An empirical study of one of the simplest causal prediction algorithms

Joris M. Mooij
Informatics Institute
University of Amsterdam
The Netherlands

Jerome Cremers
Graduate School of Informatics
University of Amsterdam
The Netherlands

Abstract

We study one of the simplest causal prediction algorithms that uses only conditional independences estimated from purely observational data. A specific pattern of four conditional independence relations amongst a quadruple of random variables already implies that one of these variables causes another one without any confounding. As a consequence, it is possible to predict what would happen under an intervention on that variable without actually performing the intervention. Although the method is asymptotically consistent and works well in settings with only few (latent) variables, we find that its prediction accuracy can be worse than simple (inconsistent) baselines when many (latent) variables are present. Our findings illustrate that violations of strong faithfulness become increasingly likely in the presence of many latent variables, and this can significantly deterioriate the accuracy of constraint-based causal prediction algorithms that assume faithfulness.

1 Introduction

One of the central tasks in causal inference is to predict the changes resulting from interventions [Pearl, 2000, Spirtes et al., 2000], where by intervention we mean a perturbation of a system by some external cause. An example of such a causal prediction task from biology is to predict the expression of some gene when another gene is knocked out (i.e., its expression is artificially reduced). This causal prediction task is more challenging than the "noncausal" prediction tasks mostly considered in statistics and machine learning (e.g., estimating the expression of some gene given a measurement of the expression of another gene, as in regression). Note that the crucial difference between the two (i.e., causal and noncausal) prediction tasks reflects the difference between (passive) observation and (active) intervention. Deriving theory and designing algorithms for causal prediction is one of the key challenges in the field of causal inference. A very challenging task in this context is to predict the effect of interventions from purely observational data (i.e., measurements from an unperturbed system), without any knowledge of the causal structure of the system. This is the setting that we study in this work.

When the causal structure of the system is known, causal predictions can be made for instance by parent adjustment, the back-door criterion, and the front-door criterion [Pearl, 2000]. More generally, Pearl's do-calculus [Pearl, 2000] can be employed, for example using the algorithm by Tian and Pearl [2002]. This allows one to identify all causal effects from purely observational data given the true causal structure [Shpitser and Pearl, 2006, Huang and Valtorta, 2006].

However, the causal structure of the system is often not known. Estimating the full causal structure from data is typically not possible without making strong assumptions. However, it is possible under weaker assumptions to estimate the *Markov equivalence class* from purely observational data, i.e., the set of all causal structures that are compatible with the observed conditional independences in the data. This can be done for instance using constraint-based causal discovery algorithms, like the PC algorithm [Spirtes et al., 2000] that assumes causal sufficiency, or the FCI algorithm [Spirtes et al., 2000] that allows for latent confounders and selection bias. These causal discovery algorithms output a compact representation of all the causal structures that are compatible with the data, under certain assumptions.

A brute-force approach to causal prediction would then consist of enumerating all possible causal graphs in the estimated Markov equivalence class, and estimating the causal effects for each of these graphs, thereby yielding a set of causal effects that are compatible with the data. Smarter approaches [Spirtes et al., 2000, Zhang, 2008, Maathuis et al., 2010, Maathuis and Colombo, 2015, Hyttinen et al., 2015] avoid this brute-force enumeration. What all these approaches have in common is that they separate the problem into two parts: first, estimate the set of all causal structures that are compatible with the data, then obtain causal predictions for all these (classes of) causal structures.

A bottleneck in those approaches is the estimation of the Markov equivalence class. This is a difficult statistical task, especially in high-dimensional settings. Constrained-based causal discovery algorithms typically perform a sequence of conditional independence tests, and which tests are performed depends on the results of previous tests. Therefore, statistical errors of conditional independence tests may propagate when estimating the Markov equivalence class, leading to wrong predictions, especially when a large number of these tests have to be performed. By first estimating the Markov equivalence class, we may be attacking a more difficult problem than necessary, and thereby introduce undesired variance into the causal effect estimates. Alternative approaches that do not require estimation of the Markov equivalence class have been proposed [VanderWeele and Shpitser, 2011, Entner et al., 2013], but these rely on partial background knowledge regarding causal relations.

In this work, we investigate a simple alternative method for predicting causal effects that is sound and consistent (also in the presence of confounders). The method effectively avoids estimating the (equivalence class of the) complete causal structure of all observed variables and focusses on small subsets of four variables instead. In this way, it *minimizes* the number of conditional independence tests necessary to reach a nontrivial causal prediction, thereby hopefully improving the accuracy of that particular prediction, as there is less possibility for statistical errors to accumulate. The main motivations behind our approach are (i) we would like to trade completeness for reliability, and (ii) focussing on a simple algorithm makes it easier to analyse its statistical properties.

We first sketch a general approach to causal reasoning, and then focus on a simple special case with four variables that leads to nontrivial conclusions. That special case is closely related to an existing method to detect so-called Y-structures [Mani et al., 2006]. Our main contributions are (i) an alternative derivation that offers straightforward ways to generalize and extend the method, and (ii) an empirical study of the performance of the algorithm and its building blocks. We conclude that the method, though simple and elegant, performs poorly on simulated data. In particular, violations of strong faithfulness become increasingly problematic as the number of (latent) variables increases, and deteriorate causal prediction accuracy so severely that the method does not even outperform simple noncausal baselines already for $p = 50$ variables in our simulations.

2 Theory

Given a set of random variables[1] $\boldsymbol{V}$, we can express their direct causal relationships by means of a causal graph, which has a directed edge $X \to Y$ if and only if $X \in \boldsymbol{V}$ is a direct cause of $Y \in \boldsymbol{V}$. A directed path (sequence of head-to-tail directed edges) corresponds to an indirect causal relationship, or *ancestral* relation. We denote the set of all indirect causes (ancestors) of a variable $X \in \boldsymbol{V}$ according to causal graph $\mathcal{G}$ by $\mathrm{An}_{\mathcal{G}}(X)$ (we adopt here the convention that this includes X itself). For a set of variables $\boldsymbol{X} \subseteq \boldsymbol{V}$, we define $\mathrm{An}_{\mathcal{G}}(\boldsymbol{X}) = \bigcup_{X \in \boldsymbol{X}} \mathrm{An}_{\mathcal{G}}(X)$. Therefore, $X \in \mathrm{An}_{\mathcal{G}}(\boldsymbol{Y})$ means that X is an (indirect) cause of some $Y \in \boldsymbol{Y}$ according to the causal graph $\mathcal{G}$, and $X \notin \mathrm{An}_{\mathcal{G}}(\boldsymbol{Y})$ means that X is not an (indirect) cause of any $Y \in \boldsymbol{Y}$ according to the causal graph $\mathcal{G}$. In addition to directed edges, the causal graph $\mathcal{G}$ may contain bidirected edges to denote confounders, i.e., latent common causes.

From now on, we assume that there is a causally sufficient set of variables $\boldsymbol{V} = \boldsymbol{O} \dot{\cup} \boldsymbol{L}$, of which we observe only the variables in $\boldsymbol{O}$, the variables in $\boldsymbol{L}$ being latent, and that the causal graph on $\boldsymbol{O} \cup \boldsymbol{L}$ is a directed acyclic graph (DAG). In particular, this means that we assume that there is no causal feedback and that there are no confounders of the variables $\boldsymbol{O} \cup \boldsymbol{L}$. Note that when considering only the observed variables $\boldsymbol{O}$, the latent variables in $\boldsymbol{L}$ may act as confounders for variables in $\boldsymbol{O}$, so we do *not* assume that the variables in $\boldsymbol{O}$ are causally sufficient on their own. Furthermore, we assume that there is no selection bias, i.e., we are not implicitly conditioning on (common effects of) the variables in $\boldsymbol{O} \cup \boldsymbol{L}$. Finally, an important assumption is faithfulness, i.e., each conditional independence $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}$ in the joint distribution of the random variables $\boldsymbol{O} \cup \boldsymbol{L}$ corresponds to a d-separation $X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \boldsymbol{Z}$ in the DAG $\mathcal{G}$. In other words, conditional independences in the distribution reflect properties of the causal *structure* rather than accidental cancellations due to very specific choices of the parameters of the causal model.

The approach we study here is a straightforward combination of two ingredients: causal discovery rules by Claassen and Heskes [2011] and a causal prediction rule by Entner et al. [2013]. We will begin by describing these causal reasoning rules.

2.1 Causal discovery rules

Claassen and Heskes [2011] pointed out a correspondence between what they call *minimal conditional (in)dependences* and ancestral relations. We adopt the same notation for these *minimal* conditional (in)dependences here. Claassen and Heskes [2011] define a *minimal con-*

[1] We denote sets of variables in boldface.

ditional independence by:

$$X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup [\boldsymbol{Z}] \iff \begin{cases} X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup \boldsymbol{Z}, \text{ and} \\ \forall \boldsymbol{Z'} \subsetneq \boldsymbol{Z} : X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup \boldsymbol{Z'} \end{cases}$$

Similarly, they define a *minimal conditional depedendence* by;

$$X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup [\boldsymbol{Z}] \iff \begin{cases} X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup \boldsymbol{Z}, \text{ and} \\ \forall \boldsymbol{Z'} \subsetneq \boldsymbol{Z} : X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup \boldsymbol{Z'} \end{cases}$$

The square brackets express that the variables in $\boldsymbol{Z}$ are necessary to obtain the (in)dependence, in the context of $\boldsymbol{W}$. The minimal conditional (in)dependences relate directly to ancestral relations in the DAG $\mathcal{G}$, as shown by Claassen and Heskes [2011]. In particular, they give the following inference rules:

Lemma 1 *For disjoint sets* $\{X\}, \{Y\}, \{Z\}, \boldsymbol{W} \subseteq \boldsymbol{O}$*:*

1. $X \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup [Z] \implies Z \in \mathrm{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{W})$
2. $X \not\perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup [Z] \implies Z \notin \mathrm{An}_{\mathcal{G}}(\{X, Y\} \cup \boldsymbol{W})$.

In addition, the following obvious rules for ancestral relations in a DAG $\mathcal{G}$ hold:

Lemma 2 *For* $X, Y, Z \in \boldsymbol{O}$*:*

1. $X \in \mathrm{An}_{\mathcal{G}}(Y) \wedge Y \in \mathrm{An}_{\mathcal{G}}(Z) \implies X \in \mathrm{An}_{\mathcal{G}}(Z)$*;*
2. $X \in \mathrm{An}_{\mathcal{G}}(Y) \wedge Y \in \mathrm{An}_{\mathcal{G}}(X) \implies X = Y$.

These rules express the transitivity and acyclicity of indirect causal relations.

2.2 Causal prediction rule

Under the same assumptions that we made above, Entner et al. [2013] show that:

Lemma 3 *For disjoint sets* $\{X\}, \{Y\}, \{Z\}, \boldsymbol{W}$*: if*

$$\begin{cases} Y \notin \mathrm{An}_{\mathcal{G}}(\{X\} \cup \boldsymbol{W} \cup \{Z\}) \\ X \notin \mathrm{An}_{\mathcal{G}}(\boldsymbol{W} \cup \{Z\}) \\ Z \perp\!\!\!\perp Y \,|\, \boldsymbol{W} \cup [X] \end{cases}$$

then $\boldsymbol{W}$ *is sufficient for adjustment of* X *on* Y*, i.e.,*

$$p(Y \,|\, \mathrm{do}(X = x)) = \int p(Y | X = x, \boldsymbol{W}) p(\boldsymbol{W}) \, d\boldsymbol{W}.$$

Here, $p(Y \,|\, \mathrm{do}(X = x))$ denotes the interventional distribution of Y under a perfect intervention on X that sets X to the value x [Pearl, 2000]. The proof uses the backdoor criterion [Pearl, 2000]. Entner et al. [2013] also provide rules for inferring no causal effect (i.e., $p(Y \mid \mathrm{do}(X = x)) = p(Y)$), but we do not reproduce those here as we are mostly interested in predicting nontrivial causal effects.

2.3 (Extended) Y-structures

The causal discovery rules by Claassen and Heskes [2011] allow to derive ancestral relations from conditional independence relations, and the causal prediction rule by Entner et al. [2013] allows to infer a sufficient adjustment set from a particular combination of ancestral and conditional independence relations. By combining these rules, sufficient adjustment sets can be found from conditional independence relations alone. In this way, we can easily arrive at causal predictions from purely observational data that even hold in the presence of confounders.

In our context, one of the simplest combinations of conditional independences that yields nontrivial causal predictions on four variables is the following:

Proposition 1 *For a quadruple* $\langle X, Y, Z, U \rangle \in \boldsymbol{O}^4$ *of different observed variables, if*

$$\begin{cases} Z \perp\!\!\!\perp Y \,|\, [X] \\ Z \not\perp\!\!\!\perp U \,|\, [X] \end{cases} \tag{1}$$

then $X \in \mathrm{An}_{\mathcal{G}}(Y)$ *and* $p(Y \mid \mathrm{do}(X)) = p(Y \,|\, X)$.

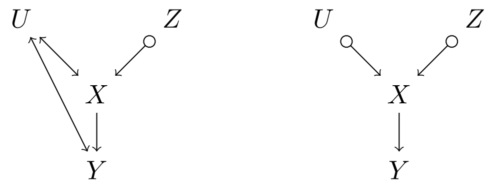
Proof. From $Z \not\perp\!\!\!\perp U \,|\, [X]$ and Lemma 1.2 it follows that $X \notin \mathrm{An}_{\mathcal{G}}(\{Z, U\})$, and therefore $X \notin \mathrm{An}_{\mathcal{G}}(Z)$. From $Z \perp\!\!\!\perp Y \,|\, [X]$ and Lemma 1.1 it follows that $X \in \mathrm{An}_{\mathcal{G}}(\{Z, Y\})$. Combining these two results, we conclude that $X \in \mathrm{An}_{\mathcal{G}}(Y)$. By acyclicity, this implies $Y \notin \mathrm{An}_{\mathcal{G}}(X)$. Further, $Y \in \mathrm{An}_{\mathcal{G}}(Z)$ would lead to $X \in \mathrm{An}_{\mathcal{G}}(Z)$ by transitivity, which contradicts $X \notin \mathrm{An}_{\mathcal{G}}(Z)$. Applying Lemma 3 with $\boldsymbol{W} = \emptyset$ immediately gives that $p(Y \mid \mathrm{do}(X)) = p(Y \,|\, X)$. □

In this simple context where $\boldsymbol{W} = \emptyset$, the causal prediction rule from Entner et al. [2013] reduces to a special case that was already known for a long time under the name *Local Causal Discovery* (LCD) [Cooper, 1997]. Therefore, we can also interpret Proposition 1 as a special case of LCD where the necessary ancestral preconditions are provided by employing the rules of [Claassen and Heskes, 2011].

The Markov equivalence class of $\mathcal{G}$ can be represented by a Partial Ancestral Graph (PAG) [Zhang, 2008] on the observed variables $\boldsymbol{O}$. Each PAG represents a collection of Maximal Ancestral Graphs (MAGs) [Richardson and Spirtes, 2002], and each MAG represents infinitely many DAGs. Each DAG (on some set of variables that contains all observed variables $\boldsymbol{O}$, and possibly more variables) represented by a PAG on $\boldsymbol{O}$ satisfies the same conditional independence relations on the observed variables $\boldsymbol{O}$.

Proposition 2 *There are two PAGs on* $\{X, Y, Z, U\}$ *that satisfy the relations in (1). They are depicted in Figure 1.*

Proof. Z and Y are not adjacent because $Z \perp\!\!\!\perp Y \,|\, X$. Z and U are not adjacent because $Z \perp\!\!\!\perp U$. We distinguish



"Extended Y-structure" "Y-structure"

Figure 1: All PAGs compatible with (1). Circles represent edge marks that can be either a tail or an arrowhead. Therefore, these two PAGs represent six MAGs.

two cases: U and Y are nonadjacent ("Y-structure") and U and Y are adjacent ("Extended Y-structure"). In both cases, three arrowheads follow from the ancestral relations $Y \notin \mathrm{An}_{\mathcal{G}}(X)$, $X \notin \mathrm{An}_{\mathcal{G}}(U)$, $X \notin \mathrm{An}_{\mathcal{G}}(Z)$, and one tail follows from $X \in \mathrm{An}_{\mathcal{G}}(Y)$, Note that if there is an edge between U and Y, then U must be a collider. Indeed, the path $Z \cdots X \cdots U \cdots Y$ must be blocked when conditioning on X. But then the edge between U and Y must have an arrowhead at Y, otherwise X would be ancestor of U. It is easy to check that each of the six MAGs corresponding with the two PAGs is compatible with the constraints (1). □

We can obtain symmetry between U and Z by adding another minimal conditional independence test (only satisfied by the Y-structures):

$$\begin{cases} Z \perp\!\!\!\perp Y \mid [X] \\ U \perp\!\!\!\perp Y \mid [X] \\ Z \not\perp\!\!\!\perp U \mid [X] \end{cases} \tag{2}$$

As we assume faithfulness, all other conditional independence relations on $\{X, Y, Z, U\}$ can now be read off from the PAGs.

Corollary 1 *Under faithfulness, the only conditional independences that hold in an Extended Y-structure are the two in (1), i.e., $Z \perp\!\!\!\perp Y \mid X$ and $Z \perp\!\!\!\perp U$. The only conditional independences that hold in a Y-structure are the three in (2), i.e., $Z \perp\!\!\!\perp Y \mid X$, $Z \perp\!\!\!\perp U$ and $U \perp\!\!\!\perp Y \mid X$, and in addition $U \perp\!\!\!\perp Y \mid \{X, Z\}$ and $Y \perp\!\!\!\perp Z \mid \{U, X\}$.*

Y-structures have been studied before by Mani et al. [2006], who showed that they can be identified by using a Bayesian scoring method (even in the presence of latent variables). [Mani and Cooper, 2004, Mani, 2006] also provide empirical results about the performance of Bayesian scoring methods for detecting Y-structures. To the best of our knowledge, Extended Y-structures have not been studied before.

Algorithm 1 Extended Y-structure search

```
Input:
  $\boldsymbol{O}$   set of observed variables
  $\mathcal{D}$   i.i.d. sample of $p(\boldsymbol{O})$
Output:
  $\mathcal{L}$   set of Extended Y-structures;
Algorithm:
  $\mathcal{L} \leftarrow \emptyset$
  for all $\langle X, Y, Z, U \rangle \in \boldsymbol{O}^4$ do
    if $\#\{X, Y, Z, U\} = 4$ then
      if $Z \not\perp\!\!\!\perp_{\mathcal{D}} Y$ and $Z \perp\!\!\!\perp_{\mathcal{D}} Y \mid X$ and $Z \perp\!\!\!\perp_{\mathcal{D}} U$ and
      $Z \not\perp\!\!\!\perp_{\mathcal{D}} U \mid X$ then
        $\mathcal{L} \leftarrow \mathcal{L} \cup \{\langle X, Y, Z, U \rangle\}$
      end if
    end if
  end for
Predictions:
$\forall \langle X, Y, Z, U \rangle \in \mathcal{L}: \quad p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$
```

2.4 Algorithms

One of the simplest algorithms that makes nontrivial causal predictions from purely observational data using the ideas above is given in Algorithm 1. It is a brute-force search over all quadruples in $\boldsymbol{O}$ that satisfy the Extended Y-structure conditional independences in (1). Any conditional independence test can be used when testing for conditional independences of the form $X \perp\!\!\!\perp_{\mathcal{D}} Y \mid \boldsymbol{Z}$ in the data $\mathcal{D}$. For each of the quadruples $\langle X, Y, Z, U \rangle$ output by the algorithm, the causal prediction is that $p(Y \mid \mathrm{do}(X = x)) = p(Y \mid X = x)$ for all x. In words: the interventional distribution of Y when setting X to the value x coincides with the conditional distribution of Y given $X = x$.

It follows directly from Proposition 1 that Algorithm 1 is sound. When using consistent conditional independence tests, it is also consistent: as the number of samples in $\mathcal{D}$ grows, the probability for an erroneous conclusion converges to 0. This directly follows from the consistency of the independence tests. However, the algorithm is not *uniformly* consistent. In practice, we do not know *a priori* how many samples we need to be confident about the correctness of the result [Robins et al., 2003]. Intuitively, as a dependence can be arbitrarily weak, we may need an arbitrarily high number of data points to be able to distinguish it from an independence. Furthermore, Cornia and Mooij [2014] showed that for LCD in a linear-Gaussian setting, it is impossible to derive a confidence interval on the causal prediction error without making strong assumptions. Their result also applies to Algorithm 1, as it makes a similar causal prediction as LCD does. Summarizing:

Proposition 3 *Algorithm 1 is sound and consistent when using consistent independence tests. However, it is not uniformly consistent and it is impossible to derive a confidence*

```
Algorithm 2 Conditional Independence Pattern search
Input:
  $\boldsymbol{O}$   set of observed variables
  $n$   pattern size
  $\pi$   pattern of conditional independences
  $\mathcal{D}$   i.i.d. sample of $p(\boldsymbol{O})$
Output:
  $\mathcal{L}$   set of $n$-tuples in $\boldsymbol{O}^n$ matching pattern $\pi$
Algorithm:
  $\mathcal{L} \leftarrow \emptyset$
  for all $T \in \boldsymbol{O}^n$ do
    if $\#T = n$ and $\pi(T)$ in $\mathcal{D}$ then
      $\mathcal{L} \leftarrow \mathcal{L} \cup \{T\}$
    end if
  end for
```

interval on the prediction error without making additional assumptions in the linear-Gaussian setting.

We have spelled out Algorithm 1 for clarity, even though it is a special case of the more general Algorithm 2 that performs a brute-force search for certain conditional independence patterns by testing whether all relations in the pattern simultaneously hold in the data. For example, using the following pattern for testing an Extended Y-structure in Algorithm 2 we recover Algorithm 1:

$$\texttt{extY}(\langle X,Y,Z,U\rangle) = Z \perp\!\!\!\perp Y \,|\, [X] \wedge Z \not\perp\!\!\!\perp U \,|\, [X].$$

In the next section, we will study also the following patterns on quadruples of variables:

$$\texttt{Y}(\langle X,Y,Z,U\rangle) = \texttt{extY}(\langle X,Y,Z,U\rangle) \wedge U \perp\!\!\!\perp Y \,|\, [X].$$

$$\begin{aligned}\texttt{Y1}(\langle X,Y,Z,U\rangle) = {} & \texttt{Y}(\langle X,Y,Z,U\rangle) \\ & \wedge Z \not\perp\!\!\!\perp X \wedge X \not\perp\!\!\!\perp Y \wedge X \not\perp\!\!\!\perp U \wedge Y \not\perp\!\!\!\perp U \\ & \wedge X \not\perp\!\!\!\perp U \,|\, Y \wedge X \not\perp\!\!\!\perp Z \,|\, Y \wedge U \not\perp\!\!\!\perp Z \,|\, Y \\ & \wedge X \not\perp\!\!\!\perp Y \,|\, U \wedge X \not\perp\!\!\!\perp Z \,|\, U \wedge Y \not\perp\!\!\!\perp Z \,|\, U \\ & \wedge X \not\perp\!\!\!\perp Y \,|\, Z \wedge X \not\perp\!\!\!\perp U \,|\, Z \wedge U \not\perp\!\!\!\perp Y \,|\, Z.\end{aligned}$$

$$\begin{aligned}\texttt{Y2}(\langle X,Y,Z,U\rangle) = {} & \texttt{Y1}(\langle X,Y,Z,U\rangle) \\ & \wedge U \not\perp\!\!\!\perp Z \,|\, \{X,Y\} \wedge U \not\perp\!\!\!\perp X \,|\, \{Z,Y\} \\ & \wedge Z \not\perp\!\!\!\perp X \,|\, \{U,Y\} \wedge X \not\perp\!\!\!\perp Y \,|\, \{U,Z\} \\ & \wedge U \perp\!\!\!\perp Y \,|\, \{X,Z\} \wedge Z \perp\!\!\!\perp Y \,|\, \{X,U\}.\end{aligned}$$

The patterns `Y`, `Y1` and `Y2` all test for a Y-structure. `Y` uses the minimal number of tests, `Y1` also tests for all (asymptotically redundant) tests up to conditioning set size 1, and `Y2` adds all (asymptotically redundant) tests up to conditioning set size 2.

3 Experiments

We performed simulation experiments to study the performance of Algorithms 1 and 2.

3.1 Simulations

For the simulations, we created random causal DAGs $\mathcal{G}$ with $p = |\boldsymbol{V}|$ variables.[2] For $i = 1, \dots, p$, we chose the parents $\mathrm{pa}(i) \subseteq \{1, \dots, i-1\}$ for variable X_i randomly (using 0,1,2,3 parents with probability $1/8, 1/2, 1/4, 1/8$, respectively). In this way, the random graph is guaranteed to be a DAG. After generating the random causal graph, we drew random weights $\tilde{B}_{ji} \sim \mathcal{N}(0,1)$ independently from a standard normal distribution for constructing linear structural equations

$$X_i = \sum_{j \in \mathrm{pa}(i)} \tilde{B}_{ji} X_j + \tilde{\epsilon}_i$$

with i.i.d. error terms $\tilde{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$ having a normal distribution with standard deviation $\sigma = 0.1$. After sampling all weights in this way, we applied rescaling transformations to all structural equations (of the form $(\tilde{B}_{ji}, \tilde{\epsilon}_i) \mapsto (B_{ji}, \epsilon_i) = (\alpha_i \tilde{B}_{ji}, \alpha_i \tilde{\epsilon}_i)$) sequentially for $i = 1, \dots, p$ such that $\mathbb{V}\mathrm{ar}(X_i) = 1$ for all $i = 1, \dots, p$. Without the rescaling, variances could easily diverge and $\mathbb{V}\mathrm{ar}(X_i)$ could depend strongly on i, thereby already revealing the causal order.

We sampled $N = 3000$ samples from $p(\boldsymbol{X})$ to simulate the observational data $\mathcal{D}$. We also simulated perfect interventions on different targets as follows. For each intervention, we chose its target i uniformly from $\{1, \dots, p\}$. Under the intervention $\mathrm{do}(X_i = \xi_i)$, the structural equation for X_i is changed into $X_i = \xi_i$, while the other structural equations and the distribution of the noise terms remain invariant under this intervention. We used a constant value $\xi_i = -2$ throughout. We then generated one sample from the intervened structural causal model. In this way, we generated 1000 interventional data points, each one corresponding to an intervention on a particular randomly chosen target variable. We used this interventional data to validate the causal predictions.

We considered four settings for the number of variables, $p = 10, p = 30, p = 50$ and $p = 70$.

3.2 Independence tests

Because we simulated linear-Gaussian data, for the (conditional) independence tests we simply calculate the (partial) correlations and their p-values by using a Student's t distribution for a transformation of the (partial) correlation. Small p-values indicate strong evidence against the null hypothesis of independence. On the other hand, for large p-values it is not clear whether there is a weak dependence or an independence. Nevertheless, following common practice in the field, we will use large p-values as evidence in

[2] In our simulations, we used $\boldsymbol{V} = \boldsymbol{O}$, i.e., all variables are observed.

favor of independence. We use two thresholds on the p-value p to distinguish three possible independence test results:

$$\begin{aligned} p < \alpha_{\mathrm{lo}} &\implies \text{dependence}, \\ \alpha_{\mathrm{lo}} \leq p \leq \alpha_{\mathrm{hi}} &\implies \text{unknown}, \\ p > \alpha_{\mathrm{hi}} &\implies \text{independence}. \end{aligned}$$

We used fixed values $\alpha_{\mathrm{lo}} = 10^{-4}$ and $\alpha_{\mathrm{hi}} = 10^{-1}$ throughout the experiments. When testing for combinations (conjunctions) of (in)dependences, we use a three-valued (false, unknown, true) logic when combining conditional independence test results with logical operators.

3.3 Discovering conditional independence patterns

We studied the performance of Algorithm (1) and Algorithm (2) with patterns `Y`, `Y1` and `Y2` on simulated data. In addition, we studied the performance of some of their building blocks: pairwise (in)dependence tests, conditional (in)dependence tests when conditioning on a single variable, and minimal conditional (in)dependence tests when conditioning on a single variable. The ground truth is provided by testing the patterns directly in the causal graph by using the Bayes Ball algorithm [Shachter, 1998] as an independence oracle.

We report precision and recall, defined as:

$$\text{precision} = \frac{TP}{TP + FP}, \quad \text{recall} = \frac{TP}{TP + FN + PU}$$

where TP are true positives, FP are false positives, FN are false negatives and PU are unknowns that are positives according to ground truth. Here, we are more interested in high precision than high recall, because being able to predict with high confidence a few strong causal effects would already be of great practical interest in applications.

The results are reported in Table 1 for $p = 10$ and $p = 50$ variables. First, note that the recall of the conditional and pairwise independence test is at $1 - \alpha_{hi}$ as it should be. Also, note that the precision of the conditional and pairwise dependence tests are very close to 1, reflecting that it is easy to recognize a strong (conditional) dependence as such. The elementary tests are not perfect, but precision and recall are within a reasonable range. However, when combining two elementary tests into a minimal test, precision may drop significantly. When combining two minimal tests into an extended Y-structure test, the precision drops even further. On the other hand, when adding another minimal conditional independence test to test for a Y-structure, precision increases. Adding more tests (patterns `Y1` and `Y2`) does not make much of a difference. To put everything into perspective, the precisions should be compared with the baseline of random guessing (indicated in angular brackets). The pattern search algorithm outperforms random guessing considerably, achieving precisions that are a few orders of magnitude higher.

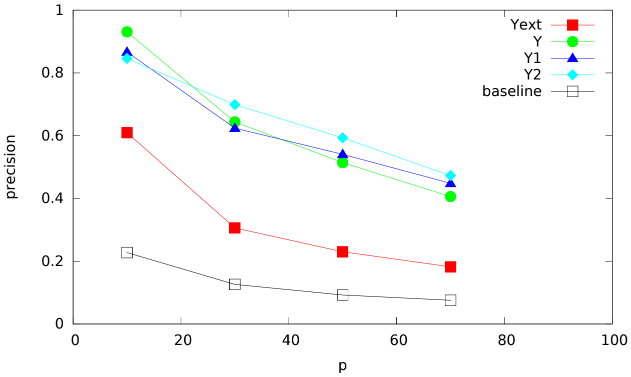


Figure 2: Precision of causal discovery $X \in \mathrm{An}(Y)$ on simulated data. The baseline is random guessing.

We conclude that for the task of detecting (Extended) Y-structures, the pattern search algorithm performs much better than random guessing. Furthermore, we observed that errors of elementary tests combine in unexpected ways into errors of compound tests. Sometimes the probability of an error of a compound test is much higher than the probability of error of its constituent tests, in other cases errors seem to cancel out and combining multiple tests results in "error correction". The reasons for this behavior of the precision are unclear. Recall has a more consistent behavior: the more tests are combined, the lower the recall.

3.4 Discovery of indirect causal relations

The evaluation measure used in the previous subsection is rather strict: the precision reflects how accurately a specific pattern can be detected from observational data. When we are only interested in using the (Extended) Y-structure patterns as a causal discovery method, i.e., as a way to detect whether $X \in \mathrm{An}_{\mathcal{G}}(Y)$ (X is an indirect cause of Y), the picture changes considerably. For this task, we define the "positives" to be the $\langle X, Y \rangle$ pairs that are contained in a quadruple $\langle X, Y, Z, U \rangle$ that satisfies the pattern of interest. The results are reported in Table 2 for $p = 10$ and $p = 50$. In both cases, the pattern search algorithms still outperform the baseline of random guessing, but not as much as for the task considered in the previous subsection. Precision again decreases as the number of variables increases. Figure 2 illustrates how the precisions depend on p, the number of (latent) variables.

We conclude that according to this performance measure, the simple causal discovery algorithm that searches for (Extended) Y-structures can outperform random guessing when used to find (indirect) causal relations. Detecting Y-structures works significantly better than detecting Extended Y-structures in this setting. However, precision decreases as the number of (latent) variables increases.

Table 1: Evaluation of Algorithm 2 for different patterns. Aggregates over 100 random models are shown. The second column gives the number of n-tuples of variables that are considered in the brute-force search, with n the number of variables that the pattern depends on. (a) $p = 10$ variables; (b) $p = 50$ variables. TP = true positives, FP = false positives, TN = true negatives, FN = false negatives, PU = unknowns that are positive according to ground truth, NU = unknowns that are negative according to ground truth. The baseline for precision is random guessing.

	Pattern	Total #	TP	FP	TN	FN	PU	NU	Recall	Precision (baseline)
	$X \perp\!\!\!\perp Y$	4500	886	47	3423	0	94	50	0.90	0.95 (0.22)
	$X \not\perp\!\!\!\perp Y$	4500	3423	0	886	47	50	94	0.97	1.00 (0.78)
	$X \perp\!\!\!\perp Y \mid Z$	36000	8917	1416	23476	0	936	1255	0.91	0.86 (0.27)
	$X \not\perp\!\!\!\perp Y \mid Z$	36000	23476	0	8917	1416	1255	936	0.90	1.00 (0.73)
(a)	$X \perp\!\!\!\perp Y \mid [Z]$	36000	2515	1037	30082	49	344	1973	0.86	0.71 (0.08)
	$X \not\perp\!\!\!\perp Y \mid [Z]$	36000	698	111	34112	53	144	882	0.78	0.86 (0.025)
	`extY`	504000	180	154	500716	11	99	2840	0.62	0.54 (0.0006)
	`Y`	504000	130	48	500797	14	146	2865	0.45	0.73 (0.0006)
	`Y1`	504000	46	28	500815	56	188	2867	0.16	0.62 (0.0006)
	`Y2`	504000	44	24	500815	56	190	2871	0.15	0.65 (0.0006)

	Pattern	Total #	TP	FP	TN	FN	PU	NU	Recall	Precision (baseline)
	$X \perp\!\!\!\perp Y$	122500	35576	3341	75778	1	3826	3978	0.90	0.91 (0.32)
	$X \not\perp\!\!\!\perp Y$	122500	75778	1	35576	3341	3978	3826	0.91	1.00 (0.68)
	$X \perp\!\!\!\perp Y \mid Z$	5880000	1705034	367019	3305326	91	190724	311806	0.90	0.82 (0.32)
	$X \not\perp\!\!\!\perp Y \mid Z$	5880000	3305326	91	1705034	367019	311806	190724	0.83	1.00 (0.68)
(b)	$X \perp\!\!\!\perp Y \mid [Z]$	5880000	108107	190968	5051117	7647	20179	501982	0.80	0.36 (0.023)
	$X \not\perp\!\!\!\perp Y \mid [Z]$	5880000	81052	33617	5383597	26773	23603	331358	0.62	0.71 (0.022)
	`extY`	552720000	54600	228284	546108050	22868	28476	6277722	0.52	0.19 (0.00019)
	`Y`	552720000	45320	51884	546247144	24017	35765	6315870	0.43	0.47 (0.00019)
	`Y1`	552720000	15376	20000	546261276	38173	51553	6333622	0.15	0.43 (0.00019)
	`Y2`	552720000	13538	14268	546262038	38209	53355	6338592	0.13	0.49 (0.00019)

Table 2: Evaluation of Algorithm 2 with different patterns for the task of predicting whether $X \in \mathrm{An}_{\mathcal{G}}(Y)$. Aggregates over 100 random models are shown. (a) $p = 10$ variables; (b) $p = 50$ variables. The baseline for precision is random guessing.

	Test pattern	Total #	TP	FP	TN	FN	Recall	Precision (baseline)
	`extY`	9000	50	32	6920	1998	0.024	0.61 (0.23)
(a)	`Y`	9000	27	2	6950	2021	0.013	0.93 (0.23)
	`Y1`	9000	13	2	6950	2035	0.0063	0.87 (0.23)
	`Y2`	9000	11	2	6950	2037	0.0054	0.85 (0.23)

	Test pattern	Total #	TP	FP	TN	FN	Recall	Precision (baseline)
	`extY`	245000	6627	22189	200191	15993	0.29	0.23 (0.09)
(b)	`Y`	245000	2486	2348	220032	20134	0.11	0.51 (0.09)
	`Y1`	245000	1155	981	221399	21465	0.051	0.54 (0.09)
	`Y2`	245000	1062	728	221652	21558	0.047	0.59 (0.09)

3.5 Causal predictions

The evaluation measure used in the previous subsection is a natural one when simulating data, but when using real data, it is often not known whether a variable is an indirect cause of another. Instead, interventional data may be available. In that context, we may be more interested in how accurately we predict the effects of interventions. When detecting an (Extended) Y-structure pattern for a quadruple $\langle X, Y, Z, U \rangle$, we can conclude that $p(Y \mid \mathrm{do}(X = x)) = p(Y \mid X = x)$. Using linear regression of Y on X we estimate $\mathbb{E}(Y \mid X = x)$ and use this as our prediction for the value of Y under the intervention $X = x$. In our setting, a natural measure for the causal prediction error of Y under an intervention $\mathrm{do}(X = x)$ is

$$\left|\mathbb{E}(Y \mid X = x) - \mathbb{E}(Y \mid \mathrm{do}(X = x))\right| .$$

We report the average error (ℓ_1 error) over all (X, Y) pairs in patterns found by the algorithm, all simulated interventions and all models. In addition, we report the corresponding root-mean-square error (ℓ_2 error).

For comparison, we also report results of two simple baselines. The first baseline predicts $p(Y \mid \mathrm{do}(X = x)) = p(Y)$ for all pairs $X \neq Y$ (i.e., complete absence of causal effects). The second baseline predicts $p(Y \mid \mathrm{do}(X = x)) = p(Y \mid X = x)$ for all pairs $X \neq Y$ (i.e., no difference between correlation and causation). Note that these baselines are provably inconsistent.

The results for these evaluation measures are reported in Table 3, for $p = 10$ and $p = 50$. Figure 3 shows how the ℓ_1 error depends on p, the number of (latent) variables. For $p = 10$ variables, most methods outperform the simple baselines. Unfortunately, that does not hold for $p = 50$ variables: in that case the simple baseline that always predicts that nothing will change due to an intervention outperforms all causal prediction methods. The reason is that even though the pattern search algorithm obtains a low error on the true positives (as expected), this is compensated by an error that is considerably higher than average on the false positives.

4 Conclusions and Discussion

We have studied a simple causal discovery and prediction method that focusses on quadruples of variables and only makes a prediction when it detects a certain pattern of conditional independences amongst those variables. The method is sound and consistent, and works well if the number of variables is not too large. However, like most constraint-based methods that rely only on conditional independences and the faithfulness assumption, it is not uniformly consistent [Robins et al., 2003]. Our empirical observations show that the accuracy of causal predictions deteriorates as more (latent) variables are present. This manifests itself quite clearly already in rather low-dimensional settings (50 variables, 3000 observations), where the simple causal prediction method does not even outperform naïve (inconsistent) baselines in terms of prediction error in our simulations.

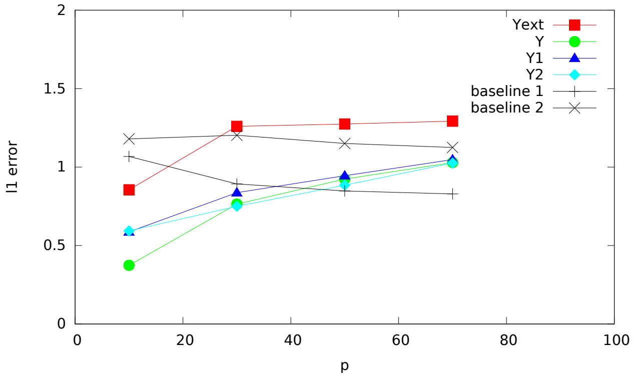


Figure 3: Causal prediction error ("ℓ_1 error") on simulated data. Baseline 1 always predicts $p(Y \mid \mathrm{do}(X)) = p(Y)$, baseline 2 always predicts $p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$.

Even though in our simulations the distribution on *all* variables $\boldsymbol{V} = \boldsymbol{O} \cup \boldsymbol{L}$ is faithful to the DAG, when only looking at a small subset of variables $\boldsymbol{Q} = \{X, Y, Z, U\}$, the marginal distribution on $\boldsymbol{Q}$ can become close-to-unfaithful to its PAG on $\boldsymbol{Q}$. One explanation for this might be that the more (latent) paths between the variables in $\boldsymbol{Q}$ there are, the higher the probability that these paths will cancel each other in some way when the edge weights are chosen randomly. This may then lead to near-faithfulness violations on $\boldsymbol{Q}$, and hence to to false-positive detections of the (Extended) Y-structure algorithms. However, the problem is not necessarily related to *cancellation* of paths, as we observe qualitatively similar behavior when we restrict all the edge weights to be positive (not reported here).

Note that the observed bad prediction performance occurs even though individual tests have relatively low probability of making an error in our simulation setting, and we only combine a few of these individual tests. In other words, there are strong dependences between test results that cannot be ignored. Adding more tests and thereby restricting the pattern to Y-structures helped to improve performance. On the other hand, adding more "redundant" tests did not significantly change accuracy. Thus, the original idea that *minimizing* the number of conditional independence tests would maximize accuracy turns out to be overly simplistic.

Problems with the faithfulness assumption have been pointed out before. Robins et al. [2003] showed that it is possible to create a sequence of faithful distributions that comes arbitrarily close to an unfaithful distribution. One possible way out would be to make a stronger assumption, like *strong faithfulness* [Zhang and Spirtes, 2003] for the linear-Gaussian case, requiring that nonzero partial corre-

Table 3: Evaluation of how well certain patterns found by Algorithm 2 predict the effect on Y of an intervention on X. Averages over 100 random models are shown. The errors are decomposed into two components: the error on the true positives (TP) and the error on the false positives (FP). Two simple noncausal baselines have been used for comparison. (a) $p = 10$ variables; (b) $p = 50$ variables.

		ℓ_1 error			ℓ_2 error		
	Method	all	only TP	only FP	all	only TP	only FP
(a)	extY	0.86	0.29	1.08	1.31	0.49	1.52
	Y	0.37	0.21	0.57	0.66	0.37	0.89
	Y1	0.59	0.32	0.74	0.88	0.47	1.04
	Y2	0.59	0.32	0.79	0.90	0.47	1.12
	$p(Y \mid \mathrm{do}(X)) = p(Y)$	1.07	-	-	1.31	-	-
	$p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$	1.18	-	-	1.74	-	-

		ℓ_1 error			ℓ_2 error		
	Method	all	only TP	only FP	all	only TP	only FP
(b)	extY	1.27	0.27	1.30	1.70	0.44	1.72
	Y	0.92	0.25	1.03	1.36	0.41	1.45
	Y1	0.94	0.33	1.05	1.36	0.48	1.46
	Y2	0.88	0.33	1.00	1.30	0.48	1.41
	$p(Y \mid \mathrm{do}(X)) = p(Y)$	0.85	-	-	1.09	-	-
	$p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$	1.15	-	-	1.59	-	-

lations are bounded away from zero, and in that way obtain uniform consistency. However, Uhler et al. [2013] show that the Lebesgue measure of distributions that do not satisfy strong faithfulness can be surprisingly large, and may grow quickly with the number of variables p. Their bounds are not directly applicable to our setting, as we are only interested in very specific conditional independence tests, and they only derived lower bounds for specific classes of DAGs that do not include the ones we used in our simulations. Nevertheless, using the techniques described in [Uhler et al., 2013], it may be possible to derive asymptotic results for the setting that we are interested in here. In our experiments we observed that by creating random linear-Gaussian causal models with a reasonably large number of variables, enough violations of strong faithfulness occur for prediction accuracy to suffer greatly.

We conclude that faithfulness violations can be very problematic for causal inference, even when individual independence tests have a low probability of error and we only combine a few of them to draw causal conclusions. The severity of this effect surprised us: one would probably need enormous amounts of observations for strong faithfulness to hold. Indeed, already for $p = 50$ variables, $N = 3000$ observations is not enough to outperform inconsistent noncausal baselines. In addition, we observed that prediction accuracy deteriorates as p becomes larger.

A related pattern search amongst quadruples of variables was proposed recently by Tsamardinos et al. [2012]. They search for a pattern amongst quadruples of variables that allows one to conclude that two of the four variables are dependent. To make this more interesting, they consider the situation that one has two datasets, each containing observations regarding only 3 out of 4 variables, and the two variables that are predicted to be dependent have not been simultaneously observed within a single dataset. Interestingly, that particular pattern search performs very well, also on high-dimensional real-world data, as reported by Tsamardinos et al. [2012]. This raises the question why certain patterns apparently lead to reliable predictions, whereas for other (superficially similar) patterns, the predictions turn out to be unreliable in high dimensions because of faithfulness violations. We leave this question for future research.

Acknowledgments

JM was supported by NWO (VIDI grant 639.072.410). We thank Oliver Stegle, Barbara Rakitsch, Tom Claassen and Ioannis Tsamardinos for helpful discussions. We thank Philip Versteeg for proofreading the manuscript.

References

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

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition, 2000.

J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, page 567573. AAAI Press, 2002.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226. AAAI Press, 2006.

Y. Huang and M. Valtorta. Identifiability in causal bayesian networks: a sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1149–1154. AAAI Press, 2006.

J. Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

M.H. Maathuis, D. Colombo, M. Kalisch, and P. Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7:247–248, 2010.

M. H. Maathuis and D. Colombo. A generalized backdoor criterion. *Annals of Statistics*, pages 1060–1088, 2015.

A. Hyttinen, F. Eberhardt, and M. Järvisalo. Do-calculus when the true graph is unknown. In *Proceedings of the 31th Conference on Uncertainty in Artificial Intelligence*, 2015.

T. J. VanderWeele and I. Shpitser. A new criterion for confounder selection. *Biometrics*, 67:1406–1413, 2011.

Doris Entner, Patrik O. Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2013)*, volume 31 of *Journal of Machine Learning Research Workshop and Conference Proceedings*, 2013.

Subramani Mani, Peter Spirtes, and Gregory F. Cooper. A theoretical study of Y structures for causal discovery. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 314–323. AUAI Press, 2006.

Tom Claassen and Tom Heskes. A logical characterization of constraint-based causal discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pages 135–144, 2011.

G. F. Cooper. A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery*, 1: 203–224, 1997.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Stat.*, 30(4):962–1030, 2002.

S Mani and GF Cooper. Causal discovery using a bayesian local causal discovery algorithm. *Studies in Health Technology and Informatics*, 107:731–735, 2004.

Subramani Mani. *A Bayesian Local Causal Discovery Framework*. PhD thesis, University of Pittsburg, March 2006. URL `http://d-scholarship.pitt.edu/10181/`.

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

Nicholas Cornia and Joris M. Mooij. Type-II errors of independence tests can lead to arbitrarily large errors in estimated causal effects: An illustrative example. In Joris M. Mooij, Dominik Janzing, Jonas Peters, Tom Claassen, and Antti Hyttinen, editors, *UAI 2014 Workshop Causal Inference: Learning and Prediction*, number 1274 in CEUR Workshop Proceedings, pages 35–42, Aachen, 2014. URL `http://ceur-ws.org/Vol-1274/uai2014ci_paper7.pdf`.

Ross D. Shachter. Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 480–487. AUAI Press, 1998.

Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 2013.

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

J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pages 632–639. Morgan Kaufmann, San Francisco, CA, 2003.

Ioannis Tsamardinos, Sofia Triantafillou, and Vincenzo Lagani. Towards integrative causal analysis of heterogeneous data sets and studies. *Journal of Machine Learning Research*, 13:1097–1157, April 2012.

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