Learning Adjustment Sets from Observational and Limited Experimental Data

Sofia Triantafillou
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA

Gregory Cooper
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA

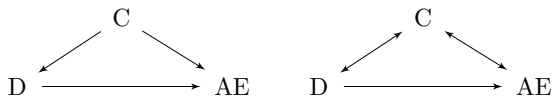
Abstract

Estimating causal effects from observational data is not always possible due to confounding. Identifying a set of appropriate covariates (adjustment set) and adjusting for their influence can remove confounding bias; however, such a set is typically not identifiable from observational data alone. Experimental data do not have confounding bias, but are typically limited in sample size and can therefore yield imprecise estimates. Furthermore, experimental data often include a limited set of covariates, and therefore provide limited insight into the causal structure of the underlying system. In this work we introduce a method that combines large observational and limited experimental data to identify adjustment sets and improve the estimation of causal effects. The method identifies an adjustment set (if possible) by calculating the marginal likelihood for the experimental data given observationally-derived prior probabilities of potential adjustmen sets. In this way, the method can make inferences that are not possible using only the conditional dependencies and independencies in all the observational and experimental data. We show that the method successfully identifies adjustment sets and improves causal effect estimation in simulated data, and it can sometimes make additional inferences when compared to state-of-the-art methods for combining experimental and observational data.

1 Introduction

Covariate adjustment is the main method for estimating causal effects from observational data. There is a lot of work on identifying the correct sets for covariate adjustment in the fields of potential outcomes and causal Bayesian networks. For the latter, sound and complete graphical criteria have recently been proven [van der Zander et al., 2014, Shpitser et al., 2012, Perkovic et al., 2017]. These criteria allow the identification of all the variable sets that lead to unbiased estimates of post-interventional probabilities through covariate adjustment when the causal graph is known. Typically,however, the true causal graph is unknown. Several causal discovery methods try to identify the causal graph for a set of variables based on the causal Markov and faithfulness assumption [Pearl, 2000]. Typically, multiple graphs fit the data equally well and are called Markov equivalent (ME). Thus, the correct sets for covariate adjustment are often not identifiable from observational data alone. In contrast, experimental data are the gold standard for estimating unbiased causal effects, but are often limited in terms of sample size and measured covariates. This situation can lead to wide confidence intervals of the estimated parameters, and provides limited insight into the causal structure of the system under study.

This paper introduces a method for identifying adjustment sets by combining large observational and limited experimental data. We are motivated by the following common scenario: Assume that a researcher is interested in quantifying the magnitude of an adverse effect (AE) of a drug (D) on a population and has access to a large collection of electronic health records of patients who take the drug or not, along with a large set of covariates. The researcher also has the published results of a randomized control trial (RCT) conducted in a random sample of the same population, which reports the estimated causal effect $\hat{P}(AE|do(D))$ and deems it significant; thus, D *causes* AE. The researcher suspects that this causal relationship is also confounded by another condition (C), not included in the RCT, that is highly correlated with both D and AE in the observational data (pairwisely and conditional on the remaining



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$

Figure 1: **Markov equivalent graphs imply different post-intervention distributions** In $\mathcal{G}_1$ C is an adjustment set for D, AE and $P(AE|do(D = d)) = \sum_c P(AE|d, c)$ for d =0, 1. In $\mathcal{G}_2$, $\emptyset$ is an adjustment set for X, Y and $P(AE|do(D = d)) = P(AE|d)$ for all d. Notice that even though the two graphs are indistinguishable based on the conditional (in) dependencies in observational data over $\{D, AE, C\}$ and experimental data over $\{D, AE\}$, they entail a different distribution for AE under $do(D)$.

variable). She wants to know if C is an adjustment set for AE and D, and if so, use C to improve the estimate of $P(AE|do(D))$ through covariate adjustment.

Figure 1 shows two different graphs that are both consistent with this scenario. The graphs cannot be distinguished based on the conditional dependencies and independencies in the observational and experimental data. However, the two graphs imply a different Interventional Distribution (ID) $P(AE|do(D))$. These two distributions can be computed from the observational data based on the appropriate covariate adjustment, which differs for the two graphs.

In this work, we present a method for using predictions of these ID estimates (from observational data) as priors in deriving the marginal likelihood of the experimental data in order to score adjustment sets and find the most probable one. By using the observational and experimental data in this way, our method can make inferences that are not possible based on conditional (in) dependence constraints alone, like identifying that C is an adjustment set for D, AE in Fig. 1, even though C is not measured in the experimental data. To our knowledge, this method is the first one described in the literature that can make this inference.

2 PRELIMINARIES

We use the framework of Semi-Markov Causal Models (SMCMs). The causal structure of SMCMs is described with mixed graphs. Mixed graphs include two types of edges: A directed edge, which denotes a direct causal relationship among the measured variables, and a bi-directed edge, which denotes the presence of a latent confounder. Each pair of variables is joined by at most two edges, one directed and one bidirected. No self-loops are allowed. Moreover, SMCMs are acyclic graphs. An SMCM with no bidirected edges is a directed acyclic graph (DAG).

We use the terms node and variable interchangeably. We use bold to denote variable sets, uppercase letters to denote single variables, and lowercase letters to denote variable values. We use standard graph-theoretic terminology. A mixed graph $\mathcal{G}$ is a triple $\langle \mathbf{V}, \mathbf{E}_D, \mathbf{E}_B \rangle$, where $\mathbf{V}$ is the set of nodes, and $\mathbf{E}_D, \mathbf{E}_B$ are the sets of directed and bidirected edges, respectively. A path between X and Y consists of an ordered sequence of distinct nodes $\langle X = V_1, \ldots, Y = V_n \rangle$ and a sequence of edges $\langle E_1, \ldots, E_{n-1} \rangle$ such that for $1 \leq i \leq n$, E_i is an edge between V_i and V_{i+1}. If every E_i is directed, the path is a directed or causal path. If there is a directed path from X to Y, then X is an ancestor of Y. We use the set $Pa_{\mathcal{G}}(X), An_{\mathcal{G}}(X), De_{\mathcal{G}}(X)$ to denote the set of parents, ancestors, and descendants of X in graph $\mathcal{G}$, respectively.

A SMCM graph is connected to the joint probability distribution (JPD) of the variables through the Causal Markov and the Causal Faithfulness Conditions [CMC, CFC Pearl, 2000]. CMC states that every variable is independent of its non-effects given its direct causes. The CMC entails a set of conditional independencies that hold in the joint probability distribution of variables in the graph. CFC states that only conditional independencies that stem from the CMC hold in the JPD. Under these two assumptions, all conditional independencies that hold in the JPD can be read off of the graph using the criterion of m-separation.

If we know the causal structure $\mathcal{G}$, a hard intervention of where a treatment X is set to x can be represented with the do-operator, $do(X = x)$. The post intervention distribution of an outcome Y given $do(X = x)$ is denoted $P(Y|do(X = x))$. In the corresponding SMCM, this is equivalent to removing all incoming edges into X, while keeping all other mechanisms intact. We use $\mathcal{G}_{\overline{X}}$ to denote the graph stemming from $\mathcal{G}$ after removing edges into X. We use $\mathcal{G}_{\underline{X}}$ to denote the graph stemming from $\mathcal{G}$ after removing edges out of X.

To estimate $P(Y|do(X))$ from observational data, we need to control for confounding bias and make sure we do not introduce additional bias (e.g., m-bias [Greenland, 2003]). This process is called **covariate adjustment**, which involves selecting a proper set of variables $\mathbf{Z}$ and "adjusting" for their effect to obtain the post-intervention distribution:

$$P(y|do(X = x)) = \sum_{\mathbf{z}} P(y|x, \mathbf{z}) P(\mathbf{z}) \quad \forall x \qquad (1)$$

Eq. 1 is called the **adjustment formula**, and set $\mathbf{Z}$ is an

adjustment set for X and Y. Pearl [2000] showed that sets of variables that satisfy a graphical criterion known as the backdoor criterion are valid adjustment sets. For DAGs, the set of parents (direct causes) of X or the set of all confounders of X and Y are examples of sets that satisfy the backdoor criterion. The criterion is sound, but not complete. Subsequent research proved that all valid adjustment sets satisfy a graphical criterion known as the **adjustment criterion** [Shpitser et al., 2012]:

Definition 1. $\mathbf{Z}$ *satisfies the adjustment criterion relative to* (X, Y) *in* $\mathcal{G}$ *if*

- *No element in* $\mathbf{Z}$ *is a descendant in* $\mathcal{G}$ *of any node that lies on a proper*[1] *causal path from* X *to* Y.
- *All non-causal paths in* $\mathcal{G}$ *from* X *to* Y *are blocked by* $\mathbf{Z}$.

The criterion is sound, meaning that any set $\mathbf{Z}$ that satisfies the adjustment criterion for (X, Y) is an adjustment set for (X, Y) in all distributions that induce $\mathcal{G}$. It is also complete, meaning that, for any $\mathbf{Z}$ that does not satisfy the adjustment criterion, there exists a distribution $\mathcal{P}$ inducing $\mathcal{G}$ where $\mathbf{Z}$ is not an adjustment set for (X, Y).

3 SCORING ADJUSTMENT SETS

The adjustment criterion allows us to identify all the adjustment sets for X and Y in an SMCM $\mathcal{G}$, if they exist. If we find an adjustment set, we can use it to estimate the post-intervention distribution $P(Y|do(X))$ from the pre-intervention distribution $P(\mathbf{V})$, for any treatment X and outcome Y. In this paper, we are interested in reverse engineering the adjustment sets for (X, Y) using the empirical observational JPD $\hat{P}(\mathbf{V})$ and an estimate of $P(Y|do(X))$ from limited experimental data.

Specifically, we assume that we have the following setting: There exists a SMCM over a set of variables $\mathbf{V}$ and a JPD $\mathcal{P}$ over the same variables such that $\mathcal{G}$ and $\mathcal{P}$ are faithful to each other. The variables include a treatment X and outcome Y caused by X.[2] We present our results for discrete variables following a multinomial distribution, although the results can be extended to other distributions in which marginal likelihoods can be computed in closed form exactly or approximated. We assume that we have the following data:

- Observational data D_{obs} measuring all variables in $\mathbf{V}$ on N samples.
- Experimental data D_{exp} that consist of an estimate $\hat{P}_{exp}(Y|do(X = c))$ over $N_{do(c)}$ for each possible value c of X.

Notice that no other variables in $\mathbf{V}$ are measured in D_{exp}; therefore the only (in) dependence statement that holds (asymptotically) in D_{exp} is the pairwise dependence of X and Y. Also notice that $\hat{P}_{exp}(Y|do(X = c)), c = 1, \ldots, C-1$ and N_{exp} are typically included and can be extracted from the publication that presents an RCT (without the raw data). We argue that this is a very common scenario in biology and medicine.

Intuitively, our method is based on the following observation: Different causal graphs, consistent with the conditional (in) dependence constraints in the data, may entail different adjustment sets for (X, Y), which in turn may lead to different predicted post-intervention distributions $\bar{P}(Y|do(X))$. In addition, there may be cases where no adjustment set exists among the set of observed variables, and therefore the observational data cannot be used to identify the post intervention distribution. By (implicitly) comparing $\bar{P}(Y|do(X))$ and $\hat{P}_{exp}(Y|do(X))$, we can identify sets that are more probable to be adjustment sets for (X, Y), and use them to improve the estimate for $P(Y|do(X))$.

To do so, we need something similar to faithfulness for the adjustment criterion. Specifically, we will assume that the adjustment sets for (X, Y) are exactly those for which the adjustment criterion holds. We call this assumption **adjustment faithfulness**:

Definition 2. *Let* $\mathcal{G}$ *be a causal SMCM and* $\mathcal{P}$ *a distribution faithful to* $\mathcal{G}$. *Then for all disjoint sets of variables* X, Y, $\mathbf{Z}$, $\mathbf{Z}$ *is an adjustment set for* (X, Y) *in* $\mathcal{P}$ *(according to Equation 1) only if* $\mathbf{Z}$ *satisfies the adjustment criterion for* (X, Y) *in* $\mathcal{G}$.

This assumption rules out distributions where $\mathbf{Z}$ is an adjustment set for (X, Y) without satisfying the adjustment criterion. In the rest of this paper, we use the following notation: Let $\mathcal{H}_{\mathbf{z}}$ be a binary variable that is true if $\mathbf{Z}$ is an adjustment set for (X, Y), and let $\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}$ denote that $\mathbf{Z}$ satisfies the adjustment criterion for (X, Y) in $\mathcal{G}$. If adjustment faithfulness holds, the two are equivalent: $\mathcal{H}_{\mathbf{z}}$ is true if and only if $\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}$.

We are interested in identifying the most likely adjustment set for X, Y. Unless otherwise mentioned, when we say that $\mathbf{Z}$ is an adjustment set, we mean it is so for X, Y. Thus, we want to find the set that maximizes the

[1] A causal path from $\mathbf{X}$ to $\mathbf{Y}$ is called proper if it does not intersect $\mathbf{X}$ except at the endpoint. The notation applies to sets of exposures. For singletons (which we consider in this work), it is identical to a directed path. However, we keep the term proper for consistency.

[2] The assumption that X is causally influencing Y is not necessary, but identifying adjustment sets for non cause-effect pairs is a simple task.

Algorithm 1: ScoreExp

input : $X, Y, \mathbf{Z}, D_{obs}, D_{do(c)}, nS$

output: $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}})$, $\hat{P}(Y|do(X))$

if $\mathbf{Z} == \nexists$ **then**

  $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\nexists}) = \prod_e \frac{\prod_i \Gamma(N^e_{do(c)})}{\Gamma(N_{do(c)}+E)}$;

  $\hat{P}(Y|do(X)) =$ NA;

else

  $\langle \mathcal{G}, f(\phi_{i|pa_i}|\mathcal{G}, D_{obs})\rangle \leftarrow$ `LearnBN`(D_{obs});

  foreach $iS = 1, \ldots, nS$ **do**

    Sample $\widetilde{\phi}_{i|pa_i} \sim f(\boldsymbol{\phi}_{i|pa_i}|\mathcal{G}, D_{obs})$;

    $\widetilde{\theta}_{y|c,k}, \widetilde{\theta}_k \leftarrow$ `BayesInf`$(\mathcal{G}, \widetilde{\phi}_{i|pa_i})$;

    $\widetilde{p}(iS) \leftarrow \sum_k \widetilde{\theta}_{y|c,k}\widetilde{\theta}_k$

    $\widetilde{p}Z(iS) = \prod_e (\sum_k \widetilde{\theta}_{y|c,k}\widetilde{\theta}_k)^{N^e_{do(c)}}$;

  $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}}) = \overline{pZ}$;

  $\hat{P}(Y|do(X)) = \overline{p}$;

the posterior

$$P(\mathcal{H}_{\mathbf{z}}|D_{exp}, D_{obs}) = \frac{P(D_{exp}, D_{obs}|\mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}})}{P(D_{exp}, D_{obs})} = \frac{P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs})}{P(D_{exp}|D_{obs})} \quad (2)$$

The score decomposes into (a) the probability of the experimental data given the observational data and given that $\mathbf{Z}$ is an adjustment set (or $\mathcal{H}_{\mathbf{z}}$ is true), (b) the probability of the observational data given that $\mathbf{Z}$ is an adjustment set, and (c) a normalizing constant $P(D_{exp}|D_{obs})$.

3.1 Estimating $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})$

D_{exp} includes data for each independent atomic intervention $P(Y|do(X = c))$, therefore $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})$ estimates $P(D_{exp}|D_{obs}, \mathcal{G})$ as

$$P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}}) = \prod_c P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}}) \quad (3)$$

For each c, we can derive $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}})$ on the basis of the adjustment formula: Under $\mathcal{H}_{\mathbf{z}}$, the adjustment formula connects the post-interventional to the observational distribution. Let $\boldsymbol{\theta}_{y|do(c)} = \{\theta_{e|do(c)}\}_{e=0}^{E-1}$ be set the parameters representing the probabilities $P(y{=}e|do(x{=}c))$ for a given value c of X. Clearly, $P(D_{do(c)}|\boldsymbol{\theta}_{y|do(c)}, D_{obs}, \mathcal{H}_{\mathbf{z}}) = P(D_{do(c)}|\boldsymbol{\theta}_{y|do(c)})$. Integrating over $\boldsymbol{\theta}_{y|do(c)}$, we have that

$$P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}}) = \int_{\boldsymbol{\theta}_{y|do(c)}} P(D_{do(c)}|\boldsymbol{\theta}_{y|do(c)}) f(\boldsymbol{\theta}_{y|do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}}) d\boldsymbol{\theta}_{y|do(c)} \quad (4)$$

$f(\boldsymbol{\theta}_{y|do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}})$ represents the prior of the parameters of the post-intervention distribution $P(Y|do(x{=}c))$ from the observational data, if $\mathbf{Z}$ satisfies the adjustment criterion. Let $\mathbf{Z}$ have k unique configurations, $k \in \{1, \ldots, K\}$. For simplicity, we use $\mathbf{z} = k$ to denote that $\mathbf{z}$ takes its k-th configuration. We use θ_k to denote the parameter for $P(\mathbf{z}{=}k)$.

Given space limitations, in the remainder of this section we will assume that y is a binary variable which has a Beta distribution. Extension to a multinomial distribution is straightforward and can be seen in Algorithm 1.

Let $\boldsymbol{\theta}_{0|c,k}$, be the k parameters for $P(y{=}0|x{=}c, \mathbf{z}{=}k)$. Under $\mathcal{H}_{\mathbf{z}}$, $P(y|do(x{=}c)){=}\sum_k P(y|x{=}c, \mathbf{z}{=}k)P(\mathbf{z}{=}k)$, or equivalently, $\theta_{y|do(c)} = \sum_k \theta_{0|c,k}\theta_k$ for $y = 0, 1$. Let $N^e_{do(c)}$ be the counts where $Y{=}e$ in $D_{do(c)}$ for $e{=}0, 1$. We can now recast Eq. 4 to include only observational parameters, as follows:

$$\int_{\boldsymbol{\theta}_{0|c,k}} \int_{\boldsymbol{\theta}_k} (\sum_k \theta_{0|c,k}\theta_k)^{N^0_{do(c)}} (\sum_k (1 - \theta_{0|c,k}\theta_k))^{N^1_{do(c)}} \times \prod_{k=1}^{K} f(\boldsymbol{\theta}_{0|c,k}, \theta_k|D_{obs}, \mathcal{H}_{\mathbf{z}}) d\boldsymbol{\theta}_{0|c,k} d\boldsymbol{\theta}_k \quad (5)$$

where we use the notation $\int_{\boldsymbol{\theta}_i}() d\boldsymbol{\theta}_i$ to denote multiple integration $\int_{\theta_1} \ldots \int_{\theta_I}() d\theta_1 \ldots d\theta_I$. Eq. 5 captures the proximity of the interventional distribution of Y in $D_{do(c)}$ to the interventional distribution we can estimate from D_{obs} using $\mathbf{Z}$ as an adjustment set for X, Y. $f(\boldsymbol{\theta}_{y|c,k}|D_{obs}, \mathcal{H}_{\mathbf{z}}) = f(\boldsymbol{\theta}_{y|c,k}|D_{obs})$ is the posterior density for the parameters $\boldsymbol{\theta}_{y|c,k}$ given D_{obs}. These are parameters in the observational distribution, and therefore independent of $\mathcal{H}_{\mathbf{z}}$ given D_{obs}.

While Eq. 5 cannot be computed in closed form, we can approximate it using a simple sampling procedure, presented in Algorithm 1. In this procedure, we sample from the posterior distribution of $f(\boldsymbol{\theta}_{y|c,k}, \boldsymbol{\theta}_k|D_{obs})$ and take the average over nS samples. To do so, we learn a Bayesian network graph $\mathcal{G}$ over the observed variables (in fact, we only need to learn a network over $X, Y, \mathbf{Z}$), and estimate the posterior of its parameters $f(\boldsymbol{\phi}_{i|pa_i}|\mathcal{G}, D_{obs})$. We then sample from this posterior distribution of the network, and use Bayesian inference to get a sample from the posterior $f(\boldsymbol{\theta}_{y|c,k}, \boldsymbol{\theta}_k|D_{obs})$. We use these parameters to derive a sample p for $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})$. We estimate $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{z}})$ as the average of these samples.

Under causal sufficiency, an adjustment set always exists. Under causal insufficiency, it is possible that no subset of $\mathbf{V}$ is an adjustment set. Thus, we also need

to consider the case where no adjustment set exists. We denote this hypothesis as $\mathcal{H}_{\not\exists}$ (note that this is different than $\mathcal{H}_{\emptyset}$, which states that the empty set is an adjustment set for (X, Y)). Under $\mathcal{H}_{\not\exists}$, we can not use the adjustment formula to connect D_{obs} to the post-intervention distribution, and thus

$$P(D_{do(c)}|D_{obs}, \mathcal{H}_{\not\exists}) = \int_{\boldsymbol{\theta}_{y_c}} P(D_{do(c)}|\boldsymbol{\theta}_{y_c}) f(\boldsymbol{\theta}_{y_c}) d\boldsymbol{\theta}_{y_c}. \tag{6}$$

We can compute $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\not\exists})$ in closed form as $\prod_e \frac{\prod_i \Gamma(N^e_{do(c)})}{\Gamma(N_{do(c)}+E)}$ for a uniform prior over the probability distributions. If $P(D_{do(c)}|D_{obs}, \mathcal{H}_{\not\exists}) > P(D_{do(c)}|D_{obs}, \mathcal{H}_{\mathbf{Z}})$, then $\mathbf{Z}$ does not give an estimate closer to the experimental data than using a weak uniform prior.

Notice that even if no adjustment set exists, it may still be possible to use the observational data to obtain an estimate of the post-intervention distribution $P(y|do(c))$ (e.g., through the front-door criterion). Our method only makes inferences that based on the back-door adjustment, and therefore Eq. 6 complements the space of possible hypotheses $\{\mathcal{H}_{\mathbf{z}}\}$ with respect to the backdoor criterion.

3.2 Estimating $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$

To estimate Eq. 2 we also need to estimate the probability that $\mathcal{H}_{\mathbf{z}}$ is true, based on the observational data: $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$. Under adjustment faithfulness, we can consider $P(\mathcal{H}_{\mathbf{z}}|D_{obs})$ in the space of possible SMCMs:

$$P(\mathcal{H}_{\mathbf{z}}|D_{obs}) = \sum_{\mathcal{G}} P(\mathcal{H}_{\mathbf{z}}|D_{obs}, \mathcal{G}) P(\mathcal{G}|D_{obs}). \tag{7}$$

For a given graph $\mathcal{G}$, $\mathcal{H}_{\mathbf{z}}$ is true if $\mathbf{Z}$ satisfies the adjustment criterion in $\mathcal{G}$ ($\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}$), and false otherwise. Thus, $P(\mathcal{H}_{\mathbf{z}}|D_{obs}, \mathcal{G}) = 1$ if $\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}$, and 0 otherwise. Eq. 7 becomes

$$P(\mathcal{H}_{\mathbf{z}}|D_{obs}) = \sum_{\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}} P(\mathcal{G}|D_{obs}). \tag{8}$$

Eq. 8 requires exhaustive enumeration of all possible graphs, and a method for obtaining the posterior probability of an SMCM given the data, both of which are complicated. Below we describe a method for approximating $\sum_{\mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}} P(\mathcal{G}|D_{obs})$.

First, we will assume that our observational sample size is large enough such that $P(\mathcal{G}|D_{obs}) = 1$ for a SMCM $\mathcal{G}$ if and only if $\mathcal{G}$ satisfies all the conditional dependence and independence constraints that hold in the data, according to the criterion of m-separation. Let $[\hat{\mathcal{G}}]$ denote this set of Markov equivalent graphs. Assuming all adjustments are equally likely, a priori, Eq. 8 is equal to the fraction of ME graphs in $[\hat{\mathcal{G}}]$ where $\mathbf{Z}$ satisfies the adjustment criterion:

$$P(\mathcal{H}_{\mathbf{z}}|D_{obs}) = \frac{|\{\mathcal{G} \in [\hat{\mathcal{G}}] : \mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}\}|}{|[\hat{\mathcal{G}}]|} \tag{9}$$

3.2.1 Identifying all Markov equivalent SMCMS with a logic-based approach

Identifying all the ME graphs D_{obs} can be done with a logic-based method for learning causal structure [Hyttinen et al., 2014, Triantafillou and Tsamardinos, 2015]. Logic-based methods encode path constraints imposed by the observed (in) dependence relationships in the data into a logic formula, expressed using the graph features (typically directed and bidirected edges) as the underlying boolean variables. Due to the exponential number of constraints, the scalability of logic-based approaches is limited. To face this problem, [Triantafillou and Tsamardinos, 2015] utilize the relationship of SMCMs with Maximal Ancestral Graphs [Richardson et al., 2002]. Markov equivalence classes of MAGs can be learned with FCI, a scalable algorithm that uses faithfulness to limit the number of possible independence tests. FCI is sound and complete in the sample limit [Zhang, 2008].

MAGs are ancestral mixed graphs, meaning that they contain no directed or almost directed cycles, where an almost directed cycle occurs if $X \leftrightarrow Y$ and X causes Y. Each pair of variables X, Y in an ancestral graph is joined by at most one edge. The orientation of this edge represents (non) causal ancestry: A bi-directed edge X↔Y denotes that X does not cause Y, and Y does not cause X, but the two share a latent confounder. $X \rightarrow Y$ denotes causal ancestry: X is a causal ancestor of Y. Thus, if X causes Y (not necessarily directly in the context of observed variables), and they are also confounded, there is an edge $X \rightarrow Y$ in the corresponding MAG. For an SMCM $\mathcal{G}$, there exists a unique MAG $\mathcal{M}$ over the same variables [Triantafillou and Tsamardinos, 2015]. In contrast, for a given MAG $\mathcal{M}$, there may exist more than one Markov equivalent SMCMs with the same ancestral relationships.

MAGs have the following attractive property: An m-separation (conditional independence for faithful distributions) between X and Y given a set $\mathbf{Z}$ corresponds to a missing edge between X and Y. Thus, all members of a Markov equivalence class of MAGs share the same skeleton, the same unshielded colliders, and the same discriminating colliders. This is not true for SMCMs. Specifically, it is possible that a pair of variables X, Y are not adjacent in an SMCM $\mathcal{G}$, even though they cannot

Condition (a) in Theorem 1:
$oncp(W) : -ancestor(x, W), ancestor(W, y).$
$decp(\mathbf{Z}) : -oncp(W), Q \in \mathbf{Z}, ancestor(W, Q).$
$conda(\mathbf{Z}) : -not\ decp(\mathbf{Z}).$
Condition (b) in Theorem 1:
$foncp(Z) : -directed(x, Z), ancestor(Z, y).$
$dirpbg(Q, Z) : -directed(Q, Z), notfoncp(Z).$
$condb(\mathbf{Z}) : -msepPG(x, y, \mathbf{Z}).$
Adjustment Sets:
$adjSet(\mathbf{Z}) : -conda(\mathbf{Z}), condb(\mathbf{Z}).$

Figure 2: ASP encoding for conditions (a) and (b).

be m-separated given any set of observed variables. As an example, consider an SMCM where $X \rightarrow Y \rightarrow Z$, and $Y \leftrightarrow Z$. X and Z are not m-separated given $\emptyset$ or Y, but they are not adjacent. The corresponding MAG is $X \rightarrow Y \rightarrow Z$, $X \rightarrow Z$. Edges in a MAG correspond to the existence of a primitive inducing path in any Markov equivalent SMCM. A primitive inducing path between X and Y is a path where every variable is a collider and an ancestor of either X or Y in $\mathcal{G}$ [Triantafillou and Tsamardinos, 2015].

Based on these results, we used the following approach to identify all Markov equivalent SMCMs that satisfy all conditional independence constraints in D_{obs}: We first use FCI to identify the Markov equivalence class of $[\hat{\mathcal{G}}]$. We then encoded the invariant features of $[\hat{\mathcal{G}}]$ in Answer Set Programming (ASP). Specifically, we used the representation presented in Zhalama et al. [2019] for the presence or absence of an inducing path in $\mathcal{G}$, and the inference rules in [Triantafillou and Tsamardinos, 2015] for definite non-colliders and discriminating colliders.

To obtain all possible SMCMs where $\mathcal{H}_{\mathbf{z}}$ holds, we encoded the adjustment criterion in ASP. To do so, we make use of some results presented in van der Zander et al. [2014] for constructing and testing adjustment sets in DAGs and MAGs. Specifically, to test if a set $\mathbf{Z}$ is an adjustment set relative to X, Y in a DAG $\mathcal{G}$, the authors construct the proper backdoor graph $\mathcal{G}^{pbd}$ by removing the first edge of each (proper) causal path from X to Y. We extend this operation to SMCMs. Then the following holds:

Theorem 1. *Let $\mathcal{G}$ be an SMCM, and let $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ be disjoint sets of variables. Then $\mathbf{Z}$ satisfies the adjustment criterion relative to ($\mathbf{X}$, $\mathbf{Y}$) if and only if:*

(a) No descendant of a node that lies on a proper causal path from X to Y is in $\mathbf{Z}$

(b) $\mathbf{Z}$ m-separates $\mathbf{X}$ and $\mathbf{Y}$ in the proper back-door graph $\mathcal{G}^{pbd}_{\mathbf{XY}}$.

Proof. The proof is identical to that of Theorem 4.4. in van der Zander et al. [2014], using m-separation in place of d-separation. □

Thus, graphs where $\mathbf{Z}$ is an adjustment set for X, Y are graphs where conditions (a) and (b) hold. We encoded these constraints in ASP, as shown in Figure 2. We use x, y to denote the treatment and outcome, respectively. These are given as inputs to the ASP, and we show them in lowercase to indicate that they are not variables. We also use the following predicates:

- $ancestor(\mathbf{W}, Q)$: $\mathbf{W}$ includes an ancestor of Q.
- $oncp(W)$: W is lies on a causal path from x to y.
- $decp(\mathbf{Z})$: $\mathbf{Z}$ includes a descendant of a node that lies on a causal path from x to y.
- $conda(\mathbf{Z})$: condition (a) holds for $\mathbf{Z}$.
- $foncp(Z)$: Z is the first node on a causal path from x to y.
- $dirpbg(Q, Z)$: edge $Q \rightarrow Z$ exists in $\mathcal{G}^{pbd}$.
- $msepPG(x, y, \mathbf{Z})$: x, y are m-separated given $\mathbf{Z}$ in $\mathcal{G}^{pbd}$. To make this inference, we encoded of the Bayes-Ball algorithm, using the rules in Borboudakis and Tsamardinos [2016], with $dirpbg$ in place of directed edges.
- $adjSet(\mathbf{Z})$: $\mathbf{Z}$ satisfies the adjustment criterion for x, y in a graph $\mathcal{G}$.

3.3 Finding Optimal Adjustment Sets

We can use Alg. 1 and Eq. 9 to compute $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs})$ for different candidate adjustment sets $\mathbf{Z}$ and identify the *most probable* adjustment set given D_{exp}, D_{obs}, i.e. the set $\mathbf{Z}^{\star} = argmax_{\mathbf{z}} P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs})$. Notice that the adjustment hypotheses are not necessarily mutually exclusive; multiple sets can be adjustment sets for (X, Y), and explain the observational data equally well; thus $argmax_{\mathbf{z}} P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}}|D_{obs})$ may have multiple optimal solutions.

Algorithm 2 describes the process of selecting an optimal adjustment set: The algorithm takes as input a set of observational data D_{obs} over variables $\mathbf{V}$ and a collection of experimental data D_{exp} that measure the Y under different manipulations $do(X)$. Additional parameters include the number of samples for the sampling approximation of $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})$, algorithm `FindPAG` for finding the ME class $[\hat{\mathcal{G}}]$ of graphs from D_{obs}, and algorithm `BayesInf` for performing Bayesian inference.

Algorithm 2: findOptimalAdjustmentSet

input : $Y, D_{obs}, D_{exp} = \{D_{do(c)}\}_{c=0}^{C-1}, nS,$ `FindPAG`
output: Adjustment set $\mathbf{Z}^\star$:
$[\hat{\mathcal{G}}] \leftarrow$ `FindPAG`(D_{obs});
PosAdjSetVars$\leftarrow$ Variables that lie on a path between X and Y in $[\hat{\mathcal{G}}]$;
foreach *subset* $\mathbf{Z}$ *of* PosAdjSetVars *and* $\notin$ **do**
  $P(\mathcal{H}_{\mathbf{z}} | D_{obs}) = \frac{|\{\mathcal{G} \in [\hat{\mathcal{G}}] : \mathcal{G} \vdash \mathcal{H}_{\mathbf{z}}\}|}{|[\hat{\mathcal{G}}]|}$;
  $P(D_{exp} | D_{obs}, \mathcal{H}_{\mathbf{z}}) \leftarrow$
   $\prod_{c=0}^{C-1}$ `scoreExp`$(X, Y, \mathbf{Z}, D_{obs}, D_{do(c)}, nS)$;
$\mathbf{Z}^\star \leftarrow argmax_{\mathbf{z}} P(D_{exp} | D_{obs}, \mathcal{H}_{\mathbf{z}}) P(\mathcal{H}_{\mathbf{z}} | D_{obs})$;

The algorithm initially learns $[\hat{\mathcal{G}}]$ and forms PosAdjSetVars, a superset of a possible adjustment set, that consists of all variables that lie on a path from X to Y. It is straightforward to show that if an adjustment set exists, it is a subset of PosAdjSetVars: If any open non-causal path exists between X and Y, they can only be blocked (if blockable) by variables on these paths. Conditioning on other variables can only violate conditions (a) or (b) in Definition 1, therefore if the paths can be blocked, they can be blocked by conditioning on a subset of PosAdjSetVars. Thus, the algorithm is asymptotically guaranteed to compute the score of at least one true adjustment set, if one exists, under our assumptions. Subsequently, the algorithm obtains $P(D_{exp}|D_{obs}, \mathcal{H}_{\mathbf{z}})P(\mathcal{H}_{\mathbf{z}} | D_{obs})$ for all subsets of PosAdjSets, as well as $\mathcal{H}_{\notin}$, and returns the optimal adjustment set, or NA if $\mathcal{H}_{\notin}$ has the highest score.

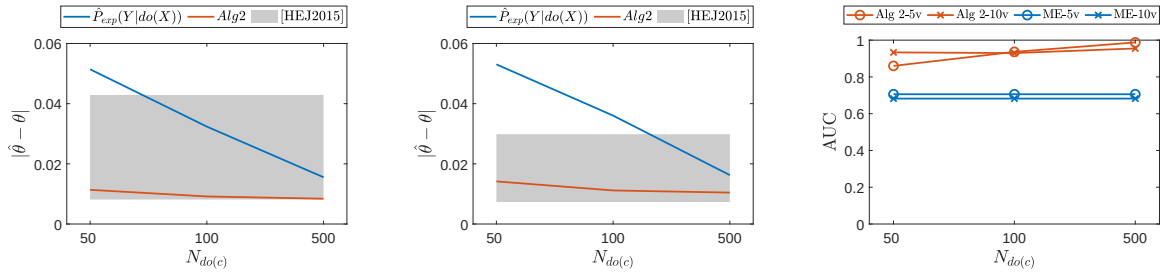
The complexity of the algorithm is exponential in the number of variables, since `FindPAG`, `BayesInf` and enumerating all possible SMCMs are NP-hard problems. Notice, however, that `BayesInf` and enumerating all possible SMCMs can be computed based on marginal graphical models, including only X, Y, and $\mathbf{Z}$ for every $\mathbf{Z}$. In practice, the enumeration of SMCMs will become infeasible even for small numbers of variables. One solution to this problem could be to use advances in ASP to sample uniformly from the space of all solutions [Smith and Mateas, 2011]. However, this is not supported in our current implementation.

4 RELATED WORK

In this work, we assume that we have a large observational data set D_{obs}, and a more limited experimental data set D_{exp} that measures Y under the randomization of X. From D_{obs} we can estimate $P(\mathbf{V})$ (or assess the set conditional (in) dependencies over $\mathbf{V}$), and from D_{exp} we get an estimate the marginal interventional distribution $P(Y|do(X))$ (or assess the dependence of X and Y in the interventional distribution). Given this information, several methods exist for selecting adjustment sets from D_{obs} alone, but require some additional prior knowledge on the relationship of the covariates to the treatment and outcome [VanderWeele and Shpitser, 2011, Entner et al., 2013]. Other methods like IDA and extensions [Nandy et al., 2017, Malinsky and Spirtes, 2017] compute bounds on causal effects from D_{obs} alone by computing causal effects in all graphs that are Markov equivalent to the highest-scoring graph. These methods do not output the most likely adjustment set. Finally, graphical methods for identifying all adjustment sets in a Markov equivalence class of graphs exist [Perkovic et al., 2017, Jaber et al., 2019] also exist. These methods return an adjustment set only if it is identifiable in the Markov equivalence class of graphs that are consistent with the observational data. When experimental data are also available, several methods try to learn the causal structure or estimate causal effects based on all available data. In the context of potential outcomes, Kallus et al. [2018] propose a method improving conditional treatment effect estimates by combining observational and experimental data. The method requires some overlap of covariates between observational and experimental data, a binary treatment and continuous covariates and outcome. For continuous data and linear relationships, observational data and limited experimental data can be combined to learn linear cyclic models [Eberhardt et al., 2010]. Algorithms for learning causal graphs from multiple experiments with discrete data exist [Cooper and Yoo, 1999, Hauser and Bühlmann, 2012], but require experimental data on all the observed variables, and are therefore not suitable for limited experimental data that only measure treatment and effect.

In the area of constraint-based causal discovery, logic-based approaches can use the conditional independences in D_{obs} and D_{exp} to learn causal graphs that are consistent with all the corresponding m-separation and m-connection constraints [Hyttinen et al., 2013, Triantafillou and Tsamardinos, 2015]. [Hyttinen et al., 2015, henceforth HEJ2015], use the resulting causal graphs to identify bounds in causal effects. This method can return can also identify marginal interventional distributions that are not identifiable with the adjustment criterion (e.g., distributions that are identifiable with the front-door criterion). However, the method will output a single estimate for $P(Y|do(X))$ only if *all the graphs that are consistent with the m-separations and m-connections in D_{obs} and D_{exp} imply the same estimate.*

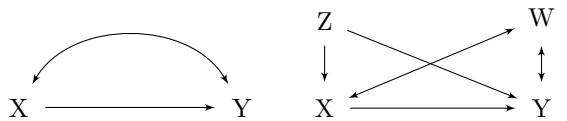
As an example, graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ in Fig. 1 are consistent with the same path constraints (m-connections and m-separations) in D_{exp} and D_{obs}, but imply different



(a) 5 variables (b) 10 variables (c) AUCs, 5 and 10 variables

Figure 3: (a, b) **Alg. 2 improves causal effect estimates.** Difference of true interventional distribution from the corresponding predicted distribution using (i) the frequentist estimate in D_{exp} (blue), (ii) Alg. 2 (orange), and the (iii) range of estimates for [HEJ2015] (gray). Our method improves the estimate of the ID for all sample sizes, when the causal effect is not identifiable from (in) dependence constraints in D_{obs} and D_{exp}. (c) **Alg. 2 correctly identifies adjustment sets.** Areas under the curve for predicting if $\mathbf{Z}$ is an adjustment set, using Alg. 2, and using the probability $\mathcal{H}_{\mathbf{z}}$ holds in the ME class of graphs.

$P(AE|do(D))$. Moreover, $\mathcal{G}_1$ with an additional bidirected edge $D \leftrightarrow AE$ is also consistent with the same path constraints, but $P(AE|do(D))$ is not identifiable. In that case, [HEJ2015] would return all possible quantities: $\{P(AE|D), \sum_c P(AE|D,c)P(c), NA\}$ as possible estimates for $P(AE|do(D))$. In contrast, our method generates a higher score for the estimate that is closer to the sample estimate $\hat{P}_{exp}(AE|do(D))$, and uses this score to select the most likely adjustment set. Thus, if the causal effect is identifiable from conditional (in) dependence constraints alone, our method and the [HEJ2015] method will yield the same results (asymptotically). However, our method can be used to improve the causal effect estimate of D_{exp} and provide insight on the causal structure among $\mathbf{V}$ even if the causal effect is not identifiable by independence constraints.



(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$

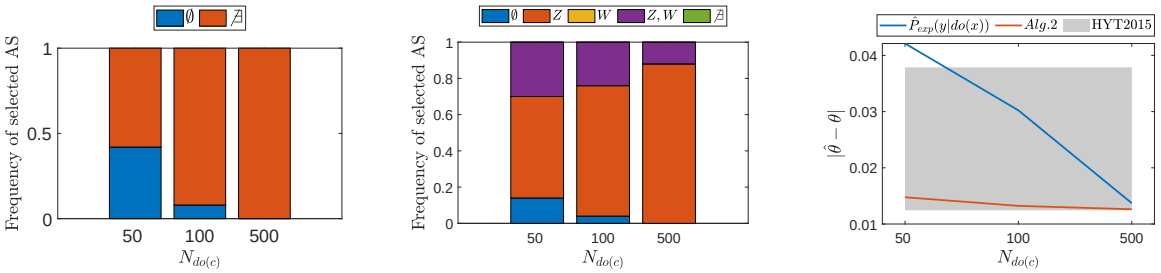
Figure 4: **SMCMs with bidirected edges used to simulate data.** In $\mathcal{G}_1$ there is no observed adjustment set for X, Y. In $\mathcal{G}_2$, Z is the only adjustment set for X, Y. In both cases, the (absence of an) adjustment set is not identifiable in the ME class defined by D_{obs} and D_{exp}.

5 EXPERIMENTS

We evaluated our methods using simulated and real data, and compared it to [HEJ2015]. Secificaally, We implemented a brute-force algorithm that identifies bounds in causal effects by running the ID algorithm [Shpitser and Pearl, 2006] on all Markov equivalent graphs (This version is also presented in [HEJ2015]). The same effects are identifiable by both algorithms. This version requires enumeration of the possible graphs over all variables, so to keep the problem tractable, we assume causal sufficiency in this part of the experiments.

We generated random DAGs with 5 or 10 variables, a mean in-degree of 3, and an ancestral relationship X to Y. We simulated discrete variables with 3 to 5 categories each and conditional probability tables $P(x|Pa(x))$ sampled from a Dirichlet distribution with priors sampled uniformly at random in [0, 1]. We used PC with an oracle to obtain the observational ME class $[\hat{\mathcal{G}}]$. For Alg. 1 in [HEJ2015], we used the subset $[\hat{\mathcal{G}}_s]$ of $[\hat{\mathcal{G}}]$ that also satisfies dependence in D_{exp}. Thus, $[\hat{\mathcal{G}}_s]$ consists of all members $\hat{G}$ of $[\hat{\mathcal{G}}]$ where X and Y are d-connected in the manipulated DAG $\mathcal{G}_{\overline{X}}$.

For each DAG, we simulated observational data D_{obs} with $N = 10,000$ samples and experimental data $D_{exp} = \{D_{do(c)}\}$, where we measured $N_{do(c)} = 50, 100$, and 500 samples of Y under $do(X = c)$ for each c. In all experiments, we used $nS = 500$ sampling iterations, and we used exact inference on a Bayesian network learned with FGES on D_{obs} as `BayesInf`. We used Alg.2 and [HEJ2015] to make a prediction (or multiple, in the case of [HEJ2015]) $P(Y|do(X))$. We estimated the absolute difference of the predicted vs the true interventional distribution, $|\hat{\theta}_{y|do(x)} - \theta_{y|do(x)}|$ averaged over all parameters $\theta_{y|do(x)}$. Thus, this is the average error each method makes per distribution parameter (denoted as $|\hat{\theta} - \theta|$ on the y-axes of the figures). For [HEJ2015], we present a range between the minimum and maximum error among all estimates.



(a) Data from $\mathcal{G}_1$ in Fig. 4 (b) Data from $\mathcal{G}_2$ in Fig. 4 (c) Data from $\mathcal{G}_2$ in Fig. 4

Figure 5: **Performance of Alg. 2 in causally insufficient systems.** (a, b) Fraction of times where Alg. 2 returns each possible adjustment set for data simulated from the SMCMs in Fig. 4. For $\mathcal{G}_1$, our method correctly identifies that there is no observed adjustment set the majority of times for all sample sizes. In data from $\mathcal{G}_2$, our method identifies that we must control for Z, but sometimes includes W in $\mathbf{Z}^*$, when the bias from adjusting for a collider is small. (c). Difference of the predicted vs true distribution using (i) the frequentist estimate in D_{exp}, (ii) Alg. 2, and the (iii) range of estimates from [HEJ2015]. In all cases, our method improves the estimation of $P(Y|do(X))$.

Fig. 3 shows that Alg. 2 allows us make inferences that are not possible using independence constraints. Fig. 3(a) shows $|\hat{\theta} - \theta|$ for Alg. 2 (orange line), the empirical estimate in $\hat{P}_{exp}(Y|do(X))$ in D_{exp} (blue line), and the range of estimates returned by [HEJ2015] (grey area), in 50 networks where $P(Y|do(X))$ is not identifiable in $[\hat{\mathcal{G}}_s]$(for identifiable causal effects, the behavior of our algorithm is the same, while [HEJ2015] will output single estimate, close to the lower bound of the grey area). We also show the same metric for the empirical distribution estimated in the experimental data, $P_{exp}(Y|do(X))$. Alg 2 improves the estimation of $P(Y|do(X))$ in all cases.

In addition, Alg. 2 correctly identifies adjustment sets. In the same networks, we used scores obtained by Alg. 2 to rank candidate adjustment sets, and computed the corresponding AUCs (Fig. 3c, orange line). As a baseline, we computed AUCs using the probability that $\mathbf{Z}$ is an adjustment set in the ME class $[\hat{\mathcal{G}}_s]$, computed as the number of graphs in which $\mathcal{H}_{\mathbf{z}}$ is true divided by the size of the ME class (Fig. 3c, blue line). Our method succesfully identifies adjustment sets that are not identifiable by (in) dependence constraints.

Alg. 2 can also identify that there is no adjustment set: We simulated data from the causally insufficient graph $\mathcal{G}_1$ in Fig. 4. $\mathcal{G}_1$ includes a latent confounder for X, Y, and therefore no adjustment set exists. Fig 5 (a) shows the percentage of times our algorithm selects $\mathcal{H}_{\not\exists}$ (orange bar) and $\mathcal{H}_{\emptyset}$ (blue bar). Our algorithm succesfully predicts that there is no adjustment set for $\mathcal{G}_1$, particularly for larger experimental sample sizes.

Finally, Alg. 2 can identify that the presence of m-bias. To show this, we simulated data from $\mathcal{G}_2$ in Fig. 4. Fig 5b shows the shows the percentage of times our algorithm selects each possible adjustment set. The correct adjustment set is selected the majority of times (orange bar), particularly for larger sample sizes. However, our method often includes the collider in the optimal adjustment set (purple bar). This happens because the effect of m-bias (i.e., bias in the causal effect estimation when conditioning on a collider) is often very small. Similar results have been reported in the literature [Greenland, 2003]. This is also seen in Fig. 5c, where we show the distribution distance for Alg. 2 (orange), D_{exp} (blue) and the ranges returned by [HEJ2015] (grey area). Our method improves causal effect estimation for all settings.

6 DISCUSSION

We present a method for learning adjustment sets and improving the estimation of causal effects by combining large observational and limited experimental data (e.g., combining electronic health records and RCTs), a scenario that is very common. Our method currently cannot handle selection, which can be a source of bias, particularly in RCTs. In addition, the scalability of the method is currently limited, particularly since it relies in the exhaustive enumeration of possible causal structures. Directions for future work include extensions of the method that address these shortcomings. Nevertheless, the method currently can make inferences accurately that are not possible with other state-of-the-art algorithms.

References

Giorgos Borboudakis and Ioannis Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 14351444, 2016.

Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 116–125. Morgan Kaufmann Publishers Inc., 1999.

Frederick Eberhardt, Patrik O Hoyer, Richard Scheines, et al. Combining experiments to discover linear cyclic models with latent variables. *Journal of Machine Learning Research*, 2010.

Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 256–264, 2013.

Sander Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt, and Matti Jarvisalo. Discovering cyclic causal models with latent variables: A general sat-based procedure. *arXiv preprint arXiv:1309.6836*, 2013.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *UAI*, pages 340–349, 2014.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. In *UAI*, pages 395–404. Citeseer, 2015.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under markov equivalence: Completeness results. In *International Conference on Machine Learning*, pages 2981–2989, 2019.

Nathan Kallus, Aahlad Manas Puli, and Uri Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems*, pages 10888–10897, 2018.

Daniel Malinsky and Peter Spirtes. Estimating bounds on causal effects in high-dimensional and possibly confounded systems. *International Journal of Approximate Reasoning*, 88:371–384, 2017.

Preetam Nandy, Marloes H Maathuis, Thomas S Richardson, et al. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2): 647–674, 2017.

J Pearl. *Causality: Models, Reasoning and Inference*, volume 113 of *Hardcover*. Cambridge University Press, 2000.

Emilija Perkovic, Johannes Textor, Markus Kalisch, and Marloes H Maathuis. Complete graphical characterization and construction of adjustment sets in markov equivalence classes of ancestral graphs. *The Journal of Machine Learning Research*, 18(1):8132–8193, 2017.

Thomas Richardson, Peter Spirtes, et al. Ancestral graph markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.

I Shpitser and J Pearl. Identification of conditional interventional distributions. In *Uncertainty in Artificial Intelligence*, 2006.

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*, 2012.

Adam M Smith and Michael Mateas. Answer set programming for procedural content generation: A design space approach. *IEEE Transactions on Computational Intelligence and AI in Games*, 3(3):187–200, 2011.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

Benito van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *CI@ UAI*, pages 11–24, 2014.

Tyler J VanderWeele and Ilya Shpitser. A new criterion for confounder selection. *Biometrics*, 67(4):1406–1413, 2011.

Zhalama, Jiji Zhang, Frederick Eberhardt, Wolfgang Mayer, and Mark Junjie Li. Asp-based discovery of semi-markovian causal models under weaker assumptions. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 1488–1494. ijcai.org, 2019. doi: 10.24963/ijcai.2019/206. URL `https://doi.org/10.24963/ijcai.2019/206`.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008.