Inference of Graphical Causal Models: Representing the Meaningful Information of Probability Distributions

Jan Lemeire JAN.LEMEIRE@VUB.AC.BE

Kris Steenhaut KRIS.STEENHAUT@VUB.AC.BE

Dept. of Electronics and Informatics (ETRO),
Vrije Universiteit Brussel (VUB) - Interdisciplinary Institute for Broadband Technology (IBBT),
Pleinlaan 2, 1050 Brussels, Belgium

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

This paper studies the feasibility and interpretation of learning the causal structure from observational data with the principles behind the Kolmogorov Minimal Sufficient Statistic (KMSS). The KMSS provides a generic solution to inductive inference. It states that we should seek for the minimal model that captures all regularities of the data. The conditional independencies following from the system's causal structure are the regularities incorporated in a graphical causal model. The meaningful information provided by a Bayesian network corresponds to the decomposition of the description of the system into Conditional Probability Distributions (CPDs). The decomposition is described by the Directed Acyclic Graph (DAG). For a causal interpretation of the DAG, the decomposition should imply modularity of the CPDs. The CPDs should match up with independent parts of reality that can be changed independently. We argue that if the shortest description of the joint distribution is given by separate descriptions of the conditional distributions for each variable given its effects, the decomposition given by the DAG should be considered as the top-ranked causal hypothesis. Even when the causal interpretation is faulty, it serves as a reference model. Modularity becomes, however, implausible if the concatenation of the description of some CPDs is compressible. Then there might be a kind of meta-mechanism governing some of the mechanisms or either a single mechanism responsible for setting the state of multiple variables.

1. Introduction

Causal inference is an ambitious research field, as it tries to learn how the world is put together from observations only. The algorithms for causal inference are based on the conditional independencies implied by the causal structure of the system. The theory of graphical causal models, as developed by Pearl et al., gives a probabilistic view on causation and is based on the theory of Bayesian networks. The Directed Acyclic Graph (DAG) of a Bayesian network can be regarded as a representation of the conditional independencies of a probability distribution. A causal model gives a causal interpretation to the edges of a Bayesian network. The causal interpretation is based on manipulability; the model exhibits the structure of the system such that it is able to predict changes to the system. Hausman and Woodward (1999) show that this interventionist interpretation of causality is tightly linked to modularity. They also defend the equivalence of modularity and the causal Markov condition (Hausman and Woodward, 1999, p. 554).

The causal interpretation of a Bayesian network is often criticized (Freedman and Humphreys, 1999; Cartwright, 2001; Williamson, 2005; Hausman and Woodward, 1999). This paper would like to contribute to the discussion by analyzing causal inference through the concept of the *Kolmogorov Minimal Sufficient Statistic* (KMSS). The idea is that patterns or regularities in the observed data do not happen by accident. They teach us the important properties of the system. We say that the regularities constitute the *meaningful information* of the data. The KMSS allows a formal separation of meaningful and random information, based on the Kolmogorov complexity of objects. The application of Kolmogorov complexity to inductive inference has given rise to different methods, such as Minimum Message Length (MML) (Wallace and Boulton, 1968) and Minimum Description Length (MDL) (Rissanen, 1978). These methods are used for selecting the best model from a given set of models. The choice of model class, however, determines the regularities under consideration. During our discussion, we will not stick to an a priori chosen set of regularities, but search for the relevant regularities. Regularities will show up to be of key importance for testing the validity of causal inference.

This paper puts forward that the meaningful information of a Bayesian network is the decomposition of the system's description into separate components, the Conditional Probability Distributions (CPDs). The correctness of the causal interpretation of this decomposition relies on whether the CPDs correspond to independent mechanisms. We will analyze the correctness by looking at the regularities not incorporated by the Bayesian network.

In Section 2, the concept of KMSS is introduced. In Section 3, we will give a survey of graphical causal model theory and the learning algorithms. The link between a Bayesian network and the KMSS of a probability distribution is discussed in Section 4. In Section 5 we will argue that causal inference is plausible if the Bayesian network gives the KMSS. Section 6 discusses the 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 Turing machine 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 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 call the model, and one part containing the remaining random noise. Such a description is called a *two-part code*. This results in generic approaches for inductive inference, such as *Minimum Description Length* (MDL). According to MDL we have to pick the model M_{mdl} from model class $\mathcal{M}$ where M_{mdl} is the model which minimizes the sum of the 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 and nothing more than the regularities of the data. The KMSS provides a formal separation of meaningful and meaningless information. We limit the introduction of 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 string x and the objects that share x's regularities. With $|S|$ the size of set S, all elements of a set S can be enumerated with a binary index of length $\log_2 |S|$. We 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. The index is constructed by index enumerating all elements of the set, its length is thus $\log_2 |S|$. Atypical elements have regularities that are not shared by most of the set's members and can therefore be described by a shorter description. Most elements of S are typical, since, by counting arguments, only a small portion of it can be described shorter than $\log_2 |S|$.

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

$$p^* = arg\ min_p \{l(p)\ \mid \mathcal{U}(p) = S,\ x \in S,\ K(S) + \log_2 |S| \leq K(x)\} \tag{4}$$

Program p^* minimally describes the meaningful information present in x and nothing else. This can be understood as follows. By the inequality, the two-part description is at least as short as the Kolmogorov complexity of x. Since we seek for the simplest S (minimal p), we will only describe regularities by p. Regularities compress the description and greatly reduce the size of S. Putting random information in S would also reduce $\log_2 |S|$, but would increase $K(S)$ equally.

$K(x)$ depends on the chosen Turing machine $\mathcal{U}$, or, in practice, on the chosen description language. Minimality and compressibility are thus partly dependent on the choice of language. Another problem for directly applying the definitions is the intractability of $K(x)$. It falls out of the scope of this paper to address these problems, consult (Li and Vitányi, 1997) for an in-depth analysis. We will apply the *principles* behind Kolmogorov complexity, MDL and KMSS in order to better understand causal inference, its interpretation and feasibility.

3. Graphical Causal Models

This chapter will introduce graphical causal models and the learning algorithms (Pearl, 2000; Spirtes et al., 1993; Tian and Pearl, 2002).

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) $P(X \mid Pa(X))$, where $Pa(X)$ is the set of parent nodes of X in the graph and constitute the direct causes of the variable. The combination of the CPDs results in the system's joint probability distribution:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid Pa(X_i)) \tag{5}$$

The right hand side is called a *factorization* of the joint probability distribution.

3.2 The Effect of Changes to the System

We attribute a causal interpretation to the edges of the graph, but what does this 'causal interpretation' signify? The approach of Pearl and many others is to draw a connection between causation and manipulability (Hausman and Woodward, 1999). The causal interpretation is defined by the model's capacity to predict the effect of changes to the system. Changes are defined by Pearl as *interventions*. An intervention is defined as an atomic operation that fixates a set of variables to some given states and eliminates the corresponding factors (CPDs) from the factorization (Eq. 5) (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 if 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. Note that they relate each CPD to a structural equation.

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

3.3 Representation of Independencies

The evidence for causal inference is the conditional independencies entailed by the system's causal structure. 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 (Spirtes et al., 1993). These independencies are irrespective of the nature of the mechanisms, of the exact parameterization of the conditional probability distributions $P(X_i \mid Pa(X_i))$. All independencies following from the causal Markov condition can be retrieved from the causal graph by the d-separation criterion. A causal graph is called *faithful* if all conditional independencies from the distribution follow from the causal Markov condition.

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 only concerned with offering a dense and manageable representation of joint distributions. A joint distribution over n variables can be *factorized* relative to a chosen variable ordering $(X_1, \ldots, X_n)$ as follows:

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

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:

$$X_j \perp\!\!\!\perp X_i \mid X_1 \ldots X_{j-1}, X_{j+1} \ldots X_{i-1} \tag{7}$$

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: $P(X_i \mid Pa(X_i))$. 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.

Causal models attribute a causal interpretation to the edges of the graph of a Bayesian network and are therefore called *causally interpreted Bayesian networks*. Bayesian networks are just dense descriptions of probability distributions and offer an explicit representation of dependencies and independencies. The link is that the causal Markov condition follows from the correctness of the factorization (Hausman and Woodward, 1999, p. 532).

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. All networks represent the probabilities just as well, except that some are more complex than others. We call the *minimal Bayesian networks* the Bayesian networks which have the least number of edges in their DAGs.

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), an entropy based method (Herskovits, 1991) and one based on the Minimum Description Length (Suzuki, 1996). Irrespective of the exact definition of the scoring criteria, we can say that the algorithms are looking for the minimal Bayesian network.

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 are based on the presence or absence of certain independencies. The algorithms assume minimality, faithfulness and the causal Markov condition (Spirtes et al., 1993). We are also searching for the minimal Bayesian network, since a faithful Bayesian network is minimal (Lemeire et al., 2009).

4. Bayesian Networks as Minimal Descriptions of Distributions

In this section we will draw the connection between Bayesian networks and the KMSS of probability distributions. Let us apply the principle of KMSS to inductive inference of multivariate data that are independently and identically distributed (i.i.d.). Following the principle, the inferred model should capture the regularities of the data. The type of regularity we have to consider 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. The KMSS of the distribution should be a minimal description of the distribution's regularities. In this section we will consider the description given by a Bayesian network, other regularities will be considered in Section 6.

From the theory of Bayesian networks, 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 (given by Eq. 7). This leads to the description of the joint distribution by a factorization: $P(X_1,\ldots,X_n) = \prod CPD_i$, with CPD_i the CPD of variable X_i, defining $P(X_i \mid Pa(X_i))$, a distribution over X_i conditional on a subset of some other variables. A two-part description of a joint distribution is then:

$$descr(P(X_1 \ldots X_n)) = descr(\{Pa(X_1),\ldots,Pa(X_n)\}) + descr(CPD_1) + \cdots + descr(CPD_n) \quad (8)$$

With $descr()$ denoting a description. The parents' lists can be described very compact by a DAG. The descriptive size of the CPDs is determined by the number of variables in the conditioning sets, the number of free parameters for describing the distributions and the chosen accuracy. Eq. 8 corresponds to the description of a Bayesian network. If this results in an incompressible description, the DAG gives the meaningful information and the KMSS. This is proven by the following theorem.

Theorem 2 *Given a set of probability distributions $\mathcal{P}$ defined over a set of n variables $X_1,\ldots,X_n$. Consider a probability distribution $P \in \mathcal{P}$ which can be decomposed by a factorization based on parents' lists $Pa(X_1),\ldots,Pa(X_n)$. Consider that the factorization is described (see Eq. 8) by a minimal code which is able to describe all $P' \in \mathcal{P}$ that can be described by a factorization based on the same parents' lists. If such a description results in an incompressible string, then the first part (the description of the parents' lists) is the Kolmogorov minimal sufficient statistic of P.*

Proof The parents' lists describe a subset $\mathcal{P}' \subset \mathcal{P}$ which includes all elements that can be decomposed by the same factorization. We assume that this description is incompressible, therefore its length corresponds to $K(\mathcal{P}')$ up to a constant [1]. The code used to describe the CPDs allows a description of all elements of $\mathcal{P}'$. It is a minimal code, so its length equals to $\log_2 |\mathcal{P}'|$. The total description is incompressible, so its length equals, up to a constant, to $K(P)$. We have found a set $\mathcal{P}'$, for which $K(\mathcal{P}') + \log_2 |\mathcal{P}'| \leq K(P)$.

Next, we have to prove that there is no other set, $\mathcal{P}''$, which has a shorter description and for which the inequality holds (see Eq. 4). Assume that such a $\mathcal{P}''$ exists. If $\mathcal{P}'' \subset \mathcal{P}'$, the description of the CPDs would be compressible. $\mathcal{P}''$ is a smaller set, indicating that there are regularities in P which are not described by $\mathcal{P}'$. If $\mathcal{P}' \subset \mathcal{P}''$, then, similarly, the description of $K(\mathcal{P}'') + \log_2 |\mathcal{P}''|$ would be compressible. It follows that if such a $\mathcal{P}''$ exists, both sets contain exclusive elements that do not belong to the other set. This implies that the descriptions of both

[1] Two minimal descriptions of x based on different codes or turing machines are equal up to a constant that is independent of x. Since the descriptions are minimal, they are incompressible.

sets exploit regularities of P that are not exploited by the other one. Therefore neither of the descriptions is minimal and incompressible. This proves that such a $\mathcal{P}''$ does not exist.

■

The DAG thus minimally describes the dependencies among the variables, the model's complexity is reduced by conditional independencies.

It must be noted that if there exists a faithful and minimal Bayesian network, it is not necessarily unique. Multiple minimal 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. Thus, all models of a Markov-equivalence class have the same complexity.

5. Correspondence of Decomposition to Independent Mechanisms

Causal inference from observations is based on finding the minimal Bayesian network (Sec. 3.5) and attaching a causal interpretation to it (Sec. 3.2). In this section we will discuss the case in which, for a given set of observations, there is exactly one minimal Bayesian network which is also the minimal description of the data. This means that there are no other regularities than the conditional independencies the model represents. The DAG is then the KMSS of the data and minimally represents all regularities. We argue that description minimality can be linked to causality.

Note that for not overloading the discussion we will assume causal sufficiency: there are no unknown variables that affect more than one known variable.

5.1 Correspondence of Bayesian Networks to a Decomposition

Let us first analyze what the meaningful information of a Bayesian network exactly represents. A Bayesian network describes a joint probability distribution by DAG G and a list of CPDs as given by Equation 8. The description is decomposed into individual CPDs. The decomposition is described by the DAG, it tells us which CPDs we have to consider. The DAG G contains the meaningful information. It describes a model set of distributions $\mathcal{P}_G$, all sharing the conditional independencies following from the Markov condition. The distributions that only have these independencies are the typical elements of $\mathcal{P}_G$. G is faithful to them. Distributions having other independencies, are atypical, their description based on G is compressible (Lemeire et al., 2009, Theorem 6). We will consider them in the next section. We will first assume that the description is unique and minimal.

A Bayesian network thus describes a decomposition and 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 (Theorem 2). 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. If the minimal Bayesian network is not unique, the Markov-equivalence class indicates exactly which parts are undecided, namely the orientation of some edges. So, we know exactly for which parts of the model we have not enough information to decide upon the decomposition.

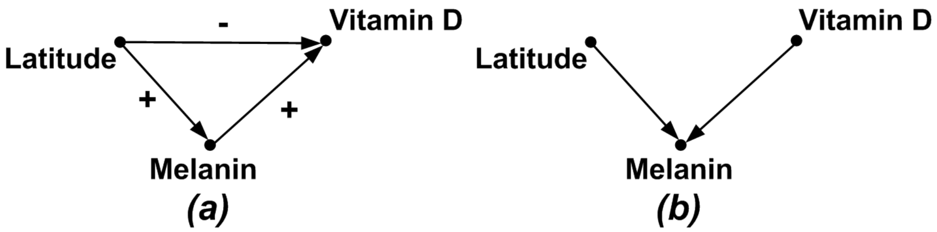

Figure 1: The relation between vitamin D creation and latitude: true causal model (a) and minimal Bayesian network (b).

For the remainder of the model, the decomposition is known. When we speak of the minimal Bayesian network, we actually mean the class of closely-related minimal Bayesian networks.

5.2 Correspondence of the Decomposition to Mechanisms

Let us now investigate whether the CPDs of the minimal Bayesian network of a probability distribution can be matched up with the mechanisms of the underlying system. The minimal Bayesian network provides modularity in the descriptive sense: the description consists of components. But does this also imply modularity in the causal sense: do the descriptive components correspond to independent parts of reality? In other words, does the model learned by observations reveal the underlying system?

We have found the simplest model. Following Occam's razor this is the model we should 'select'. But does Occam's razor also guarantees that this model tells us something about the real system? Yet, we did not only find the minimal Bayesian networks in the set of all Bayesian networks, we also found the minimal model in the general sense. The model is the KMSS and has extracted all regularities from the data. Moreover, the model is unique. From these facts we argue that:

In absence of background knowledge, experiments with interventions or other information, given that a minimal Bayesian network is the KMSS of the data, the top-ranked hypothesis is that each CPD represents an independent part of reality.

Before explaining what we exactly mean by 'top-ranked' hypothesis, let us consider two counter examples.

5.3 Counter Examples

Consider people living at different latitudes and the amount of vitamin D creation. Melanin is a pigment that protects us against harmful UV radiation. On the other hand, we need a limited amount of UV radiation to produce a necessary amount of vitamin D. To ensure this, evolution has given humans a different amount melanin, which is reflected by skin color, relative to the amount of sun they are exposed to. The latter is mainly affected by the latitude. This results in a nearly constant amount of vitamin D creation independent from the latitude we live at. Figure 1 shows the real causal model (a) and the minimal Bayesian network (b). Evolution has controlled the $Latitude \rightarrow Melanin$ relation such that the parameters were calibrated until the influences from $Latitude$ on $Vitamin\ D$ neutralized. This is a counter example of Occam's razor; the simplest model does not give us the true model. There is a *meta-mechanism*, namely evolution, controlling the mechanisms.

Next, consider the coder-decoder example, taken from Spirtes et al. (1993, Figure 3.23), shown by Figure 2(a). Variable Y encodes the values of both R and X, and Z decodes Y to

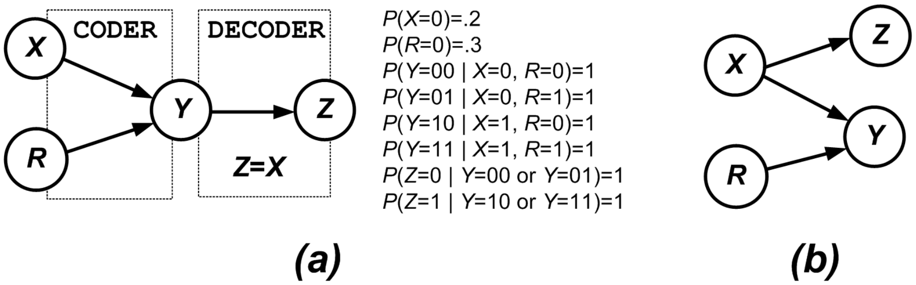

Figure 2: Coder-decoder system, taken from Spirtes et al. (1993, Figure 3.23), in which Z equals to X. Description of the system (a) and minimal model (b).

match the value of X. This is possible because the first bit of Y corresponds to the value of X. The coder-decoder system is designed to exhibit the specific behavior that Z equals to X. The model describing such a system, shown in Figure 2(a), is clearly not minimal. In the minimal description of the system given by Figure 2(b), Z is directly related to X. Occam's razor is violated. The CPD components are part of a greater mechanism and are engineered to match each other in such a way that the desired functionality is realized. The causal interpretation of the minimal model is incorrect. If we intervene on Y by manually setting the value of Y to a certain value which is not controlled by X, then Z becomes independent of X.

5.4 Conclusions about the Causal Interpretation of the Decomposition

The examples illustrate that the real system can be more complex than suggested by the complexity of the observations. Does this invalidate our claim about the minimal Bayesian network? It shows that Occam's razor cannot always be trusted. The minimal model may be incorrect, in the sense that the causal interpretation should be considered with care. But even when faulty, the minimal Bayesian network tells us two things about the system.

First, the DAG of the minimal Bayesian network describes the qualitative behavior of the system. This is true for both counter examples. From Figure 1(b) we see that the amount of vitamin D creation is independent from the latitude we live in. From Figure 2(b) we immediately see that Z only depends on X and is totally independent from R and Y. The structure of the real causal models does not reveal these independencies. We have to carefully study the parameterization to understand these independencies. Moreover, these are not accidental independencies. For the first example, it is the result of an evolution triggered by the evolutionary fitness. For the second example, it was the deliberate intention of the engineer to give the system this specific behavior. This corresponds to the rationale behind Occam's razor: regularities are most likely not accidental, but indications of a kind of mechanism. Only the occurrence of accidental (in)dependencies in the data, due to a limited sample size for example, makes the minimal model not correctly describing the system's behavior.

Secondly, we argue that the minimal Bayesian network at least serves as a reference model. The correspondence of the CPDs to real mechanisms might be untrue due to a meta-mechanism controlling the configuration of the system. The likelihood of the occurrence of such a mechanism must be estimated from background knowledge. In that case, experiments with interventions will have to be performed in order to reveal the true causal model (Korb and Nyberg, 2006). But even then will the minimal Bayesian network show its value. It can be used as a *reference model*, which will be compared with the model learned after the application of the interventions. This comparison would reveal the meta-mechanism.

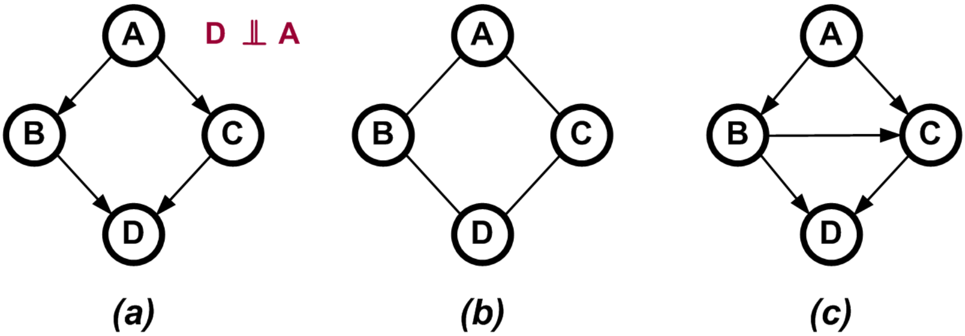

Figure 3: O-structure in which A is independent from D (a). A Markov network (b) and one of the minimal Bayesian networks describing the same system (c).

6. Compressibility of the Minimal Bayesian Network

In this section we will dig deeper. We will investigate the consequences for cases in which the minimal Bayesian network does not describe the KMSS. Then, the DAG is not the only meaningful information. There exists a simpler description of the distribution than given by the minimal Bayesian network. First we will consider the compressibility of an individual CPD and then we consider the compressibility of several CPDs taken together.

6.1 Compressibility of a Single CPD

Compressibility of individual CPDs is called *local structure* (Friedman and Goldszmidt, 1996). In this terminology, the DAG describes the global structure, the CPDs the local structure. On top of the independencies following from the causal structure the individual CPDs exhibits additional regularities. For discrete models for example, the conditional probability tables can be described shorter by decision trees when so-called context-specific independencies appear (Boutilier et al., 1996).

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 it disjunctive causes. On top of the decomposition given by Equation 8, the parts of the decomposition can be further decomposed.

In these cases, the decomposition according to Equation 8 is still valid. Modularity is still valid, in the descriptive and causal sense. The same conclusions as those of the previous section apply.

6.2 Compressibility of the Concatenation of CPDs

When the description of some CPDs together can be compressed, the regularity indicates that the CPDs are in some way related. The following counter examples show that this often invalidates the causal interpretation.

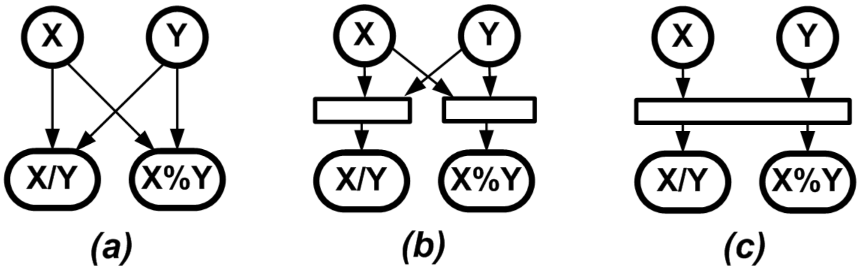

Figure 4: Bayesian network (a) of a system that calculates the quotient and remainder of two integers. The decomposition it represents (b) and a single mechanism calculating both outputs together (c).

6.2.1 O-STRUCTURE

The most-known counter example of causal inference is when in the model of Figure 3(a), 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. This cancellation is similar to the vitamin D example of Section 5.3. Except that in this case the DAG of the minimal Bayesian network corresponds to the structure of the model. The CPDs and the mechanisms are, however, not independent. Pearl considers the exact cancellation of the parameters as a measure zero event, since the probability of such a coincidence can therefore be regarded as nearly zero (Pearl, 2000, p. 48). This is true as long as there not a kind of *meta-mechanism* controlling the mechanisms such that the parameters are calibrated until they neutralize. This confirms our conclusions of the previous section that the existence of such mechanisms must be taken into consideration. Then, the likelihood of a cancellation is not zero.

Modularity becomes invalid when the meta-mechanism acts instantly. In the vitamin D case, evolution works slowly. On short term, the mechanisms are independent. It is only on the long term that the calibration will be reestablished.

6.2.2 MARKOV NETWORK

Consider a system that is minimally described by a Markov network, as shown in Figure 3 (b). Variables which are connected by a path in the network are dependent, unless each path is blocked by one of the conditioning variables. So is $B \not\perp\!\!\!\perp C \mid A$, but $B \perp\!\!\!\perp C \mid \{A, D\}$. For describing the same network with a DAG, we have to orient the edges of the network. For acyclicity, we have to create at least one v-structure. We can choose for example $B - D - C$. But then, for keeping the same dependencies, we have to add an edge, as shown in Figure 3 (c). Without $B \rightarrow C$ we would have $B \perp\!\!\!\perp C \mid A$. Clearly, this Bayesian network is not minimal; the description is longer than that of a Markov network. The parameterizations of the CPDs contain redundancies. In the model of 3 (c), the parameterizations must ensure that $B \perp\!\!\!\perp C \mid \{A, D\}$, an independency which is not captured by the DAG. When such a distribution is observed, the causal interpretation of the CPDs of the minimal Bayesian networks is incorrect.

6.2.3 MULTIPLE-OUTPUT FUNCTIONS

Consider a system that calculates the quotient and remainder of two integers. Figure 4(a) shows the minimal Bayesian network of the system. The model describes the system as two different mechanisms, one for calculating the quotient and one for the remainder, shown in Figure 4(b). Both mechanisms, however, are related; there is a lot of overlap in calculating the quotient and the remainder. A model describing the system by one component which calculates both outputs together, as shown in Figure 4(c), is more compact than a Bayesian network which only allows components with single outputs. In that case, the CPDs of the minimal Bayesian network cannot be considered as independent. If the output variables are clearly separated quantities, a mechanism setting the values of multiple variables should be taken into consideration.

6.2.4 OBJECT-ORIENTED NETS

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 necessarily result in a dependence of the CPDs in terms of manipulability. It depends on the meta-mechanism responsible for the regularities in the system. The system could, for example, be designed by an engineer, such as a digital circuit. Then, modularity holds; one mechanism can be replaced by another without affecting the rest of the model. *Object-Oriented nets* provide a representation format that explicitly capture similarities of mechanisms (Koller and Pfeffer, 1997).

7. Conclusions

We showed that the meaningful information described by a Bayesian network about a probability distribution is the decomposition of the distribution into CPDs. We argue that if the shortest description of the joint distribution is given by separate descriptions of conditional distributions, it is the *top-ranked causal hypothesis*:

(1) The Bayesian network gives a correct description of the behavior of system. The qualitative properties, namely the conditional independencies, are incorporated in the DAG. This only becomes invalid by accidental (in)dependencies due to for instance a limited sample size.

(2) The likelihood that the CPDs correspond to independent mechanisms of the system (modularity) depends on the likelihood of a kind of meta-mechanism. A meta-mechanism could result in a system which is more complex than suggested by its behavior. In that case, the behavior of the system when subjected to an external intervention will not be correctly predicted by the minimal model. Nonetheless, the minimal Bayesian network can serve as a reference model that has to be compared with the behavior of the system after applying the interventions.

By applying the principle of KMSS, we did not only look for the minimal Bayesian network in the set of all Bayesian networks. We also took into consideration the presence of other regularities than the conditional independencies following from the system's causal structure. These regularities might invalidate the above conclusions. If such regularities appear in individual CPDs, the mapping of CPDs onto independent mechanisms still holds (in the above sense). The DAG of the Bayesian network does not have to be the KMSS. The decomposition should be correct. It becomes incorrect if the concatenation of the description of the CPDs is compressible. Then the CPDs are no longer independent and modularity might become invalid. The dependence might be caused by a meta-mechanism that governs the dependent mechanisms, or a mechanism affecting the state of multiple variables.

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.

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.

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

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, Kris Steenhaut, and Abdellah Touhafi. When are graphical causal models not good models? In *Causality in the sciences, J. Williamson, F. Russo and P. McKay, editors*, 2009.

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

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

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.

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 D. M. Boulton. 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.