Augmented Bayesian Networks for Representing Information Equivalences and Extensions to the PC Learning Algorithm.

Jan Lemeire, Sam Maes, Erik Dirkx, Kris Steenhaut, Ann Nowé

August 8, 2007

Abstract

Data containing deterministic relations entail conditional independencies that cannot be represented by a faithful graph, due to violations of the intersection condition. Such data can not be handled by current constraint-based learning algorithms. More generally, these violations are characterized by information equivalence of two sets of variables with respect to a target variable. We argue that deterministically related variables contain valuable information and should not be eliminated from the data. This paper proposes augmented Bayesian networks that explicitly model such information equivalences. For attaining minimality, only the set which has the simplest relation with the target variable is connected to it. Under the assumption that complexity does not increase along a causal path, this selection criterion results in consistent models. Under weak transitivity, faithfulness of the graph is reestablished by using the generalized definition of the d-separation criterion, called D_{eq}-separation, 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

Bayesian networks are widely used as dense representations of probability distributions. They consist of a Directed Acyclic Graph (DAG) and a Conditional Probability Distribution (CPD) for each variable. The DAG provides an explicit characterization of the conditional independencies present in the distribution that follow from the Markov condition.

If the DAG represents *all* conditional independencies, it is called *faithful* to the distribution. Faithfulness is an important property. First, it enables qualitative reasoning based on the DAG, to answer questions like 'if we know variable X, does Y have information about Z'. Second, the rationale of the causal interpretation of a Bayesian network is that when a model is capable of explaining all qualitative properties - the conditional independencies - observed in the data, the model must come close to reality. Finally, the existence of a faithful graph is required by the constraint-based learning algorithms that are able to learn causal models from data.

One of the necessary conditions for faithfulness is the intersection condition (Pearl, 1988):

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W}, \boldsymbol{Y} \;\&\; \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W}, \boldsymbol{X} \;\Rightarrow\; \boldsymbol{X}, \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W} \tag{1}$$

where the notation $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C}$ stands for the independency of $\boldsymbol{A}$ and $\boldsymbol{B}$ by conditioning on $\boldsymbol{C}$. Single stochastic variables are denoted by capital letters, sets of variables by boldface capital letters. The condition states that if two variables render the other irrelevant with respect to a third variable, neither of both can depend on that variable. This condition is violated when 2 variables contain the same information about a third variable, Z. We call them *information equivalent* with respect to Z. While both are marginally dependent on Z, either becomes conditionally independent from Z by conditioning on the other variable.

A well-known case for which the intersection condition is invalid, is when the data contains deterministic relations. Take model $X \rightarrow Y \rightarrow Z$ with $Y = f(X)$. The model implies that $X \perp\!\!\!\perp Z \mid Y$, but from the function it follows that also $Y \perp\!\!\!\perp Z \mid X$; X contains all information about Y. This cannot be represented by a faithful graph and poses problems for algorithms that try to learn the model from data. Since X and Y contain the same information about Z, it is not clear which of the two should be connected to Z. Connecting both to Z would represent redundant information. We propose to connect the one having the simplest relation with Z.

The approach developed in this paper is to reestablish the faithfulness of Bayesian networks as representation of conditional independencies by characterizing information equivalences and integrating them into an augmented model.

The next section introduces causal models, it is followed by a related work section. Section 4 defines information equivalence and the augmented models. Section 5 discusses how minimality can lead to a selection criterion for information equivalent relations and section 6 shows how faithfulness can be reestablished for data containing deterministic relations. Finally, section 7 extends the PC algorithm so that it can learn augmented models and section 8 reports on experimental verification of the extended learning algorithm.

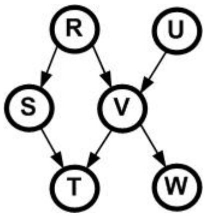

Figure 1: Example causal model.

2 Bayesian Networks

Bayesian networks offer dense representations of joint distributions. The Directed Acyclic Graph (DAG) of Fig. 1 corresponds to the following factorization:

$$P(R,S,T,U,V,W) = P(R)P(S \mid R)P(T \mid S,V)P(U)P(V \mid R,U)P(W \mid V) \tag{2}$$

The structure of the DAG implies conditional independencies. The independence of U and W conditional on V, written as $U \perp\!\!\!\perp W \mid V$, is a qualitative relational property, defined by

$$P(U \mid v,w) = P(U \mid v) \;\; whenever \;\; P(v,w) > 0 \tag{3}$$

The knowledge of W does not provide additional information about U once V is known. By an independency, the conditional distribution can be rewritten as:

$$U \perp\!\!\!\perp W \mid V \Leftrightarrow P(U \mid w) = \sum_{v \in V} P(U \mid v)P(v \mid w) \;\; whenever \;\; P(v,w) > 0 \tag{4}$$

The information shared by U and W is also present in V. It is a consequence of the model of Fig. 1.

The graphical d-separation criterion allows us to retrieve the conditional independencies from the graph that follow from the *Markov condition*. It states that a node becomes independent from all its non-descendants by conditioning on its parents. Let p be a path between a node U and a node W of a DAG G. Path p is called *blocked* given subset $\boldsymbol{V}$ of nodes in G if there is a node v on p satisfying one of the following conditions:

1. v has converging arrows (along p) and neither v nor any of its descendants are in $\boldsymbol{V}$, or

2. v does not have converging arrows (along p) and v is in $\boldsymbol{V}$.

$\boldsymbol{V}$ is said to **d-separate** U from W in G, denoted $U \perp W \mid \boldsymbol{V}$, iff members of $\boldsymbol{V}$ block every path from U to W. In the model of Fig. 1, U gets d-separated from W by V; R from T by S and V. R and U are d-separated, but are not d-separated if V is given. $R \rightarrow V \leftarrow U$ is called a *v-structure*. Conditioning unblocks a v-structure in a path, whereas it blocks non-v-structures.

A Bayesian network is called an **I-map** since all independencies it represents are present in the distribution. A Bayesian network not necessarily represents all independencies. The **faithfulness** property insists that each conditional independency in the distribution corresponds to a d-separation in the graph; that for all disjoint subsets $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$:

$$\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B} \mid \boldsymbol{C} \Leftrightarrow \boldsymbol{A} \perp \boldsymbol{B} \mid \boldsymbol{C} \tag{5}$$

A Bayesian network is minimal in the sense that no edge can be destroyed without destroying its I-mapness. Multiple Bayesian networks exist representing the same distribution though, depending on the chosen variable ordering when constructing the network. Oliver and Smith define the conditions for sound transformations of Bayesian networks, where sound means that the transformation does not introduce extraneous independencies (Oliver & Smith, 1990). No edge removal is permitted, only reorientation and addition of edges. Such transformations however eliminate some independencies represented by the original graph. Thus, if a faithful Bayesian network exists, it is the edge-minimal Bayesian network. All other Bayesian networks have more edges and represent less independencies.

Multiple faithful models can exist for a distribution though. These models represent the same set of independencies and are therefore statistically indistinguishable. They define a *Markov-equivalence class*. It is proved that they share the same v-structures and only differ in the orientation of the edges (Pearl, 2000).

3 Related Work

Recent research has developed methods for performing inferences in Bayesian Networks with functional dependencies (Cowell, Dawid, Lauritzen, & Spiegelhalter, 1999; Cobb & Shenoy, 2004). Dechter and Mateescu introduce mixed networks for expressing probabilistic and deterministic information in the form of constraints (Dechter & Mateescu, 2004), whereas we view deterministic relations as proper causal relations. Geiger, Spirtes et al. extended the d-separation criterion for retrieving the dependencies entailed by deterministic relations, which they called *D-separation* (Geiger, 1990; Spirtes, Glymour, & Scheines, 1993).

Pearl uses *stability* as the main motivation for the faithfulness of causal models (Pearl, 2000). Consider the model of Fig. 1. In general, T depends on R. T and

R are independent only when the stochastic parameterization is such that the influences via paths $R \rightarrow S \rightarrow T$ and $R \rightarrow V \rightarrow T$ cancel out exactly . This system is called unstable because a small change in the parameterization will result in a dependency. The unhappy balancing act is a measure zero event, the chance of such a coincidence can therefore be regarded as having zero probability. Deterministic relations, however, appear in nature and are not coincidences.

Current constraint-based learning algorithms fail for data containing functionally determined variables (Spirtes et al., 1993), they require that such variables are eliminated from the input data (Scheines, Spirtes, Glymour, Meek, & Richardson, 1996). The argument is that such variables are not essential to the model since they contain redundant information. In section 4.2 we show, however, that such variables provide insight in the underlying mechanisms and often reduce the complexity of the model. Moreover, determinism is not always known a priori.

For the faithfulness of graphical models, many conditions should hold (Pearl, 1988). Therefore, other representation schemes of independency information were developed, such as the imsets of Studeny (Studeny, 2001), which can model any conditional independence structure. Our approach claims that if violations of faithfulness come from local properties, these properties should be integrated into the causal modeling framework.

4 Augmented Bayesian Networks

A necessary condition for the existence of a faithful graph is the intersection condition. This condition is violated by information equivalence.

4.1 Information equivalence

Definition 1 (information equivalence) *$\boldsymbol{X}$ and $\boldsymbol{Y}$ are called **information equivalent** with respect to Z, the target variable, if*

- $\boldsymbol{X} \not\perp\!\!\!\perp Z$ *and* $\boldsymbol{Y} \not\perp\!\!\!\perp Z$
- $\boldsymbol{Y} \perp\!\!\!\perp Z \mid \boldsymbol{X}$
- $\boldsymbol{X} \perp\!\!\!\perp Z \mid \boldsymbol{Y}$

Knowledge of either $\boldsymbol{X}$ or $\boldsymbol{Y}$ is completely equivalent from the viewpoint of Z.

A variable Y is **determined** by a set of variables $\boldsymbol{X}$ if the relation between Y and the set $\boldsymbol{X}$ is a function, written as $Y = f(\boldsymbol{X})$. A functional relation implies that the variables $\boldsymbol{X}$ contain all information about Y. If Y is correlated to a third variable Z, all information from Y about Z is also present in $\boldsymbol{X}$. If additionally

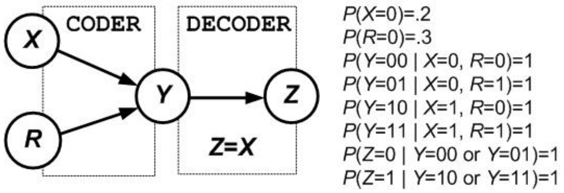

Figure 2: Causal model of SGS (Fig. 3.23) in which Z equals X

$\boldsymbol{X} \perp\!\!\!\perp Z \mid Y$ holds, meaning that $\boldsymbol{X}$ contains no additional information about Z, then Y and $\boldsymbol{X}$ are information equivalent with respect to Z. For a bijection, $Y = g(X)$ and $X = g^{-1}(Y)$, *each* variable dependent on X or Y implies an information equivalence.

Consider the coder-decoder example, taken from SGS (Fig. 3.23) (Spirtes et al., 1993), shown in Fig. 2. Variable Y encodes the values of both R and X, and Z decodes Y to match the value of X. This is possible because it is the first bit of Y that corresponds to the value of X. X is therefore deterministically related to Z, though not adjacent in the graph. Both X and Y are information equivalent with respect to Z.

Information equivalences follow from a broader class of relations than just deterministic ones. By applying Eq. 3 on both conditional independencies of equivalence $\boldsymbol{X}$ and $\boldsymbol{Y}$ for Z it follows that

$$P(Z \mid \boldsymbol{x}) = P(Z \mid \boldsymbol{y}) \;\; whenever \;\; p(\boldsymbol{x},\boldsymbol{y}) > 0 \tag{6}$$

This shown in Fig. 3. The domain of Y is partitioned into subsets grouping the y values having the same $P(Z \mid y)$. 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}.

If, besides the independencies of Definition 1, additionally $X \perp\!\!\!\perp Y \mid Z$ holds, then all three variables contain the same information about each other and are information equivalent. We call this a *multi-node equivalence*.

4.2 Example

Fig. 4 represents a causal model of performance related data of a quicksort algorithm. The overall performance is measured by the computation time (T_{comp}). $\#op$ represents the number of basic compare-swap operations, which is affected by the $array\ size$. The time to compute one compare-swap operation (T_{op}) depends on the number of processor cycles for one operation (C_{op}) and the processor's clock frequency (f_{clock}). The number of cycles consists of the cycles spent executing

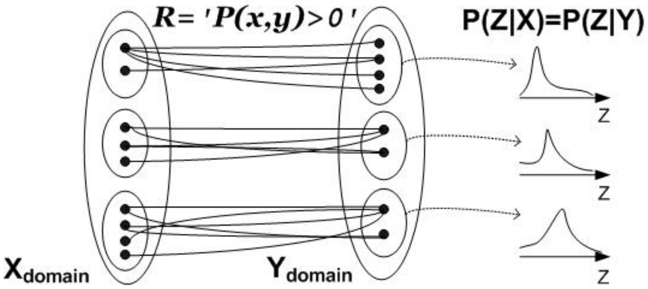

Figure 3: Variables X and Y are information equivalent for Z. $P(x, y)$ is only strictly positive for values that affect $P(Z)$ similarly. These values are related by relation R.

the instructions of one operation ($\#instr_{op}$) and the cycles spent waiting due to memory accesses, which are triggered by the $cache\,misses$. These are determined by the $memory$ capacity, the $array\ size$ and $element\ size$, which is the size in bytes of the elements of the array. Finally, the data type of the elements (variable $element\ type$) affects C_{op}, and determines $\#instr_{op}$ and the $element\ size$. The causal interpretation of the edges should be read as “a change of the state of A causes a change of the state of B”.

Deterministic variables are depicted with double-bordered circles, they are determined by their parents in the graph. Even the other non-input variables are quasi-determined by their parents, since a serial computer is deterministic. One can argue that these deterministic variables can be omitted and that the computation time can directly be expressed in function of the parameters. The intermediate variables, however, are of great explanatory importance and extend the predictive power of the model. The cache misses, for example, are only determined by the size of the data, not by the data type. This knowledge enables the prediction of the cache misses for new data types. Moreover, $element\ size$ is a countable variable, whose 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 unknown sizes. The relation of the discrete variable $element\ type$ and the $cache\ misses$ is a table without predictive capacities.

4.3 Assumptions

In order to be able to investigate the effect of an information equivalence on the conditional independencies between other variables, we will assume a condition that expresses a kind of transitivity.

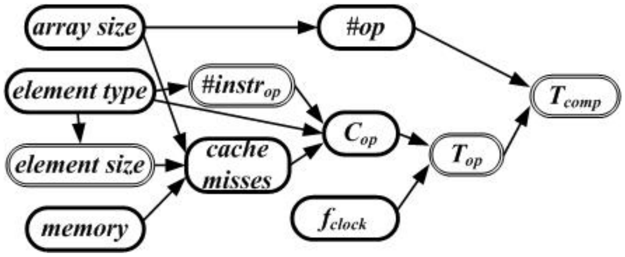

Figure 4: Causal model of the performance of the quicksort algorithm.

Assumption 1 *Weak Transitivity*

$$\boldsymbol{T}\perp\!\!\!\perp\boldsymbol{V} \mid \boldsymbol{W} \;\;\&\;\; \boldsymbol{T}\perp\!\!\!\perp\boldsymbol{V} \mid \boldsymbol{W}, \boldsymbol{U} \;\;\Rightarrow\;\; \boldsymbol{T}\perp\!\!\!\perp\boldsymbol{U} \mid \boldsymbol{W} \;\; \textit{or} \;\; \boldsymbol{U}\perp\!\!\!\perp\boldsymbol{V} \mid \boldsymbol{W} \tag{7}$$

Or, put the other way around:

$$\boldsymbol{T}\not\perp\!\!\!\perp\boldsymbol{U} \mid \boldsymbol{W} \;\;\&\;\; \boldsymbol{U}\not\perp\!\!\!\perp\boldsymbol{V} \mid \boldsymbol{W} \;\;\Rightarrow\;\; \boldsymbol{T}\not\perp\!\!\!\perp\boldsymbol{V} \mid \boldsymbol{W} \;\; \textit{or} \;\; \boldsymbol{T}\not\perp\!\!\!\perp\boldsymbol{V} \mid \boldsymbol{W}, \boldsymbol{U} \tag{8}$$

It is one of the necessary conditions for the existence of a faithful graph (Pearl, 1988). Eq. 8 says that if T depends on U and U depends on V, it implies that either T depends on V (e.g. as in model $T \rightarrow U \rightarrow V$) or becomes dependent by conditioning on U (e.g. as by v-structure $T \rightarrow U \leftarrow V$)

Take again the coder-decoder example shown in Fig. 2. X determines the first bit of Y and R the second. The decoding of Z is determined by the first bit of Y. This model, however, violates the weak transitivity condition. Y depends on X, R and Z; but R is independent from X and Z, also after conditioning on Y. The values of variable Y reflect two separate quantities, one that is determined by X and one by R. Each value of Y combines both quantities. The coder-decoder system is designed to exhibit this specific behavior.

The following assumptions are introduced to simplify the discussion. They exclude some exotic cases in which the probability distributions exhibit very specific regularities. For the experiments presented in section 8 these assumptions hold.

Assumption 2

$$\boldsymbol{X}\not\perp\!\!\!\perp\boldsymbol{Z} \;\;\&\;\; \boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Z} \mid \boldsymbol{Y} \;\;\&\;\; \boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Z} \mid \boldsymbol{C} \;\;\Rightarrow\;\; \boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Z} \mid \boldsymbol{Y}, \boldsymbol{C} \tag{9}$$

Assumption 3 *If X and Y are information equivalent with respect to a variable Z, it follows that*

$$X \perp\!\!\!\perp Y \mid D \Rightarrow X, Y \perp\!\!\!\perp Z \mid D \tag{10}$$

If D contains all information shared by X and Y, D also contains the information that X and Y have about Z.

4.4 Properties of Information Equivalences

The following properties prove that information equivalences can be reduced to a set of fundamental equivalences. An example model where these properties hold for the variables with the same names as in the property descriptions is given in Fig. 8. The first property shows that an information equivalence remains for all variables related to the information equivalent variables via the target variable.

Property 1 *If X and Y are information equivalent with respect to a variable Z and for variable A it holds that $A \not\perp\!\!\!\perp Z$ and Z screens off X and Y from A ($X \perp\!\!\!\perp A \mid Z$ and $Y \perp\!\!\!\perp A \mid Z$), then X and Y are information equivalent with respect to A.*

Proof:
From $X \perp\!\!\!\perp A \mid Z$ it follows that (using Eq. 4)

$$\begin{aligned} P(A \mid X) &= \sum_{z \in Z} P(A \mid z).P(z \mid X) \\ &= \sum_{z \in Z} P(A \mid z).P(z \mid Y) = P(A \mid Y) \end{aligned} \tag{11}$$

The last step is true by $Y \perp\!\!\!\perp A \mid Z$. ■

The next property says something about the conditional independencies implied by information equivalences. If among information equivalent variables, one contains all information that a variable has about the target variable, the other equivalent variables also contain this information.

Property 2 *If X and Y are information equivalent with respect to a variable Z, it follows that*

$$Z \perp\!\!\!\perp B \mid X \Leftrightarrow Z \perp\!\!\!\perp B \mid Y \tag{12}$$

Proof:
We partition de domain of X into subsets X^k_{dom} for which $P(Z \mid x)$ is the same, namely $P(Z \mid k)$. There are two such subsets, since $P(Z \mid x) \neq P(Z)$.

$$\begin{aligned} P(Z \mid B) &= \sum_{x \in X_{dom}} P(Z \mid x).P(x \mid B) = \sum_k \sum_{x \in X^k_{dom}} P(Z \mid x).P(x \mid B) & (13) \\ &= \sum_k P(Z \mid k) \sum_{x \in X^k_{dom}} P(x \mid B) & (14) \\ &= \sum_k P(Z \mid k) \sum_{x \in X^k_{dom}} \sum_{y \in Y_{dom}} P(x \mid y, B).P(y \mid B) & (15) \end{aligned}$$

Each subset X^k_{dom} maps to a subset Y^l_{dom} for which $P(Z \mid k) = P(Z \mid l)$. By Eq. 6, $P(x \mid y, B)$ is only positive whenever $x \in X^k_{dom}$ and $y \in Y^l_{dom}$, thus:

$$P(Z \mid B) = \sum_k P(Z \mid k) \sum_{y \in Y^l_{dom}} P(y \mid B) \sum_{x \in X^k_{dom}} P(x \mid y, B) \qquad (16)$$

The equivalence also implies that $\sum_{x \in X^k_{dom}} P(x \mid y, B) = 1$ if $y \in Y^l_{dom}$, so:

$$\begin{aligned} P(Z \mid B) &= \sum_l P(Z \mid l) \sum_{y \in Y^l_{dom}} P(y \mid B) & (17) \\ &= \sum_{y \in Y_{dom}} P(Z \mid y).P(y \mid B) \;\Leftrightarrow\; Z \perp\!\!\!\perp B \mid Y & (18) \end{aligned}$$

$\blacksquare$

The following property proves that an information equivalence can be reduced to another if there is a variable which makes the target independent from the information equivalent variables.

Property 3 *If X and Y are information equivalent with respect to a variable Z and assumptions (1), (2) and (3) hold, it follows that*

$$X \perp\!\!\!\perp Z \mid C \Leftrightarrow Y \perp\!\!\!\perp Z \mid C \qquad (19)$$

If additionally $C \not\perp\!\!\!\perp Z \mid X$, then X and Y are also information equivalent for C. Otherwise, C is together with X and Y information equivalent for Z.

Proof:
By assumption 2, $X \perp\!\!\!\perp Z \mid C, Y$ holds. Weak transitivity (Eq. 7) then demands that

$X \perp\!\!\!\perp Y \mid C$ or $Y \perp\!\!\!\perp Z \mid C$ holds. This proves the second independency, because if the first is true, the second follows from assumption 3. Then, by assumption 2 again, $Y \perp\!\!\!\perp Z \mid C, X$ holds, which means that the information equivalence remains under conditioning on C.

For proving the equivalence, we have to show that (a) $Y \perp\!\!\!\perp C \mid X$ and (b) $X \perp\!\!\!\perp C \mid Y$. From $Y \perp\!\!\!\perp Z \mid X$ and $Y \perp\!\!\!\perp Z \mid C, X$, weak transitivity demands that $Y \perp\!\!\!\perp C \mid X$ or $C \perp\!\!\!\perp Z \mid X$. The second independence is false, which proves the first independence and thus (a). Independence (b) is proved with the same arguments. $C \not\perp\!\!\!\perp Z \mid Y$ holds, because $C \perp\!\!\!\perp Z \mid Y$ would mean that C and Y are information equivalent with respect to Z. But then, by transitivity of information equivalences (follows directly from Eq. 6), C and X would be information equivalent for Z, contradicting the given $C \not\perp\!\!\!\perp Z \mid X$.

Finally, if $C \perp\!\!\!\perp Z \mid X$, then, by transitivity, X, C and Y are equivalent for Z. ▮

Property 4 *Under assumptions (1), (2) and (3), information equivalent variables are adjacent and at least one of the variables is adjacent to the target variable of a basic information equivalence.*

Proof:
Take X and Y information equivalent for Z. X and Y are dependent, so there is a path connecting both. If there is another variable on the path, for example D, making X and Y independent: $X \perp\!\!\!\perp Y \mid D$. By assumption 3, $X \perp\!\!\!\perp Z \mid D$ follows. Then, by property 3, either D is also information equivalent for Z or X and Y are information equivalent with respect to D. With the given $X \perp\!\!\!\perp Y \mid D$, the three variables form a multi-node equivalence.

By the dependency of the equivalent variables with the target variable, there must be a path connecting them. If this path is blocked by a variable C, by property 3, variable C is also equivalent for Z; or X and Y form a basic information equivalence for C, while X and Y do not for Z. ▮

4.5 Definition and Notation

An information equivalence is called *basic* if no variable exists that contains more information about the target variable than the equivalent variables. The previous section showed that the basic information equivalences suffice to augment the model. Other equivalences can easily be derived from it. Deterministically related variables, however, possibly generate multiple equivalences. Since for $Y = f(\boldsymbol{X})$,

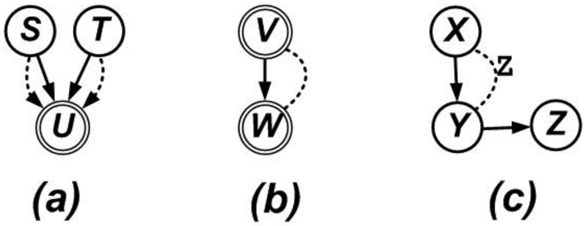

Figure 5: Augmented Causal Model with a Functional Relation (a), a Bijection (b) and an Information Equivalence (c).

$\boldsymbol{X}$ is equivalent for all variables related to Y. A deterministic relation is thus more fundamental and is added to the model instead of all equivalences that follow from it.

Definition 2 *An information equivalence augmented Bayesian network consists of a DAG G over variables $\boldsymbol{V}$, the conditional probability distributions $P(V_i \mid parents(V_i))$, the deterministic relations $Deterministic(\boldsymbol{V})$ and the information equivalences $Equivalences(\boldsymbol{V})$. $Deterministic(\boldsymbol{V})$ 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 $V_1, \ldots, V_{n-1}$. $Equivalences(\boldsymbol{V})$ is the set of ordered tuples of sets of variables in $\boldsymbol{V}$, where for each tuple $\langle \boldsymbol{W}_1, \ldots, \boldsymbol{W}_n \rangle$, the sets $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{n-1}$ are information equivalent with respect to $\boldsymbol{W}_n$.*

We propose the following notation. Deterministic nodes are depicted with double-bordered circles with dashed edges coming from the determining variables, as shown in Fig. 5 (a). If the parents comprise all the determining variables, the dashed edges may be omitted. Two variables related by a bijection are linked with an unoriented dashed edge (Fig. 5 (b)). Information equivalent variables are connected by a dashed edge annotated with the target variable (Fig. 5 (c)). We do not provide a notation for equivalences of sets of variables, this would require hyperedges. Such equivalences can be added as text to the graph. Besides, they rarely occur in practice.

5 The Complexity Criterion

Modeling is based on the minimality criterion, i.e. we should seek, in spirit of Occam's Razor, the simplest model that is able to describe the data. In the context of causal models, basically, dependent variables are connected with an edge, whereas variables that become independent when conditioned on others are not directly related. For a basic information equivalence, X and Y for Z, there is no other variable that makes X and Y independent from Z. This implies that they should be related. For a faithful representation however, the two independencies, $Y \perp\!\!\!\perp Z \mid X$ and $X \perp\!\!\!\perp Z \mid Y$, suggest that neither X or Y is adjacent to Z. On the other hand, including both edges would disrupt the minimality condition, since both variables have the same information about the target. Relating one of both with Z suffices to model the information they contain about Z. Except if additionally $X \perp\!\!\!\perp Y \mid Z$, when the three variables form a multi-node equivalence. Then, two of the three possible edges between the three variables are sufficient to reflect the dependencies. In the coder-decoder model of Fig. 2, X, Y and Z reflect this case.

5.1 Complexity of Relations

The relations of information equivalent variables with the target variable represent the same 'information transfer'. We therefore need criteria, different from conditional independencies, to select among information equivalent relations. In absence of background knowledge, the only objective criterion is the *complexity* of the relations, according to which simpler relations should be preferred over complex ones. The choice between two equivalent variables X and Y for being adjacent to the target node Z is decided upon which relation, $Z - X$ or $Z - Y$, is the simplest.

Shannon's mutual information, defined by the decrease in entropy (uncertainty) of a variable due to knowledge of another, measures the information one variable conveys about the other. But it does not take the complexity of the relation into account. Therefore we will rely on the analogous *algorithmic mutual information*, denoted as $I_A(x : y)$, defined as the decrease in Kolmogorov complexity (Grünwald & Vitányi, 2003):

$$I_A(x : y) = K(x) - K(x \mid y) \tag{20}$$

It reflects the additional compression of x thanks to the knowledge of y, where simpler relations lead to higher values. This measure is proved to be symmetric (Grünwald & Vitányi, 2003). The complexity of an individual object x is measured by its Kolmogorov complexity $K(x)$, defined as the length of the shortest program that prints the object and then halts. 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. The simplicity of relation $X - Y$ 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 sample size n.

When the complexities of the relations match, we have to decide upon other criteria. If, for example, X and Y are related by a linear bijection, the relation with any other variable will be similar. We will then connect the target to the variable(s) which is/are cause(s) of the other equivalent variable(s).

5.2 Practical Complexity Measurement

For a relation among continuous variables, 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

$$f_{best} = arg\, min_{f \in \mathcal{F}} \{K(f) + K(e^n)\} \tag{21}$$

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. This method guarantees a trade-off between hypothesis complexity and goodness-of-fit on the data. This approach corresponds to the Minimum Description Length (MDL) approach, according to which we should pick the model which minimizes the sum of the description length of the hypothesis, and the description length of the data encoded with the help of the hypothesis (Grünwald, Myung, & Pitt, 2005). The set of functions $\mathcal{F}$ is filled with functions appropriate for the system under study. We added the polynomial functions up to degree 5, the inverse, power, square root and step function, and a limited number of combinations of these basic functions. The complexity of the functions is calculated as the sum of the complexities of the functions parameters and the function type, for which we count 1 byte for each operation (addition, subtraction, product, power and condition) in the function [1]. A floating-point value is encoded with d bits, whereas an integer value i requires $\log(i)$ bits.

It has been shown that the optimal precision d for each parameter is given by $d = 1/2 \log n + c$, with n the sample size and c a constant (Rissanen, 1989). Thus

$$K(f) = \#parameters. \frac{\log(n)}{2} + 8.\#operations + C \tag{22}$$

with C a constant term that does not depend on H and that therefore does not play a role in finding the minimal description. The second part of the description reflects the goodness-of-fit of the curve $y = f(\mathbf{x})$. By choosing the normal distribution as

[1] This choice of complexity measurement attributes shorter description lengths for simpler functions, but nevertheless is somewhat arbitrary. The objectivity of Kolmogorov complexity is based on the *Invariance Theorem*. The shortest program that outputs a given string, when written in different universal computer languages will be of equal length *up to a certain constant* (Li & Vitányi, 1997). A complete objective measure is thus not possible.

probability distribution of the errors (the deviances of the data with respect to the curve), $L(D \mid H)$ equals the sum of squared errors:

$$K(e^n) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{23}$$

where the values of x_i and y_i are encoded with precision p. For calculating $I_A(x^n : y^n)$, we assume that x_i is randomly drawn from $[x_{min}, x_{max}]$, so that $K(x^n) = n.(x_{max} - x_{min})/p$.

The regression analysis has to minimize the sum of Eq. 22 and Eq. 23. The Java library, written by Dr Michael Thomas Flanagan (`http://www.ee.ucl.ac.uk/~mflanaga`), is used for calculating the closest fit of each function. If some of the parents are discrete variables, several distinct curves are considered, one for each value combination of the discrete variables. The total complexity is then the summation over all individual functions, except that equal parameter values or function types are only counted once.

For discrete variables, the conditional distributions $P(x_i \mid parents(x_i))$ are described by discrete distributions. The number of probabilities (written with precision d) in the probability table determine its complexity.

We will assume that if one of two information equivalent sets has fewer elements, the relation with the target variable is simpler.

5.3 Complexity Increase

The complexity criterion makes sense by making the following assumptions:

Assumption 4 *The Complexity Increase assumption:*

$$\begin{aligned} &\boldsymbol{A} \not\perp\!\!\!\perp \boldsymbol{C} \;\&\; \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B} \\ &\qquad \Rightarrow\; I_A(\boldsymbol{A} : \boldsymbol{C}) \leq I_A(\boldsymbol{A} : \boldsymbol{B}) \;\&\; I_A(\boldsymbol{A} : \boldsymbol{C}) \leq I_A(\boldsymbol{B} : \boldsymbol{C}) \qquad (24) \\ &\boldsymbol{A} \not\perp\!\!\!\perp \boldsymbol{D} \;\&\; \boldsymbol{A} \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{B} \;\&\; \boldsymbol{C} \not\perp\!\!\!\perp \boldsymbol{D} \;\&\; \boldsymbol{C} \perp\!\!\!\perp \boldsymbol{D} \mid \boldsymbol{B} : \\ &\qquad I_A(\boldsymbol{A} : \boldsymbol{B}) < I_A(\boldsymbol{B} : \boldsymbol{C}) \;\Leftrightarrow\; I_A(\boldsymbol{A} : \boldsymbol{D}) < I_A(\boldsymbol{C} : \boldsymbol{D}) \qquad (25) \end{aligned}$$

This assumption implies that if $X \to Y \to Z$ is the true model and X and Y are information equivalent for Z, connecting the nodes having the simplest relation gives the correct model. Fig. 6 illustrates both cases of the assumption. The complexities of the relations do not decrease when variables are more distant in a causal model. This would happen by correspondences of the relations and neutralization of its complexities. Note the similarity with the *data processing inequality*, which

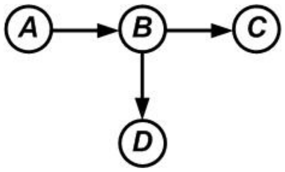

Figure 6: Example causal model with $I_A(A : B) < I_A(B : C)$.

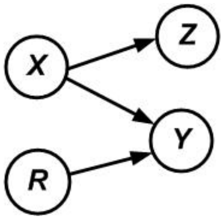

Figure 7: Simplest, but incorrect model for the system of Fig. 2.

states that, if $A \perp\!\!\!\perp C \mid B$, the mutual information of A and C cannot be higher than that of A and B.

In case of independent causal mechanisms, complexity increase is what we 'normally' can expect. It will only rarely lead to cancellation of complexities. Except in specially-designed systems, such as the coder-decoder model of Fig. 2. In which X and Y are equivalent for Z, but the relation $X - Z$ is simpler than $Y - Z$. The complexity increase assumption is violated, due to a complete dependence of the decoding relation $Y \rightarrow Z$ on both $X \rightarrow Y$ and $R \rightarrow Y$. Hence, a learning algorithm would consider the $X - Z$ relation as a direct one and not the more complex $Y - Z$ relation, as shown in Fig. 7. In the context of learning, choosing the simplest model is the best strategy (Grünwald et al., 2005). It overcomes overfitting and even if the learned model deviates from the true model, it will give good predictions about the behavior of the system. The model of Fig. 7 will correctly predict the behavior of the coder-decoder and presumably represent the aim of the system.

6 Faithfulness

Faithful models provide a compact representation of all independencies of a distribution. To capture the independencies that follow from information equivalences, causal models are augmented. The complexity criterion determines which independencies are considered to build the graph.

6.1 Conditional Independence and Simplicity

An information equivalence cannot be modeled faithfully. Therefore we restrict the conditional independencies that are shown graphically with the requirement that the conditioning set should provide a simpler relation in case of information equivalence.

Definition 3 *(Conditional independence and simplicity)* ***Conditional independence and simplicity*** *between two sets of variables* $\boldsymbol{X}$, $\boldsymbol{Y}$ *and a conditioning set* $\boldsymbol{Z}$, *written as* $\boldsymbol{X} \perp\!\!\!\perp_S \boldsymbol{Y} \mid \boldsymbol{Z}$, *occurs when*

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

6.2 D_{eq}-separation

When there are deterministic relationships among variables, there are conditional independencies that are not entailed by the Markov condition alone. SGS (Spirtes et al., 1993), based on the work of Geiger (Geiger, 1990), enlarged the concept of d-separation to create a graphical condition for retrieving all conditional independencies from a graph and a set of deterministic relations. They called it D-separation. We enlarge the criterion to also capture independencies following from information equivalences.

Definition 4 *(*D_{eq}*-separation) Let* p *be a path between a node* U *and a node* W *of a DAG* G*. Path* p *is called blocked given subset* $\boldsymbol{V}$ *of nodes in* G *and a set of deterministic relations and information equivalences if there is a node* v *on* p *satisfying one of the following conditions:*

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

$\boldsymbol{V}$ *and the set of deterministic relations and information equivalences is said to* D_{eq}***-separate*** U *from* W *in* G*, denoted* $U \perp_{eq} W \mid \boldsymbol{V}$*, iff members of* $\boldsymbol{V}$ *block every path from* U *to* W *or there is an equivalence of* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *with respect to* Z *such that*

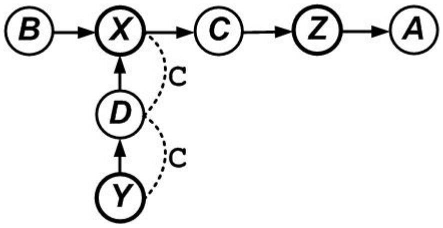

Figure 8: Model with X and Y information equivalent for Z and additional nodes depicting the possible consequences.

1. *$\boldsymbol{Y} \subset \boldsymbol{V}$, and*
2. *the members of $(\boldsymbol{V} \setminus \boldsymbol{Y}) \cup \boldsymbol{X}$ block every path from U to W, and*
3. *the members of $(\boldsymbol{V} \setminus \boldsymbol{Y}) \cup \{Z\}$ block every path from U to W, and*
4. *members of $(\boldsymbol{V} \setminus \boldsymbol{Y})$ do not unblock a path between U and W that is not blocked by $\boldsymbol{X}$.*

Take the model of Fig. 8, A and B are d-separated by X, but not by Y. If, however, X and Y are information equivalent with respect of C, A and B are D_{eq}-separated by conditioning on Y.

6.3 Faithfulness Revisited

Given the additional independencies that information equivalences entail, the definition of faithfulness should be reconsidered. In cases of information equivalence, the independencies depicted graphically are restricted by the definition of conditional independency and simplicity ($\perp\!\!\!\perp_s$). On the other hand, the extended d-separation criterion ($\perp_{eq}$) makes it possible to retrieve the independencies following from information equivalences.

Definition 5 *A causal model is called **faithful**$_{eq}$ to a probability distribution containing information equivalences if*

$$\boldsymbol{X} \perp_{eq} \boldsymbol{Y} \mid \boldsymbol{Z} \;\Leftrightarrow\; \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \tag{26}$$

$$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \;\Leftrightarrow\; \boldsymbol{X} \perp\!\!\!\perp_s \boldsymbol{Y} \mid \boldsymbol{Z} \tag{27}$$

We show that the definition makes sense by proving that the consequences of combinations of an information equivalence and conditional independencies that follow from the Markov condition can be captured by a model that is $faithful_{eq}$. Take X and Y equivalent for Z and $X - Z$ the simplest relation, $I_A(X : Z) > I_A(Y : Z)$. Consider the conditional independence statements containing at least two of the three variables of the information equivalence. There are ten possible combinations.

1. $X \not\perp\!\!\!\perp A$ and $X \perp\!\!\!\perp A \mid Z$:

   - If also $Y \perp\!\!\!\perp A \mid Z$, by property 1 it follows that X are Y equivalent for A. The second part of the Complexity Increase Assumption assures that $I_A(X : A) > I_A(Y : A)$, thus $Y \perp\!\!\!\perp_S A \mid X$, but $X \not\perp\!\!\!\perp_S A \mid Y$. An example is shown in Fig. 8.
   - On the other hand, if $Y \not\perp\!\!\!\perp A \mid Z$, then Y is connected to A via an alternative path and has more information about A than X.

2. $Z \not\perp\!\!\!\perp B$ and $Z \perp\!\!\!\perp B \mid X$:
   Independency $Z \perp\!\!\!\perp B \mid Y$ follows property 2. Then, there are two possibilities:

   - If B has less information about Z ($Z \not\perp\!\!\!\perp X \mid B$), it is related to Z via X, as shown in Fig. 8. By D_{eq}-separation the conditional independency $Z \perp\!\!\!\perp B \mid Y$ can be retrieved from the graph.
   - If on the contrary variable B contains as much information about Z as X, all three nodes are equivalent for Z. This is shown by node D in Fig. 8. The node having the simplest relation with Z is related to Z, which is X in the figure.

3. $Z \not\perp\!\!\!\perp C$ and $X \perp\!\!\!\perp Z \mid C$:
   By property 3, Y also gets independent, $Y \perp\!\!\!\perp Z \mid C$. Then, there are two possible cases:

   - If $C \perp\!\!\!\perp Z \mid X$, then C is also information equivalent with respect to Z, which is discussed in the previous case.
   - If $C \not\perp\!\!\!\perp Z \mid X$, then C has more information about Z. Property 3 proves that X or Y are information equivalent for C as well. This case is shown by node C in Fig. 8. By the second part of the Complexity Increase Assumption, $X - C$ must be simpler than $Y - C$, thus $Y \perp\!\!\!\perp_S C \mid X$.

4. $X \perp\!\!\!\perp Y \mid D$:
   By assumption 3, it follows that $X \perp\!\!\!\perp Z \mid D$, which is discussed by case 3.

5. Independency $X \perp\!\!\!\perp E \mid Y$ only interferes with the equivalence if there is an independence with Z. This is discussed by the previous cases.

The 5 remaining cases, $Y \perp\!\!\!\perp A \mid Z$, $Z \perp\!\!\!\perp B \mid Y$, $X \perp\!\!\!\perp Z \mid C$, $Y \perp\!\!\!\perp Z \mid D$ and $Y \perp\!\!\!\perp D \mid X$, are equivalent to respectively cases 1, 2, 3, 4 and 5.

7 Constraint-based Learning Algorithms

The constraint-based learning algorithms are based on the pioneering work of Spirtes, Glymour and Scheines (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 two nodes are adjacent if they remain dependent by conditioning on every set of nodes that does not include both nodes. The algorithm starts with a complete undirected graph and removes edges for each independency that is found. The number of nodes in the conditioning set is gradually increased up to a certain maximal number, called the *depth* of the search. The orientation step is based on the identification of v-structures of the form $X \rightarrow Y \leftarrow Z$, for which X and Z are independent, but become dependent conditional on Y. Recall that for all three other possible orientations of $X - Y - Z$ the opposite is true, X and Z are initially dependent, but become independent by conditioning on Y.

Besides 6 general assumptions (Scheines et al., 1996), the PC algorithm requires *causal sufficiency*, which means that all common causes should be known: variables that are the direct cause of at least two variables. Basically, if the data is random and faithful to at least one graph, the PC algorithm leads to a set of observationally indistinguishable models, which all describe the same conditional independencies. These models have the same undirected graph and v-structures, they only differ in the orientation of edges, which could not be directed due to the absence of v-structures.

7.1 Equivalence Detection

Information equivalences pose a problem for the constraint-based algorithms. Take X and Y equivalent for Z, by $Y \perp\!\!\!\perp Z \mid X$ the algorithm would remove the $Y - Z$ edge and $X \perp\!\!\!\perp Z \mid Y$ deletes the $X - Z$ edge. 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. Furthermore, equivalences imply independencies. For equivalence X and Y for Z, any independency $Y \perp\!\!\!\perp Z \mid X, U$ would be a consequence of the information equivalence. Such tests can thus be skipped in the procedure. This results in the following algorithm.

Algorithm 1 Information equivalence detection during adjacency search of PC algorithm

For each test $U \perp\!\!\!\perp V \mid \boldsymbol{W}$ during the adjacency search:

1. Skip test if an equivalence $\boldsymbol{U}^{+}$ and $\boldsymbol{W}^{-}$ for V, or $\boldsymbol{V}^{+}$ and $\boldsymbol{W}^{-}$ for U has been found previously. $\boldsymbol{U}^{+}$ is defined as a set containing U and some other nodes and $\boldsymbol{W}^{-}$ denotes a subset of $\boldsymbol{W}$.

2. If the independence test turns out positive, check for equivalences $\boldsymbol{U}^{*}$ and $\boldsymbol{W}$ for V, and $\boldsymbol{V}^{*}$ and $\boldsymbol{W}$ for U, with $\boldsymbol{U}^{*}$ and $\boldsymbol{V}^{*}$ sets containing U and V respectively and some nodes adjacent to V and U respectively such that they have the same number of elements as $\boldsymbol{W}$.

3. If an equivalence is found, it is added to the model. Unless there was already an equivalence found of one of both equivalent nodes sets with for the same target, then the other set is added to that equivalence.

4. If both equivalences hold, $\boldsymbol{U}^{*}$, $\boldsymbol{V}^{*}$ and $\boldsymbol{W}$ form a multi-node equivalence.

5. Edge $U - V$ is not removed from the graph.

7.2 Equivalence selection

The second step of the extended PC algorithm alternates selection among equivalent relations, given by algorithm 2, with the original orientation step until no more equivalences or undirected edges can be resolved. For orientation, the original orientation rules can be applied on the graph. If for an information equivalence relation $X - Z$ is considered simpler than relation $Y - Z$, node Y has to be regarded as separated from Z by X, while X is not separated from Z by Y. This d-separation information is used by the orientation rules. For the remaining equivalences, the equivalent node set that are causes of the other equivalent node set are chosen as adjacent to the target.

Algorithm 2 Edge selection among information equivalences by the complexity criterion

To evaluate information equivalences of $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$ with respect to Z, compare the complexities of the following functions f_i with i from 1 to n, only if the nodes $W_{i,j} \in \boldsymbol{W}_i$ are still connected to Z.

1. if all edges connecting the equivalent nodes and the target are oriented:

   (a) if the edges are oriented toward the target, consider the following functions: $Z = f_i(\boldsymbol{W}_i$*, other nodes with edges oriented to target node* Z*)*.

   (b) if the edges are oriented from the target, in order to evaluate $\boldsymbol{W}_i$, count up the complexities of the functions $W_{i,j} = f_i(Z$*, other nodes with incoming edges to* $W_{i,j}$*)*, for all $W_{i,j} \in \boldsymbol{W}_i$.

2. if some edges are not oriented, consider the following functions: $Z = f_i(\boldsymbol{W}_i$*)*.

The complexity of the functions is estimated as explained in section 5.1. If the complexity of the simplest function differs by at least 8 bits (this threshold corresponds to 1 operation) with the complexities of the other functions, the corresponding equivalent nodes can be related with the target, the edges of the other equivalent nodes with the target are removed. For *multi-node equivalences*, the simplest edges should remain in the graph such that all nodes are connected.

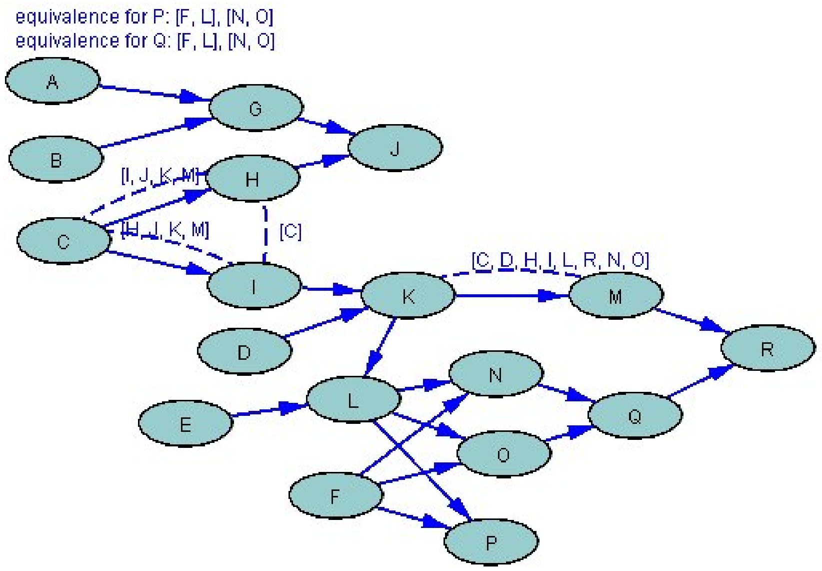

Figure 9: Model learned from random data generated by a set of structural equations.

8 Experiments

We report on learning models based on generated data. 150 data points were generated using the following structural equations:

$$\begin{array}{l|l|l} G = A + B & K = D + 0.4I + err & O = 10 + 1.3L - F \\ H = 4C^2 & L = 1.5 + 0.35K + 0.35E + err & P = L.F + err \\ I = 0.4C^3 & M = 1 + 0.04K^2 & Q = 10N/O + err \\ J = G + 0.4H & N = 10 + L + F & R = M + 0.4Q + err \end{array}$$

Variables A, B, C, D, E and F are randomly chosen between 0 and 10 and err is an error term, reflecting a random disturbance with maximal size 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 system 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).

The extended PC algorithm with default options is applied onto the data. By default, Tetrad uses the Pearson correlation coefficient for continuous, linearly related variables. To handle data with non-linear relations, our independence test is based on the conditional mutual information $I(U : V \mid W)$ (Cover & Thomas, 1991). To calculate the entropies in the definition, the underlying probability distribution of the data is estimated using kernel density estimation (Lemeire, Dirkx, & Verbist, 2007). The augmented Bayesian network learned by the extended PC algorithm, depicted in Fig. 9, provides a correct model. Remark that the algorithm does not check for deterministic relations, only for information equivalences. This strategy was followed to verify that the algorithm works.

We also verified the correctness of the assumptions. Assumptions 2 and 3 hold for the generated data. However, weak transitivity was violated in 31 out of the 1022 cases in which Eq. 7 applies. The cases are characterized by a failure of the independence test in detecting dependencies between 'distant' variables, when the influence of a variable on another variable happens via multiple variables so that the random disturbances dominate the influence. For example, the test classified A and M as independent, while weak transitivity expects dependence (or $A \not\perp\!\!\!\perp M \mid J$, which is also not true), since $A \not\perp\!\!\!\perp J$ and $J \not\perp\!\!\!\perp M$. The test also returned $E \perp\!\!\!\perp K$ and $E \perp\!\!\!\perp K \mid N$, but $E \not\perp\!\!\!\perp N$ and $K \not\perp\!\!\!\perp N$ demand a dependence. Note that the learning algorithm relies on the correctness of the independence test. The contribution of the dependent variables in the structural equations must be sufficiently large in order to correctly learn the model.

9 Conclusions

The theory of causal models and the accompanying learning algorithms are based on the faithfulness property. The existence of a faithful graph is not guaranteed when the intersection condition is violated. This violation can be characterized by an information equivalence, when two sets of variables in some sense have the same information about another variable. Under weak transitivity and two other assumptions, information equivalences can be characterized by basic information equivalences, which are added to an augmented Bayesian network. To retrieve the conditional independencies that follow from information equivalences and the Markov condition from the graph, the d-separation criterion was enlarged. To ensure minimality of the model, the complexity of the relations was introduced to determine adjacency among information equivalent relations. Faithfulness can then be reestablished by enlarging the definition of conditional independency with the requirement of simplicity. The complexity criterion leads to consistent models under the assumption that the complexity of the relations increases for more distant variables (Complexity Increase Assumption). The PC algorithm could easily be extended for learning the augmented models. Experiments with generated data show that the assumptions hold and correct models are learned.

A Appendix: Experimental Data

We have provided the learning module together with the experimental data on the web. They can be accessed at `http://parallel.vub.ac.be`.

References

Cobb, B. R., & Shenoy, P. P. (2004). Inference in hybrid bayesian networks with deterministic variables. In *in P. Lucas (ed.), Proceedings of the Second European Workshop on Probabilistic Graphical Models (PGM-04)*, pp. 57–64.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley & Sons, Inc.

Cowell, R., Dawid, A., Lauritzen, S., & Spiegelhalter, D. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Dechter, R., & Mateescu, R. (2004). Mixtures of deterministic-probabilistic networks and their and/or search space. In *AUAI '04: Proc. of the 20th conf. on Uncertainty in artificial intelligence*, pp. 120–129, Arlington, Virginia, United States. AUAI Press.

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

Grünwald, P., Myung, I., & Pitt, M. (2005). *A Tutorial Introduction to the Minimum Description Length Principle*. MIT Press.

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

Lemeire, J., Dirkx, E., & Verbist, F. (2007). Causal analysis for performance modeling of computer programs. To appear in Scientific Programming, IOS Press.

Li, M., & Vitányi, P. M. B. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag.

Oliver, R. M., & Smith, J. Q. (1990). *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley.

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

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

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

Scheines, R., Spirtes, P., Glymour, C., Meek, C., & Richardson, T. (1996). *TETRAD 3: Tools for Causal Modeling - User's Manual*. http://www.phil.cmu.edu/projects/tetrad/tet3/master.htm.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search* (2nd edition). Springer Verlag.

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