Learning Causal Markov Boundaries with Mixed Observational and Experimental Data

Konstantina Lelova TEMP74@MATH.UOC.GR
Department of Mathematics and Applied Mathematics, University of Crete, Greece

Gregory F. Cooper GFC@PITT.EDU
Department of Biomedical Informatics, University of Pittsburgh, USA

Sofia Triantafillou SOF.TRIANTAFILLOU@UOC.GR
Department of Mathematics and Applied Mathematics, University of Crete, Greece

Editors: J.H.P. Kwisthout & S. Renooij

Abstract

A frequent goal in healthcare is to estimate personalized causal effects in order to select the best treatment for a patient from observational or experimental (RCT) data (or both), where "best" is defined in terms of maximizing the expectation of the desired outcome. The first task in estimating personalized effects is selecting the optimal set of personalization covariates (causal feature selection). This set of covariates is the Markov Boundary of the outcome in the experimental distribution, also known as the Interventional Markov Boundary (IMB), and can be identified from RCT data using methods for finding Markov Boundaries. However, most RCT data are very limited in sample size and do not work well with these methods. In this work, we develop methods that combine limited experimental and large observational data to identify the IMB, and improve the estimation of conditional (personalized) causal effects. These methods extend recent results (Triantafillou et al., 2021), which were limited to discrete data, to mixed data with binary and ordinal outcomes. The methods are based on Bayesian regression models. In simulated data, we show that our methods identify the correct IMB and improve causal effect estimation.

Keywords: Causal prediction, causal graphical models, Bayesian causal 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 (MB) of Y, MB(Y), and can be identified from data using statistical methods (Yu et al., 2018). Such methods are typically based on conditional independence tests and require large sample sizes. If the causal graph is known, the MB can be identified using graphical criteria (Pearl, 2000). For example, in a causal Bayesian network $\mathcal{G}$, the MB of a target variable Y is the set of parents, children, and spouses of Y in $\mathcal{G}$. 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 healthcare, we are often interested in predicting the outcome O *after* we intervene on a treatment T (post-intervention). For personalized post-intervention prediction, we also

want to condition on a set of pre-treatment covariates $\mathbf{V}$. If we have access to data from the post-intervention distribution (i.e., RCT data D_e), we can estimate the post-intervention distribution $P(O|do(T), \mathbf{V})$. To optimize our prediction[1], we ideally want to condition for the minimal set of covariates that provide maximal information of the post-intervention outcome. These correspond to the covariates in the MB of O in the *mutilated* causal graph $\mathcal{G}^T$. This graph has no incoming edges into T, and corresponds to the post-intervention distribution. To differentiate from the original MB of Y in the observational distribution, this set is called Interventional Markov Boundary (IMB) of O with respect to treatment T ($\text{IMB}_O(T)$). However, RCT data are typically limited in sample size, in which case IMBs might not be found reliably using MB-learning algorithms.

Unlike RCT data, observational data are often plentiful, but may be biased for causal estimation. Ideally, we would use D_o to estimate $P(O|do(T), \text{IMB}(O) \setminus T)$. However, if there are latent confounders, $P(O|do(T), \text{IMB}(O) \setminus T)$ may not be identifiable from observational data (e.g., in Fig 1, $P(O|do(T), Z)$ cannot be identified from observational data if C is unobserved). Triantafillou et al. (2021) showed that a minimal most informative set of covariates for which the conditional post-intervention distribution is identifiable from D_o is the set of covariates in the Causal Markov Boundary (CMB) of O with respect to T. For a causal Markov Boundary, $P(O|do(T), \text{CMB(O)} \setminus T) = P(O|T, \text{CMB}(O) \setminus T)$, so we can just use the observational data for estimation of the post-intervention distribution. CMBs can be identified using graphical criteria from a causal graph, but they are not unique, and cannot always be identified solely from observational data with statistical methods.

To summarize, the best predictive model for $O|do(T)$ should include the pre-treatment covariates in:

- The interventional MB $\text{IMB}_T(O)$ of the outcome with respect to the treatment, when the prediction is based on experimental data D_e.
- A causal MB $\text{CMB}_T(O)$ of the outcome with respect to the treatment, when the prediction is based only on observational data D_o.

A very common scenario in healthcare is the following: Researchers have access to large observational data D_o (e.g., Electronic Health Records) and small experimental data D_e (i.e., an RCT measuring the average treatment effect of T on O). The ground truth causal graph is unknown. The researchers want to predict the post-intervention outcome $P(O|do(T), \mathbf{V} = \mathbf{v})$ for a patient with a set of pre-treatment covariates $\mathbf{V} = \mathbf{v}$, to optimize their treatment assignment.

Notice that, by definition, $P(O|do(T), \mathbf{V} = \mathbf{v}) = P(O|do(T), \text{IMB}_T(O) \setminus T)$ [2]. Conditioning on the smaller set $\text{IMB}_T(O) \setminus T$ instead of all the covariates improves our estimator, so we ideally want to estimate $P(O|do(T), \text{IMB}_T(O) \setminus T)$ instead. However, $\text{IMB}_T(O)$ is unknown. The RCT data are not powered to identify $IMB_T(O)$ or estimate P correctly, and $P(O|do(T), \text{IMB}_T(O) \setminus T)$ may not be identifiable from D_o. We present a method for combining D_e and D_o to (a) identify the $\text{IMB}_T(O)$ (b) decide if $P(O|do(T), \text{IMB}_T(O) \setminus T)$ is identifiable from D_o and (c) get a Bayesian estimate for $P(O|do(T), \text{IMB}_T(O) \setminus T)$. Our method extends the algorithm (`FindIMB`) proposed in Triantafillou et al. (2021), which is limited to categorical data, where closed-form marginal likelihoods exist. Most applica-

1. get an unbiased, minimal variance estimator
2. To be consistent with the definitions of (Triantafillou et al., 2021), we include the treatment T in the IMB and the CMB and assume that there is a causal effect of T on O

tions in healthcare have mixed data, with the outcomes very commonly being binary (e.g., 30-day mortality) or ordinal (e.g., hospital-free days). *In this work, we use Bayesian regression models (Bayesian logistic and Bayesian ordinal regression) and approximate inference methods (MCMC sampling) to extend findIMB to ordinal and binary outcomes, binary treatments, and mixed covariates.*

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 a set of treatments 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. We argue that combining observational and experimental data can improve prediction of the most effective treatments for individual patients, than either type of data alone.

2. Methodology

2.1. 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 $O|do(T)$ to denote a variable O after the hard intervention on variable T. If we know the SMCM $\mathcal{G}$, a hard intervention in which a treatment T is set to t can be represented with the do-operator, $do(T{=}t)$. In the corresponding graph, this is equivalent to removing all incoming edges into T, while keeping all other mechanisms intact(denoted by $\mathcal{G}_{\overline{T}}$).

A large body of work focuses on identifying if a post-intervention distribution can be computed from observational probabilities (and hence, observational data). When the causal graph is known, Shpitser and Pearl (2006a,b) and Tian and Shpitser (2003) provide sound and complete identifiability results. These methods take as input a causal graph and a specific marginal or conditional post-intervention probability of interest p, and they return a formula to derive p using only observational probabilities, if the query is identifiable, and N/A otherwise.

2.2. Markov Boundaries

Observational Markov Boundary: A Markov blanket of a variable O in a set of variables $\mathbf{V}$ is a subset $\mathbf{Z}$ of $\mathbf{V}$ conditioned on which other variables are independent of O: $O \perp\!\!\!\perp \mathbf{V} \backslash \mathbf{Z} | \mathbf{Z}$. The Markov boundary of O is the Markov blanket that is also minimal (Pearl, 2000). For faithful distributions, the Markov boundary of a variable O is unique (Pearl, 1988). To distinguish from other types of Markov boundaries defined in this work, we use the terminology **observational Markov boundary (OMB)** to denote the Markov boundary of a variable. OMBs can be identified graphically for SMCMs(Richardson, 2003; Pellet and Elisseeff, 2008). 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). Knowing

the OMB allows a more parsimonious representation of the conditional distribution of O given $\mathbf{V}$, since $P(O|\mathbf{V}) = P(O|\text{MB}(O))$.

In this work, we are interested in the model that gives the optimal prediction of the post-intervention distribution, with the goal of optimizing treatment assignments. For this reason, we make the following two assumptions: (a) The set of covariates $\mathbf{V}$ only includes pre-treatment variables and (b) Treatment T has a causal effect on outcome O. The first assumption reflects the fact that the values of post-treatment covariates are unknown prior to the treatment, and cannot be used for deciding the optimal treatment. It also simplifies the expressions for the Markov Boundaries, because we no longer need to consider children of O and their districts. The second assumption is not really necessary, but the methods presented here are only of interest when it holds (if a treatment has no effect, there is no point in optimizing it).

Interventional Markov Boundary: Our goal is to identify the set of variables that lead to the optimal model for the post-intervention distribution of an outcome O relative to a specific treatment T. This set is the **interventional Markov boundary (IMB)** of O relative to T, denoted $\text{IMB}_T(O)$. Obviously, $\text{IMB}_T(O) \subseteq \text{MB(O)}$. 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 T. However, experimental data are often limited in sample sizes, while OMB identification methods typically rely on conditional independence tests and may require large sample sizes.

If we know the causal graph $\mathcal{G}$, the post-intervention distribution with respect to T is induced by the manipulated graph $\mathcal{G}_{\overline{T}}$. The IMB of O is then the OMB of O in $\mathcal{G}_{\overline{T}}$, and can be identified using the definition of the Markov boundary above. However, the post-intervention distribution $P(O|do(T), \text{IMB}_T(O) \setminus T)$, may not be identifiable from the observational distribution. We then want to answer the following question: *What is the best model for predicting $O|do(T)$ from the observational distribution?*

Causal Markov Boundary: The answer to the question above is the **Causal Markov Boundary (CMB)**. Intuitively, a CMB is a minimal set of covariates that are maximally informative for the post-intervention outcome, for which the post-intervention distribution of T is identifiable from observational distributions. Formally, a CMB is defined as follows:

Definition 1 *Let $\mathbf{Z} \subseteq (\mathbf{V} \cup T)$, and $\mathbf{Z}^* = \mathbf{Z} \setminus T$. Then $\mathbf{Z}$ is a causal Markov boundary (CMB) for O relative to T iff:*

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

Condition (1) ensures identifiability. Condition (2) ensures that the set is maximally informative: any additional covariate that is informative for the post-intervention outcome leads to non-identifiability. Condition (3) ensures minimality: No subset of a CMB is a CMB. Further details and examples of causal Markov Boundaries can be found in Triantafillou et al. (2021). A CMB is not necessarily unique; it is possible that multiple sets satisfy

Definition 1. When $\mathbf{V}$ includes only pre-treatment covariates, CMBs have been shown to satisfy the following properties:

1. CMBs satisfy the backdoor criterion.
2. CMBs, like IMBs, are subsets of the OMB.
3. If $\text{IMB}_T(O)$ is a CMB, then $\text{IMB}_T(O) = \text{MB}(O)$.

These results enable more efficient algorithms for finding CMBs, limiting the types of estimators and the number of variable sets that we need to consider.

2.3. Algorithm `FindIMB`.

`FindIMB` is an algorithm that takes as input observational and experimental data (D_o and D_e, respectively) and outputs a Bayesian prediction model for $O|do(T)$. The main idea is the following: The best prediction (asymptotically) for $O|do(T)$ is $P(O|do(T), \text{IMB}_T(O))$. When the IMB is not a CMB, we should not use D_o in the estimation of $P(O|do(T), \text{IMB}_T(O))$. But when the IMB is a CMB, we can and should use both D_e and D_o for this estimation. In fact, based on property (3) above, $P(O|do(T), \text{IMB}_T(O) \setminus T) = P(O|T, \text{CMB}_T(O) \setminus T) = P(O|do(T), \text{MB}(O))$.

This is expressed using binary variables, as follows: For every set $\mathbf{Z}$ which includes T, we can express the event in which $\mathbf{Z} = IMB_T(O)$ as the disjunction of two complementary binary variables: $H_{\mathbf{Z}} = H^c_{\mathbf{Z}} \vee H^{\bar{c}}_{\mathbf{Z}}$.

- $H^c_{\mathbf{Z}}$ is true if $\mathbf{Z} = \text{IMB}_T(O) = \text{CMB}_T(O)$, and false otherwise.
- $H^{\bar{c}}_{\mathbf{Z}}$ is true if $\mathbf{Z} = \text{IMB}_T(O) \neq \text{CMB}_T(O)$, and false otherwise.

Hence, when $H_{\mathbf{Z}}$ is true, $\mathbf{Z} \setminus T$ are the best conditioning covariates for predicting $O|do(T)$, and we therefore want to estimate $P(O|do(T), \mathbf{Z} \setminus T)$. If $H^c_{\mathbf{Z}}$ is true, we can use both D_e and D_o to estimate $P(O|do(T), \mathbf{Z} \setminus T)$. Otherwise, if $H^{\bar{c}}_{\mathbf{Z}}$ is true, we only use D_e.

The algorithm works as follows: For each possible set $\mathbf{Z}$ that includes T, we compute (a) the probability $P(H^{\bar{c}}_{\mathbf{Z}} | D_e, D_o)$ that $\mathbf{Z}$ is the IMB but not a CMB, and the corresponding estimator $\hat{P}(O|do(T), \mathbf{Z} \setminus T, D_e)$ and (b) the probability $P(H^c_{\mathbf{Z}} | D_e, D_o)$ that $\mathbf{Z}$ is the IMB and a CMB, and the corresponding estimator $\hat{P}(O|do(T), \mathbf{Z} \setminus T, D_e, D_o)$. In the end, the algorithm returns a Bayesian average of all the estimates, weighted by their probabilities $P(H^C_{\mathbf{Z}} | D_e, D_o), C = c, \bar{c}$.

The estimation of $P(H^c_{\mathbf{Z}} | D_e, D_o)$ is central to the method, as it quantifies our belief that $\mathbf{Z}$ is the IMB and (not) a CMB. It is based on the following equation:

$$P(H^c_{\mathbf{Z}} \mid D_e, D_o) = \frac{P(D_e \mid D_o, H^c_{\mathbf{Z}}) P(D_o \mid H^c_{\mathbf{Z}}) P(H^c_{\mathbf{Z}})}{\sum_{\mathbf{Z}} \sum_{C=c,\bar{c}} P(D_e \mid D_o, H^C_{\mathbf{Z}}) P(D_o \mid H^C_{\mathbf{Z}}) P(H^C_{\mathbf{Z}})} \tag{1}$$

We can similarly derive $P(H^{\bar{c}}_{\mathbf{Z}} \mid D_e, D_o)$ 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 the priors that $H^c_{\mathbf{Z}}$ and $H^{\bar{c}}_{\mathbf{Z}}$ hold, respectively, and can be set as uniform all $\mathbf{Z}$ and C (hence, $P(H^c_{\mathbf{Z}}) = P(H^{\bar{c}}_{\mathbf{Z}}) = 0.5$). $P(D_o \mid H^c_{\mathbf{Z}}), P(D_o \mid H^{\bar{c}}_{\mathbf{Z}})$ quantify how well the observational data fit with the hypotheses $H^c_{\mathbf{Z}}, H^{\bar{c}}_{\mathbf{Z}}$. Triantafillou et al. (2021) showed that, based on property (3) above, $P(D_o \mid H^{\bar{c}}_{\mathbf{Z}})$ is equal to the marginal likelihood of O in D_o, in a model that uses $\mathbf{Z}$ to predict O from the observational data. $P(D_o \mid H^c_{\mathbf{Z}})$ is the same marginal likelihood

Algorithm 1: FindIMB

input : D_o, D_e, treatment T, outcome O, pre-treatment covariates $\mathbf{V}$, number of samples N

output: Post-intervention distribution $P(O|do(T), \mathbf{V})$

```
MB(O) ← MarkovBoundary(O, D_o);
foreach subset Z of MB(O) and C = c, c̄ do
    foreach s =1 to N do
        Sample (θ_e^Z)^s from a non informative prior f(θ_e^Z) ;
        Compute the likelihood L^c̄(s) = P(D_e|(θ_e^Z)^s);
        Sample (θ_o^Z)^s from the observational posterior f(θ_o^Z|D_o) using MCMC;
        Compute the likelihood L^c(s) = P(D_e|(θ_o^Z)^s);
    end
    P(D_e|D_o, H_Z^c̄) ≈ Σ_s L^c̄(s)/N;
    P(D_e|D_o, H_Z^c) ≈ Σ_s L^c(s)/N;
    Estimate P(H_Z^C |D_e, D_o) using Eq. 1;
    Estimate P(O|do(T), V, D_e, D_o, H_Z^C ) using Eq. 9;
end
P(O|do(T), V) ← Σ_Z Σ_{C=c,c̄} P(O|do(T), V, D_e, D_o, H_Z^C )P(H_Z^C |D_e, D_o);
```

for the OMB, and zero for all subsets of the OMB. For the models we use in this work, this marginal likelihood cannot be computed in closed form, so we use a sampling-based approximation.

2.3.1. Estimating $P(D_e \mid D_o, H^c_{\mathbf{Z}}), P(D_e \mid D_o, H^{\bar{c}}_{\mathbf{Z}})$.

The core of the method is the computation of the probabilities $P(D_e \mid D_o, H^c_{\mathbf{Z}}), P(D_e \mid D_o, H^{\bar{c}}_{\mathbf{Z}})$. These probabilities quantify how likely we are to see the experimental data, given the observational data and the fact that set $\mathbf{Z}$ is (or not) the IMB and the CMB. The method builds two "priors" for predicting D_e: One based on the observational data D_o, and the other based on an uninformative prior. The idea is that the observational prior, which is quite strong due to the large sample size of D_o, will be a better prior for D_e only if $H^c_{\mathbf{Z}}$ is true. Otherwise, the weak prior will be better.

We use $\theta^{\mathbf{z}}_{\mathbf{e}}$ to denote the parameters of the interventional distributions $P(O \mid do(T), \mathbf{Z} \setminus T)$, and $\theta^{\mathbf{z}}_{\mathbf{o}}$ to denote the observational distributions $P(O \mid T, \mathbf{Z} \setminus T)$. Overall, for any given $\mathbf{Z}$ and $C = c, \bar{c}$, we can obtain $P\left(D_e \mid D_o, H^C_{\mathbf{Z}}\right)$ as the marginal likelihood, marginalizing over the experimental parameters:

$$P\left(D_e \mid D_o, H^C_{\mathbf{Z}}\right) = \int_{\theta^{\mathbf{z}}_{\mathbf{e}}} P\left(D_e \mid \theta^{\mathbf{z}}_{\mathbf{e}}\right) f\left(\theta^{\mathbf{z}}_{\mathbf{e}} \mid D_o, H^C_{\mathbf{Z}}\right) d\theta^{\mathbf{z}}_{\mathbf{e}}, \tag{2}$$

Under $H^c_{\mathbf{Z}}$, the observational and post-interventional distributions are the same, hence $\theta^{\mathbf{z}}_{\mathbf{e}}{=}\theta^{\mathbf{z}}_{\mathbf{o}}$ and $f\left(\theta^{\mathbf{z}}_{\mathbf{e}} \mid D_o, H^c_{\mathbf{Z}}\right) = f\left(\theta^{\mathbf{z}}_{\mathbf{o}} \mid D_o\right)$, i.e., the posterior of the observational parameters, based on observational data. In contrast, under $H^{\bar{c}}_{\mathbf{Z}}$, the post-intervention and observational

parameters are not the same, and the observational data may not be informative [3] for the post-intervention parameters, hence $f\left(\theta_{\mathbf{e}}^{\mathbf{z}} \mid D_o, H_{\mathbf{Z}}^{c}\right) = f\left(\theta_{\mathbf{e}}^{\mathbf{z}}\right)$.

Triantafillou et al. (2021) computes Eq. 2 in closed form for discrete data using Multinomial distributions with Dirichlet priors. In this work, we extend the method to mixed data, so no closed-form solution is available. In the sections below, we describe the statistical models and approximate inference methods we use.

Once we have computed Eq. 2, we can use it in Eq. 1 to estimate the probability that $\mathbf{Z}$ is an IMB and (not) a CMB. In the end, the algorithm returns a Bayesian weighted average over all sets $\mathbf{Z}$ and all $C = c, \bar{c}$. Obviously, computing over all subsets of a large set of covariates would not be feasible. Based on property (2) described above, we only need to look into the subsets of covariates in the OMB, which can be identified with an asymptotically correct method from the large observational data. This is implemented in line 1 of Algorithm 1.

2.3.2. Implementation for binary outcomes.

To model the relationship of a binary outcome to a binary treatment and mixed covariates, we use a Bayesian logistic regression model. Let O, T, $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$ be the outcome, treatment, and covariates of the model. Then for each patient i, $O_i \sim Bernoulli(\pi_i)$, with

$$\pi_i = \frac{e^{b_0 + b_1 Z_{i1} + \ldots + b_k Z_{ik} + b_{k+1} T_i}}{1 + e^{b_0 + b_1 Z_{i1} + \ldots + b_k Z_{ik} + b_{k+1} T_i}} \tag{3}$$

For N experimental samples, the **likelihood** of D_e given a set of parameters $\theta_{\mathbf{e}}^{\mathbf{z}} = (b_0, \ldots, b_{k+1})$ is:

$$P(D_e | \theta_{\mathbf{e}}^{\mathbf{z}}) = \prod_{i=1}^{N} \pi_i^{O_i} \left(1 - \pi_i\right)^{(1-O_i)} \tag{4}$$

where π can be computed for sample i using Eq. 3.

Our goal is to compute the marginal likelihood of D_e under the two complementary hypotheses $H_{\mathbf{Z}}^{c}$ and $H_{\mathbf{Z}}^{\bar{c}}$ (Eq.2).

Case 1: Computing the marginal likelihood $P(D_e | D_o, H_{\mathbf{Z}}^{\bar{c}})$. Under $H_{\mathbf{Z}}^{\bar{c}}$, the observational data are not informative for the post-intervention distribution, so Eq. 2 becomes

$$P\left(D_e \mid D_o, H_{\mathbf{Z}}^{C}\right) = \int_{\theta_{\mathbf{e}}^{\mathbf{z}}} P\left(D_e \mid \theta_{\mathbf{e}}^{\mathbf{z}}\right) f\left(\theta_{\mathbf{e}}^{\mathbf{z}}\right) d\theta_{\mathbf{e}}^{\mathbf{z}}. \tag{5}$$

Since we do not have information about $\theta_{\mathbf{e}}^{\mathbf{z}}$, we can simply use an uninformative prior $f(b_i)$ for each coefficient in $\theta_{\mathbf{e}}^{\mathbf{z}}$. Choosing uninformative priors for logistic regression is not straightforward. A popular choice for uninformative priors is a uniform distribution, or a Normal distribution with large variance (e.g., $\mathcal{N}(0, 1000)$.) However, these priors are not invariant under parameterization, so even if a prior is uninformative for π, it may not be uninformative for other parameters of interest, such as the log odds $log(\frac{\pi}{1-\pi})$ (Wasserman and Kass, 1996; Seaman and Stamey, 2012). We therefore used a Cauchy distribution as a weakly informative default prior distribution as proposed by Gelman et al. (2008). The authors propose

3. This claim is not absolutely accurate, since it is possible that we can derive bounds for $\theta_{\mathbf{e}}^{\mathbf{z}}$ based on $\theta_{\mathbf{o}}^{\mathbf{z}}$ with some additional assumptions; however, we cannot get derive estimates.

that this distribution is a good choice since actual effects fall within a limited range. For example, a typical change in an input variable would be unlikely to correspond to a change as large as 5 on the logistic scale (which would move the probability from 0.01 to 0.50 or from 0.50 to 0.99). Hence, we use $b_\nu \sim Cauchy(0, 2.5), \nu = 1, .., k+1, b_0 \sim Cauchy(0, 10)$ as weakly informative priors. This implies that we expect the success probability for an average case to be between 10^{-9} and $1 - 10^{-9}$ (for standardized covariates).

Eq. 5 cannot be computed in closed form, so we approximate it with a sampling sum. Specifically, we sample 1000 samples $\theta^{\mathbf{z}}_{\mathbf{e}} = (b_0, \dots b_{k+1})$ from the Cauchy prior, compute the likelihoods using Eq. 4, and then sum over them to get the marginal likelihood.

Case 2: Computing the marginal likelihood $P(D_e|D_o, H^c_{\mathbf{Z}})$**.** Under $H^c_{\mathbf{Z}}$, $\theta^{\mathbf{z}}_{\mathbf{e}} = \theta^{\mathbf{z}}_{\mathbf{o}}$ and Eq. 2 becomes

$$P\left(D_e \mid D_o, H^C_{\mathbf{Z}}\right) = \int_{\theta^{\mathbf{z}}_{\mathbf{o}}} P\left(D_e \mid \theta^{\mathbf{z}}_{\mathbf{o}}\right) f\left(\theta^{\mathbf{z}}_{\mathbf{o}}|D_o\right) d\theta^{\mathbf{z}}_{\mathbf{o}}. \tag{6}$$

$f\left(\theta^{\mathbf{z}}_{\mathbf{o}}|D_o\right)$ is the posterior of the observational parameters $\theta^{\mathbf{z}}_{\mathbf{o}} = \theta^{\mathbf{z}}_{\mathbf{e}} = (b_0, b_1, \dots, b_{k+1})$ given the observational data. Eq. 6 cannot be computed in closed form, so we approximate it with a sampling sum, but this time we sample 1000 samples from the posterior $f\left(\theta^{\mathbf{z}}_{\mathbf{o}}|D_o\right)$. We use the same Cauchy priors for the model parameters $\theta^{\mathbf{z}}_{\mathbf{o}}$, and sample from the posterior using Markov Chain Monte Carlo (MCMC) sampling. For each posterior sample $(b_0, \dots, b_{k+1}|D_o)$, we then compute the likelihood of the experimental data using Eq. 4. The marginal likelihood in Eq. 6 is then approximated as the average likelihood for all samples.

2.3.3. Implementation for ordinal outcomes.

For an ordinal outcome, binary treatment, and mixed covariates, we use a Bayesian ordinal regression model. Let O be the ordinal outcome with J ordered categories, T be the treatment, and $\mathbf{Z} = \{Z_1, \dots, Z_k\}$ be the covariates of the model. Then for each patient i, $O_i \sim Categorical(\boldsymbol{\pi_i})$, where $\boldsymbol{\pi_i} = (\pi_{i0}, \pi_{i1}, ..., \pi_{iJ})$ are the probabilities for each of the J ordered categories. The cumulative probability of the outcome being less than or equal to a specific category, j, is:

$$P(O_i \leq j) = \frac{e^{b_{j0} + b_1 Z_{i1} + ... + b_k Z_{ik} + b_{k+1} T_i}}{1 + e^{b_{j0} + b_1 Z_{i1} + ... + b_k Z_{ik} + b_{k+1} T_i}} \tag{7}$$

and the probability of the outcome being in a j category is: $\pi_{i0} = P(O_i \leq 0), \pi_{ij} = P(O_i \leq j) - P(O_i \leq j-1)$ *and* $\pi_{iJ} = 1 - P(O_i \leq J-1)$. The **proportional odds assumption** requires that slopes remain constant across categories, while intercepts may vary. For N experimental samples, the **likelihood** of D_e given parameters $\theta^{\mathbf{z}}_{\mathbf{e}} = (b_{j0}, \dots, b_{k+1})$ is:

$$P(D_e|\theta^{\mathbf{z}}_{\mathbf{e}}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \pi_{ij}^{[O_i = j]} \tag{8}$$

where $[O_i = j]$ evaluates to 1 if $O_i = j$, and 0 otherwise. π_{ij} can be computed for a sample i using Eq. 7. As with logistic regression, we use a sampling-based approach to compute the marginal likelihoods under the two competing hypotheses:

Case 1: Computing the marginal likelihood $P(D_e|D_o, H_{\mathbf{Z}}^{\bar{c}})$. Under $H_{\mathbf{Z}}^{\bar{c}}$, the observational data are not informative and we approximate Eq. 1 using uninformative priors. For the slopes, we use similar Cauchy priors described above $b_\nu \sim Cauchy(0, 2.5), \nu = 1, .., k+1$. For the intercepts, we use a Symmetric Dirichlet distribution with concentration parameter vector $\mathbf{a}$=1, which is essentially a uniform prior. We then follow the same sampling procedure described for binary outcomes.
Case 2: Computing the marginal likelihood $P(D_e|D_o, H_{\mathbf{Z}}^{c})$. Under $H_{\mathbf{Z}}^{c}$ we can approximate Eq. 6 by sampling from the posterior of $\theta_{\mathbf{o}}^{\mathbf{z}}$ given the observational data. We use the same Cauchy and Dirichlet priors for the model parameters $\theta_{\mathbf{o}}^{\mathbf{z}}$, and follow the same procedure described for binary outcomes.

2.3.4. Estimating $P\left(H_{\mathbf{Z}}^{C} \mid D_e, D_o, \right)$.

Once we have estimated the marginal likelihoods of the experimental data using both the weak prior and the strong observational prior, we can use Eq. 1 to estimate $P(H_{\mathbf{Z}}^{\bar{c}} \mid D_e, D_o)$ and $P(H_{\mathbf{Z}}^{c} \mid D_e, D_o)$. These probabilities tell how likely it is that $\mathbf{Z}$ is an IMB, and if we can include observational data in the estimation of $P(O \mid do(T), \mathbf{V})$.

2.3.5. Estimating $P(O|do(T), \mathbf{V}, D_e, D_o)$.

Having quantified the probability that each subset of the OMB is an IMB and a CMB, we can average overall all subsets $\mathbf{Z}$ and hypotheses, $H_{\mathbf{Z}}^{C}$, to compute $P(O|do(T), \boldsymbol{V})$. Let t,o, $\mathbf{V}$=$\mathbf{v}$, denote given instances of T, O, and $\mathbf{V}$, respectively. When $\mathbf{V}$= $\mathbf{v}$, we use $\boldsymbol{Z} = \boldsymbol{z_v}$ to denote the corresponding values of the variables of $\mathbf{Z}$ that are in $\mathbf{V}$. Under both $H_{\mathbf{Z}}^{c}$ and $H_{\mathbf{Z}}^{\bar{c}}$, $P(O|do(T), \boldsymbol{V}) = P(O|do(T), \mathbf{Z} \setminus T)$. Hence, for an instance of $\mathbf{V}$=$\mathbf{v}$, we have:

$$P(o \mid do(t), \mathbf{v}, De, Do) = \sum_{\boldsymbol{Z} \subset \boldsymbol{V}} \sum_{C=c,\bar{c}} P(o \mid do(t), \boldsymbol{z_v} \setminus t, D_e, D_o, H_{\mathbf{Z}}^{C}) P(H_{\mathbf{Z}}^{C} \mid D_e, D_o). \quad (9)$$

The individual probabilities $P(o \mid do(t), \boldsymbol{z_v} \setminus t, D_e, D_o, H_{\mathbf{Z}}^{C})$ can be estimated as posterior expectations of $P(O \mid do(T), \boldsymbol{Z})$ from the data. Specifically, under $H_{\mathbf{Z}}^{c}$, we use the empirical posterior expectation of the parameters $E(\theta_{\mathbf{e}}^{\mathbf{z}}|D_e, D_o)$ using both experimental and observational data in our computation. In contrast, under $H_{\mathbf{Z}}^{\bar{c}}$, we use the empirical posterior expectation of the parameters $E(\theta_{\mathbf{e}}^{\mathbf{z}}|D_e)$ using only experimental data.

3. Related Work

Causal and Interventional Markov Boundaries were introduced in Triantafillou et al. (2021). We extend their work for mixed data, using Bayesian regression models and MCMC inference. We briefly present other works that estimate post-intervention distributions in the presence of latent confounders. **Markov boundaries:** Several algorithms learn OMBs from data under causal insufficiency (Yu et al., 2018, 2020). These methods can be used to identify the IMBs from the experimental data, but require large sample sizes. **Identifiability:** When the causal graph is unknown, Hyttinen et al. (2015) and Jaber et al. (2019) provide identifiability results 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 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) is closest to `FindIMB` (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. None of these methods are implemented for mixed data.
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 T on O. 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). 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 method assumes that the hidden confounding has a linear parametric structure, and is implemented for continuous variables. Extending it to binary/ordinal outcomes is not straightforward.

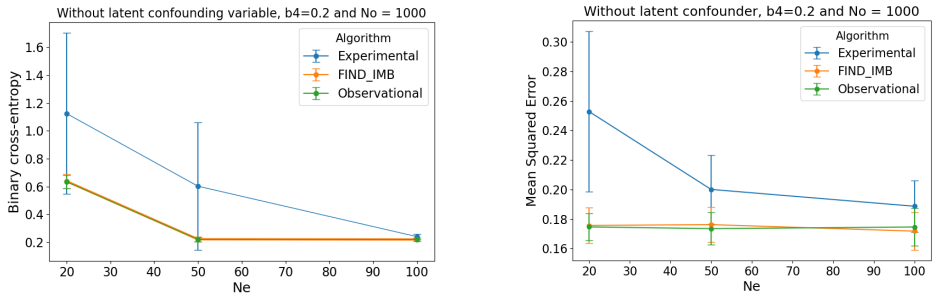
4. Experiments

In this section, we show that `FindIMB` can improve causal effect estimation in the case of mixed data, using Bayesian regression models and approximate inference. Specifically, we show that `FindIMB` can detect the presence of latent confounders, and can improve causal effect estimation by including observational data when the added bias is small. The implementation codes are available at https://github.com/n-magot/PGM_2024.

Figure 1: Causal graphs for the experimental (left) and observational (right) data, showing the causal structure among treatment T, outcome O, and pre-treatment covariates Z, C.

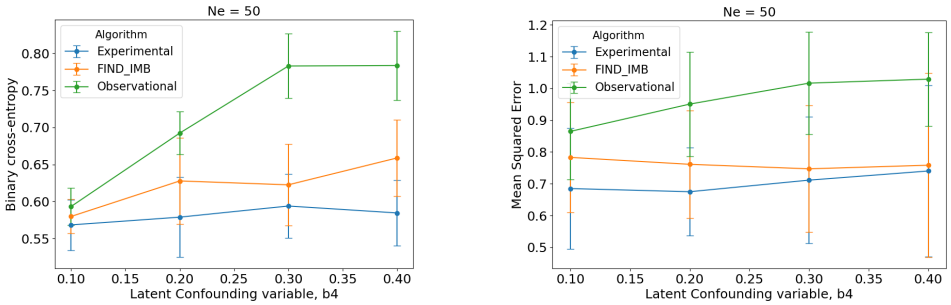
We compared `FindIMB` to the following approaches: (a) **Experimental**: using only experimental data and the ground-truth IMB. We use the posterior expectation of $P(O|do(T), \text{IMB}_T(O) \setminus T, D_e)$ as the estimator for $P(O|do(T), \mathbf{V})$. This estimator is always unbiased, but has a large variance due to low sample sizes. (b) **Observational**: using only observational data and the ground-truth IMB. We use the posterior expectation of $P(O|do(T), \text{IMB}_T(O) \setminus T, D_o)$ as the estimator for $P(O|do(T), \mathbf{V})$. Other methods using only experimental or only observational data would *at best* perform as (a) and (b) above, respectively. Other methods that combine observational and experimental data are not currently implemented for mixed data.

To illustrate the behavior of `FindIMB`, we used two scenarios: One with latent confounder, and one without. We simulated D_o from the DAG on Fig. 1 (right), and D_e from the corresponding manipulated DAG, shown in Fig. 1 (left). D_o is much larger in size than D_e, and the coefficients in the regression models are the same for both data sets. We simulated N_e experimental samples and 1000 observational samples. We used constant coefficients $(b_1, b_2, b_3, b_5) = (1.5, 0.2, 0.3, 0.4)$. We used $b_4 = 0.2$ for Scenario 1, and varied it to control confounding effect in Scenario 2. We evaluate our methods in how well they predict the outcome in a new experimental data set with 1000 samples, using binary cross-entropy for binary outcomes and root mean squared error for ordinal outcomes.



(a) Binary outcome (b) Ordinal outcome

Figure 2: Comparative performance for `FindIMB` and using only observational or experimental data, when there is no latent confounder, for increasing experimental sample size. y-axis measures prediction error (lower is better). `FindIMB` performs on par with observational data. Experimental data perform worse for small sample sizes.



(a) Binary outcome (b) Ordinal outcome

Figure 3: Comparative performance for `FindIMB`, observational and experimental, when there is a latent confounder, for increasing confounding (coefficient b_4). y-axis measures prediction error (lower is better). `FindIMB` performs similar to experimental data for most cases; observational data perform worse as confounding increases.

Scenario 1: No latent confounding: In this scenario, D_o and D_e include Z and C. Due to a much larger sample size and the lack of confounding, using D_o should improve causal prediction of the outcome, compared to $\mathbf{D_e}$. Results are shown in Fig. 2. `FindIMB` performs on par with the observational data, as expected, while the experimental data perform much worse for small sample sizes. **Scenario 2: Latent confounding:** In this scenario, C is unmeasured. By varying the coefficient of $C \to T$, we examine the behavior of `FindIMB` for increasing confounding. In this case, using D_o should lead to biased causal prediction, and using D_e only is preferable. Results for 20 repetitions can be shown in Fig. 3. Observational data lead to worse causal prediction of the outcome, compared to the experimental. The `FindIMB` performs only slightly worse than the experimental data, and both outperform the observational data. In summary, `FindIMB` can successfully identify if the IMB is also the CMB, and improve causal estimation when possible.

5. Conclusions

We present an extension of `FindIMB`, an algorithm that combines observational and experimental data to learn interventional Markov boundaries and improve causal estimation. Using Bayesian regression models and approximate inference, we show that the method improves causal estimation for ordinal/binary outcomes and mixed data. In the future, we plan to explore greedy strategies for scaling up the method to allow for more conditioning covariates, and explore non-parametric Bayesian approaches to increase model flexibility.

Acknowledgments

This work was supported by grant R01HL164835 (Individualized Prediction of Treatment Effects Using Data from Both Embedded Clinical Trials and Electronic Health Records) from the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health (NIH). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

B. Andrews, P. Spirtes, and G. 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.

D. 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.

D. C. Angus, S. Berry, R. J. Lewis, F. Al-Beidh, Y. Arabi, W. van Bentum-Puijk, Z. Bhimani, M. Bonten, K. Broglio, F. 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.

A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4), Dec. 2008. ISSN 1932-6157. doi: 10.1214/08-AOAS191. URL https://projecteuclid.org/journals/annals-of-applied-statistics/volume-2/issue-4/A-weakly-informative-default-prior-distribution-for-logistic-and-other/10.1214/08-AOAS191.full.

A. Hyttinen, F. Eberhardt, and M. 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.

A. Hyttinen, F. Eberhardt, and M. 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.

A. Jaber, J. Zhang, and E. Bareinboim. Causal identification under Markov equivalence: Completeness results. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2981–2989, 2019.

N. Kallus, A. M. Puli, and U. Shalit. Removing hidden confounding by experimental grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10888–10897, 2018.

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

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

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

J.-P. Pellet and A. 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.

E. Perkovic, J. Textor, M. Kalisch, and M. 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.

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

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

A. Rotnitzky and E. 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.

J. W. Seaman and J. D. Stamey. Hidden Dangers of Specifying Noninformative Priors. *The American Statistician*, 66(2):77–84, 2012. ISSN 0003-1305. URL https://www.jstor.org/stable/23339464. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

I. Shpitser and J. 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.

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

E. Smucler, F. Sapienza, and A. Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521*, 2020.

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

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

S. Triantafillou, F. Jabbari, and G. F. Cooper. Causal and interventional markov boundaries. In C. de Campos and M. H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 1434–1443. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/triantafillou21a.html.

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

L. Wasserman and R. E. Kass. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association*, 1996.

J. Witte, L. Henckel, M. H. Maathuis, and V. 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.