Causal Effect Identification by Adjustment under Confounding and Selection Biases

Juan D. Correa
Purdue University
correagr@purdue.edu

Elias Bareinboim
Purdue University
eb@purdue.edu

Abstract

Controlling for selection and confounding biases are two of the most challenging problems in the empirical sciences as well as in artificial intelligence tasks. Covariate adjustment (or, Backdoor Adjustment) is the most pervasive technique used for controlling confounding bias, but the same is oblivious to issues of sampling selection. In this paper, we introduce a generalized version of covariate adjustment that simultaneously controls for both confounding and selection biases. We first derive a sufficient and necessary condition for recovering causal effects using covariate adjustment from an observational distribution collected under preferential selection. We then relax this setting to consider cases when additional, unbiased measurements over a set of covariates are available for use (e.g., the age and gender distribution obtained from census data). Finally, we present a complete algorithm with polynomial delay to find all sets of admissible covariates for adjustment when confounding and selection biases are simultaneously present and unbiased data is available.

Introduction

One of the central challenges in data-driven fields is to compute the effect of interventions – for instance, how increasing the educational budget will affect violence rates in a city, whether treating patients with a certain drug will help their recovery, or how increasing the product price will change monthly sales? These questions are commonly referred as the problem of identification of causal effects. There are two types of *systematic bias* that pose obstacles to this kind of inference, namely *confounding bias* and *selection bias*. The former refers to the presence of a set of factors that affect both the action (also known as treatment) and the outcome (Pearl 1993), while the latter arises when the action, outcome, or other factors differentially affect the inclusion of subjects in the data sample (Bareinboim and Pearl 2016).

The goal of our analysis is to produce an unbiased estimand of the *causal effect*, specifically, the probability distribution of the outcome when an action is performed by an autonomous agent (e.g., FDA, robot), regardless of how the decision would naturally occur (Pearl 2000, Ch. 1). For example, consider the graph in Fig. 1(a) in which X represents a treatment (e.g., taking or not a drug), Y represents an outcome (health status), and Z is a factor (e.g., gender, age) that affects both the propensity of being treated and the outcome. The edges (Z, X) and (Z, Y) may encode the facts ”gender affects how the drug is being prescribed” and ”gender affects recovery” respectively – for example, females may be more health conscious, so they seek for treatment more frequently than their male counterparts and at the same time are less likely to develop large complications for the particular disease. Intuitively, the causal effect represents the variations of X that bring about change in Y *regardless* of the influence of Z on X, which is graphically represented in Fig. 1(b). Mutilation is the graphical operation of removing arrows representing a decision made by an autonomous agent of setting a variable to a certain value. The mathematical counterpart of mutilation is the $do()$ operator and the average causal effect of X on Y is usually written in terms of the *do*-distribution $P(y \mid do(x))$ (Pearl 2000, Ch. 1).

The gold standard for obtaining the *do*-distribution is through the use of randomization, where the treatment assignment is selected by a randomized device (e.g., a coin flip) regardless of any other set of covariates (Z). In fact, this operation physically transforms the reality of the underlying population (Fig. 1(a)) into the corresponding mutilated world (Fig. 1(b)). The effect of Z on X is neutralized once randomization is applied. Despite its effectiveness, randomized studies can be prohibitively expensive, and even unattainable in certain cases, either for technical, ethical, or technical reasons – e.g., one cannot randomize the cholesterol level of a patient and record if it causes the heart to stops, when trying to assess the effect of cholesterol level on cardiac failure.

An alternative way of computing causal effects is trying to relate non-experimentally collected samples (drawn from $P(z, x, y)$) with the experimental distribution ($P(y \mid do(x))$). Non-experimental (often called observational) data relates to the model in Fig. 1(a) where subjects decide by themselves to take or not the drug (X) while influenced by other factors (Z). There are a number of techniques developed for this task, where the most general one is known as *do-calculus* (Pearl 1995). In practice, one particular strategy from do-calculus called *adjustment* is used the most. It consists of averaging the effect of X on Y over the different levels of Z, isolating the effect of interest from the effect induced by other factors. Controlling for confounding bias by adjustment is currently the standard method for inferring causal effects in data-driven fields, and different properties and enhancements have been studied in statistics (Rubin 1974; Robinson and Jewell 1991; Pirinen, Donnelly, and Spencer 2012; Mefford and Witte 2012) and AI (Pearl 1993; 1995; Pearl and Paz 2010; Shpitser, VanderWeele, and Robins 2010; Maathuis and Colombo 2015; van der Zander, Liskiewicz, and Textor 2014).

Orthogonal to confounding, *sampling selection bias* is induced by preferential selection of units for the dataset, which is usually governed by unknown factors including treatment, outcome, and their consequences. It cannot be removed by a randomized trial and may stay undetected during the data gathering process, the whole study, or simply never be detected[1]. Consider Fig. 1(e) where X and Y represent again treatment and outcome, but S represents a binary variable that indicates if a subject is included in the pool ($S{=}1$ means that the unit is in the sample, $S{=}0$ otherwise). The effect of X on Y in the entire population ($P(y \mid do(x))$) is usually not the same as in the sample ($P(y \mid do(x), S{=}1)$). For instance, patients that went to the hospital and were sampled are perhaps more affluent and have better nutrition than the average person in the population, which can lead to a faster recovery. This preferential selection of samples challenges the validity of inferences in several tasks in AI (Cooper 1995; Cortes et al. 2008; Zadrozny 2004) and Statistics (Little and Rubin 1986; Kuroki and Cai 2006) as well as in the empirical sciences (Heckman 1979; Angrist 1997; Robins 2001).

The problem of selection bias can be addressed by removing the influence of the biased sampling mechanism on the outcome as if a random sample of the population was taken. For the graph in Fig. 1(d), for example, the distribution $P(y \mid do(x))$ is equal to $P(y \mid x, S{=}1)$ because there are not external factors that affect X and the selection mechanism S is independent of the outcome Y when the effect is estimated for the treatment X. There exists a complete non-parametric[2] solution for the problem of estimating statistical quantities from selection biased datasets (Bareinboim and Pearl 2012), and also sufficient and algorithmic conditions for recovering from selection in the context of causal inference (Bareinboim, Tian, and Pearl 2014; Bareinboim and Tian 2015).

Both confounding and selection biases carry extraneous "flow" of information between treatment and outcome, which is usually deemed "spurious correlation" since it does not correspond to the effect we want to compute on. Despite all the progress made in controlling these biases separately, we show that to estimate causal effects considering both problems requires a more refined analysis. First, note that the effect of X on Y can be estimated by blocking confounding and controlling for selection, respectively,

[1](Zhang 2008) noticed some interesting cases where detection is feasible in a class of non-chordal graphs.

[2]No assumptions about the about the functions that relates variables are made (i.e. linearity, monotonicity).

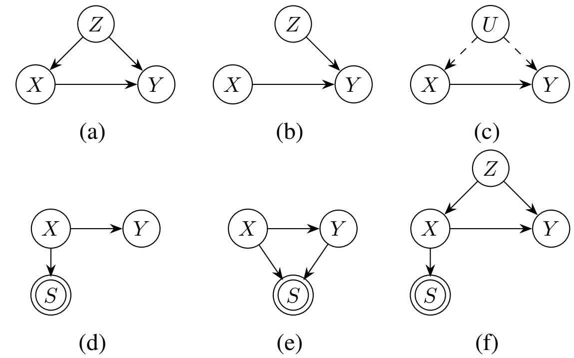


Figure 1: (a) and (d) give simple examples for confounding and selection bias respectively. (b) represents the model in (a) after an intervention is performed on X. (c) and (e) present examples where confounding and selection bias can not be removed respectively. In (f) we can control for either confounding or selection bias, but not for both unless we have external data on $P(z)$.

in Figs. 1(a) and (d). On the other hand, confounding cannot be removed in Fig. 1(c) nor it can be recovered from selection bias in Fig. 1(e). Perhaps surprisingly, Fig. 1(f) presents a scenario where either confounding or selection can be addressed separately ($P(y|do(x)) = \sum_Z P(y|x,z)P(z)$ and $P(z,y|do(x)) = P(z,y|do(x), S{=}1)$), but not simultaneously (without external data). As this example suggests, there is an intricate connection between these two biases that disallow the methods developed for these problems of being applied independently and then combined.

In this paper, we study the problem of estimating causal effects from models with an arbitrary structure that involve both biases. We establish necessary and sufficient conditions that a set of variables should fulfill so as to guarantee that the target effect can be unbiasedly estimated by adjustment. We consider two settings – first when only biased data is available, and then a more relaxed setting where additional unbiased samples of covariates are available for use (e.g., census data). Specifically, we solved the following problems:

1. **Identification and recoverability without external data**: The data is collected under selection bias, $P(\mathbf{v} \mid S{=}1)$, when does a set of covariates $\mathbf{Z}$ allow $P(\mathbf{y} \mid do(\mathbf{x}))$ to be estimated by adjusting for $\mathbf{Z}$?
2. **Identification and recoverability with external data**: The data is collected under selection bias $P(\mathbf{v} \mid S{=}1)$ and unbiased samples of $P(\mathbf{t}), \mathbf{T} \subseteq \mathbf{V}$, are available. When does a set of covariates $\mathbf{Z} \subseteq \mathbf{T}$ license the estimation of $P(\mathbf{y} \mid do(\mathbf{x}))$ by adjusting for $\mathbf{Z}$?
3. **Finding admissible adjustment sets with external data**: How can we list all admissible sets $\mathbf{Z}$ capable of identifying and recovering $P(\mathbf{y} \mid do(\mathbf{x}))$, for $\mathbf{Z} \subseteq \mathbf{T} \subseteq \mathbf{V}$?

Preliminaries

The systematic analysis of confounding and selection biases requires a formal language where the characterization of the underlying data-generating model can be encoded explicitly.

We use the language of Structural Causal Models (SCM) (Pearl 2000, pp. 204-207). Formally, a SCM M is a 4-tuple $\langle U, V, F, P(u)\rangle$, where U is a set of exogenous (latent) variables and V is a set of endogenous (measured) variables. F represents a collection of functions $F = \{f_i\}$ such that each endogenous variable $V_i \in V$ is determined by a function $f_i \in F$, where f_i is a mapping from the respective domain of $U_i \cup Pa_i$ to V_i, $U_i \subseteq U$, $Pa_i \subseteq V \backslash V_i$ (where Pa_i is the set of endogenous variables that are arguments of f_i), and the entire set F forms a mapping from U to V. The uncertainty is encoded through a probability distribution over the exogenous variables, $P(u)$. Within the structural semantics, performing an action $X{=}x$ is represented through the do-operator, $do(X{=}x)$, which encodes the operation of replacing the original equation of X by the constant x and induces a submodel M_x. For a detailed discussion on the properties of structural models, we refer readers to (Pearl 2000, Ch. 7).

We will represent sets of variables in bold. The causal effect of a set $\mathbf{X}$ when it is assigned a set of values $\mathbf{x}$, on a set $\mathbf{Y}$ when it is instantiated as $\mathbf{y}$ will be written as $P(\mathbf{y} \mid do(\mathbf{x}))$, which is a short hand notation for $P(\mathbf{Y}{=}\mathbf{y} \mid do(\mathbf{X}{=}\mathbf{x}))$. Mainly, we will operate with $P(\mathbf{v})$, $P(\mathbf{v} \mid do(\mathbf{x}))$, $P(\mathbf{v} \mid S{=}1)$, respectively, the observational, experimental, and selection-biased distributions.

Formally, the task of estimating a probabilistic quantity from a selection-biased distribution is known as *recovering* from selection bias (Bareinboim and Pearl 2012). It is not uncommon for observations of a subset of the variables over the entire population (unbiased data) to be available for use. Therefore, we will consider two subsets of $\mathbf{V}$, $\mathbf{M}, \mathbf{T} \subseteq \mathbf{V}$, where $\mathbf{M}$ contains the variables for which data was collected under selection bias, and $\mathbf{T}$ encompasses the variables observed in the overall population, without bias. The absence of unbiased data is equivalent to have $\mathbf{T} = \emptyset$.

Selection Bias with Adjustment

The main justification for the validity of adjustment for confounding comes under a graphical conditions called the "Backdoor criterion" (Pearl 1993; 2000), as shown below:

Definition 1 (Backdoor Criterion (Pearl 2000))**.** A set of variables $\mathbf{Z}$ satisfies the Backdoor Criterion relative to a pair of variables (X, Y) in a directed acyclic graph G if:

(i) No node in $\mathbf{Z}$ is a descendant of X.

(ii) $\mathbf{Z}$ blocks every path between X and Y that contains an arrow into X.

The heart of the criterion lies in cond. (ii), where the set $\mathbf{Z}$ is required to block all the backdoor paths between X and Y that generate confounding bias. Furthermore, cond. (i) forbids the inclusion of descendants of X in $\mathbf{Z}$, which intends to avoid opening new non-causal paths. For example, the empty set is admissible for adjustment in Fig. 1(e), but adding S would not be allowed since it is a descendant of X and opens the non-causal path $X \to S \leftarrow Y$. On the other hand, even though S does not open any non-causal path in Fig. 1(f), the criterion does not allow it to be used for adjustment.

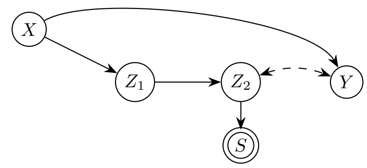


Figure 2: A graph that does not satisfy the s-backdoor criterion (respect to $\mathbf{Z}$), but the adjustment formula is recoverable and corresponds to desired causal effect.

(Bareinboim, Tian, and Pearl 2014) noticed that adjustment could be used for controlling for selection bias, in addition to confounding, which lead to a sufficient graphical condition called *Selection-Backdoor* criterion.

Definition 2 (Selection-Backdoor Criterion (Bareinboim and Tian 2015))**.** A set $\mathbf{Z} = \mathbf{Z}^+ \cup \mathbf{Z}^-$, with $\mathbf{Z}^- \subseteq De_X$ and $\mathbf{Z}^+ \subseteq \mathbf{V} \setminus De_X$ (where De_X is the set of variables that are descendants of X in G) satisfies the selection backdoor criterion (s-backdoor, for short) relative to X, Y and $\mathbf{M}, \mathbf{T}$ in a directed acyclic graph G if:

(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, $(\mathbf{Z}^- \perp\!\!\!\perp Y \mid X, \mathbf{Z}^+)$

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

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

The first two conditions echo the extended-backdoor (Pearl and Paz 2010)[3], while cond. (iii) and (iv) guarantee that the resultant expression is estimable from the available datasets. If the S-Backdoor criterion holds for $\mathbf{Z}$ relative to $\mathbf{X}, \mathbf{Y}$ and $\mathbf{M}, \mathbf{T}$ in G, then the effect $P(y \mid do(x))$ is identifiable, recoverable, and given by

$$P(y \mid do(x)) = \sum_{\mathbf{Z}} P(y \mid x, \mathbf{z}, S{=}1) P(\mathbf{z}) \tag{1}$$

We note here that the S-Backdoor is sufficient but not necessary for adjustment. To witness, consider the model in Fig. 2 where $\mathbf{Z} = \{Z_1, Z_2\}, \mathbf{M} = \{X, Y, Z_1, Z_2\}$, and $\mathbf{T} = \{Z_1, Z_2\}$. Here, $\mathbf{Z}^+ = \emptyset, \mathbf{Z}^- = \{Z_1, Z_2\}$. Condition (ii) in Def. 2 is violated, namely $(Z_1, Z_2 \not\perp\!\!\!\perp Y \mid X)$. Perhaps surprisingly, the effect $P(y \mid do(x))$ is identifiable and recoverable, as follows:

$$P(y|do(x)) = P(y|x) \tag{2}$$

$$= P(y|x) \sum\nolimits_{Z_1} P(z_1) \tag{3}$$

$$= \sum\nolimits_{Z_1} P(y|x, z_1) P(z_1) \tag{4}$$

$$= \sum\nolimits_{Z_1, Z_2} P(y|x, z_1, z_2) P(z_2|x, z_1) P(z_1) \tag{5}$$

$$= \sum\nolimits_{Z_1, Z_2} P(y|x, z_1, z_2) P(z_2|z_1) P(z_1) \tag{6}$$

$$= \sum\nolimits_{Z_1, Z_2} P(y|x, z_1, z_2, S{=}1) P(z_1, z_2) \tag{7}$$

[3]The extended-backdoor augments the backdoor criterion to allow for descendants of X that could be harmless in terms of bias.

(2) follows from the application of the second rule of do calculus and the independence $(X \perp\!\!\!\perp Y)_{G_{\underline{X}}}$. Equations (5),(6),(7) use the independences $(Y \perp\!\!\!\perp Z_1|X)$, $(Z_2 \perp\!\!\!\perp X|Z_1)$ and $(S \perp\!\!\!\perp Y|X, Z_1, Z_2)$ respectively. The final expression (7) is estimable from the available data.

Considering that $\mathbf{Z} = \emptyset$ controls for confounding, adjusting for $\mathbf{Z} = \{Z_1, Z_2\}$ seems unnecessary. As it turns out, covariates irrelevant for confounding control, could play a role when we compound this task with recovering from selection bias (where Y will need to be separated from S).

Generalized Adjustment without External Data

Let us consider the case when only biased data $P(\mathbf{v} \mid S{=}1)$ over $\mathbf{V}$ is measured. Our interest in this section is on conditions that allow $P(\mathbf{y} \mid do(\mathbf{x}))$ to be computed by adjustment without external measurements.

Consider the model G in Fig. 3(a). Note that Y and S are marginally independent in $G_{\overline{X}}$ (the graph after an intervention on X where all edges into X are not present). As for confounding, Z needs to be conditioned on, but doing so opens a path between Y and S, letting spurious correlation from the bias to be included in our calculation. It turns out that with a careful manipulation of the expression, both biases can be controlled as follows:

$$P(y \mid do(x)) = P(y \mid do(x), S{=}1) \quad (8)$$

$$= \sum_Z P(y \mid do(x), z, S{=}1) P(z \mid do(x), S{=}1) \quad (9)$$

$$= \sum_Z P(y \mid x, z, S{=}1) P(z \mid S{=}1) \quad (10)$$

Eq. (8) follows from the independence $(Y \perp\!\!\!\perp S \mid X)$ in the mutilated graph $G_{\overline{X}}$. Next we condition on Z and the (10) is valid by the application of the second rule of do-calculus to the first term and the third rule to the second in (9). Note that every term in (10) is estimable from the biased distribution.

Next we introduce a complete criterion to determine whether adjusting by a set of covariates is admissible to identify and recover the causal effect. Before that, we require the concept of *proper causal path*.

Definition 3 (Proper Causal Path (Shpitser, VanderWeele, and Robins 2010))**.** Let $\mathbf{X}, \mathbf{Y}$ be sets of nodes. A causal path from a node in $\mathbf{X}$ to a node in $\mathbf{Y}$ is called proper if it does not intersect $\mathbf{X}$ except at the end point.

Definition 4 (Generalized Adjustment Criterion Type 1)**.** A set $\mathbf{Z}$ satisfies the generalized criterion relative to $\mathbf{X}, \mathbf{Y}$ in a causal model with graph G augmented with the selection mechanism S if:

(a) No element in $\mathbf{Z}$ is a descendant in $G_{\overline{X}}$ of any $W \notin \mathbf{X}$ which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$.

(b) All non-causal $\mathbf{X}$-$\mathbf{Y}$ paths are blocked by $\mathbf{Z}$.

(c) $\mathbf{Y}$ is independent of S given $\mathbf{X}$ under intervention: $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X})_{G_{\overline{\mathbf{X}}}}$.

(d) $\mathbf{Z}$ can be partitioned in two sets $\mathbf{Z}^+, \mathbf{Z}^-$ such that $(\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}^- \mid \mathbf{X}, \mathbf{Z}^+, S)_{G_{\overline{\mathbf{X}}}}$, $\mathbf{Z}^- = \mathbf{Z} \setminus \mathbf{Z}^+$, and $\mathbf{Z}^+ = \{Z' \in \mathbf{Z} \mid (Z' \perp\!\!\!\perp \mathbf{X} \mid S)_{G_{\overline{X(S)}}}\}$.

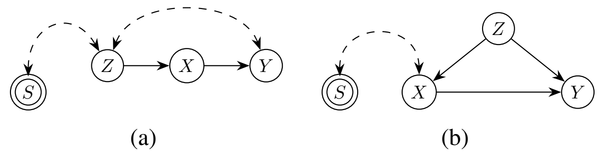


Figure 3: Models where $\mathbf{Z}$ satisfies Def. 4

$G_{\overline{X(S)}}$ is the graph where all edges into $X \in \mathbf{X} \setminus An_S)$ are removed, where An_S is the set of ancestors of the variable S in G.

Theorem 1 (Generalized Adjustment Formula Type 1)**.** *Given disjoint sets of variables $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$ in a causal model with graph G. The effect $P(\mathbf{y} \mid do(\mathbf{x}))$ is given by*

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum_{\mathbf{Z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, S{=}1) P(\mathbf{z} \mid S{=}1) \quad (11)$$

in every model inducing G if and only if they satisfy the generalized adjustment criterion type 1 (Def. 4).

Proof. Suppose $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ satisfy the criterion. Then we can decompose $\mathbf{Z}$ into $\mathbf{Z}^-$ and $\mathbf{Z}^+$. Conditions (a) and (b) imply $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z})_{G_{\underline{\mathbf{X}}}}$, but $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}, S)_{G_{\underline{\mathbf{X}}}}$ still holds, unless a path p which has S or an ancestor of it as a collider is opened in which case S and Y are d-connected in p. By condition (c), this is only possible if some $X \in \mathbf{X}$ is also a collider in p, which violate condition (b). We can obtain the causal effect as:

$$P(\mathbf{y} \mid do(\mathbf{x})) = P(\mathbf{y} \mid do(\mathbf{x}), S{=}1) \quad (12)$$

$$= \sum_{\mathbf{Z}^+} P(\mathbf{y}|do(\mathbf{x}), \mathbf{z}^+, S{=}1) P(\mathbf{z}^+|do(\mathbf{x}), S{=}1) \quad (13)$$

$$= \sum_{\mathbf{Z}^+} P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}^+, S{=}1) P(\mathbf{z}^+ \mid S{=}1) \quad (14)$$

$$= \sum_{\mathbf{Z}^+} P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}^+, S{=}1) \sum_{\mathbf{Z}^-} P(\mathbf{z}|S{=}1) \quad (15)$$

$$= \sum_{\mathbf{Z}} P(\mathbf{y} \mid do(\mathbf{x}), \mathbf{z}, S{=}1) P(\mathbf{z} \mid S{=}1) \quad (16)$$

$$= \sum_{\mathbf{Z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, S{=}1) P(\mathbf{z} \mid S{=}1) \quad (17)$$

Eq. (12) follows from cond. (c), conditioning on $\mathbf{Z}^+$ and applying the third rule of do-calculus using the definition of $\mathbf{Z}^+$ in cond. (d) we obtain (14). Conditioning the term $P(\mathbf{Z}^+)$ by $\mathbf{Z}^-$ and adding $\mathbf{Z}^-$ to the first term using condition (d) results in (16). Finally applying rule 2 of do-calculus using independence $(\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Z}, S)_{G_{\underline{\mathbf{X}}}}$ allows to remove do operator resulting in the adjustment formula in eq. (11). The necessity part of the proof is presented in the supplemental material. □

Conditions (a) and (b) echo the Extended Backdoor/Adjustment Criterion (Pearl and Paz 2010; Shpitser, VanderWeele, and Robins 2010) and guarantee that $\mathbf{Z}$ is admissible for adjustment in the unbiased distribution. Condition (c) requires the outcome $\mathbf{Y}$ to be independent of the selection mechanism S without observing any covariate $\mathbf{Z}$. Intuitively, condition (d) ensures that the influence of $\mathbf{Z}$ on S is insensitive to the intervention or than they are independent of the outcome. The model in Fig. 3(b) also satisfies Def. 4. Similarly to Fig. 3(a), if we control for confounding and try to remove the do-operator, it appears that selection bias cannot be removed since the independence $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}, \mathbf{Z})$ does not hold in G. Still, there exists a derivation strategy encapsulated in Def. 4 / Thm. 1 that allow one to recover from both selection and confounding biases.

Generalized Adjustment With External Data

A natural question that arises is whether additional measurements in the population level over the covariates can help identifying and recovering the desired causal effect. The following criterion tries to relax the previous results by leveraging the unbiased data available.

Definition 5 (Generalized Adjustment Criterion Type 2)**.** A set $\mathbf{Z}$ satisfies the generalized criterion relative to $\mathbf{X}, \mathbf{Y}$, a set of variables measured under selection bias $\mathbf{M}$ and a set of variables observed in the overall population $\mathbf{T}$ in a causal model with graph G augmented with the selection mechanism S if:

(a) No element in $\mathbf{Z}$ is a descendant in $G_{\overline{X}}$ of any $W \notin \mathbf{X}$ which lies on a proper causal path from $\mathbf{X}$ to $\mathbf{Y}$.

(b) All non-causal $\mathbf{X}$-$\mathbf{Y}$ paths in G are blocked by $\mathbf{Z}$.

(c) $\mathbf{Y}$ is independent of the selection mechanism S given $\mathbf{Z}$ and $\mathbf{X}$: $(\mathbf{Y} \perp\!\!\!\perp S \mid \mathbf{X}, \mathbf{Z})$

(d) The variables are measured with bias ($\mathbf{Z}, \mathbf{X}, \mathbf{Y} \subseteq \mathbf{M}$) and the covariates are available without bias ($\mathbf{Z} \subseteq \mathbf{T}$)

Theorem 2 (Generalized Adjustment Formula Type 2)**.** *Given disjoint sets of variables* $\mathbf{X}, \mathbf{Y}$ *and* $\mathbf{Z}$*, and sets* $\mathbf{M}, \mathbf{T}$ *in a causal model with graph G. In every model inducing G, the effect* $P(\mathbf{y} \mid do(\mathbf{x}))$ *is given by*

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum\nolimits_{\mathbf{Z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}, S{=}1) P(\mathbf{z}) \qquad (18)$$

if and only if they satisfy the criterion in Def. 5.

Proof. Suppose $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{M}, \mathbf{T}$ satisfy the criterion, by conditions (a) and (b), for every model induced by G we have:

$$P(\mathbf{y} \mid do(\mathbf{x})) = \sum\nolimits_{\mathbf{Z}} P(\mathbf{y} \mid \mathbf{x}, \mathbf{z}) P(\mathbf{z})$$

We note that S can be introduced to the first term by cond. (c), which entail Eq. (18). Cond. (d) ensures that both terms in the expression are estimable from the available distributions. The necessity part of the proof is more involved and is provided in the supplemental material (Correa and Bareinboim 2016). □

As in Def. 4, conditions (a) and (b) ensure $\mathbf{Z}$ is valid for adjustment without selection bias. Condition (c) requires that the influence of the selection mechanism in the outcome is nullified by conditioning on $\mathbf{X}$ and $\mathbf{Z}$ and condition (d) simply guarantees that the adjustment expression can be estimated from the available data. Fig. 4 presents two causal models that satisfies the previous criterion if measurements over $\mathbf{Z} = \{Z_1, Z_2, Z_3\}$ are available. To witness how the

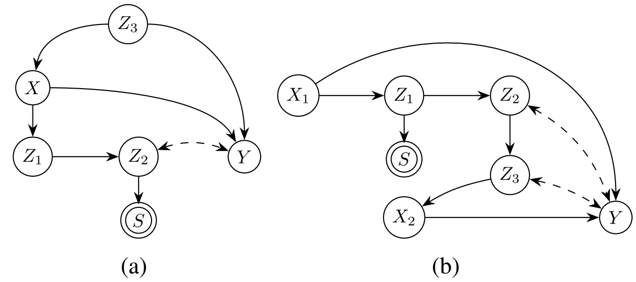


Figure 4: Models where the set $\mathbf{Z}$ satisfies Def. 5.

expression can be reached using do-calculus and probability axioms, consider Fig. 4(a):

$$P(y|do(x)) = \sum\nolimits_{Z_3} P(y|do(x), z_3) P(z_3|do(x)) \qquad (19)$$

$$= \sum\nolimits_{Z_3} P(y \mid x, z_3) P(z_3) \qquad (20)$$

$$= \sum\nolimits_{Z_1, Z_3} P(y \mid x, z_1, z_3) P(z_1, z_3) \qquad (21)$$

$$= \sum\nolimits_{\mathbf{Z}} P(y \mid x, \mathbf{z}) P(z_2 \mid x, z_1, z_3) P(z_1, z_3) \qquad (22)$$

$$= \sum\nolimits_{\mathbf{Z}} P(y \mid x, \mathbf{z}, S{=}1) P(\mathbf{z}) \qquad (23)$$

We start by conditioning on Z_3 and removing $do(x)$ using rule 3 of the do-calculus. Then conditioning the second term on Z_1, moving the summation to the left, and introducing Z_1 into the first term results in (21). Eq. (22) follows from conditioning the first term on Z_2, and finally by removing X in the second term using the independence $(Z_2 \perp\!\!\!\perp X | Z_1, Z_3)$, combining the last two distributions over the Z's and introducing the selection bias term using the independence $(Y \perp\!\!\!\perp S \mid X, \mathbf{Z})$ results in (23), which corresponds to (18).

Model in Fig. 4(b) also satisfies the type 2 criterion and illustrates how this can be applied to models where $\mathbf{X}$ and $\mathbf{Y}$ may be sets of variables.

Finding Admissible Sets for Generalized Adjustment

A natural extension to the problem is how to systematically list admissible sets for adjustment, using the criteria discussed in the previous sections. This is specially important in practice where factors such as feasibility, cost, and statistical power relate to the choosing of a covariate set.

In order to perform this kind of task efficiently, (van der Zander, Liskiewicz, and Textor 2014) introduced a transformation of the model called the *Proper Backdoor Graph* and formulate a criterion equivalent to the Adjustment Criterion:

Definition 6 (Proper Backdoor graph)**.** Let $G = (\mathbf{V}, \mathbf{E})$ be a DAG, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be pairwise disjoint subsets of variables. The proper backdoor graph, denoted as $G^{pbd}_{\mathbf{XY}}$, is obtained from G by removing the first edge of every proper causal path form $\mathbf{X}$ to $\mathbf{Y}$.

Definition 7 (Constructive Backdoor Criterion (CBD))**.** Let $G = (\mathbf{V}, \mathbf{E})$ be a DAG, and $\mathbf{X}, \mathbf{Y} \subseteq \mathbf{V}$ be pairwise disjoint subsets of variables. The set $\mathbf{Z}$ satisfies the Constructive Backdoor Criterion relative to $(\mathbf{X}, \mathbf{Y})$ in G if:

i) $\mathbf{Z} \subseteq \mathbf{V} \setminus Dpcp(\mathbf{X}, \mathbf{Y})$ and

ii) $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in the proper backdoor graph $G^{pbd}_{\mathbf{XY}}$.

Where $Dpcp(\mathbf{X}, \mathbf{Y}) = De((De_{\overline{\mathbf{X}}}(\mathbf{X}) \setminus \mathbf{X}) \cap An_{\underline{\mathbf{X}}}(Y))$

The set $Dpcp(\mathbf{X}, \mathbf{Y})$ is exactly the set of nodes forbidden by the first condition in both our generalized criteria, and $G^{pbd}_{\mathbf{XY}}$ only contain $\mathbf{X}, \mathbf{Y}$ paths that need to be blocked.

Lemma 3 (Constructive Backdoor $\implies$ Generalized Adjustment Type 2)**.** *Any set $\mathbf{Z}$ satisfying the CDB applied to $G^{pbd}_{(\mathbf{X} \cup S)\mathbf{Y}}$ and $Dpcp(\mathbf{X} \cup S, \mathbf{Y}) \cup (\mathbf{V} \setminus \mathbf{T})$ relative to $\mathbf{X}, \mathbf{Y}$ in G also satisfies the Generalized Adjustment Criterion type 2.*

Proof. By the equivalence between the CBD criterion and the adjustment criterion, we have that $Dpcp(\mathbf{X}, \mathbf{Y})$ is exactly the set of nodes forbidden by cond. (a) of the type 2 criterion, so

$$Dpcp(\mathbf{X} \cup S, \mathbf{Y}) = De((De_{\overline{\mathbf{X},\mathbf{S}}}(\mathbf{X} \cup \{S\}) \setminus (\mathbf{X} \cup S)) \cap An_{\underline{\mathbf{X},S}}(Y)) \tag{24}$$

Since S has no descendants, $De_{\overline{\mathbf{X},S}}(\mathbf{X} \cup \{S\}) = De_{\overline{\mathbf{X}}}(\mathbf{X}) \cup S$ and $An_{\underline{\mathbf{X},S}}(Y) = An_{\underline{\mathbf{X}}}(Y)$. As a consequence $Dpcp(\mathbf{X} \cup S, \mathbf{Y}) = Dpcp(\mathbf{X}, \mathbf{Y})$ implying cond. (a) of Def. 5.

$G^{pbd}_{(\mathbf{X} \cup S)\mathbf{Y}}$ has all non-causal paths from $\mathbf{X}$ to $\mathbf{Y}$ present in $G^{pbd}_{\mathbf{XY}}$, therefore, if $\mathbf{Z}$ block all non-causal paths in the former, it will do in the latter satisfying condition (b).

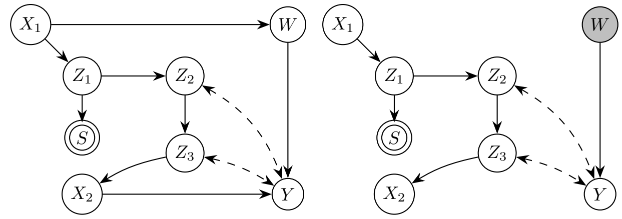
Every S – $\mathbf{Y}$ path may or may not contain $\mathbf{X}$. If not, $\mathbf{Z}$ should block it in $G^{pbd}_{(\mathbf{X} \cup S)\mathbf{Y}}$. In the latter case, the subpath from $\mathbf{X}$ to $\mathbf{Y}$ is either causal or non-causal. If it is causal $\mathbf{Z}$ will not block it, but the S-$\mathbf{Y}$ path will be blocked by $\mathbf{X}$. If the subpath is non-causal $\mathbf{Z}$ should block it, therefore, the larger path is blocked too. This argument implies condition (c). Since CBD holds for $Dpcp(\mathbf{X} \cup S, \mathbf{Y}) \cup (\mathbf{V} \setminus \mathbf{T})$ every element in $\mathbf{Z}$ must belong to $\mathbf{T}$ satisfying condition (d). ☐

Lemma 4 (Generalized Adjustment Type 2 $\implies$ Constructive Backdoor)**.** *Any set $\mathbf{Z}$ satisfying the Generalized Adjustment Criterion type 2 relative to $\mathbf{X}, \mathbf{Y}$ in G also satisfies the constructive backdoor criterion applied to $G^{pbd}_{(\mathbf{X} \cup S)\mathbf{Y}}$ and $Dpcp(\mathbf{X} \cup S, \mathbf{Y}) \cup (\mathbf{V} \setminus \mathbf{T})$.*

Proof. By lemma 3, $Dpcp(\mathbf{X} \cup S, \mathbf{Y}) = Dpcp(\mathbf{X}, \mathbf{Y})$, which combined with condition (d) implies condition (i) of the CBP.

By cond. (b) every non-causal path from $\mathbf{X}$ to $\mathbf{Y}$ is blocked by $\mathbf{Z}$ and all paths from S to $\mathbf{Y}$ (which are always non-causal when S is treated as an $\mathbf{X}$) are blocked by $\mathbf{Z}, \mathbf{X}$ by cond. (c). Those two facts together imply cond. (ii) of the CBD. ☐

Theorem 5 (Generalized Adjustment Type 2 ⇔ Constructive Backdoor)**.** *A set $\mathbf{Z}$ satisfies the Generalized Adjustment Criterion type 2 relative to $\mathbf{X}, \mathbf{Y}$ in G if and only if it satisfies the CBC applied to $G^{pbd}_{(\mathbf{X} \cup \mathbf{S})\mathbf{Y}}$ and $Dpcp(\mathbf{X} \cup S, \mathbf{Y}) \cup (\mathbf{V} \setminus \mathbf{T})$.*



(a) (b)

Figure 5: (a) shows a causal model and (b) the proper backdoor graph associated with it relative to $\mathbf{X} \cup S$ and Y. The gray nodes in (b) represents variables in *Dpcp*.

Proof. It follows immediately from lemmas 3,4. ☐

Thm. 5 allows us to use the LISTSEP procedure (van der Zander, Liskiewicz, and Textor 2014) to list all the valid sets for the generalized adjustment type 2. The algorithm guarantees $O(n(n+m))$ polynomial delay, where n is the number of nodes and m is the number of edges in G (see (Takata 2010)). That means that the time needed to output the first solution or indicate failure, and the time between the output of consecutive solutions, is $O(n(n+m))$.

To provide the reader an intuition of how the algorithm works, consider the graph in Fig. 5(a) and its associated constructive backdoor graph in (b). W is a “forbidden node” in the sense that it cannot be used for adjustment and for this example is the only element in $Dpcp(\mathbf{X}, Y)$ assuming that unbiased measurement on the covariates Z_1, Z_2 and Z_3 are available (i.e. $\{Z_1, Z_2, Z_3\} \subseteq \mathbf{T}$). The algorithm LISTSEP will output every set of variables that d-separates $X \cup S$ from Y in the proper backdoor graph that does not contain any node in $Dpcp(\mathbf{X}, Y)$.

Conclusions

We provide necessary and sufficient conditions for identification and recoverability from selection bias of causal effects by adjustment, applicable for data-generating models with latent variables and arbitrary structure in nonparametric settings. Def. 4 and Thm. 1 provide a complete characterization of identification and recoverability by adjustment when no external information is available. Def. 5 and Thm. 2 provide a complete graphical condition for when external information on a set of covariates is available. Thm. 5 allowed us to list all sets that satisfies the last criterion in polynomial-delay time, effectively helping in the decision of what covariates need to be measured for recoverability. This is especially important when measuring a variable is associated with a particular cost or effort. Despite the fact that adjustment is neither complete nor the only method to identify causal effects, it is in fact the most used tool in the empirical sciences. The methods developed in this paper should help to formalize and alleviate the problem of sampling selection and confounding biases in a broad range of data-intensive applications.

References

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 Lawrence, N., and Girolami, M., eds., *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*. La Palma, Canary Islands: JMLR. 100–108.

Bareinboim, E., and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences* 113:7345–7352.

Bareinboim, E., and Tian, J. 2015. Recovering causal effects from selection bias. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* 3475–3481.

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from selection bias in causal and statistical inference. In Brodley, C. E., and Stone, P., eds., *Proceedings of the Twenty-eighth AAAI Conference on Artificial Intelligence*, 2410–2416. Palo Alto, CA: AAAI Press. Best Paper Award.

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

Correa, J. D., and Bareinboim, E. 2016. Causal effect identification by adjustment under confounding and selection biases - supplemental material. Technical report, Purdue AI Lab, Department of Computer Science, Purdue University.

Cortes, C.; Mohri, M.; Riley, M.; and Rostamizadeh, A. 2008. Sample selection bias correction theory. In *International Conference on Algorithmic Learning Theory*, 38–53. Springer.

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

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.

Maathuis, M. H., and Colombo, D. 2015. A generalized back-door criterion. *Ann. Statist.* 43(3):1060–1088.

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

Pearl, J., and Paz, A. 2010. Confounding equivalence in causal equivalence. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*. Corvallis, OR: AUAI. 433–441.

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):669–688.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press. 2nd edition, 2009.

Pirinen, M.; Donnelly, P.; and Spencer, C. C. 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature genetics* 44(8):848–851.

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

Robinson, L. D., and Jewell, N. P. 1991. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique* 227–240.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688–701.

Shpitser, I.; VanderWeele, T. J.; and Robins, J. M. 2010. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of UAI 2010*, 527–536.

Takata, K. 2010. Space-optimal, backtracking algorithms to list the minimal vertex separators of a graph. *Discrete Applied Mathematics* 158(15):1660–1667.

van der Zander, B.; Liskiewicz, M.; and Textor, J. 2014. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of UAI 2014*, 907–916.

Zadrozny, B. 2004. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, 114. 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.