Causal inference in degenerate systems: An impossibility result

Yue Wang
Institut des Hautes Études Scientifiques

Linbo Wang
University of Toronto

Abstract

Causal relationships among variables are commonly represented via directed acyclic graphs. There are many methods in the literature to quantify the strength of arrows in a causal acyclic graph. These methods, however, have undesirable properties when the causal system represented by a directed acyclic graph is degenerate. In this paper, we characterize a degenerate causal system using multiplicity of Markov boundaries. We show that in this case, it is impossible to find an identifiable quantitative measure of causal effects that satisfy a set of natural criteria. To supplement the impossibility result, we also develop algorithms to identify degenerate causal systems from observed data. Performance of our algorithms is investigated through synthetic data analysis.

1 Introduction

Inferring causal relationships is among the most important goals in many disciplines. A formal approach to represent causal relationships uses causal directed acyclic graphs (DAGs) (Pearl, 2009), in which random variables are represented as nodes and causal relationships are represented as arrows. Besides qualitatively describing causal relationships via DAGs, it is often desirable to obtain quantitative measures of the strength of arrows therein since they provide more detailed information on causal effects. There have been many measures proposed to quantify the causal relationships between nodes in a causal DAG, such as conditional mutual information (Dobrushin, 1963), causal strength (Janzing et al., 2013) and part mutual information (Zhao et al., 2016). See Gao et al. (2016) and its reference list for more such measures.

An interesting observation is that these measures have undesirable properties when the causal system under consideration is degenerate. As a simple example, consider the confounder triangle $Z \to X \to Y$ with an edge $Z \to Y$, where $Z = X$ almost surely. In this case, the conditional mutual information $\mathrm{CMI}(X, Y \mid Z)$ is zero regardless of the influence X has on Y, while the causal strength and part mutual information for the arrow $X \to Y$ are not well-defined. Intuitively, these problems arise because it is not possible to distinguish the causal effect of X on Y from the causal effect of Z on Y.

In this paper, we generalize the observation above by providing a formal characterization of a degenerate causal system in Section 3. We first define a set of natural criteria to be expected from a reasonable measure of causal influence, and show that when the causal system is degenerate, all reasonable measures of a causal influence cannot be identified from the distribution represented by the DAG. Analysts may instead report qualitative summaries of causal relationships, such as all the causal explanations of the response variable.

Our characterization of a degenerate causal system is based on multiplicity of Markov boundaries for the response variable. The Markov boundary of a variable W in a variable set $\mathcal{S}$ is a minimal subset of $\mathcal{S}$, conditional on which all the remaining variables in $\mathcal{S}$, excluding W, are rendered statistically independent of W (Statnikov et al., 2013). In Section 4, we propose novel approaches to determine the uniqueness of Markov boundary from data. Many authors have considered methods for discovery of Markov boundaries. However, the validity of their methods often requires strong assumptions (e.g. Tsamardinos & Aliferis, 2003; Peña et al., 2007; Aliferis et al., 2010), some of which even imply that the response variable has a unique Markov boundary (e.g. de Morais & Aussem, 2010; Mani & Cooper, 2004). Furthermore, some of these methods output all the Markov boundaries (e.g. Statnikov et al., 2013), which are not necessary for our purpose. In contrast, our novel algorithms are more robust to model assumptions and computationally more tractable.

2   Background

2.1   Set-up

Consider a causal DAG Γ with vertices $\mathcal{V}$. We say X is a parent of Y if the path $X \to Y$ is present in Γ, and Y is a descendant of X if a path $X \to \cdots \to Y$ is present in Γ. A variable is a descendant of itself, but not a parent of itself. For a variable W, we use $\mathrm{DES}(W)$ to denote the set consisting of all descendants of W, and $\mathrm{PA}(W)$ to denote the set consisting of all parents of W. We assume that the probability distribution $\mathfrak{p}$ over $\mathcal{V}$ is Markov with respect to Γ in the sense that for every $W \in \mathcal{V}$, W is independent of $\mathcal{V} \setminus \mathrm{DES}(W)$ conditional on $\mathrm{PA}(W)$ (Spirtes et al., 2000).

We assume that we observe independent replications of $\mathcal{V}$. Let $\mathcal{S}$ be all the possible parents of Y, namely all the variables in $\mathcal{V}$ except for Y and those that are known not to be parents of Y. To ease presentation, in our leading case we assume no *prior* knowledge of the causal DAG so that $\mathcal{S} = \mathcal{V} \setminus \{Y\}$. Our main result also applies to settings where one has full or partial *prior* knowledge of the DAG structure. See Remark 4 for more discussions.

Let X be a possible parent of Y and we are interested in the causal effect of X on Y. Let $\mathcal{L} = \mathcal{S} \setminus \{X\}$. We denote the sample space of $X, Y, \mathcal{L}$ by $\mathbb{X}, \mathbb{Y}, \mathbb{L}$, respectively.

2.2   Measures of causal influence

We now review several measures of causal influence in the literature. We only introduce their definitions in the discrete case as they are sufficient to motivate our discussions later.

Conditional mutual information.   The conditional mutual information between X and Y conditional on a set $\mathcal{C}$ is defined as (Dobrushin, 1963)

$$\mathrm{CMI}(X, Y \mid \mathcal{C})$$
$$= \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \sum_{c \in \mathbb{C}} f(x, y, c) \log \frac{f(x, y \mid c)}{f(x \mid c) f(y \mid c)},$$

where f is the probability density function. It can be shown that $\mathrm{CMI}(X, Y \mid \mathcal{C}) = 0$ if and only if $X \perp\!\!\!\perp Y \mid \mathcal{C}$. When $\mathcal{C} = \emptyset$, CMI is known as the mutual information (MI):

$$\mathrm{MI}(X, Y) = \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X, Y),$$

where H is the Shannon entropy. More generally, we have $\mathrm{CMI}(X, Y \mid \mathcal{C}) = \mathrm{MI}(\{X\} \cup \mathcal{C}, Y) - \mathrm{MI}(Y, \mathcal{C})$.

$\mathrm{CMI}(X, Y \mid \mathcal{C})$ quantifies additional (possibly nonlinear) information contained in Y regarding X conditional on $\mathcal{C}$.

Causal strength.   When the full causal DAG (thus $\mathcal{L} = \mathrm{PA}(Y) \setminus \{X\}$) is known, the causal strength of X on Y is defined as (Janzing et al., 2013)

$$\mathrm{CS}(X \to Y)$$
$$= \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \sum_{l \in \mathbb{L}} f(x, y, l) \log \frac{f(y \mid x, l)}{\sum\limits_{x' \in \mathbb{X}} f(y \mid x', l) f(x')}.$$

CS is motivated to avoid the so-called underestimation problem of CMI: when X and Z are almost the same but have strong causal effect on Y, both $\mathrm{CMI}(X, Y \mid Z)$ and $\mathrm{CMI}(Z, Y \mid X)$ are very small.

Part mutual information.   The part mutual information between X and Y conditional on $\mathcal{C}$ is defined as (Zhao et al., 2016)

$$\mathrm{PMI}(X, Y \mid \mathcal{C})$$
$$= \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} \sum_{c \in \mathbb{C}} f(x, y, c) \log \frac{f(x, y \mid c)}{f^*(x \mid c) f^*(y \mid c)},$$

where $f^*(x \mid c) = \sum_{y \in \mathbb{Y}} f(x \mid y, c) f(y)$, $f^*(y \mid c) = \sum_{x \in \mathbb{X}} f(y \mid x, c) f(x)$.

PMI solves a similar underestimation problem of CMI, but it is also symmetric, and definition of which does not depend on knowledge of the full DAG.

2.3   Markov blanket and Markov boundary

We now formally discuss the notion of Markov blanket and Markov boundary.

Definition 1 *Suppose that $\mathcal{T}$ is a set of observed variables not containing W. A subset of $\mathcal{T}$, denoted as $\mathcal{M}$, is a Markov blanket of W within $\mathcal{T}$ if*

$$W \perp\!\!\!\perp (\mathcal{T} \setminus \mathcal{M}) \mid \mathcal{M}.$$

Using the notion of mutual information, the above condition can be written as $\mathrm{CMI}(W, \mathcal{T} \mid \mathcal{M}) = 0$, or equivalently, $\mathrm{MI}(W, \mathcal{T}) = \mathrm{MI}(W, \mathcal{M})$. This suggests that the Markov blanket $\mathcal{M}$ contains all the information of $\mathcal{T}$ on W.

Definition 2 *A Markov blanket is called a Markov boundary if none of its proper subset is a Markov blanket. In other words, a Markov boundary is a minimal Markov blanket.*

Markov boundary always exists. If $W \perp\!\!\!\perp \mathcal{T}$, then $\emptyset$ is the Markov boundary. If no proper subset of $\mathcal{T}$ is a Markov blanket, then $\mathcal{T}$ is the Markov boundary.

Even though Markov boundaries are minimal, in general they are not unique. For example, consider the causal DAG in Fig. 1. Variables X, Y, Z take value

in $\{0, 1, 2\}$, while W takes value in $\{0, 1\}$. Both $\{X, W\}$ and $\{Z, W\}$ are Markov boundaries of Y (conditioned on $\{X, W\}$, either Z or Y only takes one value, which implies independence), but the probability that $(X, W) = (Z, W)$ is less than one. The multiplicity of Markov boundary also implies unfaithfulness (discussed below).
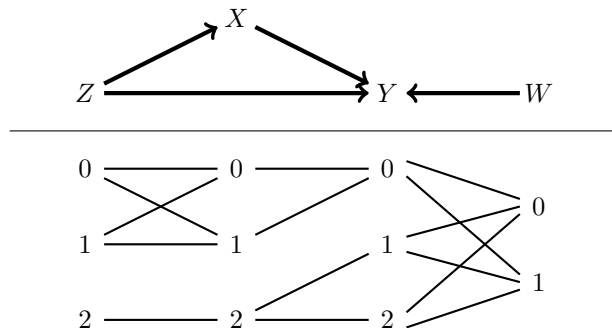

Figure 1: A causal DAG for which variable Y has multiple Markov boundaries (Statnikov et al., 2013). Combinations of values connected with lines have positive joint probabilities. For example, $\mathrm{pr}(Z = 1, X = 0, Y = 0, W = 0) > 0$, while $\mathrm{pr}(Z = 1, X = 2, Y = 1, W = 1) = 0$.

Unlike the confounder triangle example in Section 1, the two Markov boundaries of Y in Fig. 1 do not coincide almost surely. However, X and $\{Z, W\}$ are variation dependent in the sense that there exist x, z, w $(x = 2, z = 0, w = 0)$ such that $\mathrm{pr}(X = x) > 0$, $\mathrm{pr}(Z = z, W = w) > 0$, but $\mathrm{pr}(X = x, Z = z, W = w) = 0$. This variation dependence is in fact an essential property of the multiplicity of Markov boundaries.

Lemma 1 *Let* Θ *denote all Markov boundaries of* Y *in* $\mathcal{T}$, *where* $Y \notin \mathcal{T}$. *Suppose that* $X \in \underset{\mathcal{M} \in \Theta}{\cup} \mathcal{M} \setminus \underset{\mathcal{M} \in \Theta}{\cap} \mathcal{M}$, *and* $\mathcal{K} = \mathcal{T} \setminus \{X\}$. *Then* X *and* $\mathcal{K}$ *are variation dependent in that there exist* $x \in \mathbb{X}, k \in \mathbb{K}$ *such that* $f(x) > 0$, $f(k) > 0$, *but* $f(x, k) = 0$.

It is known in the literature (Pearl & Paz, 1985; Pearl, 1988) that several conditions are sufficient for the uniqueness of Markov boundary.

Definition 3 (Faithfulness) *Let* Λ *denote the collection of conditional independence relationships shared by all probability distributions that are Markov with respect to* Γ. *A probability distribution is faithful to* Γ *if and only if its conditional independence relationships are fully characterized by* Λ.

Definition 4 (Intersection) *A probability distribution on* $\mathcal{V}$ *satisfies the intersection property if and only if for any four subsets of* $\mathcal{V}$, *denoted as* $\mathcal{P}, \mathcal{Q}, \mathcal{Z}, \mathcal{W}$ *such that* $\mathcal{P} \perp\!\!\!\perp \mathcal{Z} \mid (\mathcal{Q}, \mathcal{W})$, $\mathcal{P} \perp\!\!\!\perp \mathcal{W} \mid (\mathcal{Q}, \mathcal{Z})$, *it holds that* $\mathcal{P} \perp\!\!\!\perp (\mathcal{Z}, \mathcal{W}) \mid \mathcal{Q}$.

Definition 5 (Strict positiveness) *A probability distribution on* $\mathcal{V}$ *is called strictly positive if and only if for any two disjoint subsets of variables* $\mathcal{X}$ *and* $\mathcal{Z}$ *such that* $\mathrm{pr}(\mathcal{X} = x) > 0$, $\mathrm{pr}(\mathcal{Z} = z) > 0$, *it holds that* $\mathrm{pr}(\mathcal{X} = x, \mathcal{Z} = z) > 0$.

Strict positivity allows for the expression of causal effects as conditional distributions. Nevertheless, as the proposition below shows, multiplicity of Markov boundaries implies violation of strict positivity.

Proposition 1 (Pearl & Paz, 1985; Pearl, 1988) *If a probability distribution on* $\mathcal{V}$ *(i) is faithful to* Γ, *or (ii) has the intersection property, or (iii) is strictly positive, then any variable* $Y \in \mathcal{V}$ *has a unique Markov boundary in* $\mathcal{V} \setminus \{Y\}$.

Remark 1 *None of the three conditions in Proposition 1 is necessary for the uniqueness of Markov boundary. For example, suppose* $X, Y, Z, W \in \{0, 1\}$, $pr(X = Z = Y = W = 0) = 0.5$, $pr(X = Z = 1, Y = W = 0) = 0.25$, $pr(X = Z = Y = W = 1) = 0.25$ *and* $\mathcal{S} = \{X, Y, Z, W\}$. *Since* $Y \perp\!\!\!\perp X \mid Z$, $Y \perp\!\!\!\perp Z \mid X$, $Y \not\!\perp\!\!\!\perp (X, Z)$, *the joint distribution of* (X, Z, Y, W) *does not have the intersection property and is hence not faithful (Pearl, 1988). On the other hand,* $pr(X = 0) > 0$, $pr(Z = 1) > 0$ *but* $pr(X = 0, Z = 1) = 0$. *Hence the distribution is not strictly positive. However, each variable in this example has a unique Markov boundary within the other three variables.*

2.4 Multiplicity of Markov boundaries

In practice, it often arises that the response variable of interest has multiple Markov boundaries. For instance, in breast cancer studies, several gene sets may have nearly the same effect for survival prediction (Ein-Dor et al., 2004), such that each of the gene sets is a Markov boundary of the survival indicator. In an extensive study, Statnikov et al. (2013) applied nine popular algorithms for learning multiple Markov boundaries to 13 benchmark data sets that cover a wide range of application domains, dimensionalities and sample sizes that are representative of practical settings. One response variable is identified for each data set. Across the nine algorithms, the frequency of reporting multiple Markov boundaries ranges from 46.2% (6/13) to 100%. Five out of the nine algorithms report multiple Markov boundaries in all 13 data sets. All algorithms suggest that there are multiple Markov boundaries in four out of the 13 data sets. These results suggest that a degenerate causal system (system with multiple Markov boundaries) shows up frequently in practice.

Proposition 2 provides theoretical explanation for these empirical findings. Consider n variables, each with the alphabet $\{1, ..., m\}$. The joint distribution $\mathfrak{p}$ of these n variables is randomly chosen from Dirichlet distribution $\mathrm{Dir}(1, 1, ..., 1)$.

Proposition 2 *For any $\epsilon, \delta > 0$, when m is larger than a threshold depending on ϵ, δ, and n is larger than a threshold depending on ϵ, δ, m, with probability larger than $1 - \delta$, we can find a probability distribution $\mathfrak{p}'$ with multiple Markov boundaries, such that the total variation distance between $\mathfrak{p}$ and $\mathfrak{p}'$ is smaller than ϵ.*

Remark 2 *Proposition 2 concerns the measure of distributions that are at most ϵ-distant from a degenerate distribution. A similar result is that the measure of λ-strong-faithful distributions is much less than one (Uhler et al. 2013). In fact, at most ϵ-distant from a degenerate distribution implies λ-strong-unfaithfulness for proper λ, but not vice versa.*

3 When is it possible to reasonably quantify a causal influence?

3.1 Motivation

We motivate our discussion in this section by generalizing our observation in the introduction. Specifically, we show that the causal effect measures introduced in Section 2.2 may not be reasonable when the response variable Y has multiple Markov boundaries within $\mathrm{PA}(Y)$.

Proposition 3 *If $X \in \underset{\mathcal{M} \in \Theta}{\cup} \mathcal{M} \setminus \underset{\mathcal{M} \in \Theta}{\cap} \mathcal{M}$, then (i) $\mathrm{CMI}(X, Y \mid \mathcal{L}) = 0$; (ii) $\mathrm{CS}(X \to Y)$ and $\mathrm{PMI}(X, Y \mid \mathcal{L})$ are not well-defined. Here Θ denotes all Markov boundaries of Y in $\mathcal{S}$.*

To solve problem (ii) in Proposition 3, a naive solution is to assign a value in these degenerate scenarios. However, Proposition 4 below shows that the resulting quantities cannot be continuous functions of the joint distribution of $(X, Y, \mathcal{L})$. Given a probability distribution $\mathfrak{p}'$, we use $\mathrm{CS}[\mathfrak{p}'](X \to Y)$ and $\mathrm{PMI}[\mathfrak{p}'](X, Y \mid \mathcal{L})$ to denote the corresponding causal strength and part mutual information.

Proposition 4 *If $X \in \underset{\mathcal{M} \in \Theta}{\cup} \mathcal{M} \setminus \underset{\mathcal{M} \in \Theta}{\cap} \mathcal{M}$, then there exist two sequences of distributions on $(X, Y, \mathcal{L})$, denoted as $\{\mathfrak{p}_1, \mathfrak{p}_2, ...\}$ and $\{\mathfrak{p}'_1, \mathfrak{p}'_2, ...\}$, both of which converge to $\mathfrak{p}$ under the total variation distance, but $\lim_{i \to \infty} \mathrm{CS}[\mathfrak{p}_i](X \to Y) \neq \lim_{i \to \infty} \mathrm{CS}[\mathfrak{p}'_i](X \to Y)$. The same applies to $\mathrm{PMI}(X, Y \mid \mathcal{L})$.*

Proposition 4 can be proved using Lemma 1 and the following Lemma 2.

Lemma 2 *Assume that there exist $x \in \mathbb{X}, l \in \mathbb{L}$ such that $f(x) > 0$, $f(l) > 0$, but $f(x, l) = 0$. Then there exist two real numbers $g_1 < g_2$, such that for any g with $g_1 < g < g_2$, any $\delta > 0$, there exists a probability distribution $\mathfrak{p}'$ with total variation distance $\mathrm{d}(\mathfrak{p}, \mathfrak{p}') < \delta$, such that $\mathrm{CS}[\mathfrak{p}'](X \to Y) = g$. The same result applies to $\mathrm{PMI}(X, Y \mid \mathcal{L})$.*

Lemma 2 is similar in flavor to the Picard's great theorem: if an analytic function h has an essential singularity at a point w, then on any punctured neighborhood of w, $h(z)$ takes on all possible complex values, with at most a single exception. In this sense, CS and PMI are essentially singular at the probability distribution that implies multiple Markov boundaries for Y.

3.2 Criteria for reasonable causal effect measures

Motivated by our observations in Section 3.1, we now formally describe the criteria we expect from a reasonable measure of causal influence. We focus our discussion on measures that are functionals of the joint distribution of Y and $\mathcal{S}$.

C1. The strength of $X \to Y$ is a continuous function of the joint distribution of Y and $\mathcal{S}$, under the total variation distance.

C2. If there is a unique Markov boundary $\mathcal{M}$ of Y within $\mathcal{S}$, and $X \notin \mathcal{M}$, then the strength of $X \to Y$ is 0.

C3. If there is a unique Markov boundary $\mathcal{M}$ of Y within $\mathcal{S}$, and $X \in \mathcal{M}$, then the absolute value of the strength of $X \to Y$ is at least $c(X, Y, \mathcal{M} \setminus \{X\})$. Here $c(X, Y, \mathcal{M} \setminus \{X\})$ is a positive constant, only depending on $X, Y, \mathcal{M} \setminus \{X\}$, such as $\mathrm{CMI}(X, Y \mid \mathcal{M} \setminus \{X\})$.

We now explain why these criteria are considered natural.

C1: Without continuity, a small perturbation on the observed distribution may lead to a big change in the effect measure. On the other hand, such a small perturbation on the observed distribution can be induced through a small perturbation on the causal system (e.g. coefficients in the structural equation models that generate the DAG). For *identifiable* effect measures, a perturbation on the causal system can *only* act on the effect measure through changing the observed data distribution. This suggests that a small perturbation on the underlying causal system may lead to a big change in the causal effect measure, which is undesirable.

C2: Since the unique Markov boundary contains all the information on Y from $\mathcal{S}$, it is natural to say that X has no causal effect on Y if $X \notin \mathcal{M}$.

C3: Since any variable X in the unique Markov boundary $\mathcal{M}$ of Y contains non-trivial information of Y, it is natural to assign a positive value to the absolute value of strength of $X \to Y$. Variables outside of the unique Markov boundary should not interfere with the strength of $X \to Y$.

3.3 An impossibility result

We now introduce our main result in this section, which reveals the intrinsic difficulty to define measures of causal influence satisfying C1-C3 when multiple Markov boundaries of the response variable are present.

Consider $\mathfrak{S}$, the set of probability distributions on $\mathcal{S} \cup \{Y\}$. Choose a probability distribution $\mathfrak{p} \in \mathfrak{S}$, under which Y has multiple Markov boundaries in $\mathcal{S}$, and X is in at least one, but not all of such Markov boundaries. We are looking for an identifiable measure of the strength of $X \to Y$, $f : \mathfrak{S} \to \mathbb{R}$.

Theorem 1 *In any neighborhood $\mathfrak{N}$ of $\mathfrak{p}$ in $\mathfrak{S}$, all identifiable measures of the strength of $X \to Y$ must violate at least one of the criteria in C1 – C3.*

Remark 3 *Any two criteria among C1 to C3 are compatible with each other. For example, CS and PMI satisfy C2 and C3, a naive causal effect measure that takes a large positive constant value satisfies C1 and C3, and CMI satisfies C1 and C2.*

To prove Theorem 1, we first introduce the tools that we shall use. For any random variable X, we define its perturbation X^ϵ to be a new random variable that coincides with X with probability $1 - \epsilon$, and equals an independent arbitrary noise variable U_X otherwise. For a group of variables, adding ϵ-noise on one variable in the group changes the joint distribution of the whole group by at most ϵ under the total variation distance. The following lemma shows that adding ϵ-noise to X will always decrease the information it has on Y, unless X contains no information regarding Y.

Lemma 3 (Strict Data Processing Inequality)
Let $\mathcal{S}_1$ be a group of variables not containing X or Y. If we add ϵ-noise on X to get X^ϵ, then

$$\text{CMI}(X^\epsilon, Y \mid \mathcal{S}_1) \le \text{CMI}(X, Y \mid \mathcal{S}_1), \quad (1)$$

where the equality holds if and only if

$$\text{CMI}(X, Y \mid \mathcal{S}_1) = 0. \quad (2)$$

The inequality part of Lemma 3 is a special case of the data processing inequality in information theory (Cover & Thomas, 2012). Intuitively, it states

that transmitting data through a noisy channel cannot increase information, namely: garbage in, garbage out. The original data processing inequality (Cover & Thomas, 2012) states that the equality in (1) holds if and only if

$$\text{CMI}(X, Y \mid X^\epsilon, \mathcal{S}_1) = 0. \quad (3)$$

Condition (3) relies on the concrete form of noise, and thus difficult to check. In Lemma 3, we strengthen the result by showing that (3) is equivalent to (2). This improvement is critical for the proof of Lemma 4, in which we describe how to perturb a distribution with multiple Markov boundaries for the response variable, so that in the new distribution the response variable has a unique Markov boundary.

Lemma 4 *Assume Y has multiple Markov boundaries within $\mathcal{S}$. Let $\mathcal{M}_0$ be one of them. If we add ϵ-noise on each variable in $\mathcal{S} \setminus \mathcal{M}_0$, then in the new distribution, $\mathcal{M}_0$ is the unique Markov boundary.*

Proof of Theorem 1. Assume X is in Markov boundary $\mathcal{M}_1$, but not in Markov boundary $\mathcal{M}_2$. On one hand, following Lemma 4 one may add ϵ-noise on each variable in $\mathcal{S} \setminus \mathcal{M}_1$ so that $\mathcal{M}_1$ is the unique Markov boundary of Y. Letting $\epsilon \to 0$, criteria C1 and C3 imply that the absolute value of the strength of $X \to Y$ in the original distribution should be at least $c(X, Y, \mathcal{M} \setminus \{X\})$. On the other hand, one may also add ϵ-noise on each variable in $\mathcal{S} \setminus \mathcal{M}_2$ so that $\mathcal{M}_2$ is the unique Markov boundary of Y. Letting $\epsilon \to 0$, criteria C1 and C2 then imply that the strength of $X \to Y$ in the original distribution should be zero. This constitutes a contradiction.

Remark 4 *We note that the definition of $\mathcal{S}$ depends on knowledge of the DAG, so it is possible that one may obtain consistent estimates of a reasonable causal effect measure given the structure of the underlying DAG, but may not do so without this knowledge. For example, consider causal DAG $X_1 \to X_2 \to Y$ with $X_1 = X_2$ almost surely. If the structure of the DAG is known a priori, then one may define the strength of the arrow $X_2 \to Y$ by ignoring information on X_1. If on the other hand, one has no information on the structure of the DAG, then it is impossible to distinguish the causal effect of X_1 on Y from the causal effect of X_2 on Y. In this case, Theorem 1 suggests that it is impossible to obtain a reasonable quantification of the strength of the arrow $X_2 \to Y$ from data. In general, if knowledge on the DAG implies that a variable X is not a direct cause of Y, then one can exclude X when considering the multiplicity of Markov boundaries of Y.*

Remark 5 *In the presence of multiple Markov boundaries, one can report all variables that show up in at*

least one but not all of the Markov boundaries as "potential causes" of the response variable. Accuracy of such qualitative results depends on the success of algorithms that find multiple Markov boundaries. In contrast to DAG-learning, here one only needs to learn the local structure around a target variable.

4 Tests for the uniqueness of Markov boundary

We develop a two-step procedure to test the uniqueness of Markov boundary: (i) Find a Markov boundary for the response variable Y within the observed data set $\mathcal{S}$; (ii) Decide if there exist other Markov boundaries, other than the one identified in (i).

Methods for step (i) have been discussed extensively in the literature (Tsamardinos & Aliferis, 2003; Peña et al., 2007; Aliferis et al., 2010). However, validity of existing methods typically rely on strong assumptions. For example, faithfulness is required in Aliferis et al. (2010), which implies the uniqueness of Markov boundary, and thus cannot be applied to our problem. Methods in Tsamardinos & Aliferis (2003) and Peña et al. (2007) require that the joint distribution of $\mathcal{S} \cup \{Y\}$ has the so-called composition property, that is, for any four subsets of $\mathcal{S} \cup \{Y\}$, denoted as $\mathcal{P}$, $\mathcal{Q}$, $\mathcal{Z}$, $\mathcal{W}$, such that $\mathcal{P} \perp\!\!\!\perp \mathcal{Z} \mid \mathcal{Q}$, $\mathcal{P} \perp\!\!\!\perp \mathcal{W} \mid \mathcal{Q}$, it holds that $\mathcal{P} \perp\!\!\!\perp (\mathcal{Z}, \mathcal{W}) \mid \mathcal{Q}$.

To relax these assumptions, we develop Algorithm 1 that requires no extra assumptions on the joint distribution. Let Δ be a measure of association between two random variables, with a larger value of Δ indicating a stronger association: If two variables with $\Delta = d_1$ are dependent, then another two variables with $\Delta = d_2 \geq d_1$ are also dependent. One example of Δ that we shall use in simulation studies is the conditional mutual information.

1. **Input**
   Joint distribution of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y

2. **Set** $\mathcal{M}_0 = \mathcal{S}$

3. **Repeat**
   Set $X_0 = \arg\min_{X \in \mathcal{M}_0} \Delta(X, Y \mid \mathcal{M}_0 \setminus \{X\})$
   If $X_0 \perp\!\!\!\perp Y \mid \mathcal{M}_0 \setminus \{X_0\}$
      Set $\mathcal{M}_0 = \mathcal{M}_0 \setminus \{X_0\}$
   Until $X_0 \not\perp\!\!\!\perp Y \mid \mathcal{M}_0 \setminus \{X_0\}$

4. **Output** $\mathcal{M}_0$ is a Markov boundary

Algorithm 1: An assumption-free algorithm for producing one Markov boundary

In step 3 of Algorithm 1, any tie-breaker works when there are several equal Δ.

We now turn to step (ii). The key to our approach is the following necessary and sufficient condition for the uniqueness of Markov boundary.

Definition 6 (Essential variable) *A variable $W \in \mathcal{S}$ is called an essential variable for Y if $Y \not\perp\!\!\!\perp W \mid \mathcal{S} \setminus \{W\}$. Denote the set of all essential variables by $\mathcal{E}$.*

A variable W is essential if it can provide additional information on Y, even when we have known all variables except Y. In Fig. 1, W is the only essential variable, since X and Z contain the same information on Y.

Lemma 5 *The set $\mathcal{E}$ is the intersection of all Markov boundaries of Y within $\mathcal{S}$.*

Theorem 2 *Variable Y has a unique Markov boundary within $\mathcal{S}$ if and only if $\mathcal{E}$ is a Markov boundary of Y within $\mathcal{S}$.*

Theorem 2 provides a theoretical basis for Algorithm 2 that determines if the output from Algorithm 1 is a unique Markov boundary.

1. **Input**
   Joint distribution of $\mathcal{S} = \{X_1, \ldots, X_k\}$ and Y
   An algorithm Ω which could produce one Markov boundary correctly

2. **Set** $\mathcal{M}_0 = \{X_1, \ldots, X_m\}$ to be the result of Algorithm Ω on $\mathcal{S}$

3. **For** $i = 1, \ldots, m$,
   Set $\mathcal{M}_i$ to be the result of Algorithm Ω on $\mathcal{S} \setminus \{X_i\}$

   If $Y \perp\!\!\!\perp \mathcal{M}_0 \mid \mathcal{M}_i$

      Output Y has multiple Markov boundaries
      Terminate

4. **Output** Y has a unique Markov boundary

Algorithm 2: A general algorithm for determining uniqueness of Markov boundary

Algorithm 2 is closely related to the proposal that finds all the Markov boundaries for Y in Statnikov et al. (2013). In fact, Algorithm 2 can be viewed as running the proposal in Statnikov et al. (2013) until it produces two Markov boundaries or terminates.

Proposition 5 *Algorithms 1 and 2 are sound and complete.*

Remark 6 *The test in step (3) of Algorithm 2 aims to decide if X_i is an essential variable. Alternatively, one may directly test*

$$X_i \perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}. \tag{4}$$

This results in Algorithm S1 described in the supplementary material. However, the conditional set $\mathcal{S} \setminus \{X_i\}$ is generally very large, so that the conditional independence test for (4) *may have low power.*

Remark 7 *A naive algorithm based on Theorem 2 involves first constructing the set of essential variables $\mathcal{E}$ in $\mathcal{S}$, and then testing if $Y \perp\!\!\!\perp \mathcal{S} \mid \mathcal{E}$. This results in Algorithm S2 described in the supplementary material.*

5 Simulation studies

We now evaluate the finite sample performance of the proposed methods. In our simulations, the response variable Y and ten possible parents of Y, denoted as $\mathcal{S} = \{X_1, \ldots, X_{10}\}$, are all generated from Bernoulli distributions with mean 0.5. We consider four settings that cover various scenarios regarding the uniqueness of Markov boundaries and the composition property of $\mathcal{S} \cup Y$.

Setting 1: $X_1, \ldots, X_{10}$ are independent. $\text{pr}(Y = X_1) = 0.8, \text{pr}(Y = X_2) = 0.1$ and $\text{pr}(Y = X_3) = 0.1$. In this case, Y has a unique Markov boundary $\{X_1, X_2, X_3\}$.

Setting 2: Same as Setting 1, except that $X_4 = X_2$. In this case, Y has an additional Markov boundary $\{X_1, X_3, X_4\}$. In Settings 1 and 2, the composition property holds for $\mathcal{S} \cup \{Y\}$.

Setting 3: $X_1, \ldots, X_8$ are independent. $Z = X_1 + X_2$ mod 2. $\text{pr}(Y = Z) = 0.8, \text{pr}(Y = X_3) = 0.1, \text{pr}(Y = X_4) = 0.1, \text{pr}(X_9 = X_{10} = Z) = 0.95$ and $\text{pr}(X_9 = X_{10} = 1 - Z) = 0.05$. In this case, Y has a unique Markov boundary: $\{X_1, X_2, X_3, X_4\}$.

Setting 4: $X_1, \ldots, X_7$ are independent. $Z = X_1 + X_2$ mod 2. $\text{pr}(Y = Z) = 0.8, \text{pr}(Y = X_3) = 0.1, \text{pr}(Y = X_4) = 0.1, \text{pr}(X_{10} = Z) = 0.95$ and $\text{pr}(X_{10} = 1 - Z) = 0.05$. $X_8 = X_1, X_9 = X_2$. In this case, Y has two Markov boundaries: $\{X_1, X_2, X_3, X_4\}$ and $\{X_3, X_4, X_8, X_9\}$. In Settings 3 and 4, the distribution of $\mathcal{S} \cup \{Y\}$ violates the composition property.

We compare the performance of the following algorithms that test the uniqueness of Markov boundaries for the response variable Y: (1) Alg. 2-AF: Algorithm 2, with Ω being Algorithm 1; (2) Alg. 2-KI: Algorithm 2, with Ω being the KIAMB algorithm proposed in Peña et al. (2007), which requires the composition property; (3) Alg. S1; (4) Alg. S2. The Monte Carlo

size is 500, and we report the success rates for each algorithm. In each setting, we run all four algorithms with sample size ranging from 300 to 30,000. The conditional independence test we employ is the G-test (Neapolitan, 2004) with significance level $\alpha = 0.001$. All simulations are conducted with R. Following Peña et al. (2007) and Statnikov et al. (2013), we choose the parameter K to be 0.8 in the KIAMB algorithm.

As shown in Fig. 2 and Fig. 3, both Alg. 2-AF and Alg. 2-KI have satisfactory performance under Settings 1 and 2 where the composition property holds. Alg. S1 falsely claims that there are multiple Markov boundaries for Y until the sample size approaches 10,000. This is because failure to reject the hypothesis $X_i \perp\!\!\!\perp Y \mid \mathcal{S} \setminus \{X_i\}$ leads one to conclude that Y has multiple Markov boundaries. As expected, in Settings 3 and 4 where the composition condition fails to hold, Alg. 2-AF performs much better than Alg. 2-KI. As the sample size increases, each independence test is more likely to produce correct result. When sample size is large enough, each algorithm has a high probability to produce correct final result (except Alg. 2-KI in Settings 3 and 4).

We also find that the performance of Alg. S2 is not monotonic with the number of observations. A possible explanation is that although the error rate of each single test decreases with the number of observations, certain combinations of incorrect intermediate test results might by chance, lead to a correct final result. As the number of observations increases, the power for the independence test in step (3) of Alg. S2 increases so that the size of the empirical essential variable set $\hat{\mathcal{E}}$ grows. As a result, it is more likely that $Y \perp\!\!\!\perp \mathcal{S} \mid \hat{\mathcal{E}}$ holds. On the other hand, with a larger sample size one also gains power to reject the hypothesis that $Y \perp\!\!\!\perp \mathcal{S} \mid \hat{\mathcal{E}}$. This explains the non-monotonic curves we see with Alg. S2.

On average, when the composition property holds, the performance of Alg. 2-KI is slightly better than that of Alg. 2-AF, and both are much better than Alg. S1 and Alg. S2. Furthermore, Alg. 2-KI is faster than Alg. 2-AF in computation time (results not shown). When the composition property fails, Alg. 2-KI fails to produce correct results, while Alg. 2-AF exhibits the best performance.
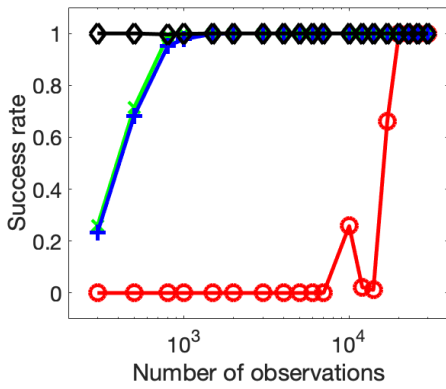
In practice, if one has a strong belief in the composition property, then we recommend Alg. 2-KI. Otherwise Alg. 2-AF is preferable.
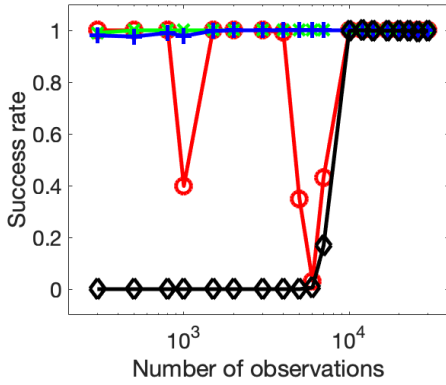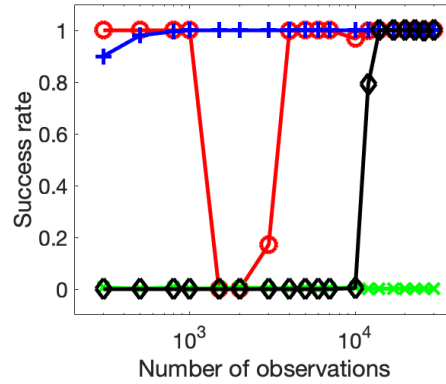
(a) Setting 1: unique Markov boundary, composition holds.
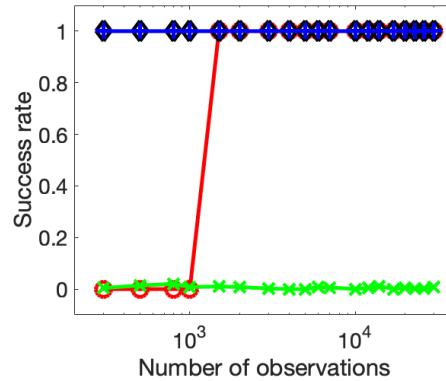

(b) Setting 2: multiple Markov boundaries, composition holds.

Figure 2: Performance of various algorithms for testing the uniqueness of Markov boundary, Settings 1, 2: proposed Alg. 2-AF (blue '+'); Alg. 2-KI (green '×'); Alg. S1 (black '◇'); Alg. S2 (red '○'). The number of observations ranges from 300 to 30,000. The x-axis is in logarithm scale.

(a) Setting 3: unique Markov boundary, composition fails.

(b) Setting 4: multiple Markov boundaries, composition fails.

Figure 3: Performance of various algorithms for testing the uniqueness of Markov boundary, Settings 3, 4: proposed Alg. 2-AF (blue '+'); Alg. 2-KI (green '×'); Alg. S1 (black '◇'); Alg. S2 (red '○'). The number of observations ranges from 300 to 30,000. The x-axis is in logarithm scale.

References

C.F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X.D. Koutsoukos. 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(Jan):171–234, 2010.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

R. L. Dobrushin. General formulation of Shannon's main theorem in information theory. *Amer. Math. Soc. Trans.*, 33:323–438, 1963.

L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: Is there a unique set? *Bioinformatics*, 21(2):171–178, 2004.

W. Gao, S. Kannan, S. Oh, and P. Viswanath. Conditional dependence via shannon capacity: Axioms, estimators and applications. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf. Quantifying causal influences. *Ann. Stat.*, 41(5):2324–2358, 2013.

S. Mani and G.F. Cooper. Causal discovery using a bayesian local causal discovery algorithm. *Medinfo*, 11(Pt 1):731–735, 2004.

R.E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

J. Pearl. *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann, San Mateo, 1988.

J. Pearl and A. Paz. *Graphoids: A Graph-based Logic for Reasoning about Relevance Relations*. University of California (Los Angeles). Computer Science Department, 1985.

J. Pearl. *Causality*. Cambridge University Press, 2009.

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

S. Rodrigues de Morais and A. Aussem. A novel Markov boundary based feature subset selection algorithm. *Neurocomput.*, 73(4):578–584, 2010.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

A. Statnikov, N.I. Lytkin, J. Lemeire, and C.F. Aliferis. Algorithms for discovery of multiple Markov boundaries. *J. Mach. Learn. Res.*, 14(Feb):499–566, 2013.

I. Tsamardinos and C.F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *Ann. Stat.*, 41(2): 436–463, 2013.

J. Zhao, Y. Zhou, X. Zhang, and L. Chen. Part mutual information for quantifying direct associations in networks. *Proc. Natl. Acad. Sci.*, 113(18):5130–5135, 2016.