The Blessings of Multiple Causes

Yixin Wang
Department of Statistics
Columbia University
yixin.wang@columbia.edu

David M. Blei
Department of Statistics
Department of Computer Science
Columbia University
david.blei@columbia.edu

June 20, 2018

Abstract

Causal inference from observational data often assumes "strong ignorability," that all confounders are observed. This assumption is standard yet untestable. However, many scientific studies involve multiple causes, different variables whose effects are simultaneously of interest. We propose the deconfounder, an algorithm that combines unsupervised machine learning and predictive model checking to perform causal inference in multiple-cause settings. The deconfounder infers a latent variable as a substitute for unobserved confounders and then uses that substitute to perform causal inference. We develop theory for when the deconfounder leads to unbiased causal estimates, and show that it requires weaker assumptions than classical causal inference. We analyze its performance in three types of studies: semi-simulated data around smoking and lung cancer, semi-simulated data around genomewide association studies, and a real dataset about actors and movie revenue. The deconfounder provides a checkable approach to estimating close-to-truth causal effects.

Keywords: Causal inference, strong ignorability, probabilistic models

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?

Suppose the producer wants to solve this problem with the potential outcomes framework (Imbens and Rubin, 2015; Rubin, 1974, 2005). Following the methodology, she associates each movie to a *potential outcome function*, $y_i(\boldsymbol{a})$. This function maps each possible cast $\boldsymbol{a}$ to its revenue if the movie i had that cast. (The cast $\boldsymbol{a}$ is a vector of indicators, with one element per actor.) The potential outcome function encodes, for example, how much money *Star Wars* would have made if Robert Redford played Han Solo, rather than Harrison Ford. In doing causal inference, the producer's goal is to estimate the population distribution of $Y_i(\boldsymbol{a})$. For example, she might consider a particular cast $\boldsymbol{a}$ and estimate the expected revenue $\mathbb{E}[Y_i(\boldsymbol{a})]$.

Classical causal inference from observational data (like Table 1) is a difficult enterprise and requires strong assumptions. The challenge is that the data is limited; it only contains the revenue of each movie at its assigned cast. But what this paper is about is that the producer's problem is not a classical causal inference. While causal inference usually considers univariate causes, e.g., whether a subject receives a drug or a control, our producer is considering a *multiple causal inference*, where each actor is a possible cause. We will show how multiple causal inference can be an easier problem than classical causal inference. Thanks to the multiplicity of causes, the producer can make valid causal inferences under weaker assumptions than the classical approach requires.

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

But there is a problem. The data in Table 1 hide *confounders*, variables that affect both how the cast is assigned to the movie and its potential outcome function. For example, every movie has a genre, such as comedy, action, or romance, and this genre has an effect on both who is in the cast and the revenue. (Action movies cast a certain set of actors and tend to make more money than comedies.) The genre induces dependence between whether an actor is in a movie and its revenue; this dependence biases the estimates, $\mathbb{E}[Y_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}] \neq \mathbb{E}[Y_i(\boldsymbol{a})]$.

Thus the main activities of classical causal inference are to identify, measure, and control for confounders. Suppose the producer measures confounders w_i for each movie. Then inference is simple: use the data (now with confounders) to take Monte Carlo estimates of the iterated expectation, $\mathbb{E}[\mathbb{E}[Y_i(\boldsymbol{a}) \mid W_i, \boldsymbol{A}_i = \boldsymbol{a}]]$. But whether this method is unbiased rests on a big and uncheckable assumption: there are no other confounders. For many applied causal inference problems, this assumption is a leap of faith.

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, … }	$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, … }	$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, … }	$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, … }	$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, … }	$1506M
Avengers: Age of Ultron	{Robert Downey Jr., Chris Hemsworth, Mark Ruffalo, Chris Evans, … }	$1405M
Frozen	{Kristen Bell, Idina Menzel, Jonathan Groff, Josh Gad, … }	$1274M
Iron Man 3	{Robert Downey Jr., Gwyneth Paltrow, Don Cheadle, Guy Pearce, … }	$1215M
Minions	{Sandra Bullock, Jon Hamm, Michael Keaton, Allison Janney, … }	$1157M
Captain America: Civil War	{Chris Evans, Robert Downey Jr., Scarlett Johansson, Sebastian Stan, … }	$1153M
Transformers: Dark of the Moon	{Shia LaBeouf, John Malkovich, Ken Jeong, Frances McDormand, … }	$1124M
The Lord of the Rings: The Return of the King	{Elijah Wood, Ian McKellen, Viggo Mortensen, Liv Tyler, … }	$1119M
Skyfall	{Daniel Craig, Judi Dench, Javier Bardem", Ralph Fiennes, … }	$1109M
Transformers: Age of Extinction	{Mark Wahlberg, Stanley Tucci, Kelsey Grammer, Nicola Peltz, … }	$1091M
The Dark Knight Rises	{Christian Bale, Michael Caine, Gary Oldman, Anne Hathaway, … }	$1085M
Toy Story 3	{Tom Hanks, Tim Allen, Ned Beatty, Joan Cusack, … }	$1067M
Pirates of the Caribbean: Dead Man's Chest	{Johnny Depp, Orlando Bloom, Keira Knightley, Stellan Skarsgerd, … }	$1066M
Pirates of the Caribbean: On Stranger Tides	{Johnny Depp, Penelope Cruz, Ian McShane, Kevin McNally, … }	$1046M
Alice in Wonderland	{Mia Wasikowska, Johnny Depp, Anne Hathaway, Helena Bonham Carter, … }	$1025M
The Hobbit: An Unexpected Journey	{Ian McKellen, Martin Freeman, Richard Armitage, Andy Serkis, … }	$1021M
⋮	⋮	⋮

Table 1: Top earning movies in the TMDB dataset

Here we develop *the deconfounder*, an alternative method for the producer who worries about missing a confounder. Here is how it works. 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 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.

The deconfounder capitalizes on the dependency structure of the observed casts, using patterns of how actors tend to appear together in movies as indirect evidence for confounders in the data. With the deconfounder, the producer replaces an uncheckable search for possible confounders with the checkable goal of building a good factor model of observed casts.

All methods for causal inference are based on assumptions. Here we make two. First, we assume that the fitted latent-variable model is a good model of the assigned causes. Happily, this assumption is testable; we will use predictive checks to assess how well the fitted model captures the data. Second, we assume 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 no unobserved confounders.

Beyond the movies, many causal inference problems, especially from observational data, also classify as multiple causal inference. Such problems come from a diversity of fields.

- **Genome-wide association studies (GWAS).** GWAS problems focus on understanding the causal connection between genetic variants and traits (Stephens and Balding, 2009; Visscher et al., 2017). The assigned causes are alleles on the genome, often encoded as either being more common ("major allele") or less common ("minor allele"), and the effect is the trait under study. Confounders, like population structure, bias naive estimates of the effect of genetic variants. We study GWAS problems in Section 3.
- **Computational neuroscience.** Neuroscientists are interested in how specific neurons or brain measurements affect behavior and thoughts (Churchland et al., 2012). The possible causes are high dimensional measurements, e.g., one per neuron, and the effect is a measured behavior. Confounders, particularly through dependencies among neural activity, bias the estimated connections between brain activity and behavior.
- **Social science.** Sociologists and policy-makers are interested in how various social programs and interventions 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 and pharmacologists are interested in how various medical treatments affect the progression of disease. In this domain, 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.

In each of these settings, we can use the deconfounder. We fit a good factor model of the assigned causes, infer substitute confounders, and then use the substitutes to perform causal inference.

Related work. This work relates to several threads of research in causal inference.

Probabilistic modeling for 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. More recently, Louizos et al. (2017) use variational autoencoders to infer unobserved confounders. Kocaoglu et al. (2017) introduce generative adversarial networks into causal inference.

With a related goal, Tran and Blei (2017) build implicit causal models. Like the GWAS example in this paper (Section 3.2), 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 multiple 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.

Our work complements all of these works. These works rest on Pearl's causal framework (Pearl, 2009); they hypothesize a causal graph with confounders, causes, and outcomes. We develop the deconfounder in the context of the potential outcomes framework (Imbens and Rubin, 2015; Rubin, 1974, 2005).

Analyzing GWAS. In GWAS, latent population structure is an important unobserved confounder. Pritchard et al. (2000b) propose a probabilistic admixture model for unsupervised ancestry inference. Price et al. (2006) and Astle et al. (2009) estimate 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, and use the kinship matrix to parameterize the random effect in a linear mixed model. Song et al. (2015) and Hao et al. (2015) rely on factor analysis and admixture models to estimate the population structure. GTEx Consortium et al. (2017) adopt a similar strategy to study the effects of genetic variants on gene expression levels across tissues. As we discuss in Section 2.7, these methods can be seen as variants of the deconfounder. 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 3.2.

Assessing the strong ignorability assumption. Rosenbaum and Rubin (1983) demonstrates that strong ignorability and a good propensity score model are sufficient to perform causal inference with observational data. Many subsequent efforts assess the plausibility of strong ignorability. For example, Robins et al. (2000); Gilbert et al. (2003); Imai and Van Dyk (2004) develop sensitivity analysis in various contexts, though focusing on data with a single cause. In contrast to these works, our work uses predictive model checks to assess unconfoundedness with multiple causes. More recently, Sharma et al. (2016) leveraged auxillary outcome data to test for confounding in time series data; Janzing and Schölkopf (2018b,a); Liu and Chan (2018) developed tests for non-confounding in multivariate linear regression. Here we work without auxillary data, focus on causal estimation, as opposed to testing, and move beyond linear models.

The (generalized) propensity score. Schneeweiss et al. (2009); McCaffrey et al. (2004); Lee et al. (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. Our work expands on this previous work by introducing latent variables into the model. As we will show, the multiplicity of causes enables us to infer these latent variables and then use them as substitutes for unobserved confounders.

Classical causal inference with multiple treatments. Lopez et al. (2017); McCaffrey et al. (2013); Zanutto et al. (2005); Rassen et al. (2011); Lechner (2001); Feng et al. (2012) extend the classical matching, subclassification, and weighting methods to multiple treatments, always assuming strong ignorability. This work complements their work by relaxing that assumption.

This paper. The rest of the paper is organized as follows. Section 2 reviews classical causal inference, sets up multiple causal inference, highlights the blessings of multiple causes, and presents the deconfounder algorithm. Section 3 presents three empirical studies, two semi-synthetic and one real. Section 4 develops the theory around the deconfounder. It justifies the deconfounder algorithm and characterizes some properties of the substitute confounder. Finally, Section 5 concludes the paper with a discussion.

2 Multiple causal inference with the deconfounder

2.1 A classical approach to multiple causal inference

Using the potential outcomes framework, we more formally describe multiple causal inference. There are *m possible causes*, encoded in a vector $\boldsymbol{a} = (a_1, \ldots, a_m)$. We can consider a variety of types: real-valued causes, binary causes, integer causes, and so on. In the movie example from the introduction, 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 ith movie, the potential outcome function maps each possible cast to the log of its revenue, $y_i(\boldsymbol{a}) : \{0,1\}^m \to \mathbb{R}$. Note $y(\cdot)$ is a function. It maps every possible cast of actors to the movie's revenue for that cast.

The goal of causal inference is to characterize the sampling distribution of the potential outcomes $Y_i(\boldsymbol{a})$ for each configuration of the causes $\boldsymbol{a}$. This distribution provides causal inferences, such as the expected outcome for a particular array of causes (a particular cast of actors) $\mu(\boldsymbol{a}) = \mathbb{E}[Y_i(\boldsymbol{a})]$ 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 and realized outcomes.[1] For each individual (movie), we observe the assigned causes $\boldsymbol{a}_i$ (the cast) and the realized outcome $y_i(\boldsymbol{a}_i)$ (its revenue); the data is $\mathcal{D} = \{(\boldsymbol{a}_i, y_i(\boldsymbol{a}_i)\}\ i = 1, \ldots, n$. Note that we only observe the outcome for the assigned causes $y_i(\boldsymbol{a}_i)$; this outcome is just one of the values of the potential outcome function. Using such data to characterize the full distribution of $Y_i(\cdot)$ is the "fundamental problem of causal inference" (Holland, 1986).

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

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

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

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

[2] Here is the notation. Capital letters denote a random variable. For example, the random variable $\boldsymbol{A}_i$ is a randomly chosen vector of assigned causes from the population. The random variable $Y_i(\boldsymbol{A}_i)$ is a a randomly chosen potential outcome from the population, evaluated at its assigned causes. A lowercase letter is a realization. For example, $\boldsymbol{a}_i$ 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 $\boldsymbol{A}_i = \boldsymbol{a}$. The right side is an expectation over the same population of the potential outcome functions, but always evaluated at the realization $\boldsymbol{a}$.

Suppose we measure all the confounders x_i. Append each data point $\mathscr{D} = \{(\boldsymbol{a}_i, x_i, y_i(\boldsymbol{a}_i)\}$ $i = 1, \ldots, n$ and estimate an iterated expectation,

$$\mathbb{E}[\mathbb{E}[Y_i(\boldsymbol{a}) \mid X_i, \boldsymbol{A}_i = \boldsymbol{a}]] = \mathbb{E}[Y_i(\boldsymbol{a})]. \tag{2}$$

Using the augmented dataset, we can estimate the left side with Monte Carlo; thus we can estimate $\mathbb{E}[Y_i(\boldsymbol{a})]$.

Equation (2) is true when X captures all the confounders. More precisely, it is true under the important assumption of *strong ignorability* (Rosenbaum and Rubin, 1983; Imai and Van Dyk, 2004). Strong ignorability says that, conditional on the confounders, the assigned causes are independent of the potential outcomes,

$$\boldsymbol{A}_i \perp\!\!\!\perp Y_i(\boldsymbol{a}) \mid X_i \qquad \forall \boldsymbol{a}. \tag{3}$$

The nuance of strong ignorability is that Equation (3) needs to hold for all possible $\boldsymbol{a}$'s, not only for the value of $Y_i(\boldsymbol{a})$ at the assigned causes. Strong ignorability is equivalent to the assumption that there are no unobserved confounders.[3]

Equation (2) underlies the practice of causal inference: find and measure all 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 strong ignorability is untestable—it fundamentally depends on the entire potential outcome function, of which we only observe one value (Holland, 1986).

2.2 Deconfounder: Multiple causal inference without strong ignorability

We now develop the *deconfounder*, an algorithm that exploits the multiplicity of causes to sidestep the search for confounders. There are three steps. First, find a good latent variable model of the assignment mechanism $p(z, a_1, \ldots, a_m)$, where z is a local factor. Second, use the model to infer the latent variable for each unit $p(z_i \mid a_{i1}, \ldots, a_{im})$. Finally, use the inferred variable as a substitute for unobserved confounders and form causal inferences. As we said above, the deconfounder replaces an uncheckable search for possible confounders with the checkable goal of building a good model of assigned causes.

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

$$\begin{aligned} Z_i &\sim p(\cdot \mid \alpha) \quad i = 1, \ldots, n, \\ A_{ij} \mid Z_i &\sim p(\cdot \mid z_i, \theta_j) \quad j = 1, \ldots, m, \end{aligned} \tag{4}$$

[3]Following Imai and Van Dyk (2004), we call Equation (3) strong ignorability. Imbens (2000) and Hirano and Imbens (2004) call it weak unconfoundedness. We 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).

where α parameterizes the distribution of Z_i and θ_j parameterizes the per-cause distribution of A_{ij}. Notice that Z_i can be multi-dimensional. Factor models encompass the probabilistic view of many common factorization methods. Examples include matrix factorization (Tipping and Bishop, 1999), mixture models (McLachlan and Basford, 1988), mixed-membership models (Pritchard et al., 2000b; Blei et al., 2003; Airoldi et al., 2008; Erosheva, 2003), and deep generative models (Neal, 1990; Ranganath et al., 2015, 2016; Tran et al., 2017; Rezende and Mohamed, 2015; Mohamed and Lakshminarayanan, 2016; Kingma and Welling, 2013). One can fit using any appropriate method, such as maximum likelihood estimation or Bayesian inference.

With the fitted factor model in hand, use it to calculate the conditional expectation of each unit's local factor weights $\hat{z}_i = \mathbb{E}_M[Z_i \mid \boldsymbol{A}_i = \boldsymbol{a}_i]$. (This expectation is from the fitted model M, as opposed to the population distribution and one can use approximate expectations.) Finally, condition on $\hat{z}_i$ as a substitute confounder (drawn from the same population) and proceed with causal inference, i.e., estimate $\mathbb{E}\left[\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid \hat{\boldsymbol{Z}}_i, \boldsymbol{A}_i = \boldsymbol{a}\right]\right]$. The main idea behind this method is that if the factor model captures the distribution of assigned causes—a testable proposition—then we can safely use $\hat{z}_i$ as a variable that contains the confounders.

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

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

Now make an additional assumption: there are no *single-cause confounders*, a variable that depends on just one of the assigned causes and on the potential outcome function. With this assumption, we will show in Section 4 that the independence statement of Equation (5) implies strong ignorability,

$$\boldsymbol{A}_i \perp\!\!\!\perp Y_i(\boldsymbol{a}) \mid Z_i. \tag{6}$$

Strong ignorability justifies causal inference.

The graphical model in Figure 1 justifies the deconfounder and reveals its assumptions.[4] 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, i.e., Equation (5). This is a contradiction; thus U_i cannot exist.

There is one nuance. The conditional independence in Equation (5) cannot rule out the existence of a single-cause confounder. (Again see Figure 1, where S_i is a single-cause confounder.) Conditional independence still holds, even if such a confounder exists.

[4]Figure 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.

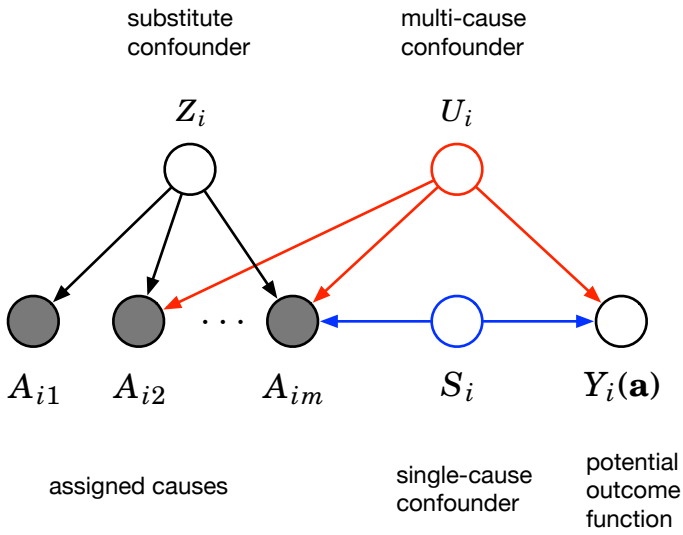


Figure 1: A graphical model argument for the deconfounder. The punchline is that if Z_i renders the A_{ij}'s conditionally independent then there cannot be a multi-cause confounder. The proof is by contradiction. Assume conditional independence holds, $p(a_{i1},...,a_{im} \mid z_i) = \prod_j p(a_{ij} \mid z_i)$; if there exists a multi-cause confounder U_i (red) then, by d-separation, conditional independence cannot hold (Pearl, 1988). Note we cannot rule out the single-cause confounder S_i (blue).

Here is the punchline. If we find a factor model that captures the population distribution of assigned causes then we have essentially discovered a variable that captures all multiple-cause confounders. The reason is that multiple-cause confounders induce dependence among the assigned causes, regardless of how they connect to the potential outcome function. Modeling their dependence, for which we have observations, provides a way to estimate variables that capture those confounders. This is the blessing of multiple causes.

Next we attend to some of the practical details of the deconfounder. The ingredients of the deconfounder are (1) a factor model of assigned causes (2) a way to check that the factor model captures their population distribution and (3) a way to estimate the conditional expectation $\mathbb{E}\left[Y_i(\boldsymbol{a}) \mid \hat{\boldsymbol{Z}}_i, \boldsymbol{A}_i = \boldsymbol{a}\right]$ for performing causal inference. We discuss each of these ingredients below (Section 2.3 and Section 2.4) and describe the full deconfounder algorithm (Section 2.5). We then answer questions that may come up for the reader (Section 2.6) and connect the deconfounder to existing methods in the research literature (Section 2.7).

2.3 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. As we described above, factor models include 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 (which we will use as a substitute confounder) and a per-cause parameter θ_j.

Example factor models. The deconfounder requires that the investigators find a factor model of the assigned causes and then use the factor model to estimate a local posterior or posterior expectation from $p(z_i \mid \boldsymbol{a}_i)$. In the simulations and studies of Section 3, 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 interpretation (Tipping and Bishop, 1999), both z_i and the per-cause parameters θ_j are real-valued K-vectors. The model is

$$\begin{aligned} Z_{ik} &\sim \mathcal{N}(0, \lambda^2), \quad k = 1, \ldots, K, \\ A_{ij} \mid Z_i &\sim \mathcal{N}\left(z_i^\top \theta_j, \sigma^2\right), \quad j = 1, \ldots, m. \end{aligned} \tag{7}$$

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

If the assigned causes are counts then Poisson factorization (PF) might be appropriate (Schmidt et al., 2009; Cemgil, 2009; Gopalan et al., 2015). PF is a probabilistic version of nonnegative matrix factorization (Lee and Seung, 1999, 2001), where z_i and θ_j are positive K-vectors. The model is

$$\begin{aligned} Z_{ik} &\sim \text{Gamma}(\alpha_0, \alpha_1), \quad k = 1, \ldots, K, \\ A_{ij} \mid Z_i &\sim \text{Poisson}(z_i^\top \theta_j), \quad j = 1, \ldots, m. \end{aligned} \tag{8}$$

PF can be fit to large datasets with efficient variational methods. Variational methods, or other forms of approximate inference, can also be used to approximate $p(z_i \mid \boldsymbol{a}_i)$.

A final example of a factor model is the deep exponential family (DEF) (Ranganath et al., 2015). A DEF is a probabilistic version of a deep neural network, generalizing on the classical sigmoid belief network of Neal (1990). For example, a two-layer DEF models each observation as

$$\begin{aligned} Z_{2,il} &\sim \text{Exp-Fam}_2(\alpha), \quad l = 1, \ldots, L, \\ Z_{1,ik} \mid Z_{2,i} &\sim \text{Exp-Fam}_1(g_1(z_{2,i}^\top \theta_{1,k})), \quad k = 1, \ldots, K, \\ A_{ij} \mid Z_{1,i} &\sim \text{Exp-Fam}_0(g_0(z_{1,i}^\top \theta_{0,j})), \quad j = 1, \ldots, m, \end{aligned} \tag{9}$$

where Exp-Fam stands for an exponential family distribution, θ_* are parameters, and $g_*(\cdot)$ are link functions. Each layer of the DEF is a generalized linear model. The DEF inherits the flexibility of deep neural networks, but uses exponential families to allow for different types of layered latent representations and data. For example, if the assigned causes are counts then Expfam_0 can be Poisson. Approximate inference in DEF can be performed with black box variational methods (Ranganath et al., 2014).

Predictive checks for the assignment model. The deconfounder rests on finding a good factor model, one that 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 the assignments that would have been observed under the model.

Checking the assignment model in this way blends a circle of related ideas around posterior predictive checks (PPCs) (Rubin, 1984), PPCs with realized discrepancies (Gelman et al., 1996), PPCs with held-out data (Gelfand et al., 1992), and stage-wise checking of hierarchical models (Dey et al., 1998; Bayarri and Castellanos, 2007). It also relates to Bayesian causal model criticism (Tran et al., 2016b) and PPCs in genome-wide association studies (GWAS) (Mimno et al., 2015).

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

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

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

$$p(\boldsymbol{a}_{i,\text{held}}^{\text{rep}} \mid \boldsymbol{a}_{i,\text{obs}}) = \int p(\boldsymbol{a}_{i,\text{held}} \mid z_i) p(z_i \mid \boldsymbol{a}_{i,\text{obs}}) \, \mathrm{d}z_i. \tag{10}$$

This distribution integrates out the local posterior $p(z_i \mid \boldsymbol{a}_{i,\text{obs}})$. (An approximate posterior also suffices; we discuss why in Section 2.6.5.)

Then compare the replicated data to the held-out data. To compare, calculate the expected log probability

$$t(\boldsymbol{a}_{i,\text{held}}) = \mathbb{E}_Z \left[\log p(\boldsymbol{a}_{i,\text{held}} \mid Z) \mid \boldsymbol{a}_{i,\text{obs}} \right], \tag{11}$$

which relates to their 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 p-value,

$$p\text{-value} = p\left(t(\boldsymbol{a}_{i,\text{held}}^{\text{rep}}) < t(\boldsymbol{a}_{i,\text{held}}) \right). \tag{12}$$

Here the randomness stems from $\boldsymbol{a}_{i,\text{held}}^{\text{rep}}$ coming from the predictive distribution in Equation (10), and we approximate the p-value with Monte Carlo.

How to interpret the p-value? A good model will produce values of the held-out causes that give similar log likelihoods to their real values—the p-value will not be extreme. A mismatched model will produce an extreme p-value, often where the replicated data has much higher log likelihood than the real data. Figure 2 illustrates a predictive check of a good assignment model. Section 3 shows predictive checks in action.

2.4 The outcome model

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

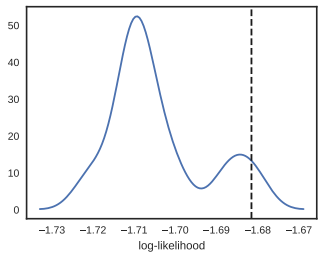


Figure 2: Predictive checks for the assignment model. The vertical dashed line shows $t(\boldsymbol{a}_{i,\text{held}})$. The blue curve shows the kernel density estimate (KDE) of $t(\boldsymbol{a}_{i,\text{held}}^{\text{rep}})$. The p-value is the area under the blue curve to the left of the vertical dashed line. The p-value here is larger than 0.5; the assignment model is good.

Suppose $p(z_i \mid \boldsymbol{a}_i, \mathcal{D})$ concentrates around a point $\hat{z}_i$. Then we can use $\hat{z}_i$ as a confounder. Follow Section 2.1 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 $\mathbb{E}[\mathbb{E}[Y_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i]]$. First, approximate the outside expectation with Monte Carlo,

$$\mathbb{E}[\mathbb{E}[Y_i(\boldsymbol{a}) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i]] \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_Y[Y_i(\boldsymbol{A}_i) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = \hat{z}_i]. \tag{13}$$

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

$$\mathbb{E}[Y_i(\boldsymbol{A}_i) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = z] = f(\boldsymbol{a}, z). \tag{14}$$

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

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

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

One outcome model we consider is a simple linear function,

$$f(\boldsymbol{a}, z) = \beta^\top \boldsymbol{a} + \gamma^\top z. \tag{15}$$

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

$$f(\boldsymbol{a}, z) = \beta^\top \boldsymbol{a} + \gamma^\top \hat{\boldsymbol{a}}(z). \tag{16}$$

It relates closely to the generalized propensity score (Imbens, 2000; Hirano and Imbens, 2004). Equation (16) can be seen as using $\hat{a}(z)$ as a proxy for the propensity score, a substitution that is used in Bayesian statistics (Laird and Louis, 1982; Tierney and Kadane, 1986; Geisser et al., 1990); this substitution is justified when higher moments of the assignment are similar across units. In both models, the coefficient β represents the average causal effect of raising each cause by one unit.

But we are not restricted to linear models. Other outcome models like random forests (Wager and Athey, 2017) and Bayesian additive regression trees (Hill, 2011) all apply here.

Note that devising an outcome model is just one approach to approximating the inner expectation of Equation (13). Another approach is again to use Monte Carlo. There are several possibilities. In one, group the confounder $\hat{z}_i$ into bins and approximate the expectation within each bin. In another, bin by the propensity score $p(a_i \mid \hat{z}_i)$ and approximate the inner expectation within each propensity-score bin (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004). A third possibility (if the assigned causes are discrete) 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).

2.5 The full algorithm, and an example

We described each component of the deconfounder. Algorithm 1 gives the full algorithm, a procedure for estimating Equation (13). The steps are: (1) find and fit a satisfactory factor model; (2) estimate $\hat{z}_i$ for each datapoint; (3) find and fit a satisfactory outcome model; (4) use the outcome model and estimated $\hat{z}_i$ to do causal inference.

Example. As an example, we 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 in humans? Here we give a brief account of how to use the deconfounder, omitting many of the details. We analyze GWAS problems extensively in Section 3.2.

Imagine we collect a dataset of $n = 5{,}000$ individuals; for each individual, we measure height and genotype, specifically the alleles at $m = 100{,}000$ locations, called the single-nucleotide polymorphisms (SNPs). Each SNP is represented by a count of 0, 1, or 2; it encodes how many of the individual's two nucleotides differ from the most common pair of nucleotides at the location. Table 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 a mixture of heterogenous populations (Pritchard et al., 2000b). We then generate the individual's height from a linear model of the SNPs (i.e. the assigned causes) and some simulated (but assumed unobserved) confounders. In this simulated data, the coefficients of the SNPs are the true causal effects; we denote them $\beta^* = (\beta_1^*, \ldots, \beta_m^*)$. See Section 3.2 for more details of the simulation.

Algorithm 1: The Deconfounder

Input: a dataset of assigned causes and outcomes $\{(\boldsymbol{a}_i, y_i)\}$, $i = 1, \ldots, n$
Output: the average potential outcome $\mathbb{E}[Y(\boldsymbol{a})]$ for any causes $\boldsymbol{a}$
repeat
  choose an assignment model from the class in Equation (4)
  fit the model to the assigned causes $\{\boldsymbol{a}_i\}$, $i = 1, \ldots, n$
  check the fitted model $\hat{M}$
until *the assignment check is satisfactory*
foreach *datapoint* i **do**
  calculate $\hat{z}_i = \mathbb{E}_{\hat{M}}[Z_i \mid \boldsymbol{a}_i]$.
end
repeat
  choose an outcome model from Equation (14)
  fit the outcome model to the augmented dataset $\{(\boldsymbol{a}_i, y_i, \hat{z}_i)\}$, $i = 1, \ldots, n$
  check the fitted outcome model
until *the outcome check is satisfactory*
estimate the average potential outcome $\mathbb{E}[Y(\boldsymbol{a})]$ by Equation (13)

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

We apply the deconfounder: model the assigned causes, infer a substitute confounder, and perform causal inference. To infer a substitute confounder, we fit a factor model of the assigned causes. Here we fit a 50-factor PF model, as in Equation (8). This fit results in estimates of non-negative factors $\hat{\theta}_j$ for each assigned cause (a K-vector) and non-negative weights $\hat{z}_i$ for each individual (also a K-vector).

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 3.2.

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

$$y_i \sim \mathcal{N}\left(\beta_0 + \beta^\top \boldsymbol{a}_i + \gamma^\top \hat{\boldsymbol{a}}(z_i), \sigma^2\right). \tag{17}$$

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 ($a_{i,9}$)	$\cdots$	SNP_100K ($a_{i,100K}$)	Height (feet) (y_i)
1	1	0	0	1	0	0	1	2	0	$\cdots$	0	5.73
2	1	2	2	1	2	1	1	0	1	$\cdots$	2	5.26
3	2	0	1	1	0	1	0	1	1	$\cdots$	2	6.24
4	0	0	0	1	1	0	1	2	0	$\cdots$	0	5.78
5	1	2	1	1	1	0	1	0	0	$\cdots$	1	5.09
6	2	2	1	0	0	2	0	1	1	$\cdots$	1	6.36
7	1	0	0	0	1	2	0	0	0	$\cdots$	2	5.51
8	1	2	0	0	1	2	0	0	0	$\cdots$	1	5.73
9	1	0	1	0	0	0	1	1	0	$\cdots$	0	6.51
10	1	1	0	0	0	2	0	0	1	$\cdots$	2	5.45
$\vdots$						$\vdots$						$\vdots$

Table 2: How do SNPs causally affect height? This table shows a portion of a dataset: simulated SNPs as the multiple causes and height as the outcome.

	w/o deconfounder	w/ deconfounder
RMSE$\times 10^2$	3.73	3.67

Table 3: Root mean squared error (RMSE) of the causal coefficients $\hat{\beta}$ with and without the deconfounder in a GWAS simulation study. We treat this RMSE as a metric of how close the estimated potential outcome function is to the truth. In this toy problem, the deconfounder produces closer-to-truth causal estimates.

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

Table 3 evaluates 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 $\boldsymbol{a}_i$.) The deconfounder produces closer-to-truth causal estimates.

2.6 A conversation with the reader

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

2.6.1 Why do I need multiple causes?

The deconfounder uses latent variables to capture dependence among the assigned causes. The theory in Section 4 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. The deconfounder requires multiple causal inference.

2.6.2 Does the deconfounder rely on assumptions?

There is no causal inference without assumptions. The deconfounder relies on *single strong ignorability*, that we observe any confounders that affect only one of the observed causes; see Figure 1. This assumption is weaker than the classical assumption of strong ignorability; we no longer need to observe all confounders.

Single strong ignorability is often plausible, and especially so when working work with many parallel causes. Consider the GWAS problem. If a confounder affects SNPs—and we observe 100,000 SNPs per unit—then the confounder is unlikely to have an affect on only one. The same reasoning can apply to other settings—medications in medical informatics data, actors in movie revenue data, neurons in neuroscience recordings, and vocabulary terms in text data.

By the same token, single strong ignorability may be compromised when we do not observe enough assigned causes. Consider a neuroscience problem where we are interested in the relationship between brain activity and animal behavior, but we only record 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.

2.6.3 Should I condition on known confounders and covariates?

Suppose we also observe X_i, known confounders and other covariates. Should we condition on them?

The deconfounder continues to maintain its good theoretical properties when we condition on observed covariates X_i as well as infer a substitute confounder Z_i. In particular, if X_i is "pre-treatment" —it does not include any mediators—then the causal estimate will be unbiased (Imai and Van Dyk, 2004) (also see Corollary 7 below). In general, it is good to condition on observed confounders, especially if they may contain 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 trade off. 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 3.1 explores this phenomenon.

2.6.4 Why does the deconfounder have two stages? Can I fit the assignment model and outcome model jointly?

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?

The main reason for two stages 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 might not have been measured.

In a related question, Algorithm 1 fits a factor model of the assigned causes and then uses the inferred variables in a model of the outcome. Should we forgo the convenience of two-stage estimation and fit these two models jointly?

We recommend that the investigator not include the outcome $y_i(\boldsymbol{a}_i)$ in the factor model. In theory, one can infer a substitute confounder $\boldsymbol{Z}_i$ that renders the assigned causes independent of each other and independent of the potential outcome. But this asks more of the model than needed: a $\boldsymbol{Z}_i$ that renders the assigned causes independent is sufficient for constructing a substitute confounder. Indeed, such a $\boldsymbol{Z}_i$ will necessarily render the assigned causes independent of the potential outcome function; if it is not then the assigned causes become conditionally dependent (again, see Figure 1).

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

In addition to these reasons, Section 4 details more technical reasons to separate the two stages.

2.6.5 Does the factor model of the assigned causes need to be the true assignment model? Do I need to be able to exactly infer the substitute confounder?

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 $\boldsymbol{Z}_i$ 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 an important step. If the deconfounder captures the population distribution—if the predictive check passes—then we can use the inferred local variables $\boldsymbol{Z}_i$ as substitute confounders.

For the same reason, the deconfounder can rely on approximate inference methods to infer the substitute confounder. The predictive check evaluates whether $\boldsymbol{Z}_i$ provides a good predictive distribution, regardless of how it was inferred. As long as the model and (approximate) inference method together give a good predictive distribution—again, 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 3.

2.6.6 Does the factor model need to be identifiable?

If a factor model captures the population distribution of the assigned causes then it can produce a substitute confounder. In Section 3, we study factor models where Z_i is identifiable—this is a byproduct of probabilistic factor models, i.e. where there is a prior on Z_i.

When the factor model identifies the latent variables Z_i then we can fit an outcome model with either the substitute confounder z (Equation (15)) or the reconstructed causes $\hat{\boldsymbol{a}}(z)$ (Equation (16)). Again, these factor models need not reflect a true mechanism in the world; see Section 2.6.5.

Suppose we choose a factor model that does not identify Z_i. D'Amour (2018) worries about causal identification, whether the causal estimate can converge to a point mass at the true causal parameter. D'Amour (2018) points out that causal identification does not hold if Z_i is not identifiable; the distribution of causal estimate can not converge to a point mass in the large-data limit. If causal identification is in question, sensitivity analysis can help (Robins et al., 2000; Gilbert et al., 2003; Imai and Van Dyk, 2004). Further, with additional assumptions, causal identification can be established; see Appendix A for an example.

In practice, the deconfounder only requires that the generalized propensity score $p(\boldsymbol{a} \mid z)$ to be identified. (Any model with an explicit likelihood will identify $p(\boldsymbol{a} \mid z)$.) In Section 3, we evaluate the uncertainty of the deconfounder estimates. The distribution of the estimates reflects how the (finite) observed data informs the causal effect.

2.7 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 2.3) and a conditional outcome model (Section 2.4).

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 2.6.5.

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 2.5.

Linear mixed models. The linear mixed model (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 $\boldsymbol{a}_{1:n}$. However, this is mathematically equivalent to posterior latent factors from a one-dimensional PCA model. Subject to its capturing the distribution of SNPs, the LMM is performing 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 Equation (7). Fitting an outcome model to its residuals is equivalent to conditioning on the reconstructed assignments, Equation (16).

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

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

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

With LFA in hand, Song et al. (2015) use inverse regression to perform association tests. Their approach is equivalent to assuming an outcome model conditional on the reconstructed assignments $a(\hat{z}_i)$, again Equation (16), 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.

3 Empirical studies

We study the deconfounder in three empirical studies. Section 3.1 and Section 3.2 involve simulations of realistic scenarios: we generate semi-synthetic data about smoking and genetics. Section 3.3 analyzes real data about actors and movie revenue. These 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 close-to-truth causal estimates.

The deconfounder requires computation at all stages—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 et al., 2014) as implemented in Edward, a probabilistic programming system (Tran et al., 2017, 2016a). (This was one choice; the deconfounder can be used with other methods for calculating the posterior and fitting models.)

3.1 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) includes information about smoking habits and medical expenses in a representative sample of the U.S. population (Imai and Van Dyk, 2004; US Department of Health and Human Services Public Health service, 1987). It contains 9,708 people and 8 variables about each. For each person, we focus on the age of starting to smoke (a_{age}), the cumulative exposure to smoking (a_{exp}), and whether he or she uses a seatbelt (a_{belt}). (The variables a_{age} and a_{exp} are positive reals; we took log transformations.)

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. Our true model is linear,

$$y_i = \beta_{\text{age}}\, a_{\text{age},i} + \beta_{\text{exp}}\, a_{\text{exp},i} + \beta_{\text{belt}}\, a_{\text{belt},i} + \varepsilon_i, \tag{18}$$

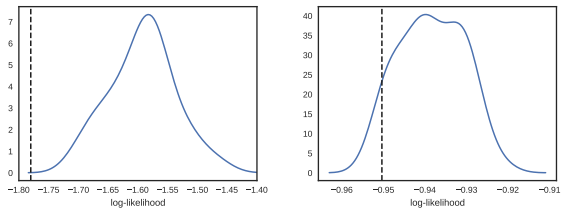
where $\varepsilon_i \sim \mathcal{N}(0,1)$. We set the true causal coefficients as

$$\beta_{\text{age}} = 0.8 \quad \beta_{\text{exp}} = 0.3 \quad \beta_{\text{belt}} = 0.1. \tag{19}$$

and from these coefficients we generate a full dataset of 9,708 tuples $(a_{\text{age},i}, a_{\text{exp},i}, a_{\text{belt},i}, y_i)$. This is semi-synthetic data: the assigned causes are from the real world, but we know the true outcome model. Note that seatbelt usage is a confounder—it is correlated to both age and exposure (each at about 0.2) and is one of the causes of the expenses.

Suppose we are interested in the causal effects of smoking age and total exposure on medical expenses. Further suppose we do not observe seatbelt usage; it is an unobserved confounder. We can use the deconfounder to solve the problem.

Modeling the assigned causes. We begin by finding a good factor model of the assigned causes $(a_{\text{age},i}, a_{\text{exp},i})$. Because there are two observed assigned causes, we consider models with a single scalar latent variable. (See Section 4.) We consider two factor models.



(a) Linear Model **(b)** Quadratic Model

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 kernel density estimate (KDE) of the test-statistic based on the predictive distribution. The dashed vertical line shows the value of the test-statistic on the observed dataset. The figure shows that the linear model mismatches the data—the observed statistic falls in a low probability region of the KDE. The quadratic factor model is a better fit to the data.

The first is a linear factor model,

$$z_{\text{line},i} \sim \mathcal{N}(0,1) \tag{20}$$

$$a_{\text{age},i} = \eta^{(1)}_{\text{age}} z_{\text{line},i} + \eta^{(0)}_{\text{age}} + \epsilon_{i,\text{age}} \tag{21}$$

$$a_{\text{exp},i} = \eta^{(1)}_{\text{exp}} z_{\text{line},i} + \eta^{(0)}_{\text{exp}} + \epsilon_{i,\text{exp}}, \tag{22}$$

where all errors are standard normal. We fit this model with variational Bayes. Then we use the predictive check to evaluate it: following Section 2.3, we hold out a subset of the assigned causes and using the expected log probability as the test statistic. The resulting p-value is 0.005, which signals a model mismatch. See Figure 3 (a).

We next consider a quadratic factor model,

$$z_{\text{quad},i} \sim \mathcal{N}(0,1) \tag{23}$$

$$a_{\text{age},i} = \eta^{(1)}_{\text{age}} z_{\text{quad},i} + \eta^{(2)}_{\text{age}} z^2_{\text{quad},i} + \eta^{(0)}_{\text{age}} + \epsilon_{i,\text{age}} \tag{24}$$

$$a_{\text{exp},i} = \eta^{(1)}_{\text{exp}} z_{\text{quad},i} + \eta^{(2)}_{\text{exp}} z^2_{\text{quad},i} + \eta^{(0)}_{\text{exp}} + \epsilon_{i,\text{exp}}, \tag{25}$$

where all errors are standard normal. We again fit this model with variational Bayes and used a predictive check. The resulting p-value is 0.18, Figure 3 (b). This value gives the green light. We use posterior estimates $\hat{z}_i = \mathbb{E}[Z \mid A = a_i]$ to form a substitute confounder in a causal inference.

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

$$a_{\text{age},i}(\hat{z}_i) = \mathbb{E}_{\text{quad}}\left[A_{\text{age}} \mid Z = \hat{z}_i\right]. \tag{26}$$

See Equation (16).

	exposure to smoking	age of starting to smoke
Truth	0.3	0.8
Control for seatbelt use (oracle)	0.293 (0.015)	0.780 (0.026)
No control	0.317 (0.034)	1.357 (0.024)
Control for z_{line}	0.360 (0.014)	0.666 (0.024)
Control for $a(z_{\text{line}})$	0.292 (0.016) ✓	0.895 (0.023)
Control for z_{quad}	0.343 (0.017)	0.725 (0.047) ✓
Control for $a(z_{\text{quad}})$	**0.297 (0.015)** ✓	**0.808 (0.011)** ✓
Control for z_{quad}, x	0.335 (0.040) ✓	0.797 (0.018) ✓
Control fot $a(z_{\text{quad}}), x$	0.310 (0.059) ✓	0.822 (0.026) ✓

Table 4: Causal coefficients of `smoking-age` and `smoking-exposure`. (The numbers are mean(std). The check mark indicates the 95% credible interval includes the truth. "Control for xxx" means we include xxx as a covariate in the linear outcome model. X represents the set of covariates that include the confounder `belt`.) Not controlling for confounders yields biased causal estimates. So does using deconfounder with a poor Z-model that fails model checking. Deconfounder with a good Z-model and a good outcome model produces unbiased causal estimates; controlling for the "reconstructed causes" $\hat{a}$ yields more efficient estimates than the substitute confounder Z. Using deconfounder along with covariates preserves the unbiasedness; yet, it inflates the variance. (The covariates include seat belt usage, gender, race, marital status, education level.)

We use predictive checks to evaluate the outcome models. Conditioning on $\hat{z}$ gives a p-value of 0.05; conditioning on $a(\hat{z})$ gives a p-value of 0.16. 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_{age} and β_{exp} of Equation (18). We emphasize that Equation (18) is the true mechanism of the simulated world, which the deconfounder does not have access to. The linear model we posit for $\mathbb{E}\left[Y(A_{\text{age}}, A_{\text{exp}}) \mid A, Z\right]$ is a functional form for the expectation we are trying to estimate.

Performance. We compare all combinations of factor model (linear, quadratic) and outcome-expectation model (conditional on $\hat{z}_i$ or $a(\hat{z}_i)$). Table 4 gives the results, reporting the estimated causal coefficients for each combination.

Table 4 also reports the true values, the estimates if we had observed the seatbelt 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; the 95% credible intervals do not include the true value. Observing the confounder corrects the problem.

	p-value	Real-valued outcome root mean squared error (RMSE)$\times 10^2$	Binary outcome RMSE$\times 10^2$
No control	—	6.55	5.75
Control for confounders*	—	6.54	5.75
(G)LMM	—	6.54	**5.74**
PPCA	0.14	6.52	**5.74**
PF	0.16	6.53	**5.74**
LFA	0.14	6.54	**5.74**
GMM	0.01	6.54	**5.74**
DEF	0.19	**6.47**	**5.74**

Table 5: GWAS simulation I: Balding-Nichols Model. ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms the LMM; DEF performs the best among the five factor models; it also outperforms the "oracle" controlling-for-confounders approach that uses the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

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. (One estimate was still biased, but the outcome check revealed this issue.) Conditioning on the reconstructed causes $a(\hat{z}_i)$ improves efficiency, showing similar efficiency to the oracle setting. Using the deconfounder along with covariates preserves the unbiasedness of the causal estimates, but it inflates the variance.

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

3.2 Many causes: Genome-wide association studies

Genome-wide association studies (GWAS) address an important problem in quantitative 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 their genome.

One benefit of GWAS for causal analysis 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

	p-value	**Real-valued outcome** RMSE×10^2	**Binary outcome** RMSE×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 6: GWAS simulation II: 1000 Genomes Project (TGP). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms the LMM; DEF performs the best among the five factor models; it also outperforms the "oracle" controlling-for-confounders approach that uses the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

be many sources of correlation; common sources include population structure, i.e., 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 Section 2.7.)

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

We generate synthetic GWAS data. Following Song et al. (2015), we simulate genotypes $\boldsymbol{a}_{1:n}$ from an array of realistic models. These include models generated from real-world fits, models that simulate mixing of populations, and models that simulate a smooth spatial mixing of populations. For each model, we produce datasets of genotypes with 100,000 SNPs for 1000-5000 individuals. Appendix H details the configurations of the simulation.

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

$$y_i = \sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i. \tag{27}$$

To introduce further confounding effects, we group the individuals by their genotypes; the ith individual is in group c_i. (Appendix H 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 our empirical study, the group indicator of each individual is an unobserved confounder.

	p-value	**Real-valued outcome** RMSE$\times 10^2$	**Binary outcome** RMSE$\times 10^2$
No control	—	9.59	5.84
Control for confounders*	—	9.52	5.84
(G)LMM	—	9.57	5.84
PPCA	0.14	9.55	5.84
PF	0.13	9.56	5.84
LFA	0.14	9.54	5.84
GMM	0.03	9.59	5.84
DEF	0.16	**9.47**	**5.83**

Table 7: GWAS simulation III: Human Genome Diversity Project (HGDP). ("Control for confounders" means including the unobserved confounders as covariates.) The deconfounder outperforms the LMM; DEF performs the best among the five factor models; it also outperforms the "oracle" controlling-for-confounders approach that uses the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

In Equation (28), 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 99% 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 impose a low signal-to-noise ratio setting; we design the intercept and random effects such that the SNPs $\sum_j \beta_j a_{ij}$ contributes 10% of the variance, the per-group intercept λ_{c_i} contributes 20% , and the error ε_i contributes 70%.

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

$$y_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp(\sum_j \beta_j a_{ij} + \lambda_{c_i} + \varepsilon_i)}\right). \tag{28}$$

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

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

Deconfounding GWAS. We apply the deconfounder with five assignment models: 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 2.3.

With the fitted assignment model, we estimate the causal effects of the SNPs on the trait. For real-valued traits, we use a linear model conditional on the SNPs and the reconstructed causes $a(\hat{z})$; see Equation (16). Each assignment model gives a different form of $a(\hat{z})$. For the binary traits, we use a logistic regression, again conditional on the SNPs and reconstructed causes. We emphasize that these are not the true model of the outcome, but rather models of the random potential outcome function.

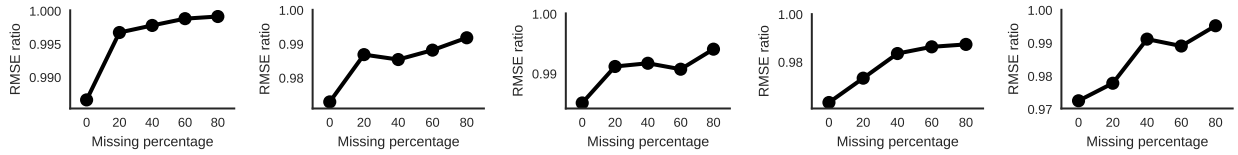
Performance. We study the deconfounder for GWAS. Tables 5 to 9 present the full results across the 11 different configurations. Each table is attached to a true assignment model and reports results across different factor models of the SNPs. For each factor model, the tables report the results of the predictive check and the root mean squared error (RMSE) of the estimated causal coefficients (for real-valued and binary- valued outcomes). Tables 5 to 9 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 LMM does not explicitly posit an assignment model so we omit the predictive check. It can be interpreted as the deconfounder though; see Section 2.7.) Notably, the deconfounder often outperforms the "oracle" regression where we include the (unobserved) confounder as a covariate; see Tables 5 to 8.

In general, predictive checks of the factor models reveal downstream issues with causal inference: better factor models of the assigned causes, as checked with 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 9). But checking the GMM signals this issue beforehand; the GMM constantly yields close-to-zero p-values 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 2.3. 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 Table 9. The spatial model produces spatially-correlated individuals; its parameter τ controls the spatial dispersion. (Consider each individual to sit in a unit square; as $\tau \to 0$, the individuals are placed closer to the corners of the unit square while when $\tau = 1$ they are distributed uniformly.) The six 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 "oracle" estimates that use the (unobserved) confounder information. Again, the predictive check hints at this issue. When the true distribution of SNPs is a spatial model, the p-values are generally more extreme (i.e., closer to zero).



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

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 mean more correction.) As the percentage of observed causes decreases, single strong ignorability is compromised; the deconfounder can no longer correct for all latent confounders.

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 *single strong ignorability*, that all single-cause confounders are observed. (See Section 2.6.2.) This assumption may be plausible when we measure all assigned causes but it may be compromised when we only observe a subset of causes—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 single strong ignorability to be compromised.

Summary. These studies provide three take-away messages: (1) the deconfounder can produce close-to-truth causal estimates, especially when we study many 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.

3.3 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.[5] It contains 901 actors (who appeared in at least five movies) and the revenue for the 2,828 movies they appeared in. The movies span 18 genres and 58 languages. (More than 60% of the movies are in English.) We focus on the cast and the log of the revenue. Note that this is a real-world observational data set. We no longer have ground truth of causal estimates.

[5] https://www.kaggle.com/tmdb

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 choose this "real world" problem in the hopes that it spawns intuitions in the reader.)

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

Deconfounded causal inference. We apply the deconfounder. We explore four assignment models: probabilistic principal component analysis (PPCA), Poisson factorization (PF), Gaussian mixture models (GMMs), and deep exponential familys (DEFs). (Each has 50 latent dimensions; the DEF has structure $[50, 20, 5]$.) We fit each model to the observed movie casts and check the models with a predictive check on held-out data; see Section 2.3.

The GMM fails its check, yielding a p-value < 0.01. The other models adequately capture patterns of actors: the checks return predictive p-values 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 lognormal regression, conditional on the observed casts and "reconstructed casts," Equation (16). While genres and languages of the movies are observed in the dataset, we rely solely on the deconfounder to debias the causal estimates.

Results: Predicting the revenue of uncommon movies. We consider testsets 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 and for observational data. In contrast, a non-causal model can produce incorrect predictions if we intervene on the causes (Peters et al., 2016). This idea of invariance has also been discussed in Haavelmo (1944); Aldrich (1989); Lanes (1988); Pearl (2009); Schölkopf et al. (2012); Dawid et al. (2010) under the terms "autonomy," "modularity," and "stability."

In one testset, we hold out 10% of non-English-language movies. (Most of the movies are in English.) Table 11 compares different models in terms of the average predictive log likelihood. The deconfounder predicts better than both the purely predictive approach (no control) and a classical approach, where we condition on the observed (pre-treatment) covariates.

In another testset, we hold out 10% of movies from uncommon genres, i.e., those that are not comedies, action, or dramas. Table 12 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 testset, 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 10 shows the average predictive log likelihood of the deconfounder and the purely predictive method. The deconfounder predicts slightly worse than the purely predictive method.

Exploratory analysis of actors and movies. We show how to use the deconfounder to explore the data, understanding the causal value of actors and movies.[6]

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

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

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

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

At the top of this list are blockbuster series. Among the top 25 include all of the *X-Men* movies, all of the *Avengers* movies, and all of the *Ocean's X* 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.

4 Theory

We develop theoretical results around the deconfounder.

[6]This section illustrates how to use the deconfounder to explore data. It is about these methods and the particular dataset that we studied, not a comment about the ground-truth quality of the actors involved. The authors of this paper are statisticians, not film critics.

First, we justify the use of factor models to find a substitute deconfounder. We show that if the factor model captures the distribution of the assigned causes then the substitute confounder renders the assignment strongly ignorable. We further show that such a factor model always exists. These results imply that the deconfounder does deconfound, and the deconfounder should always use a factor model.

Second, we establish properties of the substitute confounder. We show that it captures all multi-cause confounders and it does not capture any mediators. We define *single strong ignorability*, the assumption that all single-cause confounders are observed. We show that under this assumption the deconfounder yields unbiased causal inference.

4.1 Strong ignorability and the factor model

Does the deconfounder deconfound? Should the deconfounder always use a factor model? The answer to both questions is yes. The assigned causes are strongly ignorable given a substitute confounder Z if the assigned causes come from a factor model where Z is the local latent variable. Moreover, there always exists a factor model that describes the population distribution of the assigned causes. These results have two implications: (1) the substitute confounder renders the assigned causes strongly ignorable; (2) to find a substitute confounder, we can always fit a factor model to the assigned causes.

Recall the definition of strong ignorability, that the assigned causes are conditionally independent of the potential outcomes (Rosenbaum and Rubin, 1983).

Definition 1. *(Strong ignorability) Assigned causes are strongly ignorable given Z_i if*

$$(A_{i1},\ldots,A_{im}) \perp\!\!\!\perp Y_i(a_1,\ldots,a_m) \,|\, Z_i \tag{29}$$

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

Roughly, the assigned causes are strongly ignorable given Z_i if all confounders are captured by Z_i. More precisely, the assigned causes are strongly ignorable if all confounders are measurable with respect to the σ-algebra generated by Z_i.

To connect strong ignorability to factor models, we consider an intermediate construct, the "Kallenberg construction." The Kallenberg construction is inspired by the classical idea of randomization variables, Uniform[0,1] variables from which we can construct a random variable with an arbitrary distribution (Kallenberg, 1997). Below, we will use the Kallenberg construction of assigned causes as a bridge between the conditional independence statement in Equation (29) and the factor models of the deconfounder.

Definition 2. *(Kallenberg construction of assigned causes) The distribution of assigned causes $(A_{i1},\ldots,A_{im})$ admits a Kallenberg construction from a random variable Z_i taking values in $\mathcal{Z}$ if there exists (deterministic) measurable functions, $f_j : \mathcal{Z} \times [0,1] \to \mathcal{A}_j$ and random variables $U_{ij} \in [0,1]$ $(j = 1,\ldots,m)$ such that*

$$A_{ij} \overset{a.s.}{=} f_j(Z_i, U_{ij}), \tag{30}$$

where U_{ij} marginally follow Uniform[0,1] and jointly satisfy

$$(U_{i1},\dots,U_{im}) \perp\!\!\!\perp (Z_i, Y_i(a_1,\dots,a_m)) \tag{31}$$

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

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

Lemma 1. *(Kallenberg construction $\Leftrightarrow$ strong ignorability) The assigned causes are strongly ignorable 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.*

Proof sketch. First assume the Kallenberg construction in Equation (30). 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 $Y_i(\boldsymbol{a})$. Therefore $(A_{i1},\dots,A_{im})$ is conditionally independent of Y_i given Z_i, i.e., strong ignorability holds. Now assume that strong ignorability holds. We prove that this assumption implies a Kallenberg construction by building on the randomization variable construction of conditional distributions (Kallenberg, 1997). The full proof is in Appendix B. □

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 strongly ignorable. 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 good factor model to capture the distribution of assigned causes then we can use the fitted factor model to construct a substitute confounder.

Recall the definition of a factor model.

Definition 3. *(Factor model of assigned causes) Consider the assigned causes $\boldsymbol{A}_{1:n}$ and two independent sets of latent variables, $Z_{1:n}$ and $\theta_{1:m}$. A factor model of the assigned causes is a latent-variable model,*

$$p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n}) = p(\theta_{1:m})p(z_{1:n}) \prod_{i=1}^{n}\prod_{j=1}^{m} p(a_{ij} \mid z_i, \theta_j). \tag{32}$$

The distribution of assigned causes is the corresponding marginal,

$$p(\boldsymbol{a}_{1:n}) = \int p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n}) \,\mathrm{d}z_{1:n}\,\mathrm{d}\theta_{1:m}. \tag{33}$$

Further, a factor model of assigned causes requires that $\theta_{1:m}$ are point masses.

As we mentioned in Section 2.3, many common models from Bayesian statistics and machine learning can be written as factor models.

The next lemma connects the Kallenberg construction to factor models.

Lemma 2. *(factor models ⇒ Kallenberg construction) Under weak regularity conditions, every factor model of the assigned causes $p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n})$ admits a Kallenberg construction from $\boldsymbol{Z}_i$.*

Proof sketch. The lemma is an immediate consequence of Lemma 2.22 in Kallenberg (1997), single strong ignorability, and the following observation: $\theta_{1:m}$ are point masses, so they are *a priori* independent of the potential outcomes and the other latent variables,

$$(\theta_1, \ldots, \theta_m) \perp\!\!\!\perp (Y_i(\boldsymbol{a}), Z_i), \tag{34}$$

for any $\boldsymbol{a} \in \mathscr{A}_1 \times \ldots \times \mathscr{A}_m$. See Appendix C for the full proof. □

Lemma 1 and Lemma 2 connect strong ignorability to Kallenberg constructions and then Kallenberg constructions to factor models. The following theorem uses these results to justify the deconfounder.

Theorem 3. *(The Deconfounder) Under weak regularity conditions,*

1. *The assigned causes are strongly ignorable given a substitute confounder $\boldsymbol{Z}_i$ if the true distribution $p(\boldsymbol{a}_{1:n})$ can be written as a factor model that uses the substitute confounder, $p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n})$.*
2. *There always exists a factor model that captures the distribution of assigned causes.*

Proof sketch. The first part follows directly from Lemmas 1 and 2. The second part follows from the Reichenbach's common cause principle (Peters et al., 2017; Sober, 1976) 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. See Appendix D for the full proof. □

Theorem 3 confirms the validity of the deconfounder and justifies its use of factor models.

The first part of Theorem 3 suggests how to find a valid substitute confounder, one that renders the causes strongly ignorable. 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 directly 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 indicates that the deconfounder does deconfound.

The second part of Theorem 3 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.

4.2 Single strong ignorability and the substitute confounder

In the previous subsection, we focused on the deconfounder and its use of factor models. We now shift gears to study the substitute confounder. We prove several theoretical properties of the substitute confounder.

The first property is that the substitute confounder must capture all multi-cause confounders. If we additionally assume *single strong ignorability*—that we observe all single-cause confounders—then the substitute confounder and the observed covariates captures all confounders. This property shows that the inferred substitute confounder, together with the observed covariates, completely deconfounds causal inference.

The second property is 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.

The third property builds on the first two: conditioning the observed outcomes on both the substitute confounder and the observed covariates gives an unbiased estimate of the potential outcome function.

Throughout this subsection, we assume the substitute confounder comes from a factor model that fully captures the population distribution of the causes. (Definition 3 provides the definition of a factor model.)

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 4. *(Multi-cause confounder) A pretreatment covariate C_i is a multi-cause confounder if there exists a set of pretreament covariates V_i (possibly empty) and a set $J \subset \{1,\dots,m\}$ with $|J| \geq 2$ such that*

$$(A_{ij})_{j\in J} \perp\!\!\!\perp Y_i(a_{i1},\dots,a_{im}) | (V_i, C_i).$$

Moreover, there is no proper subset S_i of (V_i, C_i) and no proper subset J' of J such that $(A_{ij})_{j\in J'} \perp\!\!\!\perp Y_i(a_{i1},\dots,a_{im}) | S_i$.

The next proposition states that the substitute confounder must capture all multi-cause confounders.

Proposition 4. *Any multi-cause confounder C_i must be measurable with respect to the σ-algebra generated by the substitute confounder Z_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 strong ignorability. Now suppose we omitted a multi-cause confounder C_i. Then the substitute confounder Z_i could not have satisfied strong ignorability: 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 E gives a detailed proof. □

We showed that the substitute confounder Z_i includes all multi-cause confounders. How about single-cause confounders? Here we need an assumption, single strong ignorability. Single strong ignorability is the main assumption of the deconfounder. We first define single-cause confounders.

Definition 5. *(Single-cause confounder) A pretreatment covariate C_i is a single-cause confounder if there exists a set of pretreament covariates V_i (possibly empty) such that*

$$A_{ij} \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im})|(V_i,C_i), \tag{35}$$

where $j \in \{1,\ldots,m\}$. Moreover, there is no proper subset S_i of (V_i,C_i) that satisfies $A_{ij} \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im})|S_i$.

Single strong ignorability assumes all single-cause confounders are observed.

Definition 6. *(Single strong ignorability) Let X_i be the observed pretreatment covariates of unit i. Single strong ignorability requires*

$$A_{ij} \perp\!\!\!\perp Y_i(a_1,\ldots,a_m)|X_i, \tag{36}$$

for all $(a_{i1},\ldots,a_{im}) \in \mathcal{A}_1 \otimes \cdots \otimes \mathcal{A}_m$, $i = 1,\ldots,n$, and $j = 1,\ldots,m$.

Single strong ignorability is a weaker assumption than the strong ignorability. Strong ignorablity requires the joint conditional independence between *all* causes and the potential outcomes. By contrast, single strong ignorabiltiy only requires the marginal conditional independence with *individual* causes. As the number of causes increases, single strong ignorabilty becomes increasingly weak. (See the discussion in Section 2.6.2.)

As a consequence of single strong ignorabiltiy, the substitute confounder, together with the observed covariates, captures all confounders.

Corollary 5. *Under single strong ignorability, 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 single strong ignorability, a single-cause confounder must be measurable with respect to the observed covariates X_i. Because of Proposition 4, 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). □

Corollary 5 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 et al., 2010). The next proposition alleviates this concern.

Proposition 6. *Any mediator is almost surely not measurable with respect to the σ-algebra generated by the substitute confounder Z_i and the pre-treatment observed covariates X_i.*

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 F provides a detailed proof. □

Corollary 5 and Proposition 6 qualify the substitute confounder for mimicking real confounders. We can condition on substitute confounder as if they were observed covariates and proceed with causal inference as if strong ignorability holds. The next and final corollary echoes this intuition: under single strong ignorability, conditioning the observed outcome on both the substitute confounder and the observed covariates produces unbiased estimates of the average potential outcome function.

Corollary 7. *Under single strong ignorability, the deconfounder provides an unbiased estimate of the potential outcome function:*

$$\mathbb{E}[Y_i(a_1,\ldots,a_m)] = \mathbb{E}[\mathbb{E}[Y_i \mid Z_i, X_i, A_{i1} = a_1,\ldots,A_{im} = a_m]]. \tag{37}$$

Proof sketch. This corollary is an immediate consequence of Corollary 5 and Proposition 6. These results assert strong ignorability given the substitute confounders Z_i. We can thus treat the substitute confounder Z_i as if it were observed and, with strong ignorability satisfied, proceed with the classical causal inference. Appendix G gives a detailed proof. □

A note on overlap. We conclude this section with a discussion on *overlap*, roughly that any vector of assigned causes has positive probability given the substitute confounder. This assumption is often stated as the second half of strong ignorability (Imai and Van Dyk, 2004).

A substitute confounder Z_i is useful only when the assigned causes exhibit overlap,

$$p(\boldsymbol{A}_i \in \mathscr{A} \mid Z_i) > 0 \text{ for all } \mathscr{A} \subset \mathscr{A}_1 \otimes \cdots \otimes \mathscr{A}_m \text{ with positive measure.}$$

If overlap does not hold then the potential outcome function can become inestimable at values in the set $\mathscr{A}'$ where $p(\boldsymbol{A}_i \in \mathscr{A}' \mid Z_i) = 0$.

To enforce overlap, we constrain the allowable family of factor models. With continuous causes, we restrict to those models with continuous densities; if we additionally consider implicit models, we can restrict to those with a differentiable pushforward mapping from a Z_i lower-dimensional than the causes. (We assume the causes are full-rank, i.e., that no two causes are measurable with each other; if such a pair exists, merge them into a single cause.) With discrete causes, we can restrict to factor models with a continuous Z_i. For most probabilistic models, the overlap condition is easily satisfied.

Theoretically, if the model class is unconstrained, overlap is impossible to enforce. For any continuous random variables Z_i and A_i, regardless of their dimensionality, there exists a measurable function f such that $A_i \overset{a.s.}{=} f(Z_i)$. This is a consequence of Lemma 2.21 and Lemma 2.22 of Kallenberg (1997) and implies that, in theory, an exhaustive search for a good factor model might yield an "optimal" Z_i such that $Z_i \overset{a.s.}{=} A_i$. But this degeneracy rarely happens in practice.

5 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 is a blessing. Multiple causal inference liberates us from strong ignorability, providing causal inference from observational data under weaker assumptions than the classical approach requires.

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 several directions for future work. Here we focused on estimation; one direction is to develop a testing counterpart. How can we identify significant causes while still preserving family-wise error rate or false discovery rate? Here we analyzed univariate outcomes; another direction is to work with both multiple causes and multiple outcomes. Can dependence among outcomes further help causal inference?

Acknowledgements. We thank Alexander D'Amour, Alex Peysakhovich, Andrew Gelman, Barbara Engelhart, Edo Airoldi, Guido Imbens, Hal Stern, Jackson Loper, Jennifer Hill, José Zubizarreta, Léon Bottou, Qingyuan Zhao, Stefan Wager, Suresh Naidu, Victor Veitch, and Xinkun Nie for their valuable feedback on our manuscript.

		p-value	**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 8: GWAS 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 the "oracle" controlling-for-confounders approach that uses the (unobserved) confounder information. Predictive checking offers a good indication of when the deconfounder fails.

		p-value	**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 9: GWAS 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 the "oracle" controlling-for-confounders approach that uses the (unobserved) confounder information. Spatially-induced SNPs challenge many latent variable models to capture its patterns and fully deconfound causal inference. Predictive checking offers a good indication of when the deconfounder fails: GMM poorly captures the SNPs; it can amplify the error in causal estimates.

Control	Average predictive log-likelihood
No Control	**-1.1**
Control for X	**-1.1**
Control for $\hat{a}_{\text{PPCA}}$	-1.2
Control for $\hat{a}_{\text{PF}}$	-1.2
Control for $\hat{a}_{\text{DEF}}$	-1.2
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.3
Control for $(\hat{a}_{\text{PF}}, X)$	-1.2
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.2

Table 10: Average predictive log-likelihood on a holdout set of all movies. (X represents the observed covariates.) Causal models (the deconfounder) predicts slightly worse than prediction models.

Control	Average predictive log-likelihood
No Control	-2.5
Control for X	-2.1
Control for $\hat{a}_{\text{PPCA}}$	-1.6
Control for $\hat{a}_{\text{PF}}$	**-1.5**
Control for $\hat{a}_{\text{DEF}}$	**-1.5**
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.7
Control for $(\hat{a}_{\text{PF}}, X)$	**-1.5**
Control for $(\hat{a}_{\text{DEF}}, X)$	-1.6

Table 11: Average predictive log-likelihood on the holdout set of non-English movies. (X represents the observed covariates.) On a testset of uncommon movies, causal models with the deconfounder predict better than prediction models.

Control	Average predictive log-likelihood
No Control	-2.1
Control for X	-1.9
Control for $\hat{a}_{\text{PPCA}}$	-1.4
Control for $\hat{a}_{\text{PF}}$	**-1.2**
Control for $\hat{a}_{\text{DEF}}$	-1.3
Control for $(\hat{a}_{\text{PPCA}}, X)$	-1.4
Control for $(\hat{a}_{\text{PF}}, X)$	-1.3
Control for $(\hat{a}_{\text{DEF}}, X)$	**-1.2**

Table 12: Average predictive log-likelihood on the holdout set of non-drama/comedy/action movies. (X represents the observed covariates.) On a testset of uncommon movies, causal models with the deconfounder predict better than prediction models.

References

Airoldi, E., Blei, D., Erosheva, E., and Fienberg, S., editors (2014). *Handbook of Mixed Membership Models and Their Applications*. CRC Press.

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(Sep):1981–2014.

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, 41(1):15–34.

Astle, W., Balding, D. J., et al. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

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(6):1173.

Bayarri, M. and Castellanos, M. (2007). Bayesian checking of the second levels of hierarchical models. *Statistical Science*, 22:322–343.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent (d)irichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorization models. *Computational Intelligence and Neuroscience*, 2009.

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*.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405):51.

Collins, M., Dasgupta, S., and Schapire, R. E. (2002). A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624.

D'Amour, A. (2018). (Non-)identification in latent confounder models. http://www.alexdamour.com/blog/public/2018/05/18/non-identification-in-latent-confounder-models/. Accessed: 2018-05-29.

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.

Dawid, A. P., Didelez, V., et al. (2010). Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161.

Dey, D., Gelfand, A., Swartz, T., and Vlachos, P. (1998). Simulation based model checking for hierarchical models. *Test*.

Erosheva, E. A. (2003). Bayesian estimation of the grade of membership model. *Bayesian Statistics*, 7:501–510.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Resources*, 7(4):574–578.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697.

Geisser, S., Hodges, J., Press, S., and ZeUner, A. (1990). The validity of posterior expansions based on laplace method. *Bayesian and Likelihood Methods in Statistics and Econometrics*, 7:473.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Gelman, A., Meng, X., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541.

Gopalan, P., Hofman, J. M., and Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *Uncertainty in Artificial Intelligence*.

GTEx Consortium, T., 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.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, pages iii–115.

Hao, W., Song, M., and Storey, J. D. (2015). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721.

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

Heckman, J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. Technical report, National Bureau of Economic Research.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, 226164:73–84.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Imbens, G. and Rubin, D. (2015). *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.

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

Janzing, D. and Schölkopf, B. (2018b). Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810*.

Kallenberg, O. (1997). Foundations of modern probability. *Collection: Probability and Its Applications, Springer*.

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348.

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2017). CausalGAN: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*.

Laird, N. M. and Louis, T. A. (1982). Approximate posterior distributions for incomplete data problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 190–200.

Lanes, S. F. (1988). The logic of causal inference. *Causal Inference. ERI, Boston*, pages 59–75.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labor Market Policies*, pages 43–58. Springer.

Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562.

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.

Liu, F. and Chan, L. (2018). Confounder detection in high dimensional linear models using first moments of spectral measures. *arXiv preprint arXiv:1803.06852*.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3):284.

Lopez, M. J., Gutman, R., et al. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6449–6459.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, 23(19):2937–2960.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker.

Mimno, D., Blei, D. M., and Engelhardt, B. E. (2015). Posterior predictive checks to quantify lack-of-fit in admixture models of latent population structure. *Proceedings of the National Academy of Sciences*, 112(26):E3441–E3450.

Mohamed, S., Ghahramani, Z., and Heller, K. A. (2009). Bayesian exponential family pca. In *Advances in Neural Information Processing Systems*, pages 1089–1096.

Mohamed, S. and Lakshminarayanan, B. (2016). Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*.

Mooij, J. M., Stegle, O., Janzing, D., Zhang, K., and Schölkopf, B. (2010). Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695.

Morgan, S. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition.

Neal, R. M. (1990). Learning stochastic feedforward networks. *Department of Computer Science, University of Toronto*, 64(9).

Pearl, J. (1988). Probabilistic reasoning in intelligent systems.

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

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA.

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.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

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

Ranganath, R., Tang, L., Charlin, L., and Blei, D. (2015). Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771.

Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333.

Rassen, J. A., Solomon, D. H., Glynn, R. J., and Schneeweiss, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety*, 20(7):675–683.

Rezende, D. J. and Mohamed, S. (2015). Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*.

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*, pages 1–94. Springer.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schmidt, M. N., Winther, O., and Hansen, L. K. (2009). Bayesian non-negative matrix factorization. In *International Conference on Independent Component Analysis and Signal Separation*, pages 540–547. Springer.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*.

Sharma, A., Hofman, J. M., and Watts, D. J. (2016). Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv preprint arXiv:1611.09414*.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Sober, E. (1976). Simplicity.

Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554.

Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

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

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

Tran, D., Hoffman, M. D., Saurous, R. A., Brevdo, E., Murphy, K., and Blei, D. M. (2017). Deep probabilistic programming. In *International Conference on Learning Representations*.

Tran, D., Kucukelbir, A., Dieng, A. B., Rudolph, M., Liang, D., and Blei, D. M. (2016a). Edward: A library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*.

Tran, D., Ruiz, F. J., Athey, S., and Blei, D. M. (2016b). Model criticism for Bayesian causal inference. *arXiv preprint arXiv:1610.09037*.

US Department of Health and Human Services Public Health service (1987). National medical expenditure survey series (nmes).

VanderWeele, T. J. and Shpitser, I. (2013). On the definition of a confounder. *The Annals of Statistics*, pages 196–220.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted).

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.

Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73.

Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press.

Appendix

A The deconfounder with non-identifiable factor models

When the factor model identifies the latent variables Z_i then we can fit an outcome model with either the substitute confounder z (Equation (15)) or the reconstructed causes $\hat{\boldsymbol{a}}(z)$ (Equation (16)). Many factor models identify Z_i, probabilistic principal component analysis (Tipping and Bishop, 1999), mixture models (McLachlan and Basford, 1988), nonnegative matrix factorization (Lee and Seung, 1999, 2001), and tensor factorization (Anandkumar et al., 2014) for example. These models often involve constrained parameter space or prior distributions.

Suppose the factor model does not identify Z_i; this makes it complicated to condition on Z_i in the outcome model. But we can still use the deconfounder if the generalized propensity score $p(a_i \mid z_i)$ is identifiable. What we do is fit an outcome model with $\hat{\boldsymbol{a}}(z)$. Moreover, all factor models with an explicit likelihood identify the generalized propensity score (Rao, 1992; Verbeke and Molenberghs, 2017).

Causal identification with non-identifiable factor models requires extra caution. For example, when the causes are Gaussian and the outcome model is linear, see D'Amour (2018) for an example where causal effect becomes non-identifiable. In particular, the outcome model $\mathbb{E}[Y_i(\boldsymbol{A}_i) \mid \boldsymbol{A}_i = \boldsymbol{a}, Z_i = z]$ can not be identified from the observed data.

These non-identification issues can be resolved by making additional assumptions on the data generating process, for example additional independence constraints between the causes and the outcome (Miao et al., 2016) or distributional assumptions on the measurement error.

In practice, we should evaluate the uncertainty of the deconfounder estimate. The distribution of the estimate will reflect how the (finite) observed data informs the causal effect. When the causal effect is non-identifiable, the estimate will be uncertain over the ignorance region, that is the region of all possible causal effects.

In this section, we first explore the potential pathology of non- identifiable factor models. We then prove causal identification under distributional assumptions on the measurement error.

We start with restating the non-identification example from D'Amour (2018).

Denote the confounder as Z, the causes A, and the (observed) outcome Y:

$$Z_{n\times k} = \epsilon_Z, \tag{38}$$

$$A_{n\times m} = Z_{n\times k}\alpha_{k\times m} + \epsilon_A, \tag{39}$$

$$Y_{n\times 1} = A_{n\times m}\beta_{m\times 1} + Z_{n\times k}\gamma_{k\times 1} + \epsilon_Y, \tag{40}$$

where all random variables are zero mean Gaussian. In particular, the errors $\epsilon_A, \epsilon_Y, \epsilon_Z$ have diagonal covariance matrices. The above equations describe the true data generating process.

We observe the causes and the outcome; the confounder is unobserved. The goal is to estimate the coefficient β; it describes the causal effect of A on Y.

We observe the following covariance matrices:

$$\Sigma_{AA} = \alpha^\top \Sigma_{ZZ} \alpha + \sigma_A^2 \cdot I, \tag{41}$$
$$\Sigma_{AY} = \Sigma_{AA}\beta + \alpha^\top \Sigma_{ZZ}\gamma, \tag{42}$$
$$\Sigma_{YY} = \beta^\top \Sigma_{AA}\beta + \beta^\top \alpha^\top \Sigma_{ZZ}\gamma + \gamma^\top \Sigma_{ZZ}\gamma + \sigma_Y^2 \cdot I. \tag{43}$$

We can obtain two equally good solutions of $(\alpha, \beta, \gamma, Z, \sigma_A, \sigma_Y, \sigma_Z)$ by fixing γ and rescaling the latent variable Z (D'Amour, 2018).

The crux of this non-identifiability issue lies in Equation (42), where the $p \times 1$ matrix Σ_{AY} is mapped to a rank p matrix Σ_{AA} and a rank k matrix Σ_{ZZ}. This mapping is non-identifiable.

We can resolve this non-identifiability issue by assuming σ_Y is known. That is, we assume the measurement error in the outcome is $\mathcal{N}(0, \sigma_Y^2)$. This requires external information of the measurement error of the outcome. Such information could be obtained from running independent experiments on the outcome. We emphasize that this is an assumption on the data generating process.

We now prove causal identification assuming under this additional assumption on measurement error. We will write the causal parameter β in terms of observed quantities $\Sigma_{AA}, \Sigma_{AY}, \Sigma_{YY}$.

We first write the reconstructed causes $\hat{\boldsymbol{\alpha}}(z)$ as $W = Z\alpha$. Then we re-write the model Equation (40) as

$$Z_{n\times k} = \epsilon_Z, \tag{44}$$
$$W_{n\times m} = Z_{n\times k}\alpha_{k\times m} \tag{45}$$
$$A_{n\times m} = W_{n\times m} + \epsilon_A, \tag{46}$$
$$Y_{n\times 1} = A_{n\times m}\beta_{m\times 1} + W_{n\times m}\delta_{m\times 1} + \epsilon_Y, \tag{47}$$

where $\alpha\delta = \gamma$.

It leads to the following covariances:

$$\Sigma_{AA} = \Sigma_{WW} + \sigma_A^2 \cdot I \tag{48}$$
$$\Sigma_{WA} = \Sigma_{WW}, \tag{49}$$
$$\Sigma_{WY} = \Sigma_{WA}\beta + \Sigma_{WW}\delta, \tag{50}$$
$$\Sigma_{AY} = \Sigma_{AA}\beta + \Sigma_{AW}\delta. \tag{51}$$
$$\Sigma_{YY} = \beta^\top \Sigma_{AY} + \delta^\top \Sigma_{WW}\delta + \sigma_Y^2 \cdot I. \tag{52}$$

In particular, we observe $\Sigma_{AY}, \Sigma_{AA}, \Sigma_{YY}$. First of all, we can write Σ_{WW} and σ_A^2 in terms of Σ_{AA}:

$$\sigma_A^2 = \Sigma_{AA,11} - \Sigma_{AA,21} \cdot \frac{\Sigma_{AA,13}}{\Sigma_{AA,23}}, \tag{53}$$
$$\Sigma_{WW} = \Sigma_{AA} - \sigma_A^2 \cdot I. \tag{54}$$

Second, we re-arrange Equation (51):

$$\delta = \Sigma_{AW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) \tag{55}$$

Finally, we substitute Equation (55) into Equation (52)

$$\begin{aligned}
\Sigma_{YY} &= \beta^\top \Sigma_{AY} + (\Sigma_{AW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta))^\top \Sigma_{WW} \Sigma_{AW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) + \sigma_Y^2 \cdot I && (56)\\
&= \beta^\top \Sigma_{AY} + (\Sigma_{AY} - \Sigma_{AA}\beta)^\top \Sigma_{WW}^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) + \sigma_Y^2 \cdot I && (57)\\
&= \beta^\top \Sigma_{AY} + (\Sigma_{AY} - \Sigma_{AA}\beta)^\top (\Sigma_{AA} - \Sigma_{AA,11} - \Sigma_{AA,21} \cdot \frac{\Sigma_{AA,13}}{\Sigma_{AA,23}} \cdot I)^{-1}(\Sigma_{AY} - \Sigma_{AA}\beta) + \sigma_Y^2 \cdot I. && (58)
\end{aligned}$$

When σ_Y^2 is assumed known, then we can solve for the causal parameter β.

This calculation shows that we can identify the causal effect if the variance of the measurement error σ_Y^2 is known. Concretely, using outcome models like linear regression with a known measurement error or logistic regression will not result in the non-identification pathology. This aligns with our practices in the empirical studies in Section 3.

In the same vein of this calculation, causal effect will be identifiable if we observe two or more independent outcomes for each unit. We can use the off-diagonal terms of the covariance matrix Σ_{YY} to estimate β with a similar equation to Equation (58).

More generally, if (1) the distribution of measurement error is known, that is

$$P(Y \mid A, \hat{A}) = p_\epsilon(Y \mid f(A, \hat{A}))$$

with p_ϵ known and (2) f can be identifiable from $P(f(A, \hat{A}))$ and $P(A, \hat{A})$, then the mean potential outcome function can be identifiable.

Finally, we recommend performing sensitivity analysis of the hyperparameters in factor models when causal identification can potentially be an issue (Robins et al., 2000; Gilbert et al., 2003; Imai and Van Dyk, 2004). Sensitivity analysis evaluates downstream causal inference under different settings of assumptions, for example the hyperparameters in factor models: Do two factor models, both passing the predictive check, yield different causal inferences? Ideal causal inferences should be robust to changes in these assumptions.

B Proof of Lemma 1

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

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

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

$$(U_1, \ldots, U_m) \perp\!\!\!\perp (Z, Y(a_1, \ldots, a_m)) \tag{59}$$

for all $(a_1,\dots,a_m)$. By Proposition 5.18 in Kallenberg (1997), Equation (59) 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 $(a_1,\dots,a_m) \in \mathscr{A}_1 \otimes \cdots \otimes \mathscr{A}_m$. 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 $f_j : \mathscr{Z} \times [0,1] \to \mathscr{A}_j$ 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

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

Moreover, we have

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

with the same argument as the above necessity part.

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

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

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

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

In particular, we have

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

and

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

Since

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

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

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

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

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

Moreover, by

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

we have

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

there exists some measurable function $g_1' : \mathscr{Z} \times [0,1] \to [0,1]$ such that

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

and

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

In other words, we have

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

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

$$\begin{aligned}&(Y(a_1,\ldots,a_m), Z, U_1, U_2)\\ \stackrel{d}{=}\ &(Y(a_1,\ldots,a_m), Z, g_1'(Z, \hat{U}_1), g_2(Y(a_1,\ldots,a_m), Z, \hat{U}_1, \hat{U}_2))\end{aligned}$$

and

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

Again by

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

we have a measurable function $g_2' : \mathscr{Z} \times [0,1]^2 \to [0,1]$ that satisfies

$$\begin{aligned}&(Y(a_1,\ldots,a_m), Z, U_1, U_2)\\ \stackrel{d}{=}\ &(Y(a_1,\ldots,a_m), Z, g_1'(Z, \hat{U}_1), g_2'(Z, \hat{U}_1, \hat{U}_2)).\end{aligned}$$

Repeating these steps m times, we have

$$\begin{aligned}&(Y(a_1,\ldots,a_m), Z, U_1, U_2, \ldots, U_m)\\ \stackrel{d}{=}\ &(Y(a_1,\ldots,a_m), Z, g_1'(Z, \hat{U}_1), g_2'(Z, \hat{U}_1, \hat{U}_2), \ldots, g_m'(Z, \hat{U}_1, \hat{U}_2, \ldots, \hat{U}_m))\end{aligned}$$

with

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

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

$$(g_1'(Z, \hat{U}_1), g_2'(Z, \hat{U}_1, \hat{U}_2), \ldots, g_m'(Z, \hat{U}_1, \hat{U}_2, \ldots, \hat{U}_m)) \perp\!\!\!\perp_Z Y(a_1,\ldots,a_m).$$

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

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

□

C Proof of Lemma 2

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, $\mathscr{A}_j$, $j = 1,\ldots,m$ are Borel subsets of compact intervals. Without loss of generality, we could assume $\mathscr{A}_j = [0,1]$, $j = 1,\ldots,m$.

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

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

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

Equation (32) implies that $\omega_{ij}, i = 1,\ldots,n, j = 1,\ldots,m$ 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 $U_{ij} := F_{ij}^{-1}(h_{ij}(\theta_i, \omega_{ij}))$.

In particular, U_{ij} satisfies

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

It is because $\theta_{1:m}$ are point masses; they satisfy $(\theta_1,\ldots,\theta_m) \perp\!\!\!\perp (Z_i, Y_i(a_1,\ldots,a_m))$.

Moreover, $\omega_{ij} \overset{iid}{\sim} \text{Uniform}[0,1]$. We thus have

$$(\omega_{i1},\ldots,\omega_{im}) \perp\!\!\!\perp Y_i(a_1,\ldots,a_m) \,|\, Z_i.$$

It is because we assume no single-cause confounders: a single-cause confounder can induce dependence between one of ω_{ij} and $Y_i(a_1,\ldots,a_m)$; a multi-cause confounder cannot induce dependence between $(\omega_{i1},\ldots,\omega_{im})$ and $Y_i(a_1,\ldots,a_m)$ because ω_{ij}'s are independent.

More precisely, no single-cause confounder implies

$$\omega_{ij} \perp\!\!\!\perp Y_i(a_1,\ldots,a_m), j = 1,\ldots,m.$$

Because $\omega_{ij}, j = 1,\ldots,m$ are jointly independent, we have $(\omega_{i1},\ldots,\omega_{im})$ and $Y_i(a_1,\ldots,a_m)$. In particular, for $m = 2$, we have

$$p(Y_i(a_1,\ldots,a_m), \omega_{i1}, \omega_{i2})$$

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

This implies

$$
(\omega_{i1},\ldots,\omega_{im}) \perp\!\!\!\perp Y_i(a_1,\ldots,a_m).
$$

The last equality is because ω_{i2} is independent of ω_{i1} and $Y_i(a_1,\ldots,a_m)$. Given Z_i is inferred without any knowledge of $Y_i(a_1,\ldots,a_m)$), we have $(\omega_{i1},\ldots,\omega_{im}) \perp\!\!\!\perp Y_i(a_1,\ldots,a_m) \mid Z_i$.

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. □

D Proof of Theorem 3

Proof. 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 A_i, we consider a special case when $A_i \overset{a.s.}{=} Z_i$. We have

$$
p(a_{i1},\ldots,a_{im} \mid z_i) = \delta_{z_i} = \prod_{j=1}^{m} \delta_{z_{ij}} = \prod_{j=1}^{m} p(a_{ij} \mid z_i) \tag{60}
$$

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

$$
\begin{aligned}
p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n}) &= p(\theta_{1:m}) p(z_{1:n} \mid \theta_{1:m}) p(\boldsymbol{a}_{1:n} \mid z_{1:n}, \theta_{1:m}) && (61)\\
&= p(\theta_{1:m}) p(z_{1:n}) p(\boldsymbol{a}_{1:n} \mid z_{1:n}) && (62)\\
&= p(\theta_{1:m}) p(z_{1:n}) \prod_{i=1}^{n} \prod_{j=1}^{m} p(a_{ij} \mid z_i) && (63)
\end{aligned}
$$

The second equality is due to $Z_i \perp\!\!\!\perp \theta_{1:m}$ and $A_i \perp\!\!\!\perp \theta_{1:m} \mid Z_i$. They are because θ_j's are point masses. The third equality is due to the SUTVA assumption and Equation (60).

Choosing $Z_i \overset{a.s.}{=} A_i$, that is letting the substitute confounder Z_i be the same as the assigned causes A_i, does not help with causal inference; see a related discussion on overlap in Section 4.2.

This result is only meant to exemplify the large capacity of factor models. Finally, this $Z_i \overset{a.s.}{=} A_i$ example also illustrates the fact that a factor model capturing $p(\boldsymbol{a}_i)$ 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

$$
F(x,\lambda) := P(X < x) + \lambda P(X = x). \tag{64}
$$

Then if we construct U variables as

$$U := F(X, V), \tag{65}$$

then we have

$$U = F(X-) + V(F(X) - F(X-)), \tag{66}$$

$$U \overset{d}{=} Uniform(0,1), \tag{67}$$

$$X \overset{a.s.}{=} F^{-1}(U). \tag{68}$$

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

$$p(\theta_{1:m}, z_{1:n}, \boldsymbol{a}_{1:n}) = p(\theta_{1:m}) p(z_{1:n} \mid \theta_{1:m}) p(\boldsymbol{a}_{1:n} \mid z_{1:n}, \theta_{1:m}) \tag{69}$$

$$= p(\theta_{1:m}) p(z_{1:n}) p(\boldsymbol{a}_{1:n} \mid z_{1:n}, \theta_{1:m}) \tag{70}$$

$$= p(\theta_{1:m}) p(z_{1:n}) \prod_{i=1}^{n} \prod_{j=1}^{m} p(a_{ij} \mid z_i, \theta_j) \tag{71}$$

The second equality is due to $\theta_{1:m}$ being point masses; $\theta_j, j = 1, \ldots, m$ 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.

□

E Proof of Proposition 4

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 Equation (33).

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 nonconstant 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 $Y_i(a_{i1},\ldots,a_{im}) = h(W_{i,bad},\epsilon)$, where $\epsilon \perp\!\!\!\perp W_{i,bad}$.

These two statements implies that

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

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

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

It contradicts the fact that Z_i comes from the factor model (Equation (32)) with $(U_{i1},\ldots,U_{im}) \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im})|Z_i$. Therefore, there does not exist such a multi-cause confounder. □

F Proof of Proposition 6

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 $M_i(\boldsymbol{a}_i)$ is almost surely not measurable with respect to Z_i.

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

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

G Proof of Corollary 7

Proof. Lemma 1 and Lemma 2, together with single strong ignorability, ensures that the substitute confounder Z_i and the observed covariate X_i satisfies

$$(A_{i1},\ldots,A_{im}) \perp\!\!\!\perp Y_i(a_{i1},\ldots,a_{im})|Z_i,X_i. \tag{72}$$

Therefore, we have

$$\begin{aligned}
\mathbb{E}_{Z,X}[\mathbb{E}_Y[Y_i(\boldsymbol{A}_i)|\boldsymbol{A}_i=\boldsymbol{a},Z_i,X_i]] &= \mathbb{E}_{Z,X}[\mathbb{E}_Y[Y_i(\boldsymbol{a})|\boldsymbol{A}_i=\boldsymbol{a},Z_i,X_i]] &(73)\\
&= \mathbb{E}_{Z,X}[\mathbb{E}_Y[Y_i(\boldsymbol{a})|Z_i,X_i]] &(74)\\
&= \mathbb{E}[Y_i(\boldsymbol{a})], &(75)
\end{aligned}$$

where the multiple causes write $\boldsymbol{a} = (a_{i1},\ldots,a_{im})$. The first equality is due to SUTVA. The second equality is due to Equation (72). The last equality is due to the tower property.

□

H Details of Section 3.2

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 13 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.

First, without loss of generality, we set the first 1% of the m SNPs to be the true causal SNPs ($\beta_j \neq 0, \beta_j \overset{iid}{\sim} \mathcal{N}(0, 0.5)$. We set $\beta_j = 0$ for the rest of the SNPs.

Notice that the SNPs are affected by some latent population structure. We simulate the confounder λ_i and the random effects ϵ_i so that they depend on the latent population structure as well.

For the confounder λ_i, we first perform K-means clustering on the columns of S with $K = 3$ using Euclidean distance. This assigns each individual i to one of three mutually exclusive cluster sets $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$, where $\mathcal{S}_k \subset \{1, 2, \ldots, n\}$. Set $\lambda_j = k$ if $j \in \mathcal{S}_k, k = 1, 2, 3$.

We then simulate the random effects ϵ_i. Let $\tau_1^2, \tau_2^2, \tau_3^2 \overset{iid}{\sim} \text{InvGamma}(3,1)$, and set $\sigma_i^2 = \tau_k^2$ for all $j \in \mathcal{S}_i, k = 1, 2, 3$. Draw $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$.

We control the signal-to-noise ratio (SNR) to mimic the highly noisy nature of genome-wide association studies (GWAS) data sets: We let the causal signals $\sum_{j=1}^{m} \beta_j a_{ij}$ contribute to $\nu_{gene} = 0.1$ of the variance, the confounder λ_i contribute $\nu_{conf} = 0.2$, and the random effects ϵ_i contribute $\nu_{noise} = 0.7$.

We set

$$\lambda_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^{m} \beta_j a_{ij}\}_{i=1}^{n}}{\sqrt{\nu_{gene}}} \right] \left[\frac{\sqrt{\nu_{conf}}}{s.d.\{\lambda_i\}_{i=1}^{n}} \right] \lambda_i, \tag{76}$$

$$\epsilon_i \leftarrow \left[\frac{s.d.\{\sum_{j=1}^{m} \beta_j a_{ij}\}_{i=1}^{n}}{\sqrt{\nu_{gene}}} \right] \left[\frac{\sqrt{\nu_{noise}}}{s.d.\{\epsilon_i\}_{i=1}^{n}} \right] \epsilon_i. \tag{77}$$

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

$$y_{i,quant} = \sum_{j=1}^{m} \beta_j a_{ij} + \lambda_i + \epsilon_i, \tag{78}$$

$$y_{i,binary} \sim \text{Bernoulli}\left(\frac{1}{1 + \exp(\sum_{j=1}^{m} \beta_j a_{ij} + \lambda_i + \epsilon_i)} \right). \tag{79}$$

Model	Simulation details
Balding-Nichols Model (Balding-Nichols)	Each row i of Γ has i.i.d. three independent and identically distributed draws from the Balding- Nichols model: $\gamma_{ik} \overset{iid}{\sim} \text{BN}(p_i, F_i)$, where $k \in \{1,2,3\}$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HapMap data set, calculating its observed allele frequence 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. We simulate $n = 100000$ SNPs and $m = 5000$ individuals.
1000 Genomes Project (TGP)	The matrix Γ was generated by sampling $\gamma_{ik} \overset{iid}{\sim} 0.9 \times \text{Uniform}(0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. 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. We simulate $m = 100000$ and $n = 1500$, where m was determined by the number of individuals in the TGP data set.
Human Genome Diversity Project (HGDP)	Same as TGP but generating S with the HGDP genotype matrix.
Pritchard-Stephens-Donnelly (PSD)	Each row i of Γ has i.i.d. three independent and identically distributed draws from the Balding- Nichols model: $\gamma_{ik} \overset{iid}{\sim} \text{BN}(p_i, F_i)$, where $k \in \{1,2,3\}$. The pairs (p_i, F_i) are computed by randomly selecting a SNP in the HGPD data set, calculating its observed allele frequence and estimating its F_{ST} value using the Weir & Cockerham estimator (Weir and Cockerham, 1984). The estimator requires each individual to be assigned to a subpopulation, which were made according to the $K = 5$ subpopulations from the analysis in Rosenberg et al. (2002). The columns of S were sampled $(s_{1j}, s_{2j}, s_{3j} \overset{iid}{\sim} \text{Dirichlet}(\alpha, \alpha, \alpha)$ for $j = 1, \ldots, m, \alpha = 0.01, 0.1, 0.5, 1$. We simulate $m = 100000$ and $n = 5000$.
Spatial	The matrix Γ was generated by sampling $\gamma_{ik} \overset{iid}{\sim} 0.9 \times \text{Uniform}(0, 0.5)$, for $k = 1, 2$ and setting $\gamma_{i3} = 0.05$. The first two rows of S correspond to coordinates for each individual on the unit square and were set to be independent and identically distributed samples from $\text{Beta}(\tau, \tau), \tau = 0.1, 0.25, 0.5, 1$, while the third row of S was set to be 1, i.e. an intercept. As $\tau \to 0$, the individuals are placed closer to the corners of the unit square, while when $\tau = 1$, the individuals are distributed uniformly. We simulate $m = 100000$ and $n = 5000$.

Table 13: Simulating allele frequencies.

References

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832.

D'Amour, A. (2018). (Non-)identification in latent confounder models. http://www.alexdamour.com/blog/public/2018/05/18/non-identification-in-latent-confounder-models/. Accessed: 2018-05-29.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Kallenberg, O. (1997). Foundations of modern probability. *Collection: Probability and Its Applications, Springer*.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788.

Lee, D. D. and Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*, volume 84. Marcel Dekker.

Miao, W., Geng, Z., and Tchetgen, E. T. (2016). Identifying causal effects with proxy variables of an unmeasured confounder. *arXiv preprint arXiv:1609.08816*.

Rao, B. P. (1992). *Identifiability in stochastic models: characterization of probability distributions*. Academic Press.

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*, pages 1–94. Springer.

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.

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

Verbeke, G. and Molenberghs, G. (2017). Modeling through latent variables. *Annual Review of Statistics and Its Application*, 4:267–282.

Weir, B. S. and Cockerham, C. C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.