

ASP-based Discovery of Semi-Markovian Causal Models under Weaker Assumptions

Zhalama^{1*}, Jiji Zhang², Frederick Eberhardt³, Wolfgang Mayer¹ and Mark Junjie Li⁴

¹University of South Australia

²Lingnan University

³California Institute of Technology

⁴Shenzhen University

zhalama@mymail.unisa.edu.au, jijizhang@ln.edu.hk, fde@caltech.edu, wolfgang.mayer@unisa.edu.au, jj.li@szu.edu.cn

Abstract

In recent years the possibility of relaxing the so-called Faithfulness assumption in automated causal discovery has been investigated. The investigation showed (1) that the Faithfulness assumption can be weakened in various ways that in an important sense preserve its power, and (2) that weakening of Faithfulness may help to speed up methods based on Answer Set Programming. However, this line of work has so far only considered the discovery of causal models without latent variables. In this paper, we study weakenings of Faithfulness for constraint-based discovery of semi-Markovian causal models, which accommodate the possibility of latent variables, and show that both (1) and (2) remain the case in this more realistic setting.

1 Introduction

Causal inference is of great interest in many scientific areas, and automated discovery of causal structure from data is drawing increasingly more attention in the field of machine learning. One of the standard approaches to automated causal discovery, known as the constraint-based approach, seeks to infer from data statistical relations among a set of random variables, and translate those relations into constraints on the underlying causal structure so that features of the causal structure may be determined from the constraints [Spirtes *et al.*, 2000; Pearl, 2000]. In this approach, the most commonly used constraints are in the form of conditional (in)dependence, which can serve as constraints on the causal structure due in the first place to the well known causal Markov assumption. The assumption states roughly that a causal structure, as represented by a directed acyclic graph (DAG), entails a certain set of conditional independence statements. With this assumption, a conditional dependency found in the data constrains the causal DAG.

The causal Markov assumption is almost universally accepted by researchers on causal discovery. However, by itself the assumption is too weak to enable interesting causal inference [Zhang, 2013]. It is therefore usually supplemented with

an assumption known as Faithfulness, which states roughly that unless entailed by the causal structure according to the Markov assumption, no conditional independence relation should hold. With this assumption, conditional independence relations found in the data also constrain the causal DAG.

Unlike the causal Markov assumption, the Faithfulness assumption is often regarded as questionable. The standard defense of the assumption is that violations of Faithfulness involve fine-tuning of parameters (such as two causal pathways balancing out exactly), which is very unlikely if we assume parameter values are somehow randomly chosen. However, parameter values may not be randomly chosen, especially in situations where balancing of multiple causal pathways may be part of the design. More importantly, even if the true distribution is faithful to the true causal structure, with finite data, “apparent violations” of faithfulness can result from errors in statistical tests, when a false hypothesis of conditional independence fails to be rejected. Such apparent violations of faithfulness cannot be reasonably assumed away [Uhler *et al.*, 2013] and will bring troubles to causal discovery that assumes Faithfulness [Meek, 1996; Robins *et al.*, 2003].

For these reasons, in recent years the possibility of relaxing the Faithfulness assumption has been investigated [Ramsey *et al.*, 2006; Zhang and Spirtes, 2008; Zhang, 2013; Spirtes and Zhang, 2014; Raskutti and Uhler, 2014; Forster *et al.*, 2017]. This line of work made it clear that in the context of learning causal models with no latent variables, the Faithfulness assumption can be weakened or generalized in a number of ways while retaining its inferential power, because in theory these assumptions all reduce to the Faithfulness assumption when the latter happens to hold.

On a more practical note, causal discovery algorithms have also been developed to fit some of these weaker assumptions, most notably the Conservative PC algorithm [Ramsey *et al.*, 2006] and the greedy permutation-based algorithms [Wang *et al.*, 2017; Solus *et al.*, 2017]. More systematically, Zhalama *et al.* [2017] implemented and compared a number of weakenings of Faithfulness in the flexible approach to causal discovery based on Answer Set Programming (ASP) [Hyttinen *et al.*, 2014]. Among other things, they found, rather surprisingly, that some weakenings significantly boost the time efficiency of ASP-based algorithms. Since the main drawback of the ASP-based approach lies with its feasibility, this find-

*Contact Author

ing is potentially consequential for the further development of this approach.

However, neither the theoretical investigation nor the ASP-based practical exploration went beyond the limited (and unrealistic) context of learning causal models in the absence of latent confounding, also known as causal discovery with the assumption of *causal sufficiency* [Spirtes *et al.*, 2000]. Since latent confounding is ubiquitous, it is a serious limitation to restrict the study to causally sufficient settings. And it is especially unsatisfactory from the perspective of the ASP-based approach, which boasts the potential to deal with a most general search space that accommodates the possibility of latent confounding and that of causal loops [Hytinen *et al.*, 2013].

In this paper, we make a step towards remedying this limitation by generalizing the aforementioned investigation in a setting where latent confounding is allowed (but not causal loops; we remark on a complication that will arise in the presence of causal loops in the end.) Since the investigation appeals to the ASP-based platform, we will follow previous work on this topic to use semi-Markovian causal models to represent causal structures with latent confounders. Among other things, we show that it remains the case that (1) the Faithfulness assumption can be weakened in various ways that in an important sense preserve its power, and (2) weakening of Faithfulness may help to speed up ASP-based methods.

The remainder of the paper will proceed as follows. In Section 2, we introduce terminologies and describe the basic setup. In Section 3, we review a few ways to relax the Faithfulness assumption that have been proposed in the context of causal discovery with causal sufficiency and have been proved to be *conservative* in a sense we will specify. Then, in Section 4, we discuss the complications that arise with semi-Markovian causal models, and establish generalizations of the results mentioned in Section 3. This is followed by a discussion in Section 5 of how to implement the weaker assumptions in the ASP platform. Finally, we report some simulation results in Section 6 that demonstrate the speed-up mentioned above, and conclude in Section 7.

2 Preliminaries

In this paper, the general graphical representation of a causal structure is by way of a *mixed graph*. The kind of mixed graph we will use is a triple $(\mathbf{V}, \mathbf{E}_1, \mathbf{E}_2)$, where $\mathbf{V}$ is a set of vertices (each representing a random variable), $\mathbf{E}_1$ a set of directed edges ($\rightarrow$) and $\mathbf{E}_2$ a set of bi-directed edges ($\leftrightarrow$). In general, more than one edge is allowed between two vertices, but no edge is allowed between a vertex and itself. Two vertices are said to be *adjacent* if there is at least one edge between them. Given an edge $X \rightarrow Y$, X is called a *parent* of Y and Y a *child* of X . We also say the edge has a *tail* at X and an *arrowhead* at Y . An edge $X \leftrightarrow Y$ is said to have an arrowhead at both X and Y . A *path* between X and Y consists of an ordered sequence of distinct vertices $\langle X = V_1, \dots, V_n = Y \rangle$ and a sequence of edges $\langle E_1, \dots, E_{n-1} \rangle$ such that for $1 \leq i \leq n-1$, E_i is an edge between V_i and V_{i+1} . Such a path is a *directed path* if for all $1 \leq i \leq n-1$, E_i is a directed edge from V_i to V_{i+1} . X is an *ancestor* of Y and Y a *descendant* of X , if either $X = Y$ or

there is a directed path from X to Y . A *directed cycle* occurs when two distinct vertices are ancestors of each other.

If a mixed graph does not contain any directed cycle, we will call it a *semi-Markovian causal model* (SMCM), also known as an *acyclic directed mixed graph* (ADMG). Intuitively a directed edge in an SMCM represents a direct causal relationship, and a bi-directed edge represents the presence of latent confounding. A directed acyclic graph (DAG) is a special case where no bi-directed edge appears. A DAG can be thought of as representing a causal model over a *causally sufficient* set of random variables, which may be referred to as a Markovian causal model (MCM).

The conditional independence statements entailed by a graph can be determined graphically by a separation criterion. One statement of this criterion is *m-separation*, which is a natural generalization of the celebrated *d-separation* criterion for DAGs [Pearl, 1988]. Given any path in a mixed graph G , a non-endpoint vertex V on the path is said to be a *collider* on the path if both edges incident to V on the path have an arrowhead at V . Otherwise it is said to be a *non-collider* on the path.

Definition 1 (m-connection and m-separation). Given a mixed graph G over $\mathbf{V}$ and $\mathbf{Z} \subseteq \mathbf{V}$, a path in G is *m-connecting given $\mathbf{Z}$* if every non-collider on the path is not in $\mathbf{Z}$ and every collider on the path has a descendant in $\mathbf{Z}$.

For any distinct $X, Y \notin \mathbf{Z}$, X and Y are *m-separated by $\mathbf{Z}$* in G (written as $X \perp_G Y \mid \mathbf{Z}$) if there is no path between X and Y that is m-connecting given $\mathbf{Z}$. Otherwise X and Y are said to be *m-connected by $\mathbf{Z}$* .

For any $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ that are pairwise disjoint, $\mathbf{X}$ and $\mathbf{Y}$ are m-separated by $\mathbf{Z}$ in G if every vertex in $\mathbf{X}$ and every vertex in $\mathbf{Y}$ are m-separated by $\mathbf{Z}$.

This definition obviously reduces to that of d-connection and d-separation in the case of DAGs. It is well known that in a DAG, two vertices are adjacent if and only if no set of other vertices d-separates them. The ‘only if’ direction holds for SMCMs, but the ‘if’ direction does not. For example, in the simple SMCM in Figure 1, V_1 and V_3 are not adjacent, but neither the empty set nor the set $\{V_2\}$ m-separates them. This motivates the following definition.

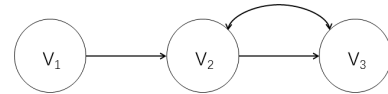


Figure 1: An inducing path between two non-adjacent vertices

Definition 2 (inducing path). A path between X and Y is an *inducing path* if every non-endpoint vertex on the path is a collider and also an ancestor of either X or Y .

For example, in Figure 1, the path $V_1 \rightarrow V_2 \leftrightarrow V_3$ is an inducing path between V_1 and V_3 . In general, two vertices in an SMCM are not m-separated by any set of other variables if and only if there is an inducing path between them [Verma, 1993]. Note that an edge between two vertices constitutes an inducing path. Following Richardson [1997], we call two ver-

tices *virtually adjacent* if there is an inducing path between them. Adjacency entails virtual adjacency, but not vice versa.

3 Faithfulness and its weakening for learning causal models without latent variables

We now review some proposals of weakening the Faithfulness assumption in the context of learning (acyclic) causal structures in the absence of latent confounding. In such a case, the target is a DAG over the given set of random variables $\mathbf{V}$, in which each edge represents a direct causal relation relative to $\mathbf{V}$ [Spirtes *et al.*, 2000]. Let G denote the unknown true causal DAG over $\mathbf{V}$, and P denote the true joint probability distribution over $\mathbf{V}$. The causal Markov assumption can be formulated as:

Causal Markov assumption For every pairwise disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, if $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$, then $\mathbf{X} \perp_P \mathbf{Y} \mid \mathbf{Z}$.

where ' $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$ ' means that $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in G , and ' $\mathbf{X} \perp_P \mathbf{Y} \mid \mathbf{Z}$ ' means that $\mathbf{X}$ and $\mathbf{Y}$ are independent conditional on $\mathbf{Z}$ according to P .

The converse is the Faithfulness assumption:

Causal Faithfulness assumption For every pairwise disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, if $\mathbf{X} \perp_P \mathbf{Y} \mid \mathbf{Z}$, then $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$.

As mentioned earlier, the Faithfulness assumption is regarded as much more questionable than the Markov assumption, and the literature has seen a number of proposals to relax it. In this paper, we focus on the following three.¹

Adjacency-faithfulness assumption For every distinct $X, Y \in \mathbf{V}$, if X and Y are adjacent in G , then $X \not\perp_P Y \mid \mathbf{Z}$, for every $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$.

Number-of-Edges(NoE)-minimality assumption: G is NoE-minimal in the sense that no DAG with a smaller number of edges than G satisfies the Markov assumption with P .

Number-of-Independencies(NoI)-minimality assumption: G is NoI-minimal in the sense that no DAG that entails a greater number of conditional independence statements than G does, satisfies the Markov assumption with P .

Under the Markov assumption, these assumptions are all weaker than the Faithfulness assumption. In words, Adjacency-faithfulness says that two variables that are adjacent in the causal structure are dependent given any conditioning set. It was first introduced in Ramsey *et al.* [2006] and motivated the CPC (conservative PC) algorithm. NoE-minimality says that the true causal structure has the least number of edges among all structures that satisfy the Markov

¹Another two proposals are known as 'Triangle-Faithfulness plus SGS-minimality' [Spirtes and Zhang, 2014] and 'P-minimality' [Zhang, 2013]. It is not yet clear how to implement the latter in ASP, and the former did not seem to help much with ASP-based methods [Zhalama *et al.*, 2017].

assumption. It underlies the novel permutation-based algorithms that were developed recently [Raskutti and Uhler, 2014; Wang *et al.*, 2017; Solus *et al.*, 2017]. NoI-minimality says that the true causal structure entails the greatest number of conditional independence statements among all structures that satisfy the Markov assumption. In the ASP-based methods, the 'hard-deps' conflict resolution scheme in Hyttinen *et al.* [2014] happened to implement this minimality constraint.

Theoretically these assumptions are particularly interesting because although they are weaker than Faithfulness (given the Markov assumption), they are in a sense strong enough to preserve the inferential power of Faithfulness. It has been shown that when Faithfulness happens to hold, all these weaker assumptions become equivalent to Faithfulness [Zhalama *et al.*, 2017]. In other words, while they are weaker than Faithfulness and therefore still hold in many cases when Faithfulness does not, they rule out exactly the same causal graphs as Faithfulness does when the latter happens to be satisfied. We propose to call this kind of weakening *conservative*, for it retains the inferential power of Faithfulness whenever Faithfulness is applicable. The choice between a stronger assumption and a weaker one usually involves a trade-off between risk (of making a false assumption) and inferential power, but there is no such trade-off if the weakening is conservative.

In addition to this theoretical virtue, both Adjacency-faithfulness and NoE-minimality, and especially Adjacency-faithfulness, have been shown to significantly improve the time efficiency of ASP-based causal discovery methods, without significant sacrifice in performance [Zhalama *et al.*, 2017]. We aim to extend these findings to the much more realistic setting where latent confounding may be present.

4 Weakening Faithfulness for learning semi-Markovian causal models

When the set of observed variables $\mathbf{V}$ is not causally sufficient, which means that some variables in $\mathbf{V}$ share a common cause or confounder that is not observed, it is no longer appropriate to represent the causal structure in question with a DAG over $\mathbf{V}$. One option is to explicitly invoke latent variables in the representation and assume the underlying causal structure is properly represented by a DAG over $\mathbf{V}$ plus some latent variables $\mathbf{L}$. Another option is to suppress latent variables and use bi-directed edges to represent latent confounding. The use of SMCMs exemplifies the latter approach.²

As Verma [1993] showed, for every DAG over $\mathbf{V}$ and set of latent variables $\mathbf{L}$, there is a unique projection into an SMCM over $\mathbf{V}$ that preserves both the causal relations among $\mathbf{V}$ and the entailed conditional independence relations among $\mathbf{V}$. Moreover, as Richardson [2003] pointed out, the original causal DAG with latent variables and its projection into an SMCM are equivalent regarding the (nonparametric) identification of causal effects. These facts justify using SMCMs to represent causal structures with latent confounding.

So let us suppose the underlying causal structure over $\mathbf{V}$ is properly represented by an SMCM G and let P denote

²Another important example is the use of ancestral graph Markov models [Richardson and Spirtes, 2002], which we describe in the appendices.

the true joint distribution over $\mathbf{V}$. In this setting, the causal Markov and Faithfulness assumptions can be formulated as before (in Section 3), except that the separation criterion is now understood as the more general m-separation. Next we examine the proposals of weakening Faithfulness.

Regarding Adjacency-faithfulness, it is easy to see that it remains a logical consequence of Faithfulness. If two variables are adjacent in an SMCM, then given any set of other variables, the two are m-connected (any edge between them constitutes a m-connecting path). Thus, if Faithfulness holds, then they are not independent conditional on any set of other variables, exactly what is required by Adjacency-faithfulness. Since Adjacency-faithfulness does not entail Faithfulness in the case of DAGs and DAGs are special cases of SMCMs, Adjacency-faithfulness remains weaker than Faithfulness.

However, it is now too weak to be a conservative weakening of Faithfulness. Here is a very simple example. Suppose the true causal structure over three random variables is a simple causal chain $V_1 \rightarrow V_2 \rightarrow V_3$, and suppose the joint distribution is Markov and Faithful to this structure. So we have $V_1 \perp\!\!\!\perp V_3 \mid V_2$. Then the distribution is not Faithful to the structure in Figure 1, because that structure does not entail that V_1 and V_3 are m-separated by V_2 . Still, Adjacency-faithfulness is satisfied by the distribution and the structure in Figure 1, for the only violation of Faithfulness occurs with regard to V_1 and V_3 , which are not adjacent. Therefore, in this simple case where Faithfulness happens to hold, if we just assume Adjacency-faithfulness, we are not going to rule out the structure in Figure 1, which would be ruled out if we assumed Faithfulness.

This simple example suggests that we should consider the following variation:

V(virtual)-adjacency-faithfulness assumption: For every distinct $X, Y \in \mathbf{V}$, if X and Y are virtually adjacent in G (i.e., if there is an inducing path between X and Y in G), then $X \not\perp\!\!\!\perp Y \mid \mathbf{Z}$, for every $\mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\}$.

V-adjacency-faithfulness is obviously stronger than Adjacency-faithfulness, but we can prove that it remains weaker than Faithfulness. More importantly, it is strong enough to be a conservative weakening of Faithfulness.

How about NoE-minimality? Since more than one edge can appear between two vertices, NoE-minimality (as it is formulated in Section 3) is no longer a consequence of Faithfulness. To see this, just suppose the true structure over two random variables is simply $V_1 \rightarrow V_2$ together with $V_1 \leftrightarrow V_2$ (i.e., V_1 is a cause of V_2 but the relation is also confounded), and suppose the distribution is Markov and Faithful to this structure. NoE-minimality is violated here, for taking away either (but not both) of the two edges still results in a structure that satisfies the Markov assumption.

So NoE-minimality is not a weakening of Faithfulness. Note that in the case of DAGs, minimization of the number of edges is equivalent to minimization of the number of adjacencies. If we replace the former with the latter, the above example is taken care of (for taking away the adjacency in that example will result in a structure that fails the Markov assumption). However, it is also easy to construct an example

where an adjacency in an SMCM can be taken away without affecting the independence model [Richardson and Spirtes, 2002], so adjacency-minimality also fails to be a weakening of Faithfulness. The right generalization of NoE-minimality is unsurprisingly the following:

V(virtual)-adjacency-minimality assumption: G is V-adjacency-minimal in the sense that no SMCM with a smaller number of virtual adjacencies than G satisfies the Markov assumption with P .

Finally, since NoI-minimality is concerned with entailed conditional independence statements, it is straightforwardly generalized to the setting of SMCMs (just replace ‘DAG’ with ‘SMCM’ in the original formulation), and remains a conservative weakening of Faithfulness. Here then is the main result of this section (a proof of which is given in Appendix C).

Theorem. Given the causal Markov assumption, the V-adjacency-faithfulness assumption, V-adjacency-minimality assumption, and NoI-minimality assumptions are all conservative weakenings of the Faithfulness assumption, in the following sense: for each of the three assumptions **AS**,

- (a) **AS** is entailed by, but does not entail, Faithfulness.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists an SMCM that satisfies both Markov and Faithfulness assumptions with P , then for every SMCM G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies **AS** with P .

5 ASP-based Causal Discovery of SMCMs

We instantiated causal discovery algorithms, which adopt V-adjacency-faithfulness and V-adjacency-minimality, using the framework of Hyttinen *et al.* [2014]. This framework offers a generic constraint-based causal discovery method based on Answer Set Programming (ASP). The logic is used to define Boolean atoms that represent the presence or absence of a directed or bi-directed edge in an SMCM. In addition, conditional independence/dependence statements (CI/CDs) obtained from tests on the input data are encoded in this logic. Finally, background assumptions, such as Markov and Faithfulness, are written as logical constraints enforcing a correspondence between the encoded test results and the underlying Boolean atoms (the edges of the SMCM). Solutions, which are truth-value assignments to the Boolean atoms, satisfying such a correspondence are found using off-the-shelf solvers. The set of solutions specifies the set of SMCMs that satisfy all the input CI/CDs and the background assumptions. Given that the results of the statistical tests may conflict with the background assumptions, there may be no solution, i.e. there is no SMCM that satisfies all the input CI/CDs and background assumptions. For that case Hyttinen *et al.* [2014] introduced the following optimization to resolve the conflict:

$$G^* \in \arg \min_{G \in \mathcal{G}} \sum_{k \in \mathbf{K} \text{ s.t. } G \neq k} w(k) \quad (1)$$

In words, an output graph G^* minimizes the weighted sum of input CI/CDs, which it does not satisfy given the encoded

background assumptions. Hyttinen *et al.* [2014] adopted three weighting schemes for the weights $w(\cdot)$: (1) “constant weights” (CW) assigns a weight of 1 to each CI and CD constraint. (2) “hard dependencies” (HW/NoI-m) assigns infinite weight to any observed CD, and a weight of 1 to any CI. (3) “log weights” (LW) is a pseudo-Bayesian weighting scheme, where the weights depend on the log posterior probability of the CI/CDs being true (see their Sec. 4).

To encode V-adjacency-faithfulness and V-adjacency-minimality, we need to encode in ASP what it is for an SMCM to have an inducing path and a virtual adjacency, and then replace the encoding of the Faithfulness assumption in Hyttinen *et al.* [2014] with its weaker versions. Figure 2 summarizes the ASP-encoding of V-adjacency-faithfulness and V-adjacency-minimality. We briefly explain the predicates:

- $edge(X, Z)$ and $conf(X, Z)$: $X \rightarrow Z$ and $X \leftrightarrow Z$, respectively, are in the SMCM.
- $ancestors(Z, X, Y)$: Z is an ancestor of X or Y in the SMCM.
- $h(X, Z, Y)$: There is a path between X and Z which is into Z , and if the path consists of two or more edges, every non-endpoint vertex on the path is a collider and every vertex on the path is an ancestor of either X or Y .
- $t(X, Z, Y)$: It differs from $h(X, Z, Y)$ only in that the path between X and Z is out of Z . Together, $t(\cdot)$ and $h(\cdot)$ are used to specify the possible inducing paths.
- $vadj(X, Y)$: X and Y are virtually adjacent.
- $indep(X, Y, \mathbf{C}, w)$: X and Y are independent conditional on $\mathbf{C}$, given as input fact, with weight w .

For V-adjacency-faithfulness, we encode that any CI statement $X \perp\!\!\!\perp Y \mid \mathbf{C}$ implies that X and Y are not virtually adjacent. For V-adjacency-minimality, we employ the minimization of the number of virtual-adjacencies. By encoding the weaker assumptions in the framework of Hyttinen *et al.* [2014], we then have the following algorithms (Hyttinen *et al.*’s algorithm based on the ‘hard dependencies’ weights is equivalent to one based on NoI-minimality):

- **VadjF**: Virtual-adjacency-faithfulness + Markov
- **VadjM**: Virtual-adjacency-minimality + Markov

6 Simulations

We report two types of simulation, one using an independence oracle that specifies the true CI/CDs of the causal model, and one that uses the CI/CDs inferred from the sample data.

For both simulations we followed the model generation process of Hyttinen *et al.* [2014] for causally insufficient models: We generated 100 random linear Gaussian models over 6 vertices with an average edge degree of 1 for directed edges. The edge coefficients were drawn uniformly from $[-0.8, -0.2] \cup [0.2, 0.8]$. The error covariance matrices (which also represent the confounding) were generated using the observational covariance matrix of a similarly constructed causally sufficient model (with its error covariances sampled from $N(0.5, 0.01)$).

Inference rules for virtual-adjacency:

$$\begin{aligned} h(X, Z, Y) &:- edge(X, Z), ancestors(Z, X, Y). \\ h(X, Z, Y) &:- conf(X, Z), ancestors(Z, X, Y). \\ h(X, Z, Y) &:- h(X, U, Y), conf(Z, U), ancestors(Z, X, Y). \\ t(X, Z, Y) &:- h(X, U, Y), edge(Z, U), ancestors(Z, X, Y). \\ vadj(X, Y) &:- h(X, Y, Y). \\ vadj(X, Y) &:- t(X, Y, Y). \\ vadj(X, Y) &:- edge(Y, X). \end{aligned}$$

Virtual-adjacency-faithfulness (violations):

$$\begin{aligned} \forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}, \\ &:- \text{not } vadj(X, Y), indep(X, Y, \mathbf{C}, w) \end{aligned}$$

Virtual-adjacency-minimality (optimization of weak constraints):

$$\begin{aligned} \forall X \forall Y > X, \\ fail(X, Y, w = 1) &:- vadj(X, Y). \\ &:\sim fail(X, Y, w). [w] \end{aligned}$$

(Variables are in an arbitrary order so that $indep(X, Y, \mathbf{C}, w)$ and $dep(X, Y, \mathbf{C}, w)$ are considered only if $Y > X$, in order to avoid double counting.)

Figure 2: ASP Encoding of V-adjacency-faithfulness and V-adjacency-minimality

In the oracle setting, we randomly generated 100 linear Gaussian models with latent confounders over 6 variables and then input the independence oracles implied by these models. We observed that the algorithms based on V-adjacency-faithfulness, on V-adjacency-minimality, and on NoI-minimality (which is equivalent to using ‘hard dependencies’ weighting) all returned the exact same results as the algorithm based on Faithfulness did, which is consistent with the theoretical results in Section 4 and confirms the correctness of our encoding.

In the finite sample case we generated five data sets with 500 samples from each of the 100 models. We used correlational t-tests and tried 10 threshold values for rejecting the null hypothesis (0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2, 0.25). The test results formed the input for the algorithms. We also used the log-weighting scheme and tried 10 values for the free parameter of the Bayesian test (0.05, 0.09, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9).

For each algorithm we output all possible solutions and compared the d-connections common to all the output graphs against those of the true data generating graph. In all the $100(\text{models}) \times 5(\text{datasets}) \times 10(\text{parameters}) = 5,000$ runs, Faithfulness was satisfied in only 367/5000 of the cases while V-adjacency-faithfulness was satisfied in 2065/5000 of the cases. This shows that V-adjacency-faithfulness is indeed significantly weaker than Faithfulness and greatly reduces the number of conflicts. By definition, V-adjacency-minimality can always be satisfied. Figure 3 plots the ROC curves for the inferred d-connections. Under “constant weighting” (CW) “hard dependencies weighting” (HW), using V-adjacency-faithfulness achieves comparable accuracy to using faith-

fulness, with some trade-offs between false-positive rates and true-positive rates. Under “log weighting” (LW), however, using Faithfulness seems slightly more accurate than using V-adjacency-faithfulness, though using V-adjacency-minimality seems to generally yield the lowest false-positive rates. How to adapt the “log weighting” to fit V-adjacency-faithfulness better is an interesting question for future work.

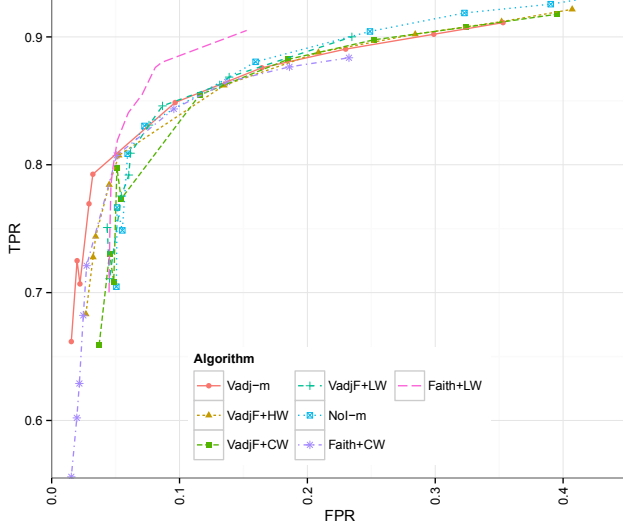


Figure 3: ROC of d-connections

Finally, to explore the efficiency gain of the weakened faithfulness assumptions, we generated 100 random linear Gaussian models with latent confounders over 8 variables and generated one data set with 500 samples from each model. For each algorithm, we only required that one graph be found. Figure 4 shows the sorted solving times for the different background assumptions (with maximum time budget of 5,000s). As in the causally sufficient case, we see a significant improvement in solving times when using the weakened faithfulness assumptions.

7 Conclusion

We have shown how to extend the results on weakening Faithfulness in the context of learning causal DAGs to the more realistic context of learning SMCMs that allow for the representation of unmeasured confounding. We identified generalizations of some proposals of weakening Faithfulness in the literature and showed that they continue to be what we call conservative weakenings. Moreover, we implemented ASP-based algorithms for learning SMCMs based on these weaker assumptions. The simulation results suggest that some of these weaker assumptions, especially V-adjacency-faithfulness, help to save solving time in ASP-based algorithms to a significant extent.

In this connection, a direction of future work is to explore how the apparent advantage of using weaker assumptions may be realized on top of other ASP-based causal discovery methods, such as ETIO in [Borboudakis and Tsamardinos, 2016] and ACI in [Magliacane *et al.*, 2016].

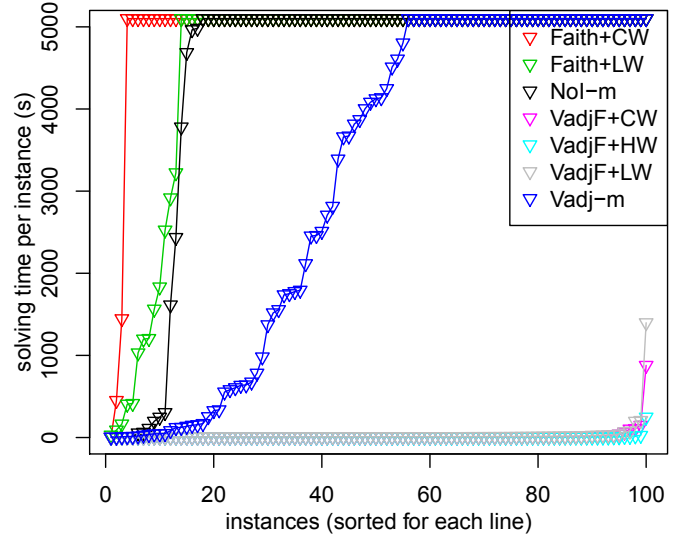


Figure 4: Sorted Solving Times for 8 Variables (time-out at 5,000s)

One great appeal of the ASP-based approach is that the background assumptions that determine the search space can be flexibly adjusted to include causal models with both latent confounding and causal feedback. We close with an illustration of a (further) complication that arises in cyclic causal models. Suppose the true causal structure is the cyclic one in Figure 5(a), which entails that $V_1 \perp V_2$ and $V_1 \perp V_2 \mid \{V_3, V_4\}$. Suppose the true distribution is Markov and Faithful to this structure and hence features exactly two non-trivial conditional independencies. Then the distribution is not Faithful to the structure in Figure 5(b) (for that structure does not entail $V_1 \perp V_2 \mid \{V_3, V_4\}$), but it is still V-adjacency-faithful (for V_1 and V_2 are not virtually adjacent).

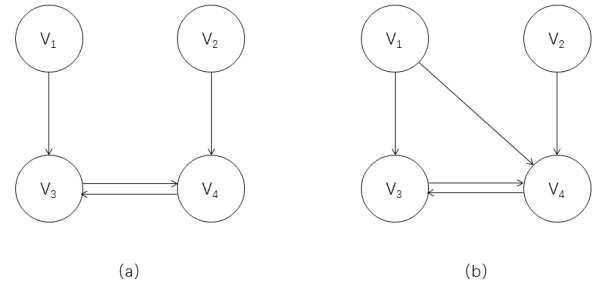


Figure 5: Illustration of a complication in cyclic models

This means that even V-adjacency-faithfulness is not a conservative weakening of Faithfulness when causal feedback is allowed. Whether it can be strengthened into a useful conservative weakening for the purpose of learning cyclic models is worth further investigation.

Acknowledgments

JZ was supported by GRF LU13602818 from the RGC of Hong Kong. FE was supported by NSF grant 1564330.

Appendices

Below we prove the theorem stated in the paper. The result is restated separately as three theorems in Section C below, one for each of the three weakenings of the Faithfulness assumption we considered. The proof makes use of maximal ancestral graphs (MAGs), for which useful characterizations of Markov equivalence are available. Below we proceed as follows. In Section A, we introduce MAGs and known facts that are relevant to our arguments. In Section B, we describe a connection between SMCMs and MAGs that we will exploit. Finally, in Section C, we prove the theorems.

A Maximal Ancestral Graphs

Like an SMCM, a MAG is a graphical object designed to represent a causal structure in the presence of latent variables. An *Ancestral graph* is a *simple* mixed graph (i.e., at most one edge can appear between any two vertices), in which for any two vertices X and Y , if X is an ancestor of Y , then there is no (directed or bi-directed) edge between X and Y that is into X . A *Maximal Ancestral Graph* is an ancestral graph in which for every pair of non-adjacent vertices, there exists some set of other vertices that m -separates them. A DAG is a special MAG which does not contain bi-directed edges.

Two DAGs are Markov equivalent if and only if they share the same adjacencies and unshielded colliders. Although Markov equivalent MAGs also share the same adjacencies and unshielded colliders, these commonalities are no longer sufficient to characterize Markov equivalence between MAGs. Two Markov equivalent MAGs may also share some shielded colliders. The definitions related to this fact are given below.

Definition 3. [Richardson and Spirtes, 2002] A path $\pi = (V_0, V_1, \dots, V_m = X, Z, Y)$ ($m \geq 1$) is a *discriminating path* for $\langle X, Z, Y \rangle$ in a MAG, if V_0 and Y are not adjacent and every vertex V_i , $1 \leq i \leq m$, is a collider on π and a parent of Y .

Definition 4. An *inducing path* π between X and Y is a path on which every non-endpoint vertex is a collider and an ancestor of either X or Y .

Definition 5. [Ali *et al.*, 2009]: Call $\langle X, Z, Y \rangle$ a triple if X, Z are adjacent and Y, Z are adjacent. The order of such a triple in a MAG is defined recursively as follows:

Order 0. A triple $\langle X, Z, Y \rangle$ has order 0 if X and Y are not adjacent.

Order i . A triple $\langle X, Z, Y \rangle$ has order i , if it does not have any order less than i , and there is a discriminating path $\langle V_0, V_1, \dots, V_m = X, Z, Y \rangle$ or $\langle V_0, V_1, \dots, V_m = Y, Z, X \rangle$, on which every collider triple centered at V_j ($1 \leq j \leq m$) has order at most $i - 1$ (and at least one of them has order $i - 1$).

A discriminating path for $\langle X, Z, Y \rangle$ is said to *have order i* if except for $\langle X, Z, Y \rangle$, every collider triple on the path has order less than i and at least one of them has order $i - 1$. Note that some triples in a graph may not have an order. Note also that the order (if any) of a shielded triple is the minimum of all discriminating paths with order for that triple [Ali *et al.*,

2009]. Colliders with order ≥ 1 are those shielded colliders that are present in all Markov equivalent MAGs.

Proposition 1 [Ali *et al.*, 2009] Two MAGs are Markov equivalent if and only if they have the same adjacencies and the same colliders with order.

B SMCMs and MAGs

SMCMs and MAGs are both generalizations of DAGs. Like in DAGs, directed cycles are not allowed in SMCMs or MAGs. But unlike DAGs, they can contain bi-directed edges ($\leftrightarrow$) in addition to the directed edges ($\rightarrow$). However, almost directed cycles — where one endpoint of a bi-directed edge is an ancestor of the other endpoint — are allowed in SMCMs but not in MAGs. For this reason at most one edge is allowed between any two variables in a MAG, while in SMCMs up to two edges (one directed and one bi-directed) are allowed. Besides, the interpretation of an edge in an SMCM is different from that in a MAG. A directed edge $X \rightarrow Y$ in an SMCM means that X is a direct cause of Y relative to $\mathbf{V}$. A bi-directed edge $X \leftrightarrow Y$ means that X and Y are confounded by a latent variable. In contrast, in a MAG a directed edge $X \rightarrow Y$ represents that X is a causal ancestor of Y , and a bi-directed edge $X \leftrightarrow Y$ means that X is not an ancestor of Y and Y is not an ancestor of X (which then imply that they are confounded by a latent variable). If X is a causal ancestor of Y (that is not mediated by any other observed variable), and they also have a common cause which is latent, then only a directed edge $X \rightarrow Y$ is present in the corresponding MAG.

For a causally insufficient system $\mathbf{V}$, we assume that there is a causal DAG G over $\mathbf{V} \cup \mathbf{L}$ (where $\mathbf{L}$ is a set of latent variables) that satisfies the causal Markov assumption. The set of m -separations (d-separations) entailed by G is called the independence model associated with G and denoted as $\mathbf{J}(G)$. The marginal independence model of $\mathbf{J}(G)$ over $\mathbf{V}$ after leaving out the set of latent variables $\mathbf{L}$ is the subset of m -separations (d-separations) entailed by G which do not involve any variables in $\mathbf{L}$. We denote it as $\mathbf{J}(G)|_{\mathbf{L}}$. Given G over $\mathbf{V} \cup \mathbf{L}$, there is a unique MAG M and a unique SMCM S over $\mathbf{V}$, which represent some causal relations among $\mathbf{V}$ in G and the marginal independence model of G over $\mathbf{V}$.

A DAG over $\mathbf{V}$ with latent variables $\mathbf{L}$ can be projected into such an SMCM in the following way [Verma, 1993; Tian and Pearl, 2002]:

Input: a DAG G over $\mathbf{V} \cup \mathbf{L}$

Output: an SMCM S_G over $\mathbf{V}$

1. Add each variable in $\mathbf{V}$ as a node of S_G .
2. For each pair of variables $X, Y \in \mathbf{V}$, if there is an edge between them in G , add the edge to S_G .
3. For each pair of variables $X, Y \in \mathbf{V}$, if there is a directed path from X to Y in G such that every mid node on the path is in $\mathbf{L}$, add edge $X \rightarrow Y$ to S_G , if it does not exist yet.
4. For each pair of variables $X, Y \in \mathbf{V}$, if there exist a directed path from a variable $L_i \in \mathbf{L}$ to X and a directed path from L_i to Y in G such that every mid node on the

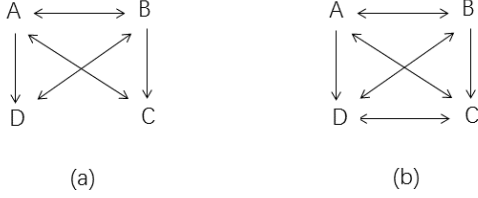


Figure 6: An SMCM (left) and its corresponding MAG (right) in which C and D are not m-separated by any subset of $\{A, B\}$.

paths is in $\mathbf{L}$, add edge $X \leftrightarrow Y$ to S_G , if it does not exist already.

The conversion of a DAG with latent variables to a MAG is given in Richardson and Spirtes [2002].

Given a set of variables $\mathbf{V}$, For every SMCM over $\mathbf{V}$, there is also a unique MAG that corresponds to it, such that they entail the same set of conditional independence statements (and the causal relations represented by the MAG are compatible with those represented by the SMCM). The following is a procedure to transform an SMCM into its corresponding MAG. It is adapted from the algorithm presented in Zhang [2008], which is used to project DAGs with latent variables to MAGs.

Conversion from an SMCM to a MAG

- S 1: For each pair of variables X and Y , X and Y are adjacent in the output MAG M , if there is an inducing path between them in the input SMCM S .
- S 2: For each pair of adjacent variables X and Y in the output MAG M ,
- (i) If X is an ancestor of Y in S , orient the edge as $X \rightarrow Y$.
 - (ii) If Y is an ancestor of X in S , orient the edge as $Y \rightarrow X$.
 - (iii) else, orient the edge as $X \leftrightarrow Y$.

It is worth noting that in MAGs, two variables are adjacent if and only if they are not m-separated by any set of other variables. This is not true for SMCMs, as shown in Figure 6. However, in SMCMs it holds that there is an inducing path between two variables if and only if the two variables are not m-separated by any set of other variables. For example, in Figure 6(a), there is an inducing path (C, A, B, D) between C and D . We write the property as a proposition for later reference.

Proposition 2 If S is an SMCM over $\mathbf{V}$, M is the corresponding MAG of S over $\mathbf{V}$, and $X, Y \in \mathbf{V}$, then the following statements are equivalent:

- 1) There is an inducing path between X and Y in S .
- 2) X and Y are m-connected given any subset of $\mathbf{V} \setminus \{X, Y\}$.
- 3) X and Y are adjacent in M .

Proof. A proof of the equivalence of 2) and 3) is given in the proof of Theorem 4.2 in Richardson and Spirtes [2002].

The equivalence of 1) and 3) can be proved in the same way as the equivalence of (ii) and (iii) of Theorem 4.2 in Richardson and Spirtes [2002].

C Conservative Weakenings of Faithfulness

In this section, we prove the theorem in the paper, via generalizations of the weakenings of Faithfulness to MAGs and establishing related results. In MAGs, it holds that if two vertices X and Y are adjacent, then there is no subset of $\mathbf{V} \setminus \{X, Y\}$ that m-separates them. Hence, the formulation of the Adjacency-faithfulness established for DAGs can be directly applied to MAGs, except that the definition of adjacency allows for bi-directed edges. Likewise, the formulation of Number-of-Edges(NOE)-minimality carries over to MAGs except that bi-directed edges are allowed in MAGs. The Number-of-Independences(NoI)-minimality assumption, which is given in terms of conditional independence statements, can be directly extended to MAGs. In the following, we prove that for MAGs, all of the generalized assumptions remain conservative weakenings of Faithfulness. Based on these results, we prove the corresponding theorems for SMCMs.

Lemma 1 Given the causal Markov assumption, the Adjacency-faithfulness assumption is a conservative weakening of the Faithfulness assumption in the case of MAGs, in the following sense:

- (a) the Adjacency-faithfulness assumption is entailed by, but does not entail, the Faithfulness assumption.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists a MAG that satisfies both Markov and Faithfulness assumptions with P , then for every MAG G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies the Adjacency-faithfulness assumption with P .

Proof. (a) Let G_F be a MAG over $\mathbf{V}$ to which P is both Markov and faithful. If two variables X and Y are adjacent in G_F , they are not m-separated given any set $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$. Then, since P is faithful to G_F , X and Y are not independent given any set $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ in P . Thus, G_F satisfies Adjacency-faithfulness. Therefore, the Adjacency-faithfulness assumption is entailed by the Faithfulness assumption.

We show that Adjacency-faithfulness does not entail Faithfulness with an example. Suppose that $\mathbf{V} = \{V_1, V_2, V_3\}$ and the conditional independence relations satisfied by the distribution are $V_1 \perp\!\!\!\perp V_3$ and $V_1 \perp\!\!\!\perp V_3 \mid V_2$ (see Zhang, 2013 for an example of such a distribution.) Then given Markov, the structure $V_1 \rightarrow V_2 \leftrightarrow V_3$ satisfies Adjacency-faithfulness but not Faithfulness with the distribution.

(b) Suppose there exists a MAG G_F that is both Markov and faithful to P . The “only if” direction has already been proved in (a), so we just need to prove that, in this case, a MAG satisfies the Faithfulness assumption with P if it satisfies Adjacency-faithfulness assumption with P . In other words, we just need to prove that in this case, every MAG

G_A that satisfies Markov Adjacency-faithfulness with P is Markov equivalent to G_F .

By Proposition 1, the MAGs that are Markov equivalent have the same adjacencies and colliders with order. It is easy to see that G_A shares the same adjacencies with G_F . The proof given for DAGs [Ramsey *et al.*, 2006] is directly applicable to MAGs. So we only need to prove that G_A and G_F have the same colliders with order. For this purpose, we first prove the following claim.

Claim: Let $\pi = (V_0, V_1, \dots, V_m = X, Z, Y) (m \geq 1)$ be a discriminating path for Z in a MAG G_F which satisfies the Markov and the Faithfulness assumption with P . If the corresponding path π^* forms a discriminating path in a MAG G_A that satisfies Markov with P , then Z is a collider on π in G_F if and only if Z is a collider on π^* in G_A .

Proof. As every vertex $V_i, 1 \leq i \leq m$ is a collider on path $(V_0, V_1, \dots, V_m = X, Z, Y)$ and a parent of Y , any set that m -separates V_0 and Y must include all of them.

If (X, Z, Y) is a collider in G_F , the path $(V_0, V_1, \dots, V_m = X, Z, Y)$ m -connects given any superset of $\mathbf{S} = \{V_1, \dots, V_m\} \cup \{Z\}$. As G_F satisfies the Faithfulness assumption with P , V_0 and Y are dependent given any superset of $\mathbf{S}$ in P . If (X, Z, Y) is not a collider in G_A , there must be some superset of $\mathbf{S}$ that m -separates V_0 and Y . But then G_A would violate the Markov assumption, which is a contradiction. Hence, (X, Z, Y) is also a collider in G_A .

If (X, Z, Y) is not a collider in G_F , then the path $(V_0, V_1, \dots, V_m = X, Z, Y)$ m -connects given any superset of $\{V_1, \dots, V_m\}$ which does not include Z . By Faithfulness, V_0 and Y are dependent given any superset of $\{V_1, \dots, V_m\}$ which does not include Z in P . If (X, Z, Y) is a collider in G_A , any set that m -separates V_0 and Y must be a superset of $\{V_1, \dots, V_m\}$ which does not include Z . But then G_A would violate the Markov assumption, which is a contradiction. Thus, (X, Z, Y) is not a collider in G_A . $\square$

Now we can prove that a triple (X, Z, Y) is a collider with order r in G_F if and only if (X, Z, Y) is a collider with order r in G_A .

Let r be the order of (X, Z, Y) .

If $r = 0$, (X, Z, Y) is an unshielded collider. The proof given for DAGs in Ramsey *et al.* [2006] is directly applicable here. The only difference is that in MAGs, colliders and non-colliders admit more edge configurations than they do in DAGs.

When $r > 0$, assume the result holds for all $s < r$.

If (X, Z, Y) is a triple with order r in $G_F(G_A)$, by the definition of ordered triple, there exists a discriminating path $\pi = (V_0, V_1, \dots, V_m = X, Z, Y)$ (or $\pi = (V_0, V_1, \dots, V_m = Y, Z, X)$) in $G_F(G_A)$ such that, except (X, Z, Y) , every triple $V_i (1 \leq i \leq m)$ on π is a collider and has order less than r . And since G_F and G_A have the same adjacencies, the sequence of vertices forming the discriminating path in $G_F(G_A)$, also forms a path in $G_A(G_F)$. Let π^* be the corresponding path in $G_A(G_F)$. By the inductive hypothesis, in

$G_A(G_F)$, each collider $V_i (1 \leq i \leq m)$ is also a collider with the same order as in $G_F(G_A)$ on the corresponding path π^* . We claim that the corresponding path π^* is also a discriminating path in $G_A(G_F)$ for (X, Z, Y) . Since we have $V_0? \rightarrow V_1 \leftrightarrow \dots \leftrightarrow V_m \leftarrow ?Z$ in $G_A(G_F)$, it suffices to show that $V_j \rightarrow Y (1 \leq j \leq m)$ in $G_A(G_F)$.

Triple (V_0, V_1, Y) is a noncollider with order 0 in $G_F(G_A)$, because V_0 and Y are not adjacent. Hence, (V_0, V_1, Y) is also a noncollider with order 0 in $G_A(G_F)$. Further, as $V_0? \rightarrow V_1$, by the definition of MAGs, $V_1 \rightarrow Y$ in $G_A(G_F)$. Arguing inductively, assume $V_i \rightarrow Y (1 < i < m)$ in $G_A(G_F)$, so that $(V_0, V_1, \dots, V_m, Y)$ forms a discriminating path with order at most r for (V_{m-1}, V_m, Y) in both G_A and G_F . As a consequence, as (V_{m-1}, V_m, Y) is a noncollider on $(V_0, V_1, \dots, V_m, Y)$ in $G_F(G_A)$, (V_{m-1}, V_m, Y) is a noncollider on $(V_0, V_1, \dots, V_m, Y)$ in $G_A(G_F)$, based on the claim established above. Since $V_{m-1} \rightarrow V_m$, by the definition of MAGs, $V_m \rightarrow Y$ in $G_A(G_F)$. Hence, π^* also forms a discriminating path in $G_A(G_F)$ for (X, Z, Y) . Again, based on Lemma 4, (X, Z, Y) is a collider in G_F if and only if (X, Z, Y) is a collider in G_A .

By the definition of an ordered triple, (X, Z, Y) has order at most r in G_A . However, if (X, Z, Y) has order less than r in G_A , by the inductive hypothesis, (X, Z, Y) will have order less than r in G_F , which is a contradiction. Thus, (X, Z, Y) has order r in both graphs.

To summarize, G_A and G_F are Markov equivalent since they have the same adjacencies and colliders with order. $\square$

Theorem 1 Given the causal Markov assumption, the V-adjacency-faithfulness assumption is a conservative weakening of the Faithfulness assumption in the case of SMCMs, in the following sense:

- (a) V-adjacency-faithfulness is entailed by, but does not entail, Faithfulness.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists an SMCM that satisfies both Markov and Faithfulness assumptions with P , then for every SMCM G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies the V-adjacency-faithfulness with P .

Proof. (a) Let S_F be an SMCM which satisfies Markov and Faithfulness with P . By Proposition 2, if two variables X and Y are virtually adjacent in S_F , they are m -connected given any subset of $\mathbf{V} \setminus \{X, Y\}$. Then by Faithfulness, X and Y are dependent conditional on any subset of $\mathbf{V} \setminus \{X, Y\}$. So S_F satisfies V-adjacency-faithfulness. Hence, the V-adjacency-faithfulness assumption is entailed by the Faithfulness assumption.

However, V-adjacency-faithfulness does not entail Faithfulness. It can be illustrated with the same example used in the proof of Lemma 1, since syntactically MAGs are special cases of SMCMs.

(b) Now we prove the “if” direction, since the “only if” direction has already been proved in (a). Let S_I be an SMCM which satisfies Markov and V-adjacency-faithfulness with P and M_I be the unique MAG corresponding to S_I .

If X and Y are adjacent in M_I , there is an inducing path between X and Y in S_I , based on Proposition 2. Then, X and Y are not independent given any subset of $\mathbf{V} \setminus \{X, Y\}$, since S_I satisfies V-adjacency-faithfulness. Thus, M_I satisfies Adjacency-faithfulness. And, if P is faithful to some SMCM, it is faithful to the corresponding MAG of that SMCM, which means that there exists a MAG that satisfies both Markov and Faithfulness with P . Further, by Lemma 1, M_I satisfies the Faithfulness assumption. It follows that S_I satisfies the Faithfulness assumption, since M_I and S_I entail exactly the same CI statements. $\square$

Lemma 2 Given the causal Markov assumption, the NOE-minimality assumption is a conservative weakening of the Faithfulness assumption in the case of MAGs, in the following sense:

- (a) the NOE-minimality assumption is entailed by, but does not entail, the Faithfulness assumption.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists a MAG that satisfies both Markov and Faithfulness assumptions with P , then for every MAG G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies the NOE-minimality assumption with P .

Proof. (a) Let G_F be a MAG to which P is both Markov and faithful. Then removing any edge from G_F will either violate the maximality or introduce an independence which is not satisfied by P , which constitutes a violation of the Markov assumption. Hence, there is no MAG with a smaller number of edges than G_F , which satisfies Markov. So G_F satisfies NOE-minimality. Thus, NOE-minimality assumption is entailed by the Faithfulness assumption.

Forster *et al.* [2017] showed that NOE-minimality is weaker than Faithfulness for DAGs. Thus, in the case of MAGs, NOE-minimality does not entail Faithfulness, since DAGs are special cases of MAGs.

(b) Suppose there exists a MAG G_F that is both Markov and faithful to P , and suppose G_{NoE} is a MAG that satisfies Markov and NOE-minimality with the distribution P . As we did previously, to prove the “if” direction, we only need to prove that G_F and G_{NoE} have the same adjacencies and colliders with order.

G_F and G_{NoE} have the same number of edges since G_F also satisfies NOE-minimality. Now we prove that G_F and G_{NoE} not only have the same number of edges but also the same adjacencies: For a contradiction, if we assume that G_{NoE} has one different edge than G_F does, then one edge that is present in G_F is removed in G_{NoE} , since they share the same number of edges. As already mentioned, removing any edge that is present in G_F will result in a violation of either maximality or Markov. Thus, G_F and G_{NoE} not only have the same number of edges but also the same adjacencies.

Next, we prove that G_F and G_{NoE} have the same unshielded colliders. Since G_F and G_{NoE} have the same adjacencies, a triple is unshielded in G_F if and only if it is unshielded in G_{NoE} . If an unshielded triple (X, Z, Y) is an unshielded collider in G_F , then X and Y are dependent given

any set that includes Z in P , because the distribution P is faithful to G_F . Then, as P and G_{NoE} satisfy the Markov assumption, (X, Z, Y) is also an unshielded collider in G_{NoE} . Similarly, if (X, Z, Y) is an unshielded non-collider in G_F , then it is an unshielded non-collider in G_{NoE} .

The fact that G_F and G_{NoE} have the same colliders with order can be proved in the same way as we did in the proof of Lemma 1, since we have already proved that G_F and G_{NoE} have the same adjacencies and unshielded colliders. $\square$

Theorem 2 Given the causal Markov assumption, the V-adjacency-minimality assumption is a conservative weakening of the Faithfulness assumption in the case of SMCMs, in the following sense:

- (a) V-adjacency-minimality is entailed by, but does not entail, Faithfulness.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists an SMCM that satisfies both Markov and Faithfulness assumptions with P , then for every SMCM G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies V-adjacency-minimality with P .

Proof. (a) Let S_{AF} be an SMCM which satisfies Markov and V-adjacency-faithfulness with P . Then if two variables X and Y are virtually adjacent in S_{AF} , they are dependent given any subset of $\mathbf{V} \setminus \{X, Y\}$ in the distribution P . Then, by Proposition 2, taking away the virtual adjacency between X and Y will introduce a new conditional independence, which is not satisfied by the distribution P and would thus result in a violation of the Markov assumption. So S_{AF} satisfies V-adjacency-minimality, which means that V-adjacency-minimality is entailed by V-adjacency-faithfulness. Further, V-adjacency-minimality is entailed by Faithfulness since V-adjacency-faithfulness is entailed by Faithfulness.

However, V-adjacency-minimality does not entail the Faithfulness assumption because V-adjacency-faithfulness does not entail Faithfulness.

(b) Now we only need to prove the “if” direction of (b) since the “only if” direction has already been proved above. Let S_{VADJ} be an SMCM which satisfies Markov and V-adjacency-minimality and M_{VADJ} be the unique MAG corresponding to S_{VADJ} .

By Proposition 2, there is an inducing path in S_{VADJ} if and only if there is an edge in M_{VADJ} . If M_{VADJ} does not satisfy NOE-minimality, there must be some MAG M , which has fewer edges than M_{VADJ} and still satisfies Markov. The SMCMs corresponding to M then have fewer virtual-adjacencies than S_{VADJ} and also still satisfy Markov. This violates our initial assumption that S_{VADJ} satisfies V-adjacency-minimality. Thus, M_{VADJ} satisfies NOE-minimality. And, when P is faithful to some SMCM, it is faithful to the corresponding MAG of this SMCM, which means that there exists a MAG that satisfies both Markov and Faithfulness with P . Further, by Lemma 2, M_{VADJ} satisfies faithfulness, since it satisfies NOE-minimality. Hence S_{VADJ} satisfies faithfulness, since it entails the same exact CIs with M_{VADJ} . $\square$

Lemma 3 Given the causal Markov assumption, the NOI-minimality assumption is a conservative weakening of the Faithfulness assumption in the case of MAGs, in the following sense:

- (a) the NOI-minimality assumption is entailed by, but does not entail, the Faithfulness assumption.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists a MAG that satisfies both Markov and Faithfulness assumptions with P , then for every MAG G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies the NOI-minimality assumption with P .

Proof. (a) The proof that was given for DAGs [Zhalama *et al.*, 2017] is directly applicable to MAGs.

(b) Let G_F be one of the graphs to which P is both Markov and faithful and G_{NoI} be a graph that satisfies Markov and NOI-minimality with P . Since G_F also satisfies NOI-minimality, G_F and G_{NoI} entail the same number of conditional independence statements (CIs). Now we prove that they entail exactly the same CIs. For a contradiction, let's assume that G_{NoI} entails one CI that is not entailed by G_F . Because G_F satisfies faithfulness and Markov with P , the CIs entailed by G_F are exactly the ones satisfied by P . But since G_{NoI} entails one CI that is not satisfied by P , it then must violate Markov, which is a contradiction. Therefore, G_F and G_{NoI} entail the exact same CIs, which means that G_{NoI} also satisfies Markov and faithfulness with P . $\square$

Theorem 3 Given the causal Markov assumption, the NOI-minimality assumption is a conservative weakening of faithfulness in the case of SMCMs, in the following sense:

- (a) NOI-minimality is entailed by, but does not entail, Faithfulness.
- (b) For every joint probability distribution P over $\mathbf{V}$, if there exists an SMCM that satisfies both Markov and Faithfulness assumptions with P , then for every SMCM G that satisfies the Markov assumption with P , G satisfies Faithfulness if and only if G satisfies NOI-minimality with P .

Proof. The proof of Lemma 3 can be directly extended to SMCMs. $\square$

References

- R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B):2808–2837, 10 2009.
- G. Borboudakis and I. Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of KDD*, pages 1435–1444, 2016.
- M. Forster, G. Raskutti, R. Stern, and N. Weinberger. The frugal inference of causal relations. *British Journal for the Philosophy of Science*, 2017.
- A. Hyttinen, P.O. Hoyer, F. Eberhardt, and M. Järvisalo. Discovering cyclic causal models with latent variables: A general SAT-based procedure. In *Proceedings of UAI*, pages 301–310. AUAI Press, 2013.
- A. Hyttinen, F. Eberhardt, and M. Järvisalo. Constraint-based causal discovery: Conflict resolution with Answer Set Programming. In *Proceedings of UAI*, 2014.
- S. Magliacane, T. Claassen, and J.M. Mooij. Ancestral causal inference. In *Advances In Neural Information Processing Systems*, pages 4466–4474, 2016.
- C. Meek. *Graphical Causal Models: Selecting Causal and Statistical Models*. PhD thesis, Department of Philosophy, Carnegie Mellon University, 1996.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.
- J. Pearl. *Causality*. Oxford University Press, 2000.
- J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of UAI*, pages 401–408, 2006.
- G. Raskutti and C. Uhler. Learning directed acyclic graphs based on sparsest permutations. *arXiv:1307.0366v3*, 2014.
- T. Richardson and P. Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962–1030, 2002.
- T. Richardson. A characterization of markov equivalence for directed cyclic graphs. *International Journal of Approximate Reasoning*, 17(2-3):107 – 162, 1997.
- T. Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003.
- J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.
- L. Solus, Y. Wang, L. Matejovicova, and C. Uhler. Consistency Guarantees for Permutation-Based Causal Inference Algorithms. *arXiv e-prints*, Feb 2017.
- P. Spirtes and J. Zhang. A uniformly consistent estimator of causal effects under the k-triangle-faithfulness assumption. *Statistical Science*, 29(4):662–678, 2014.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, 2 edition, 2000.
- J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *UAI 2002, Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 519–527. Morgan Kaufmann, 2002.
- C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of faithfulness assumption in causal inference. *The Annals of Statistics*, 41:436–463, 2013.
- T. Verma. Graphical aspects of causal model. Technical report, Computer Science Department, University of California, Los Angeles, 1993.
- Y. Wang, L. Solus, K. Yang, and C. Uhler. Permutation-based causal inference algorithms with interventions. In *NIPS*, pages 5824–5833, 2017.
- Zhalama, J. Zhang, F. Eberhardt, and W. Mayer. Sat-based causal discovery under weaker assumptions. In *UAI*, 2017.

- J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2):239–271, 2008.
- J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.
- J. Zhang. A comparison of three Occam’s razor for Markovian causal models. *British Journal for the Philosophy of Science*, 64(2):423–448, 2013.