Causal Models as Minimal Descriptions of Multivariate Systems

JAN LEMEIRE AND ERIK DIRKX

ABSTRACT. By applying the minimality principle for model selection, one should seek the model that describes the data by a code of minimal length. Learning is viewed as data compression that exploits the regularities or qualitative properties found in the data, in order to build a model containing the meaningful information. The theory of causal modeling can be interpreted by this approach. The regularities are the conditional independencies reducing a factorization and the v-structure regularities. In the absence of other regularities, a causal model is faithful and offers a minimal description of a probability distribution. The causal interpretation of a faithful Bayesian network is motivated by the canonical representation it offers and faithfulness. A causal model decomposes the distribution into independent atomic blocks and is able to explain all qualitative properties found in the data. The existence of faithful models depends on the additional regularities in the data. Local structure of the conditional probability distributions allow further compression of the model. Interfering regularities, however, generate conditional independencies that do not follow from the Markov condition. These regularities has to be incorporated into an augmented model for which the inference algorithms are adapted to take into account their influences. But for other regularities, like patterns in a string, causality does not offer a modeling framework that leads to a minimal description.

1 Introduction

This paper intends to broaden the view on causal modeling theory by interpreting it with the principles of minimality, regularity extraction and meaningful information. The principles follow from the two-part code version of the MML/MDL approach to model selection. A good model offers a minimal description that consists of the description of the model and the data with the help of the model. Both, causal modeling and MML/MDL, share the general aim of statistical inference, to ‘understand’ the system behind the studied phenomena via the observed data. MML/MDL answers the question of how one should decide among competing explanations of data by providing an objective definition of complexity, so that *Occam’s Razor* can be applied. The theory of causal models gives a probabilistic view on causation and is based on Bayesian networks. It intends to give a causal interpretation to the edges of a Bayesian network. It is based on the minimality principle, the causal Markov condition and the faithfulness property [SGS93]. The causal interpretation of the edges and the validity of faithfulness are often criticized [FH99, Car01, Wil05]. We will argue that faithfulness is an appealing principle in modeling and will study the conditions under which it is violated. Our analysis is based on the concept of *regularities*, which are defined as properties allowing data compression. They are regarded as meaningful information and should be put in the model part of the two-part code. The regularities can be interpreted as *qualitative properties*, as opposed to the quantitative information that is put in the data-to-model part. For a Bayesian network, the directed acyclic graph (DAG) is the model and describes the qualitative properties of the distribution, i.e. the conditional independencies. The conditional probability distribution (CPDs) contain the quantitative information. The faithfulness property can then be interpreted as the ability of the model to explain all qualitative properties of the data.

The following two sections review the theories of MML/MDL and causal models. Section 4 discusses related work and section 5 connects both theories. Section 6 shows that faithfulness is correct in absence of other regularities and section 7 explains the canonical representation that causal models offer. Finally, section 8 discusses the implications of additional regularities to the validity of the causal modeling framework and faithfulness.

2 Two-part code

Several related methods, such as Minimum Message Length (MML) [WD68, Wal05] and Minimum Description Length (MDL) [Ris78], provide a generic solution to the model selection problem: one has to decide among competing explanations of data given limited observations. The central idea is that learning can be equated with finding *regularities* in data. An objective property of 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 more regularities there are, the more the data can be compressed. Learning has to be viewed as data compression. A good model should capture the regularities. This results in a *two-part code*: the first part describing the model and the second part describing the data with the help of the model [WD99, Ris89]. The total description length is then:

$$description\ length = L(model) + L(data \mid model) \tag{1.1}$$

By minimizing the total length of both descriptions, this approach inherently protects against overfitting and trade-offs goodness-of-fit on the observed data, quantified by the second part, with complexity of the model, the first part. A minimal code contains no redundancy, every bit is information. However, only the regularities are regarded as *meaningful information*. The random string '110100100111010' is incompressible, but contains no meaningful information. The regular string '001001001001001' can be described as 5 times repeating '001'. The repetition is the regularity and has to be described by the model, whereas the sequence '001' should be put in the data-to-model part. Consult the work of Vitanyi et al. for a formal treatment of the distinction between meaningful and random information [Vit02, Vit05, GTV01].

Learning has thus to be viewed as a process of building a model by squeezing out all regularities from the data. The model then defines a set of objects all sharing the same regularities, which we call the *model set*. The set defined by the model of string '001001001001001' contains the 8 strings that repeat a 3-bit substring. The string is called *typical* for this set, because it shares all its regularities with the other elements of the set [LV97](sec. 1.9). It is however not typical for the model set of a random string. This set contains all 2^{15} string of 15 bits. For learning the two-part code, one should pick the minimal model for which the data is typical and which results in a minimal two-part code. Atypicalness is measured by the *randomness deficiency* of the element with respect to the set [Vit05]. To identify an element from the set of random strings, we need an index of 15 bits (the logarithm of the size of the set). The regular string can, however, be described with a shorter code than 15 bits and is therefore not typical for the set. A random string cannot be described with a shorter code. The large majority of the elements of a model set is incompressible, while only a few exhibit additional regularities that allow further compression. When picking an element from the set, it will be a typical element with high probability. Thus, if the observed data is typical, the minimal model corresponds to the correct model. This is a necessary condition for the correctness of learning methods.

3 Causal Models

We elaborate the theory of causal models in three steps. First, we show how a Bayesian network describes a probability distribution. Secondly, a faithful model is defined as describing all independencies of a distribution. Ultimately, a causal interpretation is given to the network.

3.1 Representation of Distributions

A causal model is fundamentally a Bayesian network, which offers a dense representation of a *joint distribution*. A joint distribution is defined over a set of stochastic variables $X_1 \ldots X_n$ and defines a probability ($P \in [0, 1]$) for each possible state $(x_1 \ldots x_n) \in X_{1,dom} \times \cdots \times X_{n,dom}$, where $X_{i,dom}$ stands for the domain of variable X_i. The joint distribution can be *factorized* relative to a variable ordering $(X_1, \ldots, X_n)$ as follows:

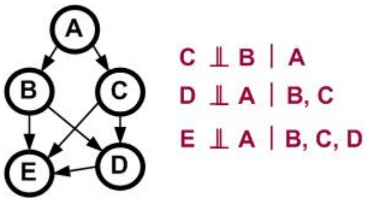

Figure 1.1. Factorization based on variable ordering (A, B, C, D, E) and reduction by three independencies.

$$P(X_1, \ldots, X_n) = \prod_i^n P(X_i \mid X_1, \ldots, X_{i-1}) \tag{1.2}$$

Variable X_j can be removed from the conditioning set if it becomes conditionally independent from X_i by conditioning on the rest of the set: $P(X_i \mid X_1 \ldots X_{i-1}) = P(X_i \mid X_1 \ldots X_{j-1}, X_{j+1} \ldots X_{i-1})$. 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 all variables conditioned on its parents, $P(X_i \mid parents(X_i))$. A *Bayesian network* is a factorization that is minimal, in the sense that no edge can be deleted without destroying the correctness of the factorization.

Although a Bayesian network is edge-minimal, it depends on the chosen variable ordering. Some orderings lead to the same networks, but others result in different topologies. Take 5 stochastic variables A, B, C, D and E. Fig. 1.1 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. The Bayesian network based on ordering (A, B, C, E, D) depicted in Fig. 1.2, however, contains 2 edges less because of 5 useful independencies.

3.2 Representation of Independencies

Pearl, Verma, and others started to interpret the DAG of a Bayesian network as a representation of the conditional independencies of a joint distribution [Pea88]. They constructed a graphical criterion, called d-separation, for retrieving independencies from the graph that follow from the *Markov condition*, which states that a node becomes independent of its non-descendants by conditioning on its parents.

DEFINITION 1.1. (d-separation) Let p be a path between a node X and a node Y of a DAG G. (By a path we mean any succession of edges, regardless of their

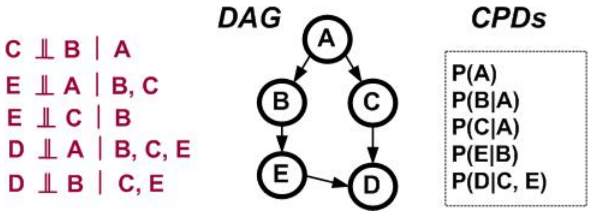

Figure 1.2. Bayesian network based on variable ordering (A, B, C, E, D) and five independencies.

directions.) Path p is called *unblocked* given subset **Z** of nodes in G if every node w on p satisfies:

1. if w has two converging arrows along p, w or any of its descendants is in **Z**.

2. and if w has no converging arrows, it is not an element of **Z**.

X and Y are called *d-connected* given **Z**, if there is an unblocked path between them in G. Conversely, **Z** is said to *d-separate* X from Y in G, denoted $X \perp Y \mid$ **Z**, iff **Z** blocks every path from X to Y. ***Z** blocks* a path if the above condition is not valid for one of the nodes on the path.

Take the graph of Fig. 1.2. The d-separation criterion tells us that variable B separates A from E, since B blocks the path $A \rightarrow B \rightarrow E$. On the other hand, the path $A \rightarrow C \rightarrow D \leftarrow E$ is blocked by $C \rightarrow D \leftarrow E$, which is called a $v - structure$. This path gets unblocked given D.

A graph is an *Independence Map*, or *I-map* for short, of a joint distribution if every independency found in the graph appears in the distribution. The DAG of a Bayesian network is a minimal I-map, removing an edge from the graph destroys its I-mapness. It is called a *perfect map* if it represents all conditional independencies of the distribution. The Bayesian network is then called *faithful* to the distribution.

3.3 Representation of Causal Relations

Where Bayesian networks are mainly concerned with offering a dense and manageable representation of a joint distribution, causal models intend to describe graphically the structure of the underlying physical mechanisms governing a system under study. In a causal model 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 the parent variables in the graph. The model then corresponds to a joint distribution defined over the variables and results in a close connection between causal and probabilistic dependence [Spo01]. For a causal model, the *Causal Markov condition* tells us how variables depend on each other: each variable is probabilistically independent of its non-effects conditional on its direct causes. The probabilistic aspect is similar to the Markov condition. Hence, a causal model can be regarded a Bayesian network in which all edges are interpreted as representing causal influences between the corresponding variables.

4 Related Work

Related work use the MML and MDL approach for selecting the best model from a given set of models. Scoring-based algorithms for learning causal models from data, for example, use them to give an objective score to the models of the model class [CD03, CD05, LB94]. The choice of the model class determines the regularities that are considered. During our discussion, we try not to stick to an a priori chosen set of regularities, but search for the relevant regularities.

By Theorem 1.2.4 of [Pea00], Pearl describes for which distributions faithful graphs exist and can be learned: ”the absence of *d*-separation implies dependence in almost all distributions compatible with the graph G. The reason is that a precise tuning of the parameters is required to generate independency along an unblocked path in the diagram, and such tuning is unlikely to occur in practice.” Pearl solves this problem by imposing a *stability* restriction on the distribution [Pea00](sec. 2.4). The occurrence of any independency must remain invariant to any change in the distributional parametrization of the graph. This corresponds with regularities in the CPDs, as will be proved by theorem 1.3. A change of the CPDs would break the regularity. Pearl claims that there exists at least 1 distribution faithful with the model, while we show that all typical models of the DAG model set 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 more general framework, called *imsets* [Stu01]. We advocate a different approach. Instead of looking for a broader representation form or a set of conditions (like Pearl), we consider that some of the violations of faithfulness are due to local regularities that can be added to an augmented model.

5 MDL Models of Multivariate Systems

Now that we have reviewed the theory, we will apply it in designing minimal models of multivariate systems. The MDL approach tells us to exploit the regularities of the data. The primary regularities in a distribution are the dependencies among the variables and exploiting them will result in a Bayesian network. A code for optimally describing the state of a stochastic variable X with domain X_{dom} and distribution $P(X)$ will have an average code length that can be no shorter than Shannon’s *entropy*:

$$H(X) = -\sum_{x \in X_{dom}} P(x).log_2(P(x)) \tag{1.3}$$

The Huffman code comes optimally close to the minimal code length [CT91]. The same is valid for a set of variables and the corresponding joint distribution. A minimal code can be constructed by a factorization, described by Eq. 1.2. The expected code length of X_i will be the conditional entropy $H(X_i \mid X_1 \dots X_{i-1})$. By the dependence of $P(X_i)$ on the values of its parents $X_1 \dots X_{i-1}$, the average code length is shorter than the entropy of X_i. This reduction in entropy is called the *mutual information*:

$$I(X_i; X_1 \dots X_{i-1}) = H(X_i) - H(X_i \mid X_1 \dots X_{i-1}) \tag{1.4}$$

The code tells us how to encode the data optimally. The model, however, should also include the description of the code. It contains the factorization ordering and the codes used for describing X_i with the help of $X_1 \dots X_{i-1}$, which are based on $P(X_i \mid X_1 \dots X_{i-1})$. The sizes of the conditioning sets of the CPDs thus greatly determines the complexity of the model. These can be reduced by conditional independencies as discussed in the previous section. A conditional independency indicates that a variable is unprofitable for further compression of the description of another. Omitting such variables from the factorization reduces the complexity of the model. We therefore have to look for a factorization ordering that leads to a Bayesian network with a minimal number of edges.

The resulting model is then:

$$model = DAG + CPDs + encoded\ data \tag{1.5}$$

where the first two terms define the distribution. The dependencies and conditional independencies are the regularities that compress the data.

The following theorem proves the minimality of a faithful Bayesian network.

THEOREM 1.2. *If a faithful Bayesian network exists for a distribution, it is the minimal factorization.*

Proof. Oliver and Smith define the conditions for sound transformations of Bayesian networks, where sound means that the transformation does not introduce extraneous independencies [OS90]. No edge removal is permitted, only reorientation and addition of edges. Additionally, if a reorientation destroys a v-structure or creates a new one, an edge should be added connecting the common parents in the former or in the newly created v-structure. Such transformations however eliminate some independencies represented by the original graph. Assume the existence of a Bayesian network based on a different variable ordering that has fewer edges than the faithful network. It must be possible to transform one into the other. But since the network has fewer edges it must have more independencies, which is impossible because the faithful network represents all independencies. ■

The graph of Fig. 1.2 can be transformed into that of Fig. 1.1 by reversing the edge $D \rightarrow E$. However, this destroys the v-structure $E \rightarrow D \leftarrow C$ and creates a new one, $B \rightarrow E \leftarrow D$. To assure the I-mapness of the new graph, C should be connected with E and B with D.

Multiple faithful models can exist for a distribution though - models that represent the same set of independencies - that are therefore statistically indistinguishable, they define a Markov-equivalent class. It is proved that they only differ in the orientation of the edges, but share the same v-structures [Pea00]. They thus all have the same complexity.

6 Faithfulness of a Random Bayesian Network

The main idea behind this paper is that models should capture the regularities observed in the data. The relational regularities, appearing as (conditional) (in)dependencies, where used to construct a Bayesian network. The following theorem shows that, although faithfulness cannot be guaranteed, d-separation tells which independencies can be 'normally' expected from a DAG of a Bayesian network, where normal means that it holds for the typical elements of the model set of the DAG. This set is created by generating distributions with randomly chosen CPDs.

THEOREM 1.3. *A Bayesian network with unrelated, random conditional probability distributions (CPDs) 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 holds that

$$P(X_i \mid pa_{i,1}, \dots pa_{i,j}, \dots pa_{i,k}) \neq P(X_i \mid pa_{i,1}, \dots pa_{i,j-1}, pa_{i,j+1}, \dots pa_{i,k}) \tag{1.6}$$

The proof will show that any two variables that are d-connected are dependent, unless the probabilities of the CPDs are related. We 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 nodes D and E of the Bayesian network of Fig. 1.2. For not overloading the proof, we will demonstrate that $P(D \mid E) \neq P(D)$, but the proof can easily be generalized. The first term can be written as:

$$P(D \mid E) = P(D \mid E, c_1).P(c_1) + P(D \mid E, c_2).P(c_2) + \dots \tag{1.7}$$

with c_1 and $c_2 \in C_{dom}$. C is also a parent of D, thus, by Eq. 1.6, there are at least two values of C_{dom} for which $P(D \mid E, c_i) \neq P(D \mid E)$ [2]. Take c_1 and c_2 being

[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 individual probability $P(D \mid E, c_1)$ is different than this average, let's say higher, than there is at least one value lower than the average.

such values, thus $P(D \mid E, c_1) \neq P(D \mid E, c_2)$. There are also at least 2 such values of E_{dom}, take e_1 and e_2. Eq. 1.7 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:

$$\begin{aligned} & P(D \mid e_1, c_1).P(c_1) + P(D \mid e_1, c_2).P(c_2) \\ = \; & P(D \mid e_2, c_1).P(c_1) + P(D \mid e_2, c_2).P(c_2) \end{aligned} \tag{1.8}$$

Note that the equation can not be reduced, the conditional probabilities are not equal to $P(D)$ nor to each other.

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. 1.2, independence of A and E requires that

$$P(E \mid a) = \sum_{b \in B} P(E \mid b).P(b \mid a) = P(E) \quad \forall a \in A. \tag{1.9}$$

and this also results 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. 1.6. Finally, if there are multiple unblocked paths connecting two variables, then independence of both variables implies a regularity, too (d). Take A and D in Fig. 1.2:

$$P(D \mid A) = \sum_{b \in B} \sum_{c \in C} \sum_{e \in E} P(D \mid c, e).P(c \mid A).P(e \mid b).P(b \mid A). \tag{1.10}$$

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. 1.6, so that the equation only equals to $P(D)$ if there is a relation among the CPDs. ■

Table 6 gives an example distribution of $P(D \mid E, C)$ for which Eq. 1.8 holds assuming $P(C = 0) = P(C = 1) = 0.5$. It results in the independence of D and E.

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

Table 1.1. Example of a CPD for which $P(D \mid E) = P(D)$, assuming $P(C = 0) = P(C = 1) = 0.5$.

A distribution can be described by several distinct Bayesian networks though, which are based on different variable orderings (see subsection 3.2). The networks build with a non-optimal variable ordering are however not minimal, the CPDs are related. Take the distribution described by the Bayesian network of Fig. 1.2, $E \perp\!\!\!\perp C | B$ holds. But this independency does not follow from the graph of Fig. 1.1. It follows from the exact cancellation of the influences through the paths $C \rightarrow E$ and $C \rightarrow D \rightarrow E$, given by equation:

$$P(E \mid B, C) = \sum_{d \in D} P(E \mid B, C, d).P(d \mid B, C) = P(E \mid B). \tag{1.11}$$

The regularities come from differences in v-structures, as explained in the proof of theorem 1.2. We call them *v-structure regularities*. A causal model is thus an factorization that has eliminated the v-structure regularities. The model is faithful in absence of other regularities.

On the other hand, we want to know the conditions under which a joint probability distribution can be represented by a faithful model. In [Pea88](p. 128), Pearl developed a set of necessary, but not sufficient, conditions, but adds that he doubts if there exists an exhaustive list of conditions that can guarantee faithfulness [Pea88](p. 131). This is approved by the theorem. Any dependency among non-adjacent variables that follows from the Markov condition can be turned into an independency by properly chosen values for the CPDs.

7 Canonical Representation

A Bayesian network decomposes a joint distribution into submodels of the form $P(X_i \mid parents(X_i))$. As shown in the previous section, the CPDs of the non-minimal networks are related. The model is therefore not a *canonical representation*, in which the decomposition is unique, minimal and the elements are atomic and independent. A causal model, based on a faithful Bayesian network, offers a canonical representation. The submodels are independent and correspond to the physical mechanisms that generate the data [Pea00]. Causal models thus have more explanatory power than Bayesian networks. Each submodel represents a stochastic process by which the values of X_i are chosen in response to the values of $parents(X_i)$, and the stochastic variation of this assignment is assumed independent of the variations in all other assignments. Each assignment process remains invariant to possible changes in assignment processes that govern other variables in the system. This *modularity assumption* enables the prediction of the effect of *interventions*, which are defined as specific modifications of some factors in the product of the factorization (Eq. 1.2). Moreover, all consequences of causality, like inference and identifiability, solely depend on these building blocks. The digital circuit of Fig. 1.3 represents a causal model. It not only defines the binary functions *c=f(x,y,z)* and *s=f(x,y,z)*, but makes it possible to reason about changes

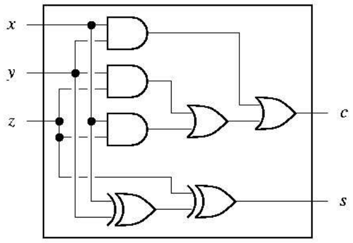

Figure 1.3. Digital Circuit.

to the system. The scheme tells us how the function will change if a digital component is replaced by another. It can also serve for finding out which components should be changed to become a desired function.

If the system is indeed build up by causal processes, they can be assumed to be non-related. Following theorem 1.3, the model will be faithful. The chance of an exact correspondence of probabilities so that two variables become independent is quite small, as expressed by the stability property used by Pearl (see sec. 4).

All Bayesian networks representing the same joint distribution (build starting from different variable orderings) can be used for predicting unknown variables quantitatively. They are however not suitable for *qualitative reasoning*: for answering questions about conditional independencies, such as 'does A affect B when C is known'. A non-faithful model can only answer these questions correctly by quantitative calculation of the probabilities. The model of Fig. 1.1 suggest that B, C and E should be known to have maximal information about D. But from the graph of Fig. 1.2 we know that B has no additional information about D, once we know C and E. Qualitative reasoning based on the model only demands that the model contains all meaningful information, that all qualitative properties can be inferred from it without needing the quantitative information. This is exactly what the faithfulness property stands for.

Following the MML/MDL principle, according to which modeling should be based on the regularities, faithfulness motivates the ability to learn causal models from observations: if a unique, minimal model has the power to foresee all consequences, it must come close to reality. The causal interpretation cannot be guaranteed, but offers, in absence of counterexamples or background knowledge, a plausible hypothesis. When we refer to the correct structure, we actually mean the correct Markov-equivalence class, as the specific structure within the equivalence class cannot be distinguished based on observational data solely. The constraint-based learning algorithms are able to learn this equivalence class, which can be represented by a graph in which some edges are not oriented. Thus, although the model is not unique, we know which parts of the model are undecided.

8 Additional Regularities

It follows from theorem 1.3 that if the minimal Bayesian network is not faithful, there are regularities among the CPDs. This section discusses the implications of such regularities. Following the MML/MDL principle, the regularities indicate that a less complex model exists. We distinguish three possible implications:

1. The model remains faithful, the regularities do not 'interfere' with the conditional independencies. They can thus be regarded as regularities of a lower level. A well-known example is when the description of individual CPDs can be further compressed. This regularity is called *local structure* [BFGK96] and appears inside a building block. Another regularity is the repetition of structures in the model. Identical components appear at different locations in the model. This is partly covered by Object-Oriented nets [KP97].

2. The causal model is unfaithful, but is still applicable. The unfaithfulness is a result of interference of the regularities with the conditional independencies - the regularities generate independencies not resulting from the Markov condition alone. The distributions can however be described minimally by a causal model augmented with a description of the additional regularities. We argue that it is indispensable to know these regularities for being able to learn the correct model and for performing the right inferences from it. Examples are given in the next subsection.

3. Causal models do not provide a minimal description. The minimality principle views causality as describing a type of regularities. This does not exclude that other regularities need other modeling frameworks, like in the model of a pattern. Take the set of strings of n bits for which m consecutive bits are 1 and the others are 0. Every bit can be regarded as a discrete variable. By picking elements from the set randomly, the joint distribution is observed. The correct model for a (n, m) pair contains a latent variable, denoting the start position of the non-zero bit sequence, which determines all bits. The graph is thus trivial, highly compressible and adds no meaningful information to the model. For minimality, the graph should not be explicitly added in the description. Next, it is questionable if the latent variable can be interpreted as the root cause of the bits.

8.1 Interfering Regularities

Regularities among the CPDs result in independencies that are not entailed by the Markov condition. The causal model is still an I-map, ie. all independencies implied by the Markov condition are present in the distribution, but the models are no longer faithful.

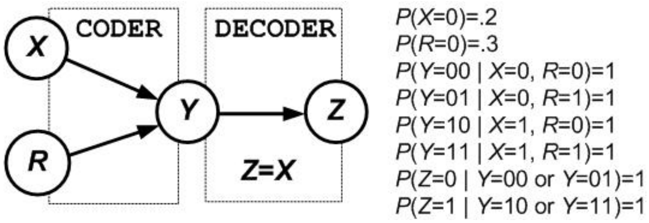

Figure 1.4. Causal model in which Z equals X (Taken from SGS [SGS93], Fig. 3.23)

The most-known example of unfaithfulness is when in the model of Fig. 1.2 A and D appear to be independent [SGS93]. This happens when the influences along the paths $A \to B \to E \to D$ and $A \to C \to D$ exactly balance, so that they cancel each other out and the net effect is an independence. The dependency of the paths forms a regularity that is not 'expected' by the causal model as explained by case (d) in the proof of theorem 1.3.

A similar phenomenon appears in the coder-decoder example shown by Fig. 1.4, taken from SGS (Fig. 3.23) [SGS93]. Variable Y encodes the values of both R and X, and Z decodes Y to match the value of X. X is therefore deterministically related to Z, though not adjacent. X is related to Z through Y. It however does not represent the shortest description of the corresponding distribution nor the model that would follow logically from observation. Z would be related to X and not to Y. It is impossible to learn the correct model from observations only. This is a general and inevitable problem of learning, since one must rely on the available data. A too simple model is learned if the data does not exhibit the total complexity of the system. Note that the exact correspondence of $X \to Y$ and $Y \to Z$ is not accidental, but intentionally build in by the engineer.

Variables in *pseudo-independent models* are pairwise independent but collectively dependent [XWC96]. Take for example three variables, X_1, X_2 and X_3, that are pairwise independent, but become dependent by conditioning on the third variable. Such distributions exhibit strict regularities. Yet, pseudo-independent models fit in the reductionist approach of causal models. We still can try to find out which variables are the causes and which the effects. Possibly, the model $X_1 \to X_3 \leftarrow X_2$ generates a pseudo-independent distribution if only the knowledge of X_1 and X_2 together says something about X_3. This is examined by case (a) in the proof of theorem 1.3.

A deterministic or functional relation between variable Y and set $\boldsymbol{X}$ reflects a non-random, thus atypical, conditional distribution, in which $P(Y \mid \boldsymbol{X})$ is 1 for exactly one state of $\boldsymbol{X}$ and 0 for the complement. Such relations among variables imply conditional independencies that cannot be represented by a faithful graph [SGS93]. $\boldsymbol{X}$ has all information about Y. If Y is related to another variable, they become independent by conditioning on $\boldsymbol{X}$. Take the model of Fig. 1.4. X and Z are related by a bijection. Either of them becomes independent of Y by conditioning on the other. We say that both variables contain *equivalent information* about Y. This cannot be represented by a faithful graph. In [LMMD06] we show that this is related to the violation of the *intersection condition*, one of the conditions that Pearl imposes on a distribution in the elaboration of causal theory and algorithms [Pea88]. The solution is to incorporate this information in an augmented causal model. We propose that in cases when several variables contain equivalent information about a target variable, we should connect the variable with the least complexity with the target [LMMD06]. This however does not solve the problem for the equality $Z = X$ in the model of Fig. 1.4 since the relation with other variables, like Y, will be identical. But by the equality, both variables behave completely identical and without background knowledge no distinction can be made between both variables. Finally, the definition of d-separation has to be extended to find the independencies that are implied by the deterministic relations [Gei90].

The *weak-transitivity condition* is also violated in the model of Fig. 1.4. R depends on Y and Y depends on Z, but R is independent of Z and does not become dependent by conditioning on Y (the latter would imply v-structure $R \rightarrow Y \leftarrow Z$). Its violation implies that the information that Y shares with R is independent of the information that Y shares with Z. Inspection of the CPDs show that R only influences the first bit of Y, but that Z is only determined by the second bit.

The previous examples provide no arguments against the causal interpretation of the graph. This, however, means that such regularities can coexist with causation. It can not be excluded that they appear in real data.

9 Conclusions

This paper confronted the principles of causal modeling theory, in which faithfulness plays an important role, with those of the two-part code, according to which the model should capture all the regularities of the data. We proved that without regularities in the DAG and the CPDs, a Bayesian network is faithful and minimal. The causal model corresponds to the edge-minimal Bayesian network. Unfaithfulness comes from a non-minimal factorization variable ordering or regularities ‘interfering’ with the independencies that follow from Markov. Additional regularities of a lower level can easily be added to the model, while for interfering regularities the model should be augmented to incorporate their effect on the independencies. But for other regularities, like patterns in a string, Bayesian networks do not offer a modeling framework that provides a minimal description. We argued that those regularities cannot be neglected.

On the other hand, the faithfulness property states that a model, the part of the two-part code containing the meaningful information, should represent all qualitative properties of the data. Faithfulness, together with minimality, guarantees that the model offers a canonical representation that is able to explain all observable regularities. This motivates the causal claim that the model represents the underlying physical mechanisms by which the data is generated. The theory of causal models is therefore audacious, it claims to know something about the inner by just observing the outer. A claim attacked by many. But isn't it a problem of all scientific research that tries to understand the world? A problem we therefore may not neglect.

BIBLIOGRAPHY

[BFGK96] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. In *Uncertainty in Artificial Intelligence*, pages 115–123, 1996.

[Car01] Nancy Cartwright. What is wrong with bayes nets? *The Monist*, pages 242–264, 2001.

[CD03] Joshua W. Comley and David L. Dowe. General bayesian networks and asymmetric languages. In *Proc. 2nd Hawaii International Conference on Statistics and Related Fields*, 2003.

[CD05] Joshua W. Comley and David L. Dowe. Minimum message length and generalised bayesian networks with asymmetric languages. In *In Advances in Minimum Description Length: Theory and Applications, P.D. Grünwald, I.J. Myung, and M.A. Pitt, Eds*. MIT Press, 2005.

[CT91] Thomas M. Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.

[FH99] D.A. Freedman and P. Humphreys. Are there algorithms that discover causal structure? *Synthese*, 121:2954, 1999.

[Gei90] D. Geiger. *Graphoids: A Qualitative Framework for Probabilistic Inference*. PhD thesis, University of California, Los Angeles, 1990.

[GTV01] P. Gacs, J. Tromp, and P. Vitanyi. Algorithmic statistics. *IEEE Trans. Inform. Theory*, 47(6):2443–2463, 2001.

[KP97] 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.

[LB94] W. Lam and F. Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 10:269–293, 1994.

[LMMD06] Jan Lemeire, Sam Maes, Stijn Meganck, and Erik Dirkx. The representation and learning of equivalent information in causal models. Technical Report IRIS-TR-0099, Vrije Universiteit Brussel, 2006.

[LV97] Ming Li and Paul Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 1997.

[OS90] Robert M. Oliver and James Q. Smith. *Influence Diagrams, Belief Nets and Decision Analysis*. Wiley, 1990.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, Morgan Kaufman Publishers, 1988.

[Pea00] J. Pearl. *Causality. Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[Ris78] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[Ris89] J. Rissanen. *Stochastic Complexity in Statistical Enquiry*. World Scientific, Singapore., 1989.

[SGS93] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Springer Verlag, 2nd edition, 1993.

[Spo01] Wolfgang Spohn. Bayesian nets are all there is to causal dependence. In *In Stochastic Causality, Maria Carla Galaviotti, Eds*. CSLI Lecture Notes, 2001.

[Stu01] 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.

[Vit02] 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.

[Vit05] P. Vitanyi. Algorithmic statistics and kolmogorov's structure functions. In *In Advances in Minimum Description Length: Theory and Applications, P.D. Grünwald, I.J. Myung, and M.A. Pitt, Eds.* MIT Press, 2005.

[Wal05] Chris S. Wallace. *Statistical and Inductive Inference by Minimum Message Length.* Springer, 2005.

[WD68] Chris S. Wallace and David L. Dowe. An information measure for classification. *Computer Journal*, 11(2):185–194, 1968.

[WD99] Chris S. Wallace and David L. Dowe. Minimum message length and kolmogorov complexity. *Computer Journal*, 42(4):270–283, 1999.

[Wil05] Jon Williamson. *Bayesian Nets And Causality: Philosophical And Computational Foundations*. Oxford University Press, 2005.

[XWC96] 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.

Jan Lemeire
jan.lemeire@vub.ac.be
Erik Dirkx
erik.dirkx@vub.ac.be
ETRO dept., Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussel, Belgium.