When are Graphical Causal Models not Good Models?

Jan Lemeire JAN.LEMEIRE@VUB.AC.BE

Kris Steenhaut KRIS.STEENHAUT@VUB.AC.BE

ETRO Dept., Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium

Editor:

Abstract

The principle of Kolmogorov Minimal Sufficient Statistic (KMSS) states that a model should capture all regularities of the data. The conditional independencies following from the causal structure of the system are the regularities incorporated in a graphical causal model. We prove that for joint probability distributions, the KMSS is described by the Directed Acyclic Graph (DAG) of the minimal Bayesian network if this results in an incompressible description. We prove that a Bayesian network that is the KMSS is faithful. In that case it can be learned from observations and modularity is the most plausible hypothesis. From modularity follows the ability to predict the effect of interventions. On the other hand, if the minimal Bayesian network is compressible, and thus not the KMSS, the above implications cannot be guaranteed. When the non-minimality of the description is due to the compressibility of an individual Conditional Probability Distribution (CPD), the true causal model is an element of the set of minimal Bayesian networks and modularity is still plausible. Faithfulness cannot be guaranteed though. When the concatenation of the descriptions of the CPDs is compressible, the true causal model is not necessarily an element of the set of minimal Bayesian networks. Also modularity may become implausible. This suggests that either there is a kind of meta-mechanism governing some of the mechanisms or either a single mechanism responsible for setting the state of multiple variables.

Keywords: Causal Models, Causal Inference, Kolmogorov Complexity, Meaningful Information, Faithfulness

1. Introduction

This paper analyzes the theory of graphical causal models and causal inference with the concept of Kolmogorov Minimal Sufficient Statistic (KMSS). The theory of graphical causal models is based on the conditional independencies that follow from a system's causal structure. The Directed Acyclic Graph (DAG) of a Bayesian network is considered as a representation of these conditional independencies. Algorithms for causal inference are based on these conditional independencies. The causal interpretation is based on modularity and manipulability; one part of the model can be manipulated without affecting the rest.

KMSS is an extension of the concept of Kolmogorov Complexity (Vitányi, 2002; Gács et al., 2001). The central idea is that modeling has to be equated with finding the patterns or regularities in observational data and constructing the minimal model which is able to explain the regularities. Regularities make up the *meaningful information* of data; they enable prediction of future data. A regularity is identified by its ability to compress the data, i.e. to describe the data using fewer symbols than the number of symbols needed to describe the data literally. The concept of KMSS is based on the assumption that patterns or regularities in observations are - most likely - not coincidences, but give us valuable information about the system under study. The concept of Kolmogorov complexity has given rise to different methods for inductive inference, such as Minimum Message Length (Wallace and Dowe, 1968) and Minimum Description Length (Rissanen, 1978). These methods are used for selecting the best model from a given set of models, the model class. The choice of model class, however, determines the regularities under consideration.

In the context of causal inference, the dependencies among the variables are the regularities that allow compression of the data and the conditional independencies are the regularities that determine the model's complexity. For causal inference, the set of Bayesian networks is used as a model class. The DAG of a Bayesian network gives a minimal description of the conditional independencies following from a causal structure. A system can, however, contain other regularities. Then, the assumptions and implications of causal model theory, such as faithfulness, modularity and the correctness of causal inference, may become invalid. For example, it can give rise to other independencies so that the DAG becomes unfaithful; the DAG does not represent all independencies.

In Section 2, we will introduce the concept of KMSS. In Section 3, we will give a survey of graphical causal model theory and the learning algorithms. Section 4 lists related work. In Section 5 we apply the principle of KMMS to inductive inference and show that a Bayesian network captures dependencies between variables. Section 6 establishes the link between minimality of Bayesian networks, compressibility and faithfulness. In Section 7 we will argue that causal inference is plausible if the minimal Bayesian network is the KMSS. Section 8 discusses various cases in which the minimal Bayesian network does not provide the minimal description.

2. Meaningful Information

Kolmogorov Complexity provides an objective measure of simplicity so that Occam's razor can be applied. The *Kolmogorov Complexity* of a string x is defined to be the length of the shortest computer program that prints the string and then halts (Li and Vitányi, 1997):

$$K(x) = \min_{p:\mathcal{U}(p)=x} l(p) \tag{1}$$

with $\mathcal{U}$ a universal computer and $l(p)$ the size in bits of program p. Patterns in the string allow for its compression, i.e. to describe the data using fewer symbols than the number of symbols to describe the data literally. The string "00010001000100010001000100100010001000100010001" can be described shorter by program `REPEAT 11 TIMES "0001"`. But not all bits of this program can be regarded as containing *meaningful information*. We consider meaningful information as the properties of the string that allow for its compression (Vitányi, 2002). Such properties are called patterns or *regularities*. The regularity of the example string is the repetition. The number of repetitions (`"11"`) or the substring `"0001"` is random information. A random string, which is incompressible has no meaningful information at all.

For inductive inference, we will look for a minimal description in 2 parts, one containing the regularities of the data, which we put in the model, and one part containing the remaining random noise. Such a description is called a *two-part code*. This results in a generic approach for inductive inference, called *Minimum Description Length* (MDL), according to which we have to pick out the model M_{mdl} from model class $\mathcal{M}$ where M_{mdl} is the model which minimizes the sum of the

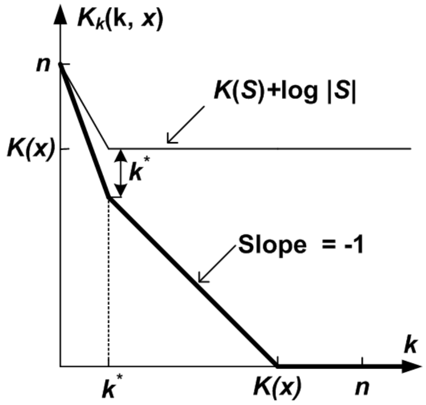

Figure 1: Kolmogorov structure function for n-bit string x, k^* is the KMSS of x.

description length of M and of the data D encoded with the help of M (Grünwald, 1998):

$$M_{mdl} = arg\ min_{M \in \mathcal{M}} \{L(M) + L(D \mid M)\} \tag{2}$$

with $L(.)$ the description length.

The MDL approach relies on the a priori chosen model class. It does not tell us how to make sure the models capture all regularities of the data. The KMSS provides a formal separation of meaningful and meaningless information. We limit the introduction to KMSS to models that can be related to a finite set of objects, called the *model set*. In the context of learning, we are interested in a model set S that contains x and the objects that share x's regularities. All elements of a set S can be enumerated with a binary index of length $\log_2 |S|$ with $|S|$ the size of set S. We therefore say that x is *typical* for S if

$$K(x \mid p_S) \geq \log_2 |S| - \beta \tag{3}$$

with p_S the shortest program that describes S and β an agreed upon constant. Given set S, x cannot be described shorter than by the set's index. Atypical elements have regularities that are not shared by most of the set's members and can therefore be described by a shorter description. Note that most elements of a set are typical, since, by counting arguments, only a small portion of it can be described shorter than $\log_2 |S|$.

The construction of S can be understood with the *Kolmogorov structure function* K_k. $K_k(k, x)$ of x is defined as the $\log_2$-size of the smallest set including x which can be described with no more than k bits (Cover and Thomas, 1991):

$$K_k(k, x) = \min_{\substack{p: l(p) \leq k \\ \mathcal{U}(p) = S \\ x \in S}} \log_2 |S| \tag{4}$$

A typical graph of the structure function is illustrated in Figure 1. By taking $k = 0$, the only set that can be described is the entire set $\{0, 1\}^n$ containing 2^n elements, so that the corresponding log set size is n. By increasing k, the model can take advantage of the regularities of x in such way that each bit reduces the set's size more than halving it. The slope of the curve is smaller than -1.

When k reaches k^*, all regularities are exploited. There are no more patterns in the data that allow for further compression. From then on each additional bit of k reduces the set by half. We proceed along the line of slope -1 until $k = K(x)$ and the smallest set that can be described is the singleton $\{x\}$. The curve $K(S) + \log_2 |S|$ is also shown on the graph. It represents the descriptive complexity of x by using the two-part code. With $k = k^*$ it reaches its minimum and equals to $K(x)$. When $k < k^*$, S is too general and is not a typical set for x. Only for no smaller values than k^* is x typical for S. For random strings the curve starts at $\log_2 |S| = n$ for k=0 and drops with a slope of -1 until reaching the x-axis at $k = n$. Each bit reveals one of the bits of x, and halves the model set.

The *Kolmogorov Minimal Sufficient Statistic* (KMSS) of x is defined as the program p^* which describes the smallest set S^* such that the two-stage description of x is as good as the minimal single-stage description of x (Gács et al., 2001):

$$p^* = arg\ min_p \{l(p)\ \ | \mathcal{U}(p) = S^*,\ x \in S^*,\ K(S^*) + \log_2 |S^*| \leq K(x)\} \tag{5}$$

The descriptive complexity of S^* is then k^*. Program p^* minimally describes the meaningful information present in x and nothing else. The definition ensures that x is a typical element of S^*.

The concepts of Kolmogorov complexity are not directly applicable for inductive inference since there exists no algorithm that computes the shortest program for a string. They are therefore used for giving preference within a given set of models. It is still up to the expert to determine the regularities that have to be considered and construct an adequate model class. We will show that the incompleteness of the model class affects the validity of current causal inference algorithms.

3. Graphical Causal Models

This chapter will introduce graphical causal models, following the theory built up by Pearl et al., and the accompanying learning algorithms (Pearl, 2000; Spirtes et al., 1993).

3.1 Representation of Causal Relations

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. All variables that influence the outcome of the process are called *causes* of the outcome variable. An *indirect cause* produces the state of the effect indirectly, through another variable. If there is no intermediate variable among the known variables, the cause is said to be a *direct cause*.

Each process represents a physical mechanism. In it most general form it can be described by a conditional probability distribution (CPD) $P(X \mid parents(X))$, where $parents(X)$ is the set of parent nodes of X in the graph and constitute the direct causes of the variable. A causal model consists of a DAG over all variables and a CPD for each variable. The combination of the CPDs results in a joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i)) \tag{6}$$

For a discrete variable, the CPD is encoded by means of a tabular representation: for each possible assignment of values to the parents of X_i, we need to specify a distribution over the values that X_i can take. This is called a conditional probability table. For continuous variables, one often relies on prior knowledge or assumptions about the structure of the distribution. If one assumes linearly-related variables, the CPDs can be described by the following *structural equations*:

$$P(X_i \mid parents(X_i)) = \sum_{X_j \in parents(X_i)} a_{i,j}.X_j + U_i + c_i \tag{7}$$

where U_i represent the stochastic variations which cannot be explained by the model and c_i a constant term. One often assumes that U_i is normally distributed.

3.2 Modularity and the Effect of Changes to the System

A causal model represents a collection of processes that could account for the generation of the observed data. Each process is a stable and autonomous physical mechanism. It is then conceivable to change one such relationship without changing the others. This *modularity* permits one to predict the effect of external interventions or local reconfigurations of the mechanisms (Pearl, 2000). An *intervention* is defined as an atomic operation that fixates a variable to a given state and eliminates the corresponding factor (CPD) from the factorization (Eq. 6) (Pearl, 2000). Applied on a causal graph, an intervention on variable X sets the value of X and breaks all of the edges in the graph directed into X and preserves all other edges in the graph, including all edges directed out of X. This is called the Manipulation Theorem by Spirtes et al. (1993, p. 51). Intervening on a variable only affects its effects. Causes have to be regarded as they were levers which can be used to manipulate their effects.

This approach does not directly define causality, but defines the implications of having a thorough knowledge of the mechanisms that make up a system. Manipulability puts a constraint of independentness on the mechanisms. The accuracy of the mutilated model relies on autonomy or modularity; a mechanism can be replaced by another without affecting the rest of the system. It is defined by Hausman and Woodward (1999, p. 545) as follows. They relate each CPD to a structural equation (Eq. 7).

Definition 1 *(Modularity) For all subsets **Z** of the variable set **V**, there is some non-empty range **R** of values of members of **Z** such that if one intervenes and sets the value of the members of **Z** within **R**, then all equations except those with a member of **Z** as a dependent variable (if there is one) remains invariant.*

3.3 Representation of Independencies

The key for causal inference are the conditional independencies entailed by the system's causal structure. They are based on the property of Markov chains and v-structures. If X is affected by Y and Z, then we do not expect that X is independent of Y conditional on Z, except if Y affects X *via* Z. This is represented by a Markov chain. Random variables X, Z, Y are said to form a *Markov chain* in that order, denoted by $X \rightarrow Z \rightarrow Y$, if the joint probability mass function can be written as

$$P(X, Z, Y) = P(X).P(Z \mid X).P(Y \mid Z) \tag{8}$$

which is equivalent to the conditional independence of X and Y given Z. *Conditional independence* of X and Y given Z, written as $X \perp\!\!\!\perp Y \mid Z$, is defined as

$$P(X, Y \mid Z) = P(X \mid Z).P(Y \mid Z) \tag{9}$$

The conditional independence expresses that learning the value of X does not provide additional information about Y once the state of Z is known. We say that Z 'screens off' X from Y. Once the state of Z is observed, the state of Y does not longer depend on that of X. For a *v-structure* on the other hand, for example $X \rightarrow Z \leftarrow Y$, X and Y are independent, but become dependent when conditioned on Z.

For a causal model, the *Causal Markov Condition* gives us the independencies that follow from the causal structure: each variable is probabilistically independent of its non-effects conditional on its direct causes. This condition is defined by Spirtes et al. (1993) as follows:

Definition 2 *(Causal Markov Condition) Let G be a causal graph with vertex set **V** and P be a probability distribution over the vertices in **V** generated by the causal structure represented by G. G and P satisfy the Causal Markov Condition if and only if for every W in **V**, W is independent of $\mathbf{V} \setminus \boldsymbol{Descendants}(W) \setminus \boldsymbol{Parents}(W)$ given $\boldsymbol{Parents}(W)$.*

These independencies follow from the causal structure. They are irrespective of the nature of the mechanisms, of the exact parameterization of the conditional probability distributions $P(X_i \mid parents(X_i))$. Pearl and Verma constructed a graphical criterion, called d-separation, for retrieving, from the causal graph, all independencies following from the Causal Markov Condition.

A graph is called *faithful* to a distribution if all conditional independencies of the distribution correspond to a d-separation in the graph and vice versa. In other words, faithfulness means that if a graph represents a causal structure, all conditional independencies follow from the system's causal structure.

3.4 Correspondence with Bayesian networks

Graphical causal models provide a probabilistic account of causality (Spohn, 2001). This resulted in a close correspondence with Bayesian networks. In contrast to causal models, Bayesian networks are mainly concerned with offering a dense and manageable representation of joint distributions. A joint distribution over n variables can be *factorized* relative to a variable ordering $(X_1, \ldots, X_n)$:

$$P(X_1, \ldots, X_n) = \prod_i^n P(X_i \mid X_1, \ldots, X_{i-1}) \tag{10}$$

Variable X_j can be removed from the conditioning set of variable X_i if it becomes conditionally independent from X_i by conditioning on the rest of the set:

$$P(X_i \mid X_1 \ldots X_{i-1}) = P(X_i \mid X_1 \ldots X_{j-1}, X_{j+1} \ldots X_{i-1}). \tag{11}$$

Such conditional independencies reduce the complexity of the factors in the factorization. The conditioning sets of the factors can be described by a Directed Acyclic Graph (DAG), in which each node represents a variable and has incoming edges from all variables of the conditioning set of its factor. The joint distribution is then described by the DAG and the conditional probability distributions (CPDs) of the variables conditional on their parents. A *Bayesian network* is a factorization that is edge-minimal, in the sense that no edge can be deleted without destroying the correctness of the factorization.

Although edge-minimality of a Bayesian network, the graph depends on the chosen variable ordering. Some orderings lead to the same networks, while others result in different topologies. Take

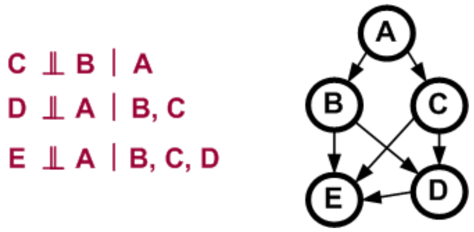

Figure 2: Factorization based on variable ordering (A, B, C, D, E) and reduction by three independencies.

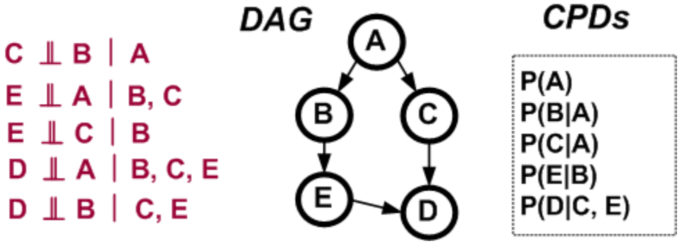

Figure 3: Bayesian network based on variable ordering (A, B, C, E, D) and five independencies.

5 stochastic variables A, B, C, D and E. Fig. 2 shows the graph that was constructed by simplifying the factorization based on variable ordering (A, B, C, D, E) by the three given conditional independencies. However, the Bayesian network, describing the same distribution, but based on ordering (A, B, C, E, D), depicted in Fig. 3 contains 2 edges less because of 5 useful independencies. Both networks represent the probabilities just as well, except that the first one is more complex. We call the *minimal factorization* as the factorization which has the least total number of variables in the conditioning sets. The corresponding Bayesian network is called the *minimal Bayesian network* of a probability distribution.

Analogue to the Causal Markov Condition, the Markov Condition gives the conditional independencies that follow from the structure of a Bayesian network: each variable is independent from all its non-descendants by conditioning on its parents in the graph. The equivalence of the Markov Condition follows and factorizability can be proven (Hausman and Woodward, 1999, p. 532). This ensures the correspondence: causal models are also Bayesian networks. The difference lies in the causal component, causal models attribute a causal interpretation to the edges of the graph and are therefore called *causally interpreted Bayesian networks*.

3.5 Causal Inference

The goal of causal inference is to learn the causal structure of a system based on observational data. Causal structure learning algorithms fall apart in two categories: scoring-based and constraint-based algorithms.

Scoring-based algorithms are based on an optimized search through the set of all possible models, which tries to find the minimal model that best describes the data. Each model is given a score that is a trade-off between model complexity and goodness-of-fit. Different scoring criteria have been applied in these algorithms, such as a Bayesian scoring method (Cooper and Herskovits, 1992)(Heckerman et al., 1994), an entropy based method (Herskovits, 1991) and one based on the Minimum Description Length (MDL) (Suzuki, 1996).

Constraint-based learning algorithms rely on the conditional independencies detected that follow from the system's causal structure. It is a kind of evidence-based construction, the decisions to include an edge and on the edge's orientation is based on the presence or absence of certain independencies. The algorithms assume the existence of a faithful graph, i.e. that all independencies follow from the causal structure. They also assume that the correct model is the minimal model. Minimality, faithfulness and the Causal Markov Condition give the 3 assumptions that ensure correct learning (Spirtes et al., 1993). Finally it must be noted that some algorithms, such as the PC algorithm, also require *causal sufficiency*, i.e. that all common causes should be known: variables that are the direct cause of at least two variables. More sophisticated learning algorithms exist that are capable of detecting latent common causes. For now we will not take the presence of latent variables into consideration and discuss the consequences of this in Section 8, case 5.

4. Related Work

The causal interpretation of a Bayesian network and the validity of faithfulness are often criticized (Freedman and Humphreys, 1999; Cartwright, 2001; Williamson, 2005; Hausman and Woodward, 1999). This paper would like to contribute to the discussion by giving an additional viewpoint through the concept of the KMSS. Some of the examples on which criticism on the possibility of causal inference is based will be discussed in Section 8. Hausman and Woodward (1999) on the other hand are strong defenders of linking the causal interpretation of models to modularity. They defend the equivalence of modularity and the Causal Markov Condition (Hausman and Woodward, 1999, p. 554). We will contribute to the discussion by motivating why and when modularity is a valid assumption, and showing the limitations of assuming faithfulness.

Pearl and others use *stability* as the main motivation for the faithfulness of causal models (Pearl, 2000, p. 48). Consider the model of Fig. 3. In general, one expects A to depend upon D. A and D are independent only if the stochastic parameterization is such that the influences via paths $A \rightarrow B \rightarrow E \rightarrow D$ and $A \rightarrow C \rightarrow D$ cancel out exactly. This system is called unstable because a small change in the parameterization results in a dependency. The unhappy balancing act is a measure zero event, the probability of such a coincidence can therefore be regarded as zero. Pearl claims that there exists at least 1 distribution faithful with a Bayesian network (Pearl, 2000, p. 18), while we will show that all typical - hence the majority of - distributions compliant with a DAG are faithful.

Milan Studeny was one of the first to point out that the 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 (referring to regularities) for the presence of conditional independencies not coming from the system's causal structure.

5. Minimal Description of Distributions

In this section we will draw the connection between Bayesian networks and the KMSS of probability distributions. Following KMSS, we have to model the regularities of a joint probability distribution. The type of regularity we have to consider here is a dependency between variables; knowing one variable gives information about the state of another variable. The knowledge about the state of a single stochastic variable is captured by a probability distribution over it. Dependency information is captured by the joint probability distribution defined over the variables of interest.

From the theory of Bayesian networks (Section 3.4), we know that a joint distribution can be described shorter by a factorization relative to a certain variable ordering that is reduced by conditional independencies of the form of Eq. 11. The minimal factorization leads to $P(X_1, \ldots, X_n) = \prod CPD_i$, with CPD_i the CPD of variable X_i. The descriptive size of the CPDs is determined by the number of variables in the conditioning sets. The total number of conditioning variables thus defines the shortest factorization. A two-part description of a joint distribution is then:

$$descr(P(X_1 \ldots X_n)) = \{parents(X_1), \ldots, parents(X_n)\} + \{CPD_1, \ldots, CPD_n\} \qquad (12)$$

Note that the parents' lists can be described very compact and correspond to the description of a DAG. The description of a random DAG defined over n nodes needs $n/2.(n-1)$ bits.

The following theorems show that the first part offers the minimal model if the descriptions of the CPDs are incompressible.

Lemma 3 *The parents' lists, $\{parents(X_1) \ldots parents(X_n)\}$, in the two-part code given by Eq. 12 contain meaningful information of a probability distribution if the description of the Bayesian network is shorter than the description of the joint distribution.*

Proof If the description of the Bayesian network is shorter than that of the joint distribution, the reduction in descriptive size by the elimination of variables from the conditioning sets of the factors outweigh the description of the parents' lists. This reduction is due to conditional independencies, which are exactly the regularities that are described by the parents' lists. ■

In the case of discrete variables, the size of the conditional probability table is reduced by conditional independencies. In the case of continuous variables, the CPDs contain fewer free parameters. In the linear case for example, each independency which eliminates a variable of a CPD's conditioning set eliminates a coefficient of a structural equation (Eq. 7).

Lemma 4 *If the two-part code description of a probability distribution, given by Eq. 12, results in an incompressible string, the first part is the Kolmogorov minimal sufficient statistic.*

Proof If a more compact description of the distribution would exist, the two-part description would contain redundant bits. Lemma 3 showed that the first part contains meaningful information. By hypothesizing incompressibility of the whole description, the second part of the description, consisting of the CPDs, is also incompressible. The CPDs thus do not contain meaningful information. Hence, the first part, described minimally, is the KMSS. ■

The graph thus minimally describes the dependencies among the variables. The conditional independencies determine the model's complexity.

6. Minimality of Bayesian Networks

The following two theorems show that the Bayesian network corresponding to the minimal factorization is the KMSS and faithful if its DAG and CPDs are random and incompressible.

Theorem 5 *If a faithful Bayesian network exists for a distribution, it is the minimal factorization.*

Proof Recall that the absence of an edge between two variables X and Y in a Bayesian network implies that there exists a set of variables S not containing X and Y that makes X and Y conditionally independent: $X \perp\!\!\!\perp Y \mid S$. In case of faithfulness, the presence of an edge forbids the existence of such a set. Let A be a graph that has fewer edges then the faithful graph B. It follows that B contains an edge between two variables X and Y that A does not contain. The absence of the edge in A implies that X and Y become independent by conditioning on some set of the other variables. But this contradicts with the faithfulness of B which implies that X and Y cannot become independent. ■

The DAG of a Bayesian network corresponds to a set of conditional independencies. Intuitively we would expect that two variables are dependent if they are not d-separated. This is, however, not true for all probability distributions that are compatible with the DAG. The next theorem proves that two variables that are not d-separated are dependent unless there is a constraint between the probabilities. Consider the example parameterization of $P(D \mid E, C)$ in the model of Fig. 3 given in Table 1. For these probabilities variables D and E become independent, assuming that $P(C = 0 \mid E) = P(C = 1 \mid E) = 0.5$. This independency is, however, not true in general, only for a special parameterization of $P(D \mid C, E)$ which is consistent with Eq. 16 of the forthcoming proof.

E	C	$P(D \mid C, E)$
0	0	0.4
0	1	0.3
1	0	0.2
1	1	0.5

Table 1: Conditional Probability Table of $P(D \mid C, E)$.

Theorem 6 *A Bayesian network for which the concatenation of the descriptions of the conditional probability distributions (CPDs) is incompressible, is faithful.*

Proof Recall that a Bayesian network is a factorization that is edge-minimal. This means that for each parent $pa_{i,j}$ of variable X_i:

$$P(X_i \mid pa_{i,1}, \ldots pa_{i,j}, \ldots pa_{i,k}) \neq P(X_i \mid pa_{i,1}, \ldots pa_{i,j-1}, pa_{i,j+1}, \ldots pa_{i,k}) \qquad (13)$$

Variables cannot be eliminated from the factors of the factorization. The proof will show that any two variables that are not d-separated are dependent. Unless the probabilities of the CPDs are ‘related’, in the sense that some probabilities can be calculated from others and the set of CPDs is compressible. We derive the relations for discrete variables. For continuous variables, the analysis results in relations among the free parameters of the CPDs.

We have to consider the following possibilities. The two variables can be adjacent (a), related by a Markov chain (b) [1], a v-structure (c), a combination of both or connected by multiple paths (d).

First we prove that a variable marginally depends on each of its adjacent variables (a). Consider adjacent nodes D and E of the Bayesian network of Fig. 3. We will demonstrate that $P(D \mid E) = P(D)$ results in a regularity. We expand the first term with all other parents of D:

$$P(D \mid E) = \sum_{c \in C_{dom}} P(D \mid E, c).P(c \mid E) \tag{14}$$

C is also a parent of D, thus, by Eq. 13, there are at least two values of C_{dom} for which $P(D \mid E, c) \neq P(D \mid E)$ [2]. Take c_1 and c_2 being such values for which

$$P(D \mid E, c_1) \neq P(D \mid E, c_2). \tag{15}$$

There are also at least 2 such values of E_{dom}, take e_1 and e_2. Eq. 14 should hold for all values of E and equal to $P(D)$ to get an independency. This results in the following relation among the probabilities:

$$\begin{aligned} &P(D \mid e_1, c_1).P(c_1 \mid e_1) + P(D \mid e_1, c_2).P(c_2 \mid e_1) \\ &= P(D \mid e_2, c_1).P(c_1 \mid e_1) + P(D \mid e_2, c_2).P(c_2 \mid e_1) \end{aligned} \tag{16}$$

Note that the equation cannot be algebraically simplified: the conditional probabilities are not equal to $P(D)$ (Eq. 13) nor to each other (Eq. 15). The proof can easily be generalized for variables having more parents.

Next, by the same arguments it can be proved that variables connected by a Markov chain are by default dependent (b). Take $A \rightarrow B \rightarrow E$ in Fig. 3, independence of A and E requires that

$$P(E \mid a) = \sum_{b \in B_{dom}} P(E \mid b).P(b \mid a) = P(E) \quad \forall a \in A. \tag{17}$$

and this would also result in a regularity among the CPDs.

In a v-structure, both causes are dependent when conditioned on their common effect (c), for $C \rightarrow D \leftarrow E$, $P(D \mid C, E) \neq P(D \mid E)$ is true by Eq. 13. Finally, if there are multiple unblocked paths connecting two variables, then independence of both variables implies a regularity as well (d). Take A and D in Fig. 3:

$$P(D \mid A) = \sum_{b \in B_{dom}} \sum_{c \in C_{dom}} \sum_{e \in E_{dom}} P(D \mid c, e).P(c \mid A).P(e \mid b).P(b \mid A).$$

Note that $P(c, e \mid A) = P(c \mid A).P(e \mid A)$ follows from the independence of C and E given A. All factors from the equation satisfy Eq. 13, so that, again, the equation only would equal to $P(D)$ if there is a relation among the probabilities. ■

[1] Recall that a Markov chain is a path not containing v-structures.

[2] $P(D \mid E)$ is a weighted average of $P(D \mid E, C)$. If one probability $P(D \mid E, c_1)$ is different than this average, let's say higher, than there must be at least one value lower than the average, thus different.

From the theorem it follows that a Bayesian network with random CPDs is the minimal factorization. Bayesian networks not based on a minimal factorization, such as the one of Fig. 2, are compressible, namely by the regularities among the CPDs that follow from the independencies not represented by the graph. Pearl hypothesizes that there is no bounded set of conditions that would ensure the existence of a faithful graph (Pearl, 1988, p. 131). Indeed, as shown by the theorem, every dependence can be turned into an independence by a balanced parameterization of some CPDs.

It must be noted that if there exists a faithful Bayesian network, it is not necessarily unique. Multiple faithful models can exist for a distribution. 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 skeleton and v-structures. They only differ in the orientation of some edges (Pearl, 2000). This set can be represented by a partially-directed acyclic graph in which some of the edges are not oriented. The corresponding factorizations have the same number of conditioning variables and thus all models of a Markov-equivalence class have the same complexity.

7. The Minimal Bayesian Network is the KMSS

In this section we will discuss the case in which there is exactly one minimal Bayesian network which is also the minimal description. This means that there are no other regularities and no other independencies than the conditional independencies the model represents. The DAG is then the KMSS and minimally represents all regularities. It is also faithful.

A Bayesian network decomposes the description of a joint distribution into a list of CPDs. We assume in this section that the description is unique and minimal. This means that the minimal description of the system is a concatenation of descriptions, namely the description of the individual CPDs. In other words, we have found a unique and minimal decomposition of the model. This brings us to modularity and manipulability. We have discovered that the minimal description is a concatenation of unrelated components. The CPDs are independent; the concatenation of their descriptions cannot be compressed. Then, among all possible explanations, the simplest is that *each CPD corresponds to an independent part of reality*. Thus, following Occam's Razor, modularity is the most likely hypothesis about the system under study. The correctness of Occam's razor cannot be proven, the principle must be interpreted as the most effective *strategy* for deciding among competing explanations (Grünwald, 1998). Modularity of the minimal Bayesian network must be regarded as the top-ranked hypothesis, which can be verified with background knowledge or experiments with interventions. Thus, the three conditions for causal inference are valid (Section 3.5): minimality and faithfulness are fulfilled, and the Causal Markov Condition follows from modularity. We argue that description minimality can be linked to causality through modularity.

Occam's razor is contradicted when the real system is more complex than suggested by the complexity of the observations. Take the case of the neutral contraceptive pills, due to Hesslow (1976). Two variables, $Pill$ and $Thrombosis$, are directly related, but also indirectly related through a third variable, as shown in Fig. 4(a). The strengths along the two paths from $Pill$ to $Thrombosis$ exactly balance, so that there is no net correlation between $Pill$ and $Thrombosis$. The minimal Bayesian network of Fig. 4(b) is faithful, incompressible and simpler than the true model. From observations alone, one cannot find indications for the more complex true model.

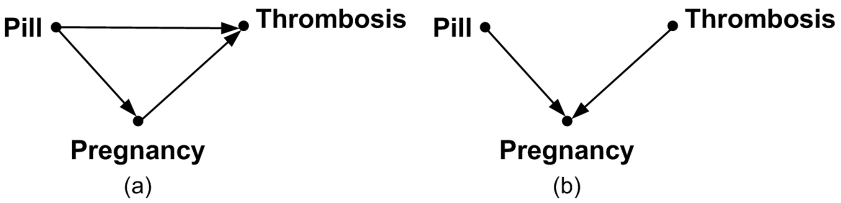

Figure 4: Hesslow's example: true causal model with $Pill \perp\!\!\!\perp Thrombosis$ (a) and minimal Bayesian network (b).

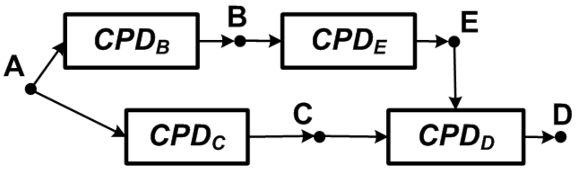

Figure 5: Decomposition of the system represented by the causal model of Fig. 3 into independent components.

The CPD of a variable is also called the variable's *causal Markov kernel*. Note that by representing a causal model with a graph, the representation suggests that the edges - instead of the CPDs - are the basic components. This is however not true. A graphical model can therefore be misleading, a better representation is shown in Fig. 5. It represents the same system as the causal model of Fig. 3, but emphasizes that CPDs are the basic components.

Decomposition and thus also causality matches with a *reductionist* view, according to which the world can be studied in parts. Indeed, if the system cannot be decomposed, if there are no conditional independencies that simplify a factorization, then the DAG does not contain meaningful information (Lemma 3). We end up with a Holist system in which everything depends on everything.

It must be noted that the assumption of a unique minimal Bayesian network is not essential. As discussed in the previous section, if the minimal Bayesian network is not unique, the Markov-equivalence class indicates exactly which parts are undecided (the orientation of some edges). So, we know exactly for which parts of the model we do not have enough information to decide upon the decomposition.

8. Cases in which the Minimal Bayesian Network is not the KMSS

To study the validity of faithfulness and the modularity property, we will in this section not assume incompressibility of the minimal Bayesian networks. They are denoted with $\boldsymbol{BN_{min}}$. Instead, we will study a wide variety of cases, appearing throughout literature, in which regularities appear that are not described by a Bayesian network. We will analyze the properties of the True Causal Model (TCM) and those of the $\boldsymbol{BN_{min}}$.

Table 2 gives an overview of the answers for the next questions, which will be discussed in the following.

- Is the TCM compressible? If so, is the compressibility due to the compressibility of the description of a single CPD or the compressibility of the concatenation of the descriptions of multiple CPDs?
- Is the compressibility of the minimal Bayesian networks due to the compressibility of the description of a single CPD or the compressibility of the concatenation of the descriptions of multiple CPDs?
- Is the TCM present in $\boldsymbol{BN_{min}}$? The answer to this and the next question determines the feasibility of causal inference.
- Is there a unique $\boldsymbol{BN_{min}}$?
- Is the true causal model faithful to the system?
- Are the minimal Bayesian networks faithful to the system?
- Does modularity holds for the true causal model?

	Compress. TCM	Compress. $\boldsymbol{BN_{min}}$	TCM $\in$ $\boldsymbol{BN_{min}}$	Unique $\boldsymbol{BN_{min}}$	TCM faithful	$\boldsymbol{BN_{min}}$ faithful	Modular TCM
1. Local structure	single	single	Yes	Yes	Yes	Yes	Yes
2. Pseudo-indep.	single	single	Yes	No	No	No	Yes
3. Deterministic	single	single	Yes	No	No	No	Yes
4. Unfaithful model	concat.	concat.	Yes	Yes	No	No	No
5. Latent variables	No*	concat.	No*	No	Yes*	No	Yes*
6. Particle decay	concat.	concat.	No	Yes*	No	Yes*	No
7. OO-nets	concat.	concat.	Yes	Yes	Yes	Yes	Yes

Table 2: Answers to questions for the different case studies. An asterisk (*) indicates that Bayesian networks with latent variables are considered.

8.1 Compressibility of a single CPD

First we consider cases in which the description of an individual CPD is compressible. Faithfulness and the uniqueness of the minimal Bayesian network are not guaranteed, but the cases show that the modularity assumption still holds. The CPDs are independent.

Case 1. When individual CPDs can be compressed, we call this type of regularity *local structure* (Friedman and Goldszmidt, 1996). For discrete variables, the conditional probability tables are exponential in the number of parents of a variable X: for each possible assignment of values to the parents of X, we need to specify a distribution over the values X can take. When regularities among the probabilities appear these tables can be described shorter. For example by decision trees. The regularities to construct the tree are called context-specific independencies (Boutilier et al., 1996). On top of the independencies following from the causal structure the system exhibits additional regularities. But the model remains faithful and the decomposition is correct.

A specific type of local structure is the decomposition of a CPD into independent components. In general, a CPD describes the mechanism by which all direct causes together produce the state of a single variable. Various authors report on independent cause-effect relations. They study representations in which the causal influences of the direct causes of a variable are independent, for example by a factorized representation of a CPD (Madsen and D'Ambrosio, 2000). Hausman and Woodward (1999, p.547) call them disjunctive causes.

Case 2. Variables in *pseudo-independent models* are pairwise independent but collectively dependent (Xiang et al., 1996). For example, consider a binary variable X_3 that is determined by two other binary variables X_1 and X_2 by an *exclusive or* relation: $X_3 = X_1 \texttt{ EXOR } X_2$. This system can be represented by causal model $X_1 \rightarrow X_3 \leftarrow X_2$. Because of the pairwise independencies $X_3 \perp\!\!\!\perp X_1$ and $X_3 \perp\!\!\!\perp X_2$, the model is not faithful. There are three minimal Bayesian networks: besides the correct $X_1 \rightarrow X_3 \leftarrow X_2$, also $X_1 \rightarrow X_2 \leftarrow X_3$ and $X_2 \rightarrow X_1 \leftarrow X_3$. The CPD $P(X_3 \mid X_1, X_2)$ exhibits a strict regularity. Yet, pseudo-independent models fit in the reductionist approach of causal models. The only problem is that the conditional independencies do not provide enough information to conclude about the causal connections.

Case 3. Deterministic or functional relations among variables result in CPDs with a very specific form. Distributions with deterministic relations cannot be represented by a faithful graph (Spirtes et al., 1993). Consider the system $X \rightarrow Y \rightarrow Z$ in which Y is a function of X: $Y = f(X)$. From the model (Markov chain) it follows that $X \perp\!\!\!\perp Z \mid Y$. By the functional relation variable X got all information about Y, which implies $Y \perp\!\!\!\perp Z \mid X$. Both independencies imply a violation of the *intersection condition*, one of the conditions that Pearl imposes on a distribution in the elaboration of causal theory and its algorithms (Pearl, 1988). In (Lemeire, 2007) we call X and Y *information equivalent* with respect to Z, both variables have in some sense the same information of Z. Then, the set of minimal Bayesian networks contains graphs that connect X with Z and graphs that connect Y with Z. From the information about the conditional independencies alone we cannot decide upon which variable, X or Y, directly relates to Z. The solution we proposed for causal inference is to connect the variables that have the simplest relation (Lemeire, 2007). We defined an augmented causal model which also incorporates information of deterministic relations. Then, faithfulness of the model could be reestablished by characterizing the influence of deterministic relations on conditional independencies. We limit the conditional independencies that are graphically represented by the simplicity condition and we enlarged the d-separation criterion to capture the conditional independencies following from deterministic relations. This approach fits into the KMMS approach. We have added the regularity of a functional relation to an augmented model.

8.2 Compressibility of a set of CPDs

When the description of some CPDs taken together can be compressed, the CPDs are in some way related.

Case 4. The most-known example of unfaithfulness is when in the model of Fig. 6, A and D appear to be independent (Spirtes et al., 1993). This happens when the influences along the paths $A \rightarrow B \rightarrow D$ and $A \rightarrow C \rightarrow D$ exactly balance, so that they cancel each other out and the net effect results in an independence. For continuous variables this happens when an exact correspondence of the free parameters is fulfilled. The model is not faithful. This balancing act can give an indication of a global mechanism or *meta-mechanism*, such as evolution (Korb and Nyberg,

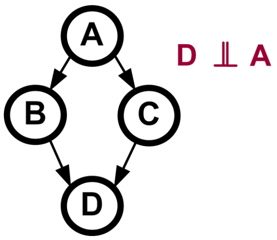

Figure 6: Causal model in which A is independent from D.

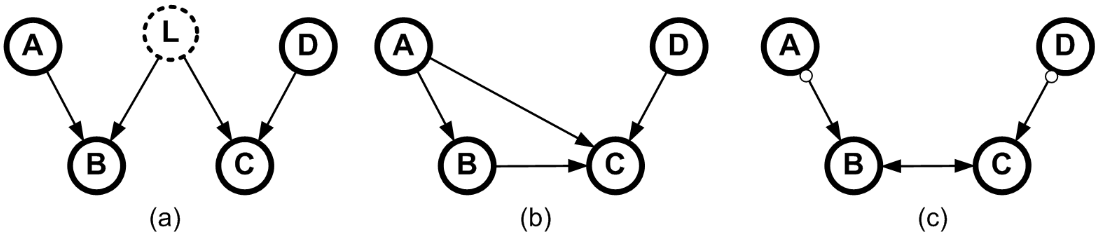

Figure 7: Learning the model of a system with a latent variable L (a). One of the minimal Bayesian networks (b) and the learned Partially-oriented Acyclic Graph (c)

2006), controlling the mechanisms such that the parameters are calibrated until they neutralize. Modularity and autonomy of the CPDs is violated.

Case 5. Causal sufficiency, the knowledge of all common causes, is an important property for correct causal learning. Take the system depicted in Fig. 7(a) in which L is an unknown variable which is the cause of B and C. This gives rise to multiple minimal Bayesian networks, none of which models the system correctly. One of them is depicted in Fig. 7(b). B and C are correlated, but none of the other known variables is the cause of both, so either B should be oriented towards C or vice versa. A should be connected to C to reflect dependency $A \not\perp\!\!\!\perp C \mid B$. But $A \perp\!\!\!\perp C$, thus there is a dependency between $P(B \mid A)$ and $P(C \mid A, B, D)$. The Bayesian network is therefore compressible and not faithful ($A \perp\!\!\!\perp C$ is not represented). The solution is to look for an alternative model class. Spirtes et al. (1995) propose the use of a *Partially-oriented Acyclic Graph* (PAG). They show that the model class of PAGs give shorter descriptions than Bayesian networks. The PAG that is learned for the system is shown in Fig. 7(c). An arrow in a PAG means that the node towards which the arrow points is not the cause of the opposite node. An “o” on an end of an edge means that we cannot decide whether the edge end should be an arrow or not. The relation $B \leftrightarrow C$ indicates that none of both variables is the cause of the other. Hence, that there must be a latent common cause. Consult Spirtes et al. (1995) for details.

Case 6. For some systems, the edges emanating from a single node are dependent. Take the example of particle decay, one of the counterexamples of the Causal Markov Condition reported by Williamson (2005, p. 55), taken from Fraassen (1980, p. 29):

> Suppose that a particle decays into 2 parts, that conservation of total momentum obtains, and that it is not determined by the prior state of the particle what the momentum of each part will be after the decay. By conservation, the momentum of one part will be determined by the momentum of the other part. By indeterminism, the prior

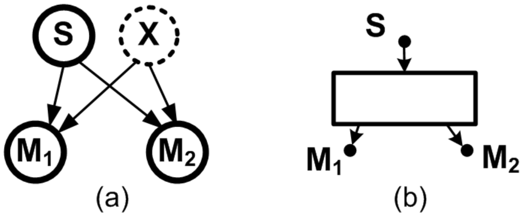

Figure 8: Particle with state S decays into 2 parts with momenta M_1 and M_2: faithful model with hidden variable (a), represented as a single mechanism (b).

> state of the particle will not determine what the momenta of each part will be after the decay. Thus there is no prior screener off.

The prior state S fails to screen off the momenta. But by symmetry, neither of the two parts' momenta M_1 and M_2 can be considered as the cause of the other. The system can be represented faithfully by adding a hidden common cause X, as shown in Fig. 8 (a). This solution must, however, be refuted for two reasons. First, X does not correspond to a known quantity in physics. Secondly, it is difficult to uphold that M_1 and M_2 are produced by distinct mechanisms. The mechanism producing M_1 is related to that producing M_2. This is reflected by the fact that there is no conceivable way of interfering in the mechanism that produces M_1 while leaving the mechanism that produces M_2 undisturbed or vice versa (Hausman and Woodward, 1999, p.551). We must conclude that basic components do not always consist of one effect variable, but can affect multiple variables at once. Such a mechanism with two outputs is depicted in Fig. 8 (b). It is characterized by dependent outgoing arrows emanating from a node. This can be observed by a strong correlation between the CPDs of Fig. 8 (a). This, however, contradicts the Causal Markov Condition. To defend the Causal Markov Condition, Hausman and Woodward (1999, p. 528) argue that M_1 and M_2 should not be regarded as two distinct events, but as just one event that is the effect of a single mechanism. A lot of similar examples in which the effect event is decomposable into parts can be found in literature, such as a cue ball colliding with two other billiard balls (Hausman and Woodward, 1999, p. 528).

Case 7. Another regularity is the repetition of similar mechanisms in a system. This results in a causal model in which identical CPDs appear. The model is therefore compressible. The compressibility does not, however, result in a dependence of the CPDs in terms of manipulability. One mechanism can still be replaced by another without affecting the rest of the model. Modularity still holds. *Object-Oriented nets* provide a representation format that explicitly capture similarities of mechanisms (Koller and Pfeffer, 1997).

9. Conclusions

If a Bayesian network provides the Kolmogorov Minimal Sufficient Statistic (KMSS) of a system, it can be interpreted as a causal model of the system. This interpretation is based on the assumption that the conditional independencies which the Bayesian network represents follow from the causal structure (Causal Markov Condition). The conditional independencies are the regularities on which the KMSS is based. For causal inference, modularity of the Conditional Probability Distributions (CPDs) of an incompressible minimal Bayesian network provides the top-ranked hypothesis about the system under study. This follows from the decomposition of the description which the Bayesian network provides and Occam's razor. It is then plausible, but not guaranteed to be correct, to match up the CPDs with mechanisms. Decomposition reflects the causal component of graphical causal models.

However, if the minimal Bayesian network of a probability distribution is compressible and is thus not the KMSS, the above conclusions may become invalid. If the description of a single CPD is compressible, this can result in unfaithfulness of the causal model. Causal inference is still possible, since the true model is an element of the set of minimal Bayesian networks and modularity is plausible. If on the other hand the concatenation of the CPDs is compressible, then the CPDs are no longer independent and the mapping of CPDs onto independent mechanisms becomes invalid. This can be due to a kind of meta-mechanism which governs other mechanisms, or a mechanism affecting multiple effect variables.

The principle of KMSS is, however, not directly applicable, since there does not exist an algorithm that computes the shortest description of an object. We are therefore limited to selecting the minimal model from an a priori chosen model class. The results of this paper make clear that the regularities under consideration must be carefully chosen. Bayesian networks only represent the conditional independencies following from a causal structure. Other regularities, such as the ones of the cases studied here, must also be taken into account for correct causal inference.

In the light of the concept of KMSS, the faithfulness property can be interpreted in a broader sense as *the ability of a model to explicitly capture all regularities of the data*. Faithfulness of a graph ensures the capacity of deriving answers to qualitative questions (in terms of conditional independencies) from the DAG only. This was one of the goals that Pearl put forward while building up causal model theory (Pearl, 1988, p. 79).

References

Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in Bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

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

G.F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

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

D.A. Freedman and P. Humphreys. Are there algorithms that discover causal structure? *Synthese*, 121:2954, 1999.

N. Friedman and M. Goldszmidt. Learning bayesian networks with local structure. In *In Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence*, 1996.

Péter Gács, J. Tromp, and Paul M. B. Vitányi. Algorithmic statistics. *IEEE Trans. Inform. Theory*, 47(6):2443–2463, 2001.

P. Grünwald. *The Minimum Description Length Principle and Reasoning under Uncertainty, ILLC Dissertation series 1998-03*. PhD thesis, University of Amsterdam, 1998.

Daniel M. Hausman and James Woodward. Independence, invariance and the causal Markov condition. *Bristish Journal For the Philosophy Of Science*, 50(4):521–583, 1999.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. Technical Report MSR-TR-94-09, Microsoft Research, 1994.

E.H. Herskovits. *Computer-Based Probabilistic Network Construction*. PhD thesis, Medical information sciences, Stanford University, CA, 1991.

G. Hesslow. Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science*, 43:290–2, 1976.

Daphne Koller and Avi Pfeffer. Object-oriented Bayesian networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence (UAI-97)*, pages 302–313, 1997.

Kevin B. Korb and Erik Nyberg. The power of intervention. *Minds and Machines*, 16(3):289–302, 2006.

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.

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

Anders L. Madsen and Bruce D'Ambrosio. A factorized representation of independence of causal influence and lazy propagation. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 8(2):151–165, 2000.

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

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

J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proc. of the 11th Conference on Uncertainty in Artificial Intelligence, ed P. Besnard and S. Hanks*, pages 499–506. Morgan Kaufmann, 1995.

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.

J. Suzuki. Learning Bayesian belief networks based on the MDL principle: An efficient algorithm using the branch and bound technique. In *Procs of the International Conf. on Machine Learning*, Bally, Italy, 1996.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *AAAI/IAAI*, pages 567–573, 2002.

Paul M. B. Vitányi. Meaningful information. In Prosenjit Bose and Pat Morin, editors, *ISAAC*, volume 2518 of *Lecture Notes in Computer Science*, pages 588–599. Springer, 2002.

Chris S. Wallace and David L. Dowe. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

Jon Williamson. *Bayesian Nets And Causality: Philosophical And Computational Foundations*. Oxford University Press, 2005.

Yang Xiang, S. K. Wong, and N. Cercone. Critical remarks on single link search in learning belief networks. In *Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence (UAI-96)*, pages 564–571, San Francisco, CA, 1996. Morgan Kaufmann Publishers.