Recovering from Selection Bias in Causal and Statistical Inference

Elias Bareinboim
Cognitive Systems Laboratory
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

Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles
Los Angeles, CA. 90095
judea@cs.ucla.edu

Abstract

Selection bias is caused by preferential exclusion of units from the samples and represents a major obstacle to valid causal and statistical inferences; it cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies. In this paper, we provide complete graphical and algorithmic conditions for recovering conditional probabilities from selection biased data. We also provide graphical conditions for recoverability when unbiased data is available over a subset of the variables. Finally, we provide a graphical condition that generalizes the backdoor criterion and serves to recover causal effects when the data is collected under preferential selection.

Introduction

Selection bias is induced by preferential selection of units for data analysis, usually governed by unknown factors including treatment, outcome, and their consequences, and represents a major obstacle to valid causal and statistical inferences. It cannot be removed by randomized experiments and can rarely be detected in either experimental or observational studies.[1] 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.

This preferential selection challenges the validity of inferences in several tasks in AI (Cooper 1995; Elkan 2001; Zadrozny 2004; Cortes et al. 2008) and 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)).

To illuminate the nature of preferential selection, consider the data-generating model in Fig. 1(a) in which X represents an action, Y represents an outcome, and S represents a binary indicator of entry into the data pool ($S = 1$ means that the unit is in the sample, $S = 0$ otherwise). If our goal is to compute the population-level conditional distribution $P(y|x)$, and the samples available are collected under selection, only $P(y, x|S = 1)$ is accessible for use. [2] Given that in principle these two distributions are just loosely connected, the natural question to ask is under what conditions $P(y|x)$ can be recovered from data coming from $P(y, x|S = 1)$. In this specific example, both action and outcome affect the entry in the data pool, which will be shown not to be recoverable (see Corollary 1) – i.e., there is no method capable of unbiasedly estimating the population-level distribution using data gathered under this selection process.

The bias arising from selection differs fundamentally from the one due to *confounding*, though both constitute threats to the validity of causal inferences. The former bias is due to treatment or outcome (or ancestors) affecting the inclusion of the subject in the sample (Fig. 1(a)), while the latter is the result of treatment X and outcome Y being affected by a common omitted variables U (Fig. 1(b)). In both cases, we have unblocked extraneous "flow" of information between treatment and outcome, which appear under the rubric of "spurious correlation," since it is not what we seek to estimate.

It is instructive to understand selection graphically, as in Fig. 1(a). The preferential selection that is encoded through conditioning on S creates spurious association between X and Y through two mechanisms. First, given that S is a collider, conditioning on it induces spurious association between its parents, X and Y (Pearl 1988). Second, S is also a descendant of a "virtual collider" Y, whose parents are X and the error term U_Y (also called "hidden variable") which is always present, though often not shown in the diagram.[3]

Related work and Our contributions

There are three sets of assumptions that are enlightening to acknowledge if we want to understand the procedures avail-

[1] Remarkably, there are special situations in which selection bias can be detected even from observations, as in the form of a non-chordal undirected component (Zhang 2008).

[2] In a typical AI task such as classification, we could have X being a collection of features and Y the class to be predicted, and $P(y|x)$ would be the classifier that needs to be trained.

[3] See (Pearl 2000, pp. 339-341) and (Pearl 2013) for further explanations of this bias mechanism.

able in the literature for treating selection bias – qualitative assumptions about the selection mechanism, parametric assumptions regarding the data-generating model, and quantitative assumptions about the selection process.

In the data-generating model in Fig. 1(c), the selection of units to the sample is treatment-dependent, which means that it is caused by X, but not Y. This case has been studied in the literature and $Q = P(y|x)$ is known to be non-parametrically recoverable from selection (Greenland and Pearl 2011). Alternatively, in the data-generating model in Fig. 1(d), the selection is caused by Y (outcome-dependent), and Q is not recoverable from selection (formally shown later on), but is the odds ratio[4] (Cornfield 1951; Whittemore 1978; Geng 1992; Didelez, Kreiner, and Keiding 2010). As mentioned earlier, Q is also not recoverable in the graph in Fig. 1(a). By and large, the literature is concerned with treatment-dependent or outcome-dependent selection, but selection might be caused by multiple reasons and embedded in more intricate realities. For instance, a driver of the treatment Z (e.g., age, sex, socio-economic status) may also be causing selection, see Fig. 1(e,f). As it turns out, Q is recoverable in Fig 1(e) but not in (f), so different qualitative assumptions need to be modelled explicitly since each topology entails a different answer for recoverability.

The second assumption is related to the parametric form used by recoverability procedures. For instance, one variation of the selection problem was studied in Econometrics, and led to the celebrated method developed by James Heckman (Heckman 1979). His two-step procedure removes the bias by leveraging the assumptions of linearity and normality of the data-generating model. A graph-based parametric analysis of selection bias is given in (Pearl 2013).

The final assumption is about the probability of being selected into the sample. In many settings in Machine learning and Statistics (Elkan 2001; Zadrozny 2004; Smith and Elkan 2007; Storkey 2009; Hein 2009; Cortes et al. 2008), it is assumed that this probability, $P(S = 1|Pa_s)$, can be modelled explicitly, which often is an unattainable requirement for the practitioner (e.g., it might be infeasible to assess the differential rates of how salaries are reported).

Our treatment differs fundamentally from the current literature regarding these assumptions. First, we do not constrain the type of data-generating model as outcome- or treatment-dependent, but we take arbitrary models (including these two) as input, in which a node S indicates selection for sampling. Second, we do not make parametric assumptions (e.g. linearity, normality, monotonicity) but operate non-parametrically based on causal graphical models (Pearl 2000), which is more robust, less prone to model misspecifications. Third, we do not rely on having the selection's probability $P(S = 1|Pa_s)$, which is not always available in practice. Our work hinges on exploiting the qualitative knowledge encoded in the data-generating model for yielding recoverability. This knowledge is admittedly a demanding requirement for the scientist, but we now understand formally its necessity for *any* approach to recoverability – any procedure aiming for recoverability, implicitly or explicitly, relies on this knowledge (Pearl 2000). [5]

[4]The odds ratio (OR) is a commonly used measure of association and has the form $\big(P(y|x)P(\overline{y}|\overline{x})\big)/\big(P(\overline{y}|x)P(y|\overline{x})\big)$. The symmetric form of the OR allows certain derivations.

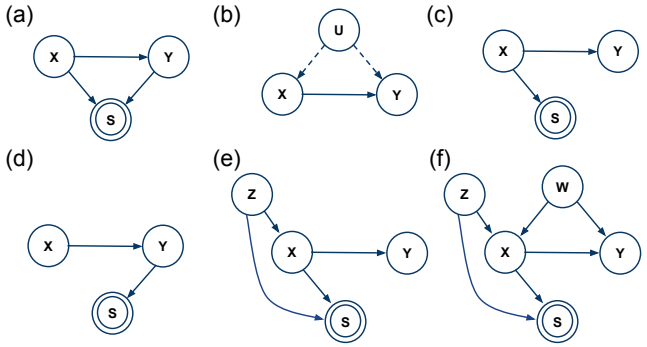

Figure 1: (a,b) Simplest examples of selection and confounding bias, respectively. (c,d) Treatment-dependent and outcome-dependent studies under selection, $Q = P(y|x)$ is recoverable in (c) but not in (d). (e,f) Treatment-dependent study where selection is also affected by driver of treatment Z (e.g., age); Q is recoverable in (e) but not in (f).

The analysis of selection bias requires a formal language within which the notion of data-generating model is given precise characterization, and the qualitative assumptions regarding how the variables affect selection can be encoded explicitly. The advent of causal diagrams (Pearl 1995; Spirtes, Glymour, and Scheines 2000; Pearl 2000; Koller and Friedman 2009) provides such a language and renders the formalization of the selection problem possible.

Using this language, (Bareinboim and Pearl 2012) provided a complete treatment for selection relative to the OR[4]. We generalize their treatment considering the estimability of conditional distributions and address three problems:

1. **Selection without external data:** The dataset is collected under selection bias, $P(\mathbf{v}|S = 1)$; under which conditions is $P(y|\mathbf{x})$ recoverable?
2. **Selection with external data:** The dataset is collected under selection bias, $P(\mathbf{v}|S = 1)$, but there are unbiased samples from $P(\mathbf{t})$, for $\mathbf{T} \subseteq \mathbf{V}$; under which conditions is $P(y|\mathbf{x})$ recoverable?
3. **Selection in causal inferences:** The data is collected under selection bias, $P(\mathbf{v}|S = 1)$, but there are unbiased samples from $P(\mathbf{t})$, for $\mathbf{T} \subseteq \mathbf{V}$; under which conditions is the interventional distribution $P(y|do(\mathbf{x}))$ estimable?

We provide graphical and algorithmic conditions for these problems without resorting to parametric assumptions nor selection probabilities. Furthermore, the solution for selection without external data is complete, in the sense that whenever a quantity is said not to be recoverable by our conditions, there exists no procedure that are able to recover it (without adding assumptions). In estimating the effects of interventions, we generalize the *backdoor criterion* for when data is collected under selection.

[5]A trivial instance of this necessity is Fig. 1(c,d) where the odds ratio is recoverable, yet $P(y|x)$ is recoverable in 1(c) but not in (d).

Recoverability without External Data

We first introduce the formal notion of recoverability for conditional distributions when data is under selection. [6]

Definition 1 (s-Recoverability)**.** *Given a causal graph G_s augmented with a node S encoding the selection mechanism (Bareinboim and Pearl 2012), the distribution $Q = P(y \mid \mathbf{x})$ is said to be s-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, $P_1(\mathbf{v} \mid S = 1) = P_2(\mathbf{v} \mid S = 1) > 0$ implies $P_1(y \mid \mathbf{x}) = P_2(y \mid \mathbf{x})$.* [7]

Consider the graph G_s in Fig. 1(c) and assume that our goal is to establish s-recoverability of $Q = P(y|x)$. Note that by d-separation (Pearl 1988), X separates Y from S, (or $(Y \perp\!\!\!\perp S|X)$), so we can write $P(y|x) = P(y|x, S = 1)$. This is a very special situation since these two distributions can be arbitrarily distant from each other, but in this specific case G_s constrains Q in such a way that despite the fact that data was collected under selection and our goal is to answer a query about the overall population, there is no need to resort to additional data external to the biased study.

Now we want to establish whether Q is s-recoverable in the graph G_s in Fig. 1(d). In this case, S is not d-separated from Y if we condition on X, so $(S \perp\!\!\!\perp Y|X)$ does not hold in at least one distribution compatible with G_s, and the identity $P(y|x) = P(y|x, S = 1)$ is not true in general. One may wonder if there is another way to s-recover Q in G_s, but this is not the case as formally shown next. That is, the assumptions encoded in G_s imply a universal impossibility; no matter how many samples of $P(x, y|S = 1)$ are accumulated or how sophisticated the estimation technique is, the estimator of $P(y|x)$ will never converge to its true value.

Lemma 1. *$P(y|x)$ is not s-recoverable in Fig. 1(d).*

Proof. We construct two causal models such that P_1 is compatible with the graph G_s in Fig. 1(d) and P_2 with the subgraph $G_2 = G_s \setminus \{Y \to S\}$. We will set the parameters of P_1 through its factors and then computing the parameters of P_2 by enforcing $P_2(\mathbf{V} \mid S = 1) = P_1(\mathbf{V} \mid S = 1)$. Since $P_2(\mathbf{V}|S = 1) = P_2(\mathbf{V})$, we will be enforcing $P_1(\mathbf{V}|S = 1) = P_2(\mathbf{V})$. Recoverability should hold for any parametrization, so we assume that all variables are binary. Given a Markovian causal model (Pearl 2000), P_1 can be parametrized through its factors in the decomposition over observables, $P_1(X), P_1(Y|X), P_1(S = 1|Y)$, for all X, Y.

We can write the conditional distribution in the second causal model as follows:

$$P_2(y|x) = P_1(y|x, S = 1) = \frac{P_1(y, x, S = 1)}{P_1(x, S = 1)} \qquad (1)$$

$$= \frac{P_1(S = 1|y)P_1(y|x)}{P_1(S = 1|y)P_1(y|x) + P_1(S = 1|\overline{y})P_1(\overline{y}|x)}, \qquad (2)$$

where the first equality, by construction, should be enforced, and the second and third by probability axioms. The other parameters of P_2 are free and can be chosen to match P_1.

Finally, set the distribution of every family in P_1 but selection variable equal to $1/2$, and set the distribution $P_1(S = 1|y) = \alpha, P_1(S = 1|\overline{y}) = \beta$, for $0 < \alpha, \beta < 1$ and $\alpha \neq \beta$. This parametrization reduces eq. (2) to $P_2(y|x) = \alpha/(\alpha+\beta)$ and $P_1(y|x) = 1/2$, the result follows. □

Corollary 1. *$P(y|x)$ is not s-recoverable in Fig. 1(a).*

The corollary follows immediately noting that lack of s-recoverability with a subgraph (Fig. 1(d)) precludes s-recoverability with the graph itself since the extra edge can be inactive in a compatible parametrization (Pearl 1988) (the converse is obviously not true). Lemma 1 is significant because Fig. 1(d) can represent a study design that is typically used in empirical fields known as case-control studies. The result is also theoretically instructive since Fig. 1(d) represents the smallest graph structure that is not s-recoverable, and its proof will set the tone for more general and arbitrary structures that we will be interested in (see Theorem 1).

Furthermore, consider the graph in Fig. 1(e) in which the independence $(S \perp\!\!\!\perp Y|X)$ holds, so we can also recover Q from selection ($P(y|x, S = 1) = P(y|x)$). However, $(S \perp\!\!\!\perp Y|X)$ does not hold in Fig. 1(f) – there is an open path passing through X's ancestor W (i.e. $S \leftarrow Z \rightarrow X \leftarrow W \rightarrow Y$) – and the natural question that arises is whether Q is recoverable in this case. It does not look obvious whether the absence of an independence precludes s-recoverability since there are other possible operators in probability theory that could be used leading to the s-recoverability of Q. To illustrate this point, note that it is not the case in causal inference that the inapplicability of the backdoor criterion (Pearl 2000, Ch. 3), which is also an independence constraint, implies the impossibility of recovering certain effects.

Remarkably, the next result states that the lack of this independence indeed precludes s-recoverability, i.e., the probe of one separation test in the graph is sufficient to evaluate whether a distribution is or is not s-recoverable.

Theorem 1. *The distribution $P(y|\mathbf{x})$ is s-recoverable from G_s if and only if $(S \perp\!\!\!\perp Y|\mathbf{X})$.* [8]

In words, Theorem 1 provides a powerful test for s-recoverability without external data, which means that when it disavows s-recoverability, there exists no procedure that would be capable of recovering the distribution from selection bias (without adding assumptions). Its sufficiency part is immediate, but the proof of necessity is somewhat involved since we need to show that for *all* graphical structures in which the given d-separation test fails, each of these structures does not allow for s-recoverability (i.e., a counterexample can always be produced showing agreement on $P(\mathbf{v}|S = 1)$ and disagreement on $P(y|\mathbf{x})$).

The next corollary provides a test for s-recoverability of broader joint distributions (including Y alone):

[6] This definition generalizes G-admissibility given in (Bareinboim and Pearl 2012).

[7] We follow the conventions given in (Pearl 2000). We use typical graph notation with families (e.g., children, parents, ancestors). We denote variables by capital letters and their realized values by small letters. We use bold to denote sets of variables. We denote the set of all variables by $\mathbf{V}$, except for the selection mechanism S.

[8] Please refer to the Appendix 2 in the full report for the proofs (Bareinboim, Tian, and Pearl 2014).

Corollary 2. *Let* $\mathbf{Z} = An(S) \setminus An(Y)$ *including* S*, and* $\mathbf{A} = Pa(\mathbf{Z}) \cap (An(Y) \setminus \{Y\})$*.* $P(Y, An(Y) \setminus (\mathbf{A} \setminus \{Y\}) | \mathbf{A})$ *is s-recoverable if and only if* Y *is not an ancestor of* S*.*

This result can be embedded as a step reduction in an algorithm to s-recover a collection of distributions in the form of the corollary. We show such algorithm in (Bareinboim, Tian, and Pearl 2014) [9]. The main idea is to traverse the graph in a certain order s-recovering all joint distributions with the form given in the corollary (updating S along the way). If the algorithm exits with failure, it means that the distributions of its predecessors are not s-recoverable.

Recoverability with External Data

A natural question that arises is whether additional measurements in the population level over certain variables can help recovering a given distribution. For example, $P(age)$ can be estimated from census data which is not under selection bias.

To illustrate how this problem may arise in practice, consider Fig. 2 and assume that our goal is to s-recover $Q = P(y|x)$. It follows immediately from Thm. 1 that Q cannot be s-recovered without additional assumptions. Note, however, that the parents of the selection node $\mathbf{Pa_s} = \{W_1, W_2\}$ separates S from all other nodes in the graph, which indicates that it would be sufficient for recoverability to measure $\mathbf{T} = \{W_1, W_2\} \cup \{X\}$ from external sources. To witness, note that after conditioning Q on W_1 and W_2, we obtain:

$$\begin{aligned} P(y|x) &= \sum_{w_1,w_2} P(y|x, w_1, w_2) P(w_1, w_2|x) \\ &= \sum_{w_1,w_2} P(y|x, w_1, w_2, S = 1) P(w_1, w_2|x), \quad (3) \end{aligned}$$

where the last equality follows from $(Y \perp\!\!\!\perp S \mid X, W_1, W_2)$. That is, Q can be s-recovered and is a combination of two different types of data; the first factor comes from biased data under selection, and the second factor is available from external data collected over the whole population.

Our goal is to understand the interplay between measurements taken over two types of variables, $\mathbf{M}, \mathbf{T} \subseteq V$, where $\mathbf{M}$ are variables collected under selection bias, $P(\mathbf{M}|S = 1)$, and $\mathbf{T}$ are variables collected in the population-level, $P(\mathbf{T})$. In other words, we want to understand when (and how) can this new piece of evidence $P(\mathbf{T})$ together with the data under selection ($P(\mathbf{M}|S = 1)$) help in extending the treatment of the previous section for recovering the true underlying distribution $Q = P(y|\mathbf{x})$.[10]

Formally, we need to redefine s-recoverability for accommodating the availability of data from external sources.

Definition 2 (s-Recoverability)**.** *Given a causal graph* G_s *augmented with a node* S*, the distribution* $Q = P(y \mid \mathbf{x})$ *is said to be* s*-recoverable from selection bias in* G_s *with external information over* $\mathbf{T} \subseteq \mathbf{V}$ *and selection biased data over* $\mathbf{M} \subseteq \mathbf{V}$ *(for short, s-recoverable) if the assumptions embedded in the causal model render* Q *expressible in terms of* $P(\mathbf{m} \mid S = 1)$ *and* $P(\mathbf{t})$*, both positive. Formally, for every two probability distributions* P_1 *and* P_2 *compatible with* G_s*, if they agree on the available distributions,* $P_1(\mathbf{m} \mid S = 1) = P_2(\mathbf{m} \mid S = 1) > 0$*,* $P_1(\mathbf{t}) = P_2(\mathbf{t}) > 0$*, they must agree on the query distribution,* $P_1(y \mid \mathbf{x}) = P_2(y \mid \mathbf{x})$*.*

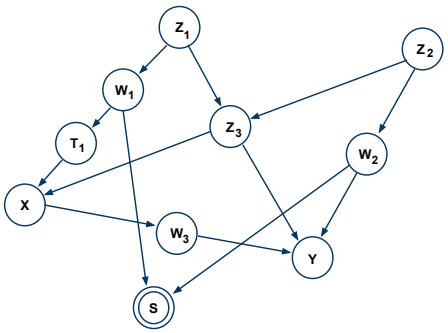

Figure 2: Causal model in which $Q = P(y|x)$ is not recoverable without external data (Thm. 1), but it is recoverable if measurements on the set $\mathbf{Pa_s} = \{W_1, W_2\}$ are taken (Thm. 2). Alternatively, even if not all parents of S are measured, any set including $\{W_2, Z_3\}$ would yield recoverability of Q.

The observation leading to eq. (3) provides a simple condition for s-recoverability when we can choose the variables to be collected. Let $\mathbf{Pa_s}$ be the parent set of S. If measurements on the set $\mathbf{T} = \mathbf{Pa_s} \cup \{X\}$ can be taken without selection, we can write $P(y|x) = \sum_{\mathbf{pa_s}} P(y|x, \mathbf{pa_s}, S = 1) P(\mathbf{pa_s}|x)$, since S is separated from all nodes in the graph given its parent set. This implies s-recoverability where we have a mixture in which the first factor is obtainable from the biased data and the second from external sources.

This solution is predicated on the assumption that $\mathbf{Pa_s}$ can be measured in the overall population, which can be a strong requirement, and begs a generalization to when part of $\mathbf{Pa_s}$ is not measured. For instance, what if in Fig. 2 W_1 cannot be measured? Would other measurements over a different set of variables also entail s-recoverability?

This can be expressed as a requirement that subsets of $\mathbf{T}$ and $\mathbf{M}$ can be found satisfying the following criterion:

Theorem 2. *If there is a set* $\mathbf{C}$ *that is measured in the biased study with* $\{\mathbf{X}, Y\}$ *and in the population level with* $\mathbf{X}$ *such that* $(Y \perp\!\!\!\perp S | \{\mathbf{C}, \mathbf{X}\})$*, then* $P(y|\mathbf{x})$ *is s-recoverable as*

$$P(y|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c}, S = 1) P(\mathbf{c}|\mathbf{x}). \quad (4)$$

In the example in Fig. 2, it is trivial to confirm that any (pre-treatment) set $\mathbf{C}$ containing W_2 and Z_3 would satisfy the conditions of the theorem. In particular, $\{W_2, Z_3\}$ is such a set, and it allows us to s-recover Q without measuring W_1 ($W_1 \in \mathbf{Pa_s}$) through eq. (4). Note, however, that the set $\mathbf{C} = \{W_2, Z_1, Z_2\}$ is not sufficient for s-recoverability. It fails to satisfy the separability condition of the theorem since

[9] This listing is useful when one needs to examine properties of the collection of distributions, analogously to the list of all back-door admissible sets by (Textor and Liskiewicz 2011)).

[10] This problem subsumes the one given in the previous section since when $\mathbf{T} = \emptyset$, the two problems coincide. We separate them since they come in different shades in the literature and also just after solving the version without external data we can aim to solve its more general version; we discuss more about this later on.

conditioning on $\{X, W_2, Z_1, Z_2\}$ leaves an unblocked path between S and Y (i.e., $S \leftarrow W_1 \rightarrow T_1 \rightarrow X \leftarrow Z_3 \rightarrow Y$).

It can be computationally difficult to find a set satisfying the conditions of the theorem since this could imply a search over a potentially exponential number of subsets. Remarkably, the next result shows that the existence of such a set can be determined by a single d-separation test.

Theorem 3. *There exists some set* $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ *such that* $Y \perp\!\!\!\perp S|\{\mathbf{C}, \mathbf{X}\}$ *if and only if the set* $(\mathbf{C}' \cup \mathbf{X})$ *d-separates* S *from* Y *where* $\mathbf{C}' = [(\mathbf{T} \cap \mathbf{M}) \cap An(Y \cup S \cup \mathbf{X})] \setminus (Y \cup S \cup \mathbf{X})$.

In practice, we can restrict ourselves to minimal separators, that is, looking only for minimal set $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ such that $(Y \perp\!\!\!\perp S|\{\mathbf{C}, \mathbf{X}\})$. The algorithm for finding minimal separators has been given in (Acid and de Campos 1996; Tian, Paz, and Pearl 1998).

Despite the computational advantages given by Thm. 3, Thm. 2 still requires the existence of a separator $\mathbf{C}$ measured in both the biased study ($\mathbf{M}$) and in the overall population ($\mathbf{T}$), and it is natural to ask whether this condition can be relaxed. Assume that all we have is a separator $\mathbf{C} \subseteq \mathbf{M}$, but $\mathbf{C}$ (or some of its elements) is not measured in population $\mathbf{T}$, and therefore $P(\mathbf{c}|\mathbf{x})$ in eq. (4) still needs to be s-recovered. We could s-recover $P(\mathbf{c}|\mathbf{x})$ in the spirit of Thm. 2 as

$$P(\mathbf{c}|\mathbf{x}) = \sum_{\mathbf{c_1}} P(\mathbf{c}|\mathbf{x}, \mathbf{c_1}, S = 1) P(\mathbf{c_1}|\mathbf{x}), \qquad (5)$$

if there exists a set $\mathbf{C_1} \subseteq \mathbf{M} \cap \mathbf{T}$ such that $(S \perp\!\!\!\perp \mathbf{C}|\mathbf{X}, \mathbf{C_1})$. Now if this fails in that we can only find a separator $\mathbf{C_1} \subseteq \mathbf{M}$ not measured in $\mathbf{T}$, we can then attempt to recover $P(\mathbf{c_1}|\mathbf{x})$ in the spirit of Thm. 2 by looking for another separator $\mathbf{C_2}$, and so on. At this point, it appears that Thm. 2 can be extended.

We further extend this idea by considering other possible probabilistic manipulations and embed them in a recursive procedure. For $\mathbf{W}, \mathbf{Z} \subseteq \mathbf{M}$, consider the problem of recovering $P(\mathbf{w}|\mathbf{z})$ from $P(\mathbf{t})$ and $P(\mathbf{m}|S = 1)$, and define procedure $RC(\mathbf{w}, \mathbf{z})$ as follows:

1. If $\mathbf{W} \cup \mathbf{Z} \subseteq \mathbf{T}$, then $P(\mathbf{w}|\mathbf{z})$ is s-recoverable.
2. If $(S \perp\!\!\!\perp \mathbf{W}|\mathbf{Z})$, then $P(\mathbf{w} \mid \mathbf{z})$ is s-recoverable as $P(\mathbf{w} \mid \mathbf{z}) = P(\mathbf{w} \mid \mathbf{z}, S = 1)$.
3. For minimal $\mathbf{C} \subseteq \mathbf{M}$ such that $(S \perp\!\!\!\perp \mathbf{W}|(\mathbf{Z} \cup \mathbf{C}))$, $P(\mathbf{w}|\mathbf{z}) = \sum_c P(\mathbf{w}|\mathbf{z}, \mathbf{c}, S = 1) P(\mathbf{c}|\mathbf{z})$. If $\mathbf{C} \cup \mathbf{Z} \subseteq \mathbf{T}$, then $P(\mathbf{w}|\mathbf{z})$ is s-recoverable. Otherwise, call $RC(\mathbf{c}, \mathbf{z})$.
4. For some $\mathbf{W}' \subset \mathbf{W}$, $P(\mathbf{w}|\mathbf{z}) = P(\mathbf{w}'|\mathbf{w} \setminus \mathbf{w}', \mathbf{z}) P(\mathbf{w} \setminus \mathbf{w}'|\mathbf{z})$. Call $RC(\mathbf{w}', \{\mathbf{w} \setminus \mathbf{w}'\} \cup \mathbf{z})$ and $RC(\mathbf{w} \setminus \mathbf{w}', \mathbf{z})$).
5. Exit with FAIL (to s-recover $P(\mathbf{w}|\mathbf{z})$) if for a singleton $\mathbf{W}$, none of the above operations are applicable.

Now, we define recoverability based on this procedure:

Definition 3. *We say that* $P(\mathbf{w}|\mathbf{z})$ *is C-recoverable if and only if it is recovered by the procedure* $RC(\mathbf{w}, \mathbf{z})$.

Remarkably, the manipulations considered in $RC()$ are not actually more powerful than Thm. 2, as shown next.

Theorem 4. *For* $\mathbf{X} \subseteq \mathbf{T}$, $Y \notin \mathbf{T}$, $Q = P(\mathbf{y}|\mathbf{x})$ *is C-recoverable if and only if it is recoverable by Theorem 2, that is, if and only if there exists a set* $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ *such that* $(Y \perp\!\!\!\perp S|\{\mathbf{C}, \mathbf{X}\})$ *(where* $\mathbf{C}$ *could be empty). If s-recoverable,* $P(y|\mathbf{x})$ *is given by* $P(y|\mathbf{x}) = \sum_{\mathbf{c}} P(y|\mathbf{x}, \mathbf{c}, S = 1) P(\mathbf{c}|\mathbf{x})$.

This result suggests that the constraint between measurement sets cannot be relaxed through ordinary decomposition and Thm. 2 captures the bulk of s-recoverable relations. (See proof in (Bareinboim, Tian, and Pearl 2014).) Importantly, this does not constitute a proof of necessity of Thm. 2.

Now we turn our attention to some special cases that appear in practice. Note that, so far, we assumed X being measured in the overall population, but in some scenarios Y's prevalence might be available instead. So, assume $Y \in \mathbf{T}$ but some variables in $\mathbf{X}$ are not measured in the population-level. Let $\mathbf{X^0} = \mathbf{X} \cap \mathbf{T}$ and $\mathbf{X^m} = \mathbf{X} \setminus \mathbf{X^0}$, we have

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x^m}|y, \mathbf{x^0}) p(y|\mathbf{x^0})}{\sum_y P(\mathbf{x^m}|y, \mathbf{x^0}) p(y|\mathbf{x^0})} \qquad (6)$$

Therefore, $P(y|\mathbf{x})$ is recoverable if $P(\mathbf{x^m}|y, \mathbf{x^0})$ is recoverable. We could use the previous results to recover $P(\mathbf{x^m}|y, \mathbf{x^0})$. In particular, Theorems 2 and 3 lead to:

Corollary 3. $P(y|x)$ *is recoverable if there exists a set* $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ *(*$\mathbf{C}$ *could be empty) such that* $(\mathbf{X^m} \perp\!\!\!\perp S|\{\mathbf{C} \cup \mathbf{Y} \cup \mathbf{X^0}\})$. *If recoverable,* $P(y|\mathbf{x})$ *is given by Eq. (6) where*

$$P(\mathbf{x^m}|y, \mathbf{x^0}) = \sum_c P(\mathbf{x^m}|y, \mathbf{x^0}, \mathbf{c}, S = 1) P(c|y, \mathbf{x^0}) \qquad (7)$$

Corollary 4. $P(y|\mathbf{x})$ *is recoverable via Corollary 3 if and only if the set* $(\mathbf{C}' \cup Y \cup \mathbf{X^0})$ *d-separates* S *from* $\mathbf{X^m}$ *where* $C' = [(\mathbf{T} \cap \mathbf{M}) \cap An(Y \cup S \cup \mathbf{X})] \setminus (Y \cup S \cup \mathbf{X})$.

For example, in Fig. 2, assuming $\mathbf{M} = \{X, Y, W_1, W_3, Z_3\}$ and $\mathbf{T} = \{Y, W_1, W_3, Z_3\}$, we have $S \perp\!\!\!\perp X|\{Y, W_1, W_3, Z_3\}$, therefore we can s-recover

$$P(x|y) = \sum_{w_1, w_3, z_3} P(x|y, w_1, w_3, z_3, S = 1) P(w_1, w_3, z_3|y), \qquad (8)$$

as well as $P(y|x)$ by substituting back eq. (8) in eq. (6).

Furthermore, it is worth examining when no data is gathered over $\mathbf{X}$ or Y in the population level. In this case, $P(y|\mathbf{x})$ may be recoverable through $P(\mathbf{x}, y)$, as shown in the sequel.

Corollary 5. $P(y|\mathbf{x})$ *is recoverable if there exists a set* $\mathbf{C} \subseteq \mathbf{T} \cap \mathbf{M}$ *such that* $(\{Y\} \cup \mathbf{X} \perp\!\!\!\perp S|\mathbf{C})$. *If recoverable,* $P(y, \mathbf{x})$ *is given by* $P(y, \mathbf{x}) = \sum_{\mathbf{c}} P(y, \mathbf{x}|\mathbf{c}, S = 1) P(\mathbf{c})$.

For instance, $P(x, y)$ is s-recoverable in Fig. 2 if $\mathbf{T} \cap \mathbf{M}$ contains $\{W_2, T_1, Z_3\}$ or $\{W_2, T_1, Z_1\}$ (without $\{X, Y\}$) .

Recoverability of Causal Effects

We now turn our attention to the problem of estimating causal effects from selection biased data. [11]

Our goal is to recover the effect of X on Y, $P(y|do(x))$ given the structure of G_s. Consider the graph G_s in Fig. 3(a), in which X and Y are not confounded, hence, $P(y|do(x)) = P(y|x)$ and, based on Theorem 1, we conclude that $P(y|do(x))$ is not recoverable in G_s. Fig. 3(b) and 3(c), on the other hand, contains covariates W_1 and W_2 that may satisfy conditions similar to those in Theorem

[11] We assume the graph G_s represents a causal model, as defined in (Pearl 2000; Spirtes, Glymour, and Scheines 2000).

1 that would render $P(y|do(x))$ recoverable. These conditions, however, need to be strengthened significantly, to account for possible confounding between X and Y which, even in the absence of selection bias, might require adjustment for admissible covariates, namely, covariates that satisfy the backdoor condition (Pearl 1993). For example, $\{W_2\}$ satisfies the backdoor condition in both Fig. 3(b) and (c), while $\{W_1\}$ satisfies this condition in (b) but not in (c).

Definition 4 below extends the backdoor condition to selection bias problems by identifying a set of covariates $\mathbf{Z}$ that accomplishes two functions. Conditions (i) and (ii) assure us that $\mathbf{Z}$ is backdoor admissible (Pearl and Paz 2013) [12], while conditions (iii) and (iv) act to separate S from Y, so as to permit recoverability from selection bias.

Definition 4 (Selection-backdoor criterion)**.** *Let a set $\mathbf{Z}$ of variables be partitioned into $\mathbf{Z}^+ \cup \mathbf{Z}^-$ such that $\mathbf{Z}^+$ contains all non-descendants of X and $\mathbf{Z}^-$ the descendants of X. $\mathbf{Z}$ is said to satisfy the selection backdoor criterion (s-backdoor, for short) relative to an ordered pairs of variables (X, Y) and an ordered pair of sets $(\mathbf{M}, \mathbf{T})$ in a graph G_s if $\mathbf{Z}^+$ and $\mathbf{Z}^-$ satisfy the following conditions:*

(i) $\mathbf{Z}^+$ blocks all back door paths from X to Y;

(ii) X and $\mathbf{Z}^+$ block all paths between $\mathbf{Z}^-$ and Y, namely, $(\mathbf{Z}^- \perp\!\!\!\perp Y | X, \mathbf{Z}^+)$;

(iii) X and $\mathbf{Z}$ block all paths between S and Y, namely, $(Y \perp\!\!\!\perp S | X, \mathbf{Z})$;

(iv) $\mathbf{Z} \cup \{X, Y\} \subseteq \mathbf{M}$, and $\mathbf{Z} \subseteq \mathbf{T}$.

Consider Fig. 3(a) where $\mathbf{Z}^- = \{W\}, \mathbf{Z}^+ = \{\}$ and $\mathbf{Z}^-$ is *not* separated from Y given $\{X\} \cup \mathbf{Z}^+$ in G_s, which means that condition (ii) of the s-backdoor is violated. So, despite the fact that the relationship between X and Y is unconfounded and $(Y \perp\!\!\!\perp S | \{W, X\})$, it is improper to adjust for $\{W\}$ when computing the target effect.

For the admissible cases, we are ready to state a sufficient condition that guarantees proper identifiability and recoverability of causal effects under selection bias:

Theorem 5 (Selection-backdoor adjustment)**.** *If a set $\mathbf{Z}$ satisfies the s-backdoor criterion relative to the pairs (X, Y) and $(\mathbf{M}, \mathbf{T})$ (as given in def. 2), then the effect of X on Y is identifiable and s-recoverable and is given by the formula*

$$P(y|do(x)) = \sum_{z} P(y|x, \mathbf{z}, S = 1) P(\mathbf{z}) \qquad (9)$$

Interestingly, X does not need to be measured in the overall population when the s-backdoor adjustment is applicable, which contrasts with the expression given in Theorem 2 where both X and $\mathbf{Z}$ (equivalently $\mathbf{C}$) are needed.

Consider Fig. 3(b) and assume our goal is to establish $Q = P(y|do(x))$ when external data over $\{W_2\}$ is available in both studies. Then, $\mathbf{Z} = \{W_2\}$ is s-backdoor admissible and the s-backdoor adjustment is applicable in this case. However, if $\mathbf{T} = \{W_1\}$, $\mathbf{Z} = \{W_1\}$ is backdoor admissible, but it is *not* s-backdoor admissible since condition (iii) is violated (i.e., $(S \perp\!\!\!\perp Y | \{W_1, X\})$ does not hold in G_s). This is interesting since the two sets $\{W_1\}$ and $\{W_2\}$ are c-equivalent (Pearl and Paz 2013), having the same potential for bias reduction in the general population. To understand why c-equivalence is not sufficient for s-recoverability, note that despite the equivalence for adjustment, $\sum_{w_1} P(y|x, w_1)P(w_1) = \sum_{w_2} P(y|x, w_2)P(w_2)$, the r.h.s. is obtainable from the data, while the l.h.s. is not.

[12] These two conditions extend the usual backdoor criterion (Pearl 1993) to allow descendants of X to be part of Z.

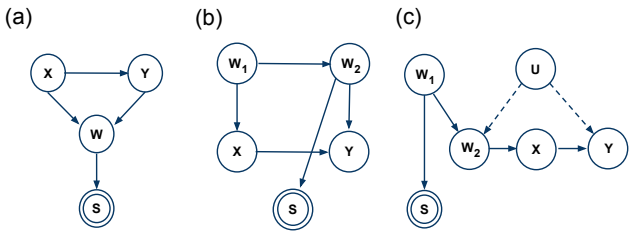

Figure 3: (a) Causal diagram in which $(S \perp\!\!\!\perp Y | \{X, W\})$ but $P(y|do(x))$ is not s-backdoor admissible. (b) $P(y|do(x))$ is s-recoverable through $\mathbf{T} = \{W_2\}$ but not $\{W_1\}$. (c) $\{W_2\}$ does not satisfy the s-backdoor criterion but $P(y|do(x))$ is still recoverable.

Now we want to recover $Q = P(y|do(x))$ in Fig. 3(c) (U is a latent variable) with $\mathbf{T} = \{W_2\}$. Condition (iii) of the s-backdoor fails since $(S \perp\!\!\!\perp Y | \{X, W_2\})$ does not hold. Alternatively, if we discard W_2 and consider the null set for adjustment ($\mathbf{Z} = \{\}$), condition (i) fails since there is an open backdoor path from X to Y ($X \leftarrow W_2 \leftarrow U \rightarrow Y$). Despite the inapplicability of the s-backdoor, $P(y|do(x))$ is still s-recoverable since, using do-calculus, we can show that $Q = P(y|do(x), S = 1)$, which reduces to $\sum_{w_2} P(y|x, w_2, S = 1)P(w_2|S = 1)$, both factors s-recoverable without the need for external information.

The reliance on the do-calculus in recovering causal effects is expected since even when selection bias is absent, there exist identifiability results beyond the backdoor. Still, this criterion, which is generalized by the s-backdoor criterion, is arguably the most used method for identifiability of causal effects currently available in the literature.

Conclusions

We provide conditions for recoverability from selection bias in statistical and causal inferences applicable for arbitrary structures in non-parametric settings. Theorem 1 provides a complete characterization of recoverability when no external information is available. Theorem 2 provides a sufficient condition for recoverability based on external information; it is optimized by Theorem 3 and strengthened by Theorem 4. Verifying these conditions takes polynomial time and could be used to decide what measurements are needed for recoverability. Theorem 5 further gives a graphical condition for recovering causal effects, which generalizes the backdoor adjustment. Since selection bias is a common problem 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 another aspect of the generalization problem in which causal effects are transported among differing environments (Bareinboim and Pearl 2013a; 2013b).

Acknowledgments

The authors would like to thank the reviewers for their comments that help improve the manuscript. This research was supported in parts by grants from NSF #IIS-1249822 and #IIS-1302448, and ONR #N00014-13-1-0153 and #N00014-10-1-0933.

References

Acid, S., and de Campos, L. 1996. An algorithm for finding minimum d-separating sets in belief networks. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence*, 3–10. San Francisco, CA: Morgan Kaufmann.

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

Bareinboim, E., and Pearl, J. 2012. 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. 2013a. Meta-transportability of causal effects: A formal approach. In *Proceedings of The Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, 135–143. JMLR (31).

Bareinboim, E., and Pearl, J. 2013b. Causal transportability with limited experiments. In desJardins, M., and Littman, M. L., eds., *Proceedings of The Twenty-Seventh Conference on Artificial Intelligence (AAAI 2013)*, 95–101.

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. Technical Report R-425, Cognitive Systems Laboratory, Department of Computer Science, UCLA.

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

Cornfield, J. 1951. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* 11:1269–1275.

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*, ALT '08, 38–53. Berlin, Heidelberg: Springer-Verlag.

Didelez, V.; Kreiner, S.; and Keiding, N. 2010. Graphical models for inference under outcome-dependent sampling. *Statistical Science* 25(3):368–387.

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.

Geng, Z. 1992. Collapsibility of relative risk in contingency tables with a response variable. *Journal Royal Statistical Society* 54(2):585–593.

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.

Greenland, S., and Pearl, J. 2011. Adjustments and their consequences – collapsibility analysis using graphical models. *International Statistical Review* 79(3):401–426.

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

Hein, M. 2009. Binary classification under sample selection bias. In Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N., eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press. 41–64.

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

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

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

Little, R. J. A., and Rubin, D. B. 1986. *Statistical Analysis with Missing Data*. New York, 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 Paz, A. 2013. Confounding equivalence in causal equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*, 433–441. Corvallis, OR: AUAI. Also: Technical Report R-343w, Cognitive Systems Laboratory, Department of Computer Science, UCLA.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. 1993. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, 391–401.

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.

Pearl, J. 2013. Lindear models: A useful "microscope" for causal analysis. *Journal of Causal Inference* 1:155–170.

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.

Smith, A. T., and Elkan, C. 2007. Making generative classifiers robust to selection bias. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, 657–666. New York, NY, USA: ACM.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press, 2nd edition.

Storkey, A. 2009. When training and test sets are different: characterising learning transfer. In Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N., eds., *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press. 3–28.

Textor, J., and Liskiewicz, M. 2011. Adjustment criteria in causal diagrams: An algorithmic perspective. In Pfeffer, A., and Cozman, F., eds., *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 681–688. AUAI Press.

Tian, J.; Paz, A.; and Pearl, J. 1998. Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles, CA.

Whittemore, A. 1978. Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, Series 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.

Zhang, J. 2008. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.* 172:1873–1896.