Journal of Machine Learning Research 17 (2016) 1-20 Submitted 10/14; Revised 8/15; Published 4/16

Conditional Independencies under the Algorithmic Independence of Conditionals.

Jan Lemeire JAN.LEMEIRE@VUB.AC.BE
Vrije Universiteit Brussel, INDI Dept, ETRO Dept. Pleinlaan 2, B-1050 Brussels, Brussels, Belgium
iMinds, Dept. of Multimedia Technologies, Gaston Crommenlaan 8, B-9050 Ghent, Belgium

Editor: Isabelle Guyon and Alexander Statnikov

Abstract

In this paper we analyze the relationship between faithfulness and the more recent condition of algorithmic Independence of Conditionals (IC) with respect to the Conditional Independencies (CIs) they allow. Both conditions have been extensively used for causal inference by refuting factorizations for which the condition does not hold. Violation of faithfulness happens when there are CIs that do not follow from the Markov condition. For those CIs, non-trivial constraints among some parameters of the Conditional Probability Distributions (CPDs) must hold. When such a constraint is defined over parameters of different CPDs, we prove that IC is also violated unless the parameters have a simple description. To understand which non-Markovian CIs are permitted we define a new condition closely related to IC: the Independence from Product Constraints (IPC). The condition reflects that CIs might be the result of specific parameterizations of individual CPDs but not from constraints on parameters of different CPDs. In that sense it is more restrictive than IC: parameters may have a simple description. On the other hand, IC also excludes other forms of algorithmic dependencies between CPDs. Finally, we prove that on top of the CIs permitted by the Markov condition (faithfulness), IPC allows non-minimality, deterministic relations and what we called proportional CPDs. These are the only cases in which a CI follows from a specific parameterization of a single CPD.

Keywords: faithfulness, causality, independence of conditionals, Kolmogorov complexity

1. Introduction

Algorithmic Independence of Conditionals (IC) has been put forward for causal inference and its relation with Faithfulness (FF) was analyzed by Lemeire and Janzing (2013). We showed that both conditions often lead to the same causal conclusions and are motivated by similar grounds: that the CPDs are independently chosen. But we argued that IC is more fundamental than FF: we can trust IC more whenever the conclusions are different from those of FF. Moreover, IC goes beyond FF. IC has led to successful causal inferences in cases that FF cannot decide on the causal orientations, see for instance Janzing and Schölkopf (2010); Janzing and Steudel (2010); Daniusis et al. (2010); Janzing et al. (2012); Chen et al. (2014). In this paper we establish the link between FF and IC regarding the Conditional Independencies (CIs) both conditions admit.

Faithfulness is based on the CIs among the observed variables entailed by a system. Faithfulness assumes that all CIs come from the system's causal structure, described by a Directed Acyclic Graph (DAG), and hold for all parameterizations of the DAG. These CIs are defined by the Markov condition applied on the DAG and can be identified with the d-separation criterion. Lemeire and Janzing (2013) have shown that some CIs are rejected by causal faithfulness but have to be accepted

by the IC condition. This is true for deterministic relations for example, since a deterministically related variable becomes independent from *all* other variables when conditioned on its determiner, also those that are not d-separated.

FF is motivated by the Lebesgue measure zero argument, saying that if the system's parameters were randomly chosen, the probability of having a configuration following a specific constraint has Lebesgue measure zero (Meek, 1995). In the case of deterministic relations, all but one probability is zero for each input state. This is very unlikely to occur by chance, hence receives Lebesgue measure zero. IC follows a different reasoning. Its justification is based on Solomonoff's Universal Prior (Solomonoff, 1964) which assigns non-zero probability to those points in parameter space that have a finite description (Lemeire and Janzing, 2013). Points reflecting some regularity (allowing compression of the description) will receive a high probability. The Universal Prior favors simple CPDs, and therefore respects Occam's razor. The reasoning is that patterns or regularities are likely to occur; patterns have to be expected. Which is not the case if parameters are randomly chosen according to for instance a uniform distribution. On the other hand, IC follows FF by excluding 'non-generic' parameter configurations. IC assumes that CPDs correspond to independent mechanisms which were 'chosen' independently. While FF excludes *all* non-trivial constraints among some parameters of CPDs, IC will in general exclude CIs following from specific parameter matches between *different* conditionals. The latter will be expressed formally by the novel condition Independence from Product Constraints (IPC). The condition is introduced to analyze the relation between IC and FF. In this paper we analyze the relation between the 3 conditions.

We first recall the definitions. Then we discuss CIs that do not follow from the Markov condition and we introduce the new IPC criterion. Then, we analyze when CIs violate the IC condition. Section 5 establishes the link between IC/IPC and faithfulness. Next we prove the link between IC and IPC. Before concluding, we discuss the practical implications of these results.

2. Definitions

A Bayesian network consists of a Directed Acyclic Graph (DAG) and a set of Conditional Probability Distributions (CPDs) defined over variables $X_1, \ldots, X_n$ such that the joint probability distribution equals the following factorization:

$$P(X_1, \ldots, X_n) = \prod_i P(X_i | Parents(X_i)) \tag{1}$$

with $Parents(X_i)$ the parent nodes of node X_i in the DAG. A Bayesian network is edge-minimal (MIN) in the sense that no edge can be removed from the DAG without violating the correctness of the factorization.

In a causal model represented by a Bayesian network, all edges correspond to direct causal relations and each CPD corresponds to an independent and autonomous mechanism of the system (Hausman and Woodward, 1999; Lemeire et al., 2011). Throughout the paper we assume that the CPD of each node X_j is described by a parameter vector $\boldsymbol{\lambda}_j \in \mathbb{R}^{k_j}$. Although the finite dimension of the parameter space restricts the conditionals for continuous variables already, we believe that this is appropriate because inference from finite data requires strong assumptions or approximations anyway.

Conditional Independence (CI) is defined as

$$U \perp\!\!\!\perp W \mid \boldsymbol{V} \Leftrightarrow$$
$$\forall \boldsymbol{v} \in \boldsymbol{V}_{dom},\, w \in W_{dom} : P(U \mid \boldsymbol{v}, w) = P(U \mid \boldsymbol{v}) \quad \text{whenever} \quad P(\boldsymbol{v}, w) > 0. \tag{2}$$

where X_{dom} is the domain of variable X. Single random variables are denoted by capital letters and sets of variables by boldface capital letters. Values of variables are denoted by lowercase letters. Note that the conditional distribution $P(U \mid \boldsymbol{v}, w)$ is only defined in points where $P(\boldsymbol{v}, w) > 0$.

The Markov condition gives all conditional independencies following from the above factorization (Hausman and Woodward, 1999, p. 532): Every variable is conditionally independent of its non-descendants (except for itself), given its parents. These Markovian independencies hold for all parameterizations of the CPDs. These independencies can be identified graphically by 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 $. \perp . \mid .$:

$$\boldsymbol{X} \perp \boldsymbol{Y} \,|\boldsymbol{Z}\,.$$

A Bayesian network is said to be *faithful* if the Markovian CIs are the only independencies present in the joint probability distribution; in other words, there are no CIs not following from Markov. We call them *non-Markovian CIs*. Where the Markovian CIs occur for every parameterization of the CPDs, non-Markovian CIs only occur for specific parameterizations of the model. As we will see, specific parameter constraints should be met.

Next we provide the definition of the IC condition. It is based on Kolmogorov complexity or algorithmic information of a binary string s, denoted by $K(s)$. 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. Based on Kolomogorov complexity we can define algorithmic independence.

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

$$K(s_1, \dots, s_n) \stackrel{+}{=} \sum_{i}^{n} K(s_i)\,. \tag{3}$$

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.

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. For fixed strings, we have to interpret $\stackrel{+}{=}$ in the sense of 'equality 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.

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

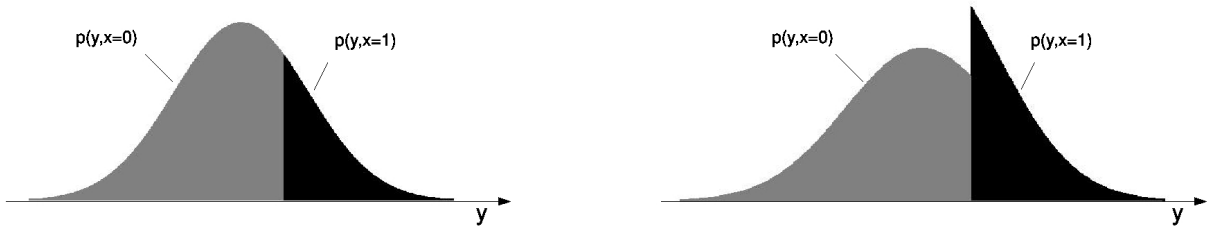

Figure 1: Binary variable X is determined by Gaussian variable Y by a thresholding mechanism, i.e., $X = 1$ for all $y > y_0$, and $X = 0$ otherwise. This is shown on the left. The causal hypothesis $Y \to X$ is plausible: the conditional $P(X|Y)$ corresponds to setting $X = 1$ for all Y above a certain 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).

Assumption 1 (CPDs have finite description length) *We assume that each parameter vector $\boldsymbol{\lambda}_j \in \mathbb{R}^{k_j}$ has finite description length. To be precise, there is a program that computes the lth component of $\boldsymbol{\lambda}_j$ up to the precision of d digits if it gets the input (l, j).*[2] *Then, $K(\boldsymbol{\lambda}_j)$ denotes the length of the shortest program of this type.*

Now we are ready to define the IC condition (Lemeire and Janzing, 2013):

Definition 2 (Independence of Conditionals)
The conditional probability densities $CPD_1, \ldots, CPD_n$ corresponding to a DAG G with n nodes are said to satisfy the Algorithmic Independence of Conditionals, or Independence of Conditionals (IC) for short, if the corresponding parameter vectors $\boldsymbol{\lambda}_j$ satisfy

$$K(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n) \stackrel{+}{=} \sum_{j=1}^{n} K(\boldsymbol{\lambda}_j)\,, \tag{4}$$

The uncomputability of Kolmogorov complexity hinders the applicability of these concepts. Applications rely on some approximative measure of algorithmic complexity or, as in the following example, on an approximative measure of 'correlation' between two distributions.

An example of the usage of the IC condition for causal inference is given by Fig. 1 (Janzing et al., 2009). Consider that Y causes X: $Y \to X$. Let Y be a Gaussian variable with zero mean and standard deviation 1 (i.e., described by a zero-dimensional parameter space). Let X be a binary variable deterministically determined by Y by a thresholding mechanism, i.e., $X = 1$ for all $y > y_0$ where $y_0 \in \mathbb{R}$ is some threshold, and $X = 0$ otherwise. Here, $P(X|Y)$ is described by 1 parameter, namely y_0. We now describe the joint distribution $P(X, Y)$ in the wrong causal direction, i.e. with $P(X)$ and $P(Y|X)$. We observe that the set of possible $P(X)$ is not restricted, i.e., we have the one-dimensional parameter $\theta_1 = P(X = 0)$. The set of possible conditionals $P(Y|X)$ obtained by

2. Since the input consists of two strings and the output of k_j strings plus extra information specifying the position of the comma, we use some canonical bijection between $\{0,1\}^*$ and $(\{0,1\}^*)^d$ for appropriate d for input and output.

the above model class is also determined by a one-dimensional parameter $\theta_2 = y_0$ that determines the cutoff. We then observe that θ_1 and θ_2 are related by

$$\theta_1 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta_2} e^{-y^2/2} dy = \mathrm{erf}(\theta_2)\,.$$

Hence, IC is violated whenever $K(\theta_1) \stackrel{+}{>} K(\mathrm{erf})$:

$$\begin{aligned} K(\boldsymbol{\lambda}_X) + K(\boldsymbol{\lambda}_Y) &\geq K(\theta_1) + K(\theta_2) \\ &\stackrel{+}{>} K(\mathrm{erf}) + \mathrm{K}(\theta_2) \\ &\stackrel{+}{=} K(\theta_1|\theta_2) + K(\theta_2) \\ &\geq K(\theta_1, \theta_2) \\ &\stackrel{+}{=} K(\boldsymbol{\lambda}_X + \boldsymbol{\lambda}_Y) \end{aligned}$$

Note that IC is not violated for $y_0 = 0$: then $K(\theta_1) = K(1/2) \stackrel{+}{=} 0$ and $K(\theta_2) = K(0) \stackrel{+}{=} 0$.

3. Non-Markovian conditional independencies

We start the investigation by analyzing the CIs that do not follow from the Markov condition. Non-trivial polynomial constraints must be satisfied for non-Markovian conditional independencies. This is shown for discrete Bayesian networks by Meek (1995) and the linear case for distributions over continuous variables by Spirtes et al. (1993).

For a given DAG G and an independence, we define the **Independence Parameter Subspace** as the set of all parameterizations $\boldsymbol{\lambda}$ of a DAG G for which the independence holds. For Markovian CIs this is the complete space $\boldsymbol{S} := \times_{j=1}^{n} \boldsymbol{S}_j$, where $\boldsymbol{S}_j$ is the set of possible parameter vectors $\boldsymbol{\lambda}_j$. For non-Markovian CIs this is a subspace of $\boldsymbol{S}$.

As already pointed out in the introduction and Lemeire and Janzing (2013), deterministic relations between some variables may induce conditional independencies that do not follow from the Markov condition. For $Y \perp\!\!\!\perp Z|X$ in the example $X \to Y \to Z$, a sufficient condition is the function $Y = f(X)$. The independence parameter subspace for $Y \perp\!\!\!\perp Z|X$ can be represented by Fig. 2(a) in the case of deterministic relation $Y = f(X)$. If $\boldsymbol{\lambda}_Y \subset \boldsymbol{R}_Y$ where $\boldsymbol{R}_Y$ represents all functions, then we have the conditional independence in $P(X, Y, Z)$. $\boldsymbol{\lambda}_Y$ denotes the parameter vector of the CPD of variable Y, $\boldsymbol{\lambda}_Z$ that of Z, and so on.

As another example, 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. Assume that all CPDs are given by linear structure equations:

$$\begin{aligned} B &= \alpha A + U_B\,, \\ C &= \beta A + U_C\,, \\ D &= \gamma B + \delta C + U_D\,, \end{aligned}$$

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

$$\alpha\gamma = -\beta\delta\,, \tag{5}$$

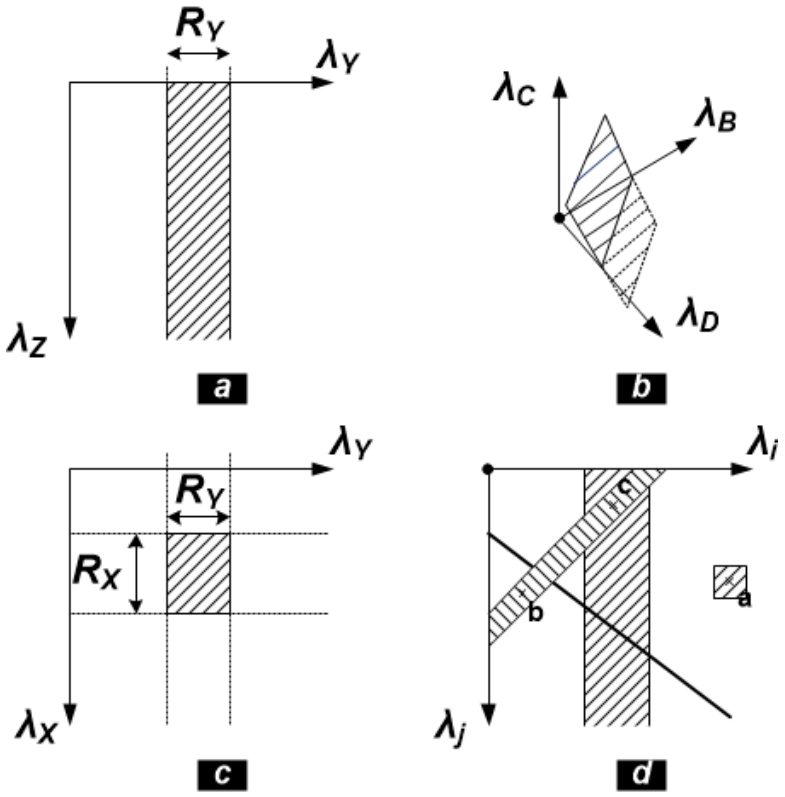

Figure 2: Independence parameter subspaces where each axis represents the possible parameter vectors of a single CPD. In (d) points c and a are permitted by IPC, while b is not.

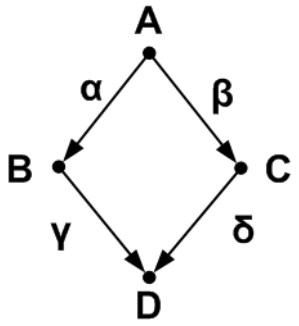

Figure 3: Causal model with linear influences described by parameters $\alpha, \beta, \gamma, \delta$.

such that $A \perp\!\!\!\perp D$. Obviously, FF rejects the causal DAG of Fig. 3 because A and D are not d-separated by the empty set and thus should not be independent. The independence parameter subspace for $A \perp\!\!\!\perp D$ can be represented by Fig. 2(b). Note that γ and δ are represented together by dimension $\boldsymbol{\lambda}_D$.

As a third example, consider the causal model $W \rightarrow X \rightarrow Y \rightarrow Z$ with $X = g(W)$ and $Y = f(X)$. It follows that $Y \perp\!\!\!\perp Z|W$. Fig. 2(c) represents the parameter subspace for this CI: both $\boldsymbol{\lambda}_X$ and $\boldsymbol{\lambda}_Y$ are restricted for the CI to hold. But both restrictions do not depend on each other: the parameter subspace can be described by the product $\boldsymbol{R}_X \times \boldsymbol{R}_Y$. The forthcoming IPC criterion will reject the independence following from Eq. 5 but accepts the last example because the latter parameter sub space is a so-called product subspace.

To formalize the analysis of the independence parameter subspaces, we need the following postulate based on the results of (Meek, 1995) and (Spirtes et al., 1993).

Postulate 1 *There exists a complexity measure $C(.)$ on polynomial equations such that for a given DAG the presence of any conditional independence can be identified by a unique minimal set of undecomposable polynomial constraints on the parameterization of the DAG. With 'undecomposable' we mean that a constraint c cannot be written as $(c_1$ or $c_2)$ (one of both constraints should be true) with $C(c) \geq C(c_1) + C(c_2)$. By 'minimal' we mean that there is no smaller such set.*

Note that Kolmogorov complexity is not appropriate as complexity measure since it attributes 0 complexity to polynomial functions with simple coefficients. A measure such as AIC or BIC is more appropriate to capture the complexities of polynomial equations. However, it would lead us too far to prove for a complexity measure that it leads to a unique decomposition for non-Markovian CIs. Hence the postulate.

The postulate implies that the parameter sub space of a CI can be described in a unique, 'canonical' way as a union of areas, where each of the areas reflects 1 basic polynomial constraint. Fig. 2(d) shows a general example of a parameter sub space. The sub space can be decomposed into 4 basic areas which cannot be further decomposed without increasing its descriptive complexity.

Some areas can be described by a product of parameter constraints and some areas can't. The latter means that the CI does not follow from a product constraint on the parameters of different CPDs, but from a constraint that is defined over different CPDs. Those parameterizations will be refuted: parameter configuration b in Fig. 2(d) is refuted while a and c are permitted. This is expressed by the Independence from Product Constraints (IPC) condition: a parameterization will be rejected if it gives rise to a CI which is part of an area in the parameter subspace which cannot be described by a product of subspaces.

Definition 3 (Independence from Product Constraints (IPC)) *Let a Bayesian network be described by the parameters $\boldsymbol{\lambda} := (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)$. Then it is said to satisfy the Independence from Product Constraints Condition if for every independence that holds true for $\boldsymbol{\lambda}$, the parameterization satisfies a constraint of the minimal set of undecomposable constraints of the independence which is a product constraint. A product constraint is a constraint c that can be written as $(c_1$ and $c_2 \ldots$ and $c_l)$ where each c_j is a constraint on exactly one $\boldsymbol{\lambda}_j$.*

IC, however, will only reject non-product constraints if it results in a compression of the parameters. This is investigated in the next section.

4. Conditional independencies resulting in violations of IC

Although the IPC condition reflects the principle that 'non-generic' relations among CPDs are rejected, not all non-Markovian CIs are rejected by IC. To analyze this in detail, we have to recall the following theorem on the violation of IC (Lemeire and Janzing, 2013, Theorem 3):

Theorem 4 *For a given DAG G, let the set of possible CPDs $P(X_j|Parents(X_j))$ be parameterized by some parameter set $\boldsymbol{\lambda}_j := \{\lambda_j^1, \ldots, \lambda_j^{k_j}\}$ of parameters. Assume that the parameter values for some specific choice $CPD_1, \ldots, CPD_n$ of conditional probability 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 $\boldsymbol{\lambda}_j$. Assume furthermore that θ_1 corresponds to CPD_1 (without loss of generality). Then the following condition implies violation of IC:*

$$K(f) \stackrel{+}{<} K(\theta_1 | CPD_1^{\setminus \theta_1, *}) \,, \tag{6}$$

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

The theorem states that a constraint results in a violation of IC provided that the parameters are sufficiently complex compared to the complexity of the constraint.

Applied on the example of Fig. 3, the constraint defined by Eq. 5 leads to a violation of IC if α, β, γ and δ have complex values. Describing the JPD by separate descriptions of $P(A)$, $P(B|A)$, $P(C|A)$ and $P(D|B,C)$ is redundant because the parameter γ in $P(D|B,C)$ can be computed from the parameters of the other CPDs via Eq. 5. The constraint is an unlikely coincidence if all real-valued parameters are chosen independently (according to some continuous distribution on $\mathbb{R}$).

Next, consider causal structure $X \rightarrow Y \leftarrow Z$ over binary variables X, Y and Z. $P(Z)$ is parameterized with $P(Z=0) = \alpha$ and $P(Z=1) = 1-\alpha$, and $P(Y|X,Z)$ with 4 parameters:

$P(Y=0\|X,Z)$	$Z=0$	$Z=1$
$X=0$	a	b
$X=1$	c	d

Then, non-Markovian independence $X \perp\!\!\!\perp Y$ holds when

$$\begin{aligned}
& P(Y=0|X=0) = P(Y=0|X=1) \\
\Leftrightarrow\ & P(Z=0).P(Y=0|X=0,Z=0) + P(Z=1).P(Y=0|X=0,Z=1) \\
& = P(Z=0).P(Y=0|X=1,Z=0) + P(Z=1).P(Y=0|X=1,Z=1) \\
\Leftrightarrow\ & \alpha.a + (1-\alpha).b = \alpha.c + (1-\alpha).d \,.
\end{aligned}$$

Note that this constraint does not depend on the parameterization of $P(X)$. This equation can be rewritten in the following constraint between the parameters of $P(Z)$ and $P(Y|X,Z)$ with T a constant:

$$\frac{\alpha}{1-\alpha} = \frac{d-b}{a-c} = T \tag{7}$$

This equation holds for the following particular parameterization:

- $P(Z=0) = P(Z=1) = 0.5$
- $P(Y|X,Z)$ is a noisy exclusive or (with real-valued $E \in]0,1[$ representing the noise):

$P(Y\|X,Z)$	$Z=0$	$Z=1$
$X=0$	E	$1-E$
$X=1$	$1-E$	E

It can easily be verified that Eq. 7 holds for all values of E and that $P(Y=1|X=x) = P(Y=1) = 0.5$ for all values of X. Although in this example the parameters of both CPDs are tightened by the constraint, one can hardly say that the parameter of one CPD helps the description of some parameter of the other CPD. $T=1$ and therefore simple. Also α has constant complexity. Even if E is complex and $K(a)$ therefore too, $K(a|b,c,d)$ has low complexity. Eq. 6 of Theorem 4 does not hold: IC is not violated. It is only violated for complex values of T. Then α will have high Kolomogorov complexity but its description length has constant complexity with the help of a, b, c and d.

Concluding, factorizations having non-Markovian CIs will only be rejected for complex parameter values. The rationale is that we consider simple parameter settings to appear with a much greater probability than when they would be taken randomly from a uniform distribution over the parameter space. We prefer the Universal Prior (Solomonoff, 1964): the probability of a parameter configuration is high for simple parameters, where simple is defined by their Kolmogorov complexity. The probability decreases for increasing parameter complexities. As such we will refute only constraints leading to a compression. Simple parameters will occur and as such make it highly probable that a coincidental CI occurs. In our case, the uniform distribution for $P(Z)$ and xor configuration of $P(Y|X,Z)$ are simple and can be expected. As such, the CI can be expected and must not be excluded.

Finally, note that if a parameter is incompressible with respect to the other CPD parameters, then $K(\theta_1) = K(\theta_1|CPD_1^{\setminus\theta_1,*})$ and the constraint of Theorem 4 becomes simply:

$$K(f) \overset{+}{<} K(\theta_1)\,. \tag{8}$$

As we will see, the polynomial constraints for non-Markovian CIs have simple coefficients which make that the functions have constant complexity. Thus, complex parameters lead to violations of IC and can be detected by an appropriate approximative measure of Kolmogorov complexity.

5. The IPC and IC conditions and Faithfulness.

Now we infer our most important result: the relation between IC and Faithfulness. We first establish the link between IPC and faithfulness, and then we show that the results also apply for IC when the parameters are sufficiently complex.

Faithfulness implies that variables that are d-connected are (conditionally) dependent: there are unblocked paths. We will prove that under IPC they yield dependence apart from 3 cases permitted by IPC. The theorem is constructing by decomposing the unblocked paths and proving dependency for each component. The basic components of the paths are adjacent variables and v-structures. These are considered first in 2 lemmas together with a lemma on conditioning variables that are not

on one of the paths. All proofs are given in the appendix. We prove it for discrete variables. We believe that a similar approach can be used to prove it for continuous variables.

For using IPC we need the following postulate:

Postulate 2 *The non-trivial polynomial constraints among different CPDs responsible for non-Markovian CIs cannot be decomposed into product constraints without increasing their complexity, i.e.* $C(c) < C(c_1) + \ldots + C(c_n)$ *for any decomposition of non-product constraint* c *into product constraints* $c_1, \ldots$ *and* c_n*.*

Since the constraints apply on parameters of different CPDs it seems reasonable to believe that the constraints cannot be decomposed into constraints on individual parameters without increasing the descriptive complexity.

We first define formally what we mean by excluding deterministic relations.

Definition 5 (INDET) *A Bayesian network is said to satisfy INDET if there is no variable* X_j *that can be written as*

$$X_j = f(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)\,.$$

The first lemma shows that 2 adjacent variables are dependent under IPC and edge-minimality (MIN).

Lemma 6 *Given a Bayesian network defined over discrete variables satisfying MIN and IPC, any two adjacent variables* X *and* Y*, where* X *is a parent of* Y*, are dependent conditioned on any subset of the other parents of* Y*.*

The next lemma states that conditionally dependent variables X and Y cannot get conditionally independent via extending the conditioning set by variables that are not on any path between X and Y unless deterministic variables are present.

Lemma 7 *Assume that MIN and IPC holds for some Bayesian network defined over discrete variables. Let* $X \not\perp\!\!\!\perp Y \mid \mathbf{Z'}$ *where* X, Y *are arbitrary nodes and* $\mathbf{Z'}$ *is a set of nodes. Let* $\mathbf{U}$ *denote the set of all variables on the non-blocked paths between* X *and* Y*, including* X *and* Y*. Then, for every set* $\mathbf{Z''} \subset \overline{\mathbf{Z'} \cup \mathbf{U}}$*, when*

$$X \perp\!\!\!\perp Y \mid \mathbf{Z'}, \mathbf{Z''}$$

holds, there are deterministic relationships.

In the third lemma we deal with the case in which X and Y are connected via a v-structure $X \rightarrow Z \leftarrow Y$ and one conditions on Z or one of its descendants, such that X and Y are d-connected. It appears that there is a special parameterization in which X and Y are conditionally independent without violating IPC, which we call a *proportional conditional probability distribution*.

Definition 8 (proportional conditional probability distribution (pCPD))
Let Z, Y *be variables and* $\mathbf{X}$ *be a set of variables. The CPD* $P(Z|Y, \mathbf{X})$ *is said to have a proportional conditional probability distribution if* Y *can only attain two values* y_1 *and* y_2 *and we have*

$$\frac{P(Z \mid y_1, \boldsymbol{x})}{P(Z \mid y_2, \boldsymbol{x})} = \alpha \quad \forall \boldsymbol{x} \in \boldsymbol{X}_{dom} \tag{9}$$

with α *a constant only depending on* Z *and* $\boldsymbol{X}_{dom}$ *the domain of* X*.*

The probability distribution of Table 1 shows an example of a pCPD for distributions over discrete variables, for which $\frac{P(z=1|y=1,x_i)}{P(z=1|y=0,x_i)} = 1.5$ and $X \perp\!\!\!\perp Y \mid Z$ holds, which is a non-Markovian CI.

X	$Y=0$	$Y=1$
0	0.3	0.45
1	0.6	0.9
2	0.4	0.6

Table 1: Conditional Probability Table of $P(z = 1 \mid X, Y)$ for which Eq. 9 holds and therefore $X \perp\!\!\!\perp Y \mid Z$.

An example for continuous variables in which Eq. 9 is satisfied is the following class of models

$$p(z|y, x_1, \ldots, x_k) = e^{-c(z)y - \sum_j \gamma_j(z) x_j} ,$$

where $c(z)$ and $\gamma_j(z)$ are arbitrary functions.

pCPDs imply non-Markovian CIs in v-structures as shown by the following lemma.

Lemma 9 *If MIN, INDET, IPC holds for a factorization of a joint probability distribution defined over discrete variables, then for any v-structure $X \to Z \leftarrow Y$ in the DAG and for all $\boldsymbol{W}$ not containing X, Y or Z:*

$$X \perp\!\!\!\perp Y \mid Z, \boldsymbol{W} \Leftrightarrow \; X \perp\!\!\!\perp Y \;\; and \;\; P(Z \mid X, Y) \;\; is \;\; a \;\; proportional \;\; CPD$$
$$\forall U \;\; descendants \;\; of \;\; Z : X \not\perp\!\!\!\perp Y \mid U, \boldsymbol{W}$$

The absence of pCPDs is denoted as the NOpCPD condition.

It can easily be verified that violation of MIN or INDET or the presence of a pCPD are permitted by IPC. These 3 cases define constraints on single CPDs. We have to exclude them on top of IPC for achieving faithfulness. The following theorem expresses the relation between IPC and faithfulness:

Theorem 10 *If for a given factorization of a JPD the conditions IPC, MIN, INDET and NOpCPD are met, then faithfulness holds:*

$$\forall \; disjoint \; subsets \; \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \;\subset\; \boldsymbol{V}: \quad \boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \;\Leftrightarrow\; \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

with $\boldsymbol{V}$ the set of all variables under consideration.

In other words, there are only 3 cases in which a non-Markovian CI comes from a specific parameterization of a single CPD: deterministic relationships, non-minimality and proportional CPDs.

To identify the relation between FF and IC we have to define what we mean by sufficiently complex parameters.

Definition 11 (COMPLEX) *A parameterization of a Bayesian network is said to satisfy COMPLEX if for all parameters θ_i of all parameter vectors λ_j:*

$$K(\theta_i | CPD_j^{\setminus \theta_i, *}) \overset{+}{>} 0 ,$$

In the 3 previous lemmas and Theorem 10, the IPC condition can be replaced with IC and COMPLEX.

Theorem 12 *If for a given factorization of a JPD defined over discrete variables the conditions IC, COMPLEX, MIN, INDET and NOpCPD are met, then faithfulness holds:*

$$\forall \ disjoint \ subsets \ \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \ \subset \ \boldsymbol{V}: \quad \boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \ \Leftrightarrow \ \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

with $\boldsymbol{V}$ *the set of all variables under consideration.*

6. The relation between IC and IPC.

IC and IPC partially overlap. IC rules out all atypical constraints on parameters between different CPDs. IPC only rules out the constraints leading to conditional independencies, while IC allows conditional independencies when matches are to be expected because of low complexity of the parameters.

The theorem on the relation between IC and IPC is based on the fact that non-trivial polynomial constraint must be satisfied for non-Markovian conditional independencies.

Theorem 13 *Given P a distribution over n variables and a DAG G with CPDs parameterized by $\theta_1, \ldots, \theta_n$ describing a factorization of P.*
If the parameter vectors satisfy IPC, there exists a parameterization $\theta'_1, \ldots, \theta'_n$ of the CPDs which has the same independencies as P and for which IC holds.
For discrete Bayesian networks and the linear case for continuous distributions, if the factorization satisfies IC and COMPLEX (the parameters θ_i^j of the parameter vectors $\theta_1, \ldots, \theta_n$ have non-constant complexity: $K(\theta_i^j | CPD_i^{\backslash \theta_i^j, *}) \stackrel{+}{>} 0$), then IPC holds as well.

The proof is given in the appendix. We believe that it is provable that under IPC, *most* parameterizations satisfy IC as well. The proof, however, need some quite technical details to be worked out.

7. Practical application

Despite the uncomputability of Kolmogorov complexity, several novel approaches to causal inference are based on the notion of IC, see for instance Janzing and Schölkopf (2010); Janzing and Steudel (2010); Daniusis et al. (2010); Janzing et al. (2012); Chen et al. (2014). As more algorithms will pop up, a philosophical and theoretical underpinning is important. This paper tries to contribute to this endeavour.

Although we advocate that IC is more fundamental than FF (Lemeire and Janzing, 2013), independence-based learning remains a powerful approach. However, we believe that non-Markovian CIs should be taken into account and deserve an in-depth study. Accordingly, independence-based learning algorithms have been adopted in the past to incorporate the presence of deterministic relationships (Lemeire et al., 2012) and pCPDs (violations of orientation faithfulness) (Ramsey et al., 2006).

The theoretical results of this paper might help to understand the nature of the CIs. We showed that the appearance of CIs happens at different levels:

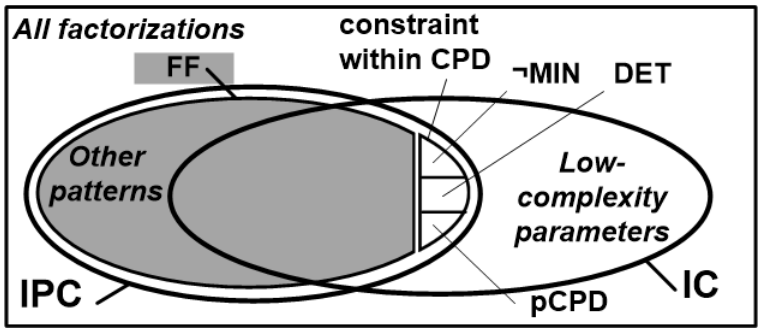

Figure 4: Relation between FF, the IC and IPC condition. It shows the factorizations for which FF, IC and/or IPC holds. The area in which FF holds is shown in gray.

1. some conditional independencies arise from the causal structure (given by the Markov condition),
2. some CIs arise from a specific parameterization of a single CPD,
3. some CIs arise from specific simple parameterizations of different CPDs,
4. some CIs arise from specific complex parameterizations of different CPDs.

Only level 1 is permitted by faithfulness. IPC also permits level 2, while the IC condition allows level 3 as well. Note that some examples of level 4 where given in Lemeire and Janzing (2013) (so-called metamechanisms).

8. Conclusions

Both Faithfulness (FF) and Independence of Conditionals (IC) express a condition on a factorization. They have been extensively used for learning a system's causal structure from observational data. Non-causal factorizations are refuted when one of both conditions is violated. In this paper we investigated more deeply the relation between both conditions.

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. The relation between the conditions is shown by Fig. 4. The sets represent the factorizations for which the conditions hold. We defined the Independence from Product Constraints (IPC) which allows non-Markovian CIs following from specific parameterizations of individual CPDs but not from constraints among parameterizations of different CPDs. We proved that those CIs come from non-minimality (¬MIN), deterministic relationships (DET) or so-called proportional CPDs (pCPDs). They are allowed by IPC.

In contrast to IPC, IC is not violated for constraints on low-complexity parameters, since such constraints can be expected to appear by chance. As such, IC can only be applied for sufficiently complex parameters, e.g. when sufficient data is present. As IC goes beyond CIs, IC provides a way to select the causal structure in the Markov equivalence class (the set of DAGs having the same CIs) based on other patterns among the parameters of different CPDs. This has been shown by several recently developed algorithms.

Acknowledgments

I have to thank Dominik Janzing from the MPI for Intelligent Systems of Tübingen, Germany for his insightful comments and discussion. He came up with the IPC condition. He shows me true science. I am also grateful to the anynomous reviewers whose comments helped me to improve the quality of the paper.

Appendix A. Appendix: Proofs

Theorem 12, which is proven here, starts with a factorization and states that if 4 conditions are met, the DAG corresponding to the factorization is faithful to the joint probability distribution. We denote CPD_X as the CPD for X, given its parents, i.e., $P(X \mid Parents(X))$. To prove faithfulness we will write the down the left-hand side of Eq. 2 as function of the factors of the factorization. The IPC condition rules out independencies following from non-trivial constraints that relate parameters that correspond to different CPDs.

Recall that **conditional independence** is defined as

$$\begin{aligned} &U \perp\!\!\!\perp W \mid V \Leftrightarrow \\ &\forall v \in V_{dom},\, w \in W_{dom} : P(U \mid v, w) = P(U \mid v) \quad \text{whenever} \quad P(v, w) > 0. \end{aligned}$$

Note that the conditional distribution is only defined in points where $P(v, w) > 0$. Since the right hand side is independent of w, the left hand side must give the same result for all values of W. Hence:

$$U \perp\!\!\!\perp W \mid V \quad \Leftrightarrow \quad P(U \mid w_1, V) = P(U \mid w_2, V) \;\; \forall\, w_1, w_2 \in W_{dom} \tag{10}$$

This is the main equation that will be used throughout the proofs. Conditional dependence is therefore defined as $P(U \mid V, W) \neq P(U \mid V)$. It means that there is at least one value for U, V and W for which the negation holds. But since $P(U \mid V) = \sum_w P(U \mid V, w) P(w \mid V)$, there are at least two values of W for which $P(U \mid V, w) \neq P(U \mid V)$. This can be understood by noting that $P(U \mid V)$ is a weighted average of $P(U \mid V, W)$. If the probability for one value of W is higher than the average $P(U \mid V)$, there must be at least one value for which the probability is lower.

For completeness we restate the postulate which is necessary to use IPC.

Postulate 2 *The non-trivial polynomial constraints among different CPDs responsible for non-Markovian CIs cannot be decomposed into product constraints without increasing their complexity, i.e.* $C(c) < C(c_1) + \ldots + C(c_n)$ *for any decomposition of non-product constraint* c *into product constraints* $c_1, \ldots$ *and* c_n.

Whenever we encounter a non-trivial constraint among different CPDs, we may assume that they are represented by a non-product constraint in the minimal decomposition (see definition of IPC). The parameterization could still satisfy a product constraint as well, but we will show that there are only 3 cases of such product constraints leading to CIs. They will be ruled out explicitly.

Lemma 6 *Given a Bayesian network defined over discrete variables satisfying MIN and IPC, any two adjacent variables* X *and* Y*, where* X *is a parent of* Y*, are dependent conditioned on any subset of the other parents of* Y.

Proof We will prove that $X \not\perp\!\!\!\perp Y \mid \boldsymbol{W}$ with $\boldsymbol{W} \subset Parents(Y) \setminus X$. We denote all other parents of Y by $\boldsymbol{U}$ (= $Parents(Y) \setminus X \setminus \boldsymbol{W}$). If $\boldsymbol{U}$ is empty, the dependency follows from MIN (otherwise the edge between X and Y can be removed without jeopardizing the factorization). Otherwise:

$$P(Y \mid X, \boldsymbol{W}) = \sum_{\boldsymbol{u}} P(\boldsymbol{u} \mid X, \boldsymbol{W}) P(Y \mid X, \boldsymbol{u}, \boldsymbol{W})$$

Independence $X \perp\!\!\!\perp Y \mid \boldsymbol{W}$ would imply (Eq. 10) that $\forall\, x_1, x_2$:

$$\begin{aligned} & P(\boldsymbol{u}_1 \mid x_1, \boldsymbol{W}) P(Y \mid x_1, \boldsymbol{u}_1, \boldsymbol{W}) + P(\boldsymbol{u}_2 \mid x_1, \boldsymbol{W}) P(Y \mid x_1, \boldsymbol{u}_2, \boldsymbol{W}) + \ldots \\ = & P(\boldsymbol{u}_1 \mid x_2, \boldsymbol{W}) P(Y \mid x_2, \boldsymbol{u}_1, \boldsymbol{W}) + P(\boldsymbol{u}_2 \mid x_2, \boldsymbol{W}) P(Y \mid x_2, \boldsymbol{u}_2, \boldsymbol{W}) + \ldots \end{aligned}$$

In this equation, $P(Y \mid x_i, \boldsymbol{u}_j, \boldsymbol{W}) \neq P(Y \mid x_k, \boldsymbol{u}_j, \boldsymbol{W})$ for at least one set of indices i, j, k since, by MIN, $P(Y \mid Parents(Y)) \neq P(Y \mid Parents(Y) \setminus X)$. The equation would therefore only hold when there is a constraint satisfied between CPD_Y and the CPDs defining $P(\boldsymbol{U} \mid X, \boldsymbol{W})$, which is ruled out by IPC. The latter follows from the postulate and the exclusion of non-minimality, the only case in which a product constraint can lead to independence $X \perp\!\!\!\perp Y$. ■

Lemma 7 *Assume that MIN and IPC holds for some Bayesian network defined over discrete variables. Let $X \not\perp\!\!\!\perp Y \mid \boldsymbol{Z'}$ where X, Y are arbitrary nodes and $\boldsymbol{Z'}$ is a set of nodes. Let $\boldsymbol{U}$ denote the set of all variables on the non-blocked paths between X and Y, including X and Y. Then, for every set $\boldsymbol{Z''} \subset \overline{\boldsymbol{Z'} \cup \boldsymbol{U}}$, when*

$$X \perp\!\!\!\perp Y \mid \boldsymbol{Z'}, \boldsymbol{Z''}$$

holds, there are deterministic relationships.

Proof We have to check when $P(Y \mid X, \boldsymbol{Z'}, \boldsymbol{Z''}) = P(Y \mid \boldsymbol{Z'}, \boldsymbol{Z''})$. Writing $P(Y \mid X, \boldsymbol{Z'})$ in function of the factors of the factorization results in a function containing the CPD_U's for all $U \in \boldsymbol{U}$ and factors describing $P(parU)$ with $parU \in Parents(U)$ for each U. Dependency $Y \not\perp\!\!\!\perp X \mid \boldsymbol{Z'}$ means that the distribution $P(Y \mid X, \boldsymbol{Z'})$ depends on X. Conditioning on $Z'' \in \boldsymbol{Z''}$ results in a new distribution for every value of Z''. We want to know when this can result in a distribution that becomes independent from X. We have to consider 3 different types of variables of $\boldsymbol{Z''}$ by looking how they participate in the CPDs containing variables of $\boldsymbol{U}$.

(1) Consider that Z'' is a member of a $Parents(U)$ set but no other parent of that U is in $\boldsymbol{U}$. Conditioning on Z'' will only change $P(U)$, which cannot affect the dependence between X and Y under IPC unless U is determined by Z'' which is discussed below.

(2) Consider that Z'' is a member of a $Parents(U)$ and there is another parent of U in $\boldsymbol{U}$. Then Z'' participates in a CPD of the form $P(U_1|U_2, Z'')$. Conditioning on Z'' results in a new CPD $P(U_1|U_2)$. This can only affect the dependence between X and Y if it makes U_1 independent from U_2. This is ruled out by Lemma 6 and MIN.

(3) If Z'' is a descendant of some U, X or Y, then we apply Bayes' theorem:

$$P(Y \mid X, \boldsymbol{Z'}, \boldsymbol{Z''}) = P(Y \mid X, \boldsymbol{Z'}) \frac{P(\boldsymbol{Z''} \mid X, Y, \boldsymbol{Z'})}{P(\boldsymbol{Z''} \mid X, \boldsymbol{Z'})} \tag{11}$$

Z" is not present in the first factor which depends on X. Therefore, for independence of X, a constraint between the three factors of Eq. 11 must be met. By construction, this will result in a constraint between the parameters of several CPDs, which is excluded by IPC or a deterministic relation must be present. This is proved in the following.

To get independence without a constraint among different CPDs, a factor of the form $P(K \mid L, \boldsymbol{M})$ must become equal to $P(K \mid \boldsymbol{M})$. This means that

$$P(K \mid l_1, \boldsymbol{M}) = P(K \mid l_2, \boldsymbol{M}) \ \ \forall l_1, l_2 \in L_{dom}$$

The conditional distribution of the left hand side is undefined whenever $P(l, \boldsymbol{m}) = 0$. If we rule out constraints leading to independence (IPC), independence can only happen when

$$P(l_1, \boldsymbol{m}) = 0 \ \ or \ \ P(l_2, \boldsymbol{m}) = 0 \ whenever \ P(K \mid l_1, \boldsymbol{m}) \neq P(K \mid l_2, \boldsymbol{m}). \tag{12}$$

The condition can only apply for one CPD (IPC), so it must be a condition on CPD_L or CPD_M which holds for all parameterizations of the other CPDs. Then Eq. 12 only holds if $P(l, \boldsymbol{m}) \neq 0$ for exactly one value of l given $\boldsymbol{m}$. It follows that $L = f(\boldsymbol{M})$. ■

Lemma 9 *If MIN, INDET, IPC holds for a factorization of a joint probability distribution defined over discrete variables, then for any v-structure $X \rightarrow Z \leftarrow Y$ in the DAG and for all $\boldsymbol{W}$ not containing X, Y or Z:*

$$\begin{aligned} &X \perp\!\!\!\perp Y \mid Z, \boldsymbol{W} \Leftrightarrow \ X \perp\!\!\!\perp Y \ \ and \ \ P(Z \mid X, Y) \ \ is \ \ a \ \ proportional \ \ CPD \\ &\forall U \ \ descendants \ \ of \ \ Z : X \not\perp\!\!\!\perp Y \mid U, \boldsymbol{W} \end{aligned} \tag{13}$$

Proof Assume that Y is not an ancestor of X (otherwise swap X and Y in the following). (1) First we prove the first equation for an empty set $\boldsymbol{W}$ and $U \neq Z$.

$$\begin{aligned} P(Y \mid X, Z) &= \frac{P(Z \mid X, Y).P(Y \mid X)}{P(Z \mid X)} \\ &= \frac{P(Z \mid X, Y).P(Y \mid X)}{\sum_{y'} P(Z \mid y', X) P(y' \mid X)} \end{aligned}$$

It follows that (using $A/B = 1/(B/A)$)

$$P(y \mid X, Z) = \frac{1}{1 + \frac{\sum_{y' \neq y} P(Z|y', X) P(y'|X)}{P(Z|y, X).P(y|X)}} \ \ \forall y$$

Then, for $X \perp\!\!\!\perp Y | Z$ Eq. 10 gives:

$$\frac{P(Z \mid y, x_1).P(y|x_1)}{\sum_{y' \neq y} P(Z \mid y', x_1) P(y'|x_1)} = \frac{P(Z \mid y, x_2).P(y|x_2)}{\sum_{y' \neq y} P(Z \mid y', x_2) P(y'|x_2)} \ \ \forall x_1, x_2$$

This equation defines a constraint between $P(Z \mid X, Y)$ and $P(Y \mid X)$ unless $Y \perp\!\!\!\perp X$ and the domain of Y contains only two values, y_1 and y_2. Then we have $P(Y \mid X) = P(Y)$ and $P(y_2) =$

$1 - P(y_1)$, which results in a constraint on $P(Z \mid X, Y)$:

$$\frac{P(Z \mid y_1, x_1).P(y_1)}{P(Z \mid y_2, x_1)(1 - P(y_1))} = \frac{P(Z \mid y_1, x_2).P(y_1)}{P(Z \mid y_2, x_2)(1 - P(y_1))} \ \forall x_1, x_2$$
$$\Rightarrow$$
$$\frac{P(Z \mid y_1, x_1)}{P(Z \mid y_2, x_1)} = \frac{P(Z \mid y_1, x_2)}{P(Z \mid y_2, x_2)} = \ldots = \frac{P(Z \mid y_1, x_n)}{P(Z \mid y_2, x_n)} \tag{14}$$

Which is a relation only depending on the parameterization of $P(Z \mid X, Y)$. In function of CPD_Z it gives:

$$P(Z|x, y) = \sum_{\boldsymbol{o}} P(\boldsymbol{o}|x, y)P(Z|x, y, \boldsymbol{o}) \; with \; \boldsymbol{o} \in domain \; of \; Parents(Z) \setminus X \setminus Y$$

Since Eq. 14 must hold independent of $P(\boldsymbol{o}|x, y)$ it follows that CPD_Z must be a *proportional CPT* (Eq. 9).

Next, we consider that the $\boldsymbol{W}$ is not empty. (2) For the other parents of Z, $\boldsymbol{W'} \subset \boldsymbol{W}$ it leads in a similar way to a pCPD:

$$\frac{P(Z \mid y_1, x_1, \boldsymbol{W'})}{P(Z \mid y_2, x_1, \boldsymbol{W'})} = \frac{P(Z \mid y_1, x_2, \boldsymbol{W'})}{P(Z \mid y_2, x_2, \boldsymbol{W'})} = \ldots = \frac{P(Z \mid y_1, x_n, \boldsymbol{W'})}{P(Z \mid y_2, x_n, \boldsymbol{W'})}.$$

(3) Conditioning on some other variables $\boldsymbol{W''} \subset \boldsymbol{W}$ cannot lead to independence by lemma 7.

(4) For proving the second equation (Eq. 13), we first prove it for an empty set. We write the equation for $P(Y \mid X, U)$ in function of $P(Z \mid X, Y)$ with Bayes' theorem:

$$P(Y \mid X, U) = \frac{P(U \mid X, Y).P(Y \mid X)}{P(U \mid X)}$$
$$P(Y \mid X, U) = \frac{\sum_z P(U \mid z, X, Y).P(z \mid X, Y).P(Y \mid X)}{\sum_z P(U \mid z, X).P(z \mid X)}$$

Since Y is not an ascendant of X, each of the equation's factors depends on different CPDs of the factorization. Independence of X cannot be obtained by a constraint on one CPD unless $Z = f(U)$ which is excluded by DET.

(5) If $X \not\perp\!\!\!\perp Y \mid U$ holds unconditionally, conditioning on some other variables $\boldsymbol{W}$ cannot lead to independence by lemma 7. ■

Theorem 10 *If for a given factorization of a JPD defined over discrete variables the conditions IPC, MIN, INDET and NOpCPD are met, then faithfulness holds:*

$$\forall \; disjoint \; subsets \; \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \; \subset \; \boldsymbol{V}: \quad \boldsymbol{X} \not\perp \boldsymbol{Y} \mid \boldsymbol{Z} \; \Leftrightarrow \; \boldsymbol{X} \not\perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

with $\boldsymbol{V}$ *the set of all variables under consideration.*

Proof We prove that $X \not\perp Y|\boldsymbol{Z} \Rightarrow X \not\perp\!\!\!\perp Y|\boldsymbol{Z}$. The proof recursively cuts the non-blocked paths between X and Y into subpaths until we end up with a basic connection that follows under lemma 6 or lemma 9. With these lemmas it will be proven that the variables at the outer ends of each subpath

are dependent under the 4 conditions. Secondly, we will prove that combining subpaths results in a dependence. To illustrate this consider $X \rightarrow U \rightarrow Y$. The path between X and Y is cut by U. First, $X \not\perp\!\!\!\perp U$ and $U \not\perp\!\!\!\perp Y$ are proven and, secondly, that the concatenation of subpaths $X \rightarrow U$ and $U \rightarrow Y$ leads to a $X \not\perp\!\!\!\perp Y$.

Without loss of generality, we may consider that X is not a descendant of Y. If X and Y are adjacent, Lemma 6 proves the dependency. If X and Y are part of a v-structure $X \rightarrow Z \leftarrow X$ of which Z or descendants of Z are in $\boldsymbol{Z}$, Lemma 9 proves the dependency. If they are part of several such v-structures, the same approach as in the proof of Lemma 9 can be used to prove dependency. Otherwise, take a minimal ordered cutset $\boldsymbol{U} \subset \boldsymbol{V} \setminus \{X, Y\}$ such that $X \perp Y \mid \boldsymbol{Z}, \boldsymbol{U}$. Minimal means that omitting any element of the set would lead to a d-connection. Ordered means that $U_i \in \boldsymbol{U}$ is not a descendant (in the DAG corresponding to the factorization) of $U_j \in \boldsymbol{U}$ whenever $i < j$. From $X \perp Y \mid \boldsymbol{Z}, \boldsymbol{U}$ follows $X \perp\!\!\!\perp Y \mid \boldsymbol{Z}, \boldsymbol{U}$, thus:

$$P(Y \mid X, \boldsymbol{Z}) = \sum_{u_1} \dots \sum_{u_n} P(Y \mid \boldsymbol{Z}, u_1, \dots, u_n).P(u_1 \dots u_n \mid X, \boldsymbol{Z})$$

$$= \sum_{u_1} \dots \sum_{u_n} P(Y \mid \boldsymbol{Z}, u_1, \dots, u_n).P(u_1 \mid X, \boldsymbol{Z}) \prod_{i=2}^{n} P(u_i \mid X, \boldsymbol{Z}, u_1 \dots u_{i-1}) \qquad (15)$$

We will prove that (1) no u_i can be removed from the first factor ($Y \not\perp\!\!\!\perp U_i \mid \boldsymbol{Z}, \boldsymbol{U} \setminus U_i$) and that X cannot be removed from the following factors ($X \not\perp\!\!\!\perp U_i \mid \boldsymbol{Z}, U_{i+1} \dots U_n$ for all values of i). Under these dependencies we will prove (2) that the right hand side of the equation does not lead to a distribution which is independent of X.

(1) The dependencies are of the form $K \not\perp\!\!\!\perp L \mid \boldsymbol{M}$. The d-connection of all dependencies, $K \not\perp L \mid \boldsymbol{M}$, follows from the minimality of the set $\boldsymbol{U}$. That the dependencies follow from the d-connections is proven in the same way as $X \perp Y \mid \boldsymbol{Z}$. The non-blocked paths between K and L are cut into subpaths until the nodes K and L are adjacent or connected by a v-structure. In that case lemmas 6 and 9 apply and prove the dependency.

(2) Because of the dependencies, Eq. 15 cannot be reduced. The first factor depends on CPD_Y, the following on CPD_{U_i}. Therefore, independence of X and Y can only come from a specific parameterization of at least three CPDs. ■

Theorem 12 *If for a given factorization of a JPD defined over discrete variables the conditions IC, COMPLEX, MIN, INDET and NOpCPD are met, then faithfulness holds:*

$$\forall \ disjoint\ subsets\ \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \ \subset \ \boldsymbol{V}: \quad \boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z} \ \Leftrightarrow \ \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

with $\boldsymbol{V}$ *the set of all variables under consideration.*

Proof In the previous proofs, constraints among parameters from different CPDs where ruled out by IPC and postulate 2. Conditions COMPLEX and IC will rule out the same constraints by Theorem 4. ■

Theorem 13 *Given P a distribution over n variables and a DAG G with CPDs parameterized by $\theta_1, \dots, \theta_n$ describing a factorization of P.*

If the parameter vectors satisfy IPC, there exists a parameterization $\theta'_1, \ldots, \theta'_n$ of the CPDs which has the same independencies as P and for which IC holds.
For discrete Bayesian networks and the linear case for continuous distributions, if the factorization satisfies IC and COMPLEX (the parameters θ_i^j of the parameter vectors $\theta_1, \ldots, \theta_n$ have non-constant complexity: $K(\theta_i^j | CPD_i^{\backslash \theta_i^j, *}) \stackrel{+}{>} 0$), then IPC holds as well.

Proof
The first statement, *if the parameter vectors satisfy IPC, there exists a parameterization $\theta'_1, \ldots, \theta'_n$ of the CPDs which has the same independencies as P and for which IC holds* (**IPC** $\rightarrow$ **IC**).
By IPC no constraint should hold among the parameters of different CPDs for the independencies of P. It could be that IC is violated for the given parameterization $\theta_1, \ldots, \theta_n$, but it is not necessary for any of the conditional independencies in P to hold. By IPC, the parameter subspace of every independence in P can cover the complete subspace or be a product subspace of the form $R_1 \times \cdots \times R_n$ where every R_i is an arbitrary subset of S_i, the set of possible parameter vectors θ_j. By Theorem 12 the subsets R_i have a generic form: some values of parameter vector θ_i are zero or similar (violation of MIN or presence of deterministic relationships) or determined by the other parameters (pCPD), but the remaining values of the CPDs are not constrained. Therefore we can choose the parameters independently in the subsets R_i such that the same independencies hold but IC is not violated.
Now the second statement, *if the factorization satisfies IC and COMPLEX, then IPC holds as well* (**IC and COMPLEX** $\rightarrow$ **IPC**).
Non-Markovian CIs require non-trivial polynomial constraints (Spirtes et al., 1993; Meek, 1995) among some parameters, which can be either parameters of a single CPD or among parameters of different CPDs. The first case is not a problem for IPC, while the latter would imply a violation of IC if COMPLEX, since Theorem 4 applies: polynomial constraints correspond to holomorphic functions having zeroes for a finite number of values only. As such, they have a constant description length, while we assumed that all parameters of the CPDs have non-constant complexity. ■

References

Gregory Chaitin. On the length of programs for computing finite binary sequences. *J. Assoc. Comput. Mach.*, 13:547–569, 1966.

Gregory Chaitin. A theory of program size formally identical to information theory. *J. Assoc. Comput. Mach.*, 22:329–340, 1975.

Zhitang Chen, Kun Zhang, Laiwan Chan, and Bernhard Schoelkopf. Causal discovery via reproducing kernel hilbert space embeddings. *Neural Computations*, 26:1484–1517, 2014.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. In *Procs of UAI-2010*, 2010.

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

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

Dominik Janzing and Bastian Steudel. Justifying additive noise model-based causal discovery via algorithmic information theory. *Open Syst. Inform. Dynam.*, pages 189–212, 2010.

Dominik Janzing, Xiaohai Sun, and Bernhard Schölkopf. Distinguishing cause and effect via second order exponential models. `http://arxiv.org/abs/0910.5561`, 2009.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniusis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 56(10):5168–5194, 2012.

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

Jan Lemeire and Dominik Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, 23(2):227–249, 2013. ISSN 0924-6495.

Jan Lemeire, Kris Steenhaut, and Abdellah Touhafi. When are graphical causal models not good models? In *Causality in the sciences, J. Williamson, F. Russo and P. McKay, editors, Oxford University Press*, 2011.

Jan Lemeire, Stijn Meganck, Francesco Cartella, and Tingting Liu. Conservative independence-based causal structure learning in absence of adjacency faithfulness. *Int. J. Approx. Reasoning*, 53(9):1305–1325, 2012.

Christopher Meek. Strong completeness and faithfulness in Bayesian networks. In *Procs of UAI-1995*, pages 411–418, 1995.

Joseph Ramsey, Jiji Zhang, and Peter Spirtes. Adjacency-faithfulness and conservative causal inference. In *Procs of UAI-2006*, pages 401–408, 2006.

Ray Solomonoff. A preliminary report on a general theory of inductive inference. *Technical report V-131*, Report ZTB-138 Zator Co., 1960.

Ray Solomonoff. A formal theory of inductive inference. *Information and Control, Part II*, 7(2): 224–254, 1964.

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