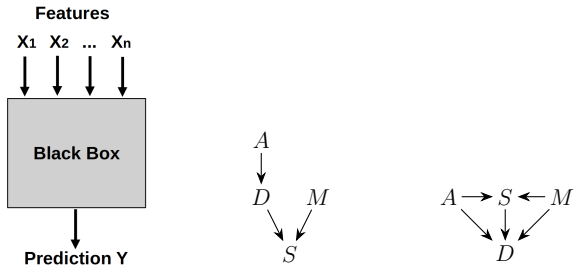
Modeling and Discovering Direct Causes for Predictive Models

Yizuo Chen[12*], **Amit Bhatia**[2†]

[1]University of California, Los Angeles, USA
[2]RTX Technology Research Center, Berkeley, USA

Abstract

We introduce a causal modeling framework that captures the input-output behavior of predictive models (e.g., machine learning models) by representing it using causal graphs. The framework enables us to define and identify features that *directly cause* the predictions, which has broad implications for data collection and model evaluation. We show two assumptions under which the direct causes can be discovered from data, one of which further simplifies the discovery process. In addition to providing sound and complete algorithms, we propose an optimization technique based on an independence rule that can be integrated with the algorithms to speed up the discovery process both theoretically and empirically.



(a) setup (b) G (c) G'

Figure 1: The setup and causal graph for a predictive model.

1 Introduction

Predictive models have become increasingly prevalent in decision-making over the past few decades. In essence, a predictive model is a function that maps a set of features (often available from data) to a set of outcomes; see, e.g., (MacKenzie 2013; Neilson et al. 2019; Ellis 2012). For instance, a predictive model can be used to forecast the future weather based on data from the past ten days. Machine learning models are a common class of predictive models whose parameters are learned from data, e.g., support vector machines (Cortes and Vapnik 1995), decision trees (Breiman et al. 1984), and more recently, neural networks (Bishop 1995; Goodfellow, Bengio, and Courville 2016). Other types of predictive models that do not involve machine learning include statistical models such as linear regression (Freedman 2005), rule-based expert systems (Buchanan and Shortliffe 1984), and probabilistic models constructed from domain knowledge (Pearl 1988; Darwiche 2009).

In this work, we consider a setup (in Figure 1a) where the predictive models are treated as "black boxes", meaning that their behaviors are not interpretable by humans. This happens, for instance, when the model parameters are not publicly available or when the models (e.g., deep neural networks) are too complex to be transparent; see, e.g., (Lipton 2018; Caruana et al. 2015; Lada Kohoutová et al. 2020). To model the input-output behavior of predictive models under this setup, we introduce a class of causal graphs that represent the predictive models as causal mechanisms.[1] This type of modeling appears to be different from the conventional approach yet effectively captures the data-generating process of the predictions. To illustrate the subtlety, consider an example where a predictive model is used to predict a patient's Disease (D) based on their Age (A), Symptom (S), and Medication (M). Without bearing in mind that D is a prediction from $\{A, S, M\}$, one may model the interactions among variables using the graph G shown in Figure 1b. However, G does not capture the data generating process of D, and it becomes erroneous if we use G to answer causal queries. For example, we would falsely conclude that an intervention on Symptom (S) does not affect the predictions on Disease (D). Instead, Figure 1c depicts the true causal graph in which the predictive model is converted into a causal mechanism for D. As we will see later, this conversion technique can be applied to model all predictive models.

Once we represented the predictive models as graphs, the *direct causes* for predictions on the outcome Y become exactly the parents of Y in the causal graph. Understanding the direct causes for model predictions has a wide range of applications. First, it provides insights into which features contribute to the predictions, which has vast implica-

*Email: yizuo.chen@ucla.edu. This work was done during the author's internship at RTX Technology Research Center.

†Email: amit.bhatia2@rtx.com.

[1]The idea of treating machine learning models as causal mechanisms was mentioned briefly in (Darwiche 2020). This work formalizes the idea and studies the problem of discovering direct causes based on the formulation. We allow causal mechanisms to exhibit uncertainties in this work.

tions for model explainability and fairness; see, e.g., (Ali et al. 2023; Ribeiro, Singh, and Guestrin 2016; Darwiche and Hirth 2020; Barocas, Hardt, and Narayanan 2023; Zafar et al. 2017). Second, identifying features that do not directly cause the predictions allows us avoid unnecessary data collections, which reduces the cost on data acquisition; see, e.g., (Coffey and Elliott 2023; Trask et al. 2012). Our main question becomes: how can we discover these direct causes from data (on the features and outcome)? To answer this question, we first propose two assumptions on the data distribution that ensure the direct causes are discoverable (uniquely determined). Under both assumptions, the direct causes form a *Markov boundary* of the outcome — a notion introduced in (Pearl 1988) that has been studied extensively since then. This immediately offers us sound and complete methods for discovering direct causes based on leveraging the existing algorithms for discovering Markov boundaries. One of these assumptions also simplifies the procedure of discovering direct causes, which improves the computational efficiency of the discovery process. Moreover, we introduce an independence rule that can be integrated with the existing algorithms to further speed up the discovery process as we demonstrate both theoretically and empirically.

The paper is structured as follows. We start with some technical preliminaries in Section 2. In Section 3 we introduce the causal modeling for predictive models and formally define the notion of direct causes in this context. Section 4 provides two assumptions under which direct causes can be discovered from data, along with algorithms for discovering these direct causes. We then show an independence rule that can be integrated into the discovery algorithm to further improve the efficiency in Section 5. Section 6 presents empirical results to demonstrate the effectiveness of the independence rule. We close with some concluding remarks in Section 7. All proofs are included in the Appendix.

2 Technical Preliminaries

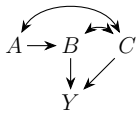
We assume all variables are discrete, though all the results can be extended to continuous domains. Single variables are denoted by uppercase letters (e.g., X) and their states are denoted by lowercase letters (e.g., x). Sets of variables are denoted by bold, uppercase letters (e.g., $\mathbf{X}$) and their instantiations are denoted by bold, lowercase letters (e.g., $\mathbf{x}$).

2.1 Causal Models and Interventions

In this work, we consider causal graphs in the form of acyclic directed mixed graphs (ADMGs) (Richardson 2003).

Definition 1. *An acyclic directed mixed graph (ADMG) is a graph that contains directed edges ($\rightarrow$) and bidirected edges ($\leftrightarrow$) and in which directed edges do not form any cycles.*

Figure 2a depicts an ADMG over four variables. Let X, Y be two variables in an ADMG, we say that X is a *parent* of Y, and Y a *child* of X if $X \rightarrow Y$. Moreover, we say that X is an *ancestor* of Y, and Y a *descendant* of X if there is a directed path from X to Y. We say that X is a *sibling* of Y if $X \leftrightarrow Y$, and a *spouse* of Y if X and Y share a same child. We say that X is a *neighbor* of Y if it is a parent, child, or sibling of Y. A variable V is called a *collider* on a path if $\rightarrow V \leftarrow$, $\leftrightarrow V \leftrightarrow$, $\rightarrow V \leftrightarrow$, or $\leftrightarrow V \leftarrow$ appears on the path and is called a *non-collider* otherwise.[2]



(a) ADMG G

A	B	C	Y
1	0	1	1
0	2	3	0
0	1	5	0
1	2	0	1
...	...	...	...

(b) data

Figure 2: A causal graph and empirical data.

Intervention is a standard technique for studying the causal relations among events. By definition, an intervention fixes a variable to a specific state, which is different from naturally observing the state of a variable. For example, instructing (intervening) a patient to take a drug yields a different effect than seeing (observing) a patient taking a drug. We write $do(X = x)$, or simply $do(x)$, if an intervention fixes a variable X to the state x. A variable X has a *causal effect* on variable Y if an intervention on X changes the distribution of Y. The causal effect can only happen if X is an ancestor of Y in the causal graph (Pearl 2009).

2.2 Independences in Graphs and Distributions

(Conditional) independence is a central notion in the domain of causal inference and discovery. In fact, the goal of discovery is to identify causal graphs consistent with the independencies present in a given data distribution. We next review the definitions of independence for both causal graphs and distributions and discuss the interplay between the two.

The independence relations in a causal graph (ADMG) are characterized by the notion of *m-separation* (Richardson 2003). By definition, let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint variables sets in an ADMG G, $\mathbf{X}$ and $\mathbf{Y}$ are said to be m-separated by $\mathbf{Z}$, denoted $\text{msep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$, iff every path between $\mathbf{X}$ and $\mathbf{Y}$ satisfy the following property: (1) all the non-colliders on the path are in $\mathbf{Z}$; and (2) none of the colliders on the path is an ancestor of $\mathbf{Z}$. In Figure 2a, A and Y are m-separated by $\{B, C\}$ but are not m-separated by $\{B\}$.

Let $\Pr$ be a distribution and $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint variable sets, we say that $\mathbf{X}$ and $\mathbf{Y}$ are independent conditioned on $\mathbf{Z}$ if $\Pr(\mathbf{x}|\mathbf{y}, \mathbf{z}) = \Pr(\mathbf{x}|\mathbf{z})$ for all instantiations $\mathbf{x}, \mathbf{y}, \mathbf{z}$. We adopt the notations in (Darwiche 2009) and write $\mathcal{I}_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ if the independence relation holds and $\overline{\mathcal{I}_{\Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ otherwise. In practice, the distribution $\Pr$ is often provided in the form of data as shown in Figure 2b. Popular methods for testing independences from data include χ^2-test (Pearson 1900) and G-test (Sokal and Rohlf 2013). These independence tests, however, suffer from two bottlenecks as pointed out in (Spirtes, Glymour, and Scheines 2000, Ch. 5). The first is *computational inefficiency*, which occurs when independence tests are overused

[2] A directed path from X to Y is a path in the form of $X \rightarrow \cdots \rightarrow Y$. Each $X \leftrightarrow Y$ represents a hidden confounder U, i.e., $X \leftarrow U \rightarrow Y$.

since the time required by each independence test is at least linear in the sample size. The second is *sample inefficiency*, which occurs when we test independence with a large conditioned set $\mathbf{Z}$, since the number of samples required for testing independence stably is exponential in $|\mathbf{Z}|$.[3]

The notions of m-separations and independencies are connected with the following concepts of independence map (I-MAP), dependency map (D-MAP), and perfect map (P-MAP) as defined in (Pearl 1988; Darwiche 2009).

Definition 2. *Let G be a causal graph and $\Pr$ be a distribution over a same set of variables. We say that G is an I-MAP of $\Pr$ iff $\text{msep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ implies $\mathcal{I}_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ (for all $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$); G is a D-MAP of $\Pr$ iff $\mathcal{I}_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ implies $\text{msep}_G(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$; and G is a P-MAP of $\Pr$ iff G is both an I-MAP and a D-MAP of $\Pr$.*

We may sometimes say "$\Pr$ is an I-MAP of G" to mean that "G is an I-MAP of $\Pr$", similarly for D-MAP and P-MAP. D-MAP is also known as *faithfulness* in the causal discovery literature. The notion of P-MAP is usually required by the existing causal discovery algorithms (such as PC (Spirtes, Glymour, and Scheines 2000), FCI (Spirtes, Glymour, and Scheines 2000), etc.) to ensure that the true causal graph can be discovered from data.

2.3 Markov Boundary

As we will see later, the discovery of direct causes for model predictions can be reduced to the discovery of Markov boundary in some scenarios. Therefore, we also review the notion of Markov boundary along with some discovery algorithms in this section. We start with the definition of Markov boundary in (Pearl 1988) with a slight rephrasing.

Definition 3. *Let $\Pr$ be a distribution over variables $\mathbf{X}, Y$. The Markov boundary for Y, denoted MB(Y), is the minimal subset of $\mathbf{X}$ such that $\mathcal{I}_{\Pr}(Y, \text{MB}(Y), \mathbf{X} \setminus \text{MB}(Y))$.*[4]

That is, Y is independent of other features when conditioned on its Markov boundary. Suppose a causal graph G is a P-MAP of the distribution $\Pr$, then the Markov boundary of Y is unique and is equivalent to the *Markov blanket* of Y in G (Pearl 1988); see Appendix A for a review of the formal definitions and discovery algorithms for Markov blankets.

One key subroutine (procedure) used extensively by most Markov blanket discovery algorithms is the *adjacency search*, which identifies the neighbors of Y in the causal graph G. The procedure is based on the following observation: variables X, Y are adjacent to each other in G iff they are always dependent in $\Pr$ regardless of the conditioned variables. To check whether two variables are adjacent, the adjacency search algorithm enumerates all possible conditioned sets $\mathbf{Z} \subseteq \mathbf{X}$ with an increasing size and removes a variable X from the neighbors of Y if $\mathcal{I}_{\Pr}(X, \mathbf{Z}, Y)$. Consider the causal graph G in Figure 2a that is a P-MAP of some distribution $\Pr$. The adjacency search algorithm initializes all features $\{A, B, C\}$ to be the neighbors of Y. It then starts enumerating the conditioned sets $\mathbf{Z}$ with an increasing size. When $\mathbf{Z} = \{B, C\}$, we find that $\mathcal{I}_{\Pr}(A, \mathbf{Z}, Y)$ and therefore remove A from the neighbors of Y. The algorithm finally concludes that the neighbors of Y are $\{B, C\}$ after it enumerates all feasible conditioned sets.

In the worst case, the number of independence tests required by adjacency search is exponential in the number of variables. As we will see later, one main focus of this paper is to optimize the efficiency of the adjacency search, which is key to speeding up the discovery of direct causes.

3 Causal Modeling for Predictive Models

We present a class of causal graphs called *predictive graphs* to represent the input-output behavior of predictive models. Given a predictive model that takes a set of input features $\mathbf{X}$ and predicts an outcome Y, we construct a predictive graph that satisfies the following constraints: (1) Y cannot be a cause of any $X \in \mathbf{X}$; and (2) there is no hidden confounder between a feature X and Y. These constraints follow naturally from the data generating process of Y: intervening on predictions can never modify the input features, and the only possible causal factors for the predictions are the input features. We formally define the notion of predictive graphs.

Definition 4. *Let $\mathbf{X}$ be a set of features and Y be an outcome. A predictive graph is an ADMG over $\mathbf{X}, Y$ where the only possible edge between $X \in \mathbf{X}$ and Y is $X \to Y$.*

We will use $G(\mathbf{X}, Y)$ to denote a predictive graph wrt features $\mathbf{X}$ and outcome Y. Figure 1 depicts a predictive graph $G(\{A, S, M\}, D)$. One key observation is that the predictive model is translated into the causal mechanism for Y in the predictive graph; that is, the causal mechanism (which involves Y and its parents) captures the input-output behavior of the predictive model. From now on, we shall assume that the data distribution $\Pr(\mathbf{X}, Y)$ from a predictive model is always induced by some *true* predictive graph $G(\mathbf{X}, Y)$.[5]

In a predictive graph, the parents of outcome Y are exactly the *direct causes* of the mode predictions. In practice, however, the predictive graph is often not available and we do not know the direct causes. Hence, our goal is to discover the direct causes from data. This leads to two key questions: (1) when are the direct causes discoverable (uniquely determined)? (2) how can we find these direct causes efficently if they are indeed discoverable? Before addressing these questions, we consider the formal definition of direct causes from (Woodward 2004) with a rephrasing.

Definition 5. *A variable X is a direct cause of Y if $\Pr(Y|do(x), do(\mathbf{x}')) \neq \Pr(Y|do(\mathbf{x}'))$ for some state x of X and instantiation $\mathbf{x}'$ of $\mathbf{X} \setminus \{X\}$.*

That is, variable X is a direct cause of Y iff an intervention on X affects the distribution of Y while fixing the states of other variables. The definition suggests that discovering

[3] To illustrate, suppose $|\mathbf{Z}| = 100$, then there are 2^{100} distinct instantiations over $\mathbf{Z}$ so we need at least 2^{100} samples to ensure that each instantiation appears at least once.

[4] Minimal subset here means that no proper subset of MB(Y) satisfies the property.

[5] A distribution $\Pr$ is induced by some causal graph G iff it is generated by some parameterizations of G. Moreover, G is guaranteed to be an I-MAP of $\Pr$ but may not be a D-MAP of $\Pr$.

direct causes requires conducting interventions, which cannot be inferred from observational studies in general.[6] However, when the distribution is induced by a predictive graph, the direct causes can actually be defined without interventions as we show next.

Proposition 6. *Let $G(\mathbf{X}, Y)$ be a predictive graph that induces a distribution* $\Pr$ *where* $\Pr(\mathbf{X}) > 0$.[7] *Then $X \in \mathbf{X}$ is a direct cause of Y by Definition 5 iff $\overline{\mathcal{I}_{\Pr}}(X, \mathbf{X} \setminus \{X\}, Y)$.*

We shall assume that the true predictive graph is consistent with Definition 5 so the parents of Y in the true predictive graph satisfy the condition in Proposition 6. From the proposition, we immediately attain a method for discovering direct causes by simply checking whether $\overline{\mathcal{I}_{\Pr}}(X, \mathbf{X} \setminus \{X\}, Y)$ for each feature X. This method, however, is not sample-efficient since the conditioned set may become extremely large when we have more features; see our earlier discussion about sample efficiency in Section 2.2. We next propose some assumptions under which the direct causes can be discovered with sample-efficient algorithms.

4 Assumptions for Discovering Direct Causes

We propose two assumptions under which the direct causes of the predictions are discoverable. In both cases, we show that the direct causes become equivalent to the Markov boundary (Definition 3) so we can leverage methods for discovering Markov boundaries for discovering direct causes.

4.1 Canonicalness

We start with the first assumption called *canonicalness*, which is commonly assumed by existing algorithms for discovering Markov blankets.

Definition 7. *A distribution* $\Pr$ *is said to be canonical if it is a P-MAP of some causal graph G.*

Note that the causal graph G in Definition 7 may not be the true predictive graph; in fact, G can be any ADMG, which makes the assumption quite general. The following result shows that the direct causes are discoverable when the data distribution is canonical.

Theorem 8. *Let $G(\mathbf{X}, Y)$ be a predictive graph that induces a canonical distribution* $\Pr$ *. Then the direct causes of Y in G form a unique Markov boundary of Y in* $\Pr$ *.*

Let G be the causal graph that is a P-MAP of the distribution $\Pr$, then the Markov boundary for Y in $\Pr$ is exactly the Markov blanket of Y in G. That is, the problem of discovering direct causes in a predictive graph can be reduced to to the problem of discovering the Markov blanket when the given distribution is canonical. To illustrate, suppose a distribution $\Pr$ induced by a predictive graph G is a P-MAP of the causal graph G' in Figure 3a, then the set of direct causes of Y in G is $\{A, B, C, D, E, F\}$, which is exactly the Markov blanket of Y in G'. This result provides a method for discovering direct causes that is based on adopting the existing algorithms for discovering Markov blankets under ADMGs such as the M3B algorithm (Yu et al. 2018).

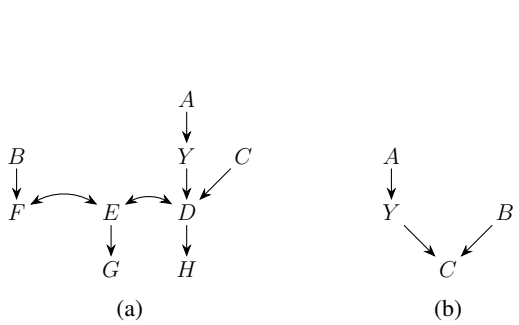


Figure 3: Causal graphs for different assumptions.

4.2 Weak Faithfulness

Our second assumption is a weaker type of faithfulness that imposes constraints on the distributions induced by the true predictive graph. As we will show later, the assumption not only makes the direct causes discoverable but also leads to an improvement on the computational efficiency.

Definition 9. *Consider a predictive graph $G(\mathbf{X}, Y)$. A distribution* $\Pr$ *is weakly faithful wrt G if $X \in \mathbf{X}$ is a parent of Y in G only if $\overline{\mathcal{I}_{\Pr}}(X, \mathbf{Z}, Y)$ for all $\mathbf{Z} \subseteq \mathbf{X} \setminus \{X\}$.*[8]

In practice, the weak faithfulness requires that the model predictions always depend on the direct causes regardless of the conditioned set on other features. This assumption is likely to hold, for instance, when the predictive model is a linear or polynomial regressor. To see an example where the assumption is violated, consider a canonical distribution $\Pr$ that is a P-MAP of the causal graph in Figure 3b. $\Pr$ is not weakly faithful since $\mathcal{I}_{\Pr}(Y, A, B)$ even though B is a direct cause of Y by Theorem 8. The following result states that the direct causes are discoverable under weak faithfulness.

Theorem 10. *Let $G(\mathbf{X}, Y)$ be a predictive graph and* $\Pr$ *be a distribution that is weakly faithful wrt G. The direct causes of Y in G form a unique Markov boundary of Y in* $\Pr$ *.*

Another advantage brought by the weak faithfulness assumption is that it enables a faster discovery of direct causes when compared to the existing Markov *blanket* discovery algorithms. First, the direct causes of Y coincide with the definition of neighbors of Y under the weak faithfulness. Hence, all the direct causes can be found through a single adjacency search for Y, which avoids the additional independence tests for discovering non-neighbor variables (e.g., spouses) as required by Markov blanket discovery. Second, since $\Pr$ is induced by a predictive graph, Y is independent of all other features conditioned on its parents by Markov assumption. This allows us skip the "symmetry correction" step (Tsamardinos, Brown, and Aliferis 2006) in adjacency search, which further simplifies the discovery process.[9]

[6] See (Pearl and Mackenzie 2018) for a discussion on different layers of causal hierarchy.

[7] This positivity assumption ensures that $\Pr(Y|\mathbf{X})$ is well-defined.

[8] Weak faithfulness is a type of *adjacency faithfulness* in (Ramsey, Zhang, and Spirtes 2006) which is weaker than faithfulness.

[9] Symmetry correction is required by Markov blanket discovery algorithms for the sake of correctness. To conclude that a feature X is a neighbor of the outcome Y, the symmetry correction further checks whether Y is adjacent to X in addition to checking that X is adjacent to Y.

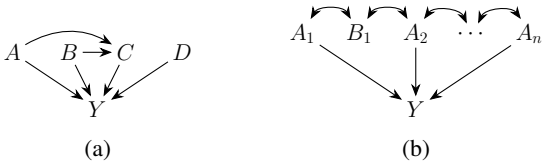


Figure 4: Example predictive graphs.

Algorithm 1 depicts the details of adjacency search. Under the weak faithfulness assumption, the direct causes of model predictions can be discovered by calling ADJ-SEARCH in Algorithm 1 while skipping lines 14-18.[10]

Before showing another technique for optimizing the discovery process, we note that a distribution $\Pr$ can be both canonical and weakly faithful. This happens, for example, when $\Pr$ is a P-MAP of the true predictive graph.

5 Optimization with an Independence Rule

We next introduce a novel independence rule that can be integrated with the adjacency search to accelerate the discovery process when $\Pr$ is canonical. This result can be combined with the optimization technique mentioned in the previous section if $\Pr$ is also weakly faithful.

We start with the following main theorem which introduces an important independence rule.

Theorem 11. *Let* $\Pr$ *be a distribution over disjoint variable sets* $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$. *If* $\mathcal{I}_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y})$ *and* $\mathcal{I}_{\Pr}(\mathbf{X} \cup \mathbf{Z}, \emptyset, \mathbf{W})$*, then* $\mathcal{I}_{\Pr}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$.

This result allows us to skip the independence test on $\mathcal{I}_{\Pr}(\mathbf{X}, \mathbf{Z} \cup \mathbf{W}, \mathbf{Y})$, which involves a larger conditioned set, if we know that $\overline{\mathcal{I}_{\Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ and $\mathcal{I}_{\Pr}(\mathbf{X} \cup \mathbf{Z}, \emptyset, \mathbf{W})$, which involve smaller conditioned sets. This method can be applied widely to skip independence tests in adjacency search, where the independence tests are conducted with increasingly larger conditioned sets. Again, skipping independence tests speeds up the adjacency search, therefore the discovery of direct causes, since the time complexity of discovery algorithms is dominated by the number independence tests; see Section 2.2 for our earlier discussion on this.

We next define a notion that can be used to characterize scenarios in which an independence test can be skipped.

Definition 12. *A variable set* $\mathbf{V}$ *is said to be I-decomposable wrt distribution* $\Pr$ *if* $\mathbf{V}$ *can be partitioned into non-empty sets* $\mathbf{V}_1$ *and* $\mathbf{V}_2$ *where* $\mathcal{I}_{\Pr}(\mathbf{V}_1, \emptyset, \mathbf{V}_2)$.

The notion of I-decomposability can be used as follows. Suppose we want to test $\mathcal{I}_{\Pr}(X, \mathbf{Z}, Y)$, the classical way is to apply χ^2-test (or G-test), which can be quite time consuming if the sample size is large. However, if we further know that $\mathbf{Z}' = (\mathbf{Z} \cup \{X\})$ is I-decomposable, we can immediately conclude that the independence does not hold and skip the independence test for the following reason. Given that $\mathbf{Z}'$ is I-decomposable, we can partition $\mathbf{Z}'$ into independent sets $\mathbf{Z}_1, \mathbf{Z}_2$ where $X \in \mathbf{Z}_1$. Since we test independence with an increasing size of conditioned set, we must have already tested and concluded $\overline{\mathcal{I}_{\Pr}}(X, \mathbf{Z}_1 \setminus \{X\}, Y)$, which implies that $\overline{\mathcal{I}_{\Pr}}(X, \mathbf{Z}, Y)$ by Theorem 11. To illustrate, consider a distribution $\Pr$ that is a P-MAP of the predictive graph in Figure 4a. During the adjacency search , we can skip the independence test $\mathcal{I}_{\Pr}(Y, \{A, D\}, B)$ given the fact that $\mathcal{I}_{\Pr}(B, \emptyset, \{A, D\})$ and $\overline{\mathcal{I}_{\Pr}}(Y, \emptyset, B)$. We insert this type of optimization as a precondition in line 7 of Algorithm 1 and will call it the *I-decomposability rule*.

One practical question related to I-decomposability is: How can we check efficiently whether a set $\mathbf{V}$ is I-decomposable? Our method for checking I-decomposability hinges on the convenience brought by the canonicalness assumption. In particular, when $\Pr$ is canonical, the marginal independence $\mathcal{I}_{\Pr}(\mathbf{V}_1, \emptyset, \mathbf{V}_2)$ holds iff the pairwise independence $\mathcal{I}_{\Pr}(V_1, \emptyset, V_2)$ holds for all $V_1 \in \mathbf{V}_1$ and $V_2 \in \mathbf{V}_2$.[11] We can now efficiently check I-decomposability for a variable set $\mathbf{V}$ as follows. Initialize a set $\mathbf{S} = \{V\}$ with any $V \in \mathbf{V}$. Add variables to $\mathbf{S}$ recursively: for each $V \in \mathbf{V}$ that is not already in $\mathbf{S}$, add V to $\mathbf{S}$ if $\overline{\mathcal{I}_{\Pr}}(V, \emptyset, \mathbf{S})$. The set $\mathbf{V}$ is I-decomposable iff $\mathbf{S} \neq \mathbf{X}$ when no more variable can be added to $\mathbf{S}$. In practice, we can avoid repeated independence tests by caching pairwise independences.

Algorithm 1 with the I-decomposability rule preserves adjacency searches as shown in the following theorem.

Theorem 13. *Let* $\Pr$ *be a canonical distribution over* $\mathbf{X}, Y$. *Then* ADJ-SEARCH$(\mathbf{X}, Y, \Pr)$ *in Algorithm 1 yields the same result with or without line 7.*

That is, we can integrate the I-decomposability rule into the Markov blanket discovery algorithms (such as M3B) while preserving their soundness and completeness for finding direct causes. When the distribution $\Pr$ is both canonical and weakly faithful, we can combine the I-decomposability rule with the results in Section 4.2 to further speed up the discovery of direct causes.

Algorithm 1: Adjacency Search with Symmetry Correction

```
1:  procedure NONSYM-SEARCH(Features X, Target Y, Pr)
2:      Initialize adjacent nodes C ← X
3:      depth d ← 0
4:      while d < |C| do
5:          for every W ∈ C do
6:              for every Z ⊆ (C \ {W}) where |Z| = d do
7:                  if Z ∪ {W} is I-decomposable then continue
8:                  if I_Pr(Y, Z, W) then remove W from C
9:          d ← d + 1
10:     return C
11: procedure ADJ-SEARCH(Features X, Outcome Y, Pr)
12:     neighbors(Y) ← NONSYM-SEARCH(X, Y, Pr)
13:     /* The following code is for symmetry correction */
14:     for every Z ∈ neighbors(Y) do
15:         W ← X ∪ {Y} \ {Z}
16:         neighbors(Z) ← NONSYM-SEARCH(W, Z, Pr)
17:         if Y ∉ neighbors(Z) then
18:             neighbors(Y) ← neighbors(Y) \ {Z}
19:     return neighbors(Y)
```

[10] The independence test $\mathcal{I}_{\Pr}$ will be replaced by χ^2-test (or G-test) in practice.

[11] This result does not hold for general distributions; see (Pearl 1988; Darwiche 2009) for details.

Corollary 14. *Let $G(\mathbf{X}, Y)$ be a predictive graph and* $\Pr$ *be a distribution induced by G. Suppose* $\Pr$ *is both canonical and weakly faithful wrt G, then* ADJ-SEARCH$(\mathbf{X}, Y, \Pr)$ *in Algorithm 1 (with line 7 and without lines 14-18) returns the direct causes of Y in G.*

We next analyze the time complexity of the adjacency search. In particular, we focus on the number of independence tests required by the NONSYM-SEARCH procedure since it is dominating component of adjacency search (as shown in Algorithm 1). Similar to the result in (Spirtes, Glymour, and Scheines 2000), the number of independence tests required by NONSYM-SEARCH *without the I-decomposability rule* is bounded by $O(n \cdot \sum_{k=0}^{c} \binom{n}{k})$, where n is the number of features and $c = |\mathbf{C}|$ is the number of variables returned by the procedure. It is evident that the procedure requires no more independence tests when we add the I-decomposability rule as a precondition in line 7, since it only skips independence tests.[12] But how much speedup can the I-decomposability rule provide? The next proposition shows that the I-decomposability rule can *exponentially* reduce the number of independence tests in some cases.

Proposition 15. *There exists a class of distributions* $\Pr$ *with n features where* NONSYM-SEARCH *with line 7 requires $O(n^3)$ independence tests while* NONSYM-SEARCH *without line 7 requires $O(n \cdot \exp(n))$ independence tests.*

The proof of the result is based on constructing distributions that are P-MAP of the predictive graphs in Figure 4b. To summarize, we introduced two types of optimizations that speed up the discovery of direct causes. Both optimizations are based on improving the efficiency of the discovery of Markov boundaries. The first one simplifies the discovery procedure when the distribution is weak faithful, whereas the second one allows us to skip independence tests in adjacency search when the distribution is canonical.

Before presenting empirical results to further demonstrate the effectiveness of the I-decomposability rule, we note here that the NONSYM-SEARCH in Algorithm 1 (with line 7) is *anytime*. That is, we can interrupt the search algorithm at any time without losing the true direct causes. This can be achieved by bounding the depth d in Line 4 of Algorithm 1. This result is crucial in practice under limited resources.

6 Experiments

We conduct experiments to further demonstrate the effectiveness of the I-decomposability rule. We compare the *computational efficiency* and *sample efficiency* of discovery algorithms with and without I-decomposability rule under the cases of (i) canonicalness and weak faithfulness; and (ii) canonicalness only. For case (i), we compare the performance of six different algorithms: Algorithm 1 without line 7 (ADJ), Algorithm 1 with line 7 (ALG1), Interleaved HITON-PC (Aliferis, Tsamardinos, and Statnikov 2003; Aliferis et al. 2010) (I-HITON), interleaved HITON-PC with the I-decomposability rule (I-HITON-DEC), Semi-Interleaved HITON-PC (Aliferis et al. 2010) (SI-HITON) and Semi-Interleaved HITON-PC with the I-decomposability rule (SI-HITON-DEC). For case (ii), we compare the performance of two algorithms: the M3B algorithm (Yu et al. 2018) (M3B), and M3B algorithm with the I-decomposability rule (M3B-DEC).[13]

For all algorithms, we employ χ^2-tests to test independences from data. In particular, we set a threshold of 0.2 on the p-value for all pairwise independence tests required by I-decomposability (line 7 in Algorithm 1), a threshold of 0.1 for ADJ, ALG1, I-HITON, I-HITON-DEC, SI-HITON, SI-HITON-DEC, and a threshold of 0.05 for M3B, M3B-DEC. When a discovery algorithm returns more direct causes than there actually are, we keep the direct causes that attain the lowest p-value among all independence tests conducted by the algorithm. In Algorithm 1, this can be implemented by recording the p-values for all independence tests in line 8.

For all experiments, we consider random causal models (causal Bayesian networks) that contains 100 variables. The causal graphs for these models are generated using the Erdős–Rényi method (Erdős, Paul and Rényi, Alfréd 1959) as follows. In case (i), we first generate a random ADMG over 99 features where each directed edge is added with probability 0.5 and each bidirected edge is added with probability 0.1.[14] We then randomly pick c features to be the parents the outcome Y. In case (ii), we generate a random ADMG over 100 variables where each directed edges is added with probability 0.5 and each bidirected edge is added with probability 0.01. We bound the maximal degree of variables by d. In both cases, every variable has 2 or 3 states.

Our first set of experiments compares the computational efficiency of the algorithms. We consider causal graphs with different complexity by varying the number of direct causes $c \in \{7, 8, 9, 10\}$ in case (i) and vary the maximal degree with $d \in \{7, 8, 9, 10\}$ in case (ii). In both cases, the algorithms need to discover the direct causes from 100,000 random samples generated from the true causal model. Tables 1 and 2 (in Appendix) record the average accuracy, time (in seconds), and number of independence tests of algorithms over 20 runs. It is evident that algorithms with the I-decomposability rule require fewer independence tests and hence are faster than the algorithms without the rule, with extreme cases when the integration of I-decomposability rule halves the time required by discovery, e.g., $c = 10$ in Table 1. This demonstrates that the I-decomposability rule can significantly speed up the computational efficiency of the discovery algoirthms. In general, the improvement is more significant in Table 1 than Table 2, indicating that the rule is more effective in case (i) than case (ii).

Our second set of experiments compares the sample efficiency of the algorithms. We vary the sample size from $N \in \{1000, 5000, 10000, 20000, 50000, 100000, 150000, 200000\}$

[12] The I-decomposability rule adds at most $O(n^2)$ tests for pairwise independences. This overhead, however, is negligible when the causal graph is dense.

[13] The I-decomposability rule can be incorporated into the HITON-PC algorithms and M3B algorithm, similar to Algorithm 1, as a precondition for each independence test. We also implemented symmetry correction for the M3B algorithm.

[14] We bound the maximal degree of features by 6, where the degree of a node is defined as the number of its parents and children.

while keeping $c = 8$ fixed in case (i) and $d = 7$ fixed in case (ii). Figure 5 (in Appendix) plots the average accuracy of the algorithms over 20 runs. The results show that the algorithms with and without the I-decomposability rule achieved similar accuracy under all sample sizes. This suggests that the integration of I-decomposability rule does not worsen the sample efficiency of existing algorithms.

7 Conclusion

We studied the problem of discovering features that directly cause the predictions made by predictive models, empowered by a causal modeling framework that represents the prediction process using causal graphs. We presented two conditions under which the direct causes are guaranteed to be discoverable and become equivalent to the notion of Markov boundary. In these cases, existing methods for discovering Markov boundaries can be leveraged to discover the direct causes. We further proposed a novel independence rule that can be integrated with existing algorithms to improve the computational efficiency. This work opens the door to modeling predictive models with causal tools, even when these models are non-transparent like neural networks. Potential future works include identifying more conditions under which the direct causes can be (efficiently) discovered, studying the discovery of indirect causes for model predictions, and exploring the applications of the independence rule in broader contexts of causal discovery.

Acknowledgments

This work is supported by RTX Technology Research Center. The authors would like to thank Adnan Darwiche (UCLA), and Isaac Cohen, Kishore Reddy, Adam Suarez, and Nathanial Hendler (RTX) for all the useful discussions and feedback.

References

Ali, S.; Abuhmed, T.; El-Sappagh, S. H. A.; Muhammad, K.; Alonso-Moral, J. M.; Confalonieri, R.; Guidotti, R.; Ser, J. D.; Rodríguez, N. D.; and Herrera, F. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Inf. Fusion*, 99: 101805.

Aliferis, C. F.; Statnikov, A. R.; Tsamardinos, I.; Mani, S.; and Koutsoukos, X. D. 2010. Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *J. Mach. Learn. Res.*, 11: 171–234.

Aliferis, C. F.; Tsamardinos, I.; and Statnikov, A. R. 2003. HITON: A Novel Markov Blanket Algorithm for Optimal Variable Selection. In *AMIA*. AMIA.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Bishop, C. M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth.

Buchanan, B. G.; and Shortliffe, E. H. 1984. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley series in artificial intelligence)*. USA: Addison-Wesley Longman Publishing Co., Inc. ISBN 0201101726.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD '15, 1721–1730. New York, NY, USA: Association for Computing Machinery. ISBN 9781450336642.

Coffey, S. M.; and Elliott, M. R. 2023. Optimizing Data Collection Interventions to Balance Cost and Quality in a Sequential Multimode Survey. *Journal of Survey Statistics and Methodology*, 12(3): 741–763.

Cortes, C.; and Vapnik, V. 1995. Support-Vector Networks. *Mach. Learn.*, 20(3): 273–297.

Darwiche, A. 2009. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press.

Darwiche, A. 2020. Three Modern Roles for Logic in AI. In *PODS*, 229–243. ACM.

Darwiche, A.; and Hirth, A. 2020. On the Reasons Behind Decisions. In *ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, 712–720. IOS Press.

Ellis, G. 2012. Chapter 13 - Model Development and Verification. In Ellis, G., ed., *Control System Design Guide (Fourth Edition)*, 261–282. Boston: Butterworth-Heinemann, fourth edition edition. ISBN 978-0-12-385920-4.

Erdős, Paul; and Rényi, Alfréd. 1959. On Random Graphs. *Publicationes Mathematicae Debrecen*, 6: 290–297.

Freedman, D. 2005. *Statistical Models : Theory and Practice*. Cambridge University Press.

Fu, S.; and Desmarais, M. C. 2008. Fast Markov Blanket Discovery Algorithm Via Local Learning within Single Pass. In *Canadian AI*, volume 5032 of *Lecture Notes in Computer Science*, 96–107. Springer.

Goodfellow, I. J.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. Cambridge, MA, USA: MIT Press. http://www.deeplearningbook.org.

Lada Kohoutová; Juyeon Heo; Sungmin Cha; Sungwoo Lee; Taesup Moon; Tor D. Wager; and Choong-Wan Woo. 2020. Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature protocols*, 15(4): 1399–1435.

Ling, Z.; Li, B.; Zhang, Y.; Wang, Q.; Yu, K.; and Wu, X. 2023. Causal Feature Selection With Efficient Spouses Discovery. *IEEE Trans. Big Data*, 9(2): 555–568.

Ling, Z.; Yu, K.; Wang, H.; Liu, L.; Ding, W.; and Wu, X. 2019. BAMB: A Balanced Markov Blanket Discovery Approach to Feature Selection. *ACM Trans. Intell. Syst. Technol.*, 10(5): 52:1–52:25.

Lipton, Z. C. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10): 36–43.

Liu, H.; Shi, Q.; Cai, Y.; Wang, N.; Zhang, L.; and Liu, D. 2024. Fast Shrinking parents-children learning for Markov

blanket-based feature selection. *Int. J. Mach. Learn. Cybern.*, 15(8): 3553–3566.

MacKenzie, I. S. 2013. Chapter 7 - Modeling Interaction. In MacKenzie, I. S., ed., *Human-computer Interaction*, 233–292. Boston: Morgan Kaufmann. ISBN 978-0-12-405865-1.

Neilson, A.; Indratmo; Daniel, B. K.; and Tjandra, S. 2019. Systematic Review of the Literature on Big Data in the Transportation Domain: Concepts and Applications. *Big Data Res.*, 17: 35–44.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc. ISBN 1558604790.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. USA: Cambridge University Press, 2nd edition. ISBN 052189560X.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books.

Pearson, K. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302): 157–175.

Peña, J. M.; Nilsson, R.; Björkegren, J.; and Tegnér, J. 2007. Towards scalable and data efficient learning of Markov boundaries. *Int. J. Approx. Reason.*, 45(2): 211–232.

Ramsey, J. D.; Zhang, J.; and Spirtes, P. 2006. Adjacency-Faithfulness and Conservative Causal Inference. In *UAI*. AUAI Press.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD*, 1135–1144. ACM.

Richardson, T. 2003. Markov Properties for Acyclic Directed Mixed Graphs. *Scandinavian Journal of Statistics*, 30(1): 145–157.

Richardson, T.; and Spirtes, P. 2002. Ancestral graph Markov models. *The Annals of Statistics*, 30(4): 962 – 1030.

Sokal, R.; and Rohlf, F. 2013. *Biometry : the principles and practice of statistics in biological research*.

Spirtes, P.; Glymour, C.; and Scheines, R. 2000. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press.

Trask, C.; Mathiassen, S.; Wahlström, J.; Heiden, M.; and Rezagholi, M. 2012. Data collection costs in industrial environments for three occupational posture exposure assessment methods. *BMC medical research methodology*, 12: 89.

Tsamardinos, I.; Aliferis, C. F.; and Statnikov, A. R. 2003. Time and sample efficient discovery of Markov blankets and direct causal relations. In *KDD*, 673–678. ACM.

Tsamardinos, I.; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.*, 65(1): 31–78.

Wang, H.; Ling, Z.; Yu, K.; and Wu, X. 2020. Towards efficient and effective discovery of Markov blankets for feature selection. *Inf. Sci.*, 509: 227–242.

Woodward, J. 2004. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. ISBN 9780195155273.

Yu, K.; Liu, L.; Li, J.; and Chen, H. 2018. Mining Markov Blankets Without Causal Sufficiency. *IEEE Trans. Neural Networks Learn. Syst.*, 29(12): 6333–6347.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, 962–970. PMLR.

A More Details on Markov Blankets

A.1 Markov Blankets

When the distribution $\Pr$ is a P-MAP of some directed directed acyclic graph (DAG) G over variables $\mathbf{X}, Y$, the Markov boundary of Y in $\Pr$ is exactly the Markov blanket of Y in G, which contains the parents, children, and spouses of Y in G (Pearl 1988).

More generally, when $\Pr$ is a P-MAP of some ADMG G, the Markov boundary of Y is also unique determined but its conversion to the Markov blanket is more subtle. First observe that $\Pr$ must also a P-MAP of some maximal ancestral graph (MAG)[15] since classes of ADMGs and MAGs are Markov equivalent as shown in (Richardson 2003). That is, every ADMG can be converted to a MAG with equivalent m-separations. Instead of studying the Markov blanket in ADMGs, we shall consider the notion of Markov blankets in the equivalent MAGs, which has been defined in (Yu et al. 2018). In particular, let the *district set* of Y to be all the variables that are connected to Y with a bidirected path,[16]the Markov blanket of Y in a MAG contains the following variables (Yu et al. 2018):

- $pa(Y)$: the parents of Y
- $ch(Y)$: the children of Y
- $sp(Y)$: the spouses of Y
- $dis(Y)$: the district set of Y
- $pa(dis(Y))$: all the parents of variables in $dis(Y)$
- $dis(ch(Y))$: all the district sets of $ch(Y)$
- $pa(dis(ch(Y)))$: all the parents of variables $dis(ch(Y))$

A.2 Discovery of Markov Blankets

We summarize some popular methods for discovering Markov blankets (MB) from data, which has been studied extensively in the past. We start with algorithms that discover MBs in DAGs. IAMB (Tsamardinos, Aliferis, and Statnikov 2003) was one of the earliest methods that uses greedy search for discovering MBs. Later, the *divide-and-conquer* approach was employed to improve the efficiency by finding parents and children (PC) and spouses separately; see, e.g., HITON (Aliferis, Tsamardinos, and Statnikov 2003), PCMB (Peña et al. 2007), IPCMB (Fu and Desmarais 2008), GLL (Aliferis et al. 2010), CFS (Ling et al. 2023). More recently, it was shown that the performance can be further improved by finding PC and spouses simultaneously; see, e.g., BAMB (Ling et al. 2019), EEMB (Wang et al. 2020), FSMB (Liu et al. 2024).

These algorithms can be classified into two approaches based on how the PC set is discovered. First is the *adjacency-based approach*, which initializes the PC set with all features and then removes variables from the PC set if they become independent of the target Y under some conditioned set. Algorithms that fit into this category include IPCMB and FSMB. Second is the *grow-and-shrink* approach, which initializes the PC set to be empty and iteratively adds variables to it. Each time a new variable is added to the PC set, the grow-and-shrink approach also removes variables that become non-neighbors of the target Y from the PC set. Algorithms that fit into this approach include HITON, PCMB, GLL, BAMB, EEMB. While the grow-and-shrink approach can potentially reduce the number of independence tests by maintaining a small PC set, the performance depends heavily on the order of variables being added to the PC set. It is worth noting that both approaches use the procedure of adjacency search extensively to decide whether a variables can be removed from the PC set.

Recently, the discovery of Markov blankets under causal insufficiency have been studied and the M3B algorithm is developed to discover Markov blankets in MAGs in (Yu et al. 2018). The algorithm first discovers the neighbors of Y via an adjacency search and then discover the rest of Markov blanket through a recursive search. Both steps call the procedure of adjacency search (Algorithm 1) extensively. We will also use the M3B algorithm for discovering Markov boundaries in the context of ADMGs since the classes of ADMGs and MAGs are Markov equivalent as we mentioned earlier.

[15]MAG is a subtype of ADMG that satisfies the *maximal* and *ancestral* properties; see (Richardson and Spirtes 2002) for more details.

[16]A bidirected path is a path that contains only bidirected edges.

B Proofs

Proof of Proposition 6

Proof. We first show the only-if direction. By contradiction, suppose $\mathcal{I}_{\Pr}(Y, \mathbf{X}', X)$, then $\Pr(y|x, \mathbf{x}') = C$ for all x where C is a constant. We can then compute $\Pr(y|do(\mathbf{x}'))$ as follows.

$$\begin{aligned}
&\Pr(y|do(\mathbf{x}')) = \sum_x \Pr(y|x, do(\mathbf{x}'))\Pr(x|do(\mathbf{x}')) \\
&= \sum_x \Pr(y|x, \mathbf{x}')\Pr(x|do(\mathbf{x}')) \quad \text{(Rule 2 of do-calculus)} \\
&= C\sum_x \Pr(x|do(\mathbf{x}')) = C
\end{aligned}$$

Since $\Pr(y|x, \mathbf{x}') = \Pr(y|do(x), do(\mathbf{x}'))$ by Rule 2 of do-calculus (Pearl 2009), we conclude $\Pr(y|do(\mathbf{x}')) = \Pr(y|do(x), do(\mathbf{x}')) = C$ for all x, contradiction.

Now consider the if direction. Suppose $\overline{\mathcal{I}_{\Pr}}(Y, \mathbf{X}', X)$, we can always find an instantiation $y, \mathbf{x}'$ such that $\Pr(y|x_1, \mathbf{x}') \neq \Pr(y|x_2, \mathbf{x}')$. Moreover, there must exists some state x^* that attains the largest $\Pr(y|x^*, \mathbf{x}')$. Again, we can write out the $\Pr(y|do(\mathbf{x}'))$ as follows

$$\begin{aligned}
&\Pr(y|do(\mathbf{x}')) = \sum_x \Pr(y|x, do(\mathbf{x}'))\Pr(x|do(\mathbf{x}')) \\
&= \sum_x \Pr(y|x, \mathbf{x}')\Pr(x|do(\mathbf{x}')) \quad \text{(Rule 2 of do-calculus)} \\
&< \sum_x \Pr(y|x^*, \mathbf{x}')\Pr(x|do(\mathbf{x}')) \\
&= \Pr(y|x^*, \mathbf{x}') \\
&= \Pr(y|do(x^*), do(\mathbf{x}')) \quad \text{(Rule 2 of do-calculus)}
\end{aligned}$$

We conclude $\Pr(y|do(\mathbf{x}')) \neq \Pr(y|do(x^*), do(\mathbf{x}'))$. □

Proof of Theorem 8

Proof. It suffices to check whether the result holds for the class of MAGs since every ADMG can be convert to some MAG that is Markov equivalent as discussed in Appendix A. As shown in (Yu et al. 2018), the minimal set that separates a target Y and other variables in a MAG is the Markov blanket (MB) of Y. We next show that $\text{MB}(Y)$ are the only variables that satisfy the condition in Proposition 6. First, by weak union, $\mathcal{I}_{\text{Pr}}(Y, \mathbf{X} \setminus \{X\}, X)$ for all $X \notin \text{MB}(Y)$ since $\mathcal{I}_{\text{Pr}}(Y, \mathbf{X} \setminus \text{MB}(Y), \text{MB}(Y))$ by the definition of Markov boundary. We next show that all $X \in \text{MB}(Y)$ satisfies the condition. Note that $\overline{\mathcal{I}_{\text{Pr}}}(Y, \text{MB}(Y) \setminus X, X)$ for each $X \in \text{MB}(Y)$. Otherwise, by contraction rule, $\text{MB}'(Y) = \text{MB}(Y) \setminus \{X\}$ is also a valid Markov boundary, contradicting the uniqueness of MB. Moreover, for each $X \in \text{MB}(Y)$, the active path from Y to X is still not m-separated even when we condition on more variables besides $\text{MB}(Y)$. Hence, $\overline{\mathcal{I}_{\text{Pr}}}(Y, \mathbf{X} \setminus \{X\}, X)$. □

Proof of Theorem 10

Proof. Let $\mathbf{C}$ be the causes of Y in G that satisfies the condition in Proposition 6. By m-separation, it is guaranteed that $\mathcal{I}_{\text{Pr}}(Y, \mathbf{C}, \mathbf{X} \setminus \mathbf{C})$, so $\mathbf{C}$ is a valid Markov blanket. We are left to show that $\mathbf{C}$ is unique and minimal. Suppose there exists another Markov boundary $\mathbf{W}$ that omits some variable $T \in \mathbf{C}$, then $\overline{\mathcal{I}_{\text{Pr}}}(T, \mathbf{W}, Y)$ by the definition of weak faithfulness, contradicting $\mathbf{W}$ being a Markov boundary. □

Proof of Theorem 11

Proof. First, by the rule of weak union, $\mathcal{I}_{\text{Pr}}(\mathbf{X} \cup \mathbf{Z}, \emptyset, \mathbf{W})$ implies $\mathcal{I}_{\text{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{W})$. We then have the following:

$$\begin{aligned} \Pr(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) &= \sum_{\mathbf{W}} \Pr(\mathbf{Y}, \mathbf{W}|\mathbf{X}, \mathbf{Z}) \\ &= \sum_{\mathbf{W}} \Pr(\mathbf{Y}|\mathbf{W}, \mathbf{X}, \mathbf{Z}) \Pr(\mathbf{W}|\mathbf{X}, \mathbf{Z}) \\ &= \sum_{\mathbf{W}} \Pr(\mathbf{Y}|\mathbf{W}, \mathbf{Z}) \Pr(\mathbf{W}|\mathbf{Z}) \\ &= \Pr(\mathbf{Y}|\mathbf{Z}) \end{aligned} \tag{1}$$

which implies $\mathcal{I}_{\text{Pr}}(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$. □

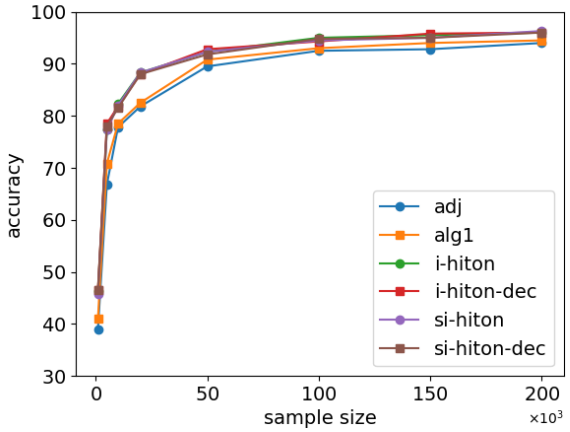
Proof of Theorem 13

Proof. It suffices to show that the output of NONSYM-SEARCH is invariant with or without the I-decomposability rule. That is, whenever $\mathbf{Z} \cup \{W\}$ is I-decomposable, it is guaranteed that $\overline{\mathcal{I}_{Pr}}(Y, \mathbf{Z}, W)$. This follows from Theorem 11. Suppose not, then $\mathcal{I}_{Pr}(Y, \mathbf{Z}, W)$ together with $\mathbf{Z} \cup \{W\}$ I-decomposable would imply that $\mathcal{I}_{\text{Pr}}(Y, \mathbf{Z}', W)$ for some $\mathbf{Z}' \subset \mathbf{Z}$. This leads to a contradiction since we should have removed W from $\mathbf{C}$ much earlier. □

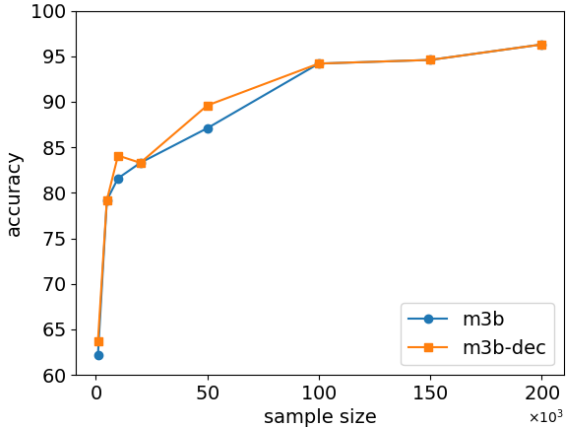
Proof of Proposition 15

Proof. Consider the class of predictive graphs G shown in Figure 4b where n can be arbitrarily large. Let Pr be the distributions that is a P-MAP of G. Algorithm 1 with line 7 will first remove all B's at $d = 2$ since $\mathcal{I}_{\text{Pr}}(B_i, \{A_i, A_{i+1}\}, Y)$. It takes $O(n^3)$ conditional independence tests for $d = 2$ since we need to enumerate $\binom{n}{2}$ conditioned variables for each of the n variables. No more conditional independence tests will be needed since any subsets of $\{A_i\}_{i=1}^n$ with size greater than 2 are I-decomposable.

We next consider the case without the I-decomposability rule. Similar to the previous case, the adjacency search removes all B's at $d = 2$. However, the algorithm will continue searching for $d = 3, \ldots, n-1$ afterwards, which takes a total of $O(n \cdot \exp(n))$ independence tests. □



(a) weak faithfulness



(b) canonicalness

Figure 5: Accuracy of algorithms for identifying direct causes under various sample sizes. Methods ALG1, I-HITON-DEC, SI-HITON-DEC, M3B-DEC are integrated the I-decomposability rule.

Method	$c = 7$			$c = 8$			$c = 9$			$c = 10$		
	Acc	Time	#CI	Acc	Time	#CI	Acc	Time	#CI	Acc	Time	#CI
ADJ	93.1	3.1	2171	93.0	4.1	2853	86.0	5.3	3630	84.8	6.5	4309
ALG1	**93.7**	**1.9**	**1497**	**93.0**	**2.5**	**1834**	**87.3**	**2.6**	**1923**	**85.2**	**3.0**	**2142**
I-HITON	**96.6**	1.9	1132	95.0	3.4	2,061	89.8	5.6	3335	89.0	7.7	4477
I-HITON-DEC	96.3	**1.2**	**685**	**95.0**	**1.9**	**1095**	**90.0**	**3.0**	**1650**	**90.2**	**3.5**	**1931**
SI-HITON	96.0	2.3	1313	**95.0**	3.5	2062	90.0	5.9	3385	88.4	8.7	4887
SI-HITON-DEC	**96.0**	**1.4**	**805**	94.8	**2.0**	**1167**	**90.2**	**2.8**	**1560**	**89.6**	**3.8**	**2035**

Table 1: Average accuracy (Acc), time (Time), and number of independence tests (#CI) of algorithms under both canonicalness and weak faithfulness. Methods ALG1, I-HITON-DEC, SI-HITON-DEC are integrated with the I-decomposability rule.

Method	$d = 7$			$d = 8$			$d = 9$			$d = 10$		
	Acc	Time	#CI	Acc	Time	#CI	Acc	Time	#CI	Acc	Time	#CI
M3B	86.9	52.1	48131	**73.8**	178.6	146322	74.7	818.5	523858	**71.3**	1866.1	1090243
M3B-DEC	**86.9**	**41.9**	**41865**	73.3	**156.5**	**129390**	**75.5**	**794.6**	**473593**	71.2	**1755.3**	**1009157**

Table 2: Average accuracy (Acc), time (Time), and number of independence tests (#CI) of algorithms under canonicalness only. Method M3B-DEC are integrated with the I-decomposability rule.