Causal Inference on Data Containing Deterministic Relations.

Jan Lemeire **Kris Steenhaut** **Sam Maes**
COMO lab, ETRO Department
Vrije Universiteit Brussel, Belgium
{*jan.lemeire, kris.steenhaut*}*@vub.ac.be*, *sammaes@gmail.com*

Abstract

Data containing deterministic relations cannot be handled by current constraint-based causal learning algorithms; they entail conditional independencies that cannot be represented by a faithful graph. Violation of the faithfulness property is characterized by an information equivalence of two sets of variables with respect to a reference variable. The conditional independencies do not provide information about which set should be connected to the reference variable. We propose to use the complexity of the relationships as criterion to determine adjacency. Correct decisions are made under the assumption that the complexity of relations does not increase along a causal path. This paper defines an augmented Bayesian network which explicitly models deterministic relations. The faithfulness property is redefined by using a generalized definition of the d-separation criterion, which also gives the conditional independencies following from deterministic relations, and by limiting the conditional independencies that are graphically described with the simplicity condition. Based on this, an extension to the PC learning algorithm is developed that allows the construction of minimal augmented Bayesian networks from observational data. Correct models are learned from data generated by a set of structural equations.

1 Introduction

Graphical causal models intend to describe with a Directed Acyclic Graph (DAG) the structure of the underlying physical mechanisms governing a system under study. The state of each variable, represented by a node in the graph, is generated by a stochastic process that is determined by the values of its parent variables in the graph. They are the *direct causes*.

Learning causal models from observations relies on the probabilistic independencies that follow from the system's causal structure. Consider causal model $X \rightarrow Y \rightarrow Z$. X influences Z, but is independent from Z when conditioned on Y. The conditional independency is written as $X \perp\!\!\!\perp Z \mid Y$. This independence is irrespective of the nature of the mechanisms $X \rightarrow Y$ and $Y \rightarrow Z$. The *Causal Markov Condition* gives us all independencies that follow from a causal structure: each variable is probabilistically independent of its non-effects conditional on its direct causes. The condition links the causal interpretation of a DAG with its probabilistic interpretation, which is defined by Bayesian networks [Spohn, 2001]. A Bayesian network describes the joint probability distribution P characterizing the behavior of a system. A DAG is called *faithful* if all conditional independencies of P follow from the Causal Markov Condition.

The constraint-based learning algorithms are based on the pioneering work of Spirtes et al. [1993]. The standard algorithm is the PC algorithm. The graph is constructed in two steps. The first step, called *adjacency search*, learns the undirected graph and the second step tries to orient the edges. The construction of the undirected graph is based on the property that if two nodes are adjacent in the true causal graph, they remain dependent by conditioning on every set of nodes that does not include both nodes. This property is called *Adjacency Faithfulness* [Ramsey et al., 2006]. The orientation step is based on the identification of v-structures, in which one node has incoming edges from two non-adjacent nodes. Take for example $X \rightarrow Y \leftarrow Z$. A v-structure differs from the three other possible orientations of $X - Y - Z$ by the property that X and Z are unconditional independent, but become dependent conditional on Y. For the three other possible orientations the opposite is true, X and Z are initially dependent, but become independent by conditioning on Y.

The adjacency search of the PC algorithm fails when there are deterministic relations present in the data. Consider model

$$datatype \rightarrow data\ size \rightarrow cache\ misses. \qquad (1)$$

The model indicates how the $datatype$ of the main data structure (integer, floating point, double-precision, ...) used in an algorithm affects the cache misses when executed. It is in essence the size of the datatype which determines the cache misses and not its specific type. The causal structure implies that

$$datatype \perp\!\!\!\perp cache\ misses \mid data\ size. \quad (2)$$

On the other hand, the relationship between $datatype$ and $datasize$ is a function: $datasize = f(datatype)$. Variable $datatype$ contains all information about variable $datasize$, thus

$$data\ size \perp\!\!\!\perp cache\ misses \mid datatype. \quad (3)$$

In other words, $datatype$ and $data\ size$ contain the same information about the $cache\ misses$. We call $datatype$ and $data\ size$ *information equivalent* with respect to $cache\ misses$, which is called the *reference variable*. The adjacency search will eliminate, from both conditional independencies (Eq. 2 and 3), the edge between $datatype$ and $cache\ misses$, and the edge between $data\ size$ and $cache\ misses$.

Since current constraint-based learning algorithms fail for data containing functionally determined variables [Spirtes et al., 1993, p. 57], it is required that such variables are eliminated from the input data [Scheines et al., 1996]. The argument is that such variables are not essential to the model since they contain redundant information. We argue, however, that such variables provide insight in the underlying mechanisms of the system and often reduce the complexity of the model. Reconsider the model of Eq. 1. Variable $data\ size$ indeed contains redundant information. But it is of great explanatory importance, it explains which property of the data structure affects the cache misses. The intermediate variable also extends the predictive power of the model. The number of cache misses is only determined by the size of the data, not by the exact datatype. This knowledge enables the prediction of the cache misses for new data types. Moreover, $data\ size$ is a countable variable; its domain is an ordered set. The relation with the cache misses can be expressed by a continuous function, which makes it possible to predict the cache misses for yet untested sizes. The relation between the discrete variable $datatype$ and the $cache\ misses$ is a table without predictive capacities.

Milan Studeny was one of the first to point out that Bayesian networks cannot represent all possible sets of independencies. He constructed a different framework, called *imsets* [Studeny, 2001], which is capable of representing broader sets of independencies. We advocate a different approach. We will not look for a different representation of conditional independencies, but stick to Bayesian networks. Yet, we will try to find explanations in the form of deterministic relations for the presence of conditional independencies not coming from the causal structure.

Our approach is organized as follows. In the next section we will define information equivalences and add the information about deterministic relations to an augmented Bayesian network. In Section 3, we introduce the complexity of relationships as the decision criterion for choosing which of information equivalent variables directly relates to the reference variable. Based on the complexity criterion we can define which independencies are represented by the model and extend the PC learning algorithm. The latter is done in Section 4. It is proven under which assumptions the algorithm returns a correct model. Section 5 shows that correct models are learned from data retrieved from structural equations that contain deterministic relations.

2 Information Equivalence

The conditional independencies that follow from deterministic relations result in violations of the *intersection condition*, one of the necessary conditions for faithfulness [Pearl, 1988]:

$$\boldsymbol{X} \perp\!\!\!\perp Z \mid \boldsymbol{W}, \boldsymbol{Y}\ \&\ \boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{W}, \boldsymbol{X}\ \Rightarrow\ \boldsymbol{X}, \boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{W}. \quad (4)$$

Note that single stochastic variables are denoted by capital letters, sets of variables by boldface capital letters. Violation of the condition is characterized by a kind of relative information equivalence: knowledge of either $\boldsymbol{X}$ or $\boldsymbol{Y}$ is completely equivalent from the viewpoint of Z.

Definition 1 (Information Equivalence)**.** *$\boldsymbol{X}$ and $\boldsymbol{Y}$ are called information equivalent with respect to Z - the reference variable - if*

$$\boldsymbol{X} \not\perp\!\!\!\perp Z,\ \ \boldsymbol{Y} \not\perp\!\!\!\perp Z,\ \ \boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{X},\ and\ \boldsymbol{X} \perp\!\!\!\perp Z \mid \boldsymbol{Y}. \quad (5)$$

If additionally, there is a set $\boldsymbol{Z}$ containing Z but disjoint with $\boldsymbol{X}$ and $\boldsymbol{Y}$, for which $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ holds, all three sets contain the same information about each other and are information equivalent. We call this a multi-node equivalence.

Information equivalences follow from a broader class of relations than just deterministic ones. Recall that independence $U \perp\!\!\!\perp W \mid V$ is defined as

$$\forall v \in V_{dom}, w \in W_{dom}: \quad P(U \mid v, w) = P(U \mid v)\ \ whenever\ \ P(v, w) > 0, \quad (6)$$

with X_{dom} the domain of variable X. By applying the definition on both conditional independencies of equivalence $\boldsymbol{X}$ and $\boldsymbol{Y}$ for Z it follows that [Lemeire, 2007]

$$\forall \boldsymbol{x} \in \boldsymbol{X}_{dom}, \boldsymbol{y} \in \boldsymbol{Y}_{dom}: \quad P(Z \mid \boldsymbol{x}) = P(Z \mid \boldsymbol{y})\ \ whenever\ \ p(\boldsymbol{x}, \boldsymbol{y}) > 0. \quad (7)$$

This is shown in Fig. 1 for an equivalence of single variables. The domain of Y, Y_{dom}, is partitioned into subsets

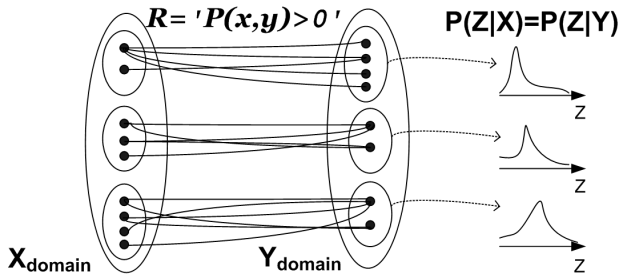

Figure 1: If X and Y are information equivalent for Z, $P(x, y)$ is only strictly positive for values of the domains of X and Y that affect $P(Z)$ similarly. These values are related by relation R.

grouping the y values having the same conditional distribution $P(Z \mid y)$. Information equivalence appears when the domain of X can be partitioned into subsets such that $P(x, y) > 0$ only if $P(Z \mid x) = P(Z \mid y)$. Each subset of Y_{dom} maps to a subset of X_{dom}.

The impact of deterministic relationships on causal models and the learning algorithms forces us to take them into account. We explicitly add the information about deterministic relations to the model.

Definition 2 (Bayesian Network$_D$)**.** *A deterministic relations augmented Bayesian network consists of a DAG G over variables $\boldsymbol{V}$, the conditional probability distributions $P(V_i \mid parents(V_i))$ and the deterministic relations $Deterministic(\boldsymbol{V})$. $Deterministic(\boldsymbol{V})$ is the set of ordered tuples of variables in $\boldsymbol{V}$, where for each tuple $\langle V_1, \ldots, V_n \rangle$, V_n is a deterministic function of $V_1, \ldots, V_{n-1}$ and is not a deterministic function of any subset of $V_1, \ldots, V_{n-1}$.*

Pearl and Verma constructed a graphical criterion, called d-separation, denoted with the symbol $\perp$, for retrieving from the causal graph all independencies following from the Causal Markov Condition. Deterministic relationships entail additional independencies. Spirtes et al. [1993], based on the work of Geiger [1990], enlarged the concept of d-separation for also retrieving them.

Definition 3. *(D-separation) Let p be a path between a node X and a node Y of a DAG G. Path p is called blocked given subset $\mathbf{Z}$ of nodes in G if there is a node w on p satisfying one of the following conditions:*

1. *w has converging arrows (along p) and neither w nor any of its descendants are in $\mathbf{Z}$, or*
2. *w does not have converging arrows (along p) and w is in $\mathbf{Z}$ or is determined by $\mathbf{Z}$.*

$\mathbf{Z}$ and the set of deterministic relations is said to D-separate X from Y in G, denoted $X \perp_D Y \mid \mathbf{Z}$, iff they block every path from X to Y.

Thus, every D-separation implies a conditional independence:

$$\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{Z} \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}. \quad (8)$$

3 The Complexity Criterion

In case of an information equivalence, the amount of information that one variable conveys about another does not give us a criterion to decide upon adjacency. We introduce the complexity of relationships as a criterion together with the Complexity Increase Assumption.

3.1 Algorithmic Mutual Information

In classical information theory, the amount of information is measured by the *mutual information* (we come back to this in Section 4.1). But this measure does not take the complexity of the relation into account. To do so, we will rely on the analogous concept of *algorithmic mutual information*, denoted as $I_A(x : y)$, defined as the decrease in Kolmogorov complexity of data sequence x when knowing y [Grünwald and Vitányi, 2003]:

$$I_A(x : y) = K(x) - K(x \mid y) \quad (9)$$

where the conditional Kolmogorov complexity $K(x \mid y)$ of x given y is the length of the shortest program that given y as input prints x and then halts. $I_A(x : y)$ expresses the additional compression of x thanks to the knowledge of y. Simpler relations lead to higher values. The complexity of the relationship between X and Y can then be quantified by estimating $I_A(x^n : y^n)$, where x^n and y^n are the vectors of the observed data, with n the sample size.

For calculating $K(x^n)$, we assume that x_i is randomly drawn from a uniform distribution in range $[x_{min}, x_{max}]$, so that $K(x^n) = n.(x_{max} - x_{min})$.

A regression analysis is used for estimating $K(x^n \mid y^n)$. It seeks the most appropriate function that fits the data, such that the function minimizes

$$f_{min} = arg\ min_{f \in \mathcal{F}} \{K(f) + K(e^n)\}, \quad (10)$$

with $\mathcal{F}$ the set of admissible functions and e^n the error vector defined as $e_i = x_i - f(y_i)$ with i from 1 to n. The model class $\mathcal{F}$ is populated with the polynomials up to degree 5, the inverse, the power, the square root and the step function. The description of the hypothesis then contains the values of the function's parameters, each needing d bits (the precision), and the function type, for which we count 1 byte for each operation (addition, subtraction, multiplication, division, power, square root and logarithm) in the function[1]. A floating-point value is encoded with d bits, whereas an integer value i requires $\log(i)$ bits.

[1]This choice of description method attributes shorter description lengths for simpler function, but nevertheless is somewhat

It is shown that the optimal precision d for each parameter is given by $d = 1/2 \log_2 n + c$, with n the sample size and c some constant [Rissanen, 1989]. Hence

$$K(f) = \#parameters.\frac{\log_2(n)}{2} + 8.\#operations + K \tag{11}$$

with K a constant term that does not depend on f. Therefore it does not play any role in finding the minimal description. The second part of Eq. 10, $K(e^n)$, reflects the goodness-of-fit of the curve $Y = f(\mathbf{X})$. By choosing the normal distribution as probability distribution of the errors (the deviances of the data with respect to the curve), $K(e^n)$ equals the sum of squared errors:

$$K(e^n) = \sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{12}$$

A regression analysis thus has to minimize the sum of Eq. 11 and Eq. 12.

3.2 Matching Complexities

If two variables X and Y are related by a linear bijection, the relation of X and Y with *any other* variable will be completely similar, qualitatively and quantitatively. Both variables contain the same information about any other variable and in the same form, so - in the absence of background knowledge - they represent equivalent quantities. The variables are indistinguishable from the perspective of the system under study. Then, they are redundant and one can be removed from the data.

3.3 Increase of Complexity

The complexity criterion makes sense by the following assumption:

Assumption 1 (Complexity Increase Assumption)**.** *Given a set of variables **V** whose causal structure can be represented by a DAG G, for all disjoint subsets **X**, **Y**, **Z** of **V** it holds that*

$$\boldsymbol{X} \perp \boldsymbol{Z} \mid \boldsymbol{Y} \; in \; G \;\Rightarrow\; I_A(\boldsymbol{X} : \boldsymbol{Z}) \leq I_A(\boldsymbol{X} : \boldsymbol{Y}). \tag{13}$$

The complexities of the relations do not decrease along a causal path. Take a system with causal structure $X \rightarrow Y \rightarrow Z$. In general the complexity of the relation $X - Z$ will not be lower than that of $X - Y$. Except if there is an exact correspondence of the relations $X - Y$ and $Y - Z$ and a neutralization of the complexities. Consider for example that $Y = X^2$ and $Z = \sqrt{Y}$, the relation between X and Z is then a simpler linear relation. The Complexity Increase Assumption is violated. This changes however when the relations are not deterministic: $Y = X^2 + \varepsilon_1$ and $Z = \sqrt{Y} + \varepsilon_2$, with ε the random disturbances. Now, in the calculation of the algorithmic mutual information between X and Z the description of the function (Eq. 11) is still smaller, but the error term (Eq. 12) will be higher due to the increase in uncertainty (a combination of ε_1 and ε_2). So with sufficiently large random disturbances and in absence of an information equivalence ($X \not\perp\!\!\!\perp Y \mid Z$ and $Z \not\perp\!\!\!\perp Y \mid X$) the algorithmic mutual information always increases along a causal path.

3.4 Redefinition of Faithfulness

Information equivalences cannot be modeled faithfully by the original definition ($\perp \Leftrightarrow \perp\!\!\!\perp$). To reestablish faithfulness, we restrict the conditional independencies that are shown graphically with the requirement that the conditioning set should provide a simpler relation.

Definition 4 (Conditional Independence and Simplicity)**.** *Conditional independence and simplicity between two sets of variables **X**, **Y** and a conditioning set **Z**, written as $\boldsymbol{X} \perp\!\!\!\perp_S \boldsymbol{Y} \mid \boldsymbol{Z}$, occurs when*

- $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$*, and*
- $I_A(\boldsymbol{X} : \boldsymbol{Z}) > I_A(\boldsymbol{X} : \boldsymbol{Y})$*, and*
- $I_A(\boldsymbol{Y} : \boldsymbol{Z}) > I_A(\boldsymbol{X} : \boldsymbol{Y})$.

The conditions about complexities of the definition only apply when information equivalences appear. If $X \perp\!\!\!\perp Y \mid Z$ and there is no information equivalence ($X \not\perp Z \mid Y$ and $Z \not\perp Y \mid X$), then we may assume that both inequalities of the definition follow, as argued in the previous section.

It trivially follows from the Causal Markov Condition ($\perp \Rightarrow \perp\!\!\!\perp$) and the Complexity Increase Assumption that d-separation entails conditional independence and simplicity: $\perp \Rightarrow \perp\!\!\!\perp_S$. We can therefore redefine the faithfulness property as follows.

Definition 5 (Faithfulness$_D$)**.** *A DAG G is called faithful$_D$ to a probability distribution P containing information equivalences if for any disjoint subsets **X**, **Y**, **Z** of **V***

$$\boldsymbol{X} \perp_D \boldsymbol{Y} \mid \boldsymbol{Z} \; in \; G \;\Leftrightarrow\; \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \; in \; P, \; and \tag{14}$$

$$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \; in \; G \;\Leftrightarrow\; \boldsymbol{X} \perp\!\!\!\perp_S \boldsymbol{Y} \mid \boldsymbol{Z} \; in \; P. \tag{15}$$

The property holds for a system if all conditional independencies of P only follow from the system's causal structure and the presence of deterministic relations.

arbitrary. The objectivity of the Kolmogorov complexity is based on the *Invariance Theorem*. The shortest programs that outputs a given string written in different universal computer languages are of equal length *up to a certain constant* [Li and Vitányi, 1997]. A complete objective measure does not exist.

4 PC$_D$ Algorithm

We first review the original PC algorithm, on which our modified algorithm will be based. Then we explain our form-free independence test which is also valid for non-linear relations. Finally, we consider the assumptions under which the algorithm returns a correct graph.

The PC algorithm [Spirtes et al., 1993] is described by Algorithm 1, where $Adj(G, X)$ denotes the set of nodes adjacent to X in graph G.

Algorithm 1 PC algorithm

- Start with the complete undirected graph U on the set of variables $\boldsymbol{V}$.

Part I ***Adjacency search***.

- $n = 0$;
- **repeat**
  - For each pair of variables A and B that are adjacent in (the current) U such that $Adj(U, A) \setminus \{B\}$ or $Adj(U, B) \setminus \{A\}$ has at least n elements, check through the subsets of $Adj(U, A) \setminus \{B\}$ and the subsets of $Adj(U, B) \setminus \{A\}$ that have exactly n variables. For all such subsets $\boldsymbol{S}$:
    - **AT** Check independency $A \perp\!\!\!\perp B \mid \boldsymbol{S}$. If independent, remove the edge between A and B in U, and record $\boldsymbol{S}$ as $Sepset(A, B)$;
  - $n = n + 1$;
- **until** for each ordered pair of adjacent variables A and B, $ADJ(U, A) \setminus \{B\}$ has less than n elements.

Part II ***Orientation.***

- Let G be the undirected graph resuling from step S2. For each unshielded triple $\langle A, B, C \rangle$ in G, orient it as $A \rightarrow B \leftarrow C$ iff B is not in $Sepset(A, B)$.
- Execute the orientation rules given in [Meek, 1995].

It is proven by Spirtes et al. [1993] that the algorithm returns a set of DAGs which contains the true DAG, if we can rely on the independence test and 4 assumptions: the Causal Markov Condition, the existence of a faithful graph for the system under study, the Minimality Condition and the Causal Sufficiency Assumption. The Minimality Condition stipulates that every proper subgraph of the real causal graph does not satisfy the Causal Markov Condition. Causal Sufficiency demands that there are no latent common causes: unknown variables which are the direct causes of at least two of the known variables.

4.1 Independence Test for Non-linear Relations

We will not make any assumption about the relationships among the variables. Most current implementations of causal inference algorithms, however, assume linear relations. Their independence test is based on Pearson's correlation coefficient. It gives a measure of how close a relation approximates linearity. Conditional independencies are measured by partial correlations, which can be calculated directly from the correlation coefficients, but only if linearity holds. Correlations can measure non-linear relations, as long as they are quasi-monotonically increasing or decreasing. Partial correlations, however, fail if the relations diverge too much from linearity. This was confirmed by our experiments.

We will use a form-free definition of dependency. The *mutual information* measures, independently from the form of the relation, the degree of association between two variables X and Y [Cover and Thomas, 1991]:

$$I(X; Y) = H(X) - H(X \mid Y). \tag{16}$$

It measures the reduction in uncertainty of X due to the knowledge of Y. The amount of uncertainty of a random variable is measured by its entropy:

$$H(X) = -\sum_{x \in X_{dom}} p(x) \log_2\ p(x), \tag{17}$$

measured in bits. For measuring the entropy of continuous variables, their distributions are discretized. This result in an analogue formula for discrete and continuous variables [Lemeire et al., 2007]. In this way, data containing a mixture of both types can be handled.

For applying the definition of mutual information, it is necessary to obtain an estimation of the underlying probability distributions. The distribution of discrete variables can be estimated by simply counting the number of occurrences of each state and dividing them by the number of data points n. For continuous variables, *kernel density estimation* makes it possible to estimate the distribution from limited sample sizes. Consult [Lemeire, 2007] and [Lemeire et al., 2007] for practical details about our implementation. It is shown for correct independency tests on data such as the one used in the experiments of Section 5 the dataset should contain at least 100 data points.

4.2 Determinism Test

Our algorithm will try to find the deterministic relationships in data. Apart from the test for information equivalences, the measurement of entropies can be used to determine functional relations. If Y is a function of $\boldsymbol{X}$, the state

of Y is known exactly when the states of all variables of set $\boldsymbol{X}$ are known. Therefore, the conditional entropy is zero:

$$Y = f(\boldsymbol{X}) \Leftrightarrow H(Y \mid \boldsymbol{X}) = 0 \tag{18}$$

Practically, we will use kernel density estimation discussed in the previous section to estimate the underlying probability distribution. From this, the conditional entropy can be calculated and if it falls below a certain threshold, Y is considered determined by $\boldsymbol{X}$. With our implementation of kernel density estimation, calibration based on experimental data returned a threshold of 0.3 bits.

4.3 PC$_D$ Algorithm

Information equivalences pose a problem for the PC algorithm. Take X and Y equivalent for Z, by $Y \perp\!\!\!\perp Z \mid X$ the adjacency search removes the edge between Y and Z and $X \perp\!\!\!\perp Z \mid Y$ deletes the edge between X and Z. Information equivalences should therefore be detected during the construction of the undirected graph. For each conditional independency that is found, it should be tested whether an equivalence can be found by swapping variables of the conditioning set with one of both arguments. The algorithm assumes that if one of two information equivalent sets has fewer elements, the relation with the reference variable is simpler.

For the PC$_D$ algorithm, replace the adjacency test step **AT** of the original algorithm with Algorithm 2.

4.4 Correctness

It is proven by Ramsey et al. [2006] that a weaker form of faithfulness is sufficient to guarantee the correctness of the PC algorithm: Adjacency Faithfulness and Orientation Faithfulness. Our modified PC algorithm will rely on an extended form of Adjacency Faithfulness, which takes deterministic relations into account.

Assumption 2 (Adjacency Faithfulness$_D$)**.** *Given a set of variables **V** whose causal structure can be represented by a DAG G, if two variables X, Y are adjacent in G, then they are dependent conditional on any subset of $\boldsymbol{V} \setminus \{X, Y\}$, except if X or Y are member of a subset of **V** which is information equivalent with another subset of **V** for respectively Y or X.*

The original definition of Orientation Faithfulness on the other hand is still correct as proven by the following theorem. Call a triple of variables $\langle X, Y, Z\rangle$ in a DAG an *unshielded triple* if X and Z are both adjacent to Y but are not adjacent to each other.

Assumption 3 (Orientation Faithfulness)**.** *Given a set of variables **V** whose causal structure can be represented by a DAG G, let $\langle X, Y, Z\rangle$ be any unshielded triple in G.*

Algorithm 2 Adjacency Test in the Presence of Deterministic Relations

- If $A \perp\!\!\!\perp B \mid \boldsymbol{S}$, check for equivalence $\boldsymbol{A}^*$ and $\boldsymbol{S}$ for B (test $\boldsymbol{S} \perp\!\!\!\perp B \mid \boldsymbol{A}^*$), and check for equivalence $\boldsymbol{B}^*$ and $\boldsymbol{S}$ for A (test $A \perp\!\!\!\perp \boldsymbol{S} \mid \boldsymbol{B}^*$), with $\boldsymbol{A}^*$ and $\boldsymbol{B}^*$ sets containing A and B respectively and some nodes adjacent to B and A respectively such that they have the same number of elements as $\boldsymbol{S}$. For every independence test, check whether the outcome can be predicted from the known deterministic relations. For example, if $A = f(\boldsymbol{S})$, then $A \perp\!\!\!\perp B \mid \boldsymbol{S}$ is true for any B.
- If no information equivalence is found, remove the edge between A and B in U, and record $\boldsymbol{S}$ as $Sepset(A, B)$.
- If an information equivalence is found, say $\boldsymbol{X}$ and $\boldsymbol{Y}$ for Z, do the following:
  - Test for deterministic relations $\boldsymbol{X} = f(\boldsymbol{Y})$ or $\boldsymbol{Y} = f(\boldsymbol{X})$. Add any deterministic relation to the augmented Bayesian network.
  - Determine $I_A(\boldsymbol{X} : Z)$ and $I_A(\boldsymbol{Y} : Z)$ as explained in section 3.1.
  - Compare both complexities. If $I_A(\boldsymbol{X} : Z) > I_A(\boldsymbol{Y} : Z)$, remove for all $Y \in \boldsymbol{Y}$ the edge between Y and Z, and record $\boldsymbol{X}$ as $Sepset(Y, Z)$. Otherwise, remove edges between $\boldsymbol{X}$ and Z, and record $\boldsymbol{Y}$ as $Sepset(Y, Z)$.
  - If $\boldsymbol{X}$ and Z are also information equivalent for $\boldsymbol{Y}$, a multi-node equivalence is found. Then, also determine $I_A(\boldsymbol{X} : \boldsymbol{Y})$ and select the simplest edges such that all nodes are connected.

- *if $X \rightarrow Y \leftarrow Z$, then X and Z are dependent given any subset of $\boldsymbol{V} \setminus \{X, Z\}$ that contains Y.*
- *otherwise, X and Z are dependent conditional on any subset of $\boldsymbol{V} \setminus \{X, Y\}$ that does not contain Y.*

Theorem 1 (Validity of Orientation Faithfulness)**.** *The conditional independencies that follow from deterministic relations do not invalidate Orientation Faithfulness when considering $\perp\!\!\!\perp_S$ as the independence test and under the Minimality Condition and Complexity Increase Assumption.*

Proof. Take $\langle X, Y, Z\rangle$ an unshielded triple in causal structure G. We will check whether independencies from deterministic relations could occur that invalidate Orientation Faithfulness.

If the unshielded triple does not form a v-structure, X and Z could become independent by conditioning on a variable X' that is information equivalent with respect to Z. By minimality, X' is related to Z via X, since X is already related to Z via Y. Then, by the Complexity Increase Assumption $I_A(X' : Z) < I_A(X : Z)$, so $X \not\perp\!\!\!\perp_S Z \mid X'$ follows and orientation faithfulness is not invalidated.

If the unshielded triple forms v-structure $X \rightarrow Y \leftarrow Z$, X and Z become independent conditional on Y when $X = f(Y)$ or $Y = f(X)$. The process of determining Y is influenced by X and Z. Both are direct causes of Y. But Z has no additional information on Y once X is known, since $Z \perp\!\!\!\perp Y \mid X$ by the deterministic relation. This violates the Minimality Condition, so this cannot happen. □

These assumptions allow us to prove the correctness of our algorithm.

Theorem 2 (Correctness of PC$_D$)**.** *Under the Causal Markov Condition, Adjacency Faithfulness$_D$, Orientation Faithfulness and the Complexity Increase Assumption the PC$_D$ algorithm is correct in the sense that given a perfect conditional independence oracle, the algorithm returns a set of DAGs that contains the true causal DAG.*

Proof. Suppose the true causal graph is G, and all conditional independencies judgments are correct. By Adjacency Faithfulness$_D$, the undirected graph resulting from step 2 has the same adjacencies as G does. If an information equivalence appears, the least complex relation is chosen. By the Complexity Increase Assumption, the remaining edge corresponds to the correct causal relation. The correct undirected graph is constructed, which has the same adjacencies as the true graph. For the second part of the algorithm, Orientation-Faithfulness guarantees that v-structures can be recognized by the independencies, as shown by Ramsey et al. [2006]. □

5 Simulation Results

In this section we report on experiments with the PC$_D$ learning algorithm based on data generated from structural equations. A model over variables $X_i \in \boldsymbol{V}$ is defined by equations of the form:

$$P(X_i \mid parents(X_i)) = f_i(parents(X_i)) + \varepsilon_i + c_i \quad (19)$$

with $parents(X_i)$ the direct causes of X_i, ε_i the stochastic variations which cannot be explained by the model and c_i a constant term. We assume that ε_i is normally distributed.

First we made a set of 8 models to test our algorithm, containing a wide variety of situations. Secondly, we tested our algorithm with 25 randomly generated structural equations models containing 10 variables. 4 variables are chosen as random input variables. The other variables have 1, 2 or 3 direct causes. One third of the variables is chosen to be deterministic ($\varepsilon_i = 0$). The functions are chosen randomly between the inverse, linear, quadratic or cubic function, the independent variables in the function are combined by a sum or a product. The coefficients, the constant and error term have to be chosen within a certain range to assure a measurable correlation of the dependent variable with each independent variable in the function. Otherwise Adjacency Faithfulness is violated. We put as a restriction that the impact of one term could not be more than 10 times lower than that of another term. The impact is defined by the range that the values of the variable take; the difference between its maximal and minimal outcome. This limitation comes from our independence test, which is capable of detecting dependencies for any form of assocation, but which has a lower resolution than Pearson's correlation coefficient.

Our implementation is an extension of the TETRAD tool (`http://www.phil.cmu.edu/projects/tetrad/`). The Java library written by Dr Michael Thomas Flanagan (`http://www.ee.ucl.ac.uk/~mflanaga`) was used for implementing the regression analysis. Applied to a dataset of 200 data points, it turned out that all learned graphs correctly represent the structural equations.

Among the test models, the following one was the most complex:

$G = A + B$	$M = 1 + 0.04K^2$
$H = 4C^2$	$N = 10 + L + F$
$I = 0.4C^3$	$O = 10 + 1.3L - F$
$J = G + 0.4H$	$P = L.F + \varepsilon$
$K = D + 0.4I + \varepsilon$	$Q = 10N/O + \varepsilon$
$L = 1.5 + 0.35K + 0.35E + \varepsilon$	$R = M + 0.4Q + \varepsilon$

The values of variables A, B, C, D, E and F are randomly chosen between 0 and 10. ε is an error term, reflecting a random disturbance with maximal value 1/8 of the variable's range (the difference between its maximal and minimal value). There are 7 deterministic variables (G, H, I, J, M, N and O). The model incorporates the different cases discussed in this paper: a bijective relation between K and M; G is determined by A and B; multi-node equivalence C, H and I; and equivalence of (L, F) and (N, O) for P and Q. The model depicted in Fig. 2 was learned from 200 data points generated by the above equations (the position of the nodes in the graph was set manually).

6 Conclusions

Deterministic relationships lead to what we have called information equivalences: two sets of variables containing information about a certain reference variable, but each of both sets getting independent from the reference variable when conditioned on the other set. This leads to violations of Adjacency Faithfulness and a failure of the constraint-

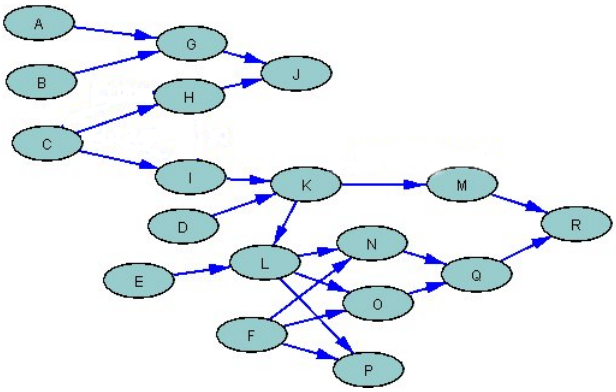

Figure 2: Model learned from random data generated by a set of structural equations.

based causal learning algorithms. It therefore makes sense to add a test for information equivalence to the algorithms, even if one does not intend to keep deterministically related variables in the data. We propose to keep deterministic variables, since they contain valuable information about the system under study, unless they are determined by a linear bijective function.

We defined an augmented Bayesian network which includes the information about deterministic relations. The PC algorithm was extended by testing for information equivalences and using the notion of complexity of relationships as a criterion to determine adjacency. The complexity of relationships is quantified by the algorithmic mutual information. The complexity criterion relies on the Complexity Increase Assumption, according to which the complexity of the relations does not decrease for more distant variables in the causal graph. This assumption, together with Orientation Faithfulness and an extended version of Adjacency Faithfulness, assure that correct models are learned. This was confirmed by our experiments with data retrieved from structural equations.

References

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

D. Geiger. *Graphoids: A Qualitative Framework for Probabilistic Inference*. PhD thesis, University of California, Los Angeles, 1990.

Peter Grünwald and Paul M. B. Vitányi. Kolmogorov complexity and information theory. with an interpretation in terms of questions and answers. *Journal of Logic, Language and Information*, 12(4):497–529, 2003.

Jan Lemeire. *Learning Causal Models of Multivariate Systems and the Value of it for the Performance Modeling of Computer Programs*. PhD thesis, Vrije Universiteit Brussel, 2007.

Jan Lemeire, Erik Dirkx, and Frederik Verbist. Causal analysis for performance modeling of computer programs. *Scientific Programming*, 15(3):121–136, 2007.

Ming Li and Paul M. B. Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proc. of the 11th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–41. Morgan Kaufmann, 1995.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, Morgan Kaufman Publishers, 1988.

Joseph Ramsey, Jiji Zhang, and Peter Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. AUAI Press, 2006.

J. Rissanen. *Stochastic Complexity in Statistical Enquiry*. World Scientific, Singapore, 1989.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. *TETRAD 3: Tools for Causal Modeling - User's Manual*. `http://www.phil.cmu.edu/projects/tetrad/tet3/master.htm`, 1996.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 2nd edition, 1993.

Wolfgang Spohn. Bayesian nets are all there is to causal dependence. In *In Stochastic Causality, Maria Carla Galaviotti, Eds*. CSLI Lecture Notes, 2001.

Milan Studeny. On non-graphical description of models of conditional independence structure. In *HSSS Workshop on Stochastic Systems for Individual Behaviours*, Louvain la Neuve, Belgium, January 2001.