An empirical study of the simplest causal prediction algorithm

Abstract

We study the simplest causal prediction algorithm that uses only conditional independences in purely observational data. A specific pattern of only four conditional independence relations amongst a quadruple of random variables already implies that one of these variables causes another 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 noncausal baselines when many (latent) variables are present. We also find that the accuracy can sometimes be improved by adding more conditional independence tests, but even then the performance need not outperform the baselines. More generally, our findings illustrate that high accuracy of individual conditional independence tests is no guarantee for high accuracy of a combination of such tests. Also, they illustrate the severity of the faithfulness assumption in practice.

1 Introduction

One of the central tasks in causal inference is to predict the changes resulting from interventions [Pearl, 2000, Spirtes et al., 2000]. An intervention is a perturbation of a system that forces it to behave differently than it would have in the absence of that intervention. An example of such a causal prediction task from biology is to predict the expression of some gene when another gene is knocked down or 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). 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 challenging task in this context is to predict the effect of interventions from purely observational data (i.e., measurements from an unperturbed system).

An interesting attempt at predicting the effects of gene knockouts from purely observational gene expression data was performed by Maathuis et al. [2010]. They analyzed micro-array data [Hughes et al., 2000] of *Saccharomyces cerevisiae*, a species of yeast. After preprocessing the data, the observational part of it consists of measurements of 5,361 gene expressions for 63 wild-type cultures, and the interventional part contains measurements of the expressions of the same 5,361 genes for 234 single-gene deletion ("knockout" or "knockdown") mutant strains that were grown under similar conditions as the wild-type cultures. The method of Maathuis et al. [2010] predicts gene pairs (A, B) for which knocking out gene A has a strong effect on the expression of gene B, using only the observational (wild-types) data. The interventional (mutants) part of the data was used for validation of the predictions.

The method employed by Maathuis et al. [2010] first uses the PC algorithm [Spirtes et al., 2000] to estimate the Markov equivalence class, i.e., the set of causal structures that are compatible with observed conditional independences in the data. From the estimated Markov equivalence class one can read off the possible parent sets of each gene, i.e., the sets of its possible direct causes. When the parent set of a gene is known, the strength of the causal effect on another gene can be estimated from observational data by parent adjustment [Pearl, 2000]. The novel idea of the method of Maathuis et al. [2010] is to calculate lower bounds on causal effect strengths by minimizing the effect strength over all possible parent sets according to the estimated Markov equivalence class in a computationally efficient way.

A bottleneck in this approach is the estimation of the Markov equivalence class, which is a difficult task in this high-dimensional setting. The PC algorithm performs 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. Indeed, the estimated Markov equivalence class turns out to be unstable in this high-dimensional setting [Colombo and Maathuis, 2014]. Another issue with the approach is that it makes the strong assumption of causal sufficiency of the 5,361 gene expression levels. In other words, it is assumed that there are no confounders, i.e., latent common causes of gene expressions that may lead to spurious dependences. It is very likely that this assumption is violated in practice.

In this work, we investigate an alternative method for predicting strong intervention effects that is sound and consistent even 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, the method *minimizes* the number of conditional independence tests necessary to reach a nontrivial causal prediction, thereby hopefully improving the accuracy of the predictions, as there is less possibility for statistical errors to accumulate.

We first sketch a general approach to causal reasoning, and then focus on the simplest 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 statistical behaviour of the method is unexpected and poorly understood, and that empirical violations of faithfulness can become increasingly problematic as the number of (latent) variables increases. Based on our simulation results, we expect that this simple method will probably not be successful when applied in high-dimensional settings like the challenging task of predicting strong effects of gene knockouts from the observational gene-expression data of Hughes et al. [2000].

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 with 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 DAG $\mathcal{G}$, 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 DAG $\mathcal{G}$. 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 with 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 conditional independence* by:

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

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

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

[1] We denote sets of variables in boldface.

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 strong 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, the simplest combination of conditional independences that yields nontrivial causal predictions involves four variables:

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 $p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$.

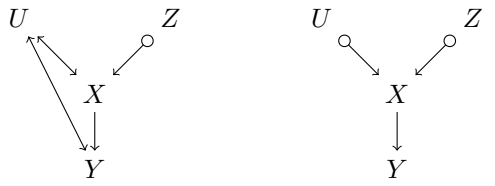
Proof. From $Z \not\perp\!\!\!\perp U \,|\, [X]$ and Lemma 1.2 it follows that $X \notin \mathrm{An}_{\mathcal{G}}(\{Z, U\})$, and therefore $X \notin \mathrm{An}_{\mathcal{G}}(Z)$. From $Z \perp\!\!\!\perp Y \,|\, [X]$ and Lemma 1.1 it follows that $X \in \mathrm{An}_{\mathcal{G}}(\{Z, Y\})$. Combining 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 $p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$. □

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] and was used by Chen et al. [2007] to infer causal relations between yeast genes from a combination of genotype and gene expression data. 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 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



"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.

$Y \notin \mathrm{An}_{\mathcal{G}}(X)$, $X \notin \mathrm{An}_{\mathcal{G}}(U)$, $X \notin \mathrm{An}_{\mathcal{G}}(Z)$, 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 *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.

2.4 Algorithms

The simplest algorithm 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

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:
$\forall \langle X, Y, Z, U \rangle \in \mathcal{L}: \quad p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$

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 [J. M. 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, 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 impossible to derive a confidence interval on the prediction error without making additional assumptions.*

We have spelled out Algorithm 1 for clarity, even though it is a special case of the more general Algorithm 2 that per-

Algorithm 2 Conditional Independence Pattern search

Input:
- $\boldsymbol{O}$ set of observed variables
- n pattern size
- π 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 π

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

forms 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:

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

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

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

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 $i = 1, \ldots, p$, we chose the parents $\mathrm{pa}(i) \subseteq \{1, \ldots, i-1\}$ 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 drawing a random causal graph, we draw random weights $\tilde{B}_{ji} \sim \mathcal{N}(0, 1)$ independently from a standard normal distribution for 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 $\sigma = 0.01$. 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 $i = 1, \ldots, p$ such that $\mathbb{V}\mathrm{ar}(X_i) = 1$ for all $i = 1, \ldots, p$. 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 $\{1, \ldots, p\}$. 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 two settings for the number of variables, $p = 10$ and $p = 50$. Considering the high signal-to-noise ratio and high number of observations, we would expect the algorithms to perform well in this setting.

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 use $\boldsymbol{V} = \boldsymbol{O}$, 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:

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

We used fixed values $\alpha_{\text{lo}} = 10^{-4}$ and $\alpha_{\text{hi}} = 10^{-1}$ 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:

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

where TP are true positives, FP are false positives, FN are false negatives and UP 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 intervention 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. The precision of the minimal conditional independence test drops to 60% for $p = 10$ variables, and all the way down to a meager 25% for $p = 50$ variables. When combining two minimal tests into an extended Y-structure test, the precision drops even further, unfortunately. On the other hand, when adding another minimal conditional independence test to test for a Y-structure, precision increases. However, when adding more tests (patterns `Y1` and `Y2`), precision decreases again. 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.

Table 1: Evaluation of Algorithm 2 for different patterns. Averages 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.

(a)

Pattern	Total #	Recall	Precision
$X \perp\!\!\!\perp Y$	45	0.89	0.98
$X \not\perp\!\!\!\perp Y$	45	0.98	1.00
$X \perp\!\!\!\perp Y \mid Z$	360	0.90	0.82
$X \not\perp\!\!\!\perp Y \mid Z$	360	0.88	1.00
$X \perp\!\!\!\perp Y \mid [Z]$	360	0.88	0.60
$X \not\perp\!\!\!\perp Y \mid [Z]$	360	0.86	0.89
`extY`	5040	0.71	0.45
`Y`	5040	0.59	0.53
`Y1`	5040	0.03	0.50
`Y2`	5040	0.01	0.33

(b)

Pattern	Total #	Recall	Precision
$X \perp\!\!\!\perp Y$	1225	0.90	0.96
$X \not\perp\!\!\!\perp Y$	1225	0.95	1.00
$X \perp\!\!\!\perp Y \mid Z$	58800	0.90	0.80
$X \not\perp\!\!\!\perp Y \mid Z$	58800	0.83	1.00
$X \perp\!\!\!\perp Y \mid [Z]$	58800	0.85	0.25
$X \not\perp\!\!\!\perp Y \mid [Z]$	58800	0.79	0.82
`extY`	5527200	0.72	0.24
`Y`	5527200	0.62	0.40
`Y1`	5527200	0.05	0.32
`Y2`	5527200	0.03	0.35

We conclude 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".

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 $X \in \text{An}_{\mathcal{G}}(Y)$ (X is an indirect cause of Y), the picture changes considerably. The results are reported in Table 2 for $p = 10$ and $p = 50$. For $p = 10$ variables, precision of Algorithm 1 for this causal discovery task is around 50%, and increases as more tests are added up to 100% for Algorithm 2 with the `Y1` and `Y2` patterns. Unfortunately, however, precision seems to decrease quickly

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

(a)

Test pattern	Total #	Recall	Precision
extY	90	0.0195	0.47
Y	90	0.0156	0.65
Y1	90	0.0020	1.00
Y2	90	0.0010	1.00

(b)

Test pattern	Total #	Recall	Precision
extY	2450	0.1890	0.22
Y	2450	0.0908	0.36
Y1	2450	0.0160	0.36
Y2	2450	0.0106	0.39

as the number of variables increases: for $p = 50$ all precisions are already lower than 40%.

We conclude that according to this performance measure, the simplest causal discovery algorithm that detects Extended Y-structures does not perform well. Detecting Y-structures works better, especially when redundant tests are added and when the total number of variables is relatively small. However, precision decreases quickly when the number of variables increases.

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 are 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 $X = x$. We define the causal prediction error of Y under an intervention $\mathrm{do}(X = x)$ as

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

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

For comparison, we also report results of two simple baselines. The first baseline always predicts $p(Y \mid \mathrm{do}(X = x)) = p(Y)$ for all pairs $X \neq Y$ (i.e., absence of causal effects). The second baseline predicts $p(Y \mid \mathrm{do}(X = x)) = p(Y \mid X = x)$ (i.e., not distinguishing correlation from causation) for all pairs $X \neq Y$. Note that these baselines are naïve and provably inconsistent.

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. Two simple noncausal baselines have been used for comparison. (a) $p = 10$ variables; (b) $p = 50$ variables.

(a)

Method	ℓ_1 error	ℓ_2 error
extY	0.85	1.45
Y	0.67	1.28
Y1	0.30	0.39
Y2	0.32	0.40
$p(Y \mid \mathrm{do}(X)) = p(Y)$	1.08	1.33
$p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$	1.72	4.99

(b)

Method	ℓ_1 error	ℓ_2 error
extY	1.23	1.77
Y	1.01	1.58
Y1	0.96	1.37
Y2	0.91	1.38
$p(Y \mid \mathrm{do}(X)) = p(Y)$	0.85	1.10
$p(Y \mid \mathrm{do}(X)) = p(Y \mid X)$	1.63	3.72

Table 3 contains the results, for $p = 10$ and $p = 50$. The error decreases as more tests are added to the (Extended) Y-structure pattern, and for $p = 10$ variables, most methods beat the simple baselines. Unfortunately, that does not hold for $p = 50$ variables, as in that case the simple baseline that always predicts that nothing will change due to an intervention outperforms all causal prediction methods.

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, but like all constraint-based methods that rely on conditional independences, is not uniformly consistent. This manifests itself quite clearly already in low-dimensional settings (50 variables, 3000 observations), where the causal prediction method cannot even outperform simple noncausal baselines.

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 MAG 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. Note that this surpris-

ing behaviour happens 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. Problems with the faithfulness assumption have been pointed out before [e.g., Lemeire and Janzing, 2013, Uhler et al., 2013]. We conclude that faithfulness violations are 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 faithfulness to hold empirically, already for $p = 50$ variables. In addition, we hypothesize that the larger p becomes, the higher the probability for accidental cancellations of paths. In other words, the probability for faithfulness violations seems to increase quickly with the number of variables. Therefore, this approach to causal discovery and prediction, simple and elegant as it is, will probably not work on the original task we had in mind, predicting strong intervention effects from purely observational micro-array data on the scale of the yeast genome ($p > 5000$, $N \sim 10^2$).

Acknowledgments

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.

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.

T.R. Hughes, M.J. Marton, A.R. Jones, C.J. Roberts, R. Stoughton, C.D. Armour, H.A. Bennett, E. Coffey, H. Dai, Y.D. He, M.J. Kidd, A.M. King, M.R. Meyer, D. Slade, P.Y. Lum, S.B. Stepaniants, D.D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S.H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, July 2000.

D. Colombo and M.H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15:3741–3782, 2014.

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.

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.

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.

L. S. Chen, F. Emmert-Streib, and J. D. Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome Biology*, 8, 2007.

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

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/`.

R. Scheines J. M. Robins, 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.