Causal Markov Boundaries

Sofia Triantafillou[1] **Fattaneh Jabbari**[1] **Gregory F. Cooper**[1]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Abstract

Feature selection is an important problem in machine learning, which aims to select variables that lead to an optimal predictive model. In this paper, we focus on feature selection for post-intervention outcome prediction from pre-intervention variables. We are motivated by healthcare settings, where the goal is often to select the treatment that will maximize a specific patient's outcome; however, we often do not have sufficient randomized controlled trial data to identify well the conditional treatment effect. We show how we can use observational data to improve feature selection and effect estimation in two cases: (a) using observational data when we know the causal graph, and (b) when we do not know the causal graph but have observational and limited experimental data. Our paper extends the notion of Markov boundary to treatment-outcome pairs. We provide theoretical guarantees for the methods we introduce. In simulated data, we show that combining observational and experimental data improves feature selection and effect estimation.

1 INTRODUCTION

Feature selection is a fundamental problem in machine learning that aims to select the minimal set of features that lead to the optimal prediction of a target variable Y. For observational distributions, this set is the Markov boundary of Y, $\text{MB}(Y)$. In causal graphical models, this set can be identified from the causal graph $\mathcal{G}$ [Pearl, 2000]. This set exhausts the predictive information for the state of a variable Y, and can be used to obtain the best (and minimal) predictive model $P(Y|\text{MB}(Y))$ for Y.

In decision making, we are often interested in finding the optimal predictive model for the post-intervention distribution of an outcome Y after we intervene on a treatment X, when we only have observational data. Ideally, we would like to include in our model the Markov boundary $\mathbf{Z}$ of Y in the post-intervention causal graph that is parameterized with the post-intervention distribution. However, under causal insufficiency in which latent confounding may exist, the conditional post-interventional distribution $P(Y|do(X), \mathbf{Z})$ may not be identifiable. For example, in Fig. 1, $P(Y|do(X), A, B)$ is not identifiable from the observational distribution alone. In this case, we are interested in identifying the optimal set $\mathbf{Z}$ for which the post-intervention distribution $P(Y|do(X), \mathbf{Z})$ is identifiable from observational data, which we call the *causal Markov boundary*.

Moreover, even when experimental data are available, they typically have much smaller sample sizes and are not powered to identify conditional distributions. In that case, we would like to combine large observational data with limited experimental data to improve interventional feature selection and effect estimation.

Our methods are heavily motivated by embedded clinical trials [Angus, 2015, Angus et al., 2020], which take place within usual clinical care. In these trials, patients who agree to participate are randomized to receive a treatment from among those considered effective for that patient. The electronic health records (EHRs) of the health system in which the trial is being conducted contains both experimental data from the trial, and observational data obtained outside (e.g., before/after) the trial, all measuring the same variables. Combining observational and experimental data has the potential to better predict the most effective treatments for individual patients, than either type of data alone.

Our contributions are the following:

- We define the interventional Markov boundary $\text{MB}_X(Y)$ and the causal Markov boundaries $\mathbf{CMB}_X(Y)$ for an outcome Y and a treatment X. These sets correspond to the minimal set of covariates $\mathbf{Z}$ that are maximally informative for $Y|do(X)$, from experimental and observational data, respectively (Sec. 3). Table 1 summarizes

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

Table 1: Different Markov boundaries discussed in this paper.

Observational Markov Boundary (OMB) of Y: MB(Y)	The Markov boundary of Y. Leads to optimal prediction of Y from observational data.
Interventional Markov Boundary (IMB) of Y relative to X: MB$_X(Y)$	The Markov boundary of Y in the post-intervention distribution P_x. Leads to optimal prediction of Y_x from experimental data.
Causal Markov Boundaries (CMB) of Y relative to X: **CMB**$_X(Y)$	Sets of measured variables that satisfy Definition 3.2. Possibly not unique, and possibly empty. If not empty, one of the CMBs leads to the optimal prediction of Y_x from observational data.

the types of Markov boundaries discussed in this paper.

- We present a Bayesian method that combines observational and experimental data to learn interventional Markov boundaries. The method provides estimates of the post-interventional distribution that are based on both observational and experimental data, when possible (Sec. 4), in which case the IMB is a CMB. In simulated data, we show that our method improves causal effect estimation (Sec. 6).

2 PRELIMINARIES

We use the framework of semi-Markovian causal models [SMCMs, Tian and Shpitser, 2003], and assume the reader is familiar with related terminology. Variables are denoted in uppercase, their values in lowercase, and variable sets in bold. We use $\mathcal{G}$ to denote a causal graph, and say $\mathcal{G}$ induces a probability distribution P if P factorizes according to $\mathcal{G}$.

We use $Y|do(X)$ or Y_X to denote a variable Y after the hard intervention on variable X. If we know the SMCM $\mathcal{G}$, a hard intervention in which a treatment X is set to x can be represented with the do-operator, $do(X{=}x)$. We use P_x to denote the interventional distribution over the same variables for $do(X{=}x)$. In the corresponding graph, 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. We use the terms $Pa_{\mathcal{G}}(Z)$ and $Ch_{\mathcal{G}}(Z)$ to denote the set of parents and children of Z in $\mathcal{G}$,

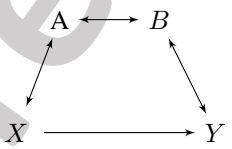

Figure 1: An example SMCM. $\{X, A, B\}$ is the Markov boundary of Y in $\mathcal{G}_{\overline{X}}$. $P(Y|do(X), A, B)$ is not identifiable from the observational distribution P, but $P(Y|do(X), A)$ and $P(Y|do(X), B)$ are. $\{X, B\}$ is the causal Markov boundary for Y_X.

respectively. The set of variables that are connected with a variable Y through a bidirected path (i.e., a path that only has bidirected edges) is called the district of Y and denoted $Dis_{\mathcal{G}}(Y)$.

3 MARKOV BOUNDARIES

A Markov blanket of a variable Y in a set of variables $\mathbf{V}$ is a subset $\mathbf{Z}$ of $\mathbf{V}$ conditioned on which other variables are independent of Y: $Y \perp\!\!\!\perp \mathbf{V}\backslash\mathbf{Z}|\mathbf{Z}$. The Markov boundary of Y is the Markov blanket that is also minimal (i.e., no subset of the Markov boundary is a Markov blanket) [Pearl, 2000]. In distributions that satisfy the intersection property (including faithful distributions), the Markov boundary of a variable Y is unique [Pearl, 1988]. To distinguish from other types of Markov boundaries defined in this work, we often use the terminology **observational Markov boundary (OMB)** to denote the Markov boundary of a variable.

For a DAG $\mathcal{G}$, the OMB of a variable Y in any distribution faithful to $\mathcal{G}$ is the set parents, children, and spouses of Y: MB$(Y) = Pa_{\mathcal{G}}(Y) \cup Ch_{\mathcal{G}}(Y) \cup Pa_{\mathcal{G}}(Ch_{\mathcal{G}}(Y))$. For SMCMs, it has been shown that the OMB of a variable Y is the set of parents, children, children's parents (spouses) of Y, district of Y and districts of the children of Y, and the parents of each node of these districts [Richardson, 2003, Pellet and Elisseeff, 2008][1].

The OMB has been shown to be the minimal set of variables with optimal predictive performance for a given distribution and response variable, given some assumptions on the learner and the loss function [Tsamardinos and Aliferis, 2003]. In this work, we are interested in the model that gives the optimal prediction of the post-intervention distribution, with the goal of designing optimal policies. For this reason, we do not include post-intervention covariates in this model, because these variables are not known prior to treatment assignment, and thus, cannot affect the assignment. In the rest of this document, we make the following assumption:

Assumption 3.1. Covariates $\mathbf{V}$ are pre-treatment.

Making this assumption also simplifies the expressions for the OMBs, because we no longer need to consider children

[1] Pellet and Elisseeff [2008] prove this for maximal ancestral graphs, but the proof can be readily adapted to SMCMs.

of Y and their districts. Knowing the OMB allows a more efficient representation of the conditional distribution of Y given $\mathbf{V}$, since the following equation holds:

$$P(Y|\mathbf{V}) = P(Y|\text{MB}(Y)). \quad (1)$$

3.1 INTERVENTIONAL MARKOV BOUNDARY

Our goal is to identify the set of variables that lead to the optimal model for the post-intervention distribution of a target Y relative to a specific treatment X. We call this set the **interventional Markov boundary (IMB)** of Y relative to X, and denote it $\text{MB}_X(Y)$. Obviously, $\text{MB}_X(Y) \subseteq \text{MB(Y)}$. When we have data from the post-intervention distribution, we can apply statistical methods for OMB identification to obtain the IMB of Y relative to X. However, experimental data are often limited in sample sizes, while OMB identification methods may require large sample sizes.

If we know the causal graph $\mathcal{G}$, the post-intervention distribution with respect to X is induced by the manipulated graph $\mathcal{G}_{\overline{X}}$. The IMB of Y is then the OMB of Y in $\mathcal{G}_{\overline{X}}$, and can be identified using the definition of the Markov boundary above. However, the post-intervention distribution $P(Y|do(X), \text{MB}_X(Y) \setminus X)$, may not be identifiable from the observational distribution. For example, in Fig. 1, $\text{MB}_X(Y) = \{X, A, B\}$, but $P(Y|do(X), A, B)$ is not identifiable from observational data. We then want to answer the following question: *What is the best model for predicting Y_X from the observational distribution, when the causal graph is known?*

3.2 CAUSAL MARKOV BOUNDARIES

To answer this question, we define the **causal Markov boundaries** of an outcome Y relative to a treatment X as follows:

Definition 3.2. Let $\mathbf{Z} \subseteq (\mathbf{V} \cup X)$, and $\mathbf{W} = \mathbf{Z} \setminus X$. Then $\mathbf{Z}$ is a causal Markov boundary (CMB) for Y relative to X if it satisfies the following properties:

1. $P(Y|do(X), \mathbf{W})$ is identifiable from $P(X, Y, \mathbf{V})$.
2. For every subset $\mathbf{W'}$ of $\mathbf{V} \setminus \mathbf{W}$ either $P(Y|do(X), \mathbf{W}, \mathbf{W'})$=$P(Y|do(X), \mathbf{W})$ or $P(Y|do(X), \mathbf{W}, \mathbf{W'})$ is not identifiable from $P(X, Y, \mathbf{V})$.
3. $\nexists \mathbf{W'} \subset \mathbf{W}$ s.t. $P(Y|do(X), \mathbf{W'})$=$P(Y|do(X), \mathbf{W})$.

Condition (1) ensures that the post-intervention conditional probability of Y_X given a CMB is identifiable. Condition (2) states that the covariates that are not in that CMB are either redundant for the prediction of Y_X given the CMB, or they make the post-intervention distribution non-identifiable. Condition (3) ensures that $\mathbf{Z}$ is additionally maximally informative for Y_X in the sense that you cannot remove any variable from $\mathbf{Z}$ without losing some information for Y_X. This condition rules out sets like $\{X, A\}$ in Fig. 1, where, while $P(Y|do(X), A)$ is identifiable from P, it is equal to $P(Y|do(X))$. Thus, conditioning on A does not improve the prediction of Y_X compared to its subset $\emptyset$.

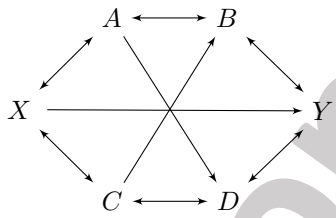

Figure 2: Causal Markov boundaries are not necessarily unique. Both $\{X, B, C\}$ and $\{X, A, D\}$ are causal Markov boundaries for Y relative to X.

Notice that this definition does not capture the spirit of Markov boundaries precisely: Markov boundaries make all remaining variables redundant for predicting Y; however, this does not necessarily hold with CMBs. For example, in Fig. 1, $\{X, B\}$ is a CMB according to the definition above, but $\{A\}$ remains relevant for predicting Y_X; however, including it with B in the CMB leads to non-identifiability.

A CMB is not necessarily unique; it is possible that multiple sets satisfy Definition 3.2. For example, assume the distribution P is induced by the SMCM shown in Fig. 2. Both $\{X, B, C\}$ and $\{X, A, D\}$, satisfy Definition 3.2. The best predictive CMB for predicting Y_X will depend on the parameters in P. We use the notation $\mathbf{CMB}_X(Y)$ to denote the set of causal Markov boundaries of Y relative to X. Thus, we will generally need to find all CMBs and then determine which of them leads to the best prediction of Y_X. Also, notice that the $\mathbf{CMB}_X(Y)$ can be empty; thus, no subset of $\mathbf{V}$ satisfies the Definition 3.2. This can happen for example if $X \to Y$ and $X \leftrightarrow Y$ in $\mathcal{G}$.

The CMB is useful in determining a minimal set of maximally predictive variables for which we can use observational data to predict post-interventional distributions. In the next section we show that, for pre-treatment covariates, CMBs satisfy the backdoor criterion and are subsets of the observational Markov boundary. These results enable more efficient algorithms for finding CMBs, limiting the types of estimators and the number of variable sets we need to consider.

3.2.1 Characterization

Given a graph $\mathcal{G}$, Shpitser and Pearl [2006a] provide a sound and complete algorithm (IDC) for estimating conditional post-intervention distributions from observational distributions induced by $\mathcal{G}$. The output of this algorithm is an expression for $P(Y|do(X), \mathbf{W})$ if the distribution is identifiable

from distribution P and $\mathcal{G}$, or N/A otherwise. Thus, we can identify CMBs in a brute-force way by running IDC for every possible subset of $\mathbf{V}$, and then check for sets that satisfy the conditions in Definition 3.2. This process is computationally expensive and would not be possible for graphs with more than a few variables.

In this section, we provide theoretical results that lead to a much easier process when all candidate conditioning variables are pre-treatment (all proofs can be found in the supplementary). For pre-treatment covariates, one obvious family of sets for which the conditional post-intervention distributions are identifiable are sets that m-separate X and Y in $\mathcal{G}_{\underline{X}}$. These sets satisfy Rule 2 of do-calculus [Pearl, 2000], so the conditional interventional distribution $P(Y|do(X), \mathbf{W})$ is equal to the observational distribution $P(Y|X, \mathbf{W})$. Sets of pre-treatment covariates that m-separate X and Y in $\mathcal{G}_{\underline{X}}$ are also known to satisfy the backdoor criterion [Van der Zander et al., 2014] and the adjustment criterion [Shpitser et al., 2010]. However, these definitions are more general to include possible post-treatment covariates, and are intended for estimating marginal post-intervention distributions (or average effects). For brevity, we will call sets that m-separate X and Y in $\mathcal{G}_{\underline{X}}$ **backdoor sets**, since they block all backdoor paths between X and Y.

One question that arises is if there are sets that are not backdoor sets that may satisfy the conditions in Definition 3.2. In that case, identifiability could stem from some sequential application of do-calculus rules. As we show next, this is not possible for pre-treatment covariates. This ensures that we only need to check CMB Conditions (2) and (3) for sets for which $P(Y|do(X), \mathbf{W}) = P(Y|X, \mathbf{W})$. This makes the identification of $P(Y|do(X), \mathbf{W})$ more straightforward than having to compute more complex probability expressions.

Theorem 3.3. *We assume that P_x and $\mathcal{G}_{\overline{X}}$ are faithful to each other. If $\mathbf{Z}$ is a CMB for Y relative to X, then $\mathbf{Z} \setminus X$ is a backdoor set for X relative to Y.*

The second theoretical result is that any CMB of Y relative to X is a subset of the OMB of Y. While this sounds intuitive, it is not completely straightforward. It could be the case that conditioning on every subset of the OMB opens some m-connecting path between X and Y, that can only be blocked by a variable that is not a member of the Markov boundary. The following theorem proves that this is not possible, allowing for more efficient search algorithms:

Theorem 3.4. *We assume that P_x and $\mathcal{G}_{\overline{X}}$ are faithful to each other. Every CMB $\mathbf{Z}$ of an outcome variable Y w.r.t a treatment variable X is a subset of the OMB* $\mathrm{MB}(Y)$.

Based on Theorem 3.4, we only need to look for CMBs within subsets of the OMB of Y. So far, we have shown that both the IMB and any CMB are subsets of the OMB. We can also show that when the IMB is a CMB, then it also coincides with the OMB:

Theorem 3.5. *If* $\mathrm{MB}_X(Y)$ *is a causal Markov boundary, then* $\mathrm{MB}_X(Y) = \mathrm{MB}(Y)$.

4 COMBINING OBSERVATIONAL AND EXPERIMENTAL DATA

When the causal graph is known, we can obtain CMBs by looking for subsets of $\mathrm{MB}(Y)$ that satisfy Definition 3.2. Unfortunately, in most real-world applications, the true graph is unknown, and selecting the causal/interventional Markov boundary is not possible from observational data alone. Experimental data may exist, but are typically much fewer than observational data, due to expense or ethical concerns. This scenario is common in embedded trials, where non-randomized patients are much more common than trial participants. In such cases, the experimental data may be underpowered to accurately estimate conditional effects. As a result, the conditional effects that can be derived from the experimental data have high variance and may not be reliable. In this case, combining all data (observational and experimental) in a Bayesian manner may help improve the prediction of Y_x.

We assume that we have observational data D_o and experimental data D_e measuring treatment X, outcome Y, and pre-treatment covariates $\mathbf{V}$. We use N_o, N_e to denote the number of samples in D_o, D_e, respectively.

We present a Bayesian method, called `FindIMB`, that uses both D_e and D_o to estimate the probability of a set being the $\mathrm{MB}_X(Y)$, and estimate $P(Y|do(X), \mathbf{V}) = P(Y|do(X), \mathrm{MB}_X(Y) \setminus X)$. The method is presented in Alg. 1. It first estimates the OMB of Y in observational data $\mathrm{MB}(Y)$ (Line 1), and then looks among subsets of $\mathrm{MB}(Y)$ for sets that are IMBs (Line 2). It uses D_e and D_o to evaluate the probability that a set is an IMB (Line 3), and then returns a weighted average for $P(Y|do(X), \mathbf{V})$ based on these probabilities (Line 5).

The enabling idea of the method is that, when the IMB is a CMB, we can use both the D_o and D_e to estimate the conditional post-intervention distribution. Otherwise, we use only D_e to derive the estimate. We use the following notation to express these hypotheses:

- $H^c_{\mathbf{Z}}$ is a binary variable denoting the hypothesis that $\mathbf{Z}$ is the IMB $\mathrm{MB}_X(Y)$, and it is also a CMB: $\mathbf{Z} = \mathrm{MB}_X(Y) \wedge \mathbf{Z} \in \mathbf{CMB}_X(Y)$.
- $H^{\bar{c}}_{\mathbf{Z}}$ is a binary variable denoting the hypothesis that $\mathbf{Z}$ is the IMB $\mathrm{MB}_X(Y)$, but it is not a CMB: $\mathbf{Z} = \mathrm{MB}_X(Y) \wedge \mathbf{Z} \notin \mathbf{CMB}_X(Y)$.

For a set $\mathbf{Z}^\star$, if either $H^c_{\mathbf{Z}^\star}$ or $H^{\bar{c}}_{\mathbf{Z}^\star}$ is true, $\mathbf{Z}^\star$ is an IMB and therefore $P(Y|do(X), \mathbf{V}) = P(Y|do(X), \mathbf{Z}^\star \setminus X)$.

Under $H^c_{\mathbf{Z}^\star}$, however, $\mathbf{Z}^\star$ is also a CMB and therefore the pre- and post- intervention distributions are the same, i.e.,

$$P(Y|do(X), \mathbf{Z}^\star \setminus X, H^c_{\mathbf{Z}^\star}) = P(Y|X, \mathbf{Z}^\star \setminus X). \quad (2)$$

In contrast, under $H^{\bar{c}}_{\mathbf{Z}^\star}$ (i.e., $\mathbf{Z}^\star$ is an IMB but not a CMB), $P(Y|do(X), \mathbf{Z}^\star \setminus X)$ is not identifiable from observational data.

Corollary 4.1. *Under $H^{\bar{c}}_{\mathbf{Z}^\star}$, $P(Y|do(X), \mathbf{Z}^\star \setminus X)$ is not identifiable from observational data.*

Corollary 4.1 stems from the fact that $H^c_{\mathbf{Z}^\star}$ is true only when $\mathbf{Z}^\star$ is the IMB, and from the definition of the CMB. Specifically, if $P(Y|do(X), \mathrm{MB}_X(Y) \setminus X)$ was identifiable from observational data, then it would satisfy all conditions in Definition 3.2, and it would therefore be a CMB. Therefore, if $H^{\bar{c}}_{\mathbf{Z}^\star}$ holds, $P(Y|do(X), \mathrm{MB}_X(Y) \setminus X)$ is not identifiable from P, and we cannot use D_o to estimate $P(Y|do(X), \mathbf{Z}^\star \setminus X)$[2]. Thus, if $H^c_{\mathbf{Z}^\star}$ holds, we can use both D_o and D_e in our estimation of $P(Y|do(X), \mathbf{Z}^\star \setminus X)$, while if $H^{\bar{c}}_{\mathbf{Z}^\star}$ holds we can only use D_e.

Based on this observation, we want to compute $P(H^c_{\mathbf{Z}} | D_e, D_o)$ and $P(H^{\bar{c}}_{\mathbf{Z}} | D_e, D_o)$ for possible IMBs $\mathbf{Z}$. These probabilities tell us both how likely it is that $\mathbf{Z}$ is an IMB (their sum), and if we can include observational data in the estimation of $P(Y|do(X), \mathbf{V})$. Using Bayes' rule, we obtain:

$$P(H^c_{\mathbf{Z}'} | D_e, D_o) = \frac{P(D_e|D_o, H^c_{\mathbf{Z}'})P(D_o|H^c_{\mathbf{Z}'})P(H^c_{\mathbf{Z}'})}{\sum_{\mathbf{Z}} \sum_{C=c,\bar{c}} P(D_e|D_o, H^C_{\mathbf{Z}})P(D_o|H^C_{\mathbf{Z}})P(H^C_{\mathbf{Z}})}. \quad (3)$$

We can similarly derive $P(H^{\bar{c}}_{\mathbf{Z}} | D_e, D_o)$ by replacing each appearance of c with $\bar{c}$ in the numerator. The denominator is the same for all sets. $P(H^c_{\mathbf{Z}'})$ and $P(H^{\bar{c}}_{\mathbf{Z}'})$ are our priors that $H^c_{\mathbf{Z}'}$ and $H^{\bar{c}}_{\mathbf{Z}'}$ hold, respectively. We set this to be uniform over both values of C and all $\mathbf{Z}$.

As Eq. 3 shows, using Bayes' rule, we can estimate the posterior probabilities for the set of hypotheses $H^c_{\mathbf{Z}}$ and $H^{\bar{c}}_{\mathbf{Z}}$ using marginal likelihoods of the experimental and observational data. In the next sections, we show how we can compute each term in Eq. 3.

We present our results for multinomial distributions, but we believe the method can be readily extended to any type of distribution with closed-form marginals. For Gaussian distributions, the equations in this paper can be computed using results provided in [Heckerman and Geiger, 1995]. Due to space constraints, the closed-form solution for each equation appearing in the remainder of the paper is presented in Supplementary Table S1.

[2]Notice however that D_o may still place some constraints on $P(Y|do(X), \mathbf{Z}^\star \setminus X)$, like for example provide bounds.

Estimating $P(D_e|D_o, H^c_{\mathbf{Z}'})$, $P(D_e|D_o, H^{\bar{c}}_{\mathbf{Z}'})$:
Let $\mathbf{W} = \mathbf{Z} \setminus X$, and let $\theta_{Y_x|\mathbf{W}}$ be a set of parameters expressing the conditional probabilities for $P(Y|do(X), \mathbf{W})$. Also, let $\theta_{Y|X,\mathbf{W}}$ denote the observational parameters for $P(Y|X, \mathbf{W})$. By integrating over all $\theta_{Y_x|\mathbf{W}}$, we obtain

$$P(D_e|D_o, H^c_{\mathbf{Z}}) = \int_{Y_x|\mathbf{W}} P(D_e|\theta_{Y_x|\mathbf{W}}) f(\theta_{Y_x|\mathbf{W}}|D_o, H^c_{\mathbf{Z}}) d\theta_{Y_x|\mathbf{W}}, \quad (4)$$

where $f(\theta_{Y_x|\mathbf{W}}|D_o, H^c_{\mathbf{Z}})$ is the posterior for $\theta_{Y_x|\mathbf{W}}$ given the observational data, when $\mathbf{Z}$ is the IMB and the CMB. In this case, $P(Y|do(X), \mathbf{W}) = P(Y|X, \mathbf{W})$, and therefore $f(\theta_{Y_x|\mathbf{W}}|D_o, H^c_{\mathbf{Z}}) = f(\theta_{Y|X,\mathbf{W}}|D_o)$. Eq. 4 can then be rewritten in terms of the observational parameters as

$$P(D_e|D_o, H^c_{\mathbf{Z}}) = \int_{\theta_{Y|X,\mathbf{W}}} P(D_e|\theta_{Y|X,\mathbf{W}}) f(\theta_{Y|X,\mathbf{W}}|D_o) d\theta_{Y|X,\mathbf{W}}. \quad (5)$$

Eq. 5 is the marginal likelihood of Y in experimental data, with parameter density $f(\theta_{Y|X,\mathbf{W}}|D_o)$ being equal to the parameter posterior given D_o. In other words, under $H^c_{\mathbf{Z}}$, the observational and experimental parameters coincide. Therefore, D_o gives us a strong "prior" for D_e. Eq. 5 can be computed in closed-form for distributions with conjugate priors.

Under $H^{\bar{c}}_{\mathbf{Z}}$, the equality of the observational and experimental parameters does not hold, and we cannot use the $\theta_{Y|X,\mathbf{W}}$ to inform $\theta_{Y_X|\mathbf{W}}$, at least not in a straightforward way. Instead, we model that $f(\theta_{Y_X|\mathbf{W}}|D_o) = f(\theta_{Y_X|\mathbf{W}})$. Then $P(D_e|D_o, H^{\bar{c}}_{\mathbf{Z}})$ corresponds to the marginal likelihood of Y in the experimental data, using a prior that we model as being non-informative.

Estimating $P(D_o|H^c_{\mathbf{Z}})$, $P(D_o|H^{\bar{c}}_{\mathbf{Z}})$:
These probabilities score how well the observational data fit with the hypotheses $H^c_{\mathbf{Z}}$, $H^{\bar{c}}_{\mathbf{Z}}$. We can derive these terms on the basis of the OMB and its connection to the IMB and the CMBs. We first need to express the hypothesis **that a set $\mathbf{U}$ is the OMB of** Y: Let $H^o_{\mathbf{U}}$ denote this hypothesis; thus, for any $\mathbf{U} \subseteq \mathbf{V} \cup X$, $H^o_{\mathbf{U}}$ is true iff $\mathbf{U}$ is the OMB for Y. Then we can write

$$P(D_o|H^C_{\mathbf{Z}}) = \sum_{\mathbf{U} \subseteq \mathbf{V} \cup X} P(D_o|H^o_{\mathbf{U}}) P(H^o_{\mathbf{U}} | H^C_{\mathbf{Z}}), \quad (6)$$

for $C = c, \bar{c}$. Under $H^c_{\mathbf{Z}}$, Theorem 3.5 implies that $P(H^o_{\mathbf{U}} | H^c_{\mathbf{Z}}) = 1$ if $\mathbf{U} = \mathbf{Z}$, and zero otherwise. Under $H^{\bar{c}}_{\mathbf{Z}}$, the IMB is not a CMB. Instead, the IMB has to be a subset of $\mathbf{U}$, therefore $P(H^o_{\mathbf{U}} | H^C_{\mathbf{Z}}) = 0$ for any $\mathbf{Z} \supset \mathbf{U}$.

$P(D_o|H^o_{\mathbf{U}})$ is the marginal likelihood of Y in D_o, under the hypothesis that $\mathbf{U}$ is the data-generating OMB for Y in the observational data. We can obtain this likelihood using a Bayesian scoring algorithm like FGES [Ramsey et al., 2017], by scoring a DAG where Y is a child of variables $\mathbf{U}$

(and no other edges are in the graph). We call this algorithm FGESMB. For discrete variables, in the large sample limit this probability will be maximum only for the true OMB:

Theorem 4.2. *Given dataset D_o that contains samples from a strictly positive distribution P, which is a perfect map for a SMCM $\mathcal{G}$, the BD score [Heckerman et al., 1995] will assign the highest score to the OMB of Y in the large sample limit.*

Eq. 6 needs to be computed for all subsets of the covariate sets. In practice, however, for large N_o, these probabilities are dominated by the true OMB. Assuming our sample is large enough, we can use an algorithm with asymptotic guarantees for identifying the true OMB. In fact, once we commit to the observational Markov boundary $\mathbf{U}^\star$, Eq. 6 leads to the following equations

$$\begin{aligned} &P(D_o|H^c_{\mathbf{Z}}) = P(D_o|H^o_{\mathbf{U}^\star}) \text{ for } \mathbf{Z}=\mathbf{U}^\star \\ &P(D_o|H^{\bar{c}}_{\mathbf{Z}}) = P(D_o|H^o_{\mathbf{U}^\star}) \text{ for all } \mathbf{Z} \subseteq \mathbf{U}^\star. \end{aligned} \tag{7}$$

For the remaining cases (i.e., $H^c_{\mathbf{Z}}$ and $\mathbf{Z} \neq \mathbf{U}^\star$, or $H^{\bar{c}}_{\mathbf{Z}}$ and $\mathbf{Z} \supset \mathbf{U}^\star$), the corresponding probabilities are zero.

Bayesian estimation of $P(Y|do(X), \mathbf{V}, D_e, D_o)$:
We now compute the $P(Y|do(X), \mathbf{V})$ using Bayesian model averaging over the hypotheses $H^C_{\mathbf{Z}}$. Let $x, y, \mathbf{V} = \mathbf{v}$, denote given instances of X, Y and $\mathbf{V}$, respectively. When $\mathbf{V} = \mathbf{v}$, we use $\mathbf{W} = \mathbf{w}_v$ to denote the corresponding values of a set $\mathbf{W} \subset \mathbf{V}$. Recall that under $H^c_{\mathbf{Z}}$ and under $H^{\bar{c}}_{\mathbf{Z}}$, $P(Y|do(X), \mathbf{V}) = P(Y|do(X), \mathbf{W})$, where $\mathbf{W} = \mathbf{Z} \setminus X$. Then for a given instance of $\mathbf{V} = \mathbf{v}$, we have

$$\begin{aligned} &P(y|do(x), \mathbf{v}, D_e, D_o) = \\ &\sum_{\mathbf{Z} \subset \mathbf{V}} \sum_{C=c,\bar{c}} P(y|do(x), \mathbf{w}_v, D_e, D_o, H^C_{\mathbf{Z}}) P(H^C_{\mathbf{Z}}|D_o, D_e). \end{aligned} \tag{8}$$

This equation computes the *expectation* of the conditional probability parameter. The individual probabilities $P(y|do(x), \mathbf{w}_v, D_e, D_o, H^C_{\mathbf{Z}})$ can be estimated as posterior expectations of $P(Y|do(X), \mathbf{W})$ from the data. Specifically, under given $H^c_{\mathbf{Z}}$, $P(Y|do(X), \mathbf{W}) = P(Y|X, \mathbf{W})$, and therefore we can use both D_e and D_o for the posterior expectation. In contrast, under $H^{\bar{c}}_{\mathbf{Z}}$, we only use D_e. Analytical equations for these probabilities for multinomial distributions can be found in Supplementary Table S1.

5 RELATED WORK

We are not aware of other methods that try to identify causal and interventional Markov boundaries. Our work has connections and builds on work from many different areas. Due to space constraints, we only focus on methods that do not require causal sufficiency.

Algorithm 1: FindIMB

input : D_o, D_e, treatment X, outcome Y, pre-treatment covariates $\mathbf{V}$
output : Post-intervention distribution $P(Y|do(X), \mathbf{V})$

```
MB(Y) ← MarkovBoundary(Y, D_o);
foreach subset Z of MB(Y) and C = c, c̄ do
    Compute P(H^C_Z | D_e, D_o) using Eq. 3;
    Compute P(Y|do(X), V, D_e, D_o, H^C_Z) using Eq. 8;
P(Y|do(X), V) ← Σ_Z Σ_{C=c,c̄} P(Y|do(X), V, D_e, D_o, H^C_Z) P(H^C_Z | D_e, D_o);
```

Markov boundaries: Several algorithms learn OMBs from data under causal insufficiency [Yu et al., 2018, 2020]. In addition, FGESMB presented in Sec. 4 is also a sound and complete method for learning OMBs from data. These methods can be used to identify the IMBs from the experimental data, but they do not combine observational and experimental data to learn IMBs.

Identifiability: Shpitser and Pearl [2006a,b] and Tian and Shpitser [2003] provide sound and complete identifiability results for post-intervention distributions from observational data when the causal graph is known. These methods can answer queries for a specific marginal or conditional probability of interest. Hyttinen et al. [2015] and Jaber et al. [2019] provide similar identifiability results when the graph is unknown, using the Markov equivalence class of graphs that are consistent with the observational data. Hyttinen et al. [2015] can provide identifiability results for graphs that are consistent with conditional independencies in both D_e and D_o. However, the method is not proven to be complete for these settings. These methods are not directly comparable with our method because they do not select features for optimal prediction. Moreover, they provide expressions for the post-intervention distributions that are based on observational data alone, not by combining D_o and D_e like FindIMB.

Combining observational and experimental data to learn causal graphs: Several causal discovery methods combine observational and experimental data to learn causal structure [Triantafillou and Tsamardinos, 2015, Hyttinen et al., 2014, Mooij et al., 2020, Andrews et al., 2020]. These methods return a summarized version of all the causal graphs that are consistent with all the independence constraints in all the data sets, observational and experimental. While these methods can be used to improve the estimation of IMBs, it is not clear that they can always provide a unique solution in this setting. Two additional drawbacks they have for the purpose of optimized target prediction are that (a) they rely on conditional independence tests that

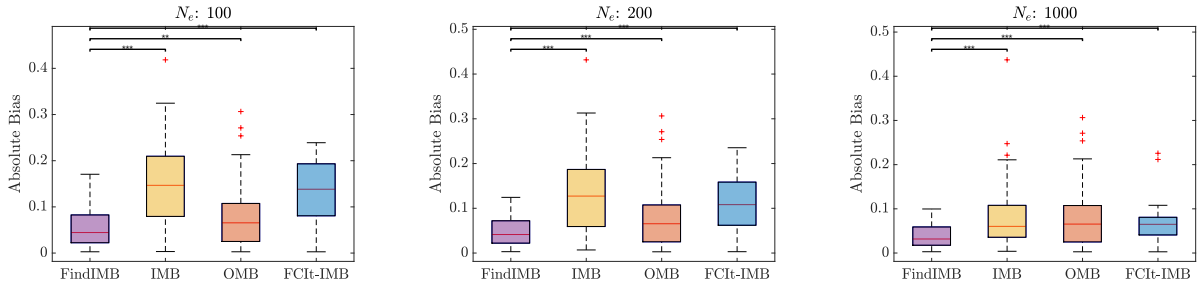

Figure 3: Boxplots of absolute bias in the estimation of $P(Y|do(X), \mathbf{V})$ using (a) `FindIMB` (b) IMB (c) OMB and (d) `FCIt-IMB`. Data were simulated from random DAGs with 10 observed and 5 latent variables. D_o included 10,000 samples, and D_e included 100 (left), 200 (middle) or 1000 (right) samples. `FindIMB` improves the estimation of $P(Y|do(X), \mathbf{V})$ particularly for smaller experimental sample sizes. Black asterisks denote statistical significance, assessed with the Wilcoxon signed-rank test. Three stars correspond to $p < 0.001$.

are unreliable when N_e is low, and (b) they learn the entire graph and do not focus on finding the neighborhood of the target variable. This can result in unreliable orientations due to error propagation. The FCItiers method introduced by Andrews et al. [2020] has the closest setting to ours ([Mooij et al., 2020] is also related, but more general, and the two are equivalent for our setting). FCItiers can learn a family of SMCMs from D_e and D_o when (a) the target of the intervention is known and (b) we specify "tiered knowledge" on the variables (e.g., we know which variables are pre-treatment). The method is complete in these settings. In the experimental section, we develop a baseline comparison method based on FCItiers.

Selecting optimal adjustment sets: Some methods seek to select optimal adjustment sets for efficient average treatment effect estimation. Given a graph (DAG/PDAG or SMCM), these methods apply a graphical adjustment criterion to identify a set of valid adjustment sets for estimating the average treatment effect of X on Y. Then, they try to identify the set that leads to the estimator with the smallest asymptotic variance among all the valid adjustment sets [Perkovic et al., 2017, Rotnitzky and Smucler, 2019, 2020, Smucler et al., 2020, Witte et al., 2020]. While these methods have a different purpose than ours, they have some connections with our work since, for pre-treatment variables, any CMB is also an adjustment set. We point out that while optimal adjustment sets and CMBs may often coincide (for example in DAGs), they are not always the same (see example Fig. S1 in the Supplementary). Moreover, these methods are not directly comparable to ours since they focus on identifying average treatment effects while our method focuses on conditional effects and combines observational with experimental data when the graph is unknown.

Potential outcomes approaches: Kallus et al. [2018] present a method for estimating conditional average treatment effects (CATEs) by combining D_o and D_e. The method assumes a binary treatment and uses the experimental data to model the effect of possibly unmeasured confounders as a function of the measured covariates. The CATE is obtained from the D_o by adding the modeled correction. The main assumption of the method is that the hidden confounding has an identifiable parametric structure. The method is implemented for continuous covariates and outcome and a linear correction function, obtained by solving a least squares optimization problem. It is not directly applicable to our settings of categorical covariates, and extending the optimization problem in these settings is not straightforward.

Transportability: Finally, our work has some connections with the field of transportability [Bareinboim and Pearl, 2013], where knowledge of the causal graph is used to determine if the results of an experimental trial apply to a different population. However, the methods require knowing the causal graph, and focus on transferring estimators across distributions rather than combining data to improve estimators.

6 EXPERIMENTS

In this section, we show the performance of `FindIMB` using simulated data. We simulated random DAGs with a varying number of discrete variables, with mean in-degree 2. Each DAG includes a binary treatment X and outcome Y, where $X \to Y$. The remaining covariates $\mathbf{V}$ are pre-treatment and are binary or ternary, and $1/3$ of the variables are set to be latent at random. The observational data D_o consist of 10,000 simulated samples from the ground truth DAGs, and do not include values for the latent variables. The experimental data D_e were simulated from the manipulated graph $\mathcal{G}_{\overline{X}}$. For sample sizes $N_e = \{100, 200, 1000\}$, we simulated $N_e/2$ samples for each value of binary treatment X. The code for the data simulation and the following experiments are available at `https://github.com/striantafillou/UAI2021`.

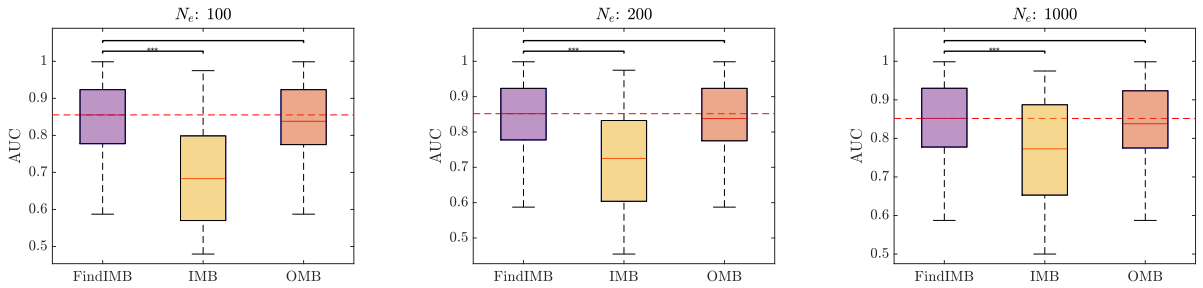

Figure 4: Boxplots of areas under the ROC curve for predicting Y_x using (a) `FindIMB` (b) IMB and (c) OMB. Data were simulated from random DAGs with 40 observed and 20 latent variables. D_o included 10,000 samples, and D_e included 100 (left), 200 (middle) or 1000 (right) samples. Black asterisks denote statistical significance, assessed with the Wilcoxon signed-rank test. Three stars correspond to $p < 0.001$, no stars denote non-significance.

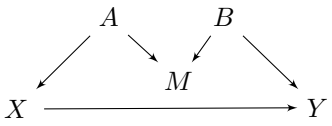

Figure 5: The m-bias graph used to simulate data for Fig. 6. A and B are unobserved. All variables are binary. Parameters are as follows: $P(A{=}1){=}0.8$, $P(B{=}1){=}0.8, P(M{=}1|A{=}1, B{=}1){=}\alpha, P(X{=}1|A{=}1){=}\alpha$, $P(Y{=}1|X{=}1, B{=}1){=}\alpha$. All other parameters $P(Y{=}1|\dots), P(M{=}1|\dots), P(X{=}1|A{=}0)$ were set to zero.

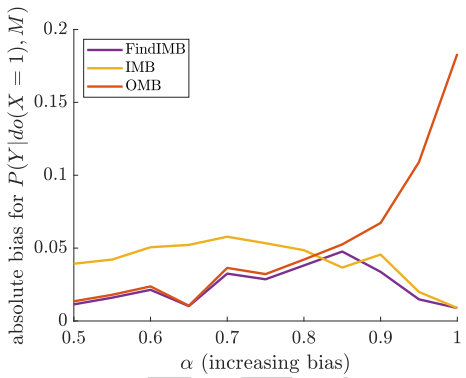

Figure 6: Performance of `FindIMB`, IMB, and OMB for estimating $P(Y|do(X{=}1), M)$ with increasing m-bias.

Comparison to other approaches: We compared `FindIMB` to the following approaches: (a) **IMB**: using only experimental data. We used D_e to identify the $\text{MB}_X(Y)$ using `FGESMB`. After identifying $\text{MB}_X(Y)$, we used the posterior expectation $P(Y|do(X), \text{MB}_X(Y) \setminus X, D_e)$ as the estimator for $P(Y|do(X), \mathbf{V})$. (b) **OMB**: using only observational data. We used `FGESMB`(D_o) to identify the OMB of Y, $\text{MB}(Y)$, and used the posterior expectation $P(Y|do(X), \text{MB}(Y) \setminus X, D_o)$ estimated on D_o as the estimator of $P(Y|do(X), \mathbf{V})$. This estimator is unbiased when conditional ignorability holds for the OMB of Y. (c) **FCIt-IMB:** using both observational and experimental data based on `FCIt-IMB`. We applied `FCIt-IMB` using as input a data set D constructed by concatenating D_e and D_o, and adding a binary variable $I_e \to X$ that corresponds to the presence or absence of manipulation of X. So, $I_e = 1$ for samples in D_e and $I_e = 0$ otherwise. `FCIt-IMB` outputs a PAG $\mathcal{P}$ representing all possible underlying SMCMs. Let $\mathcal{P}_{\overline{X}}$ denote the corresponding manipulated PAG. We then considered the Markov boundary of Y in $\mathcal{P}_{\overline{X}}$ to be the $\text{MB}_X(Y)$. After identifying $\text{MB}_X(Y)$, we tested if it is a backdoor set in $\mathcal{P}_{\overline{X}}$. If so, we used both D_e and D_o pooled together to estimate $P(Y|do(X), \text{MB}_X(Y) \setminus X)$. Otherwise, we only used D_e.

First, we tested if our `FindIMB` method improves estimation of the probability $P(Y|do(X), \mathbf{V})$. We simulated DAGs with 10 observed and 5 latent variables, and applied the methods described above. Each method outputs a set of variables $\mathbf{Z}$, that is used as an estimate for $P(Y|do(X), \mathbf{Z}, \mathbf{V} \setminus \mathbf{Z})$. Notice that even for 10 variables, the number of possible configurations of $\mathbf{Z}$ can be very large, and some of these configurations may be very rare. To avoid computing these parameters for all possible configurations, we tested the methods in a test dataset D_e^{test}, that includes 1000 treatment and 1000 control cases simulated from the manipulated ground truth graph. For each sample in D_e^{test}, we obtained an estimate $\hat{P}(Y|do(X), \mathbf{V})$ with the four methods. The ground truth probability was estimated from the original manipulated Bayesian network with the junction tree algorithm. We then computed the average absolute bias $|\hat{P}(Y|do(X), \mathbf{V}) - P(Y|do(X), \mathbf{V})|$ over all test samples. Fig. 3 shows that `FindIMB` produced the most accurate probabilities, compared to using only observational or only experimental data. Moreover, `FindIMB` outperforms `FCIt-IMB` ($p < 10^{-3}$ in all cases for a Wilcoxon signed-rank test). One reason is that `FCIt-IMB` selects much larger IMBs than `FindIMB`, possibly due to error propagation that results in many bidirected edges. Thus, the

resulting parameters are estimated based on much fewer samples.

We also tested the scalability of FindIMB, using DAGs with 40 observed and 20 latent variables. For this experiment, we could not test against FCIt−IMB because the method results in very large IMBs due to the presence of latent variables. Similarly, we could not estimate the true parameters $P(Y|do(X), \mathbf{V})$ for computational reasons, since the ground truth IMBs can also be very large and include rare configurations, and ground truth values would need to be computed using inference and marginalizing over latent variables. For this reason, instead of measuring the average absolute bias, we measured the performance of the methods to classify the samples in the test dataset D_e^{test} (2000 samples) correctly. Fig. 4 shows the area under the ROC curve (AUC) of the models based on FindIMB, IMB, and OMB. FindIMB performs on par or better than the two alternatives. Average running time for FindIMB (learning the model) was 1.71 ± 2.46 seconds for one run of the algorithm.

One interesting finding is that in all experiments, using the observational data only, performs better than using experimental data and often is close to the performance when combining D_e and D_o using FindIMB. This happens because in random graphs, the effect of variables inducing bias is often negligible [Greenland, 2003], and proxies of the unmeasured confounders are often included in the observed covariates. However, there are cases where the conditioning on an observed variable in observational data can produce heavily biased post-intervention probability estimates. A very simple example is the graph in Fig. 5, which we call the "m-bias" graph. To illustrate how m-bias can affect the prediction of Y_x from observational data, we simulated data from the m-bias graph with binary variables. We set Y, X and M to be noisy-AND functions of their parents with a parameter α. α has a monotonic relationship with the bias in estimating $P(Y|do(X = 1), M)$ using observational data: Larger α leads to a larger bias. We then varied alpha from 0.5 to 1, and we simulated D_o and D_e with 10,000 and 1000 samples, respectively. We used FindIMB, IMB and OMB to estimate $P(Y|do(X = 1), M)$ in a test data set (2000 samples). Fig. 6 shows the bias in the estimated parameter. We can see that while using D_o to estimate $P(Y|do(X = 1), M)$ leads to increasing bias, combining D_e and D_o can identify the situations where the parameter is not identifiable from observational data. We believe that noisy-AND types of distributions are not rare in biomedical data.

7 DISCUSSION

Our paper extends the concepts of Markov boundaries for predicting post-intervention distributions, and presents a method for learning such causal Markov boundaries from mixtures of observational and experimental data. The method introduced in this paper could be used in real-world applications like embedded clinical trials, where we have abundant observational and limited experimental data, to predict the most effective treatment for patients. In real-world settings, data often may include heterogeneous types of variables (i.e., a mixture of continuous, nominal, ordinal, etc.). Additionally, observational and experimental data may not measure exactly the same set of covariates, rather, data may contain overlapping covariates. Extending the method to non-singleton treatments and outcomes is another possible avenue for future research.

Acknowledgements

This research was funded in part by grant R01LM012011 from the National Institutes of Health.

References

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4002–4011. PMLR, 2020.

Derek C Angus. Fusing randomized trials with big data: The key to self-learning health care systems? *Journal of American Medical Association (JAMA)*, 314(8):767–768, 2015.

Derek C Angus, Scott Berry, Roger J Lewis, Farah Al-Beidh, Yaseen Arabi, Wilma van Bentum-Puijk, Zahra Bhimani, Marc Bonten, Kristine Broglio, Frank Brunkhorst, et al. The REMAP-CAP (randomized embedded multifactorial adaptive platform for community-acquired pneumonia) study rationale and design. *Annals of the American Thoracic Society*, 17(7):879–891, 2020.

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

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

David Heckerman and Dan Geiger. Learning Bayesian networks: A unification for discrete and Gaussian domains. In *Proceedings of the 11th conference on Uncertainty in Artificial Intelligence (UAI)*, pages 274–284, 1995.

David Heckerman, Dan Geiger, and David M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Constraint-based causal discovery: Conflict resolution with answer set programming. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 340–349, 2014.

Antti Hyttinen, Frederick Eberhardt, and Matti Järvisalo. Do-calculus when the true graph is unknown. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 395–404, 2015.

Amin Jaber, Jiji Zhang, and Elias Bareinboim. Causal identification under Markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 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 (NeurIPS)*, pages 10888–10897, 2018.

JM Mooij, S Magliacane, and T Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

J Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., 1988.

Jean-Philippe Pellet and André Elisseeff. Finding latent causes in causal networks: An efficient approach based on Markov blankets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1249–1256, 2008.

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. *Journal of Machine Learning Research*, 18(1):8132–8193, 2017.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: The fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International Journal of Data Science and Analytics*, 3(2):121–129, Mar 2017. ISSN 2364-4168. doi: 10.1007/s41060-016-0032-z. URL `https://doi.org/10.1007/s41060-016-0032-z`.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1): 145–157, 2003. ISSN 03036898.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*, 2019.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020. URL `http://jmlr.org/papers/v21/19-1026.html`.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1219–1226, 2006a.

Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 437–444, 2006b.

Ilya Shpitser, Tyler VanderWeele, and James M Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 527–536, 2010.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521*, 2020.

Jin Tian and Ilya Shpitser. On the identification of causal effects. Technical report, Cognitive Systems Laboratory, University of California at Los Angeles, 2003.

Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16(66):2147–2205, 2015.

Ioannis Tsamardinos and Constantin F Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. Citeseer, 2003.

Benito Van der Zander, Maciej Liskiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 11–24, 2014.

Janine Witte, Leonard Henckel, Marloes H Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

K. Yu, L. Liu, J. Li, and H. Chen. Mining Markov blankets without causal sufficiency. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6333–6347, 2018. doi: 10.1109/TNNLS.2018.2828982.

K. Yu, L. Liu, and J. Li. Learning Markov blankets from multiple interventional data sets. *IEEE Transactions on Neural Networks and Learning Systems*, 31(6):2005–2019, 2020. doi: 10.1109/TNNLS.2019.2927636.