Recovering Causal Effects from Selection Bias

Elias Bareinboim*
Computer Science Department
University of California, Los Angeles
Los Angeles, CA. 90095
eb@cs.ucla.edu

Jin Tian*
Department of Computer Science
Iowa State University
Ames, IA. 50011
jtian@iastate.edu

Abstract

Controlling for selection and confounding biases are two of the most challenging problems that appear in data analysis in the empirical sciences as well as in artificial intelligence tasks. The combination of previously studied methods for each of these biases in isolation is not directly applicable to certain non-trivial cases in which selection and confounding biases are simultaneously present. In this paper, we tackle these instances non-parametrically and in full generality. We provide graphical and algorithmic conditions for recoverability of interventional distributions for when selection and confounding biases are both present. Our treatment completely characterizes the class of causal effects that are recoverable in Markovian models, and is sufficient for Semi-Markovian models.

Introduction

Computing the effects of interventions is one of the fundamental problems in the empirical sciences. Whenever data collection is performed, the goal is almost invariably to evaluate causal effect relationships – for instance, what is the impact of a new taxation program, how should a robot react to unanticipated situations, will a new advertisement campaign change the propensity of users buying a product, or what is the effect of a new drug for curing cancer?

The first challenge that needs to be addressed when computing these effects is to control for confounding bias, which may arise when randomized experiments are infeasible to conduct due to costs, ethical, or technical considerations. This implies that the data is collected under an observational regime, where the population follows its natural tendency. Our goal, however, is to compute how the population reacts when its undergoes a change (intervention), following a new, compulsory protocol. For instance, one is not interested in estimating the correlation between smoking and cancer, which follows a process of self-selection (associational), but whether the incidence of cancer would decrease if smoking were banned in this population (interventional).

More formally, the *identifiability problem* is concerned with determining the effect of a treatment (X) on an outcome (Y), $P(y|do(x))$, whenever only an observational, non-experimental distribution $P(v)$ is available (where V represents all observable variables). This is not always possible, however, and the difference between $P(y|do(x))$ and its probabilistic counterpart, $P(y|x)$, is known as *confounding bias* (Pearl 2000, Ch. 3; pp. 202-212). In Fig. 1(a), for example, the effect $P(y|do(x))$ can be computed by cutting the incoming arrows towards X to simulate the intervention, but if the probabilistic estimate $P(y|x)$ is considered, the path going through the backdoor $X \leftarrow Z \rightarrow Y$ will also be present in the estimand yielding bias – the variable Z generates extraneous variations of the outcome that are not due to X. There is a well-known method for removing confounding bias in examples of this kind, which is given by the expression

$$P(y|do(x)) = \sum_z P(y|x, Z = z)P(Z = z). \quad (1)$$

This identity is known as the "adjustment formula" or "backdoor formula" (Pearl 1995) and represents a special case in which a mapping from $P(v)$ to $P(y|do(x))$ exists.

The problem of confounding has been broadly studied in the literature and a number of conditions for non-parametric identification had emerged, for instance (Spirtes, Glymour, and Scheines 1993; Galles and Pearl 1995; Pearl and Robins 1995; Halpern 1998; Kuroki and Miyakawa 1999). A general mathematical treatment was given in (Pearl 1995) and culminated in the *do-calculus*, which was shown to be complete (Tian and Pearl 2002a; Huang and Valtorta 2006; Shpitser and Pearl 2006; Bareinboim and Pearl 2012a).

Another major challenge that needs to be addressed when evaluating the effect of interventions is the problem of selection bias, which arises due to the preferential exclusion of units from the sample. For instance, in a typical study of the effect of training program on earnings, subjects achieving higher incomes tend to report their earnings more frequently than those who earn less. The data-gathering process in this case will reflect this distortion in the sample proportions and, since the sample is no longer a faithful representation of the population, biased estimates will be produced regardless of how many samples were collected.

It is instructive to depict this phenomenon graphically, so consider the model in Fig. 1(b) in which X represents a treatment, Y represents an outcome, and S represents a binary indicator of entry into the data pool ($S = 1$ indicates that

*These authors contributed equally to this paper.

the unit is in the sample, $S = 0$ otherwise). In this case, selection is affected by the outcome as represented by the arrow $Y \rightarrow S$ (e.g., achievers of higher incomes have a higher chance of reporting their earnings). (If, instead, a random sample of the population is taken, S will be independent of all variables in the analysis.) If our goal is to compute the effect $P(y|do(x))$ but the samples are collected preferentially, then only $P(y, x|S = 1)$ is accessible for use. Given that confounding bias is not present in the example in Fig. 1(b), the effect of X on Y is the same as the corresponding conditional distribution (i.e., $P(y|do(x)) = P(y|x)$). The natural question to ask is under what conditions $P(y|do(x))$ $(=P(y|x))$ can be recovered from data drawn from $P(y, x|S = 1)$, since in principle these two distributions are just loosely connected.

The selection bias problem has been studied in a wide range of scenarios, for instance, in several tasks in AI (Cooper 1995; Elkan 2001; Zadrozny 2004; Cortes et al. 2008), statistics (Whittemore 1978; Little and Rubin 1986; Jewell 1991; Kuroki and Cai 2006) as well as in the empirical sciences (e.g., genetics (Pirinen, Donnelly, and Spencer 2012; Mefford and Witte 2012), economics (Heckman 1979; Angrist 1997), and epidemiology (Robins 2001; Glymour and Greenland 2008)). These works lead to a complete treatment for recoverability of the *odds ratio* in (Bareinboim and Pearl 2012b), and culminated in a complete treatment for non-parametric recoverability of conditional distributions in (Bareinboim, Tian, and Pearl 2014).

The biases arising from confounding and selection are fundamentally different, though both constitute threats to the validity of causal inferences. The former bias is the result of treatment X and outcome Y being affected by common ancestral variables, such as Z in Fig. 1(a), while the latter is due to treatment X or outcome Y (or ancestors) affecting the inclusion of the subject in the sample, such as Y in Fig. 1(b). In both cases, we have extraneous "flow" of information between treatment and outcome, which falls under the rubric of "spurious correlation," since it is not what we seek to estimate. These problems constitute a "basis" for causal analysis, and one might appear without the other. For instance, confounding bias might still exist even when a perfect random sample of the population is drawn, while selection bias may also exist even when the treatment assignment is perfectly randomized.

The combined treatment of these biases was not discussed in its full generality until now in the literature, and in this paper we show non-trivial instances in which previous methods are not directly applicable, so no method is known to date. Building on the previous conditions developed for independently controlling these biases, this paper provides a systematic treatment for the problem of simultaneous selection and confounding, more specifically:

- We provide a necessary and sufficient graphical and algorithmic condition for recoverability from simultaneous selection and confounding in models without latent variables (i.e., Markovian);
- We construct a general algorithm and sufficient condition for recoverability from selection and confounding biases in models with latent variables (i.e., semi-Markovian);

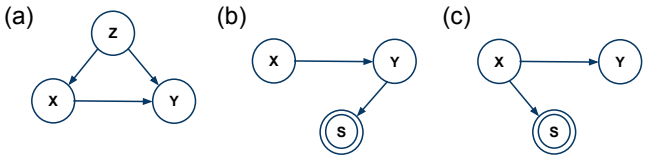

Figure 1: Simplest examples of confounding bias (a) and selection bias (b,c). The effect $Q = P_x(y)$ is recoverable from confounding through adjustment in (a), and it is not recoverable from selection in (b), but is in (c).

Models, Causal Effects, and Recoverability

We first introduce some basic machinery used throughout the paper. The basic semantical framework in our analysis rests on *structural causal models* as defined in (Pearl 2000, pp. 205), also called data-generating models. In the structural causal framework (Pearl 2000, Ch. 7), actions are modifications of functional relationships, and each action $do(\mathbf{x})$ on a causal model M produces a new model $M_{\mathbf{x}} = \langle \mathbf{U}, \mathbf{V}, \mathbf{F}_{\mathbf{x}}, P(\mathbf{U}) \rangle$, where $\mathbf{V}$ is the set of observable variables, $\mathbf{U}$ is the set of unobservable variables, and $F_{\mathbf{x}}$ is obtained after replacing $f_X \in \mathbf{F}$ for every $X \in \mathbf{X}$ with a new function that outputs a constant value x given by $do(\mathbf{x})$.

We follow the conventions given in (Pearl 2000) denoting variables by capital letters and their realized values by small letters. We use the typical graph-theoretic terminology with the corresponding abbreviations $Pa(C)$ and $An(C)$, which will denote the union of C and respectively the parents and ancestors of C. Finally, for any set $C \subseteq V$, let G_C denote the subgraph of G composed only of variables in C.

Key to the analysis in this paper is the notion of "identifiability" (Pearl 2000, pp. 77), which expresses the requirement that causal effects are computable from a combination of data and assumptions embodied in a causal graph G:

Definition 1 (Causal Effects Identifiability) *The causal effect of an action* $do(\mathbf{T} = \mathbf{t})$ *on a set of variables* $\mathbf{R}$ *is said to be identifiable from* P *in* G *if* $P_{\mathbf{t}}(\mathbf{r})$ *is uniquely computable from* $P(\mathbf{v})$ *in any model that induces* G.

Another important element used in our analysis is the notion of recoverability, which expresses the requirement that effects are computable from the available (biased) data and assumption embodied in an augmented causal graph G_s (Bareinboim and Pearl 2012b):

Definition 2 (Recoverability from Selection Bias) *Given a causal graph* G_s *augmented with a node* S *encoding the selection mechanism, the distribution* $Q = P_{\mathbf{t}}(\mathbf{r})$ *is said to be recoverable from selection biased data in* G_s *if the assumptions embedded in the causal model renders* Q *expressible in terms of the distribution under selection bias* $P(\mathbf{v} \mid S = 1)$. *Formally, for every two probability distributions* P_1 *and* P_2 *compatible with* G_s, $P^{M_1}(\mathbf{v} \mid S = 1) = P^{M_2}(\mathbf{v} \mid S = 1) > 0$ *implies* $P^{M_1}_{\mathbf{t}}(r) = P^{M_2}_{\mathbf{t}}(\mathbf{r})$.

In this paper, the analysis will basically focus on determining whether these two conditions can be simultaneously satisfied, which we will call recoverability for short.

Recoverability in Markovian Models

It is known that a causal effect $P_t(r)$ is always identifiable in terms of the distribution $P(v)$ in Markovian models (i.e., when all variables are observed) (Pearl 2000; Tian and Pearl 2002a). Further, a complete condition for recovering conditional probabilities from selection biased data has been given in (Bareinboim, Tian, and Pearl 2014) and read as follows:

Theorem 1 *The conditional distribution $P(y|t)$ is recoverable (as $P(y|t, S = 1)$) if and only if $(Y \perp\!\!\!\perp S|T)$.*

In Fig. 1(c), consider the recoverability of $P_x(y)$. We can combine this result with a do-calculus reduction [1] and write

$$P_x(y) = P(y|x) \quad (2)$$
$$= P(y|x, S = 1), \quad (3)$$

where the first equality follows from the second rule of the do-calculus (since X and Y are unconfounded), and the second equality follows from Theorem 1. Eq. (3) states that the effect of X on Y is equal to the conditional distribution of Y given X estimated from the selected biased data.

Based on this derivation, it is now immediate to state the following result that combines recoverability of conditional distributions with identifiability using the do-calculus:

Corollary 1 *The causal effect $Q = P_t(r)$ is recoverable from selection biased data if using the rules of the do-calculus, Q is reducible to an expression in which no do-operator appears, and recoverability is determined by Theorem 1.*

We want to recover $Q = P_x(y)$ in Fig. 1(b), and the same reasoning used in Eq. (2) applies here yielding $P_x(y) = P(y|x)$. Based on Theorem 1, the distribution $P(y|x)$ is marked as not recoverable, which indeed implies that Q is not recoverable in this case (formally shown in Theorem 1).

Based on this result, one might surmise that determining $P_t(r)$ by first expressing it in terms of probabilities over the observables and using Theorem 1 to determine the recoverability of each factor in the resultant expression is not only valid, but also a necessary condition for controlling both selection and confounding biases. However, this approach is somewhat misleading, because there usually exist equivalent expressions for the effect $P_t(r)$ in terms of the probability over the observables, and while each expression is equally sound for controlling confounding bias, they appear to entail different conclusions for the problem of recoverability.

To understand this subtlety, note that the effect $P_x(y)$ of a singleton X is always identifiable through adjustment for its direct causes (Pearl 2000, Theorem 3.2.2),

$$P_x(y) = \sum_{pa_x} P(y|x, pa_x)P(pa_x). \quad (4)$$

Following the strategy of Corollary 1, one would need to check the recoverability of the factors $P(y|x, pa_x)$ and $P(pa_x)$ using Theorem 1 to determine the recoverability of $P_x(y)$, which does not always work as shown in the graph in Figure 2(a). To witness, note that we can obtain using Eq. (4) the following expression (equivalently, through the back-door criterion with $\{W_1, W_2\}$ as the covariate set),

$$P_x(y) = \sum_{w_1,w_2} P(y|x, w_1, w_2)P(w_1, w_2). \quad (5)$$

It may appear now that $P_x(y)$ is not recoverable since despite the fact that $P(y|x, w_1, w_2)$ is recoverable by Theorem 1, the second factor $P(w_1, w_2)$ is not. However, we can use the back-door criterion with $\{Z\}$ as the covariate set and obtain:

$$P_x(y) = \sum_{z} P(y|x, z)P(z). \quad (6)$$

Eq. (6) constitutes another expression witnessing the identifiability of $P_x(y)$ [2], but in this case recoverable – the factors $P(y|x, z)$ and $P(z)$ are both recoverable by Theorem 1.

We see from this example that, after expressing $P_t(r)$ in terms of the observational distribution, while it is straightforward to determine $P_t(r)$ to be recoverable if all probabilities involved are recoverable, in general it is not easy to determine whether $P_t(r)$ is recoverable when some probabilities involved are not recoverable.

It can be computationally difficult to find a set satisfying the conditions of recoverability and identifiability simultaneously since this could imply a search over an exponentially large number of subsets. Next we reduce this problem to a more tractable case and show that the recoverability of $P_t(r)$ can be determined by expressing it in a "canonical" form in terms of the Markovian factors $P(v_i|pa_i)$ (i.e., probability of V_i given its parents). We then derive a complete condition for recovering $P_t(r)$ from selection biased data. First, we give a complete condition for recovering the factors $P(v_i|pa_i)$ from selection biased data.

Lemma 1 *The distribution $P(v_i|pa_i)$ for $i = 1, \ldots, n$ is recoverable if and only if V_i is not an ancestor of S. When recoverable, $P(v_i|pa_i) = P(v_i|pa_i, S = 1)$.*

Proof. This follows from Theorem 1 since V_i is independent of S given Pa_i if and only if V_i is not an ancestor of S. □

We present next a complete condition for recovering $P_t(r)$ from selection biased data.

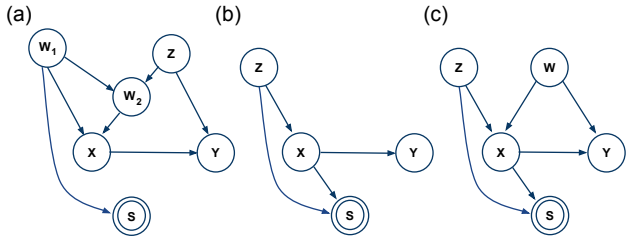

Figure 2: (a) Causal model in which the effect $Q = P_x(y)$ is identifiable with sets $\{W_1, W_2\}$ and $\{Z\}$, but recoverable just with $\{Z\}$; Q is recoverable in (b) but not recoverable in (c), due to the existence of W (common ancestor of S and Y).

[1] For more on the do-calculus, see Appendix 1 (supplemental material) or (Pearl 2000, Ch. 3).

[2] Note that both the expressions in (5) and (6) can be obtained using the back-door criterion. The two sets $\{W_1, W_2\}$ and $\{Z\}$ are called c-equivalent (Pearl and Paz 2014).

Theorem 2 *The causal effect $P_t(r)$, where T and R are arbitrary disjoint subsets of V, is recoverable if and only if for every $V_i \in D$ where $D = An(R)_{G_{V\setminus T}}$, V_i is not an ancestor of S. When recoverable, $P_t(r)$ is given by*

$$P_t(r) = \sum_{D\setminus R} \prod_{\{i:V_i\in D\}} P(v_i|pa_i, S = 1). \tag{7}$$

Proof. (if) Let $T' = V \setminus T$. We have

$$P_t(r) = \sum_{T'\setminus R} P_t(t') = \sum_{T'\setminus R} \prod_{\{i:V_i\in T'\}} P(v_i|pa_i). \tag{8}$$

All the factors $P(v_i|pa_i)$ in (8) for which $V_i \notin D$ (i.e., V_i is not an ancestor of R) will be summed out and we obtain

$$P_t(r) = \sum_{D\setminus R} \prod_{\{i:V_i\in D\}} P(v_i|pa_i). \tag{9}$$

If for every $V_i \in D$, V_i is not an ancestor of S, then the corresponding $P(v_i|pa_i)$ is recovered as $P(v_i|pa_i, S = 1)$ by Lemma 1. Therefore $P_t(r)$ is recoverable.

(only if sketch – see Appendix 2 in the supplemental material) We show that whenever there exists $V_i \in D$ such that V_i is an ancestor of S, two distributions P_1, P_2 compatible with the causal model can be constructed such that they agree in the probability distribution under selection bias, $P_1(v \mid S = 1) = P_2(v \mid S = 1)$, and disagree in the target effect $Q = P(y \mid do(x))$, i.e., $P_1(y \mid do(x)) \neq P_2(y \mid do(x))$.

Let P_1 be compatible with the graph $G_1 = G_s$, and P_2 with the subgraph G_2 where the edges pointing to S are removed (see (Tian 2002, Lemma 8)). Notice that P_2 harbors an additional independence relative $(V \perp\!\!\!\perp S)_{P_2}$, where V represents all variables in G_s but the selection mechanism S. We will set the parameters of P_1 through its factors and then compute the parameters of P_2 by enforcing $P_2(v \mid S = 1) = P_1(v \mid S = 1)$. Since $P_2(v|S = 1) = P_2(v)$, we will have $P_1(v|S = 1) = P_2(v)$. We will exploit this symmetry throughout the proof. We partition how S can be connected to Y in four topological relations, and then show that for each one of these equality in distribution under selection bias and inequality of the causal effects will follow (see Appendix 2, supplemental material). □

We demonstrate the application of Theorem 2 with a few examples. In Figure 2(a), $D = \{Y, Z\}$ and so $\{W_1, W_2\}$ can be ignored, and $P_x(y)$ is recoverable as given by Eq. (6). In Figure 2(b), note that $D = \{Y\}$ is not an ancestor of S, so $P_x(y)$ is recoverable and given by $P_x(y) = P(y|x, S = 1)$. In Figure 2(c), we have $D = \{W, Y\}$ and $P_x(y) = \sum_w P(y|w, x, S = 1)P(w)$, which is not recoverable due to the fact that $P(w)$ is not recoverable since W is an ancestor of S.

Recoverability in Semi-Markovian Models

Some relevant confounders are difficult to measure in many real-world applications (e.g., intention, mood, dna mutation), which leads to the need of modelling explicitly latent variables that affect more than one observed variable in the system[3]. These scenarios can be encoded formally as Semi-Markovian models.

In Markovian models identifiability is always attainable, and the challenge addressed in the previous section was to search through the space of admissible sets testing for recoverability from selection. The challenge in Semi-Markovian models arises due to the fact that parents of X as well as possibly other variables might not be observed, which implies that identifiability is not achievable in all experimental designs. The evaluation of identifiability itself usually goes through a non-trivial algebraic process known as the do-calculus (Pearl 2000). Recoverability then becomes more involved since the evaluation of the feasibility of identifying a quantity needs to be coupled with the search for sets yielding recoverability from selection bias.

In order to visualize the difference between recoverability in Markovian and semi-Markovian models, let us consider the simple example depicted in Fig. 3(a). Our goal is to evaluate the effect $Q = P_x(y)$ assuming the availability of selected biased data. If we apply the direct adjustment given in Eq. (4) (using the set of parents of X, $\{W_2\}$), we obtain

$$P_x(y) = \sum_{w_2} P(y|x, w_2)P(w_2). \tag{10}$$

In this case, however, none of the factors in Eq. (10) are recoverable by Theorem 1. If we enlarge our criterion to consider arbitrary back-door admissible sets, it is also the case that neither the empty set nor $\{W_1\}$ are admissible for back-door adjustment – there exists a back-door path passing through W_2 that needs to be closed (and in general, $P_x(y) \neq P(y|x)$). Furthermore, if we decide to include this set and try to adjust for $\{W_1, W_2\}$, the back-door path is indeed closed but Q is still not recoverable.

One might be tempted to believe, when taking the Markovian perspective, that there exists no set yielding recoverability of the target Q, which turns out to not be the case. To witness, invoke the first rule of the do-calculus noticing that S is independent of Y in the mutilated graph when the incoming arrow to X is cut (Fig. 3(b)), $(S \perp\!\!\!\perp Y|X)_{G_{\overline{X}}}$, which implies the equality $P_x(y) = P_x(y|S = 1)$. Perhaps surprisingly to some [4], this can be coupled with traditional adjustment by the parent set, which yields

$$P_x(y) = \sum_{w_2} P(y|x, w_2, S = 1)P(w_2|S = 1) \tag{11}$$

Note that Eqs. (10) and (11) are essentially the same except for the conditioning variable S, and the critical step leading

[3]Following the convention (Pearl 2000), the unobserved common causes are encoded implicitly in the dashed bidirected arrows.

[4]This conclusion follows organically from the logic of structural causality (Pearl 2000), but this condition is missing in other attempts for treating the selection bias problem (Angrist 1997).

to recoverability was to realize that despite the fact that Theorem 1 does not apply to neither of the sub-factors in *any* of the possible adjustment sets, the independence $(S \perp\!\!\!\perp Y|X)_{G_{\overline{X}}}$ in the mutilated graph should first be evaluated, and the adjustment step should be considered after that.

Given that our goal is to find a systematic procedure for deciding recoverability, we would expect that a positive test for separability of S, similarly to the derivation culminating in Eq. (11), combined with the identification test of the target quantity should be enough to yield recoverability. Still, this strategy does not work if applied in a naive fashion. To witness, consider the recoverability of $Q = P_x(y)$ in Fig. 3(c). We first mutilate the graph cutting the incoming arrows going towards X (Fig. 3(d)), so we have $P_x(y) = P_x(y|S = 1)$. The challenge here is that after adding S to the expression, our ability of applying the third rule of the do-calculus and obtaining the desired expression is curtailed. Despite the fact that the effect of X on Y is zero (i.e., the equality $P_x(y) = P(y)$ holds), it it not the case that $P_x(y|S = 1) = P(y|S = 1)$ is true. In fact, what follows from the analysis is the equality $P_x(y|S = 1) = P(y)$, which does not represent a viable mapping from the biased data to the target expression Q (note that the target does not appear in this equality. [5] After all, we note that even when the effect Q is identifiable and S is separable from Y in the mutilated graph, it might still be the case that the target quantity Q is not recoverable.

We next state some useful lemmas combining the understanding acquired through these examples. Following the notation in (Tian and Pearl 2002a), for any set $C \subseteq V$, we define the quantity $Q[C](v)$ to denote the following function

$$Q[C](v) = P_{v\setminus c}(c) = \sum_{u} \prod_{\{i|V_i \in C\}} P(v_i|pa_i, u^i)P(u). \quad (12)$$

In particular, we have $Q[V](v) = P(v)$. For convenience, we will often write $Q[C](v)$ as $Q[C]$. The set of variables V can be partitioned into so-called c-components by assigning two variables to the same c-component if and only if they are connected by a path composed entirely of bidirected edges. The following lemma is from (Tian and Pearl 2002b):

Lemma 2 *Let $H \subseteq V$, and assume that H is partitioned into c-components $H_1, \ldots, H_l$ in the subgraph G_H. Then we have*
(i) $Q[H]$ decomposes as

$$Q[H] = \prod_i Q[H_i]. \quad (13)$$

(ii) Let a topological order of the variables in H be $V_{h_1} < \cdots < V_{h_k}$ in G_H. Let $H^{\leq i} = \{V_{h_1}, \ldots, V_{h_i}\}$ be the set of variables in H ordered before V_{h_i} (including V_{h_i}), and $H^{>i} = H \setminus H^{\leq i}$ for $i = 1, \ldots, k$, and $H^{\leq 0} = \emptyset$. Then each $Q[H_j]$, $j = 1, \ldots, l$, is computable from $Q[H]$ and given by

$$Q[H_j] = \prod_{\{i|V_{h_i} \in H_j\}} \frac{Q[H^{\leq i}]}{Q[H^{\leq i-1}]}, \quad (14)$$

[5] This exemplifies a more fundamental matter, which is that the syntactic goal of recoverability is different than that of conditional identifiability. In the latter, the do-operator needs to be removed, and the conditioning set is not constrained in any fashion; in the former, the do-operator also needs to be removed, but this is contingent on the existence of the S-node in the final expression.

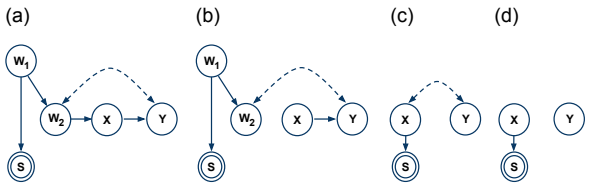

Figure 3: (a,b) Diagram and respective mutilated graph in which the effect $Q = P_x(y)$ is recoverable despite the inapplicability of directly using Theorem 1 followed by adjustment. (c, d) Model with mutilated graph showing insufficiency of evaluating recoverability and identifiability independently.

where each $Q[H^{\leq i}]$, $i = 0, 1, \ldots, k$, is given by

$$Q[H^{\leq i}] = \sum_{h^{>i}} Q[H]. \quad (15)$$

So, we can generalize Q-decompositions for functions of the joint distribution (including the selection distribution):

Lemma 3 *Let $H \subseteq V$, and assume that H is partitioned into c-components $H_1, \ldots, H_l$ in the subgraph G_H. If*

$$f(P(v|S = 1)) = \frac{P(S = 1|pa_S)}{P(S = 1)} Q[H], \quad (16)$$

where $f(P(v|S = 1))$ is some recoverable quantity, then for $j = 1, \ldots, l$, $Q[H_j]$ is recoverable if $H_j \cap An(S) = \emptyset$, that is, H_j contains no ancestors of S.

Proof. Let $V_{h_1} < \cdots < V_{h_k}$ be a topological order in G_H such that $An(S) < H \setminus An(S)$. $Q[H_j]$ is given by Eqs. (14) and (15) by Lemma 2, and from Eq. (16) $Q[H]$ is given by

$$Q[H] = \frac{P(S = 1)}{P(S = 1|pa_S)} f(P(v|S = 1)). \quad (17)$$

If H_j contains no ancestors of S, then all the variables in H_j are ordered after the variables in $An(S)$, then for each $V_{h_i} \in H_j$, $h^{>i} \cup \{V_{h_i}\}$ contains no variables in Pa_S. We obtain

$$\begin{aligned} Q[H^{\leq i}] &= \sum_{h^{>i}} \frac{P(S = 1)}{P(S = 1|pa_S)} f(P(v|S = 1)) \\ &= \frac{P(S = 1)}{P(S = 1|pa_S)} \sum_{h^{>i}} f(P(v|S = 1)), \end{aligned} \quad (18)$$

$$Q[H^{\leq i-1}] = \frac{P(S = 1)}{P(S = 1|pa_S)} \sum_{h^{>i}, v_{h_i}} f(P(v|S = 1)), \quad (19)$$

and finally

$$\frac{Q[H^{\leq i}]}{Q[H^{\leq i-1}]} = \frac{\sum_{h^{>i}} f(P(v|S = 1))}{\sum_{h^{>i}, v_{h_i}} f(P(v|S = 1))}. \quad (20)$$

Therefore, $Q[H_j]$ given by Eqs. (14) is recoverable. □

Note that a special case of Lemma 2 and 3 is when $H = V$ and we have the decomposition $P(v) = \prod_i Q[H_i]$. We next present a procedure for determining the recoverability of the

Function RC(D, P, G)
INPUT: D a set with G_D being a c-component, P a distribution, G a causal diagram over V and S node.
OUTPUT: Expression for $Q[D]$ in terms of P or FAIL

1. If $V \setminus (An(D) \cup An(S)) \neq \emptyset$,
   return RC($D, \sum_{V \setminus (An(D) \cup An(S))} P, G_{An(D) \cup An(S)}$)
2. Let $C_1, \ldots, C_k$ be the c-components of G that contains no ancestors of S, and let $C = \cup C_i$.
3. If there exists no such c-component (i.e., $C = \emptyset$), return FAIL.
4. If D is a subset of some C_i, return Identify($D, C_i, Q[C_i]$).
5. Return RC($D, \frac{P}{\prod_i Q[C_i]}, G_{V \setminus C}$).

Figure 4: Algorithm based on c-components capable of simultaneously identifying and recovering causal effects.

causal effect $P_t(r)$, where T and R are arbitrary subsets of V.

Let $T' = V \setminus T$, and $D = An(R)_{G_{T'}}$. We have

$$P_t(r) = \sum_{T' \setminus R} P_t(t') = \sum_{T' \setminus R} Q[T'] = \sum_{D \setminus R} Q[D] \quad (21)$$

$$= \sum_{D \setminus R} \prod_i Q[D_i], \quad (22)$$

where $D_1, \ldots, D_l$ are the set of c-components of the subgraph G_D. Now we call RC($D_i, P(v|S = 1), G$) for each D_i to determine the recoverability of $Q[D_i]$.

Theorem 3 *Function RC($D, P(v|S = 1), G$) is correct.*

Proof. We have

$$P(v|S = 1) = \frac{P(S = 1|pa_S)}{P(S = 1)} P(v). \quad (23)$$

The nonancestors of $D \cup S$ can be summed out from both sides without influencing the recoverability results.

$$P(An(D) \cup An(S)|S = 1) = \frac{P(S = 1|pa_S)}{P(S = 1)} P(An(D) \cup An(S)). \quad (24)$$

Now assume that nonancestors of $D \cup S$ have been summed out, now $P(v)$ can be decomposed into product:

$$P(v|S = 1) = \frac{P(S = 1|pa_S)}{P(S = 1)} Q[V \setminus C] \prod_i Q[C_i]. \quad (25)$$

From Lemma 3, if a component C_i contains no ancestors of S, then $Q[C_i]$ is recoverable. If D is a subset of some C_i, it is known that $Q[D]$ is identifiable if and only if it is identifiable from $Q[C_i]$ by the Identify($D, C_i, Q[C_i]$) algorithm in (Huang and Valtorta 2006). Therefore, if $Q[D]$ is not identifiable, it is not recoverable; if $Q[D]$ is identifiable from a recoverable $Q[C_i]$, then it is recoverable. If none of the recoverable c-components C_i contain D, then moving recoverable quantity $\prod_i Q[C_i]$ to the l.h.s. of the Eq. (25) we obtain

$$\frac{P(v|S = 1)}{\prod_i Q[C_i]} = \frac{P(S = 1|pa_S)}{P(S = 1)} Q[V \setminus C] \quad (26)$$

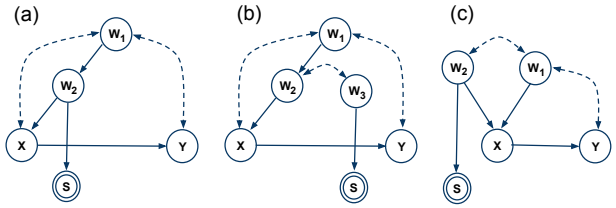

Figure 5: Non-trivial scenarios involving intrincate relationship of the counfounded structure and the S-nodes; $Q = P_x(y)$ is not recoverable in (a) but is in (b) and (c).

Now the problem of recovering $Q[D]$ is reduced to a problem of recovering $Q[D]$ in the subgraph $G_{V \setminus C}$ with distribution $P(v|S = 1)/\prod_i Q[C_i]$. □

Our procedure returns FAIL to recover $Q[Y]$ if there is a subgraph G_C that contains Y such that: all nodes in G_C are ancestors of S or Y, and every c-component of G_C contains an ancestor of S.

We demonstrate the application of our procedure RC(D, P, G) using a few examples. Remarkably, some of these quantities are non-trivial to derive manually. In Figure 5(a), $P_x(y) = Q[Y]$, and RC($\{Y\}, P(v|S = 1), G$) will return FAIL because all the nodes are ancestors of Y or S and neither c-components are recoverable by Lemma 3 (line 3). In the model in Figure 5(b), $P_x(y) = Q[Y]$, Y is in the c-component $\{Y, X, W_1\}$ which is recoverable by Lemma 3, and therefore $P_x(y)$ is recoverable (line 4). In Figure 5(c), $P_x(y) = Q[Y]$, the c-components are $\{X\}$ and $C' = \{Y, W_1, W_2\}$, and $Q[\{X\}]$ is recoverable by Lemma 3 as $Q[\{X\}] = P(x|w_1, w_2, S = 1)$ where $Q[C']$ is not. The problem is reduced to recovering $Q[Y]$ in the subgraph $G_{C'}$ by calling RC($Y, P(v|S = 1)/P(x|w_1, w_2, S = 1), G_{C'}$) (line 5). In $G_{C'}$, W_1 is not an ancestor of Y or S and can be summed out (line 1), which reduce the problem to RC($Y, \sum_{w_1} P(v|S = 1)/P(x|w_1, w_2, S = 1), G_{\{Y,W_2\}}$). In $G_{\{Y,W_2\}}$, Y is a c-component and is recoverable using Lemma 3 as $\sum_{w_1} P(v|S = 1)/P(x|w_1, w_2, S = 1)/\sum_{w_1,y} P(v|S = 1)/P(x|w_1, w_2, S = 1)$.

Conclusions

We provide conditions for recoverability from selection and confounding biases applicable for arbitrary structures in non-parametric settings. Theorem 2 provides a complete characterization of recoverability in Markovian models. Figure 4 (combined with Theorem 3) provides the most general procedure known to date for recoverability in semi-Markovian models. Verifying the conditions given in these theorems takes polynomial time and could be used to decide what measurements are needed for recoverability. Since confounding and selection biases are common problems across many disciplines, the methods developed in this paper should help to understand, formalize, and alleviate this problem in a broad range of data-intensive applications.

This paper complements recent work on *transportability* (Pearl and Bareinboim 2014; Bareinboim and Pearl 2013), which deals with transferring causal information across disparate, heterogeneous environments.

References

Angrist, J. D. 1997. Conditional independence in sample selection models. *Economics Letters* 54(2):103–112.

Bareinboim, E., and Pearl, J. 2012a. Causal inference by surrogate experiments: z-identifiability. In de Freitas, N., and Murphy, K., eds., *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI 2012)*, 113–120. AUAI Press.

Bareinboim, E., and Pearl, J. 2012b. Controlling selection bias in causal inference. In Girolami, M., and Lawrence, N., eds., *Proceedings of The Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2012)*, 100–108. JMLR (22).

Bareinboim, E., and Pearl, J. 2013. A general algorithm for deciding transportability of experimental results. *Journal of Causal Inference* 1(1):107–134.

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In Brodley, C., and Stone, P., eds., *Proceedings of the Twenty-Eight National Conference on Artificial Intelligence (AAAI 2014)*, 2410–2416. Menlo Park, CA: AAAI Press.

Cooper, G. 1995. Causal discovery from data in the presence of selection bias. *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics* 140–150.

Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, 38–53. Berlin, Heidelberg: Springer.

Elkan, C. 2001. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'01, 973–978. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Galles, D., and Pearl, J. 1995. Testing identifiability of causal effects. In Besnard, P., and Hanks, S., eds., *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann. 185–195.

Glymour, M., and Greenland, S. 2008. Causal diagrams. In Rothman, K.; Greenland, S.; and Lash, T., eds., *Modern Epidemiology*. Philadelphia, PA: Lippincott Williams & Wilkins, 3rd edition. 183–209.

Halpern, J. 1998. Axiomatizing causal reasoning. In Cooper, G., and Moral, S., eds., *Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann. 202–210.

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47(1):pp. 153–161.

Huang, Y., and Valtorta, M. 2006. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*. Menlo Park, CA: AAAI Press. 1149–1156.

Jewell, N. P. 1991. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* 59(2):227–240.

Kuroki, M., and Cai, Z. 2006. On recovering a population covariance matrix in the presence of selection bias. *Biometrika* 93(3):601–611.

Kuroki, M., and Miyakawa, M. 1999. Identifiability criteria for causal effects of joint interventions. *Journal of the Royal Statistical Society* 29:105–117.

Little, R. J. A., and Rubin, D. B. 1986. *Statistical Analysis with Missing Data*. NY, USA: John Wiley & Sons, Inc.

Mefford, J., and Witte, J. S. 2012. The covariate's dilemma. *PLoS Genet* 8(11):e1003096.

Pearl, J., and Bareinboim, E. 2014. External validity: From *do*-calculus to transportability across populations. Technical Report R-400, Cognitive Systems Lab, UCLA. *Statistical Scien*, forthcoming.

Pearl, J., and Paz, A. 2014. Confounding equivalence in causal equivalence. *Journal of Causal Inference* 2:77–93.

Pearl, J., and Robins, J. 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables. In Besnard, P., and Hanks, S., eds., *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*. San Francisco: Morgan Kaufmann. 444–453.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–710.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. Second ed., 2009.

Pirinen, M.; Donnelly, P.; and Spencer, C. 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics* 44:848–851.

Robins, J. 2001. Data, design, and background knowledge in etiologic inference. *Epidemiology* 12(3):313–320.

Shpitser, I., and Pearl, J. 2006. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*. Menlo Park, CA: AAAI Press. 1219–1226.

Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer-Verlag.

Tian, J., and Pearl, J. 2002a. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*. Menlo Park, CA: AAAI Press/The MIT Press. 567–573.

Tian, J., and Pearl, J. 2002b. On the testable implications of causal models with hidden variables. In Darwiche, A., and Friedman, N., eds., *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann. 519–527.

Tian, J. 2002. *Studies in Causal Reasoning and Learning*. Ph.D. Dissertation, Computer Science Department, University of California, Los Angeles, CA.

Whittemore, A. 1978. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, B* 40(3):328–340.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, 114–. New York, NY, USA: ACM.