On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives[1]

Alexander D'Amour
Google AI

Abstract

Unobserved confounding is a central barrier to drawing causal inferences from observational data. Several authors have recently proposed that this barrier can be overcome in the case where one attempts to infer the effects of several variables simultaneously. In this paper, we present two simple, analytical counterexamples that challenge the general claims that are central to these approaches. In addition, we show that nonparametric identification is impossible in this setting. We discuss practical implications, and suggest alternatives to the methods that have been proposed so far in this line of work: using proxy variables and shifting focus to sensitivity analysis.

1 INTRODUCTION

Estimating causal effects in the presence of unobserved confounding is one of the fundamental challenges of casual inference from observational data, and is known to be infeasible in general (Pearl, 2009). This is because, in the presence of unobserved confounding, the observed data distribution is compatible with many potentially contradictory causal explanations, leaving the investigator with no way to distinguish between them on the basis of data. When this is the case, we say that the causal quantity of interest, or estimand, is not identified. Conversely, when the causal estimand can be written entirely in terms of observable probability distributions, we say the query is identified.

A recent string of work has suggested that progress can be made with unobserved confounding in the special case where one is estimating the effects of multiple interventions (causes) simultaneously, and these causes are conditionally independent in the observational data given the latent confounder (Wang and Blei, 2018; Tran and Blei, 2017; Ranganath and Perotte, 2018). The structure of this solution is compelling because it admits model checking, is compatible with modern machine learning methods, models real-world settings where the space of potential interventions is high-dimensional, and leverages this dimensionality to extract causal conclusions. Unfortunately, this work does not establish general sufficient conditions for identification.

In this paper, we explore some of these gaps, making use of two simple counterexamples. We focus on the central question of how much information about the unobserved confounder can be recovered from the observed data alone, considering settings where progressively more information is available. In each setting, we show that the information gained about the unobserved confounder is insufficient to pinpoint a single causal conclusion from the observed data. In the end, we show that parametric assumptions are necessary to identify causal quantities of interest in this setting. This suggests caution when drawing causal inferences in this setting, whether one is using flexible modeling and machine learning methods or parametric models.

Despite these negative results, we discuss how it is still possible to make progress in this setting under minor modifications to either the data collection or estimation objective. We highlight two alternatives. First, we discuss estimation with proxy variables, which can be used to identify causal estimands without parametric assumptions by adding a small number of variables to the multi-cause setting (Miao et al., 2016; Louizos et al., 2017). Secondly, we discuss sensitivity analysis, which gives a principled approach to exploring the set of causal conclusions that are compatible with the distribution of observed data.

[1]See arXiv for an expanded version of this paper: `https://arxiv.org/abs/1902.10286`.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

2 RELATED WORK

This paper primarily engages with the young literature on multi-cause causal inference whose primary audience has been the machine learning community. This line of work is motivated by several applications, including genome-wide association studies (GWAS) (Tran and Blei, 2017), recommender systems (Wang et al., 2018), and medicine (Ranganath and Perotte, 2018). Wang and Blei (2018) include a thorough review of this line of work and application areas.

These papers can be seen as an extension of factor models to causal settings. Identification in factor models is an old topic. The foundational results in this area are due to Kruskal (1989) and were extended to a wide variety of settings by Allman et al. (2009). For more elementary results similar to those in our first counterexample, see Bollen (1989) or similar introductory texts on factor analysis.

The approach taken in this paper is an example of sensitivity analysis, which is a central technique for assessing the robustness of conclusions in causal inference. One prominent approach, due to Rosenbaum and Rubin (1983) posits the existence of a latent confounder, and maps out the causal conclusions that result when unidentified parameters in this model are assumed to take certain values. Our second counterexample takes inspiration from the model suggested in this paper.

3 NOTATION AND PRELIMINARIES

Consider a problem where one hopes to learn how multiple inputs affect an outcome. Let $A = (A^{(1)}, \ldots, A^{(m)})$ be a vector of m variables (causes) whose causal effects we wish to infer, and let Y be the scalar outcome variable of interest. We write the supports of A and Y as $\mathcal{A}$ and $\mathcal{Y}$, repsectively. For example, suppose that A corresponds a set of genes that a scientist could, in principle, knock out by gene editing, where $A^{(k)} = 1$ if the gene remains active and $A^{(k)} = 0$ if it is knocked out. In this case, the scientist may be interested in predicting a measure of cell growth Y if various interventions were applied. Formally, we represent this quantity of interest using the *do*-operator:

$$P(Y \mid do(A)),$$

which represents the family of distributions of the outcome Y when the causes A are set to arbitrary values in $\mathcal{A}$ (Pearl, 2009).

In general, it is difficult to infer $P(Y \mid do(A))$ from observational, or non-experimental, data because there may be background factors, or confounders, that drive both the outcome Y and the observed causes A. We represent these confounders with the variable U, with support $\mathcal{U}$. In the presence of such confounders, the conditional distribution $P(Y \mid A)$ may be different from the intervention distribution $P(Y \mid do(A))$.

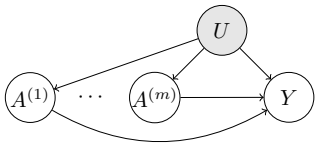

Figure 1: DAG representation of the multi-cause factor model setting (Assumption 1).

If U is observed, the following assumptions are sufficient to identify the intervention distribution:

- **Unconfoundedness**: U blocks all backdoor paths between A and Y, and
- **Positivity**: $P(A = a \mid U) > 0$ almost surely for each $a \in \mathcal{A}$.

Importantly, under these conditions, no parametric assumptions are necessary to identify the intervention distribution. We say that the intervention distribution is nonparametrically identified under these conditions.

The unconfoundedness assumption ensures that the following relation holds between the unobservable intervention distribution and observable distributions:

$$\begin{aligned} P(Y \mid do(A = a)) &= E[P(Y \mid do(A = a), U)] \\ &= E[P(Y \mid A = a, U)] \\ &= \int_{\mathcal{U}} P(Y \mid A = a, U = u) P(U = u) du. \end{aligned} \quad (1)$$

Meanwhile, the postivity assumption ensures that all pairs (A, U) are observable in the support of U, so that the integrand in (1) can be evaluated along each point on the path of the integral. The intervention distribution $P(Y \mid do(A = a))$ is identified under these conditions because (1) can be written completely in terms of observable distributions.

4 UNOBSERVED CONFOUNDING AND MULTIPLE CAUSES

When the confounder U is unobserved, the unconfoundedness and positivity assumptions are no longer sufficient for $P(Y \mid do(A = a))$ to be identified. In this case, additional assumptions are necessary because (1) is no longer a function of observable distributions.

The multi-cause approach attempts to infer (1) from the observable data (A, Y) alone under assumptions about the conditional independence structure of this distribution. Specifically, this approach incorporates

the assumption that the observed distribution of causes $P(A)$ admits a factorization by the unobserved confounder U. We group these central assumptions in Assumption 1 below, and illustrate them in Figure 1.

Definition 1. We say a variable U factorizes the distribution of a set of variables A iff

$$P(A) = \int_{\mathcal{U}} \left[\prod_{j=1}^{m} P(A^{(j)} \mid U = u) \right] P(U = u) du, \quad (2)$$

Assumption 1. There exists an unobserved variable U such that (i) U blocks all backdoor paths from A to Y and (ii) U factorizes the distribution of A.

Under this assumption, the most general form of multi-cause causal inference rests on the following identification claim.

Claim 1. *Under Assumption 1, for any variable V that factorizes A, the following relation $(*)$ holds*

$$\begin{aligned} &\int_{\mathcal{V}} P(Y \mid A = a, V = v) P(V = v) dv \quad (3) \\ &\qquad \overset{(*)}{=} \int_{\mathcal{U}} P(Y \mid A = a, U = u) P(U = u) du \\ &\qquad = P(Y \mid do(A = a)). \end{aligned}$$

If this claim were true, one could obtain an arbitrary factor model for the observed causes satisfying (2) and calculate $P(Y \mid do(A))$.

In Section 5, we present a simple counterexample that shows that this claim does not hold in general. The crux of the counterexample is that factorizations of $P(A)$ are not always unique, and differences between these factorizations can induce different values for (3).

In light of this counterexample, it is natural to ask whether identification by (1) is feasible in the special case that the factorization of $P(A)$ is unique. In this case, we say the factorization is identified. Depending on the specification, a factor model may be identified under fairly weak conditions, especially when the latent factor U is categorical; Allman et al. (2009) present a broad set of sufficient conditions.

Claim 2. *Under Assumption 1, if the factorization of $P(A)$ is identified, then the intervention distribution $P(Y \mid do(A = a))$ is identified by* (1).

Ranganath and Perotte (2018) and Tran and Blei (2017) make a variation of this claim by supposing that U can be consistently estimated as a function of A. In this case, the factorization is identified in the limit where the number of causes m grows large.

Claim 3. *Under Assumption 1, if there exists an estimator of U that is a function of A, $\hat{U}(A)$, such that*

$$\hat{U}(A) \xrightarrow{a.s.} U,$$

then the intervention distribution is identified by

$$\begin{aligned} &P(Y \mid do(A = a)) = \\ &\qquad \int_{\mathcal{U}} P(Y \mid A = a, \hat{U}(A) = u) P(\hat{U}(A) = u) du. \quad (4) \end{aligned}$$

In Section 6, we give a counterexample and a theorem showing that Claim 2 is false except in the trivial case that the observational and intervention distributions coincide; that is, $P(Y \mid do(A = a)) = P(Y \mid A = a)$. In building up to this result, we show that Claim 3 is false because its consistency premise implies that the positivity assumption is violated.

5 FACTORIZATION EXISTENCE IS INSUFFICIENT

5.1 Setup

In this section, we show that Claim 1 is false by a thorough exploration of a counterexample. Specifically, we show that, even under Assumption 1, the observed data can be compatible with many distinct intevention distributions $P(Y \mid do(A))$.

Consider a simple setting where all variables are linearly related, and all independent errors are Gaussian. Letting $\epsilon_w \sim N(0, \sigma_w^2)$ for each $w \in \{A, Y, U\}$, the structural equations for this setting are

$$\begin{aligned} U &:= \epsilon_U \\ A &:= \alpha U + \epsilon_A \\ Y &:= \beta^\top A + \gamma U + \epsilon_Y \end{aligned}$$

Here, α, β are $m \times 1$ column vectors, and γ is a scalar; ϵ_A is a $m \times 1$ random column vector, and ϵ_Y, ϵ_U are random scalars. This data-generating process satisfies Assumption 1.

Under this model, the intervention distribution has the following form:

$$P(Y \mid do(A = a)) = N(\beta^\top a, \gamma^2 \sigma_U^2 + \sigma_Y^2).$$

We will focus specifically on estimating the conditional $E[Y \mid do(A = a)]$, which is fully parameterized by β. Thus, our goal is to recover β from the distribution of observed data.

The covariance matrix can be written as

$$\Sigma_{AYU} = \begin{pmatrix} \Sigma_{UU} & \Sigma_{UA} & \Sigma_{UY} \\ \Sigma_{AU} & \Sigma_{AA} & \Sigma_{AY} \\ \Sigma_{YU} & \Sigma_{YA} & \Sigma_{YY} \end{pmatrix}$$

where Σ_{AA} is $m \times m$, $\Sigma_{AY} = \Sigma_{YA}^\top$ is $m \times 1$, and Σ_{YY} is 1×1.

The marginal covariance matrix of the observable variables (A, Y) is the bottom-right 2×2 sub-matrix of this matrix. Its entries are defined by:

$$\Sigma_{AA} = \alpha\alpha^\top\sigma_U^2 + \text{diag}(\sigma_A^2)$$
$$\Sigma_{AY} = \Sigma_{AA}\beta + \gamma\sigma_U^2\alpha$$
$$\Sigma_{YY} = (\beta^\top\alpha + \gamma)^2\sigma_U^2 + \beta^\top \text{diag}(\sigma_A^2)\beta + \sigma_Y^2$$

In these equations, the quantity on the LHS is observable, while the structural parameters on the RHS are unobservable. The goal is to invert these equations to obtain a unique value for β.

5.2 Equivalence Class Construction

When $m \geq 3$, the number of equations in this model exceeds the number of unknowns, but there still exists an equivalence class of structural equations with parameters

$$(\alpha_1, \beta_1, \gamma_1, \sigma_{U,1}^2, \sigma_{A,1}^2, \sigma_{Y,1}^2) \neq (\alpha, \beta, \gamma, \sigma_U^2, \sigma_A^2, \sigma_Y^2)$$

that induce the same observable covariance matrix, and for which $\beta_1 \neq \beta$. These parmeterizations cannot be distinguished by observed data. In this section, we show how to construct such a class.

The key to this argument is that the scale of U is not identified given A, regardless of the number of causes m. This is a well-known non-identification result in confirmatory factor analysis (e.g., Bollen, 1989, Chapter 7). In our example, the expression for Σ_{AA} does not change when α and σ_U^2 are replaced with the structural parameters $\alpha_1 := c \cdot \alpha$ and $\sigma_{U,1}^2 := \sigma_U^2/c^2$. In the following proposition, we state how the remaining structural variables can be adjusted to maintain the same observable covariance matrix when c is changed.

Proposition 1. *For any fixed vector of parameters $(\alpha, \beta, \gamma, \sigma_U^2, \sigma_A^2, \sigma_Y^2)$ and a valid scaling factor c (defined below), there exists a vector of parameters that induces the same observable data distribution.*

$$\alpha_1(c) := c \cdot \alpha; \quad \sigma_{U,1}^2(c) := \sigma_U^2/c^2;$$
$$\gamma_1(c) := \gamma; \quad \sigma_{A,1}^2(c) := \sigma_A^2;$$
$$\beta_1(c) := \beta + \Sigma_{AA}^{-1}\alpha \cdot \gamma\sigma_U^2\left(1 - \frac{1}{c}\right);$$
$$\sigma_{Y,1}^2(c) := \Sigma_{YY} - (\beta_1^\top\alpha_1 + \gamma_1)^2\sigma_{U,1}^2 - \beta_1^\top \text{diag}(\sigma_{A,1}^2)\beta_1$$

The factor c is valid if it implies positive $\sigma_{Y,1}^2(c)$.

We call the set of all parameter vectors that correspond to valid values of c the *ignorance region* in the parameter space. Parameters in the ignorance region cannot be distinguished on the basis of observed data because they all imply the same observed data distribution.

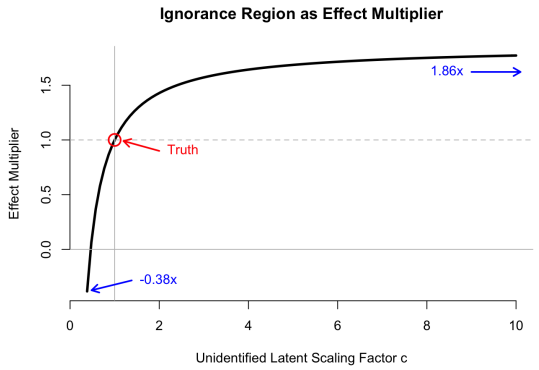

Figure 2: An illustration of the ignorance region for β in an example where causal hypotheses can be represented as a scalar effect multiplier.

We plot an illustration of the ignorance region for β from a numerical example in Figure 2. We set $\beta = b \cdot \mathbf{1}_{m\times1}$ and $\alpha = a \cdot \mathbf{1}_{m\times1}$ to be constant vectors for some constants a and b. In this case $\beta_1(c) = s(c) \cdot b \cdot \mathbf{1}_{m\times1}$ is a simple scaling of β, so the ignorance region can be represented by the value of this scalar $s(c)$. In this example, the data cannot distinguish between effect vectors that have the opposite sign of the true effect vector β, and those those that overstate the effect of A by nearly a factor of 2.

Importantly, the ignorance region does not in general disappear in the large treatment number (large-m) limit. We demonstrate this in an asymptotic extension of this example in the appendix.

6 FACTORIZATION UNIQUENESS IS INSUFFICIENT

6.1 Impossibility of Nonparametric Identification

In this section, we consider identification in the special case where the factorization of $P(A)$ by U is unique; that is, we consider the case where $P(A)$ can be decomposed uniquely into a mixing measure $P(U)$ and a conditional treatment distribution $P(A \mid U)$. In this setting, Claims 2 and 3 assert that $P(Y \mid do(A))$ is identified by (1). This claim arises as a natural response to the counterexample in the last section, where the non-uniqueness of the factorization contributes to non-identification.

As in the last section, we show via counterexample that the conditions in these claims are insufficient for identification of $P(Y \mid do(A))$. In addition, we show that, in general, parametric assumptions about the conditional outcome distribution $P(Y \mid U, A)$ are nec-

essary to identify $P(Y \mid do(A))$, except in the case where there is no confounding, i.e., the intervention distribution $P(Y \mid do(A))$ and the observed conditional distribution $P(Y \mid A)$ are equal almost everywhere. We summarize this statement in the following theorem.

Theorem 1. *Suppose that Assumption 1 holds, that $P(U, A)$ is identified, and that the model for $P(Y \mid U, A)$ is not subject to parametric restrictions.*

Then either $P(Y \mid do(A)) = P(Y \mid A)$ almost everywhere, or $P(Y \mid do(A))$ is not identified.

Put another way, our theorem states that nonparametric identification is impossible in this setting, except in the trivial case. In this section, we prove two supporting propositions for this theorem and demonstrate them in the context of our running example. The proof of the theorem, which follows almost immediately from these propositions, is included in the appendix.

6.2 Counterexample Setup

Let U be a binary latent variable, and $A := (A^{(1)}, \cdots, A^{(m)})$ a vector of m binary causes. For convenience, suppose that the individual causes $A^{(k)}$ are generated independently and identically as a function of U. Let the outcome Y be binary, and generated as a function of U and A. These assumptions yield the structural model:

$$\begin{aligned} U &:= \text{Bern}(\pi_U) \\ A^{(k)} &:= \text{Bern}(p_A(U)) \quad k = 1, \cdots, m \\ Y &:= \text{Bern}(p_Y(U, A)). \end{aligned}$$

In addition, we assume that $m \geq 3$ and that $p_A(U)$ is a non-trivial function of U. These assumptions are sufficient for the factorization of $P(A)$ by U to be unique (Kruskal, 1989; Allman et al., 2009). Thus, this example satisfies the premise of Claim 2.

Our goal is to estimate the intervention distribution for each $a \in \mathcal{A}$ using the identity in (1). Here, the intervention distribution can be summarized by the following causal parameter:

$$\begin{aligned} \pi_{Y|do(a)} &:= P(Y = 1 \mid do(A = a)) \\ &= (1 - \pi_U) p_Y(0, a) + \pi_U p_Y(1, a). \end{aligned} \quad (5)$$

Because the factorization of $P(A)$ is identified, π_U and $p_A(U)$ are identifiable. Thus, to calculate (5), it remains to recover $p_Y(U, A) = P(Y = 1 \mid U, A)$.

We will show that this conditional probability cannot be recovered from the observed data. Our approach will be to analyze the residual distribution $P(U, Y \mid A = a)$. For each value $A = a$, we can characterize

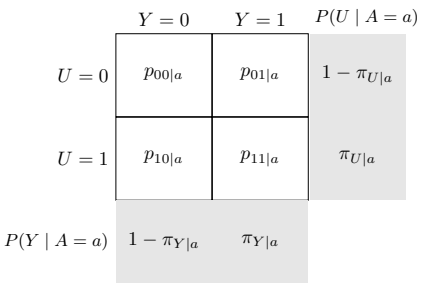

Figure 3: 2×2 table representation of $P(U, Y \mid A)$, the joint distribution of confounder U and outcome Y conditional on causes A, in the all-binary example.

$P(U, Y \mid A = a)$ by 4 probabilities in a 2×2 table (see Figure 3). We use shorthand notation $p_{uy|a} := P(U = u, Y = y \mid A = a)$ to denote each entry in this table. These entries are unobservable, but they are subject to several constraints. First, the entries are constrained to be positive and sum to 1. In addition, because $P(Y, A)$ and $P(U, A)$ are identified, the margins of the table are constrained to be equal to probabilities given by $P(Y \mid A = a)$ and $P(U \mid A = a)$. We will use the shorthand $\pi_{Y|a} := P(Y = 1 \mid A = a)$ and $\pi_{U|a} := P(U = 1 \mid A = a)$.

We consider identification in two separate cases.We first consider the case where $P(U \mid A = a)$ is non-degenerate, so that there is residual uncertainty about U after A is observed. We then consider the degenerate case, where U can be deterministically reconstructed as a function of A (the premise of Claim 3).

6.3 Non-Degenerate Case

When $P(U \mid A = a)$ is not degenerate, the table in Figure 3 is underdetermined by the data, and has one remaining degree of freedom. This degree of freedom determines the dependence between Y and U conditional on A. We parameterize this dependence in terms of $p_{11|a} := P(U = 1, Y = 1 \mid A = a)$. For purposes of interpretation, $p_{11|a}$ is linearly related to the $cor(U, Y, \mid A = a)$. $p_{11|a}$ is constrained to lie in the following range (Joe, 1997, Sec 7.1.1):

$$\max\{0, \pi_{U|a} + \pi_{Y|a} - 1\} \leq p_{11|a} \leq \min\{\pi_{U|a}, \pi_{Y|a}\}. \quad (6)$$

Fixing a value for $p_{11|a}$ in this range determines the values of all four entries in the table which, in turn, determine the causal parameter $\pi_{Y|do(a)}$ by (5).

By varying $p_{11|a}$ in the valid range (6), we can generate the ignorance region of $\pi_{Y|do(a)}$ values that are equivalently compatible with the observed data. To demonstrate, we instantiate the model with $m = 6$ causes. To simplify plotting, we specify $P(Y \mid U, A)$ so that it

only depends on $S(A) = \sum_{k=1}^{m} A^{(k)}$. This ensures that inferences about $\pi_{Y|do(a)}$ only depend on $S(a)$. In Figure 4, we plot ignorance regions for $\pi_{Y|do(a)}$ indexed by $S(a)$.

The figure demonstrates that in this simple example, $P(Y \mid do(A))$ is not identified in general. Specifically, for all vectors a where $S(a) \neq 3$, the ignorance region is non-trivial. For each $S(a)$, the extreme points of the ignorance region correspond to the extreme values of $p_{11|a}$ compatible with the margins given by $P(Y \mid A)$ and $P(U \mid A)$. The true causal parameters $P(Y = 1 \mid do(A = a))$ and the observational parameters $P(Y = 1 \mid A = a)$ are always contained in this region.

Figure 4 also shows a trivial case, where $P(Y \mid do(A = a)) = P(Y \mid A = a)$, and thus the intervention distribution is identified despite $p_{11|a}$ being underdetermined. This case arises because we defined the structural parameters such that $p_A(1) = 1 - p_A(0)$, which ensures that when $S(a) = m/2 = 3$, $P(U \mid A = a) = P(U)$, and thus $P(Y \mid do(A = a)) = P(Y \mid A = a)$.

6.4 Copula Non-Identification in General

The non-identification in the above example occurs generally when $P(Y \mid U, A)$ is nonparametric. We state this in the following supporting proposition for Theorem 1. Here, we represent the underdetermined degree of freedom by the copula density:

$$c(Y, U \mid A) := \frac{P(Y, U \mid A)}{P(Y \mid A)P(U \mid A)},$$

which specifies the dependence between Y and U given A, independently of the margins $P(Y \mid A)$ and $P(U \mid A)$ (see, e.g., Nelsen, 2007). In the following proposition, we show in a very general setting that this copula is not identified, and that this non-identification precludes identification of $P(Y \mid do(A))$.

Proposition 2. *In the setting of Theorem 1, suppose that $P(U \mid A)$ is almost surely non-degenerate. Then, the following are true*

1. *The copula density $c(Y, U \mid A)$ is not identified.*
2. *Either $P(Y \mid do(A)) = P(Y \mid A)$, or $P(Y \mid do(A))$ is not identified.*

The key idea in the proof is that only the margins $P(Y \mid A)$ and $P(U \mid A)$ are constrained by the observed data distribution (as in the previous example), leaving $c(Y, U \mid A)$ unconstrained. Using this fact, we can show constructively that there must be more than one causal hypothesis compatible with the observed data distribution, except in the trivial case. The full proof is included in the appendix.

6.5 Degenerate Case

We now continue our example and consider the case, invoked in Claim 3, where $P(U \mid A)$ is degenerate almost everywhere, i.e., where for any observable unit, U can be written as a deterministic function of A. In this case, the copula $P(U, Y \mid A)$ is trivial, so the non-identification in Proposition 2 is not an issue. However, $P(Y \mid do(A))$ remains unidentified because the degeneracy of $P(U \mid A)$ implies that the positivity assumption is violated.

To explore this case, we analyze our example as the number of causes grows large. First, we show that U can be consistently estimated from A, up to label switching. By the strong law of large numbers, for any unit with $U = u$, as m grows large,

$$\hat{p}(A) := S(A)/m \overset{a.s.}{\to} p_A(u).$$

From this fact, and our assumption that $p_A(u)$ is a non-trivial function of u, we can construct consistent estimators $\hat{U}(A)$ for U. For example, letting $I\{\cdot\}$ be an indicator function,

$$\hat{U}(A) := I\left\{\hat{p}(A) > \frac{p_A(1) + p_A(0)}{2}\right\} \tag{7}$$

is consistent as m grows large.

However, as m grows large, the causes A begin to occupy disjoint regions of the cause space $\mathcal{A}$, inducing a violation of the positivity assumption. This is the same phenomenon that drives the consistency of $\hat{U}(A)$. We illustrate in Figure 5, where we plot samples of the causes A at various values of m, projected into two dimensions and scaled. The first dimension is obtained by calculating $\hat{p}(A)$, expressed as a linear operator: $\hat{p}(A) = A^\top(m^{-1} \cdot \mathbf{1}_{m\times 1})$. The second dimension is obtained by projecting A onto a vector orthogonal to $\mathbf{1}_{m\times 1}$. In this case, we choose the vector $v_2 = m^{-1/2} \cdot (\mathbf{1}^\top_{(m/2)\times 1}, -\mathbf{1}^\top_{(m/2)\times 1})^\top$ and calculate $A^\top v_2$. When m is small, A has the same support whether $U = 0$ or $U = 1$, but as m grows, the causes A concentrate on either side of the decision boundary of our estimator $\hat{U}(A)$, ensuring consistency of the estimator, but violating positivity.

Because of this phenomenon, the conditional probabilities $P(Y = 1 \mid U = 0, \hat{U}(A) = 1)$ and $P(Y = 1 \mid \hat{U}(A) = 0, U = 1)$ are inestimable from the data in the large-m limit, and we cannot evaluate (5). At best, we can bound (5) by substituting extreme values 0 and 1 for the inestimable probabilities.

This problem manifests even for finite m, when the residual uncertainty about U after observing A is small. Consider the 2×2 table in Figure 3 corresponding to a cause vector a for which $\pi_{U|a} < \min\{\pi_{Y|a}, 1 -$

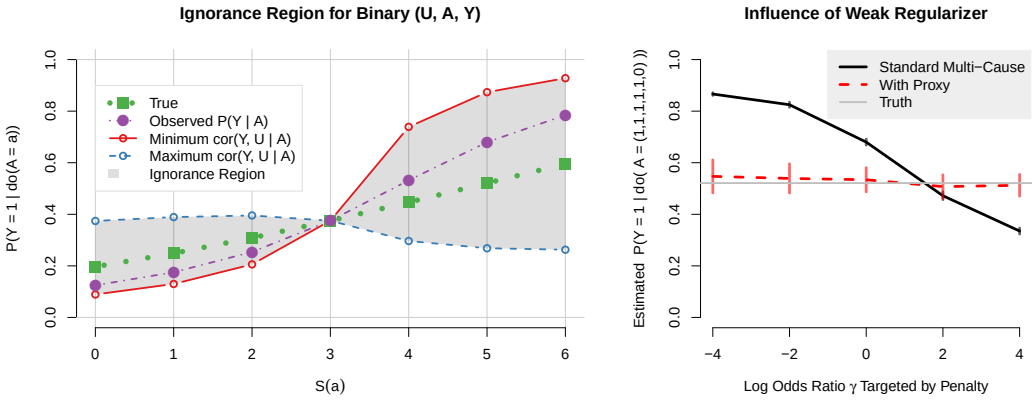

Figure 4: (Left) Ignorance regions for the causal parameters $\pi_{Y|do(a)}$, in an example where the ignorance regions only depend on $S(a) = \sum_{k=1}^{m} a^{(k)}$. All values of $\pi_{Y|do(a)}$ in the gray region are compatible with the observed data distribution. (Right) Results of estimation experiment, attempting to recover $P(Y \mid do(A = (1,1,1,1,1,0))$ using maximum likelihood estimation. In the standard multi-cause setting, a weak regularizer determines the estimate $\pi_{Y|do(a)}$ within the ignorance region. When two proxy variables are added, the weak regularizer has little effect. Vertical bars show ± 1 sd from 20 simulations.

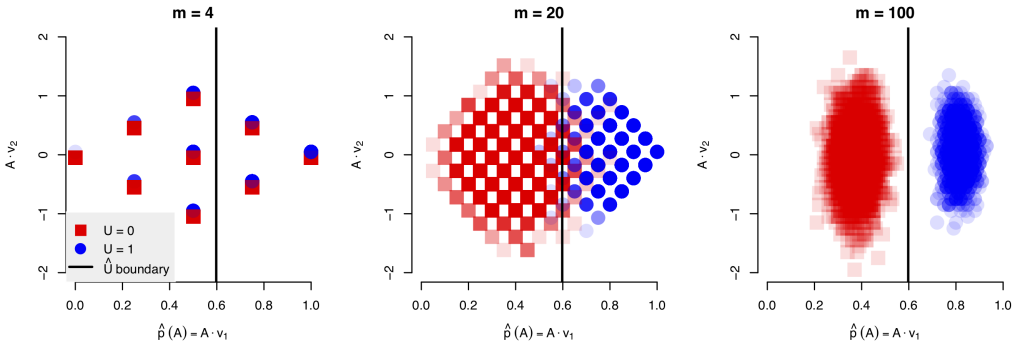

Figure 5: Scaled linear projections of sampled treatments A in the all-binary example. Color and shape indicate the value of the latent confounder U. As m, the dimension of A, increases, the treatment vectors A generated under each value of U become more distinct. This separation enables consistent estimation of U; the black boundary corresponds to the estimator (7). However, this separation also constitutes a positivity violation.

$\pi_{Y|a}\}$. The bounds in (6) imply that the ignorance region includes$P(Y = 1 \mid U = 1, A = a) \in [0, 1]$, i.e., the data provide no information about this conditional probability. In the limit, as $\pi_{U|a} \to 0$, the ignorance region has the form

$$\pi_{Y|do(a)} \in (1 - \pi_U)P(Y = 1 \mid A = a) + [0, \pi_U].$$

Similarly, for cases where $\pi_{U|a}$ approaches 1, the data provide no information about $P(Y \mid U = 0, A = a)$, and the ignorance region has width equal to $1 - \pi_U$. This regime appears in Figure 4 for small and large values of $S(a)$, where the ignorance region widths approach $\pi_U = 0.3$ and $1 - \pi_U = 0.7$, respectively. As m grows large, the probability that a sampled cause vector A falls in this regime approaches 1.

6.6 Consistency and Positivity in General

The above positivity violation occurs in any case where the latent confounder U can be reconstructed consistently from the causes A. We summarize this fact in the following supporting proposition for Theorem 1.

Proposition 3. *Suppose that Assumption 1 holds, that $P(U)$ is not degenerate, and that there exists a consistent estimator $\hat{U}(A_m)$ of U as m grows large. Then positivity is violated as m grows large.*

The central idea in the proof is that when $\hat{U}(A)$ is consistent in the large m limit, the event $U = u$ implies that A takes a value a such that $\hat{U}(a) = u$. Thus, for distinct values of the latent variable U, the observed causes A must lie in disjoint regions of the cause space, violating positivity. The proof is in the appendix.

7 DISCUSSION

7.1 Practical Implications

Theorem 1 suggests that we should generally be cautious about drawing causal inferences in the multi-cause setting. Specifically, the existence of a non-trivial ignorance region can make conclusions highly dependent on modeling choices that in other contexts would be innocuous.

On one hand, flexible, nonparametric models will yield estimates of causal quantities that are highly sensitive to regularization within the ignorance region. To demonstrate this, we plot the results of an estimation experiment on the right of Figure 4. We estimate $\pi_{y|do(a)}$ for a given vector a from 15,000 datapoints drawn from the binary example's data generating process. We perform estimation by maximum likelihood, but we add a weak L2 penalty term on the log-odds ratio $\gamma := \log\left(\frac{p_{11|a}}{p_{10|a}} \Big/ \frac{p_{01|a}}{p_{00|a}}\right)$, which pushes estimates of $p_{11|a}$ to take specific values. The penalty determines where in the ignorance region the estimate appears. This behavior can be unpredictable in complex problems, especially when applying estimation methods that involve tuning parameters or stochastic training.

On the other hand, models the incorporate parametric assumptions, e.g., that $E[Y \mid U, A]$ is linear, may yield unique solutions. However, our results still suggest caution. First, it can be difficult to confirm whether a set of parametric assumptions are sufficient to identify the causal quantities of interest, as our first counterexample suggests. In addition, because of the lack of nonparametric identification, the components of these assumptions that determine causal parameter estimates cannot be checked: the data will be indifferent to different parametric specifications that yield different estimates within the ignorance region.

7.2 Proxy Variables

Despite these negative results, there are straightforward modifications to the multi-cause setting that can yield nonparametric identification. One alternative is to incorporate *proxy variables* (a.k.a. negative controls) into the estimation procedure. A proxy variable Z is a supplementary variable that is downstream of the latent confounder U, but is causally unrelated to either the outcome Y or the cause variables A. Several authors have given sufficient conditions for causal identification when using proxy variables in the presence of unobserved confounding (see, e.g., Kuroki and Pearl, 2014; Miao et al., 2016; Shi et al., 2018). Recently, Louizos et al. (2017) applied similar ideas to estimate causal effects by calculating (1) from the posterior distribution of a variational autoencoder model.

Including proxy variables in a multi-cause problem may only require small modifications. For example, the sufficient conditions for nonparametric identification in Miao et al. (2016) require two proxy additional variables: one conditionally independent of A given U, and the other conditionally independent of Y given U. In the multi-cause setting, it is not difficult to find a proxy of the first type; each cause variable $A^{(k)}$ satisfies this condition. Similarly, it is plausible that proxies of the second type would be readily available in many contexts that could be framed as multi-cause problems, e.g., GWAS (Tran and Blei, 2017) or recommender systems (Wang et al., 2018). For example, in GWAS, where genetic variants operate as causes A for some phenotype Y, we may know *a priori* that some variants are causally unrelated to Y, but are driven by the same underlying confounder U. These unrelated variants could serve as proxies of the second type.

To demonstrate that the proxy approach addresses some of the pathologies highlighted in this paper, we repeat the estimation experiment on the right side of Figure 4, adding two proxy variables to the data. As in Section 7.1, we estimate the model parameters by maximum likelihood estimation for varying settings of a weak regularizer. In this case, the regularizer has negligible effect on the estimates of $\pi_{Y|do(a)}$.

7.3 Sensitivity Analysis

Finally, the multi-cause setting could present a fruitful use case for sensitivity analysis as an alternative to point estimation of causal estimands. In such a case, rather than focus on point-identifying a causal estimand under strong assumptions, it can be useful to relax our requirements, and attempt to identify an ignorance region under weaker assumptions (see, e.g., Manski, 2009). The multi-cause setting is particularly amenable to sensitivity analysis, because the assumptions about factorizations of $P(A)$ represent a reasonable middle ground between the strong assumption of no unobserved confounding at one extreme and the weak assumption of unstructured confounding on the other. For example, in Figure 4, Assumption 1 induces ignorance regions that are relatively narrow for some values of a. For some applications, analyses that explicitly map these sensitivity regions could be acceptably informative. Usefully, sensitivity analyses of this type can be applied *post hoc*, so that an investigator is free to find a best-fitting model for the observed data, and then reason about the causal implications of the model separately (see also Franks et al., 2018).

Acknowledgements

Thanks to Yixin Wang, Rajesh Ranganath, Dustin Tran, and David Blei for open conversations about their work. Thanks to Alexander Franks, Andrew Miller, Avi Feller, Jasper Snoek, and D. Sculley for their feedback on this project.

References

Allman, E. S., C. Matias, J. A. Rhodes, et al. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics 37*(6A), 3099–3132.

Bollen, K. A. (1989). *Structural Equation Models with Latent Variables*. New York: John Wiley and Sons.

Franks, A., A. D'Amour, and A. Feller (2018). Flexible sensitivity analysis for observational studies without observable implications. *arXiv preprint arXiv:1809.00399*.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*. Chapman and Hall/CRC.

Kruskal, J. (1989). Rank, decomposition, and uniqueness for three-way and n-way arrays. *Multiway data analysis. Elsevier Science Publisher BV (North-Holland)*.

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

Louizos, C., U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456.

Manski, C. F. (2009). *Identification for prediction and decision*. Harvard University Press.

Miao, W., Z. Geng, and E. Tchetgen (2016). Identifying causal effects with proxy variables of an unmeasured confounder.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Pearl, J. (2009). *Causality*. Cambridge university press.

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

Rosenbaum, P. R. and D. B. Rubin (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)*, 212–218.

Shi, X., W. Miao, and E. T. Tchetgen (2018). Multiply robust causal inference with double negative control adjustment for unmeasured confounding. *arXiv preprint arXiv:1808.04906*.

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

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

Wang, Y., D. Liang, L. Charlin, and D. M. Blei (2018). The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581*.