Noname manuscript No.
(will be inserted by the editor)

Replacing Causal Faithfulness with Algorithmic Independence of Conditionals.

Jan Lemeire · Dominik Janzing

Published in Minds and Machines, 2012, DOI 10.1007/s11023-012-9283-1.
The final publication is available at springerlink.com: http://www.springerlink.com/openurl.asp?id=doi:10.1007/s11023-012-9283-1

Abstract Independence of Conditionals (IC) has recently been proposed as a basic rule for causal structure learning. If a Bayesian network represents the causal structure, its Conditional Probability Distributions (CPDs) should be algorithmically independent. In this paper we compare IC with Causal Faithfulness (FF), stating that only those conditional independences hold true that are implied by the causal Markov condition. The latter is a basic postulate in common approaches to causal structure learning.

The common spirit of FF and IC is to reject causal graphs for which the joint distribution looks 'non-generic'. The difference lies in the notion of genericity: FF sometimes rejects models just because one of the CPDs is simple, for instance if the CPD describes a deterministic relation. IC does not behave in this undesirable way. It only rejects a model when there is a non-generic *relation* between different CPDs although each CPD looks generic when considered separately. Moreover, it detects relations between CPDs that cannot be captured by conditional independences. IC therefore helps in distinguishing causal graphs that induce the same conditional independences (i.e., they belong to the same Markov equivalence class).

The usual justification for FF implicitly assumes a prior that is a probability density on the parameter space. IC can be justified by Solomonoff's universal prior, assigning non-zero probability to those points in parameter space that have a finite description. In this way, it favours simple CPDs, and therefore respects Occam's razor.

Since Kolmogorov complexity is uncomputable, IC is not directly applicable in practice. We argue that it is nevertheless helpful, since it has already served as inspiration and justification for novel causal inference algorithms.

Jan Lemeire
Vrije Universiteit Brussel (VUB), ETRO Dept., Pleinlaan 2, B-1050 Brussels, Belgium
Interdisciplinary Institute for Broadband Technology (IBBT), FMI Dept., Gaston Crommenlaan 8 (box 102), B-9050 Ghent, Belgium
E-mail: jan.lemeire@vub.ac.be

Dominik Janzing
MPI for Intelligent Systems, Tübingen, Germany
E-mail: dominik.janzing@tuebingen.mpg.de

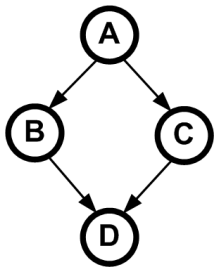

Fig. 1 Example causal model.

1 Introduction

Inferring causal relations by observation is at the heart of scientific reasoning. Although there is still some general scepticism about whether causal conclusions can be drawn from passive observations alone (e.g., without randomized interventions), the field of *causal inference* has developed postulates that link observations to causal statements and developed algorithms based on these postulates (Spirtes et al, 1993; Pearl, 2000). Here we compare two postulates, one that is widely accepted in the causal inference community and one that has been formulated by Janzing and Schölkopf (2010) after it had implicitly been stated by Lemeire and Dirkx (2006), and that we wish to advertise further. The main contribution of this paper lies in the comparison of the postulates.

Both postulates refer to the following popular scenario: given the variables $X_1, \ldots, X_n$, infer the causal relation after observing the joint distribution $P(X_1, \ldots, X_n)$. In practice, of course, one has only m points in $\mathbb{R}^n$ that are i.i.d. drawn from $P(X_1, \ldots, X_n)$, but the 'infinite sample limit' of knowing $P(X_1, \ldots, X_n)$ is often used to simplify the discussion. Following (Spirtes et al, 1993; Pearl, 2000), a causal structure is formalized by a Directed Acyclic Graph (DAG) with $X_1, \ldots, X_n$ as nodes. A basic postulate (which is also crucial and taken for granted in this article) is the causal Markov condition which states that a DAG G is only acceptable as a possible causal hypothesis if every node is statistically independent of its non-descendants conditioned on its parents. This is, up to a technical condition (Lauritzen, 1996), equivalent to saying that the Joint Probability Distribution (JPD) *factorizes* into the following conditional probabilities

$$P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P(X_j | Parents(X_j)) \,, \tag{1}$$

with $Parents(X_j) \subset \{X_1, \ldots, X_n\} \setminus X_j$ denoting the parents of variable X_j in G. The Conditional Probability Distributions (CPDs) $P(X_j|Parents(X_j))$ define the free parameters corresponding to G. For instance, the causal structure in Fig. 1 induces the factorization

$$P(A, B, C, D) = P(A)P(B|A)P(C|A)P(D|B, C) \,.$$

The DAG and the CPDs together define a *Bayesian network*. The causal Markov condition, however, is not sufficient to uniquely identify the causal DAG from the JPD because there are many different DAGs that render the JPD Markovian. In particular, every *complete DAG* (one having $\binom{n}{2}$ arrows) is consistent with any JPD. Therefore, additional postulates are required to select a smaller set from the huge set of DAGs that are allowed by the causal Markov condition.

Causal faithfulness (FF) (Spirtes et al, 1993) states that only those conditional independences that are imposed by the causal Markov condition hold. The idea is that additional independences are non-generic in the sense that they only occur for specific choices of CPDs, i.e., the free parameters of the Bayesian network. We challenge this reasoning and the implicit prior assumption on which it relies. We argue that a certain type of violation of faithfulness is quite likely to occur in nature.

In contrast, we propose the principle of algorithmic independence of CPDs, or Independence of Conditionals (IC) for short, saying that the shortest description of the JPD is given by the *causal* CPDs, i.e., the CPDs that occur in the factorization with respect to the causal DAG. Here description length is understood in the sense of Kolmogorov complexity. Like FF, IC also allows hypothetical causal DAGs to be ruled out on the basis that the observed JPD is 'non-generic' for the corresponding graphs. Despite this common spirit of both principles, the implications differ. We will explain why we trust IC more than FF, even though FF is more practical since IC relies on the uncomputable notion of Kolmogorov complexity. Since we are discussing first principles of causal inference, we cannot provide formal arguments in favour of IC, but we can describe examples which we hope will convince the reader that the implications of IC are more plausible than those of FF. We believe that IC, although it is not a practical method itself, is a useful basis for deriving new causal inference rules.

The paper is structured as follows. Sections 2 and 3 describe FF and IC, respectively, in detail. Section 4 explains the common idea behind both principles and describes distributions that are ruled out by both of them. Section 5 explores the differences between them, with respect to their justification and their implications. The examples where a factorization is accepted by IC but rejected by FF and vice versa are the crucial part of this paper (including the argument of why we think that IC yields more rational conclusions). Section 6 describes possible extensions of the fundamental idea of IC to causal models other than DAGs.

2 Faithfulness

We first recall faithfulness.

2.1 Formal statement

As already mentioned, every distribution that admits the factorization (1) with respect to a DAG G satisfies the following condition with respect to G:

Parental Markov condition: Every variable is conditionally independent of its non-descendants (except for itself), given its parents.

The Markov condition is a purely mathematical condition describing the relation between a DAG and a joint distribution of its nodes, while the *causal* Markov condition is a postulate that links statistical observations to causal semantics; it states hat only those DAGs can be causal hypotheses that render the observed distributions Markovian.

It can be shown that the conditional independences that are explicitly stated by the parental Markov condition imply other conditional independences. To describe the set of all independences that follow from Eq. (1), we need to introduce some notation and terminology. The ternary operator $.\perp\!\!\!\perp. \mid .$ denotes the statistical independence of the first two operands when conditioned on the third operand. Single random variables are denoted by

capital letters and sets of variables by boldface capital letters. Next we introduce the concept of d-separation (Lauritzen, 1996):

d-separation: A path[1] is said to be blocked by $\boldsymbol{Z}$ if it contains a collider $\rightarrow \cdot \leftarrow$ whose descendants are not in $\boldsymbol{Z}$ or a non-collider $\rightarrow \cdot \rightarrow$ or $\leftarrow \cdot \rightarrow$ or $\leftarrow \cdot \leftarrow$ that is in $\boldsymbol{Z}$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$ if every path between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is blocked by $\boldsymbol{Z}$. d-separation is denoted by the ternary operator $. \perp . \mid .$:

$$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} .$$

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not d-separated by $\boldsymbol{Z}$, we call them d-connected by $\boldsymbol{Z}$. Then we have:

Global Markov condition: For any disjoint sets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ of variables for which $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$, we have

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} .$$

The global Markov condition is equivalent to the parental one (Lauritzen, 1996). Now we are able to formally state FF:

Definition 1 (faithfulness) *A JPD is called faithful with respect to a DAG G if for all disjoint subsets $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$*

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z} \quad \Leftrightarrow \quad \boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} .$$

Causal faithfulness is the postulate that every JPD is faithful with respect to the true causal DAG.

In other words, a JPD and a DAG are faithful to one another if all and only the conditional independencies true in the JPD are entailed by the Markov condition applied to the DAG. Note that faithfulness is a purely mathematical condition that describes the relation between any JPD and any DAG, while *causal* faithfulness links *the observed probability distribution* to the *true* causal DAG. Whenever this causes no confusion, we will use the term faithfulness and the abbreviation FF also for causal faithfulness.

The set of DAGs for which the observed JPD is faithful is called the Markov equivalence class. The idea of conditional independence based causal inference algorithms is to output the Markov equivalence class as the set of possible causal hypotheses (Spirtes et al, 1993).

2.2 Justification for FF

The faithfulness condition can be thought of as the assumption that conditional independencies are due to causal structure rather than to 'accidents' of parameter values. This is motivated by the following reasoning. For variables with a finite range, there is a canonical parameterization of CPDs by describing $P(x_j|pa_j)$ for all possible values x_j of X_j and all possible values pa_j of $Parents(X_j)$. For instance, if X_j is a binary variable having k binary parents, every CPD $P(X_j|Parents(X_j))$ is defined by a point in a $(2^k - 1)$-dimensional subset of $[0,1]^{2^k}$ (specifying the probabilities $P(X_j = 1|pa_j)$ for each of the 2^k possible values of the parents). Depending on the number of parents and their ranges, every CPD $P(X_j|Parents(X_j))$ is thus specified by a point in some compact subset $S_j \subset \mathbb{R}^{m_j}$. The entire JPD is thus determined by specifying a point in the Cartesian product $S := \times_{j=1}^{n} S_j$. It has been shown that the subset of points that yield an unfaithful distribution has zero volume, i.e., it has Lebesgue measure zero (Meek, 1995). The conclusion

[1] A path is a set of consecutive edges (independent of the direction) that do not visit a vertex more than once.

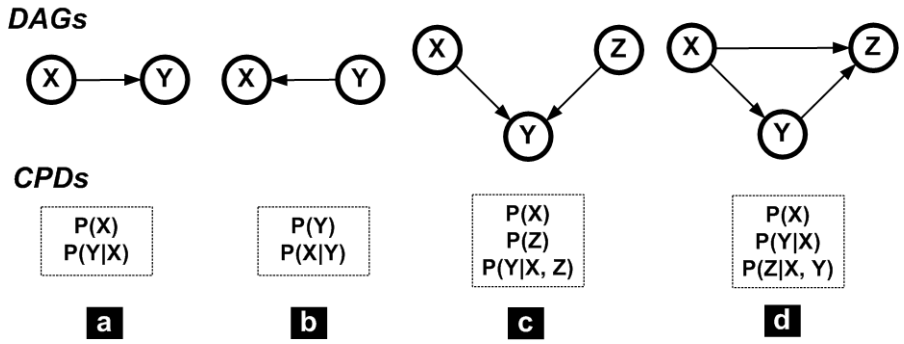

Fig. 2 Four example Bayesian networks: (a) and (b) represent $P(X, Y)$, and (c) and (d) represent $P(X, Y, Z)$.

that unfaithfulness therefore happens with zero probability is correct if one assumes that the points in S_j are chosen according to some probability *density* on S (which is, by definition, equivalent to saying that sets of Lebesgue measure zero have zero probability).

2.3 Minimality

For reasons that will become clear, we also mention a condition that is weaker than faithfulness:

Definition 2 (MIN) ***Minimality of a factorization*** *A DAG G is minimal for a JPD if the JPD is Markovian and if for every node X_i the following condition holds: for every parent p of X_i*

$$X_i \not\perp\!\!\!\perp p \mid Parents(X_i) \setminus p \tag{2}$$

The ternary operator $.\not\perp\!\!\!\perp. \mid .$ denotes statistical dependence of the first two operands when conditioned on the third operand. It is clear that for faithful distributions the DAG is minimal because no parent of X_j is d-separated from X_j by the other parents. To see that minimality is strictly weaker than faithfulness, we observe that for every variable ordering $X_1, \ldots, X_n$ there is a DAG that renders the JPD minimal. We start with the *complete DAG* containing all edges (X_i, X_j) for $i < j$. For every X_j, we remove edges pointing to X_j until every remaining parent is dependent on X_j, given the other parents of X_j.

For two dependent variables X and Y, the DAGs (a) and (b) in Fig. 2 are both minimal and they render the JPD faithful. For a case where minimality holds but faithfulness is violated, we consider a JPD generated by (c). In the generic case, we obtain $X \perp\!\!\!\perp Y$ as the only independence. Although the complete DAG shown in (d) does not render the JPD faithful because $X \perp\!\!\!\perp Y$ is not implied by this DAG, it is still minimal: one can easily check that none of the arrows can be removed without violating the Markov condition. These examples show that Def. 2 (MIN) defines a *local* minimality criterion for the number of arrows. FF can be regarded as a criterion for a *global* minimum, since it has been proven that if a faithful DAG exists, it is the Markovian DAG with the least number of edges (Lemeire et al, 2011b).

In the context of minimality, we should briefly discuss the intention of causal inference again. We have actually stated that we are interested in identifying the 'true causal DAG', but there is also a reasonable way to weaken this purpose: assume we have identified a DAG

that coincides with the true one except for containing additional arrows. We should ask whether this error is harmful. On the one hand, it may be confusing if the list of direct causes of a variable contains variables whose influence is zero (as opposed to the assumptions made by Zhang and Spirtes (2011)). On the other hand, the extended DAG together with the joint distribution still contains the information on which arrows are true ones, because each $P(X_j|Parents(X_j))$ shows which of the potential parents really influence X_j; the 'extended' causal Bayesian network makes the same predictions regarding how variables change under interventions. The semantics of the extended DAG G is no longer to represent the true causal structure but a set of possible causal structures, namely all DAGs that can be obtained by removing edges from G. Causal inference algorithms that output such a DAG G are still helpful when augmented by a statistical testing procedure that removes the additional edges. Since the latter is a purely statistical problem, one could argue that the actual causal problem consists in the first step. Sometimes it may be convenient to work with complete DAGs and leave it to the description of the CPDs to specify which direct influences are zero. For instance, if the order $X_1, \ldots, X_n$ corresponds to the time order of observations, it is quite natural to connect each X_i with every X_j corresponding to a later time instance ($j > i$) because they are *potential* effects.

3 Independence of conditionals (IC)

IC postulates that the set of CPDs corresponding to a causal DAG is generic in the sense that no CPD contains information about another one. Here information is understood in the sense of algorithmic information which is defined in terms of Kolmogorov complexity. We first introduce the basic concepts.

3.1 Introduction to algorithmic information theory

For a binary string $s \in \{0,1\}^*$ the algorithmic information $K(s)$ (or 'Kolmogorov complexity') is defined as the length of the shortest program on a universal prefix-free Turing machine that generates s and then stops (Solomonoff, 1960; Kolmogorov, 1965; Chaitin, 1966, 1975). Prefix-free means that the program has to be given with respect to an encoding where no allowed program code is the prefix of another one. Thus, the program does not require an extra symbol indicating its end. This fact has several important implications, for instance for the definition of the universal prior (see Sec. 5.1).

Two strings s, t are called algorithmically dependent whenever compressing them jointly is more economical than compressing them independently. The algorithmic mutual information is defined as (Chaitin, 1975)

$$I(s:t) := K(s) + K(t) - K(s,t)\,,$$

where the pair (s,t) is implicitly identified with a single string via some given standard bijection between $\{0,1\}^* \times \{0,1\}^*$ and $\{0,1\}^*$. We also need the additivity rule for the joint Kolmogorov complexity (Chaitin, 1975):

$$K(s,t) \stackrel{+}{=} K(t) + K(s|t^*)\,, \tag{3}$$

where $K(s|t^*)$ denotes the conditional Kolmogorov complexity of s, given the shortest compression t^* of t. As usual in algorithmic information theory, $\stackrel{+}{=}$ denotes equality up to a constant that is independent of the string s, but does depend on the Turing machine.

The above definition of independence generalizes in a straightforward way to the joint independence of n strings:

Definition 1 (Algorithmic Independence)
Binary strings $s_1 \dots s_n$ are algorithmically independent if

$$K(s_1, \dots, s_n) \stackrel{+}{=} \sum_i^n K(s_i)\,. \tag{4}$$

Note that here and throughout the paper we consider the number n of strings as a constant. Accordingly, in the following the number n of nodes will also be considered as a constant.

Here we need to comment on how to interpret the sign $\stackrel{+}{=}$. For a fixed set of strings $s_1, \dots, s_n$, the term 'up to a constant' does not make sense, it only acquires a meaning in theoretical statements such as 'if $s_1, \dots, s_n$ satisfy a certain condition, then they are algorithmically independent', because every s_j then plays the role of a string valued variable. For fixed strings, we have to interpret $\stackrel{+}{=}$ in the sense of 'up to a small number' without further specifying what 'small' means. This arbitrariness in setting a threshold is similar to the freedom of choosing the significance level in a statistical dependence test.

According to the following postulate (which is a straightforward generalization of Lemma 5 in Janzing and Schölkopf (2010) to more than two objects) every algorithmic dependence requires a causal explanation.

Causal principle: If $s_1, \dots, s_n$ are binary words that describe n objects in nature and

$$K(s_1, \dots, s_n) \ll \sum_{j=1}^n K(s_j),,$$

then at least some of the n objects are causally related.

We will see that this principle is also the basis for using IC as causal inference rule.

3.2 Formal statement of the IC condition

Since we are interested in the algorithmic dependence of CPDs, the strings in Eq. 4 are now the descriptions of the CPDs.

Definition 2 (Independence of Conditionals)
The conditional probability densities $CPD_1, \dots, CPD_n$ corresponding to a DAG G with n nodes are said to satisfy Algorithmic Independence of Conditionals, or Independence of Conditionals (IC) for short, if

$$K(CPD_1, \dots, CPD_n) \stackrel{+}{=} \sum_{i=1}^n K(CPD_i)\,, \tag{5}$$

where $K(CPD_i)$ for $CPD_i = P(X_i|Parents(X_i))$ is the length of the shortest program that computes the probability or probability density $P(x_i|pa_i)$ from the input (x_i, pa_i), with x_i and pa_i outcomes of X_i and $Parents(X_i)$ respectively.

By 'causal IC condition' we mean the postulate that IC holds for the observed JPD whenever the true causal structure can be described by the DAG G and the CPDs correspond to mechanisms that are independently 'designed' by nature.

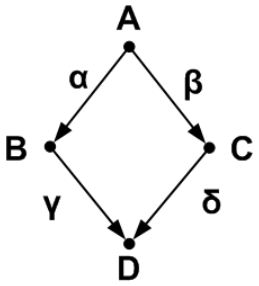

Fig. 3 Causal model with linear influences described by parameters $\alpha, \beta, \gamma, \delta$.

Like the distinction between faithfulness and causal faithfulness, causal IC is a postulate for causal reasoning whereas IC is a purely mathematical relation between a DAG G and a JPD. Whenever this causes no confusion, we will simply use IC for 'causal IC'.

Note that Definition 2 implicitly assumes that the set of parents of X_i is already known, but this amount of information is in $O(1)$ since we assume n to be a constant. Moreover, it is assumed that the JPD, and hence every CPD, has a finite description. If the variables have a finite domain, all rational probability values are allowed; irrational values are only possible if there is a finite rule defining them. For infinite domains, the finiteness of the description is certainly a stronger restriction. One could argue that the 'true' probability values will not satisfy this condition since generic real values are uncomputable, that is, they have an infinite description. The computable values define a set of Lebesgue measure zero in parameter space. However, we argue that, regardless of what the 'true' values are, science always works with finite descriptions[2].

We will later also use the following equivalent formulation of IC. Since all CPDs together describe the joint distribution (JPD) and, conversely, every CPD can be computed from the JPD if the DAG is given (which is, again, only a constant amount of information), the left hand side of (5) is always equal to $K(JPD)$. In other words, neglecting the information required to describe the DAG, IC holds if and only if the shortest description of the joint distribution is given by separate descriptions of the conditionals (Janzing and Schölkopf, 2010):

$$K(P(X_1, \ldots, X_n)) \stackrel{+}{=} \sum_{i=1}^{n} K(P(X_i|Parents(X_i)))\,. \tag{6}$$

4 Similarities between FF and IC

We first discuss the similarities between both principles.

4.1 Common spirit: rejecting non-generic adjustments

Although IC and FF sound like completely different inference principles, the common idea is to reject causal structures for which the CPDs satisfy non-generic relations. To describe this,

[2] Note that model selection procedures that are based on the minimum description length principle automatically define a probability distribution having finite description length (Grünwald, 2007).

consider the DAG in Fig. 3 and consider the case where A and D are independent because the influence via B compensates for the influence via C. To simplify the mathematics, we assume that all CPDs are given by linear structure equations:

$$B = \alpha A + U_B \,, \tag{7}$$

$$C = \beta A + U_C \,, \tag{8}$$

$$D = \gamma B + \delta C + U_D \,, \tag{9}$$

where U_B, U_C and U_D are unobserved disturbances or 'noise' terms that are jointly statistically independent and independent of A. Then the two influences of A on D cancel for

$$\alpha\gamma = -\beta\delta \,, \tag{10}$$

which is an unlikely coincidence if all real-valued parameters are chosen independently (according to some continuous distribution on $\mathbb{R}$). Obviously, FF rejects the causal DAG in Fig. 3 because A and D are not d-separated by the empty set and thus should not be independent.

To see that the DAG is also rejected by IC, we observe that the CPDs $P(B|A)$, $P(C|A)$, and $P(D|B,C)$ are described by the structure coefficients α, β, γ and δ as well as the distributions of the noise variables U_B, U_C, U_D, respectively. Describing the JPD by separate descriptions of $P(A), P(B|A), P(C|A), P(D|B,C)$ is therefore redundant because the parameter γ in $P(D|B,C)$ can be computed from the other CPDs via Eq. 10. This way of reasoning is made precise and more general by the following theorem.

Theorem 3 *For a given DAG G, let the set of possible CPDs $P(X_j|Parents(X_j))$ be parameterized by some set $S_j := \{\lambda_1^j, \ldots, \lambda_{k_j}^j\}$ of parameters. Assume that the parameter values for some specific choice $CPD_1, \ldots, CPD_n$ of conditional probality densities satisfy a functional relation in the sense that $\theta_1 = f(\theta_2, \ldots, \theta_k)$, where f is some function and $\theta_1, \ldots, \theta_k$ are parameters taken from at least two different sets S_j. Assume furthermore that θ_1 corresponds to CPD_1 (without loss of generality). Then the following condition implies violation of IC:*

$$K(f) \stackrel{+}{<} K(\theta_1 | CPD_1^{\setminus\theta_1,*}) \,, \tag{11}$$

where $CPD_1^{\setminus\theta_1}$ denotes the parameters of CPD_1 without θ_1 (recall that the asterisk denotes the shortest compression).

Note that we do not assume that the set of possible parameter combinations $(\lambda_1^j, \ldots, \lambda_{k_j}^j)$ is a cartesian product of the range of every single parameter λ_i^j. Therefore, knowing the other parameters of CPD_1 could reduce the set of possible θ_1, but Eq. (11) ensures that the description of θ_1 still requires non-negligible length.

Proof Since θ_1 can be computed from the other parameters using f we have

$$K(\theta_1 | (f, \theta_2, \ldots, \theta_k)^*) \stackrel{+}{=} 0 \,,$$

which implies

$$K(\theta_1 | (\theta_2, \ldots, \theta_k)^*) \stackrel{+}{\leq} K(f) \,. \tag{12}$$

This is due to the general rule

$$K(a|c^*) \stackrel{+}{\leq} K(a,b|c^*) \stackrel{+}{=} K(b|c^*) + K(a|(b,c)^*) \stackrel{+}{\leq} K(b) + K(a|(b,c)^*) \,,$$

where the second equality is due to $K(x,y) = K(x) + K(y|x^*)$ (Gacs et al, 2001). From Eq. 11 and Eq. 12 it follows that:

$$K(\theta_1|\theta_2,\ldots,\theta_k) \stackrel{+}{<} K(\theta_1|CPD_1^{\backslash\theta_1,*}) \tag{13}$$

Violation of IC can now be derived as follows. We first use Eq. 3. Further comments on the derivation are given after the equations.

$$\begin{aligned}
&K(CPD_1,\ldots,CPD_n) && (14)\\
&\stackrel{+}{=} K(CPD_1|(CPD_2,\ldots,CPD_n)^*) + K(CPD_2,\ldots,CPD_n) && (15)\\
&\stackrel{+}{\leq} K(CPD_1|(CPD_2,\ldots,CPD_n)^*) + \sum_{i=2}^{n} K(CPD_i) && (16)\\
&\stackrel{+}{=} K(CPD_1^{\backslash\theta_1}|(CPD_2,\ldots,CPD_n)^*) && (17)\\
&\quad + K(\theta_1|(CPD_1^{\backslash\theta_1},CPD_2,\ldots,CPD_n)^*) + \sum_{i=2}^{n} K(CPD_i) && (18)\\
&\stackrel{+}{\leq} K(CPD_1^{\backslash\theta_1}) + K(\theta_1|(\theta_2,\ldots,\theta_k)^*) + \sum_{i=2}^{n} K(CPD_i) && (19)\\
&\stackrel{+}{<} K(CPD_1^{\backslash\theta_1}) + K(\theta_1|(CPD_1^{\backslash\theta_1})^*) + \sum_{i=2}^{n} K(CPD_i) && (20)\\
&\stackrel{+}{=} K(CPD_1) + \sum_{i=2}^{n} K(CPD_i) && (21)
\end{aligned}$$

For Ineq. 17 we separate θ_1 from the remaining parameters of CPD_1. Ineq. 19 follows because we drop some information from the conditioning set. For Ineq. 20 we use Eq. 13. For Eq. 21, we recombine the description of CPD_1 (using Eq. 3). □

Due to condition (11), the theorem only states significant violation of IC if the function is significantly simpler than the information that the parameter θ_1 contains about CPD_1. If θ_1 has a simple description, like $\theta_1 = 0.5$ or $\theta_1 = \pi$, for instance, this is not the case. This already suggests the following general difference between IC and FF: while FF can already be violated if some parameters of a single CPD are non-generic, IC only excludes the case where the parameter vectors of different CPDs are related by a simple rule, although they are complex themselves. Sec. 5 will elaborate on this difference.

4.2 Both FF and IC are sanity checks of the model class

IC and FF are not only principles for selecting the best model from the set of all possible DAGs, they can also be used to check whether the true causal structure can be represented by a DAG at all. If no factorization of the JPD satisfies FF or IC, this might indicate that the system's structure must be described by a different type of model.

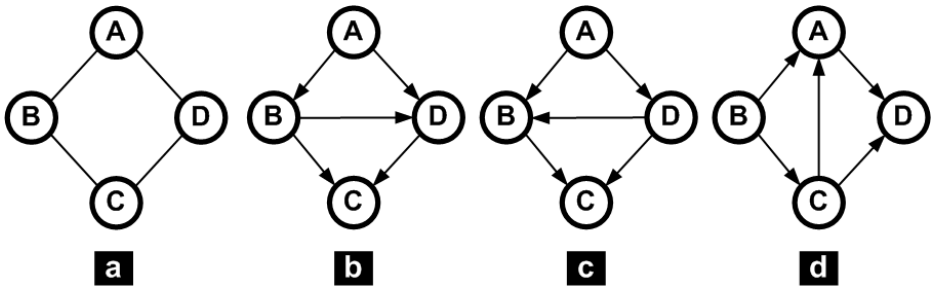

Fig. 4 A Markov network (a) and three Bayesian networks describing the same system (b-d).

As an example, consider four variables, A, B, C and D that influence each other via the following cyclic ('non-recursive') linear equations:

$$\begin{aligned} B &= \beta A + E_B \\ C &= \gamma B + E_C \\ D &= \delta C + E_D \\ A &= \alpha D + E_A \,, \end{aligned}$$

where the noise variables E_A, E_B, E_C, E_D are jointly statistically independent. Assume a dynamical evolution where the noise terms remain constant and A, B, C, D are updated according to these equations (this scenario has been referred to as 'deterministic equilibrium' by Lauritzen and Richardson (2002)). Whenever $|\alpha\beta\gamma\delta| < 1$, one can show that the unique stationary distribution is given by the structure equation

$$\begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -\alpha \\ -\beta & 1 & 0 & 0 \\ 0 & -\gamma & 1 & 0 \\ 0 & 0 & -\delta & 1 \end{pmatrix}^{-1} \begin{pmatrix} E_A \\ E_B \\ E_C \\ E_D \end{pmatrix} = \Gamma^{-1} \begin{pmatrix} E_A \\ E_B \\ E_C \\ E_D \end{pmatrix} ,$$

where the last equation holds by definition of the matrix Γ. To derive the conditional independences induced by these equations, we first consider the cross covariance matrix of (A, B, C, D). It is given by

$$\Sigma = \Gamma^{-1} \Sigma_{EE} \Gamma^{-T} \,,$$

where Γ^{-T} denotes the transpose of the inverse of Γ and Σ_{EE} the cross-covariance matrix of the noise variables $(E_A, \ldots, E_D)$. We then obtain $\Sigma^{-1} = \Gamma \Sigma_{EE}^{-1} \Gamma^T$. By assumption, Σ_{EE} is diagonal, and hence also Σ_{EE}^{-1}. One can easily check that Σ^{-1} therefore has zeros at positions that relate the variables A and C, meaning that they are conditionally independent given all the remaining variables, namely B and D. Likewise, we have zeros at positions relating B and D. We thus obtain the conditional independences

$$\begin{aligned} A &\perp\!\!\!\perp C \,|\{B, D\} \\ B &\perp\!\!\!\perp D \,|\{A, C\} \,. \end{aligned}$$

There is no DAG that faithfully represents these relations.

Fig. 4 (b) to (d) show examples of DAGs that are allowed by the Markov condition but violate FF since no DAG can represent both independences while also obeying Markov. However, the independences can be represented by a different type of model, namely Markov

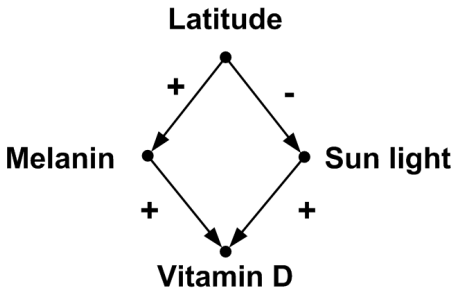

Fig. 5 Example of a meta-mechanism (evolution) which makes the latitude at which a person lives independent of its amount of vitamin D.

networks. A Markov network is an undirected graph where two nodes are adjacent if and only if they are conditionally dependent, given all other nodes. The above independences correspond to the Markov network in 4 (a).

As a reviewer correctly commented, the Markov network cannot be considered a description of the system's *causal* structure. The causal structure of the generative process should capture the system dynamics leading to the above fixpoint, as is done by dynamic Bayesian networks for instance. But the Markov network captures the relational structure among the variables better than a DAG does.

The fact that no DAG on these four variables represents the dependences faithfully correctly suggests to the observer that the underlying causal structure cannot be described by any DAG. Likewise, Theorem 3 shows that IC is violated if the CPDs satisfy atypical relations that induce additional conditional independences, provided that the CPDs themselves are sufficiently complex.

To give a second example apart from cyclic causal structures, note that also latent common causes may lead to a violation of FF and IC. So, both FF and IC can also be used as a sanity check of the model class under consideration.

4.3 Limitations of causal IC and causal FF due to meta-mechanisms

Meta-mechanisms are mechanisms that determine other mechanisms, they set the parameters of other mechanisms. In the case of Bayesian networks, meta-mechanisms might govern several CPDs and adjust their values, which may result in an atypical parameter configuration and hence a possible violation of IC or of FF. This is illustrated by the following example.

Consider people living at different latitudes and the amount of vitamin D created by the body, as shown in Fig. 5. 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 the necessary amount of vitamin D (`http://en.wikipedia.org/wiki/Melanin`). To ensure this, evolution has given humans different amounts of melanin, which is reflected by skin colour, relative to the amount of sun light to which they are exposed. The skin colour is mainly affected by latitude. This results in a nearly constant level of vitamin D creation that is independent of the latitude at which we live. If the influences can be linearly approximated, Eq. (10) describes the constraint needed to arrive at the independence. Here, the *meta-mechanism* is biological evolution, which develops the mechanisms in a joint process

rather than that each mechanism $P(X_j \mid Parents(X_j))$ has been 'chosen' independently. Evolution has controlled the $Latitude \to Melanin$ relation such that the parameters calibrated until the influences of $Latitude$ on $Vitamin\ D$ were neutralized, and hence there is unfaithfulness. Assuming that unfaithfulness is due to Eq. (10), Theorem 3 describes the conditions under which we also obtain a violation of IC. A violation of IC and FF gives an indication for the presence of a meta-mechanism.

Another meta-mechanism explaining specific parameterizations is a 'designer'. One can think of a system built by an engineer who deliberately tunes the system in such way that some related variables become independent. The causal graph is then unfaithful to the system, but the unfaithfulness points to an interesting fact, namely that there is a plan behind the system; the system is intentionally unfaithful in order to meet certain requirements. Meta-mechanisms were already considered by Korb and Nyberg (2006) to explain violations of faithfulness. The presence of meta-mechanisms in socio-economic and medical cases is also reported by Cartwright (1999).

When postulating that IC holds for the true causal DAG, we assume that the CPDs are independently 'designed', which explicitly excludes meta-mechanisms. Observing algorithmic dependences between the CPDs either shows that the corresponding DAG is not the causal one or that some meta-mechanism made the CPDs dependent. Note (as already pointed out by Janzing and Schölkopf (2010)) that this reasoning is consistent with the 'causal principle' mentioned at the end of Sec. 3.1. Although this may sound a bit unfamiliar, we now consider each CPD as an 'object'. Assume, for instance, that $X \to Y$ is the true causal DAG. Then $P(Y|X)$ describes the causal mechanism, a 'machine' that generates y-values according to $P(Y|x)$ if the input is x. Likewise, $P(X)$ corresponds to a 'machine' that generates x-values. Observing dependences between $P(X)$ and $P(Y|X)$ shows a causal relation between the two machines.

5 Differences between FF and IC

This section is devoted to the discussion of the differences between FF and IC. We start by describing the different foundations of FF and IC. Then, we discuss their differences in causal inference. Given a JPD, the conditions FF and IC accept or refute factorizations as valid causal hypotheses. This is illustrated by Fig. 6. Sec. 5.2 provides examples for causal hypotheses that are rejected by FF and accepted by IC. We argue that we consider the acceptance by IC to be more rational. Sec. 5.3 shows how causal inference can be done based on IC in structures rejected by FF. Sec. 5.4 describes examples that are accepted by FF but rejected by IC. We explain why we also trust IC in these cases, i.e., that IC correctly rules out the factorizations as valid causal hypotheses. This shows that IC sometimes rules out some of the hypotheses in a Markov equivalence class. Sec. 5.5 describes a case rejected by FF and accepted by IC, where both answers are right in an appropriate sense. The overall goal of this section is to show that the implications of IC are more convincing than those of faithfulness, with respect to those DAGs that are accepted as well as those that are rejected. A blind reviewer correctly pointed out that IC and FF cannot be strictly ordered in terms of strength. We argue, instead, that IC is more convincing than FF since it yields more plausible conclusions in all our examples. Since we are talking about first principles, there cannot be a formal argument about which one is better or more fundamental.

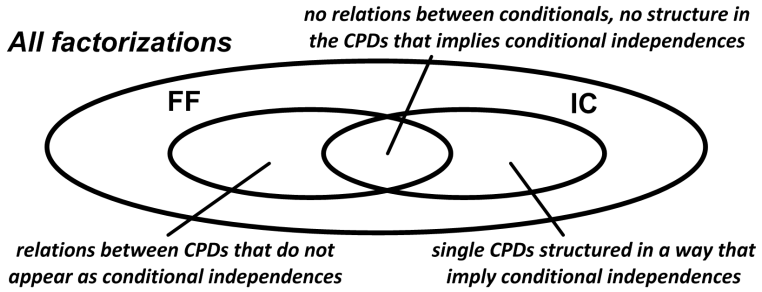

Fig. 6 The subsets FF and IC denote the factorizations for which FF and IC, respectively, hold for a given JPD.

5.1 Different foundations: uniform prior versus Solomonoff prior

Faithfulness can be justified by a generative model where first the causal DAG on the given variables $X_1, \ldots, X_n$ is randomly drawn (from some probability distribution over all DAGs with n nodes) and then each CPD is randomly drawn from a uniform distribution on the set of CPDs corresponding to the respective node and its parent set (here, the 'uniform' distribution is defined as the Lebesgue measure with respect to the canonical parameterization mentioned in Sec. 2.2). Then, unfaithful distributions occur with probability zero because they define a set of Lebesgue measure zero in parameter space. The same argument holds true if the parameters are drawn from some prior distribution that has a *density* in parameter space.

Like FF, IC can also be justified by a model where the parameter vector for each CPD is independently drawn, but not from a density in parameter space. Instead, IC uses Solomonoff's prior (Solomonoff, 1964), which we now introduce. The idea is to initialize the input band of a Turing machine by randomly setting each bit to 0 or 1 with equal probability. This random process actually generates infinite random strings, but some of them correspond to programs of finite length because the prefix-free encoding tells the Turing machine where to stop reading. Some of these programs halt and thus generate an output string s. The probability of obtaining s as output, given that the random input encodes a valid program that eventually stops, is known as Solomonoff's prior. It is also given by Levin (1974)

$$Pr(s) \stackrel{\times}{=} 2^{-K(s)} ,$$

where $\stackrel{\times}{=}$ denotes a multiplicative constant that we ignore in the sequel. Solomonoff's prior is an elegant implementation of Occam's Razor since it favours simple structures. Strings with high regularities like 10101010101010 are considered more likely than generic (random) strings of the same length. This is because the structure is not interpreted as a rare coincidence. Instead, it is generated from a *short* input via a simple process. It has been shown that Bayesian inference using Solomonoff's prior is able to learn every structure underlying the observations (Solomonoff, 1964), which has also been used as a basis for so-called universal artificial intelligence (Hutter, 2007).

Now we describe how to apply Solomonoff's prior for causal inference. Assume we observe that the JPD is Markovian with respect to the DAG G and we are supposed to infer whether G is a plausible causal hypothesis or not. Let CPD_j with $j = 1, \ldots, n$ be the conditionals corresponding to G and θ_j be a binary string encoding the parameter

vector describing CPD_j. By slightly abusing notation, let $K(CPD_j) := K(\theta_j)$ be the Kolmogorov complexity of θ_j.

If G is the true causal DAG, we assume that each CPD is independently drawn from the prior probability distribution

$$Pr(CPD_j) := \frac{1}{Z_j} 2^{-K(CPD_j)} ,$$

with

$$Z_j := \sum_{CPD'_j} 2^{-K(CPD'_j)} ,$$

where the sum runs over all binary words CPD'_j that encode a possible conditional probability distribution $P'(X_j|Parents(X_j))$. Hence, the probability of obtaining a certain JPD is given by

$$Pr(JPD) \stackrel{\times}{=} \prod_{j=1}^{n} 2^{-K(CPD_j)} .$$

On the other hand, if we do not have any information about the causal structure of the process that generated the JPD, we use the prior probability

$$Pr(JPD) \stackrel{\times}{=} 2^{-K(JPD)} ,$$

where $K(JPD)$ denotes the Kolmogorov complexity of the string that encodes JPD, given an appropriate parameterization of the set of joint distributions. This is just Solomonoff's prior applied to the set of all JPDs. Note that

$$K(JPD) \stackrel{\times}{=} K(CPD_1, \ldots, CPD_n) ,$$

because describing the JPD is equivalent to describing all CPDs. Here, we have implicitly assumed that the parameterizations of the set of all JPDs is compatible with the Cartesian product of all parameterizations of the CPDs in the sense that they can be translated into each other by a rule of negligible Kolmogorov complexity. Whenever

$$K(CPD_1, \ldots, CPD_n) \ll \sum_{j=1}^{n} K(CPD_j) ,$$

the independent generation of the CPDs is significantly less likely than a process that generated the JPD jointly. Thus, G is rejected provided that we exclude meta-mechanisms, see the discussion in Subsec. 4.3.

The following subsections provide examples that suggest that Solomonoff's prior is more plausible than the uniform one.

5.2 FF rejects many structured CPDs

This section argues that not all violations of faithfulness are unlikely. To provide the intuition, first consider again the DAG in Fig. 3 and assume $U_B = 0$ for the noise in the structural equations of Eq. 7. This implies that B is a function of A, that is, $B = f(A)$. It follows that $B \perp\!\!\!\perp D|A$ and $B \perp\!\!\!\perp C|A$ since A has all information about B. Neither of

the independencies follow from the Markov condition and both render the distribution unfaithful. To simplify the discussion, assume that the noise is a Gaussian distribution with zero mean and standard deviation σ. The conditional $P(B|A)$ is then described by the pair $(\alpha, \sigma) \in \mathbb{R} \times \mathbb{R}_0^+$, where the noiseless case corresponds to the line $\mathbb{R} \times \{0\}$, which has Lebesgue measure zero. Nevertheless, noiseless relations do occur in nature (at least ones with negligible noise). We think a higher than zero probability should be assigned to the case $\sigma = 0$ because *this is a special case*. To show that the 'Lebesgue measure zero argument' is unconvincing, we just have to observe that the set of all linear models has zero measure in the set of all possible probability distributions, after we have parameterized the latter in an appropriate way. Likewise, the set of Gaussian distributions has zero measure in the set of all distributions. Nevertheless, such kinds of special (conditional or marginal) distributions fit many observations in nature. To give a further example, let us parameterize the set of distributions of a binary variable by $\lambda \in [0, 1]$. Then we want to assign non-zero probability to special cases like $\lambda = 0$, $\lambda = 0.5$, $\lambda = 1$ because determinism is possible. Likewise, unbiased coins should not be excluded either.

One may argue that our arguments are spoiled by the fact that no mechanism in nature is *exactly* deterministic, no distribution is exactly Gaussian, and so on, which would justify assigning zero prior probability to these cases. However, knowing that every mathematical concept can at most be a useful approximation of nature, we prefer to use a prior like Solomonoff's that assigns non-zero probability to special points in parameter space (like $\lambda = 0$) instead of designing a complicated density that is peaked around this value. For the special case of the distribution of a binary variable, it has been argued by Rathmanner and Hutter (2011) why densities on the parameter values λ do not provide reasonable inferences and that Solomonoff's prior works better.

To show also that causal inference benefits from the preference of simple CPDs, consider deterministic relations. Consider, for instance, the causal DAG $X \to Y \to Z$ and assume that $Y = f(X)$ for some function f. We observe

$$Y \perp\!\!\!\perp Z \,|X\,,$$

which violates faithfulness. The reason that no DAG satisfies FF is that the CPD $P(Y|X)$ itself is 'non-generic' (Lemeire et al, 2011a), but there need not to be a non-generic *relation* between the CPDs.

5.3 Causal inference in unfaithful structures

IC, on the other hand, has the advantage that not only does it accept such a deterministic model, it also helps to identify the direct causes and causal directions in this regime. In Lemeire et al (2011a) we show that the model $Y \leftarrow X \to Z$ is equivalent to the correct model $X \to Y \to Z$ from the point of view of conditional independencies (although none of the models is faithful). The identification of the correct model is based on the principle that $P(Z|X)$ is defined by $Y = f(X)$ and $P(Z|Y)$. If $f(X)$ and $P(Z|Y)$ are set independently and causal IC holds, then $P(Z|Y)$ will in most cases be simpler than $P(Z|X)$ since the latter contains the complexity of f on top of the complexity of $P(Z|X)$. Hence, the simplest CPD can be regarded as the causal CPD. This method allows the identification of the correct direct cause of Z.

Next, for the above example $X \to Y \to Z$, IC can sometimes tell us the causal direction. This is shown by the following toy example, where we restrict the attention to the relation between X and Y. Assume that we observe that X is uniformly distributed in the

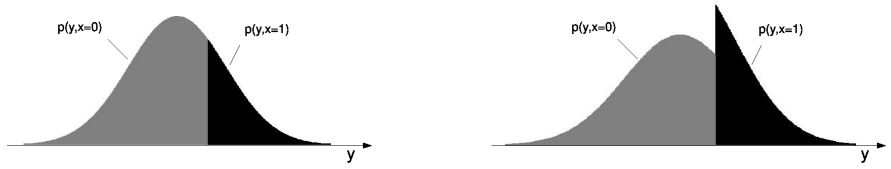

Fig. 7 Joint density $p(x, y)$ of a real-valued random variable Y and a binary variable X. The marginal distribution $p(y)$ is Gaussian. The causal hypothesis $Y \to X$ is plausible: the conditional $p(x|y)$ corresponds to setting $x = 1$ for all y above a certain threshold. On the other hand, $X \to Y$ is rejected by IC because $p(y|x)$ and $p(x)$ share algorithmic information: given $p(y|x)$, only *specific* choices of $p(x)$ reproduce the Gaussian $p(y)$, whereas generic choices of $p(x)$ would yield 'odd' densities of the type on the right. The figures are taken from (Janzing et al, 2009)

interval $[0, 1]$ and that $f : [0, 1] \to [0, 1]$ is a strictly monotonously increasing function. If f is differentiable with a differentiable inverse, the distribution of Y is given by the density

$$p(y) = \frac{d}{dy} f^{-1}(y) \,.$$

The causal direction $Y \to X$ would be prohibited by IC. Observe

$$K(P(X, Y)) \stackrel{+}{=} K(f) \,,$$

since $K(P(X)) \stackrel{+}{=} 0$. For the backward direction we obtain

$$K(P(Y)) + K(P(X|Y)) \stackrel{+}{=} K(f) + K(f^{-1}) \stackrel{+}{=} 2K(f) \stackrel{+}{>} K(P(X, Y)) \,.$$

Hence, $P(Y)$ and $P(X|Y)$ are not independent; their forms both depend on f.

As shown by Daniusis et al (2010) and Janzing et al (2012), $P(Y)$ still contains information on f^{-1} when $P(X)$ is not the uniform distribution. Intuitively speaking, this is because peaks of the density $p(y)$ often co-occur with regions of large slope of f^{-1}. An inference method is presented that is based on this observation. Of course, it does not use *algorithmic* dependence between $P(Y)$ and $P(X|Y)$, but a computable kind of dependence that has strongly been inspired by the above ideas. A modification using linear relations between high-dimensional variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ has been described by Janzing et al (2010) and Zscheischler et al (2011).

5.4 IC selects within Markov equivalence classes

One fundamental limitation of conditional independence based causal inference is given by the fact that it is impossible to distinguish causal DAGs that belong to the same Markov equivalence class. For two statistically dependent variables X and Y, for instance, both causal explanations $X \to Y$ and $Y \to X$ are allowed by causal FF. Janzing and Schölkopf (2010) and Janzing et al (2009) give examples where IC excludes one of the DAGs because $P(X)$ and $P(Y|X)$ are algorithmically dependent. We briefly report one of the most intuitive examples. Assume that $p(y)$ is the probability density of a Gaussian distribution and the supports of $p(y|x = 0)$ and $p(y|x = 1)$ are $(-\infty, y_0]$ and $[y_0, \infty)$, respectively, as shown in Fig. 7, left.

One can easily think of a causal mechanism whose output x is 1 for all inputs y above a certain threshold y_0, and 0 otherwise. Assuming $X \to Y$, we would require a mechanism that generates outputs y from inputs x according to $p(y|x)$. Given this mechanism, there is only one distribution $p(x)$ of inputs for which $p(y)$ is Gaussian (Fig. 7, right, shows what kind of output is obtained by 'detuning' $p(x)$. Hence, the description of $p(x)$ is short when $p(y|x)$ is given. Note that Janzing et al (2009) describe a causal inference method that indeed infers the correct causal direction for this case and whose justification has been inspired by IC. Janzing and Steudel (2010) show that IC provides a justification for a causal inference method that has already been successfully implemented on multiple real data sets with known ground truth (Hoyer et al, 2008; Peters et al, 2011b,a). They are based on additive noise models that we now sketch. Assume that Y is a function of X up to an additive noise term E that is statistically independent of X, that is,

$$Y = f(X) + E \quad \text{with} \quad E \perp\!\!\!\perp X\,.$$

Then Hoyer et al (2008) show that, in the generic case, there is no additive noise model in the opposite direction such that

$$X = g(Y) + \tilde{E} \quad \text{with} \quad \tilde{E} \perp\!\!\!\perp Y\,.$$

If $P(X, Y)$ has an additive noise model from X to Y then one rejects $Y \to X$. This kind of reasoning is justified by IC, provided that the complexity of $P(Y)$ is sufficiently high (this excludes, for instance, the bivariate Gaussian case, where the method obviously fails). More specifically, Janzing and Steudel (2010) show that the algorithmic mutual information between $P(Y)$ and $P(X|Y)$ is close to $K(P(Y))$ up to some small terms.

5.5 IC accepts local non-minimality

We have argued in 2.3 that it does no harm to violate minimality by drawing a DAG that contains additional edges, that is, some variables have additional causes that actually have no influence on them. For instance, if $X \perp\!\!\!\perp Y$, the DAG $X \to Y$ violates minimality, which is rejected by FF. IC, on the other hand, accepts the DAG provided that $P(X)$ and $P(Y|X)$ (which is equal to $P(Y)$) are algorithmically independent.

This behavior of IC can be tolerated provided that the corresponding causal inference algorithm is augmented by a statistical testing procedure that removes redundant parents (see our remarks in subsection 2.3). This is an example where both FF and IC yield reasonable results although they do not coincide. According to one's taste, minimality could also be used on top of IC.

6 Outlook: a more general version of IC

IC has been proposed in this text as a principle to infer causal relations (whether A causes B). Yet, a graphical causal model also provides a description of the *structure* of the underlying physical mechanisms governing a system under study (Lemeire et al, 2011b). 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. Each CPD corresponds to a separate, autonomous part of the system; processes corresponding to different CPDs do not interact. The modularity property ('model components corresponds to independent

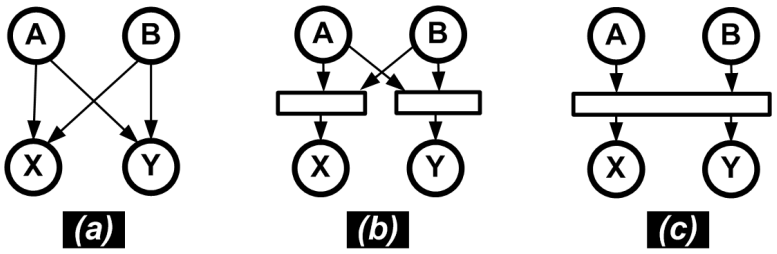

Fig. 8 Example multiple output mechanism.

parts') makes it possible to reason about changes to the system (Pearl, 2000). A mechanism in the system can be replaced by another without affecting the rest of the system. Hence, a CPD in the model can be replaced by another without affecting the rest of the model. As such IC provides a condition to verify whether model components provides a 'natural' decomposition of the system.

The following rough ideas may give a first hint in the direction of generalizing the IC condition so that it can be applied to other model classes. IC can be restated in a more abstract way by saying that *model components should be algorithmically independent*. The acronym can then be used as 'Independence of Components'. The following example of a so-called 'multiple output mechanism' shows that this principle goes beyond factorizations given by Bayesian networks. Consider a chemical reaction between substances A and B which results in two new substances X and Y. This is modeled by a causal graph as depicted in Fig. 8(a) and quantified with two CPDs as shown in Fig. 8(b). But since the chemical reaction is one indivisible mechanism producing X and Y together, Fig. 8(c) better describes the decomposition of the system, since we cannot regard the CPDs $P(X|A,B)$ and $P(Y|A,B)$ as two separate independent and autonomous mechanisms. We should model it as one indivisible mechanism. Such a *multiple output mechanism* might be identifiable by violation of the IC condition for the Bayesian network representation. In some cases, the description of $P(X \mid A,B)$ will be related to that of $P(Y \mid A,B)$. The description of $P(X,Y \mid A,B)$ will be shorter than the separate descriptions of both CPDs taken together. Note that the example relates to the chemical factory given by Cartwright (2002, p.436), but she uses it as a counterexample of the Markov condition.

In Sec. 6 we proposed Markov networks as an alternative model class. According to a Markov network, described by an undirected graph G, the JPD factorizes as (Lauritzen, 1996)

$$P(X_1,\ldots,X_n) = \prod_{C \in cl(G)} \phi(C)\,,$$

where $cl(G)$ denotes the cliques of G (that is, subsets of nodes that are fully connected) and $\phi(C)$ is a 'potential' corresponding to the clique C. Applying our abstract IC onto Markov networks - the potentials $\phi(C)$ ought to be algorithmically independent - might offer a useful condition to infer a meaningful decomposition of the system into potentials.

7 Conclusions

We have argued that the principle of algorithmically Independent Conditionals (IC) is a helpful basis for causal inference. Our examples suggest that its implications are more plausible than those of Faithfulness (FF). IC relies on a prior that not only allows simple CPDs

(as opposed to FF which often rules them out) but even prefers them. Causal inference that is based on IC therefore respects Occam's razor better than FF.

However, one should keep in mind that IC is not a practical inference method, in particular because Kolmogorov complexity is uncomputable. Instead, we consider its role more as being a 'gold standard' for causal inference from statistical data. We have argued that several practical inference algorithms have already been inspired and justified by IC; to develop further practical methods along this line is a challenging goal for the future.

Acknowledgements We would like to thank the blind reviewers in helping to structure our exposition and make our ideas clear. We would also like to thank Patrik Hoyer for providing us the example of Sec. 4.2. This work has partially been carried out within the framework of the Prognostics for Optimal Maintenance (POM) project (grant nr. 100031; `www.pom-sbo.org`) which is financially supported by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

References

Cartwright N (1999) The dappled word: a study of the boundaries of science. Cambridge University Press

Cartwright N (2002) Against modularity, the causal Markov condition and any link between the two. British Journal for the Philosophy of Science 53:41153

Chaitin G (1966) On the length of programs for computing finite binary sequences. J Assoc Comput Mach 13:547–569

Chaitin G (1975) A theory of program size formally identical to information theory. J Assoc Comput Mach 22:329–340

Daniusis P, Janzing D, Mooij J, Zscheischler J, Steudel B, Zhang K, Schölkopf B (2010) Inferring deterministic causal relations. In: Procs of the 26th Conf. on Uncertainty in Artificial Intelligence (UAI)

Gacs P, Tromp J, Vitányi P (2001) Algorithmic statistics. IEEE Trans Inf Theory 47(6):2443–2463

Grünwald P (2007) The minimum description length principle. MIT Press, Cambridge, MA

Hoyer PO, Janzing D, Mooij JM, Peters J, Schölkopf B (2008) Nonlinear causal discovery with additive noise models. In: Koller D, Schuurmans D, Bengio Y, Bottou L (eds) NIPS, MIT Press, pp 689–696

Hutter M (2007) On universal prediction and Bayesian confirmation. Theoretical Computer Science 384(1):33–48

Janzing D, Schölkopf B (2010) Causal inference using the algorithmic Markov condition. IEEE Transactions on Information Theory 56(10):5168–5194

Janzing D, Steudel B (2010) Justifying additive-noise-based causal discovery via algorithmic information theory. Open Systems and Information Dynamics 17(2):189–212

Janzing D, Sun X, Schölkopf B (2009) Distinguishing cause and effect via second order exponential models. `http://arxivorg/abs/09105561`

Janzing D, Hoyer P, Schölkopf B (2010) Telling cause from effect based on high-dimensional observations. in Procs of the Int Conf on Machine Learning (ICML), Haifa, Israel

Janzing D, Mooij J, Zhang K, Lemeire J, Zscheischler J, Daniusis P, Steudel B, Schölkopf B (2012) Information-geometric approach to inferring causal directions. Artificial Intelligence 56(10):5168–5194

Kolmogorov A (1965) Three approaches to the quantitative definition of information. Problems Inform Transmission 1(1):1–7

Korb KB, Nyberg E (2006) The power of intervention. Minds and Machines 16(3):289–302
Lauritzen SL (1996) Graphical Models. Clarendon Press, Oxford
Lauritzen SL, Richardson TS (2002) Chain graph models and their causal interpretation. Journal of the Royal Statistical Society, Series B 64:321 – 361
Lemeire J, Dirkx E (2006) Causal models as minimal descriptions of multivariate systems. http://parallel.vub.ac.be/~jan
Lemeire J, Meganck S, Cartella F, Liu T, Statnikov A (2011a) Inferring the causal decomposition under the presence of deterministic relations. In: Special session Learning of causal relations at the ESANN conference
Lemeire J, Steenhaut K, Touhafi A (2011b) When are graphical causal models not good models? In: Causality in the sciences, J. Williamson, F. Russo and P. McKay, editors, Oxford University Press
Levin L (1974) Laws of information conservation (non-growth) and aspects of the foundation of probability theory. Problems Information Transmission 10(3):206–210
Meek C (1995) Strong completeness and faithfulness in Bayesian networks. In: Procs of the Conf. on Uncertainty in Artificial Intelligence (UAI), pp 411–418
Pearl J (2000) Causality. Models, Reasoning, and Inference. Cambridge University Press
Peters J, Janzing D, Schölkopf B (2011a) Causal inference on discrete data using additive noise models. IEEE Transac Patt Analysis and Machine Int 33(12):2436–2450
Peters J, Mooij J, Janzing D, Schölkopf B (2011b) Identifiability of causal graphs using functional models. In: Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence (UAI)
Rathmanner S, Hutter M (2011) A philosophical treatise of universal induction. Entropy 13(6):1076–1136, DOI 10.3390/e13061076
Solomonoff R (1960) A preliminary report on a general theory of inductive inference. Technical report V-131 Report ZTB-138 Zator Co.
Solomonoff R (1964) A formal theory of inductive inference. Information and Control, Part II 7(2):224–254
Spirtes P, Glymour C, Scheines R (1993) Causation, Prediction, and Search, 2nd edn. Springer Verlag
Zhang J, Spirtes P (2011) Intervention, determinism, and the causal minimality condition. Synthese 182(3):335–347
Zscheischler J, Janzing D, Zhang K (2011) Testing whether linear equations are causal: A free probability theory approach. In: In Proc. of the 27th Conf. on Uncertainty in Artificial Intelligence (UAI)