SAT-Based Causal Discovery under Weaker Assumptions

Zhalama
University of South Australia
Adelaide, Australia

Jiji Zhang
Lingnan University
Tuen Mun, NT, Hong Kong

Frederick Eberhardt
Caltech
Pasadena, CA, USA

Wolfgang Mayer
University of South Australia
Adelaide, Australia

Abstract

Using the flexibility of recently developed methods for causal discovery based on Boolean satisfiability (SAT) solvers, we encode a variety of assumptions that weaken the Faithfulness assumption. The encoding results in a number of SAT-based algorithms whose asymptotic correctness relies on weaker conditions than are standardly assumed. This implementation of a whole set of assumptions in the same platform enables us to systematically explore the effect of weakening the Faithfulness assumption on causal discovery. An important effect, suggested by simulation results, is that adopting weaker assumptions greatly alleviates the problem of conflicting constraints and substantially shortens solving time. As a result, SAT-based causal discovery is potentially more scalable under weaker assumptions.

1 INTRODUCTION

In the framework of causal graphical models [Spirtes et al., 2000, Pearl, 2000], the task of causal discovery can be framed as the inference from a statistical dataset of measurements of a set of variables to the underlying causal structure that gave rise to the data. The inference is enabled by a set of assumptions or "bridge principles" that link statistical features of the data to features of the underlying causal structure. Two of the best known assumptions of this kind are the Causal Markov and Faithfulness assumptions. These two assumptions, defined and discussed below, together entail an exact correspondence between conditional independence relations that hold in the distribution (from which the data are drawn) and certain separation features in the underlying causal structure. This correspondence enabled the development of so-called constraint-based causal discovery algorithms — such as PC and FCI [Spirtes et al., 2000] — that exploit the independence structure found in the data, to recover as much as possible about the causal structure.

Although the causal Markov assumption is usually regarded as an ontological principle grounded in the nature of causation, the Faithfulness assumption is often viewed as a methodological assumption in the spirit of Occam's razor [Zhang, 2013]. Moreover, even if Faithfulness holds in the true distribution, there are often "almost violations" of Faithfulness in finite-sample settings that affect causal discovery [Meek, 1996, Robins et al., 2003, Uhler et al., 2013]. These considerations sparked investigations into the possibility of relaxing the Faithfulness assumption in constraint-based causal discovery [Ramsey et al., 2006, Zhang and Spirtes, 2008, Zhang, 2013, Spirtes and Zhang, 2014, Raskutti and Uhler, 2014, Forster et al., 2017], which generated a number of proposals for weakening Faithfulness. However, the impact of these weaker assumptions on causal discovery has not been systematically investigated on a common platform, since constraint-based algorithms have until recently always been custom-built for a specific set of assumptions. In this paper we leverage the recent development of causal discovery algorithms based on general-purpose Boolean satisfiability solvers [Triantafilou et al., 2010, Hyttinen et al., 2013, 2014, Triantafillou and Tsamardinos, 2015, Borboudakis and Tsamardinos, 2016, Magliacane et al., 2016], and show how to implement various weaker assumptions in this flexible approach. We compare the resulting algorithms on synthetic data. Among other things, the results suggest that the weaker assumptions significantly reduce the need for conflict resolution, which translates into substantial gains in solving time.

We will proceed as follows. After introducing the basic terminology in Section 2, we review the Boolean satisifiability and optimization approach to causal discovery in Section 3, focusing on the framework proposed by Hyttinen et al. [2013, 2014]. Then, in Section 4, we go

through a number of weakenings of the Faithfulness assumption and show, for all except one, how to encode constraints that correspond to the weaker assumptions. In addition, we "weaken" the causal Markov assumption by implementing constraints associated with the local Markov assumption as opposed to those with the global Markov assumption. Despite their equivalence given a perfect oracle of conditional independence, the former is in general weaker than the latter when the oracle is replaced by finite sample tests. These various constraints in different combinations yield a multitude of algorithms for causal discovery of acyclic, causally sufficient structures, and we report some findings on their empirical performance in Section 5. We conclude in Section 6.

2 PRELIMINARIES

Throughout this paper we consider the setting of a set of observed variables $\mathbf{V}$ that is causally sufficient (there are no unmeasured common causes) and that does not feature causal feedback.[1] The (unknown) causal structure over $\mathbf{V}$ can then be properly represented by a directed acyclic graph (DAG) over $\mathbf{V}$, in which an arrow $X \to Y$ means that X is a direct cause of Y relative to $\mathbf{V}$.[2]

Some basic graph terminology: Given a DAG G over $\mathbf{V}$, for any $X, Y \in \mathbf{V}$, if there is an edge between X and Y, then X and Y are said to be *adjacent*. If the edge is directed from X to Y, i.e., $X \to Y$, then X is called a *parent* of Y and Y a *child* of X. A *path* in G is a sequence of distinct vertices $(V_1, ..., V_n)$ such that for $1 \leq i \leq n-1$, V_i and V_{i+1} are adjacent in G. A path between V_1 and V_n is called a *directed path* from V_1 to V_n if V_i is a parent of V_{i+1} for $1 \leq i \leq n-1$. X is called an *ancestor* of Y and Y a *descendant* of X in G if $X = Y$ or there is a directed path from X to Y in G.

An (ordered) triple of distinct vertices (X, Y, Z) in G is called a *collider* if X, Y are adjacent, Y, Z are adjacent, and the edge between X and Y and the edge between Z and Y are both directed at Y, i.e., $X \to Y \leftarrow Z$. It is called a *non-collider* if X, Y are adjacent, Y, Z are adjacent, and the triple is not a collider. Given a path $(V_1, ..., V_n)$, V_i $(1 < i < n)$ is said to be a collider (non-collider) on the path if the triple (V_{i-1}, V_i, V_{i+1}) is a collider (non-collider).

We are concerned with the constraint-based approach to causal discovery that seeks to recover causal structure from statistically inferred conditional independence (CI) and conditional dependence (CD) statements. A fundamental principle that is almost always assumed (for causally sufficient systems) is the following condition:

(Causal) Markov Assumption: The joint distribution p of $\mathbf{V}$ is Markov to the causal DAG G over $\mathbf{V}$ in the sense that every variable is independent of its non-descendants conditional on its parents in G.

This is a formulation in terms of the *local* Markov property of DAGs. It specifies some CI statements that must be true of p according to G. These CI statements entail others by the laws of probability, all of which are captured by the notion of *d-separation* [Pearl, 1988]. Given $\mathbf{Z} \subseteq \mathbf{V}$, a path in G is blocked by $\mathbf{Z}$ (or not d-connecting given $\mathbf{Z}$) if some non-collider on the path is in $\mathbf{Z}$ or some collider on the path has no descendant in $\mathbf{Z}$. For any distinct $X, Y \notin \mathbf{Z}$, X and Y are d-separated by $\mathbf{Z}$ in G (we write $X \perp_G Y \mid \mathbf{Z}$) if every path between X and Y is blocked by $\mathbf{Z}$. For any $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$ that are pairwise disjoint, $\mathbf{X}$ and $\mathbf{Y}$ are d-separated by $\mathbf{Z}$ in G if every vertex in $\mathbf{X}$ and every vertex in $\mathbf{Y}$ are d-separated by $\mathbf{Z}$. The Markov assumption can then be reformulated as: for every disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp\!\!\!\perp_p \mathbf{Y} \mid \mathbf{Z}$. We will refer to this formulation as the *global* Markov assumption, and we call a CI statement $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$ *entailed* by G if $\mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$.

Standard constraint-based methods also assume the converse of the Markov assumption:

(Causal) Faithfulness Assumption: The joint distribution p over $\mathbf{V}$ is faithful to the causal DAG G over $\mathbf{V}$ in the sense that for every disjoint $\mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{V}$, $\mathbf{X} \perp\!\!\!\perp_p \mathbf{Y} \mid \mathbf{Z} \Rightarrow \mathbf{X} \perp_G \mathbf{Y} \mid \mathbf{Z}$.

Under these two assumptions, the true causal DAG G is determined by the CI and CD statements that are true of p up to the Markov equivalence class of G. Two DAGs over $\mathbf{V}$ are *Markov equivalent* just in case they entail the exact same CI statements.

3 BOOLEAN SATISFIABILITY AND CONSTRAINED OPTIMIZATION

(Independence-)constraint-based methods that adopt the Markov and Faithfulness assumptions seek to infer the Markov equivalence class of the underlying causal DAG based on CI and CD constraints obtained from data. Standard methods use a graphical representation of the Markov equivalence classes — known as *patterns* (or *essential graphs*) — to perform causal discovery. In contrast, recent approaches to constraint-based causal discovery have attempted to directly encode the d-separation/connection constraints implied by CI/CD

[1] We focus on the relatively simple task of inferring acyclic, causally sufficient structures, because the literature on weakening the Faithfulness assumption has been so focused. However, at least some aspects of this work are readily generalizable to a much more general setting, as we indicate in Section 6.

[2] As usual, we use "variable" and "vertex" interchangeably.

statements in terms of Boolean constraints over the set of possible causal graphs [Hyttinen et al., 2013, 2014, Triantafillou and Tsamardinos, 2015]. Such encodings enable the use of general purpose Boolean constraint solvers and, importantly for our aims here, make the inference from the CI constraints to the graphical constraints flexible: e.g., whether an observed CI only implies that the two independent variables are non-adjacent in the causal graph or whether it implies that there is no d-connecting path of any kind between them, can simply be written into the rules that the solver must respect.

While the specific encodings and the modeling languages vary among extant approaches of this type, the underlying strategy is the same: Define Boolean atoms of the form $A =$ "$X \rightarrow Y \in G$" that specify the presence of a particular edge in the true graph, encode (in a schema) what it means for two variables to be (conditionally) d-connected/separated, and define predicates for the CI/CD-statements that then imply constraints on the d-connections present or absent in the underlying graph. Observed CI/CDs are then encoded as their corresponding predicates and off-the-shelf solvers are used to find truth value assignments to the Boolean atoms that satisfy all the constraints. Any satisfying truth value assignment then describes a causal graph G that is consistent with the input CI/CDs. However, given that the CI/CDs result from a statistical process, the set of observed CI/CDs may contain errors, which may result in a set of constraints that are unsatisfiable. That is, for such a "conflicted" set of constraints, there is no causal graph that satisfies the encoded assumptions. In that case, some method for conflict resolution is required.

Hyttinen et al. [2014], whose encoding forms the basis for our approach here, encoded the standard correspondence between conditional independence and d-separation licensed by the combination of the (global) Markov assumption and the Faithfulness assumption:

$$X \perp\!\!\!\perp_p Y \mid \mathbf{Z} \quad \Leftrightarrow \quad X \perp_G Y \mid \mathbf{Z} \tag{1}$$

In the case of a conflicted set of constraints $\mathbf{K}$, they used the following optimization to determine their output G^*:

$$G^* \in \arg\min_{G \in \mathcal{G}} \sum_{k \in \mathbf{K} \text{ s.t. } G \not\models k} w(k) \tag{2}$$

That is, G^* minimizes the weighted sum of unsatisfied constraints, where the weights $w(.)$ can be set in a variety of ways. Specifically, Hyttinen et al. [2014] consider three weighting schemes: (1) "constant weights" just assigns a weight of 1 to each constraint. (2) "hard dependencies" assigns infinite weight to any observed CD and weight of 1 to any CI. Finally, (3) "log weights" 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). Other weighting schemes based on p-values have been developed by Triantafillou and Tsamardinos [2015] and Magliacane et al. [2016].

4 WEAKENING THE ASSUMPTIONS

The Faithfulness assumption has been subject to forceful criticism [Cartwright, 2001, Hoover, 2001]. Examples of violations of Faithfulness are easily constructed by, for example, two causal pathways that cancel each other's effect exactly, resulting in an independence between variables that are (doubly) causally connected. We may find such violations in cases of back-up mechanisms or control systems, where cancellations of causal pathways are part of the design. More problematically, in finite sample inference violations of faithfulness will occur when a (weak) statistical dependence is not detected due to low sample size. Uhler et al. [2013] have shown that such "almost-violations" of Faithfulness are in a measure-theoretic sense fairly common. These considerations together with the recognition that some violations of Faithfulness are in fact testable, have led to a whole set of weaker versions of Faithfulness.

Below we describe a number of weaker assumptions that have been studied in the literature. Let G denote the true causal DAG over $\mathbf{V}$ and p the true joint distribution.

Adjacency-faithfulness: For every distinct $X, Y \in \mathbf{V}$ and $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$, if X and Y are adjacent in G, then $X \not\perp\!\!\!\perp_p Y \mid \mathbf{C}$.

Adjacency-faithfulness is a consequence of Faithfulness and is used to justify the step of inferring adjacencies in algorithms like PC [Spirtes et al., 2000]. The Conservative PC algorithm [Ramsey et al., 2006], for example, assumes Adjacency-faithfulness instead of Faithfulness.

Triangle-faithfulness: For every distinct $X, Y, Z \in \mathbf{V}$ such that (X, Z, Y) is a triangle (i.e., they are pairwise adjacent) in G: (i) If (X, Z, Y) is a collider in G, then for every $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ that includes Z, $X \not\perp\!\!\!\perp_p Y \mid \mathbf{C}$; and (ii) If (X, Z, Y) is a non-collider, then for every $\mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$ that excludes Z, $X \not\perp\!\!\!\perp_p Y \mid \mathbf{C}$.

Triangle-faithfulness is much weaker than even Adjacency-faithfulness [Zhang and Spirtes, 2008, Spirtes and Zhang, 2014] and is of particular interest in conjunction with the following, very weak assumption.

SGS-minimality: G is SGS-minimal in the sense that no proper subgraph[3] of G satisfies the Markov assumption with p.

[3] A proper subgraph of G is a DAG over $\mathbf{V}$ whose set of edges is a proper subset of that of G.

SGS-minimality is so weak that given a standard, interventional interpretation of causal DAGs, it is guaranteed to be true if p is positive [Zhang and Spirtes, 2011]. Zhang and Spirtes (2008) showed that under the assumptions of Triangle-faithfulness and SGS-minimality, Faithfulness (and so Adjacency-faithfulness) are in principle testable without knowing the causal structure.

There are also a few minimality assumptions that are stronger than SGS-minimality but nonetheless weaker than Faithfulness. For example, Pearl [2000] discussed the following minimality condition:

P-minimality: G is P-minimal in the sense that no proper independence-submodel of G satisfies the Markov assumption with p.[4]

The GES algorithm [Chickering, 2002] can be viewed as aiming to find a P-minimal pattern [Zhalama et al., 2017]. Finally, there are count-based minimalities that are somewhat stronger. One of them says that the true causal DAG is sparest among all that are Markov to p.

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

Note the contrast to SGS-minimality: Both SGS-minimality and NoE-minimality are concerned with edges, but NoE-minimality is a minimality with respect to the order defined over the number of edges, whereas SGS-minimality is a minimality with respect to the (partial) order defined by the inclusion relation between sets of edges. NoE-minimality is called "Frugality" by Forster et al. [2017], who showed that it is stronger than P-minimality but weaker than Faithfulness. Raskutti and Uhler (2014) proposed a *Sparsest Permutation Algorithm* based on NoE-minimality. Sonntag et al. (2015) applied the same idea in the context of learning chain graphs.

The other minimality condition we will consider states that the true causal DAG entails the greatest number of CI statements among all DAGs that are Markov to p.

Number-of-Independencies(NoI)-minimality: 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.

NoI-minimality is to P-minimality as NoE-minimality is to SGS-minimality: NoI-minimality is a minimality with respect to the order defined over numbers of entailed CI statements, whereas P-minimality is a minimality with respect to the (partial) order defined by the inclusion relation between sets of entailed CI statements. To our knowledge, NoI-minimality has not been explicitly studied in the literature. However, as we will see in Section 4.1, one of the weighting schemes considered by Hyttinen et al. [2014], namely "hard dependencies", can be interpreted as implementing precisely this assumption.

NoI-minimality is obviously stronger than P-minimality, in the same way that NoE-minimality is stronger than SGS-minimality. On the other hand, it is weaker than Faithfulness. It is easy to see that Faithfulness entails NoI-minimality: if p is faithful to G, then any DAG that entails more CIs than G does, must entail some CI that is not satisfied by p and so does not satisfy the Markov assumption with p. Conversely, NoI-minimality does not entail Faithfulness. To show this, it suffices to note that there are distributions over $\mathbf{V}$ that are not both Markov and Faithful to any DAG over $\mathbf{V}$ [Zhang and Spirtes, 2008], but every distribution over $\mathbf{V}$ is Markov to some DAG over $\mathbf{V}$ that is NoI-minimal.

To summarize, we shall consider the following weakenings of the Faithfuless assumption: (i) Adjacency-faithfulness; (ii) the conjunction of SGS-minimality and Triangle-faithfulness; (iii) NoE-minimality; and (iv) NoI-minimality. (We do not consider P-minimality as we have not yet developed a concise encoding of this assumption.) It is worth noting that although these are weaker assumptions than Faithfulness, they become equivalent to Faithfulness for distributions that admit a DAG representation that is both Markov and Faithful. In other words, given the Markov assumption, (i)-(iv) are weaker than Faithfulness *only* because they are consistent with bigger sets of distributions than Faithfulness is [Zhang and Spirtes, 2016]. That is, assuming Faithfulness for distributions from those bigger sets, can result in "conflicted" or unsatisfiable sets of constraints even without statistical errors. Adopting (i) or (ii) significantly reduces but does not completely eliminate conflicts (with or without statistical errors), while adopting (iii) or (iv) completely eliminates conflicts. Indeed, an implementation of (iv) in the SAT approach is precisely a conflict resolution scheme considered by Hyttinen et al. [2014].

4.1 IMPLEMENTATIONS IN ASP

We now show how to implement constraints corresponding to these weaker assumptions in Answer Set Programming (ASP), the framework used in Hyttinen et al. [2014], which we take as basis. ASP is a Prolog-style modeling language expressed in terms of first-order logical rules [Baral, 2010]. None of the details of ASP, the corresponding solvers or even the details of Hyttinen et al.'s encoding are needed to understand what follows.

[4] A *proper independence-submodel* of G is a DAG over $\mathbf{V}$ such that the set of CIs it entails is a proper superset of that entailed by G. We borrow the names "SGS-minimality" (Spirtes, Glymour, and Scheines's minimality) and "P-minimality" (Pearl's minimality) from [Zhang, 2013].

One could similarly express the weakened assumptions in propositional logic and use a standard SAT-solver (or Boolean constraint optimizer) for the inference task.

As mentioned in Section 3, the Markov assumption is encoded as constraints that CD statements impose on the underlying causal structure: $X \not\perp\!\!\!\perp Y \mid \mathbf{C}$ implies that there is a d-connecting path between X and Y given $\mathbf{C}$; the Faithfulness assumption is encoded as constraints that CI statements impose on the underlying causal structure: $X \perp\!\!\!\perp Y \mid \mathbf{C}$ implies that there is no d-connecting path between X and Y given $\mathbf{C}$. Thus, the weakenings of the Faithfulness assumption will be encoded by modifying the constraints imposed by CI statements.

This is most straightforward for Adjacency-faithfulness and Triangle-faithfulness. To encode the former, we simply take a CI statement $X \perp\!\!\!\perp Y \mid \mathbf{C}$ to imply the constraint that there is no edge between X and Y. To encode the latter, we take a CI statement $X \perp\!\!\!\perp Y \mid \mathbf{C}$ to imply the constraint that there is no triangle (X, Z, Y) such that either (X, Z, Y) is a collider and $Z \in \mathbf{C}$, or (X, Z, Y) is a non-collider and $Z \notin \mathbf{C}$.

The minimality assumptions are a little less straightforward. For SGS-minimality, a reformulation will help. Given the Markov assumption, SGS-minimality can be equivalently formulated as: for every $X \in \mathbf{V}$ and every non-empty $\mathbf{P} \subseteq \mathbf{PA}_G(X)$ (where $\mathbf{PA}_G(X)$ denotes the set of parents of X in G), $X \not\perp\!\!\!\perp_p \mathbf{P} \mid \mathbf{PA}_G(X) \setminus \mathbf{P}$. According to this reformulated SGS-minimality, a CI statement $X \perp\!\!\!\perp Y \mid \mathbf{C}$ implies the constraint that $\mathbf{PA}_G(X) \neq \mathbf{C} \cup \{Y\}$ and $\mathbf{PA}_G(Y) \neq \mathbf{C} \cup \{X\}$.

For NoI-minimality and NoE-minimality we employ an optimization of (weak) constraint satisfication, since both essentially assume that the true causal DAG is "optimal" in some sense among all DAGs that satisfy the Markov assumption (that is, satisfy the constraints imposed by CD statements.) The difference is that NoI-minimality aims at maximizing the number of d-separation relations, whereas NoE-minimality aims at maximizing the number of non-adjacencies.

For NoI-minimality, we can keep the original encoding of the constraint associated with a CI statement. We assign each such CI constraint weight 1, while taking every constraint associated with a CD statement as a *hard constraint* that must be satisfied (or assigning them weight ∞). Then, given a perfect oracle of CI statements, the DAGs that minimize the total weight of unsatisfied constraints, as expressed in equation (2), are precisely those that are Markov and NoI-minimal. As we said, this implementation of NoI-minimality is exactly one of the conflict resolution schemes considered by Hyttinen et al. [2014]. Thus, this way of conflict resolution can be motivated from the perspective of weakening Faithfulness.

For NoE-minimality, we can ignore CI statements, and take each possible non-adjacency as a constraint with weight 1 (while taking every constraint associated with a CD statement as a hard constraint). Then, given a perfect oracle of CI statements, the DAGs that minimize the total weight of unsatisfied constraints are precisely those that are Markov and NoE-minimal.

In addition to encoding weaker variations on Faithfulness, we also encode the local Markov assumption in addition to the global Markov assumption. For the local Markov assumption, instead of taking a CD statement $X \not\perp\!\!\!\perp Y \mid \mathbf{C}$ to imply the constraint that X and Y are not d-separated by C (as stated by the global Markov assumption), we take it to imply the constraint that it is not the case that X's parent set is $\mathbf{C}$ and Y is not a descendant of X, or that Y's parent set is $\mathbf{C}$ and X is not a descendant of Y. Although these constraints are equivalent given a perfect oracle of CI statements, the local version is in general weaker given an imperfect one.

Figure 1 summarizes the ASP-encoding of these various assumptions, where, to improve readability, we use the following predicates that are defined in terms of more basic predicates used in Hyttinen et al's encoding.

- $indep(X, Y, \mathbf{C}, w)$: X and Y are independent conditional on C, given as input fact, with weight w.
- $dep(X, Y, \mathbf{C}, w)$: X and Y are independent conditional on C, given as input fact, with weight w.
- $dsep(X, Y, \mathbf{C})$: X and Y are d-separated given $\mathbf{C}$.
- $pa(X, \mathbf{C})$: $\mathbf{C}$ is X's (exact) parent set.
- $ismember(\mathbf{C}, X)$: X is a member of $\mathbf{C}$.
- $desc(X, Y)$: Y is X's descendant.

The constraints are coded in ASP as violation conditions. The ones that begin with :- are *hard* violations that are not allowed in the output, whereas the ones that use the predicate $fail$ are treated as *weak* violations, which are allowed but penalized according to the given weights. In Figure 1, only NoE-minimality and NoI-minimality are encoded with this device (and minimization of the total weight of the violations). In general, any constraint can be treated as a weak constraint with a certain weight for the sake of conflict resolution [Hyttinen et al., 2014].

The encoding in Figure 1 delivers a multitude of SAT-based algorithms that are asymptotically correct under different assumptions. For example, combining the encoding of (global or local) Markov and that of Adjacency-faithfulness in Hyttinen et al.'s general setup

(Variables are arbitrarily ordered so that $indep(X, Y, \mathbf{C}, w)$ and $dep(X, Y, \mathbf{C}, w)$ are considered only if $Y > X$.)

Faithfulness (violations):
$\forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
:- not $dsep(X, Y, \mathbf{C}), indep(X, Y, \mathbf{C}, w)$

Adjacency-faithfulness (violations):
$\forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
:- $edge(X, Y), indep(X, Y, \mathbf{C}, w)$.
:- $edge(Y, X), indep(X, Y, \mathbf{C}, w)$.

Triangle-faithfulness (violations):
$\forall X \forall Y > X, \forall Z \in \mathbf{V} \setminus \{X, Y\}, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
:- $edge(X, Y), edge(X, Z), edge(Y, Z), ismember(\mathbf{C}, Z), indep(X, Y, \mathbf{C}, w)$.
:- $edge(Y, X), edge(X, Z), edge(Y, Z), ismember(\mathbf{C}, Z), indep(X, Y, \mathbf{C}, w)$.
:- $edge(X, Y), edge(X, Z), edge(Z, Y),$ not $ismember(\mathbf{C}, Z), indep(X, Y, \mathbf{C}, w)$.
:- $edge(X, Y), edge(Z, X), edge(Z, Y),$ not $ismember(\mathbf{C}, Z), indep(X, Y, \mathbf{C}, w)$.
:- $edge(Y, X), edge(Z, X), edge(Y, Z),$ not $ismember(\mathbf{C}, Z), indep(X, Y, \mathbf{C}, w)$.
:- $edge(Y, X), edge(Z, X), edge(Z, Y),$ not $ismember(\mathbf{C}, Z), indep(X, Y, \mathbf{C}, w)$.

SGS-minimality (violations):
$\forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
:- $pa(X, \{Y\} \cup \mathbf{C}), indep(X, Y, \mathbf{C}, w)$.
:- $pa(Y, \{X\} \cup \mathbf{C}), indep(X, Y, \mathbf{C}, w)$.

NoE-minimality (optimization of weak constraints):
$\forall X \forall Y > X$,
$fail(X, Y, \mathbf{C}, w = 1)$:- $edge(X, Y)$.
$fail(X, Y, \mathbf{C}, w = 1)$:- $edge(Y, X)$.

$:\sim fail(X, Y, \mathbf{C}, w).\ [w]$

NoI-minimality (optimization of weak constraints):
$\forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
$fail(X, Y, \mathbf{C}, w = 1)$:- not $dsep(X, Y, \mathbf{C}), indep(X, Y, \mathbf{C}, w)$.

$:\sim fail(X, Y, \mathbf{C}, w).\ [w]$

Global Markov (violations):
$\forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
:- $dsep(X, Y, \mathbf{C}), dep(X, Y, \mathbf{C}, w)$.

Local Markov (violations):
$\forall X \forall Y > X, \forall \mathbf{C} \subseteq \mathbf{V} \setminus \{X, Y\}$,
:- $pa(X, \mathbf{C})$,not $desc(X, Y), dep(X, Y, \mathbf{C}, w)$.
:- $pa(Y, \mathbf{C})$, not $desc(Y, X), dep(X, Y, \mathbf{C}, w)$.

Figure 1: ASP Encoding of Various Assumptions

yields a SAT-based algorithm that is correct under the Markov and Adjacency-faithfulness assumptions (together with the assumptions of causal sufficiency and no feedback). Similarly, combining the encoding of Markov and those of SGS-minimality and Triangle-faithfulness yields a SAT-based algorithm that is correct under the Markov, SGS-minimality, and Triangle-faithfulness assumptions. Moreover, it is easy to see that these algorithms are "query-complete" in Hyttinen et al. [2013]'s sense, for the output implicitly contains all DAGs that are compatible with the inputted CI/CD statements given the corresponding assumptions. Consequently, any query about whether certain edge configurations are shared by all those DAGs can be easily computed. By contrast, although the Conservative PC algorithm [Ramsey et al., 2006] is asymptotically correct under the Markov and Adjacency-faithfulness assumptions, it is unknown whether its output entails answers to all such queries. For Triangle-faithfulness (plus Markov and SGS-minimality), Spirtes and Zhang [2014] proposed a Very Conservative SGS algorithm that is asymptotically correct, but it is clear that neither that algorithm nor the variations investigated by Havrilla [2015] are complete in this sense.[5]

So we have the following algorithms:

- **Adj**: Adjacency-faithfulness + Global Markov
- **Tri**: Triangle-faithfulness + SGS-minimality + Global Markov
- **NoE**: NoE-minimality + Global Markov
- **NoI**: NoI-minimality + Global Markov (which is essentially identifical to one of Hyttinen et al.'s conflict resolution algorithms[6], and will also be called **Faith + HW**.)

In each of them, Global Markov can be replaced by Local Markov and we get **AdjLM**, **TriLM**, **NoELM**, and **NoILM**, respectively.

5 SIMULATIONS

In this section, we report some findings from two types of simulations, following the setup of Hyttinen et al. [2013]

[5]For example, in the large sample limit these algorithms will never output any adjacency that violates Adjacency-faithfulness — though if there is any, they will return "don't know" and so do not err — but there are cases in which an edge that violates Adjacency-faithfulness is identifiable under Markov, SGS-minimality, and Triangle-faithfulness, which will be picked up by our algorithm.

[6]Except that NoI outputs all DAGs with the same, optimal weight, whereas Hyttinen et al.'s original, "hard dependencies" algorithm outputs one of them.

and that of Hyttinen et al. [2014], respectively. The former takes a perfect oracle of CI and CD statements as input, whereas the latter uses conditional independence tests on (Gaussian) data of a moderate sample size in place of an oracle.[7] In the experiments reported below, all conditional independence/dependence constraints are taken into account, and the output of an algorithm is the set of *all* DAGs that satisfy all the relevant constraints (with no conflict resolution) or optimize the constraint satisfaction (with conflict resolution).

5.1 SIMULATIONS WITH PERFECT ORACLES

We randomly generate 100 DAGs over 6, 8, and 10 variables, respectively; for each DAG, the average degree of a vertex is set to be 2. We use (the d-separation features of) each DAG as an oracle for CI and CD statements. That is, the input oracle satisfies both Markov and Faithfulness assumptions. We run this simulation (1) to demonstrate the claim that when the input happens to be consistent with Faithfulness (and Markov), all those algorithms based on weaker assumptions will return the exact same result as the algorithm based on Markov and Faithfulness (that is, the extent of underdetermination does not increase even though the Faithfulness assumption is weakened), and (2) to check the effect of adopting weaker assumptions on solving time when there are no conflicts, which provides an interesting contrast to the situation with conflicts.

As expected, in every case all four algorithms — $\mathbf{Adj}$, $\mathbf{Tri}$, $\mathbf{NoE}$, and $\mathbf{NoI}$ — return the exact same result (i.e., the Markov equivalence class of the true DAG) as the algorithm $\mathbf{Faith}$, which combines global Markov and Faithfulness. A comparison of the median solving times is given in Table 1.

Table 1: Median Solving Times (in seconds) given Perfect Oracles

$\lvert\mathbf{V}\rvert$	**Faith**	**Adj**	**Tri**	**NoI**	**NoE**
6	0.14	0.15	0.22	0.16	0.19
8	1.89	1.91	3.03	4.12	17.22
10	27.97	28.03	51.61	75.04	234.47

Apparently, when there is no conflict, solving times increase as the assumption is weakened. This seems to echo Zhang and Spirtes [2016]'s observation that the Faithfulness assumption may boost computational efficiency over its weaker variants. However, the situation changes dramatically when conflicts are present.

5.2 SIMULATIONS WITH FINITE SAMPLES

Following Hyttinen et al. [2014], we randomly generate 200 DAGs over 6 variables (with expected degree of 2 for each vertex) and parameterize each as a linear Gaussian model, in which every edge coefficient is uniformly drawn from $[-0.8, -0.2] \cup [0.2, 0.8]$, and the variance of each error term is $|1 + 0.1Z|$ where Z is drawn from a standard Gaussian distribution. From each model, we draw 20 i.i.d. samples of size 500, and use the CI tests employed in Hyttinen et al. [2014] to obtain, for each sample, an input of (weighted) CI and CD "facts".

For each algorithm that uses classical CI tests, we try 3 threshold values for rejecting the null hypothesis (0.005, 0.01 and 0.05). In these $200 \times 20 \times 3$ runs, $\mathbf{Faith}$ produces results (i.e., returns a non-empty set of DAGs) in about 5% (604/12000) of the cases, whereas $\mathbf{Adj}$ produces results in 44% (5306/12000) of the cases and $\mathbf{Tri}$ produces results in 66% (7917/12000) of the cases. All remaining cases are unsatisfiable under the given assumptions. So weaker assumptions significantly reduce the need of conflict resolution.

To measure the accuracy of inferred adjacencies, we calculate the true positive rate (TPR) and false positive rate (FPR) of *definite* adjacencies in the output, where a definite adjacency is an adjacency that is shared by all DAGs in the output; to measure the accuracy of inferred edge directions, we compare the definite arrows in the output to the definite arrows in the true Markov equivalence class, and calculate arrow precision (AP) and arrow recall (AR) accordingly.

Figure 2 plots the ROC curves for inferred adjacencies (with 3 points on each curve corresponding to the 3 test thresholds), counting only those cases where $\mathbf{Adj}$ produces results. Figure 3 plots the precision-recall curves for inferred arrows.[8] In addition to the four aforementioned algorithms, Hyttinen et al.'s conflict resolution algorithms $\mathbf{Faith + CW}$ (constant weights) and $\mathbf{Faith + LW}$ (log weights) are included.[9] For these "no-conflict-for-$\mathbf{Adj}$" cases, $\mathbf{Adj}$'s (or $\mathbf{Tri}$'s) performance is quite comparable even to that of the pseudo-Bayesian algorithm $\mathbf{Faith + LW}$. For adjacencies, its

[7] All experiments are done on a virtual machine running RedHat Enterprise Linux 6, with 12 virtual CPUs (Intel(R) Xeon(R) E5-2690 v4 CPU@2.60GHz). We use Clingo 4.5.4 as the ASP solver [Gebser et al., 2011].

[8] It is not hard to prove that when $\mathbf{Adj}$ returns a non-empty set of DAGs, the output of $\mathbf{NoE}$ must be identical to that of $\mathbf{Adj}$. So in Figures 2 and 3, $\mathbf{NoE}$ and $\mathbf{Adj}$ share the same curve. In addition, in Figure 2, $\mathbf{Tri}$ also shares the same curve, for it returns the same definite adjacencies as $\mathbf{Adj}$ does.

[9] $\mathbf{Faith + LW}$ uses Bayesian tests and we tried 3 values for the prior: 0.05, 0.1, and 0.2.

accuracy seems to be among the best; for orientations, its recall is lower than **Faith** + **LW** but precision is higher. Since it does not involve constraint optimization, **Adj** also runs significantly faster.

In order to resolve conflicts in the remaining cases, we try **Adj** + **CW** and **Adj** + **HW** ("hard dependencies" weights).[10] Figure 4 and Figure 5 show the comparative performance of a number of algorithms. **Adj** + **HW** and **Faith** + **HW** (i.e., **NoI**) have very similar performances, with slightly different balances between TPR and FPR, and between AP and AR. **Adj** + **CW** and **Faith** + **CW** are also fairly comparable: **Adj** + **CW** seems to output more accurate adjacencies than **Faith** + **CW**, but **Faith** + **CW** is more accurate on arrows when test threshold is 0.05. Overall, **Faith** + **LW** seems to have the best performance, and **NoELM** acheives the best arrow precision at the cost of arrow recall (in general, using LM instead of global Markov tends to increase precision at a significant cost of decreasing recall.)

Remarkably, however, **Adj** + **CW**/**HW** turns out to require much shorter solving times than **Faith** + **CW**/**HW**/**LW**. The differences become dramatic with only 8 variables. We generate 100 linear Gaussian models on 8 variables and from each model draw a sample of size 500. Figure 6 shows the sorted solving times for these 100 cases. The time saving of **Adj** + **CW**/**HW** is huge. It is also interesting to note that **Tri** + **CW** is substantially faster than **Faith** + **CW**, and **Tri** + **HW** than **Faith** + **HW**, even though when there is no conflict, **Tri** is significantly slower than **Faith**.

In addition, we generated 100 datasets on 10 variables, with a 1-hour time-out for each dataset. It turns out that **Faith** + **CW**/**HW**/**LW** fails to finish any case within the time limit, whereas **Adj** + **HW** finishes $95/100$ and **Adj** + **CW** finishes $63/100$.

6 CONCLUSION

We have shown how to encode a variety of weakenings of the Faithfulness assumption on top of the framework presented in Hyttinen et al. [2014]. The encoding results in a number of variations on their algorithm that are asymptotically correct and query-complete under assumptions that are weaker than Faithfulness. For some of the weaker assumptions, such as Triangle-faithfulness plus SGS-minimality, no other algorithm that is asymptotically correct and query-complete under them is currently known.

[10] Somehow **Adj** + **LW** does not work nearly as well as **Faith** + **LW**, and we are still investigating the reason.

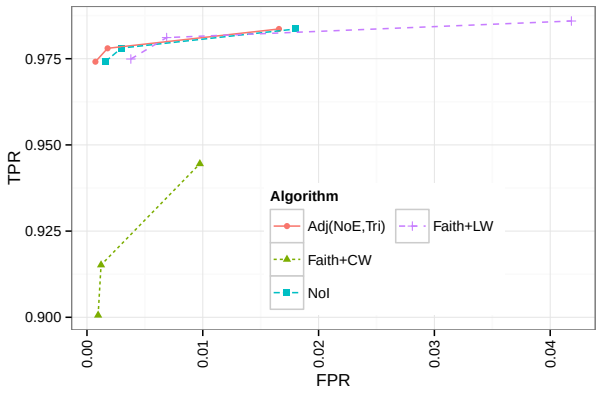

Figure 2: ROC of Adjacencies Among the Cases that **Adj** Produces Results

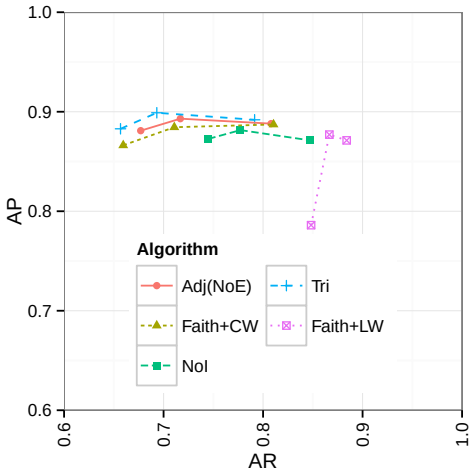

Figure 3: Arrow Precision-Recall Among the Cases that **Adj** Produces Results

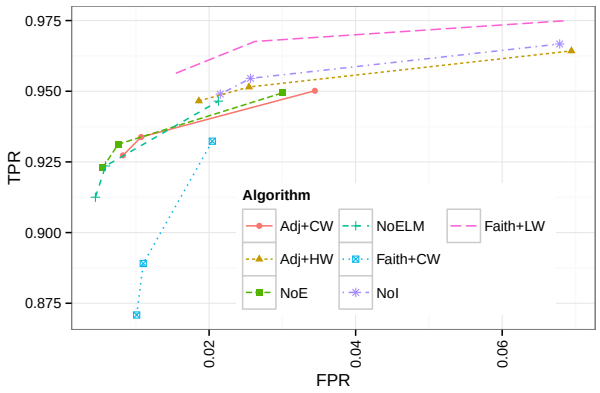

Figure 4: ROC of Adjacencies Among All Cases

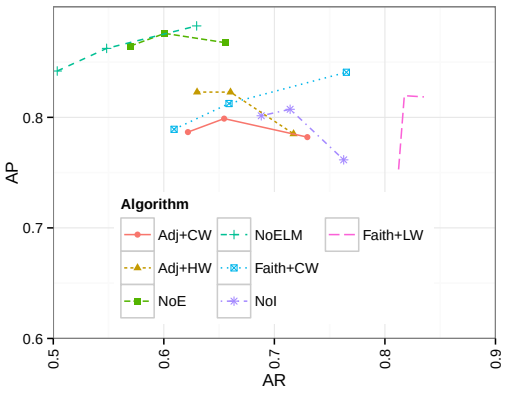

Figure 5: Arrow Precision-Recall Among All Cases

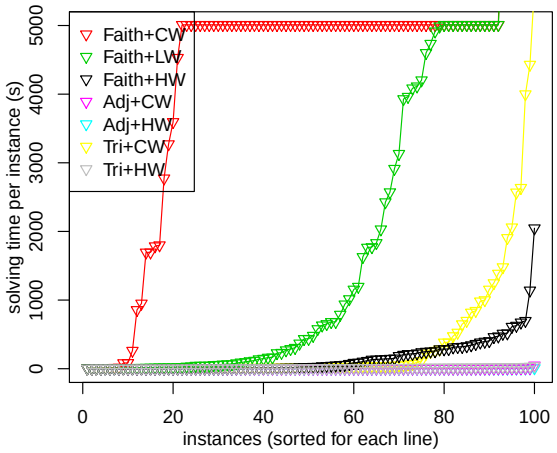

Figure 6: Sorted Solving Times for 8 Variables

These weakenings of Faithfulness are theoretically interesting not only because Faithfulness is a controversial assumption, but more importantly, also because the weaker assumptions are in a sense as inferentially powerful as Faithfulness. For any distribution that is consistent with the Markov and Faithfulness assumptions, adopting any of the weaker assumptions considered in this paper as opposed to Faithfulness does not increase the extent to which the distribution underdetermines the causal structure. This important fact is empirically illustrated by our simulations that take perfect, faithful oracles as inputs.

In this connection, an open problem is how to encode the P-minimality assumption briefly mentioned in Section 4, which is theoretically important for at least two reasons. First, it is even weaker than any of the four weakenings of Faithfulness we encoded. Second, it remains true that when Faithfulness happens to be true, adopting P-minimality does not increase the extent of underdetermination [Zhang, 2013].

The practical value of SAT-based causal discovery algorithms is currently limited by the fact that they do not yet scale well. Our algorithms are no exception. However, the finite-sample simulation results suggest that using a weaker faithfulness such as Adjacency-faithfulness not only reduces the need of conflict resolution, but also leads to substantial savings of solving time when combined with a conflict resolution scheme, compared to using Faithfulness together with the same scheme. It is thus reasonable to expect that our work here will contribute to improving the scalability of SAT-based causal discovery.

On the other hand, we also observed that some conflict resolution strategies, especially the one using "log weights", are less accurate when combined with Adjacency-faithfulness than they are when combined with Faithfulness. It is thus worth exploring whether there are alternative conflict resolution strategies that work better with the weaker constraints.

We have focused in this paper on the task of inferring acyclic, causally sufficient structures, from a single, observational distribution or dataset, because the studies on weakening Faithfulness have been so focused. However, a distinctive power of the SAT-based approach is that it can handle a very general search space, including cyclic structures with latent confounders, as well as multiple, overlapping datasets obtained from observational and/or experimental regimes [Hyttinen et al., 2013]. If the causal structure over $\mathbf{V}$ is possibly cyclic and/or causally insufficient, then in general the causal structure can be represented by a mixed graph over $\mathbf{V}$ that (1) can contain two types of edges, directed and bi-directed ($\leftrightarrow$), where a bi-directed edge between X and Y means that they are confounded by a latent variable, (2) allows multiple edges between any two variables, and (3) allows directed cycles. The notion of d-separation is readily generalized to such a graph, and so are the global Markov[11] and Faithfulness assumptions. We intend to further investigate to what extent the various results on weakening Faithfulness can be generalized to this setting. It is at least clear that Adjacency-faithfulness remains a consequence of Faithfulness, and using the simple, weaker constraint imposed by Adjacency-faithfulness may also speed things up substantially in the general setting.

Acknowledgements

JZ's research was supported by the Research Grants Council of Hong Kong under the General Research Fund LU13600715 and by a Faculty Research Grant from Lingnan University. FE's research was supported by NSF grant #1564330.

[11]The local Markov assumption often fails even in linear cyclic models [Spirtes, 1995].

References

C. Baral. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press, 2010.

G. Borboudakis and I. Tsamardinos. Towards robust and versatile causal discovery for business applications. In *Proceedings of KDD*, pages 1435–1444, 2016.

N. Cartwright. What is wrong with Bayes nets? *The Monist*, pages 242–264, 2001.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

M. Forster, G. Raskutti, R. Stern, and N. Weinberger. The frugal inference of causal relations. *British Journal for the Philosophy of Science*, 2017. doi: /10.1093/bjps/axw033.

M. Gebser, B. Kaufmann, R. Kaminski, M. Ostrowski, T. Schaub, and M. T. Schneider. Potassco: The Potsdam Answer Set Solving collection. *AI Communications*, 24(2):107–124, 2011.

N. Havrilla. Exploring very conservative search algorithms. Master's thesis, Department of Philosophy, Carnegie Mellon University, 2015.

K. D. Hoover. *Causality in Macroeconomics*. Cambridge University Press, 2001.

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.

J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.

D. Sonntag, M. Järvisalo, J. M. Pena, and A. Hyttinen. Learning optimal chain graphs with answer set programming. In *Proceedings of UAI*, pages 822–831, 2015.

P. Spirtes. Directed cyclic graphical representations of feedback models. In *Proceedings of UAI*, pages 491–498, 1995.

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.

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

S. Triantafilou, I. Tsamardinos, and I. G. Tollis. Learning causal structure from overlapping variable sets. In *Proceedings of AISTATS*, pages 860–867, 2010.

C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of faithfulness assumption in causal inference. *The Annals of Statistics*, 41:436–463, 2013.

Zhalama, J. Zhang, and W. Mayer. Weakening faithfulness: Some heuristic causal discovery algorithms. *International Journal of Data Science and Analytics*, 3: 93–104, 2017.

J. Zhang. A comparison of three Occam's razor for Markovian causal models. *British Jounrnal for the Philosophy of Science*, 64(2):423–448, 2013.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18(2): 239–271, 2008.

J. Zhang and P. Spirtes. Intervention, determinism, and the causal minimality condition. *Synthese*, 182:335–347, 2011.

J. Zhang and P. Spirtes. The three faces of faithfulness. *Synthese*, 193(4):1011–1027, 2016.