A}_i, \mathbf{Z}_i)$ and $p(Y_i | \mathbf{A}_i, \mathbf{Z}_i)$ rather than adopting Wang and Blei's two-step proof strategy.

2.1. Parametric Approach

Wang and Blei use parametric models in their empirical applications. Here, we consider a more general parametric approach. A primary advantage of the parametric approach is simplicity, whereas its major limitation is the required modeling assumptions that may not be credible in practice.

Suppose that there exists a uniquely identifiable factor model for the treatments, and that the joint distribution of (\mathbf{A}, \mathbf{Z}) is also identifiable. We assume the following additive model for the outcome variable,

$$\mathbb{E}\{Y_i(\mathbf{a}) \mid \mathbf{Z}_i\} = \sum_{j=1}^m \beta_j b_j(a_j) + \sigma g(\mathbf{Z}_i),$$

where $b_j(\cdot)$ and $g(\cdot)$ are prespecified functions. Under this setting, it can be shown that if σ is known, then the average treatment effect is identifiable so long as $(b_1(A_{i1}), \dots, b_m(A_{im}))$ is linearly independent. In contrast, if σ is unknown, then the average treatment effect is identifiable if $(b_1(A_{i1}), \dots, b_m(A_{im}), \mathbb{E}\{g(\mathbf{Z}_i) \mid \mathbf{A}_i\})$ is linearly independent. This linear independence assumption is analogous to the overlap assumption discussed earlier, but the assumption can be tested using the observed data.

To illustrate this parametric approach, consider an example, in which we have three binary treatments $m = 3$ and one binary latent factor Z_i . Further assume that we have the following outcome model,

$$\mathbb{E}\{Y_i(\mathbf{a}) \mid \mathbf{Z}_i\} = \beta_0 + \sum_{j=1}^3 \beta_j A_{ij} + \sigma Z_i.$$

Now, consider a scenario, under which A_{ij} 's are mutually independent of one another given Z_i . Then, the joint distribution $p(A_{i1}, A_{i2}, A_{i3}, Z_i) = p(Z_i) \prod_{j=1}^3 p(A_{ij} \mid Z_i)$ is identifiable based on the joint distribution of (A_{i1}, A_{i2}, A_{i3}) up to label switching (see Kruskal 1977). Note that the average treatment effects are invariant to label switching. Thus, under this condition, even if σ is unknown, β_j 's are identifiable so long as $\mathbb{E}(Z_i \mid A_{i1}, A_{i2}, A_{i3})$ is not linear in (A_{i1}, A_{i2}, A_{i3}) .

Next, consider a different case shown as the DAG in Figure 3, in which one treatment causally affects other treatments. In

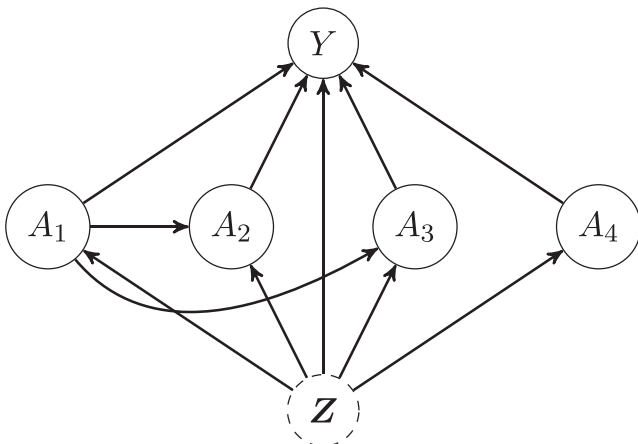


Figure 3. Directed acyclic graph in the presence of causal relations among treatments.

this case, we may focus on estimating the causal effects of (A_2, A_3, A_4) conditional on A_1 . We assume the following model for the outcome variable,

$$\mathbb{E}\{Y_i(\mathbf{a}) \mid \mathbf{Z}_i\} = \beta_0 + \sum_{j=1}^4 \beta_j A_{ij} + \sigma Z_i.$$

The joint distribution of \mathbf{A}_i and \mathbf{Z}_i under Figure 3 is given by $p(Z_i)p(A_{i1} \mid Z_i)p(A_{i2} \mid A_{i1}, Z_i)p(A_{i3} \mid A_{i1}, Z_i)p(A_{i4} \mid Z_i)$. This factorization is identifiable from the observed data (Allman, Matias, and Rhodes 2009). Then, even when σ is unknown, we can identify the parameters in the outcome model so long as $\mathbb{E}(Z_i \mid A_{i1}, A_{i2}, A_{i3}, A_{i4})$ is not linear in $(A_{i1}, A_{i2}, A_{i3}, A_{i4})$. Using these estimated parameters, we can obtain the estimates for the causal effects.

2.2. Nonparametric Approach

In the causal inference literature, many scholars first consider the problem of nonparametric identification by asking whether or not causal effects can be identified without making any modeling assumption. Only after the nonparametric identification of causal effects is established, researchers proceed to their estimation and inference. Cox and Donnelly (2011) regarded this approach as a general principle of applied statistics. They state, *If an issue can be addressed nonparametrically then it will often be better to tackle it parametrically; however, if it cannot be resolved nonparametrically then it is usually dangerous to resolve it parametrically.* (p. 96)

To enable the general nonparametric identification of causal effects in the current setting, we must introduce auxiliary variables. D'Amour (2019) considers the use of proxy variables. Here, we examine an approach based on instrumental variables. Figure 4 presents the DAG for this approach where \mathbf{W} represents a set of instrumental variables. Instrumental variables have the property that they are independent of the unobserved confounders \mathbf{Z} and influence the outcome Y only through the treatments \mathbf{A} .

For the sake of simplicity, we begin by considering the following separable model for the outcome,

$$\mathbb{E}\{Y_i(\mathbf{a}) \mid \mathbf{Z}_i\} = q(\mathbf{a}) + r(\mathbf{Z}_i),$$

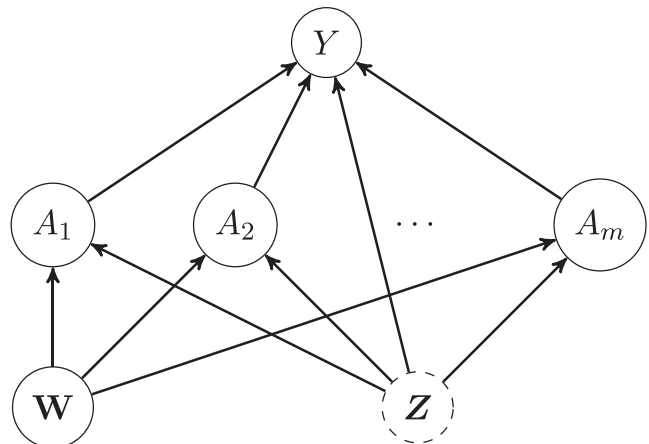


Figure 4. Directed acyclic graph for the instrumental variable approach.

where $\mathbb{E}\{r(Z_i)\} = 0$ without loss of generality. Since the instrumental variables satisfy $\mathbb{E}\{r(Z_i) \mid \mathbf{W}_i\} = \mathbb{E}\{r(Z_i)\} = 0$, we obtain,

$$\mathbb{E}(Y_i \mid \mathbf{W}_i) = \mathbb{E}\{q(A_i) \mid \mathbf{W}_i\} = \sum_{a \in \mathcal{A}} q(A_i = a)p(A_i = a \mid \mathbf{W}_i). \quad (12)$$

Since we can identify $\mathbb{E}(Y_i \mid \mathbf{W}_i)$ and $p(A_i \mid \mathbf{W}_i)$ from the observed data, the causal effects are identifiable if we can uniquely solve $q(\cdot)$ using Equation (12). Suppose that all the treatments are binary and the instrumental variable is discrete with L levels. Since there are 2^m parameters in $q(a)$, Equation (12) implies that the identification requires the $2^m \times L$ matrix $\{p(A_i \mid \mathbf{W}_i)\}$ to be full-rank. This condition is analogous to the overlap assumption discussed earlier and can be checked using the observed data. The proposed approach here, however, requires the instrumental variables to have more than 2^m levels. When m is large, it may be difficult to find instrumental variables that satisfy this condition.

The deconfounder method is closely related to the control function methods developed in the econometrics literature. The control function is a variable that, when adjusted for, renders an otherwise endogenous treatment variable exogenous (see, e.g., Wooldridge 2015). Imbens and Newey (2009) considered the nonparametric identification of the following nonseparable triangular system of equations (as before, we omit observed pretreatment confounding variables for simplicity),

$$Y_i = s_1(A_i, Z_i), \quad (13)$$

$$A_i = s_2(W_i, U_i), \quad (14)$$

where Z_i and U_i are unobserved, A_i is the endogenous treatment variable of interest, W_i is the instrumental variable with $W_i \perp\!\!\!\perp (Z_i, U_i)$, and $s_2(\cdot, \cdot)$ is a strictly monotonic function of U_i . When A_i is a vector and $U_i = Z_i$, Equations (13) and (14) become identical to the setting of the deconfounder method. Imbens and Newey show that the control function C_i is given by the cumulative distribution function of A_i given W_i , that is, $C_i = F_{A_i|W_i}(A_i, W_i)$. Like the substitute confounder, the control function unconfounds the treatment variable, that is, $Y_i(a) \perp\!\!\!\perp A_i \mid C_i$. This is because C_i is a one-to-one function of U_i , and A_i depends only on W_i conditional on U_i .

It is important to emphasize that the control function methodology requires the overlap assumption that the support of the marginal distribution of the control function, that is, $p(C_i)$, is the same as the support of the conditional distribution, that is, $p(C_i \mid A_i)$. However, unlike the case of the deconfounder method, control function is a function of both treatment and instrumental variables, making this overlap assumption more likely to be satisfied.

In sum, the nonparametric identification of causal effects in the current settings requires the existence of auxiliary variables. Here, we consider an approach based on instrumental variables. Even when such instrumental variables are available, certain overlap assumptions are needed. This point is also clearly shown for the control function methods that are closely related to the deconfounder method. As we discussed, the overlap assumptions required for these instrumental variable methods are less stringent than those required for the deconfounder method.

2.3. Stochastic Intervention Approach

Our discussion has identified the overlap assumption as a main methodological challenge for the deconfounder method. Because the estimated substitute confounder itself is a function of treatment variables, conditioning on the particular treatment values alters the support of its distribution. The parametric and nonparametric approaches introduced above address this problem through the reliance on modeling assumptions and the use of instrumental variables, respectively.

The final approach we consider is to change the causal quantities of interest using the idea of stochastic intervention. Instead of comparing two sets of fixed treatment values, we propose to contrast the two different distributions of treatments. In the movie application of the original article, one may be interested in comparing the revenue of a film featuring a typical cast for action movies with that featuring common actors for Sci-Fi movies. Stochastic intervention is a useful approach especially in the settings where inferring the average outcome under the fixed treatment values is difficult. For example, Geneletti (2007) applied it to mediation analysis, while Hudgens and Halloran (2008) proposed an experimental design with stochastic intervention to identify spillover effects. More recently, Kennedy (2019) considers the incremental interventions that shift propensity score values to avoid overlap assumption.

Specifically, we focus on the average causal effects of distributions of treatments rather than the effects of treatments themselves.

$$\delta(p_1, p_0) = \mathbb{E} \left\{ \int Y_i(a)p_1(A_i = a)da - \int Y_i(a)p_0(A_i = a)da \right\}, \quad (15)$$

where p_1 and p_0 are the prespecified distributions of treatments to be compared. Various distributions can be selected for comparison. For example, we may compare the conditional distributions of treatments given the different values of observed covariates, that is, $p_1(A_i \mid \mathbf{X}_i = \mathbf{x}_1)$ and $p_0(A_i \mid \mathbf{X}_i = \mathbf{x}_2)$. Moreover, if factors are interpretable, then we may choose the conditional distributions given some specific values of the factors, that is, $p_1(A_i \mid Z_i = z_1)$ and $p_0(A_i \mid Z_i = z_2)$. Topic models in the analysis of texts and ideal point models in the analysis of roll calls are good examples of interpretable factor models (Blei, Ng, and Jordan 2003; Clinton, Jackman, and Rivers 2004).

In the current setting, we may use the following estimator,

$$\hat{\delta}(p_1, p_0) = \sum_{i=1}^n Y_i \frac{p_1(A_i) - p_0(A_i)}{\hat{p}(A_i \mid Z_i)}, \quad (16)$$

where $\hat{p}(A_i \mid Z_i)$ is the estimated factor model. For this estimator, the required overlap assumption is that the support of $p_j(A_i)$ is a subset of the support of $p(A_i \mid Z_i)$ for $j = 0, 1$. Researchers can choose $p_1(A_i)$ and $p_0(A_i)$ so that this overlap assumption is satisfied. Furthermore, although the deconfounder method is not applicable when one treatment causally affects another, under the stochastic intervention approach one could model causal relationships among treatments by specifying $p(A_i \mid Z_i)$ provided that the model is identifiable. An example of such case is given in Figure 3.

3. Concluding Remarks

The article by Wang and Blei is an important contribution to the causal inference literature because it opens up a new research frontier. The authors study a relatively unexplored question of how to infer the causal effects of many treatments in the presence of unobserved confounders. The deconfounder method provides a novel and yet intuitive approach using familiar statistical models. A key insight is that under certain assumptions, the factorization of treatments can yield a substitute confounder as well as a practically useful diagnostic tool for checking the validity of the resulting substitute confounder.

Although the deconfounder method has advantages, as first pointed out by D'Amour (2019) and further elaborated in this commentary, the method is not free of limitations. In particular, it cannot achieve nonparametric identification without additional restrictions. We emphasized the violation of the overlap assumption due to the fact that the estimated substitute confounder is a function of observed treatments. Wang and Blei consider some restrictions on the outcome model that may overcome this limitation and enable identification. However, such restrictions may severely limit the applicability of the deconfounder method. More research is needed to investigate the consequences of these restrictions in practical settings.

We discussed three alternative approaches to the methodological problems of the deconfounder method. The first approach is based on parametric assumptions and extend the data analysis conducted in the original article. The second approach relies upon the use of instrumental variables and is related to the control function literature in econometrics. The final approach considers an alternative causal estimand based on stochastic intervention, which is particularly useful in the settings with high-dimensional treatments. We expect and hope that many researchers will follow up on the work of Wang and Blei and develop new methods for estimating the causal effects of multiple treatments in observational studies.

Acknowledgments

We thank Naoki Egami, Connor Jerzak, Michael Li, and Xu Shi for helpful discussions.

References

- Allman, E. S., Matias, C., and Rhodes, J. A. (2009), "Identifiability of Parameters in Latent Structure Models With Many Observed Variables," *The Annals of Statistics*, 37, 3099–3132. [1607,1608]
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1609]
- Clinton, J., Jackman, S., and Rivers, D. (2004), "The Statistical Analysis of Roll Call Data," *American Political Science Review*, 98, 355–370. [1609]
- Cox, D. R., and Donnelly, C. A. (2011), *Principles of Applied Statistics*, Cambridge: Cambridge University Press. [1608]
- D'Amour, A. (2019), "On Multi-Cause Causal Inference With Unobserved Confounding: Counterexamples, Impossibility, and Alternatives," arXiv no. 1902.10286. [1607,1608,1610]
- Geneletti, S. (2007), "Identifying Direct and Indirect Effects in a Non-Counterfactual Framework," *Journal of the Royal Statistical Society, Series B*, 69, 199–215. [1609]
- Hudgens, M. G., and Halloran, E. (2008), "Toward Causal Inference With Interference," *Journal of the American Statistical Association*, 103, 832–842. [1609]
- Imbens, G. W., and Newey, W. K. (2009), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. [1609]
- Kennedy, E. H. (2019), "Nonparametric Causal Effects Based on Incremental Propensity Score Interventions," *Journal of the American Statistical Association*, 114, 645–656. [1609]
- Kruskal, J. B. (1977), "Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, With Application to Arithmetic Complexity and Statistics," *Linear Algebra and Its Applications*, 18, 95–138. [1607,1608]
- Rubin, D. B. (1990), "Comments on 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9' by J. Splawa-Neyman Translated From the Polish and Edited by D. M. Dabrowska and T. P. Speed," *Statistical Science*, 5, 472–480. [1605]
- Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000), *Causation, Prediction, and Search*, Cambridge, MA: MIT Press. [1607]
- Wooldridge, J. M. (2015), "Control Function Methods in Applied Econometrics," *Journal of Human Resources*, 50, 420–445. [1609]



Comment on “Blessings of Multiple Causes”

Elizabeth L. Ogburn, Ilya Shpitser & Eric J. Tchetgen Tchetgen

To cite this article: Elizabeth L. Ogburn, Ilya Shpitser & Eric J. Tchetgen Tchetgen (2019) Comment on “Blessings of Multiple Causes”, Journal of the American Statistical Association, 114:528, 1611-1615, DOI: [10.1080/01621459.2019.1689139](https://doi.org/10.1080/01621459.2019.1689139)

To link to this article: <https://doi.org/10.1080/01621459.2019.1689139>



Published online: 03 Jan 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



Comment on “Blessings of Multiple Causes”

Elizabeth L. Ogburn^a, Ilya Shpitser^b, and Eric J. Tchetgen Tchetgen^c

^aDepartment of Biostatistics, Johns Hopkins University, Baltimore, MD; ^bDepartment of Computer Science, Johns Hopkins University, Baltimore, MD;

^cDepartment of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA

We are grateful to Wang and Blei (2019) (hereafter WB) for drawing attention to the important and increasingly popular project of using latent variable methods to control for unmeasured confounding. Prior causal inference research on this topic has not been adequately communicated or disseminated, leaving room for misconceptions, which we hope to begin to remedy in this discussion. We also appreciate that the authors have sought and been receptive to our feedback about their work. We would also like to thank the editors for giving us the opportunity to comment on this article.

However, this article has foundational errors. Specifically, the premise of the *deconfounder*, namely that a variable that renders multiple causes conditionally independent also controls for unmeasured multi-cause confounding, is incorrect. This can be seen by noting that no fact about the observed data alone can be informative about ignorability, since ignorability is compatible with any observed data distribution. Methods to control for unmeasured confounding may be valid with additional assumptions in specific settings (e.g., Price et al. 2006; Angrist and Pischke 2008; Kuroki and Pearl 2014), but they cannot, in general, provide a checkable approach to causal inference, and they do not, in general, require weaker assumptions than the assumptions that are commonly used for causal inference. While this is outside the scope of this comment, we note that much recent work on applying ideas from latent variable modeling to causal inference problems suffers from similar issues.

Causal inference aims to draw inferences about the parameters of the full data distribution—the distribution of the observed random variables and the potential outcomes—from realizations of the observed data distribution, which is generally a coarsened version of the full data distribution. For example, the full data distribution for a conditionally ignorable model with binary treatment is of the form $p(Y(1), Y(0), A, \mathbf{X})$, where the following conditional independences hold on the counterfactual outcomes $Y(1)$, and $Y(0)$, the treatment A and the set of baseline covariates \mathbf{X} : $Y(1) \perp A | \mathbf{X}$, and $Y(0) \perp A | \mathbf{X}$. The parameter of interest is often the average causal effect (ACE): $\mathbb{E}[Y(1) - Y(0)]$. The observed data distribution, on the other hand, is of the form $P(Y, A, \mathbf{X})$, where the observed outcome Y is a coarsened version of $Y(1)$ and $Y(0)$, defined by consistency as $Y(1) \cdot A + Y(0) \cdot (1 - A)$. Causal inference problems are often viewed as missing data problems, since every realization of the observed outcome Y yields exactly one of

the potential outcomes for the corresponding subject, with the other outcomes being missing data. With the *deconfounder*, WB aim to tackle settings with a vector \mathbf{A} of multiple treatments, where baseline covariates are unobserved (except for single-cause confounders, which we ignore throughout). In such cases, the observed data distribution is a marginal distribution of the form $p(Y, \mathbf{A})$, marginalized over the missing potential outcomes and the unobserved confounders.

The *deconfounder* proposal can be loosely summarized as follows:

- Suppose *ignorability* for the effect of a vector of causes \mathbf{A} on an outcome Y holds conditional on U : $\mathbf{A} \perp Y(\mathbf{a}) | U$.
- U is unobserved, but if it were observed then conditioning on and standardizing by U (*covariate adjustment*, or the *adjustment formula*) would identify causal effects of \mathbf{A} on Y , as in equation (2) of WB.
- In lieu of the unmeasured U , and in the absence of any unmeasured single-cause confounders, one can control for any variable Z such that A_1, \dots, A_m are mutually independent conditional on Z , because such a Z satisfies ignorability for all multi-cause confounders. Z is a *substitute confounder* for the true confounder U .

In addition to the above, the authors impose several additional assumptions at various points throughout the article. We describe these below. Nevertheless, the assumptions, as stated, do not imply the claimed results.

1. Conditionally Independent Causes Do Not Ensure Conditional Ignorability

The third step is the crux of the *deconfounder*. However, the criterion of conditional independence does not suffice to make Z a valid substitute confounder. This criterion does not rule out the inclusion of variables that may bias effects, nor does it ensure that all multi-cause confounders are captured by Z . Finding an observed proxy that suffices to control for all confounding via covariate adjustment is related to a body of work on *complete adjustment criteria* (Shpitser, VanderWeele, and Robins 2010; Perkovic et al. 2015). Below we give a few examples that violate these adjustment criteria, meaning that covariate adjustment is

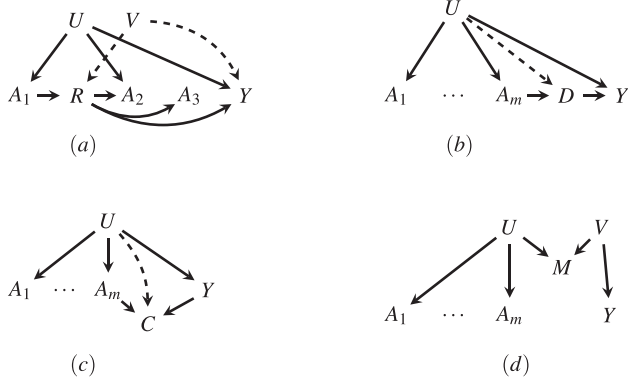


Figure 1. (a) A DAG in which A_1 and A_2 are causally dependent. (b) A DAG with a single-cause mediator. (c) A DAG with a single-cause collider. (d) A DAG with an M -bias collider.

not a valid identification strategy, but that are not excluded from the deconfounder.

1.1. The Deconfounder May Include Variables That Bias Effects

A substitute confounder constructed to render the causes mutually independent may include three types of variables that undermine the ability to identify causal effects. M -bias colliders, such as M in the directed acyclic graph (DAG) in Figure 1(d), and single-cause colliders, such as C in Figure 1(c), are variables that *induce* confounding (Cole et al. 2009; Elwert and Winship 2014), and single-cause mediators, such as R in Figure 1(a) and D in Figure 1(b), are variables that bias causal effects.

Both colliders and mediators are post-treatment variables. As WB note, it is crucial that all covariates used to identify causal effects via the formula in (2) be pretreatment variables, because conditioning on a downstream effect of \mathbf{A} may introduce bias in any direction (it need not bias effect estimates toward the null). However, Lemma 4, which states that the substitute confounder Z is guaranteed to be pretreatment, is incorrect. We first give an intuitive counterexample in which mediators would be included in a substitute confounder and then point out a problem in the proof of Lemma 4.

1.1.1. Causes Cannot Be Causally Dependent

Suppose the causes A_1, \dots, A_m can themselves have causal effects on one another, as would be expected in most of the motivating examples described in the introduction of WB (neurons may cause one another to fire; enrolling in one social program may increase the chance that someone will learn of or be referred to another social program; one medicine may be prescribed to treat side effects of another or of a procedure). Specifically, consider the case depicted in Figure 1(a), where A_1 causes A_2 , and to render them conditionally independent the deconfounder must include a variable, R , that breaks this connection. However, the effect of A_1 on Y is *through* R and therefore cannot be identified controlling for R ; depending on the relationships among A_1 , R , and Y , an estimator that controls for R could either over- or underestimate the true effect. This scenario is directly analogous to longitudinal causal inference problems with multiple time-varying treatments that contain

time-varying confounders, variables that serve as confounders for some treatments and as mediators for other treatments. If there is an unmeasured confounder for the R - Y relationship (represented by V and the dashed arrows in Figure 1(a)), then conditioning on R fails to identify the direct effects of \mathbf{A} on Y , because it opens a confounding pathway through V . See Hernan and Robins (2020) for an overview of these issues.

The answer to the question posed in Appendix B of WB, “Can the causes be causally dependent among themselves?” is therefore “no.” If they are causally dependent then the deconfounder, by dint of rendering the causes independent, breaks some of the structure among the causes \mathbf{A} , and as was originally established in the time-varying treatment setting, this undermines the identification of joint effects of \mathbf{A} on Y by standard covariate adjustment.

1.1.2. Analysis of Lemma 4

This simple argument also serves as a counterexample to Lemma 4, which states that the deconfounder does not pick up any post-treatment variables and can be treated as a pretreatment covariate. This is necessarily false whenever the causes are causally dependent among themselves, but it need not hold even if the causes are not causally dependent, see below.

The proof of Lemma 4 in Appendix I states that “Inferring the substitute confounder Z_i is separated from estimating the potential outcome. It implies that the substitute confounder is independent of the potential outcomes conditional on the causes.” The proof invokes the assumption that $Z \perp Y(\mathbf{A})|\mathbf{A}$. By the consistency property in causal inference, which defines the observed data variable Y as $\sum_{\mathbf{a}} \mathbb{I}(\mathbf{A} = \mathbf{a})Y(\mathbf{a})$, $Y(\mathbf{A})$ is equal to Y , which implies $Z \perp Y|\mathbf{A}$. This conditional independence cannot hold for any Z that satisfies ignorability, except in trivial settings. Limiting inquiry to settings in which there exists a deterministic function of the causes that suffices to identify causal effects rules out almost everything that is typically considered confounding.¹ (This is also the case replacing $Y(\mathbf{A})$ with $Y(\mathbf{a})$ in the original assumption (note the lower case \mathbf{a}), since Y is a fixed function of $Y(\mathbf{a})$ and \mathbf{A} .)

In fact, confounders confound *because* they are related to potential outcomes even conditional on the observed treatment and outcome. For example, if a person knows that their potential pain status under treatment $A = \text{tylenol}$ is preferable to their potential pain status under treatment $A = \text{notylenol}$, then they are more likely to take tylenol when they have a headache—so $Y(\mathbf{a})$ affects A . Obviously $Y(\mathbf{a})$ also affects Y , their pain status after treatment, so $Y(\mathbf{a})$ is itself a confounder. While this may be an extreme example, in general confounders are, almost by definition, intricately linked to the potential outcomes.

When the causes are not causally dependent (which is the setting for which WB recommend using the deconfounder, see Appendix B), can we ensure that a substitute confounder does not contain post-treatment variables? Any mediator or collider caused by more than one cause will be excluded from the substitute confounder, because such a variable is a collider between its causes and conditioning on it induces, rather than eliminates,

¹We updated the statement of this result to reflect the fact that different definitions exist for the presence of confounding; we are grateful to WB for drawing to our attention the fact a previous version was not entirely clear.

dependence among the causes. But single-cause mediators and colliders may be incorporated into the substitute confounder.

1.1.3. Single-Cause Mediators and Colliders

A single-cause mediator, such as D in Figure 1(b) will generally not be required to render the causes conditionally independent, and the same is true of a single cause collider, such as C in Figure 1(c). But in the absence of Lemma 4, one cannot guarantee that single-cause mediators and colliders would be excluded from substitute confounders. In particular, if the dashed arrow in Figure 1(b) is present, so that the unmeasured confounder is not independent of the mediator, then it is possible that a substitute confounder would include some or partial information about the mediator. Similarly, if the dashed arrow in Figure 1(c) is present, so that the unmeasured confounder is not independent of the collider, then it is possible that a substitute confounder would include some or partial information about the collider.

1.1.4. M-Bias Colliders

Even if one could exclude post-treatment random variables from the deconfounder, M -bias colliders, like M in Figure 1(d), can be pretreatment. They provide a counterexample to the premise that a pretreatment Z that renders the causes conditionally independent suffices to control for multi-cause confounding of A on Y , and specifically to Lemmas 1 and 2. While conditioning on U itself would suffice to control for M -bias, if, in addition to M , Z captures the part of U that affects dependence among the causes without capturing the part of U that relates A_m to M , then M -bias would remain.²

1.2. The Deconfounder Need Not Capture All Multi-Cause Confounders

We provide an example to illustrate that the deconfounder may not capture all multi-cause confounders, and then we point out a flawed premise in the proof of Lemmas 1 and 2. A related point is that the deconfounder may not be able to control for confounding even if it does capture all multi-cause confounders; this is because confounding involves the joint distribution of the causes and the potential outcomes, so in general learning a latent confounder requires dealing with this joint distribution. This is established by a copula argument in D'Amour (2019).

Conditioning on Z can render the causes mutually independent by separating a multi-cause confounder U into single-cause components, while failing to control for the relationship between the causes and the outcome. Here is an example: suppose U is a confounder of A_1 , A_2 , and Y , and suppose that, conditional on Z , $U \sim \text{Unif}(0, 1)$. Then $U|Z$ is decomposable into the sum of V and W , where V and W are independent.³ Further suppose that A_1 only depends on V and A_2 only depends on W . Then conditioning on Z renders A_1 and A_2 independent,

but there is no reason to think that it controls for confounding by U .

This counterexample to the claim that the deconfounder controls for all multi-cause confounding is pathological, but given the fact that modeling the marginal distribution of the causes can only tell us about the joint distribution of the causes and the outcome under stringent assumptions or in degenerate models, we expect counterexamples to be the rule, not the exception.

1.2.1. Analysis of Lemmas 1 and 2

The discussion above undermines the claim that the deconfounder, estimated via a factor model of the causes, suffices for ignorability to hold. The argument for this claim in WB is rather technical, but we briefly analyze it here. It is made through Lemma 1, which states that if A admits a Kallenberg construction from the deconfounder then ignorability holds conditional on the deconfounder, and Lemma 2, which states that all factor models of A admit a Kallenberg construction. However, Definition 3 misstates the Kallenberg construction for the relevant probability model. The probability model for analyzing the causal effect of A on Y subject to confounding by Z is the model for the full data distribution; the probability model that includes only the observed data is appropriate for prediction but not for causal inference. The full data comprise (in chronological/causal order) the random variables $\{Y(\mathbf{a}) : \mathbf{a} \in \mathcal{A}\}$, Z , A , and Y . Note that ignorability is a restriction on the full data distribution, not the observed data distribution (which often has no restrictions in causal inference problems). Put another way, no fact about the observed data alone can be informative about ignorability, since ignorability is compatible with any observed data distribution. Therefore, Theorem 5.10 of Kallenberg (1997) in fact implies $A_{ij} \stackrel{a.s.}{=} f_j(Z_i, \{Y(\mathbf{a}) : \mathbf{a} \in \mathcal{A}\}, U_{ij})$ rather than the construction given in equation (37) of WB, which omitted $\{Y(\mathbf{a}) : \mathbf{a} \in \mathcal{A}\}$. Thus, the Kallenberg construction used in the article cannot link factor models to ignorability. A Kallenberg construction on the full data, which could be informative about ignorability, is impossible to obtain given observed data information alone.

1.3. When Would a Latent Substitute Confounder Be Expected to Control for All Multi-Cause Confounding?

Identifying a latent substitute confounder from the observed data on A essentially requires the assumption that learning structure on the causes suffices to learn about any joint structure linking the causes with the outcome, in addition to the assumptions above.

A widely studied setting in which this would hold is when U represents unknown structure that is common to each A_k and to Y . This is likely to be the case in GWAS studies and in problems with clustered data with unknown clusters. In GWAS studies, including in WB's simulations, U represents population structure that is common across all of the causes and the outcome. For example, U might be an ancestry matrix indicating how n subjects are related to one another, and each of the causes and the outcome are expected to show dependence across the n subjects due to this same ancestry matrix. In this setting, any subset of the collection of variables with this same structure, that is any subset of $\{A_1, \dots, A_m, Y\}$, can be used to learn the

²We are grateful to WB for catching a mistake in a previous version of this counterexample.

³A random variable is decomposable if it is equal to the sum of independent random variables; a $\text{Unif}(0, 1)$ random variable is decomposable into a bernoulli random variable that takes the values 0 or 0.5 with equal probability and a uniform random variable over $(0, 0.5)$.

common underlying population structure, in particular the set $\{A_1, \dots, A_m\}$ as is commonly done in practice (Price et al. 2006).

Theorem 6 requires the deconfounder to be piecewise constant in the causes; this reduces the problem of confounding to one of clustering.

Another example when a latent substitute confounder controls for all multi-cause confounding is the fully parametric model given in Appendix C of WB.

2. Assumptions Beyond Ignorability

In this section, we assume that we are in the class of problems for which latent substitute confounders are known to perform well, for example, in the GWAS or clustering setting. We argue that even for those limited settings the assumptions required of the deconfounder are quite strong, and are not nonparametric. Below we discuss the assumptions required for the deconfounder that go beyond those required for “classical causal inference.” In exchange for the assumptions listed below, “classical causal inference” requires the sole (but strong and untestable) assumption of no unmeasured multi-cause confounders. Both the deconfounder and classical methods require no unmeasured single-cause confounders, SUTVA, and overlap (or positivity).

2.1. Nonparametric Identification

Although the terms *parametric* and *nonparametric* can mean different things to different researchers, generally a causal effect is said to be *nonparametrically identified* if either (a) the assumptions required for identification place no restrictions on the observed data distribution, except possibly up to a set of distributions of measure zero (Bickel et al. 1993), or (b) the only restrictions on the observed data distribution are those imposed by a nonparametric structural equation model. Such restrictions may include some conditional independences and inequality constraints. But causal effects cannot be nonparametrically identified (in either sense) in the setting considered in WB; identification requires assumptions that place substantial restrictions on the observed data distribution and on the structural equation models.

2.2. Semiparametric and Parametric Assumptions

Contrary to its statement, Theorem 6, which identifies the joint causal effect of all of the causes on Y , rests on the parametric assumptions that the confounding variable is a clustering indicator and that the treatment effects are constant across clusters (no treatment-confounder interaction). Furthermore, although it is not listed in the assumptions in the article, in order for $f_1(\mathbf{a}, x)$ and $f_2(z)$ to be jointly estimable even though z is a deterministic function of \mathbf{a} , Theorem 6 also requires f_1 to be more smooth than f_2 , for example, they cannot be collinear.

Theorem 7 identifies the causal effect of a subset of k out of the m causes, assuming overlap/positivity for those k causes: $P((A_1, \dots, A_k) \in \mathcal{A} | Z_i) > 0$ for any set \mathcal{A} such that $P(\mathcal{A}) > 0$. Because the conditioning event Z_i is a deterministic function of A_1, \dots, A_m , this is a stronger assumption than the classical overlap assumption, and it greatly restricts the possible functional forms for the deterministic function of \mathbf{A} that gives Z .

This restriction will be greatest when k is close to m . Two open questions are (1) whether these restrictions imply that the model for Z is degenerate as $m \rightarrow \infty$ and (2) whether they restrict the observed data distribution in addition to restricting the function of \mathbf{A} that gives Z . Neither of these concerns is addressed in the article, leaving open the possibility that the statement of the theorem might be vacuous, requiring parametric and/or additional causal assumptions in order for these conditions to be met.⁴ This framework, but with $k \ll m$ and the addition of parametric assumptions and exclusion restrictions (i.e. that most causes are null), is often used to test the effects of many SNPs in GWAS studies (e.g., Price et al. 2006; Gagnon-Bartsch, Jacob, and Speed 2013; Wang et al. 2017).

2.3. The Number of Causes Must Go to Infinity

The identification results in WB require *consistency of substitute confounders* (Definition 4 of WB), which generally holds asymptotically as the number of causes, m , goes to infinity. This is the case, for example, for probabilistic PCA and Poisson factorization, as discussed by WB and for which $(n + m) \log(nm)/(nm) \rightarrow 0$ ensures consistency. Consistency likely also requires either (a) a parametric factor model or (b) that a discrete variable with finite support suffices to control for confounding. It is not immediately clear what estimands are defined and identified in this limit, since Theorems 6–8 are written for finite m . Furthermore, it is not clear whether identification holds for any finite m . Of course, desirable frequentist properties for estimators of causal effects often require asymptotic arguments. However, in most settings that argument is required for estimation but not for identification; here an asymptotic limit in both the number of causes and the number of subjects is required for unmeasured confounding to be controlled for and therefore for identification.

However, the requirement that, in the limit, Z be a deterministic function of \mathbf{A} suggests that it cannot, in fact, control for confounding. This is because such a Z is independent of Y given \mathbf{A} , which is not true of confounders (see the analysis of Lemma 4). If causal effects are identifiable using such a Z , it must be because bias due to unmeasured confounding is estimable with a function of \mathbf{A} , and that function is not collinear with the causal effects themselves. In this case the method would have to rely for identification not on ignorability, but rather on an assumption that a biased, confounded effect and its bias are simultaneously identified.

3. Conclusion

One of the most important roles of causal inference in statistics and data science is to be transparent about the strong, usually untestable assumptions under which causal inference is possible (Pearl 2000; Robins 2001). The burden for transparency about assumptions is arguably greater in causal inference than in other areas of statistics, because it is crucial that scientists and consumers of research, for example, policy makers or doctors, have the tools to reason about whether an association is in fact causal. To that end, our best current understanding of when it is

⁴We are grateful to WB for pointing out that a previous version of this statement was imprecise.

justified to use a substitute confounder based on a factor analysis to estimate causal estimands in the presence of unmeasured confounding is under these conditions/assumptions (some of which are explicit in WB):

1. No unmeasured single-cause confounders.
2. SUTVA
3. No M structures exist between A and Y .
4. The causes are not causally dependent.
5. No post-treatment variables are captured by Z .
6. Unmeasured multi-cause confounding is due to a dependence or clustering structure that is common to each cause and to the outcome.
7. Z is consistent, which may rule out confounding altogether (see discussion above).
8. In the limit as the number of causes and the number of observations go to infinity.
9. One of the following:
 - (a) Confounding is due to a clustering indicator, treatment effects are constant in Z , continuous causes, and relative smoothness constraints on functions of the causes and of Z identify joint treatment effects of all of the causes (WB, Theorem 6).
 - (b) Overlap for some causes identifies treatment effects for those causes (WB, Theorem 7). This is at best a semiparametric assumption given the definition of Z in terms of the causes (see discussion of semiparametric and parametric assumptions above).
 - (c) Common values of Z identify conditional potential outcomes (WB, Theorem 8).

Some of these assumptions may be able to be relaxed or replaced with different assumptions, but unfortunately—we wish this were not the case!—it is impossible to identify causal effects in the presence of unmeasured confounding with nonparametric or empirically verifiable assumptions.

Acknowledgments

We would like to thank Alex D'Amour, Susan Murphy, and Zach Wood-Doughty for helpful discussions.

References

- Angrist, J. D., and Pischke, J.-S. (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, NJ: Princeton University Press. [1611]
- Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993), *Efficient and Adaptive Estimation for Semiparametric Models* (Vol. 4), Baltimore, MD: Johns Hopkins University Press. [1614]
- Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., and Poole, C. (2009), "Illustrating Bias due to Conditioning on a Collider," *International Journal of Epidemiology*, 39, 417–420. [1612]
- D'Amour, A. (2019), "On Multi-Cause Causal Inference With Unobserved Confounding: Counterexamples, Impossibility, and Alternatives," arXiv no. 1902.10286. [1613]
- Elwert, F., and Winship, C. (2014), "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable," *Annual Review of Sociology*, 40, 31–53. [1612]
- Gagnon-Bartsch, J. A., Jacob, L., and Speed, T. P. (2013), "Removing Unwanted Variation From High Dimensional Data With Negative Controls," Tech Reports from Dep Stat Univ California, Berkeley, pp. 1–112. [1614]
- Hernan, M. A., and Robins, J. M. (2020), *Causal Inference*, Boca Raton, FL: Chapman & Hall/CRC. [1612]
- Kallenberg, O. (1997), *Foundations of Modern Probability*, New York: Springer. [1613]
- Kuroki, M., and Pearl, J. (2014), "Measurement Bias and Effect Restoration in Causal Inference," *Biometrika*, 101, 423–437. [1611]
- Pearl, J. (2000), *Causality: Models, Reasoning and Inference* (Vol. 29), New York: Springer. [1614]
- Perkovic, E., Textor, J., Kalisch, M., and Maathuis, M. H. (2015), "A Complete Generalized Adjustment Criterion," in *Proceedings of the Thirty First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, AUAI Press. [1611]
- 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, 904. [1611, 1614]
- Robins, J. M. (2001), "Data, Design, and Background Knowledge in Etiologic Inference," *Epidemiology*, 12, 313–320. [1614]
- Shpitser, I., VanderWeele, T., and Robins, J. M. (2010), "On the Validity of Covariate Adjustment for Estimating Causal Effects," in *Proceedings of the Twenty Sixth Conference on Uncertainty in Artificial Intelligence (UAI-10)*, AUAI Press, pp. 527–536. [1611]
- Wang, J., Zhao, Q., Hastie, T., and Owen, A. B. (2017), "Confounder Adjustment in Multiple Hypothesis Testing," *The Annals of Statistics*, 45, 1863–1894. [1614]
- Wang, Y., and Blei, D. (2019), "Blessings of Multiple Causes," *Journal of the American Statistical Association*, 114, this issue, DOI: [10.1080/01621459.2019.1686987](https://doi.org/10.1080/01621459.2019.1686987). [1611]



The Blessings of Multiple Causes: Rejoinder

Yixin Wang & David M. Blei

To cite this article: Yixin Wang & David M. Blei (2019) The Blessings of Multiple Causes: Rejoinder, Journal of the American Statistical Association, 114:528, 1616-1619, DOI: [10.1080/01621459.2019.1690841](https://doi.org/10.1080/01621459.2019.1690841)

To link to this article: <https://doi.org/10.1080/01621459.2019.1690841>



Published online: 03 Jan 2020.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)



The Blessings of Multiple Causes: Rejoinder

Yixin Wang^a and David M. Blei^{a,b}

^aDepartment of Statistics, Columbia University, New York, NY; ^bDepartment of Computer Science, Columbia University, New York, NY

We thank all the discussants for taking the time and energy to build on this work; and we thank the editors for putting together an engaging and thought-provoking collection of discussions. After reading these contributions, we were struck that these are not mere discussions—indeed, each is an article in itself. This collection pushes forward “The Blessings of Multiple Causes” in significant ways, offering new theory, new criticism, and new application. After highlighting some of the themes of these articles, we will turn to each individually.

“The Blessings of Multiple Causes” provide assumptions, theory, and algorithms for multiple causal inference. The deconfounder algorithm involves modeling the causes, using the model to infer a substitute confounder, and then using the substitute confounder in a downstream causal inference. The deconfounder is not a black-box solution to causal inference. Rather, it is a way to use careful domain-specific modeling in the service of causal inference.

Causal inference with the deconfounder involves a number of assumptions and trade-offs, and many of the discussants highlighted these. Among them are the following. (1) There can be no unobserved single-cause confounders. (2) When we apply the deconfounder, we trade an increase in estimation variance for a reduction in confounding bias; there is no free lunch. (3) We do not recommend using the deconfounder with causally dependent causes, such as a time series; finding a substitute confounder may be too difficult in these scenarios.

There are many directions for further research, and the discussants have pointed out several of the most important ones. We need a more complete picture of identification; D’Amour (2019) and the discussions here make good progress (see Table 1). We need to understand the finite-sample properties of the deconfounder, and how to estimate uncertainty about causal inferences when using a substitute multi-cause confounder. We need rigorous methods of model criticism for assessing the validity of the substitute confounder.

Deconfounder-like methods have already been used for genome-wide association studies (e.g., Pritchard et al. 2000) and estimating peer effects in networks (Shalizi and McFowland III 2016). More broadly, the deconfounder strategy points to many applications, including in genetics, psychology, education, and marketing, where factor models are routinely fit to large-scale data. We hope that statisticians and machine learners will

continue to study multiple causality, and that scientists and other practitioners will adapt the deconfounder to help analyze and understand their observational data.

1. Athey, Imbens, and Pollmann

Athey, Imbens, and Pollmann (AIP) consider a problem in economics: how do the prices of products affect their demand? The causes are prices; the outcome is demand; and the unobserved confounders are shocks to demand that also affect price. AIP apply the deconfounder to a setting of two products, gracefully using their domain expertise to directly construct a substitute confounder. They show that the deconfounder only helps when the two products have highly correlated demand shocks, that is, when there is shared unobserved confounding. AIP’s application beautifully illustrates the importance of domain knowledge to the deconfounder.

AIP compare two methods for estimating causal effects. The deconfounder uses the average price of the two products as the substitute confounder; the instrumental variable approach uses the price of one product as an instrument for the other. It is interesting that these two strategies work best in opposite cases. The deconfounder works when the demand shocks of the two products are highly correlated. The instrumental variable works when they are not (or weakly) correlated. More precisely, both the deconfounder and the instrumental variable approach require the prices be correlated. But the deconfounder requires that the driver of this correlation also affects the outcome, while the instrumental variable approach requires that it not affect the outcome.

AIP’s method further suggests extending the deconfounder to more general structures of shared confounding. Unlike the simpler settings in the article, AIP examine multiple causal problems: each product’s price affects a different outcome, but with shared unobserved confounders. We imagine that other scientific settings bear the same parallel structure.

2. Imai and Jiang

Imai and Jiang (IJ) discuss two technical aspects of the deconfounder.

Table 1. Identification in multiple causal inference.

Causal quantity	Result	Condition	Source
$P(Y(\mathbf{a}))$	Non-ID	No conditions	D'Amour (2019)
$\mathbb{E}[Y(\mathbf{a})] - \mathbb{E}[Y(\mathbf{a}')]]$	ID	Consistent substitute confounder; Categorical substitute confounder; No confounder/cause interaction; Differentiable relationships	Theorem 6 (WB)
$\mathbb{E}_A[\mathbb{E}_Y[Y(a_{1:k}, A_{(k+1):m})]]$	ID	Consistent substitute confounder; $A_{1:k}$ satisfy overlap	Theorem 7 (WB)
$\mathbb{E}[Y(\mathbf{a}') \mathbf{A} = \mathbf{a}]$	ID	Consistent substitute confounder; \mathbf{a}' and \mathbf{a} map to same substitute	Theorem 8 (WB)
$\mathbb{E}[Y(\mathbf{a})]$	ID	$\mathbb{E}[U \mathbf{A}]$ nonlinear; $\mathbb{E}[Y \mathbf{A}, U]$ linear	Section 2.1 (IJ)
$\mathbb{E}[Y(\mathbf{a})]$	ID	Measure instrument W ; Instrument W satisfies overlap	Section 2.2 (IJ)
$\int Y(\mathbf{a})q_1(\mathbf{a}) d\mathbf{a} - \int Y(\mathbf{a})q_2(\mathbf{a}) d\mathbf{a}$	ID	$p(\mathbf{a} \mathbf{z}) > 0$ when $q_1(\mathbf{a}), q_2(\mathbf{a}) > 0$	Section 2.3 (IJ)

NOTE: ID = identifiable.

They first point out the difficulty of defining “multi-cause” and in particular of defining the assumption “no unobserved single-cause confounders.” In the DAG language, this assumption requires (1) the causal graph resides in a class where unobserved confounders must be parents of two or more causes and (2) the causal problem be faithful to the graph (Spirtes 2010).

We agree with IJ that it is difficult to simultaneously express such graphical and faithfulness conditions in the potential outcomes notation. The definition in the article attempts to express faithfulness by considering the smallest sigma algebra that renders the causes conditionally independent (see condition 2 in Definition 4). Note this definition excludes those multi-cause confounders that can be separated into multiple single-cause confounders, as illustrated in Figure 2 of IJ’s article.

IJ correctly note that it is unclear whether the definition of “no unobserved single-cause confounders” in the article is equivalent to the one we intended in the DAG language. As suggested by IJ at JSM 2019, a more precise form of “no unobserved single-cause confounders” may be: *there exist a random variable Z s.t. (1) Z satisfies $p(\mathbf{a} | \mathbf{z}) = \prod_{j=1}^m p(a_j | \mathbf{z})$ and no sigma-algebra smaller than $\sigma(Z)$ satisfies this equation; (2) $A_1, \dots, A_m \perp Y(\mathbf{a}) | Z$.* Moreover, assessing the credibility of “no unobserved single-cause confounders” may require substantial domain expertise. How to rigorously translate graphical and faithfulness conditions into the potential outcomes notation is an interesting direction of research.

The second thread of IJ’s article is about causal identification of a complete intervention $\mathbb{E}[Y(\mathbf{a})]$, and the difficulty of the deconfounder in satisfying overlap. Because the substitute confounder Z is a function of the causes \mathbf{A} , the overlap condition $P(\mathbf{A} | Z) > 0$ can be stringent. IJ consider three ways forward: parametric assumptions, instrumental variables, and stochastic interventions.

IJ’s parametric approach achieves the identification of $\mathbb{E}[Y(\mathbf{a})]$ by the incongruence between the linear outcome model and the nonlinear factor model. Related to IJ’s setting, Theorem 6 in the article achieves identification via the incongruence between the differentiability of the outcome model and the non-differentiability of the substitute confounder. IJ’s result and Theorem 6 suggest that the idea of incongruence may serve as a general approach to causal identification.

IJ’s instrumental variable approach requires an overlap condition that is weaker than the one required by the deconfounder. But, as IJ illustrate in their discrete-variable example, this overlap condition may become more stringent as the number of

causes increases. Notice there may be an increasing number of instrumental variables as the number of causes increases, though not one of them might satisfy overlap by itself. IJ’s thinking is suggestive of a direction of future investigation: how to combine multiple instrumental variables in multi-cause problems to satisfy overlap and obtain causal identification.

The final approach IJ explore is stochastic intervention. It tackles the problem of overlap by restricting the causal queries. This approach relates to Theorem 8 in the article, which restricts the causal queries to those interventions that map to the same value of the substitute confounder. But IJ’s approach is more powerful than Theorem 8 because it handles causally dependent causes. We look forward to more developments in the stochastic intervention approach of multiple causal inference.

Including IJ’s new results, Table 1 summarizes the current landscape of identification results in multiple causality.

3. D’Amour

In both his discussion here and his earlier article (D’Amour 2019), Alex D’Amour has significantly contributed to the understanding of multi-causal identification. We have enjoyed a productive conversation with him over the past years. We were glad to read that the feeling is mutual.

In his discussion, D’Amour 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, that is, that $P(\mathbf{A} | Z) > 0$. As D’Amour (2019) points out, both cannot be simultaneously satisfied.

Theorems 6–8 in the article 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 7.) The pinpointed substitute confounder is achievable thanks to the multiplicity of the causes and the consistent estimability of factor models. Going forward, how does identification fare as we move away from this extreme? Point identification might no longer be possible, but partial identification might be.

With the same assumptions as Theorems 6–8, D'Amour studies both parametric and nonparametric identification. The parametric direction is risky without strong prior knowledge. But certain applications enjoy parametric models that are worth studying. For example, when we believe causal effects are small, a structural model that is linear in the causes but nonparametric in the unobserved confounder may be reasonably close to reality, $Y = \sum_{j=1}^m \beta_j A_j + g(U) + \varepsilon$. Identification conditions for such parametric models can be convenient for practical applications.

In the nonparametric direction, D'Amour explores Theorems 7 and 8 of the article. D'Amour's Proposition 1 summarizes well the essence of the theorems. Toward a more cautious application of the deconfounder, he suggests performing conditional independence tests or sensitivity analysis. This is an important direction of investigation and could be useful in many scientific domains.

4. Ogburn, Shpitser, and Tchetgen Tchetgen

Ogburn, Shpitser, and Tchetgen Tchetgen (OSTT) provide a technical meditation on some of the theoretical aspects of the article, and a dissenting opinion about its value. 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.

The identification results in Theorems 6–8 capitalize on two requirements: (1) the distribution of the causes $p(\mathbf{a})$ can be described by a factor model and (2) the factor model pinpoints the substitute confounder Z , that is, $Z \stackrel{\text{a.s.}}{=} f_\theta(\mathbf{A})$ for some f_θ . The first requirement relies on the successful execution of the deconfounder, that is, finding a factor model that captures $p(\mathbf{a})$. 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 the “consistency of the substitute confounder.” 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.

OSTT's main concern revolves around Lemma 4, which states the substitute confounder cannot pick up information about multi-cause mediators, single-cause mediators, or any of the other graphs that OSTT put forward. Lemma 4 is correct, as is the proof in the article. But Lemma 4 might also seem surprising. Here is an alternative proof.

Restatement of Lemma 4. No post-treatment variable can be measurable with respect to a consistent substitute confounder.

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

The substitute also cannot pick up any single-cause variables. These variables include pretreatment 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 $Z \stackrel{\text{a.s.}}{=} f(\mathbf{A}; \theta)$, as is the case where there are many causes. The deconfounder further requires that the converse is not true, that is, 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 of causes increases. Further, the deconfounder requires that the factor model cannot 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,” that is, nonprobabilistic.

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 consistent substitute confounder says

$$p(z | \mathbf{a}, \theta) = p(z_s, z_m | \mathbf{a}, \theta) = \delta_{(f_s(\mathbf{a}; \theta), f_m(\mathbf{a}; \theta))}, \quad (1)$$

where $\mathbf{a} = (a_1, \dots, a_m)$ 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,

$$p(z_s | \mathbf{a}) = p(z_s | \mathbf{a}, z_m = f_m(\mathbf{a}; \theta)), \quad (2)$$

$$= p(z_s | a_1, z_m = f_m(\mathbf{a}; \theta)), \quad (3)$$

$$= \frac{p(z_s | z_m = f_m(\mathbf{a}; \theta)) \cdot p(a_1 | z_s, z_m = f_m(\mathbf{a}; \theta))}{p(a_1 | z_m = f_m(\mathbf{a}; \theta))}. \quad (4)$$

Equation (2) is due to the consistency of substitute confounder. Equation (3) is due to $Z_s \perp A_2, \dots, A_m | A_1, Z_m$. Equation (4) is due to the definition of conditional probability.

Equation (4) and Equation (1) imply that at least one of $p(z_s | z_m = f_m(\mathbf{a}; \theta))$ and $p(a_1 | z_s, z_m = f_m(\mathbf{a}; \theta))$ 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

$$p(z_m | \mathbf{a}) = p(z_m | \mathbf{a}, z_s = f_s(\mathbf{a}; \theta)), \quad (5)$$

$$= \frac{p(a_1, z_m | z_s = f_s(\mathbf{a}; \theta)) \cdot \prod_{j=2}^m p(a_j | z_m)}{p(\mathbf{a})}, \quad (6)$$

where $\prod_{j=2}^m p(a_j | z_m)$ can converge to a point mass with non-degenerate factor models and $m \rightarrow \infty$. \square

OSTT also question the random variable on which we used the Kallenberg construction in Lemmas 1 and 2. Definition 3 is the Kallenberg construction we intended, and it involves

potential outcomes (see Equation (38) in the article). Lemmas 1 and 2 link factor models of the causes to their Kallenberg construction and unconfoundedness, thanks to the consistency of the substitute confounder. Such a substitute cannot separate a multi-cause confounder into single-cause confounders, as the one in OSTT's counterexample does. OSTT claim that the article leaves open that Theorem 7 is "vacuous" because the overlap condition may be impossible to satisfy. D'Amour's discussion of the article shows how Theorem 7 can be useful.

Finally, OSTT remark that requiring a pinpointed substitute implies that the unobserved (multi-cause) confounding is effectively observed. Their intuition is correct—the multiplicity of the causes and the consistent estimability of factor models enable us to effectively observe such multi-cause confounding. It is these two features that form the basis of the deconfounder.

Acknowledgments

We thank Gemma Moran, Aaron Schein, Keyon Vafa, Victor Veitch, and Alex D'Amour for helpful discussions.

References

- D'Amour, A. (2019), "On Multi-Cause Approaches to Causal Inference With Unobserved Counfounding: Two Cautionary Failure Cases and a Promising Alternative," in *Artificial Intelligence and Statistics*, pp. 3478–3486. [1616,1617]
- Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000), "Association Mapping in Structured Populations," *The American Journal of Human Genetics*, 67, 170–181. [1616]
- Shalizi, C. R., and McFowland III, E. (2016), "Estimating Causal Peer Influence in Homophilous Social Networks by Inferring Latent Locations," arXiv no. 1607.06565. [1616]
- Spirtes, P. (2010), "Introduction to Causal Inference," *Journal of Machine Learning Research*, 11, 1643–1662. [1617]

Supplementary Material: The Blessings of Multiple Causes

A Connections to genome-wide association studies

Many methods from the research literature, especially around genome-wide association studies, can be reinterpreted as instances of the deconfounder algorithm. Each can be seen as positing a factor model of assigned causes (Section 4.1) and a conditional outcome model (Section 4.2).

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 5.

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 4.3.

Linear mixed models. The LMM is one the most popular classes of methods for analyzing GWAS (Yu et al., 2006; Kang et al., 2008; Yang et al., 2014; Lippert et al., 2011; Loh et al., 2015; Darnell et al., 2017). 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 $\mathbf{a}_{1:n}$. However, this is mathematically equivalent to posterior latent factors from a one-dimensional principal component analysis (PCA) model. Subject to its capturing the distribution of SNPs, the LMM is performing multiple causal inference with a 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 Eq. 9. Fitting an outcome model to its residuals is equivalent to conditioning on the reconstructed assignments, Eq. 21.

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

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

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 $\alpha(\hat{z}_i)$, again Eq. 21, 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 a 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 ([Pritchard et al., 2000a,b](#); [Falush et al., 2003, 2007](#)). 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.

B Can the causes be causally dependent among themselves?

When the causes are causally dependent, the deconfounder can still provide unbiased estimates of the potential outcomes. Its success relies on a valid substitute confounder.

Note there are cases where a valid substitute confounder cannot exist. 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 $Z \stackrel{a.s.}{=} A_1$ or $Z \stackrel{a.s.}{=} A_2$, because it needs to render the two causes conditionally independent. But such a Z does not satisfy overlap.

On the other hand, causal dependence among the causes does not necessarily imply the nonexistence of a valid substitute confounder. Consider a different mechanism for the causal relationship between A_1 and A_2 ,

$$\begin{aligned} A_1 &\sim \mathcal{N}(0, 1), \\ A_2 &= |A_1| + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1). \end{aligned}$$

Here $Z \stackrel{a.s.}{=} |A_1|$ is a valid substitute confounder; it satisfies overlap and renders A_1 conditionally independent of A_2 .

Empirically, it is hard to detect the nonexistence of a valid substitute confounder without knowing the functional form of how the causes are structurally dependent. Insisting on using the deconfounder in this case results in limited overlap and high variance causal estimates downstream.

To illustrate this phenomenon, we repeat the experiments in Section 6.1 with the same confounder a_{age} but three causes: $a_{\text{mar}}, a_{\text{exp}}$ and an additional cause $a_{\text{mar+}}$. We assume $a_{\text{mar+}}$ causally depend on a_{mar} , where

$$a_{\text{mar+}} = a_{\text{mar}} + \epsilon_{i,\text{mar+}}, \quad \epsilon_{i,\text{mar+}} \sim \mathcal{N}(0, 0.1^2). \quad (43)$$

	Check	Bias ² × 10 ⁻²	Variance × 10 ⁻²	MSE × 10 ⁻²
No control	–	41.89	0.01	41.90
Control for age (oracle)	–	22.57	0.01	22.57
Control for 1-dim z_{line}	✓	29.98	16.97	46.96
Control for 1-dim $a(z_{\text{line}})$	✓	28.01	18.49	46.50
Control for 1-dim z_{quad}	✓	25.10	16.70	41.80
Control for 1-dim $a(z_{\text{quad}})$	✓	27.46	15.77	43.23

Table 5: Total bias and variance of the estimated causal coefficients β_{exp} and β_{mar} when there is a third cause dependent on a_{mar} . The nonlinear factor model outperforms linear factor model. The deconfounder estimate has much higher variance than usual (e.g., Table 4) when two of the causes are dependent.

It implies that theoretically there exists no substitute confounders that can both satisfy overlap and render the causes conditionally independent.

We simulate the outcome from

$$y_i = \beta_{\text{mar}} a_{\text{mar},i} + \beta_{\text{exp}} a_{\text{exp},i} + \beta_{\text{age}} a_{\text{age},i} + \beta_{\text{mar+}} a_{\text{mar+},i} + \varepsilon_i, \quad (44)$$

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

$$\beta_{\text{mar}} \sim \mathcal{N}(0, 1) \quad \beta_{\text{exp}} \sim \mathcal{N}(0, 1) \quad \beta_{\text{age}} \sim \mathcal{N}(0, 1) \quad \beta_{\text{mar+}} \sim \mathcal{N}(0, 1). \quad (45)$$

Nevertheless, we apply the deconfounder to this data. We model the three causes with one-dimensional linear and quadratic factor model; both pass the predictive check, with a predictive score of 0.28 and 0.20. Table 5 shows the bias and variance of the deconfounder estimate of β_{mar} and β_{exp} . With causally dependent causes (Table 5), the deconfounder estimates have much larger variance than usual (Table 4); it signals that the substitute confounder we constructed is close to breaking overlap. That said, the deconfounder is still able to correct for a substantial portion of confounding bias.

Finally, we recommend applying the deconfounder to non-causally dependent causes. A valid substitute confounder is guaranteed to exist in this case; it will both satisfy overlap and render the causes conditionally independent of each other.

C 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:

$$Z = \epsilon_Z, \quad (46)$$

$$A_1 = \alpha_{10} + \alpha_{11}Z + \alpha_{12}Z^2 + \epsilon_{A1}, \quad (47)$$

$$A_2 = \alpha_{20} + \alpha_{21}Z + \alpha_{22}Z^2 + \epsilon_{A2}, \quad (48)$$

$$Y = \beta_0 + \beta_1A_1 + \beta_2A_2 + \gamma Z + \epsilon_Y, \quad (49)$$

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 :

$$\begin{aligned} Y &= (\beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20}) + (\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma) \cdot Z + (\beta_1\alpha_{12} + \beta_2\alpha_{22}) \cdot Z^2 + \beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y, \\ &= (\beta_1\alpha_{12} + \beta_2\alpha_{22}) \cdot \left(Z + \frac{\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma}{2 \cdot (\beta_1\alpha_{12} + \beta_2\alpha_{22})} \right)^2 + \beta_1\epsilon_{A1} + \beta_2\epsilon_{A2} + \epsilon_Y \\ &\quad + \left(\beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20} - \left(\frac{\beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma}{2 \cdot (\beta_1\alpha_{12} + \beta_2\alpha_{22})} \right)^2 \right) \end{aligned}$$

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:

$$B_0 \triangleq \beta_0 + \beta_1\alpha_{10} + \beta_2\alpha_{20}, \quad (50)$$

$$B_1 \triangleq \beta_1\alpha_{11} + \beta_2\alpha_{21} + \gamma, \quad (51)$$

$$B_2 \triangleq \beta_1\alpha_{12} + \beta_2\alpha_{22}. \quad (52)$$

Therefore, we have

$$Y = B_0 + B_1 \cdot Z + B_2 \cdot Z^2 + \epsilon_Y, \quad (53)$$

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

$$M_{A_1, A_2, Y}(t_1, t_2, t_3) \quad (54)$$

$$= \mathbb{E}[\exp(t_1A_1 + t_2A_2 + t_3Y)] \quad (55)$$

$$= \exp(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2) \quad (56)$$

$$\cdot \mathbb{E} \left[\exp \left((\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3) \cdot Z + (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot Z^2 \right) \right] \quad (57)$$

$$\cdot \mathbb{E} \left[t_1 \epsilon_{A1} + t_2 \epsilon_{A2} + t_3 (\beta_1 \epsilon_{A1} + \beta_2 \epsilon_{A2} + \epsilon_Y) \right] \quad (58)$$

$$= \exp \left(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2 - (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)} \right)^2 \right) \quad (59)$$

$$\cdot \mathbb{E} \left[\exp \left((\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \sigma_Z^2 \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \sigma_Z} + \frac{Z}{\sigma_Z} \right)^2 \right) \right] \quad (60)$$

$$\cdot \mathbb{E} \left[t_1 \epsilon_{A1} + t_2 \epsilon_{A2} + t_3 (\beta_1 \epsilon_{A1} + \beta_2 \epsilon_{A2} + \epsilon_Y) \right] \quad (61)$$

$$= \exp \left(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2 - (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)} \right)^2 \right) \quad (62)$$

$$\cdot \frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{1/2}} \quad (63)$$

$$\cdot \exp\left(\frac{1}{2}(t_1 + t_3 \beta_1)^2 \sigma_{A_1}^2\right) \exp\left(\frac{1}{2}(t_2 + t_3 \beta_2)^2 \sigma_{A_2}^2\right) \exp\left(\frac{1}{2} t_3 \sigma_Y^2\right) \quad (64)$$

$$= \exp \left(B_0 t_3 + \alpha_{10} t_1 + \alpha_{20} t_2 - (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \cdot \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3)} \right)^2 \right) \quad (65)$$

$$\cdot \frac{\exp(\frac{\lambda t}{1-2t})}{(1-2t)^{1/2}} \quad (66)$$

$$\cdot \exp\left(\frac{1}{2}(t_1 \sigma_{A_1}^2 + \beta_1^2 \sigma_{A_1}^2 t_3^2 + 2\beta_1 \sigma_{A_1}^2 t_1 t_3 + t_2 \sigma_{A_2}^2 + \beta_2^2 \sigma_{A_2}^2 t_3^2 + 2\beta_2 \sigma_{A_2}^2 t_2 t_3 + \sigma_Y^2 t_3)\right), \quad (67)$$

where $t = (\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \sigma_Z^2$ and $\lambda = \left(\frac{\alpha_{11} t_1 + \alpha_{21} t_2 + B_1 t_3}{2(\alpha_{12} t_1 + \alpha_{22} t_2 + B_2 t_3) \sigma_Z} \right)^2$.

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 Eqs. 65 and 66 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.

D Detailed Results of the GWAS Study

In this section, we present tables of results from the GWAS study in Section 6.2.

Tables 6 to 10 contain the result under the high SNR setting.

		Real-valued outcome	Binary outcome
	Pred. check	RMSE $\times 10^{-2}$	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 6: 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.

Tables 11 to 15 contain the result under the low SNR setting.

E Detailed Results of the Movie Study

In this section, we present tables of results from the movies study in Section 6.3.

F Proof of Lemma 1

Proof sketch. First assume the Kallenberg construction in Eq. 37. This form shows that the assigned causes (A_{i1}, \dots, A_{im}) are captured by functions of Z_i and randomization variables U_{ij} . This fact, in turn, implies that the randomness in $(A_{i1}, \dots, A_{im}) | Z_i$ comes from the randomization variables which are (by definition) independent of $Y_i(\mathbf{a})$. Therefore (A_{i1}, \dots, A_{im}) is conditionally independent of Y_i given Z_i , i.e., unconfoundedness holds. Now assume that unconfoundedness holds. We prove that this assumption implies a Kallenberg construction by building on the randomization variable construction of conditional distributions (Kallenberg, 1997). \square

Proof. For notation simplicity, we suppress the i subscript in this proof.

We assume \mathcal{Z} is a measurable space and $\mathcal{A}_j, j = 1, \dots, m$ are Borel spaces.

		Real-valued outcome	Binary outcome
	Pred. check	RMSE $\times 10^{-2}$	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 7: 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.

We first prove the necessity. Assume that $A_j = f_j(Z, U_j), j = 1, \dots, m$, where $f_j, j = 1, \dots, m$ are measurable and

$$(U_1, \dots, U_m) \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m)) \quad (68)$$

for all (a_1, \dots, a_m) . By Proposition 5.18 in [Kallenberg \(1997\)](#), Eq. 68 implies

$$(U_1, \dots, U_m) \perp\!\!\!\perp_Z Y(a_1, \dots, a_m),$$

and so

$$(Z, U_1, \dots, U_m) \perp\!\!\!\perp_Z Y(a_1, \dots, a_m)$$

by Corollary 5.7 in [Kallenberg \(1997\)](#). It implies

$$(A_1, \dots, A_m) \perp\!\!\!\perp_Z Y(a_1, \dots, a_m)$$

for all $(a_1, \dots, a_m) \in \mathcal{A}_1 \otimes \dots \otimes \mathcal{A}_m$. The last step is because A_j ’s are measurable functions of (Z, U_1, \dots, U_m) .

Now we prove the sufficiency. Assume that $Y(a_1, \dots, a_m) \perp\!\!\!\perp_Z (A_1, \dots, A_m)$. Marginalizing out all but one A_j gives

$$Y(a_1, \dots, a_m) \perp\!\!\!\perp_Z A_j, j = 1, \dots, m.$$

By Theorem 5.10 in [Kallenberg \(1997\)](#), there exists a measurable function $f_j : \mathcal{Z} \times [0, 1] \rightarrow \mathcal{A}_j$ and a Uniform[0,1] random variable \tilde{U}_j satisfying $\tilde{U}_j \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m))$ such that the random variable $\tilde{A}_j = f_j(Z, \tilde{U}_j)$ satisfies

$$\tilde{A}_j \stackrel{d}{=} A_j \text{ and } (\tilde{A}_j, Z) \stackrel{d}{=} (A_j, Z).$$

Moreover, we have

$$\tilde{A}_j \perp\!\!\!\perp_Z Y(a_1, \dots, a_m)$$

with the same argument as the above necessity part.

		Real-valued outcome	Binary outcome
	Pred. check	RMSE $\times 10^{-2}$	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 8: 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.

Hence, by Proposition 5.6 in [Kallenberg \(1997\)](#),

$$P(\tilde{A}_j \in \cdot \mid Z, Y(a_1, \dots, a_m)) = P(\tilde{A}_j \in \cdot \mid Z) = P(A_j \in \cdot \mid Z) = P(A_j \in \cdot \mid Z, Y(a_1, \dots, a_m)),$$

and so

$$(\tilde{A}_j, Z, Y(a_1, \dots, a_m)) \stackrel{d}{=} (A_j, Z, Y(a_1, \dots, a_m)).$$

By Theorem 5.10 in [Kallenberg \(1997\)](#), we may choose some random variable U_j such that

$$U_j \stackrel{d}{=} \tilde{U}_j \text{ and } (\tilde{A}_j, Z, Y(a_1, \dots, a_m), U_j) \stackrel{d}{=} (A_j, Z, Y(a_1, \dots, a_m), \tilde{U}_j).$$

In particular, we have

$$U_j \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m))$$

and

$$(A_j, f_j(Z, U_j)) \stackrel{d}{=} (\tilde{A}_j, f_j(Z, \tilde{U}_j)).$$

Since

$$\tilde{A}_j = f_j(Z, \tilde{U}_j)$$

and the diagonal in S^2 is measurable, we have

$$A_j \stackrel{a.s.}{=} f_j(Z, U_j).$$

We then show $(U_1, \dots, U_m) \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m))$. By Theorem 5.10 in [Kallenberg \(1997\)](#), there exists a measurable function $g_1 : \mathcal{Y} \times \mathcal{Z} \times [0, 1] \rightarrow [0, 1]$ and a Uniform[0,1] random variable \hat{U}_1 satisfying $\hat{U}_1 \perp\!\!\!\perp (Y(a_1, \dots, a_m), Z)$ and

$$(Y(a_1, \dots, a_m), Z, U_1) \stackrel{d}{=} (Y(a_1, \dots, a_m), Z, g_1(Y(a_1, \dots, a_m), Z, \hat{U}_1)).$$

Moreover, by

$$U_1 \perp\!\!\!\perp ZY(a_1, \dots, a_m),$$

we have

$$g_1(Y(a_1, \dots, a_m), Z, \hat{U}_1) \perp\!\!\!\perp_Z Y(a_1, \dots, a_m)$$

there exists some measurable function $g'_1 : \mathcal{Z} \times [0, 1] \rightarrow [0, 1]$ such that

$$g_1(Y(a_1, \dots, a_m), Z, \hat{U}_1) = g'_1(Z, \hat{U}_1)$$

and

$$\hat{U}_1 \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m)).$$

In other words, we have

$$(Y(a_1, \dots, a_m), Z, U_1) \stackrel{d}{=} (Y(a_1, \dots, a_m), Z, g'_1(Z, \hat{U}_1)).$$

Repeating these steps, we again have from Theorem 5.10 in [Kallenberg \(1997\)](#) that there exists a measurable function $g_2 : \mathcal{Y} \times \mathcal{Z} \times [0, 1]^2 \rightarrow [0, 1]$ and a Uniform[0,1] random variable \hat{U}_2 satisfying

$$\begin{aligned} & (Y(a_1, \dots, a_m), Z, U_1, U_2) \\ & \stackrel{d}{=} (Y(a_1, \dots, a_m), Z, g'_1(Z, \hat{U}_1), g_2(Y(a_1, \dots, a_m), Z, \hat{U}_1, \hat{U}_2)) \end{aligned}$$

and

$$\hat{U}_2 \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m), \hat{U}_1).$$

Again by

$$U_1 \perp\!\!\!\perp_Z Y(a_1, \dots, a_m),$$

we have a measurable function $g'_2 : \mathcal{Z} \times [0, 1]^2 \rightarrow [0, 1]$ that satisfies

$$\begin{aligned} & (Y(a_1, \dots, a_m), Z, U_1, U_2) \\ & \stackrel{d}{=} (Y(a_1, \dots, a_m), Z, g'_1(Z, \hat{U}_1), g'_2(Z, \hat{U}_1, \hat{U}_2)). \end{aligned}$$

Repeating these steps m times, we have

$$\begin{aligned} & (Y(a_1, \dots, a_m), Z, U_1, U_2, \dots, U_m) \\ & \stackrel{d}{=} (Y(a_1, \dots, a_m), Z, g'_1(Z, \hat{U}_1), g'_2(Z, \hat{U}_1, \hat{U}_2), \dots, g'_m(Z, \hat{U}_1, \hat{U}_2, \dots, \hat{U}_m)) \end{aligned}$$

with

$$\hat{U}_j \perp\!\!\!\perp (Z, Y(a_1, \dots, a_m), \hat{U}_1, \dots, \hat{U}_{j-1}), j = 1, \dots, m.$$

We notice that the right side of the equation have conditional independence property

$$(g'_1(Z, \hat{U}_1), g'_2(Z, \hat{U}_1, \hat{U}_2), \dots, g'_m(Z, \hat{U}_1, \hat{U}_2, \dots, \hat{U}_m)) \perp\!\!\!\perp_Z Y(a_1, \dots, a_m).$$

This implies the same property holds for the left side of the equation, that is

$$(U_1, \dots, U_m) \perp\!\!\!\perp_Z Y(a_1, \dots, a_m).$$

□

G Proof of Lemma 2

Proof sketch. The lemma is an immediate consequence of Lemma 2.22 in [Kallenberg \(1997\)](#) and “no unobserved single-cause confounders”. We also rely $p(\theta_{1:m})$ and $p(z_i | \mathbf{a}_i)$ are point masses, so they are *a priori* independent of the potential outcomes and the other latent variables. \square

Proof. For simplicity, we consider continuous random variables A_{ij}, Z_i, θ_j . Also, we assume there are no single-cause confounders. The proof can be easily extended to accommodate discrete random variables and observed single-cause confounders.

We first state the regularity condition: The domains of the causes, \mathcal{A}_j , $j = 1, \dots, m$ are Borel subsets of compact intervals. Without loss of generality, we could assume $\mathcal{A}_j = [0, 1]$, $j = 1, \dots, m$.

By Lemma 2.22 in [Kallenberg \(1997\)](#), there exists some measurable function $f_j : \mathcal{Z} \times [0, 1] \rightarrow [0, 1]$ such that $\gamma_{ij} \perp\!\!\!\perp Z_i$ and

$$A_{ij} = f_j(Z_i, \gamma_{ij}).$$

Furthermore, there exists some measurable function $h_{ij} : \Theta \times [0, 1] \rightarrow [0, 1]$ such that

$$\gamma_{ij} = h_{ij}(\theta_j, \omega_{ij}),$$

where $\omega_{ij} \perp\!\!\!\perp (Z_i, \theta_j)$ and $\omega_{ij} \sim \text{Uniform}[0, 1]$. Lastly, we write

$$U_{ij} = F_{ij}^{-1}(\gamma_{ij}) \sim \text{Uniform}[0, 1],$$

where F_{ij} is the cumulative distribution function of γ_{ij} .

Eq. 35 implies that $\omega_{ij}, i = 1, \dots, n, j = 1, \dots, m$ are jointly independent: if they were not, then $A_{ij} = f_j(Z_i, h_{ij}(\theta_j, \omega_{ij}))$ would not have been conditionally independent given Z_i, θ_j .

We thus have

$$A_{ij} = f_j(Z_i, U_{ij}),$$

where $U_{ij} := F_{ij}^{-1}(h_{ij}(\theta_j, \omega_{ij}))$.

Below we will prove that U_{ij} satisfies

$$(U_{i1}, \dots, U_{im}) \perp\!\!\!\perp (Z_i, Y_i(a_1, \dots, a_m)). \quad (69)$$

We will rely on the “no single-cause confounders” assumption and the consistency of substitute confounder assumption $p(z_i | \mathbf{a}_i) = \delta_{f_\theta(\mathbf{a}_i)}$.

First, we notice that $\theta_{1:m}$ are point masses; they satisfy $(\theta_1, \dots, \theta_m) \perp\!\!\!\perp (Z_i, Y_i(a_1, \dots, a_m))$.

Next, we notice that the “no single-cause confounders” assumption implies that there exists a random variable \tilde{Z}_i such that

$$p(a_{i1}, \dots, a_{im} | \tilde{z}_i) = \prod_{j=1}^m p(a_{ij} | \tilde{z}_i) \quad (70)$$

and

$$A_{i1}, \dots, A_{im} \perp Y_i(a_1, \dots, a_m) | \tilde{Z}_i. \quad (71)$$

Moreover, no sigma algebra smaller than \tilde{Z}_i satisfies Eq. 70. Further, the consistency of substitute confounder assumption $Z_i = f_{\theta}(\mathbf{A}_i)$ required for the factor model implies that the \tilde{Z}_i that satisfies Eq. 70 is unique, i.e. $\tilde{Z}_i \stackrel{a.s.}{=} Z_i$. The reason is that the consistency of substitute confounder assumption implies

$$p(\mathbf{a}_i, z_i) = p(\mathbf{a}_i)p(z_i | \mathbf{a}_i) = p(\mathbf{a}_i) \cdot \delta_{f_{\theta}(\mathbf{a}_i)},$$

which is a function of $p(\mathbf{a}_i)$ by construction. This is a key step that illustrates how the consistency of substitute confounder assumption interacts with the no single-cause confounder assumption to provide causal identification. Hence, Z_i also satisfies the unconfoundedness condition Eq. 71, which implies Eq. 69 and also

$$(\omega_{i1}, \dots, \omega_{im}) \perp (Y_i(a_1, \dots, a_m), Z_i)$$

or equivalently, $(\omega_{i1}, \dots, \omega_{im}) \perp Y_i(a_1, \dots, a_m) | Z_i$. In particular, for $m = 2$, we have

$$\begin{aligned} & p(Y_i(a_1, \dots, a_m), \omega_{i1}, \omega_{i2} | Z_i) \\ &= p(\omega_{i1} | Z_i) \cdot p(Y_i(a_1, \dots, a_m) | \omega_{i1}, Z_i) \cdot p(\omega_{i2} | \omega_{i1}, Y_i(a_1, \dots, a_m), Z_i) \\ &= p(\omega_{i1} | Z_i) \cdot p(Y_i(a_1, \dots, a_m) | Z_i) \cdot p(\omega_{i2} | Z_i) \end{aligned}$$

Finally, this argument illustrates how the “no single-cause confounders” assumption interacts with the consistency of substitute confounder assumption.

If all pre-treatment single-cause confounders W_i are observed, we can simply expand Z_i ; we consider $Z'_i := (Z_i, W_i)$ in the place of Z_i . The same argument applies. \square

H Proof of Lemma 3

We first define multi-cause confounders. A multi-cause confounder is a confounder that confounds two or more causes. The following definition formalizes this idea. This definition stems from Definition 4 of [VanderWeele and Shpitser \(2013\)](#).

Definition 6. (*Multi-cause confounder*) A pretreatment covariate C_i is a multi-cause confounder if there exists a set of pre-treatment covariates V_i (possibly empty) and a set $J \subset \{1, \dots, m\}$ with $|J| \geq 2$ such that $(A_{ij})_{j \in J} \perp Y_i(a_{i1}, \dots, a_{im}) | \sigma(V_i, C_i)$. Moreover, there is no proper subset S_i of $\sigma(V_i, C_i)$ and no proper subset J' of J such that $(A_{ij})_{j \in J'} \perp Y_i(a_{i1}, \dots, a_{im}) | S_i$.

Proof sketch. This proposition is a consequence of Lemma 1, Lemma 2, and a proof by contradiction. The intuition is that if a confounder affects two or more causes then the substitute confounder Z_i must have captured it. Why? Obtain the substitute confounder Z_i from a factor model; Lemma 1 ensures that it satisfies unconfoundedness. Now suppose we omitted a multi-cause confounder C_i . Then the substitute confounder Z_i could not have satisfied unconfoundedness: the omitted confounder C_i renders the causes and potential outcomes conditionally dependent, even given Z_i . Figure 1 gives the intuition with a graphical model and Appendix H gives a detailed proof. \square

Proof. Without loss of generality, we work with two-cause confounders. The proof is directly applicable to general multi-cause confounders.

We prove the proposition by contradiction. Suppose there exists such a multi-cause confounder $W_{i,bad}$ that is not measurable with respect to $\sigma(Z_i)$; we show that Z_i could not have satisfied the factor model Eq. 36.

By Lemma 2.22 in [Kallenberg \(1997\)](#), there exist some function f_j such that $A_{ij} = f_j(Z_i, U_{ij})$, where $U_{ij} \perp\!\!\!\perp Z_i$. (f_j is non-constant in Z_i .)

Then $W_{i,bad}$ being a multi-cause confounder has two implications:

1. There exist j_1, j_2 and nontrivial functions g_1, g_2 such that $U_{ij_1} = g_1(W_{i,bad}, \gamma_{ij_1})$ and $U_{ij_2} = g_2(W_{i,bad}, \gamma_{ij_2})$, where $(\gamma_{ij_1}, \gamma_{ij_2}) \perp\!\!\!\perp W_{i,bad}$;
2. There exists a nontrivial function h such that $Y_i(a_{i1}, \dots, a_{im}) = h(W_{i,bad}, \epsilon)$, where $\epsilon \perp\!\!\!\perp W_{i,bad}$.

These two statements implies that

$$(U_{ij_1}, U_{ij_2}) \not\perp\!\!\!\perp Y_i(a_{i1}, \dots, a_{im}) | Z_i,$$

because $W_{i,bad}$ is not measurable with respect to $\sigma(Z_i)$. This implies

$$(U_{i1}, \dots, U_{im}) \not\perp\!\!\!\perp Y_i(a_{i1}, \dots, a_{im}) | Z_i.$$

It contradicts the fact that Z_i comes from the factor model (Eq. 35) with $(U_{i1}, \dots, U_{im}) \perp\!\!\!\perp Y_i(a_{i1}, \dots, a_{im}) | Z_i$. Therefore, there does not exist such a multi-cause confounder. \square

Corollary 9. Under “no unobserved single-cause confounders”, any confounder must be measurable with respect to the σ -algebra generated by the substitute confounder Z_i and the observed covariates X_i .

Proof. Because of “no unobserved single-cause confounders”, a single-cause confounder must be measurable with respect to the observed covariates X_i . Because of Lemma 3, a multi-cause confounder must be measurable with respect to the substitute confounder Z_i . Thus all confounders must be measurable with respect to the union of the substitute confounders and the observed covariates (Z_i, X_i) . \square

Corollary 9 shows how the “no unobserved single-cause confounder” assumption is necessary for the deconfounder; the substitute confounder Z_i can only handle multi-cause confounders.

I Proof of Lemma 4

Proof sketch. The deconfounder separates inference of the substitute confounder from estimation of causal effects; see Algorithm 1. This two-stage procedure guarantees that the substitute confounder is “pre-treatment”; it does not contain a mediator. The reason is that a mediator is, by

definition, a post-treatment variable that affects the potential outcome. Thus it (almost surely) cannot be identified with only the assigned causes and it is not measurable with respect to the observed (pre-treatment) covariates X_i . Appendix I provides a detailed proof. \square

Proof. We prove the proposition by contradiction.

Consider a mediator M . We denote $M_i(a)$ as the potential value of the mediator M for unit i when the assigned cause is a . We show that $M_i(\mathbf{a}_i)$ is almost surely not measurable with respect to Z_i .

The deconfounder operating in two stages. Inferring the substitute confounder Z_i is separated from estimating the potential outcome. It implies that the substitute confounder is independent of the outcomes conditional on the causes \mathbf{A}_i : $Z_i \perp\!\!\!\perp Y_i(\mathbf{A}_i) | \mathbf{A}_i$. The intuition is that, without looking at $Y_i(\cdot)$, the only dependence between Z_i and Y_i must come from \mathbf{A}_i .

However, a mediator must satisfy $M_i(\mathbf{A}_i) \not\perp\!\!\!\perp Y_i(\mathbf{A}_i) | \mathbf{A}_i$; otherwise, it has no mediation effect (Imai et al., 2010). If a mediator is measurable with Z_i , then $Z_i \perp\!\!\!\perp Y_i(\mathbf{A}_i) | \mathbf{A}_i$. This contradicts the conditional independence of Z_i and $Y_i(\mathbf{A}_i)$ given \mathbf{A}_i . We ensured this conditional independence by inferring the substitute confounder Z_i based only on the causes \mathbf{A}_i . \square

As a consequence of “no unobserved single-cause confounders”, the substitute confounder, together with the observed covariates, captures all confounders.

J Proof of Proposition 5

The first part is a direct consequence of Lemmas 1 and 2.

We now prove the second part. We provide two constructions.

We start with the first trivial one. For any assigned causes \mathbf{A}_i , we consider a special case when $\mathbf{A}_i \stackrel{a.s.}{=} Z_i$. We have

$$p(a_{i1}, \dots, a_{im} | z_i) = \delta_{z_i} = \prod_{j=1}^m \delta_{z_{ij}} = \prod_{j=1}^m p(a_{ij} | z_i) \quad (72)$$

This step is due to point masses are factorizable. Therefore, we can write the distribution of \mathbf{A}_i in the form of a factor model; we set $\theta_j \stackrel{a.s.}{=} 0, j = 1, \dots, m$ and $Z_i \stackrel{a.s.}{=} \mathbf{A}_i$:

$$p(\theta_{1:m}, z_{1:n}, \mathbf{a}_{1:n}) = p(\theta_{1:m})p(z_{1:n} | \theta_{1:m})p(\mathbf{a}_{1:n} | z_{1:n}, \theta_{1:m}) \quad (73)$$

$$= p(\theta_{1:m})p(z_{1:n})p(\mathbf{a}_{1:n} | z_{1:n}) \quad (74)$$

$$= p(\theta_{1:m})p(z_{1:n}) \prod_{i=1}^n \prod_{j=1}^m p(a_{ij} | z_i) \quad (75)$$

The second equality is due to $Z_i \perp\!\!\!\perp \theta_{1:m}$ and $\mathbf{A}_i \perp\!\!\!\perp \theta_{1:m} | Z_i$. They are because θ_j ’s are point masses. The third equality is due to the SUTVA assumption and Eq. 72.

Choosing $Z_i \stackrel{a.s.}{=} \mathbf{A}_i$, that is letting the substitute confounder Z_i be the same as the assigned causes \mathbf{A}_i , does not help with causal inference; see a related discussion on overlap around Eq. 6.

This result is only meant to exemplify the large capacity of factor models. Finally, this $Z_i \stackrel{a.s.}{=} \mathbf{A}_i$ example also illustrates the fact that a factor model capturing $p(\mathbf{a}_i)$ is not necessarily the true assignment model.

We now present the second construction. It relies on copulas and the Sklar’s theorem. We follow the modified distribution function from [Rüschendorf \(2009\)](#). Let X be a real random variable with distribution function F and let $V \sim U(0, 1)$ be uniformly distributed on $(0, 1)$ and independent of X . The modified distribution function $F(x, \lambda)$ is defined by

$$F(x, \lambda) := P(X < x) + \lambda P(X = x). \quad (76)$$

Then if we construct U variables as

$$U := F(X, V), \quad (77)$$

then we have

$$U = F(X-) + V(F(X) - F(X-)), \quad (78)$$

$$U \stackrel{d}{=} \text{Uniform}(0, 1), \quad (79)$$

$$X \stackrel{a.s.}{=} F^{-1}(U). \quad (80)$$

Now we set $Z_{ij} = F_{ij}^{-1}(A_{ij})$, where F_{ij} is the modified distribution function of A_{ij} . We also set $\theta_j, j = 1, \dots, m$ as point masses. The Sklar’s theorem then implies

$$p(\theta_{1:m}, z_{1:n}, \mathbf{a}_{1:n}) = p(\theta_{1:m})p(z_{1:n} | \theta_{1:m})p(\mathbf{a}_{1:n} | z_{1:n}, \theta_{1:m}) \quad (81)$$

$$= p(\theta_{1:m})p(z_{1:n})p(\mathbf{a}_{1:n} | z_{1:n}, \theta_{1:m}) \quad (82)$$

$$= p(\theta_{1:m})p(z_{1:n}) \prod_{i=1}^n \prod_{j=1}^m p(a_{ij} | z_i, \theta_j) \quad (83)$$

The second equality is due to $\theta_{1:m}$ being point masses; $\theta_j, j = 1, \dots, m$ can be considered as parameters of the marginal distribution of A_{ij} . The third equality is due to the SUTVA assumption and the Sklar’s theorem.

This construction aligns more closely with the idea of the deconfounder; it aims to capture multi-causes confounders that induces the dependence structure, i.e. the copula. However, the deconfounder is different from directly estimating the copula; the latter is a more general (and harder) problem.

K Proof of Theorem 6

Proof sketch. Theorem 6 rely on two results: (1) “no unobserved single-cause confounders” and Lemma 3 ensure (X_i, Z_i) capture all confounders; (2) the pre-treatment nature of X_i and Lemma 4

ensure (X_i, Z_i) capture no mediators. These results assert unconfoundedness given the substitute confounders Z_i and the observed covariates X_i . They greenlight us for causal inference if the factor model admits consistent estimates of Z_i , i.e. the substitute confounder has a degenerate distribution $P(Z_i | \mathbf{A}_i) = \delta_{f(\mathbf{A}_i)}$.

Given these results, Theorem 6 identifies the average causal effect of all the causes by assuming $\nabla_{\mathbf{a}} f(a_1, \dots, a_m) = 0$ almost everywhere and a separable outcome model. These two assumptions let us identify the average causal effect without assuming overlap.

More specifically, $\nabla_{\mathbf{a}} f(a_1, \dots, a_m) = 0$ roughly requires that the substitute confounder is a step function of the all causes. In other words, we can partition all possible values of (a_1, \dots, 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 $\mathbb{E}_Y[Y_i(\mathbf{a})] - \mathbb{E}_Y[Y_i(\mathbf{a}')] is identifiable if \mathbf{a} and \mathbf{a}' belong to the same region.$

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 \mathbf{a} and \mathbf{a}' . The full proof is in Appendix K. \square

Proof. For notational simplicity, denote $\mathbf{a} = (a_1, \dots, a_m)$, $\mathbf{a}' = (a'_1, \dots, a'_m)$, and $\mathbf{A}_i = (A_{i1}, \dots, A_{im})$. We also write $f_{\theta}(\cdot) = f(\cdot)$.

We start with rewriting $\mathbb{E}_Y[Y_i(\mathbf{a})] - \mathbb{E}_Y[Y_i(\mathbf{a}')] using the unconfoundedness assumption and the separability assumption.$

First notice that

$$\mathbb{E}_Y[Y_i(\mathbf{a})] = \mathbb{E}_{Z, X}[\mathbb{E}_Y[Y_i(\mathbf{a}) | X_i, Z_i]] \quad (84)$$

$$= \mathbb{E}_X[f_1(\mathbf{a}, X_i)] + \mathbb{E}_Z[f_2(Z_i)]. \quad (85)$$

The first equality is due to the tower property. The second equality is due to the separability assumption. The third equality is due to linearity of expectations.

Hence we have

$$\mathbb{E}_Y[Y_i(\mathbf{a})] - \mathbb{E}_Y[Y_i(\mathbf{a}')] = \mathbb{E}_X[f_1(\mathbf{a}, X_i)] - \mathbb{E}_X[f_1(\mathbf{a}', X_i)] \quad (86)$$

$$= \int_{C(\mathbf{a}, \mathbf{a}')} \nabla_{\mathbf{a}} \mathbb{E}_X[f_1(\mathbf{a}, X_i)] d\mathbf{a}, \quad (87)$$

where $C(\mathbf{a}, \mathbf{a}')$ is a line where \mathbf{a} and \mathbf{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 $\nabla_{\mathbf{a}} \mathbb{E}_X[f_1(\mathbf{a}, X_i)]$ relates to the gradient of the outcome model we fit. The key idea here is that the two gradients are equal in regions $\{\mathbf{a} : f(\mathbf{a}) = c\}$ for each c .

We will rely on the consistent substitute confounder assumption. Notice that, for almost all \mathbf{a} , we have

$$\nabla_{\mathbf{a}} \mathbb{E}_X [f_1(\mathbf{a})] = \nabla_{\mathbf{a}} \mathbb{E}_X [f_3(\mathbf{a})] \quad (88)$$

It is due to two observations. The first observation is that

$$\nabla_{\mathbf{a}} \mathbb{E}_X [\mathbb{E}_Y [Y_i | Z_i = f(\mathbf{a}), A_i = \mathbf{a}, X_i]] \quad (89)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [\mathbb{E}_Y [Y_i(\mathbf{a}) | Z_i = f(\mathbf{a}), A_i = \mathbf{a}, X_i]] \quad (90)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [\mathbb{E}_Y [Y_i(\mathbf{a}) | Z_i = f(\mathbf{a}), X_i]] \quad (91)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [f_1(\mathbf{a}, X_i)] + \nabla_{\mathbf{a}} f_2(f(\mathbf{a})) \quad (92)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [f_1(\mathbf{a}, X_i)] + \nabla_{f(\mathbf{a})} f_2 \cdot \nabla_{\mathbf{a}} f(\mathbf{a}) \quad (93)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [f_1(\mathbf{a}, X_i)] \quad (94)$$

The first equality is due to SUTVA. The second equality is due to Proposition 5.1: $Y_i(\mathbf{a}) \perp \mathbf{A}_i | X_i, Z_i$. The third equality is due to the separability condition. The fourth equality is due to the chain rule. The fifth equality is due to $\nabla_{\mathbf{a}} f(\mathbf{a}) = 0$ up to a set of Lebesgue measure zero.

The second observation is that

$$\nabla_{\mathbf{a}} \mathbb{E}_X [\mathbb{E}_Y [Y_i | Z_i = f(\mathbf{a}), \mathbf{A}_i = \mathbf{a}, X_i]] \quad (95)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [f_3(\mathbf{a}, X_i)] + \nabla_{\mathbf{a}} f_4(f(\mathbf{a})) \quad (96)$$

$$= \nabla_{\mathbf{a}} \mathbb{E}_X [f_3(\mathbf{a}, X_i)] \quad (97)$$

Hence Eq. 88 is true because f_1 and f_3 are continuously differentiable.

Therefore, we have

$$\mathbb{E}_Y [Y_i(\mathbf{a})] - \mathbb{E}_Y [Y_i(\mathbf{a}')] \quad (98)$$

$$= \int_{C(\mathbf{a}, \mathbf{a}')} \nabla_{\mathbf{a}} \mathbb{E}_X [f_1(\mathbf{a}, X_i)] d\mathbf{a} \quad (99)$$

$$= \int_{C(\mathbf{a}, \mathbf{a}')} \nabla_{\mathbf{a}} \mathbb{E}_X [f_3(\mathbf{a}, X_i)] d\mathbf{a} \quad (100)$$

$$= \mathbb{E}_X [f_3(\mathbf{a}, X_i)] - \mathbb{E}_X [f_3(\mathbf{a}', X_i)] \quad (101)$$

$$= (\mathbb{E}_X [f_3(\mathbf{a}, X_i)] + \mathbb{E} [f_4(Z_i)]) - (\mathbb{E}_X [f_3(\mathbf{a}', X_i)] + \mathbb{E} [f_4(Z_i)]) \quad (102)$$

$$\begin{aligned} &= \int \mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}', X_i, Z_i] P(Z_i, X_i) dZ_i dX_i \\ &\quad - \int \mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}, X_i, Z_i] P(Z_i, X_i) dZ_i dX_i \end{aligned} \quad (103)$$

$$= \mathbb{E}_{Z, X} [\mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}, Z_i, X_i]] - \mathbb{E}_{Z, X} [\mathbb{E}_Y [Y_i | \mathbf{A}_i = \mathbf{a}', Z_i, X_i]]. \quad (104)$$

The first equality is due to Eq. 87. The second equality is due to Eq. 88. 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.

□

L Proof of Theorem 7

Proof. Lemma 1 and Lemma 2, together with “no unobserved single-cause confounders”, ensures that the substitute confounder Z_i and the observed covariate X_i satisfies

$$(A_{i1}, \dots, A_{im}) \perp\!\!\!\perp Y_i(a_{i1}, \dots, a_{im}) | Z_i, X_i. \quad (105)$$

Therefore, we have

$$\mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_{1:k}, A_{i,(k+1):m})]] \quad (106)$$

$$= \mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_1, \dots, a_k, A_{i,k+1}, \dots, A_{im})]] \quad (107)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_1, \dots, a_k, A_{i,k+1}, \dots, A_{im}) | Z_i, X_i]]] \quad (108)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(a_1, \dots, a_k, A_{i,k+1}, \dots, A_{im}) | Z_i, X_i, A_{i1} = a_1, \dots, A_{ik} = a_k]]] \quad (109)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i(A_{i1}, \dots, A_{ik}, A_{i,k+1}, \dots, A_{im}) | Z_i, X_i, A_{i1} = a_1, \dots, A_{ik} = a_k]]] \quad (110)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_{A_{(k+1):m}} [\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i1} = a_1, \dots, A_{ik} = a_k]]] \quad (111)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i1} = a_1, \dots, A_{ik} = a_k]] \quad (112)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i,1:k} = a_{1:k}]] \quad (113)$$

The first equality is an expansion of the notations. The second equality is due to the tower property. The third equality is due to Eq. 105. The fourth equality is due to $A_{i1} = a_1, \dots, A_{ik} = a_k$. The fifth equality is due to SUTVA. The sixth equality is due to the inner expectation does not depend on $A_{(k+1):m}$.

Therefore, we have

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

by the linearity of expectation.

Finally, $\mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i | Z_i, X_i, A_{i,1:k} = a_{1:k}]]$ 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 consistently estimated. \square

M Proof of Theorem 8

Proof. As with Theorem 6 and Theorem 7, Theorem 8 relies on the unconfoundedness given the substitute confounders Z_i and the observed covariates X_i due to Lemma 3 and Lemma 4.

Given this unconfoundedness, Theorem 8 identifies the mean potential outcome of an individual given its current cause assignment $A_i = (a_1, \dots, a_m)$; it only requires that the new cause assignment of interest (a'_1, \dots, a'_m) lead to the same substitute confounder estimate: $f(a_1, \dots, a_m) = f(a'_1, \dots, a'_m)$.

To prove identification, we rewrite this conditional mean potential outcome

$$\mathbb{E}_Y [Y_i(a'_1, \dots, a'_m) | A_{i1} = a_1, \dots, A_{im} = a_m] \quad (114)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i(a'_1, \dots, a'_m) | A_{i1} = a_1, \dots, A_{im} = a_m, Z_i, X_i]] \quad (115)$$

$$= \mathbb{E}_X [\mathbb{E}_Y [Y_i(a'_1, \dots, a'_m) | A_{i1} = a_1, \dots, A_{im} = a_m, Z_i = f(a_1, \dots, a_m), X_i]] \quad (116)$$

$$= \mathbb{E}_X [\mathbb{E}_Y [Y_i(a'_1, \dots, a'_m) | A_{i1} = a'_1, \dots, A_{im} = a'_m, Z_i = f(a_1, \dots, a_m), X_i]] \quad (117)$$

$$= \mathbb{E}_{Z,X} [\mathbb{E}_Y [Y_i | A_{i1} = a'_1, \dots, A_{im} = a'_m, Z_i, X_i]] \quad (118)$$

The first equality is due to the tower property. The second equality is due to the consistency requirement on the substitute confounder $P(Z_i | \mathbf{A}_i) = \delta_{f(\mathbf{A}_i)}$. The third equality is due to unconfoundedness given Z_i, X_i . The fourth equality is estimable from the data because $f(a_1, \dots, a_m) = f(a'_1, \dots, a'_m)$. Hence the nonparametric identification of $\mathbb{E}_Y [Y_i(a'_1, \dots, a'_m) | A_{i1} = a_1, \dots, A_{im} = a_m]$ is established. We note that this identification result does not require overlap. \square

N Details of Section 6.2

We follow [Song et al. \(2015\)](#) in simulating the allele frequencies. We present the full details here.

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 19 details how we generate Γ and S for each simulation setup.

For each simulation scenarios, 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 .

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 $\mathcal{S}_k \subset \{1, 2, \dots, n\}$. Set $\lambda_j = k$ if $j \in \mathcal{S}_k, k = 1, 2, 3$.

We then simulate the random effects ϵ_i . Let $\tau_1^2, \tau_2^2, \tau_3^2 \stackrel{iid}{\sim} \text{InvGamma}(3, 1)$, 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 simulate datasets of $n = 5000$ individuals and $m = 100,000$ SNPs; we let the causal signals $\sum_{j=1}^m \beta_j a_{ij}$ contribute to $v_{\text{gene}} = 0.1$ of the variance, the confounder λ_i contribute $v_{\text{conf}} = 0.2$, and the random effects ϵ_i contribute $v_{\text{noise}} = 0.7$. We set the first 10% of the m SNPs to be the true causal SNPs ($\beta_j \neq 0, \beta_j \stackrel{iid}{\sim} \mathcal{N}(0, 1)$; $\beta_j = 0$ for the rest of the SNPs. In the high SNR setting, we simulate datasets of $n = 5,000$ individuals and $m = 5,000$ SNPs; we have $v_{\text{gene}} = 0.4$, $v_{\text{conf}} = 0.4$, and $v_{\text{noise}} = 0.2$.

We set

$$\lambda_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^m \beta_j a_{ij}\}_{i=1}^n}{\sqrt{v_{\text{gene}}}} \right] \left[\frac{\sqrt{v_{\text{conf}}}}{s.d.\{\lambda_i\}_{i=1}^n} \right] \lambda_i, \quad (119)$$

$$\epsilon_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^m \beta_j a_{ij}\}_{i=1}^n}{\sqrt{v_{\text{gene}}}} \right] \left[\frac{\sqrt{v_{\text{noise}}}}{s.d.\{\epsilon_i\}_{i=1}^n} \right] \epsilon_i. \quad (120)$$

We finally generate a real-valued outcome from a linear model and a binary outcome from a logistic model:

$$y_{i,\text{real}} = \sum_{j=1}^m \beta_j a_{ij} + \lambda_i + \epsilon_i, \quad (121)$$

$$y_{i,\text{binary}} \sim \text{Bernoulli} \left(\frac{1}{1 + \exp(\sum_{j=1}^m \beta_j a_{ij} + \lambda_i + \epsilon_i)} \right). \quad (122)$$

		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 9: 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 10: 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.

		Real-valued outcome	Binary outcome
	Pred. check	RMSE $\times 10^{-2}$	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 11: 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.

		Real-valued outcome	Binary outcome
	Pred. check	RMSE $\times 10^{-2}$	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 12: 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.

		Real-valued outcome	Binary outcome
	Pred. check	RMSE $\times 10^{-2}$	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 13: 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 14: 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 15: 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}_{PPCA}	-1.2
Control for \hat{a}_{PF}	-1.2
Control for \hat{a}_{DEF}	-1.2
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.3
Control for (\hat{a}_{PF}, X)	-1.2
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.2

Table 16: 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}_{PPCA}	-1.6
Control for \hat{a}_{PF}	-1.5
Control for \hat{a}_{DEF}	-1.5
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.7
Control for (\hat{a}_{PF}, X)	-1.5
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.6

Table 17: 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}_{PPCA}	-1.4
Control for \hat{a}_{PF}	-1.2
Control for \hat{a}_{DEF}	-1.3
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.4
Control for (\hat{a}_{PF}, X)	-1.3
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.2

Table 18: 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: $\gamma_{ik} \stackrel{iid}{\sim} \text{BN}(p_i, F_i)$, 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 $\gamma_{ik} \stackrel{iid}{\sim} 0.9 \times \text{Uniform}(0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. In order 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: $\gamma_{ik} \stackrel{iid}{\sim} \text{BN}(p_i, F_i)$, 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 $(s_{1j}, s_{2j}, s_{3j}) \stackrel{iid}{\sim} \text{Dirichlet}(\alpha, \alpha, \alpha)$ for $j = 1, \dots, m, \alpha = 0.01, 0.1, 0.5, 1$.
Spatial	The matrix Γ was generated by sampling $\gamma_{ik} \stackrel{iid}{\sim} 0.9 \times \text{Uniform}(0, 0.5)$, 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 $\tau \rightarrow 0$, the individuals are placed closer to the corners of the unit square, while when $\tau = 1$, the individuals are distributed uniformly.

Table 19: Simulating allele frequencies.

References

- Airoldi, E., Blei, D., Erosheva, E., and Fienberg, S., editors (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press.
- 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.
- 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.
- 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.
- Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71.
- Kallenberg, O. (1997). Foundations of modern probability. *Collection: Probability and Its Applications*, Springer.
- 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.
- 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.
- Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsdottir, 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.
- Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition.
- 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.
- 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.

- Rüschendorf, L. (2009). On the distributional transform, Sklar’s theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927.
- Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554.
- Tran, D. and Blei, D. M. (2017). Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*.
- VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, pages 196–220.
- Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.
- 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.
- 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.