Causal Markov condition for submodular information measures

Bastian Steudel
Max Planck Institute for
Mathematics in the Sciences
Leipzig, Germany
steudel@mis.mpg.de

Dominik Janzing
Max Planck Institute for
Biological Cybernetics
Tübingen, Germany
janzing@tuebingen.mpg.de

Bernhard Schölkopf
Max Planck Institute for
Biological Cybernetics
Tübingen, Germany
schoelkopf@tuebingen.mpg.de

Abstract

The causal Markov condition (CMC) is a postulate that links observations to causality. It describes the conditional independences among the observations that are entailed by a causal hypothesis in terms of a directed acyclic graph. In the conventional setting, the observations are random variables and the independence is a statistical one, i.e., the information content of observations is measured in terms of Shannon entropy. We formulate a generalized CMC for any kind of observations on which independence is defined via an arbitrary submodular information measure. Recently, this has been discussed for observations in terms of binary strings where information is understood in the sense of Kolmogorov complexity. Our approach enables us to find computable alternatives to Kolmogorov complexity, e.g., the length of a text after applying existing data compression schemes. We show that our CMC is justified if one restricts the attention to a class of causal mechanisms that is adapted to the respective information measure. Our justification is similar to deriving the statistical CMC from functional models of causality, where every variable is a deterministic function of its observed causes and an unobserved noise term.

Our experiments on real data demonstrate the performance of compression based causal inference.

1 Introduction

Explaining observations in the sense of inferring the underlying causal structure is among the most important challenges of scientific reasoning. In practical applications it is generally accepted that causal conclusions can be drawn from observing the influence of interventions. The more challenging task, however, is to infer causal relations on the basis of non-interventional observations and research in this direction still is considered with skepticism. It is therefore important to thoroughly formalize the assumptions and discuss the conditions under which they are satisfied. For causal reasoning from statistical data, Spirtes et al. (2001) and Pearl (2000) formalized the assumptions under which the task is solvable. With respect to a causal hypothesis in terms of a directed acyclic graph (DAG) the most basic assumption is the causal Markov condition stating that every variable is conditionally independent of its non-descendants, given its parents,

$$x_j \perp\!\!\!\perp nd_j \,|pa_j \,,$$

for short. Pearl argues that this follows from a "functional model" of causality (or non-linear structure equations), where every node is a deterministic function of its parents pa_j and an unobserved noise term n_j (see Fig. 1), i.e.,

$$x_j = f_j(pa_j, n_j)\,. \tag{1}$$

The causal Markov condition is then a consequence of the statistical independence of the noise terms, which is called *causal sufficiency*. It can be justified by the assumption that every dependence between them requires a common cause (as postulated by Reichenbach (1956)), which should then explicitly appear in the causal model. From a more abstract point of view, condition (1) can be interpreted as saying that the node x_j does not add any more information that is not already contained in the parents and the noise together. If we restrict the assumption to discrete variables, the corresponding information measure can be, for instance, the Shannon entropy, but also other measures could make sense.

In (Janzing & Schölkopf, 2007) the probabilistic setting is generalized to the case where every observation is formalized by a binary string x_j (without any statistical population). The information content of an observation is then measured using Kolmogorov complexity (also "algorithmic information") which gives

rise to an algorithmic version of (conditional) mutual information. The corresponding functional model is given by a Turing machine that computes the string x_j from its parent strings pa_j and a noise n_j.
The algorithmic information theory based approach generalizes the statistical framework since the average algorithmic information content per instance of a sequence of i.i.d. observations converges to the Shannon entropy, but on the other hand observations need not be generated by i.i.d. sampling.

Unfortunately, Kolmogorov complexity is uncomputable and practical causal inference schemes must deal with other measures of information. In Section 2 we define general information measures and show that they induce independence relations that satisfy the semi-graphoid axioms (Section 3). Then, in Section 4, we phrase the causal Markov condition within our general setting and explore under which conditions it is a reasonable postulate. To this end, we formulate an information theoretic version of functional models observing that their decisive feature is that the joint information of a node, its parents and its noise is the same as the joint information of its parents and noise alone. We demonstrate with examples how these functional models restrict the set of allowed causal mechanisms to a certain class (Section 5). We emphasize that the choice of the information measure determines this class and is therefore the essential prior decision (which certainly requires domain knowledge). Thus, when applying our theory to real data, one first has to think about the causal mechanisms to be explored and then design an information measure that is sufficiently "powerful" to detect the generated dependences.
Section 6 discusses a modification for known independence based causal inference that is necessary for those information measures for which conditioning can only decrease dependences. Section 7 describes one of the most important intended applications of our theory, namely information measures based on compression schemes (e.g. Lempel-Ziv). Applications of these measures using the PC algorithm for causal inference to segments of English text demonstrate the strength of causal reasoning that goes beyond already known applications of compression for the purpose of (hierarchical) clustering.

2 General information measures

In this section we define information from an axiomatic point of view and prove properties that will be useful in the derivation of the causal Markov condition. We start by rephrasing the usual concept of measuring statistical dependences. Let $\mathcal{X}$ be a set of discrete-valued random variables and $\Omega := 2^{\mathcal{X}}$ be the set of subsets. For each $A \in \Omega$ let $H(A)$ denote the joint Shannon entropy of the variables in A. For three disjoint sets A, B, C the conditional mutual information between A and B given C then reads

$$I(A : B|C) := H(A \cup C) + H(B \cup C) - H(A \cup B \cup C) - H(C)\,. \tag{2}$$

The set of subsets constitutes a lattice $(\Omega, \vee, \wedge)$ with respect to the operations of union and intersection and H can be seen as a function on this lattice[1]. We observe that the non-negativity of (2) can be guaranteed if

$$H(D) + H(E) \geq H(D \vee E) + H(D \wedge E)\,,$$

for two sets $D, E \in \Omega$. This *submodularity* condition is known to be true for Shannon entropy (Cover & Thomas, 2006). Motivated by these remarks, we now introduce an abstract information measure defined on the elements of a general lattice. Throughout this paper let $(\Omega, \wedge, \vee)$ be a finite lattice and denote by 0 the meet of all of its elements.

Definition 1 (information measure)
We say $R : \Omega \to \mathbb{R}$ is an information measure *if it satisfies the following axioms:*

(1) normalization: $R(0) = 0\,,$

(2) monotonicity: $s \leq t$ *implies* $R(s) \leq R(t)$ *for all $s, t \in \Omega$,*

(3) submodularity: $R(s) + R(t) \geq R(s \vee t) + R(s \wedge t)$ *for all $s, t \in \Omega$.*

Note that submodular functions have been considered in different contexts, see for example (Lovász, 1983; Matus, 1994; Madiman & Tetali, 2008).
Based on R we define a conditional version for all $s, t \in \Omega$ by

$$R(s|t) := R(s \vee t) - R(t).$$

In analogy to (2), R gives rise to the following measure of independence:

Definition 2 (conditional mutual information) *For $s, t, u \in \Omega$ the conditional mutual information of s and t given u is defined by*

$$I(s : t\,|u) \quad := \quad R(s \vee u) + R(t \vee u) - R(s \vee t \vee u) - R(u).$$

We say s and t are independent given u or equivalently $s \perp\!\!\!\perp t\,|u$ *if* $I(s : t\,|u) = 0$.

[1] Also the information measures that are presented in this paper can all be rephrased as functions on the lattice of subsets it is nevertheless notationally convenient to formulate the theory with respect to general lattices.

Since the join on lattices is associative and commutative, for ease of notation we write $R(s,t,u,\ldots)$ instead of $R(s \vee t \vee u \vee \ldots)$ as well as $R(S) := R(s_1 \vee \ldots \vee s_n)$ for a subset $S = \{s_1,\ldots,s_n\} \subseteq \Omega$. Further $I(s_1,\ldots,s_n : u)$ is to be read $I((s_1 \vee \ldots \vee s_n) : u)$. The following Lemmas generalize usual information theory.

Lemma 1 (non-negativity of mutual information and conditioning) *For $s,t,u \in \Omega$ we have*

(a) $I(s : t\,|u) \geq 0$ *and* *(b)* $0 \leq R(s|t,u) \leq R(s|t)$.

Proof: (a) By definition, $I(s : t|u) \geq 0$ is equivalent to $R(s,u) + R(t,u) \geq R(s,t,u) + R(u)$. Defining $a = s \vee u$ and $b = t \vee u$ and using associativity of $\vee$ we have $a \vee b = s \vee t \vee u$. Further, using Lemma 4 in Ch.1 from (Birkhoff, 1995), in any lattice

$$a \wedge b = (s \vee u) \wedge (t \vee u) \geq u \vee (s \wedge t) \geq u$$

and hence by monotonicity of R: $R(a \wedge b) \geq R(u)$. Combining everything

$$R(s,u) + R(t,u) = R(a) + R(b) \geq R(a \vee b) + R(a \wedge b) \geq R(s,t,u) + R(u),$$

where the first inequality uses submodularity of R.
(b) The first inequality follows from (a) by $I(s : s|t,u) \geq 0$. The second inequality follows directly from (a) and the definition of I. □

Lemma 2 (chain rule for mutual information) *For $s,t,u,x \in \Omega$*

$$I(s : t \vee u\,|x) = I(s : t\,|x) + I(s : u\,|t,x)\,. \tag{3}$$

Proof: This is directly seen by using the definition of conditional mutual information on both sides.□

Lemma 3 (data processing inequality) *Given $s,t,x \in \Omega$ it holds*

$$R(s|t) = 0 \quad \Rightarrow \quad I(s : x\,|t) = 0 \quad \Rightarrow \quad I(s : x) \leq I(t : x).$$

Proof: The first implication is clear. For the second we apply the chain rule for mutual information two times and obtain

$$I(s : x) \quad = \quad I(s,t : x) - I(t : x\,|s) = I(t : x) + I(s : x\,|t) - I(t : x\,|s) \leq I(t : x)\,,$$

since the second summand is zero by assumption and conditional mutual information is non-negative. □

3 Submodular dependence measures and semi-graphoid axioms

The axiomatic approach to stochastic independence goes back to Dawid (1979) who stated four axioms of conditional independence that are fulfilled for any kind of probability distribution. Later, any relation I on triplets that satisfies the same axioms has been named semi-graphoid by Pearl (2000). It is easy to see that the function I constructed from R in the last section satisfies these axioms.

Theorem 1 (I satisfies semi-graphoid axioms) *The function I defined in the last section satisfies the semi-graphoid axioms, namely for $x,y,w,z \in \Omega$*

$$\begin{array}{llcll}
(1) & I(x : y\,|z) = 0 & \Rightarrow & I(y : x\,|z) = 0 & \textit{(symmetry)} \\
(2) & I(x : y,w\,|z) = 0 & \Rightarrow & \left\{ \begin{array}{l} I(x : y\,|z) = 0 \\ I(x : w\,|z) = 0 \end{array} \right. & \textit{(decomposition)} \\
(3) & I(x : y,w\,|z) = 0 & \Rightarrow & I(x : y\,|z,w) = 0 & \textit{(weak union)} \\
(4) & \left. \begin{array}{r} I(x : w\,|z,y) = 0 \\ I(x : y\,|z) = 0 \end{array} \right\} & \Rightarrow & I(x : w,y\,|z) = 0 & \textit{(contraction)}
\end{array}$$

Proof: Symmetry is clear and the remaining implications follow directly from the chain rule and non-negativity. □

On the contrary, if we are given a function $I : \Omega \times \Omega \times \Omega \to \mathbb{R}_+$, what axioms do we need to define a submodular information measure R from I? It turns out that the chain rule in eq. (3) together with non-negativity $I(a : b|c) \geq 0$ and symmetry $I(a : b|c) = I(b : a|c)$ already implies that $R(a) := I(a : a|0)$ is an information measure and I coincides with the dependence measure introduced in Definition 2. We omit the proof due to space constraints.

Thus we characterized the type of dependence measures that we are able to incorporate into our framework. To conclude, note that the chain rule is actually a strong restriction. As an example consider the lattice of linear subspaces of some finite vector space, where the join of two subspaces is the subspace generated by the set-theoretic union and the intersection is just the set-theoretic intersection. An independence measure can be defined by

$$I(a : b\,|c) = \dim\left(a_{|c^{\perp}}\right)_{|\left(b_{|c^{\perp}}\right)},$$

where $a_{|b}$ stands for the orthogonal projection of a onto b and $c^{\perp}$ denotes the orthogonal complement of c. This is a quantitative version of a notion of independence that satisfies the semi-graphoid axioms (Lauritzen, 1996) even though the chain rule does not hold.

4 Causal Markov condition for general information measures

In this section we define three versions of the causal Markov condition with respect to a general submodular information measure and show that they are equivalent (similar to the statistical framework). Then we discuss under which conditions we expect it to be a reasonable postulate that links observations with causality. Assume we are given observations $x_1, \dots, x_k$ that are connected by a DAG. It is no restriction to consider the observations as elements of a lattice, e.g. the lattice of their subsets.

Definition 3 (causal Markov condition (CMC), local version) *Let G be a DAG that describes the causal relations among observations $x_1, \dots, x_k$. Then the observations are said to fulfill the causal Markov condition with respect to the dependence measure I if*

$$I(nd_j : x_j\,|pa_j) = 0 \qquad \textit{for all}\;\; 1 \leq j \leq k,$$

where pa_j denotes the join of the parents of x_j and nd_j the join of its non-descendants (excluding the parents).

The intuitive meaning of the postulate is that conditioning on the direct causes of an observation screens off its dependences from all its non-effects. The following theorem generalizes results in (Lauritzen, 1996) for statistical independences and (Janzing & Schölkopf, 2007) for algorithmic independences. In particular it states that if the causal Markov condition holds with respect to a graph G, then independence relations implied by the CMC can be obtained through the convenient graph-theoretical criterion of d-separation (Pearl, 2000; Spirtes et al., 2001). Two sets of nodes A and B of a DAG are d-separated given a set C disjoint from A and B if every undirected path between A and B is blocked by C. A path that is described by the ordered tuple of nodes $(x_1, x_2, \dots, x_r)$ with $x_1 \in A$ and $x_r \in B$ is blocked if at least one of the following is true

(1) there is an i such that $x_i \in C$ and $x_{i-1} \rightarrow x_i \rightarrow x_{i+1}$ or $x_{i-1} \leftarrow x_i \leftarrow x_{i+1}$ or $x_{i-1} \leftarrow x_i \rightarrow x_{i+1}$,

(2) there is an i such that x_i and its descendants are not in C and $x_{i-1} \rightarrow x_i \leftarrow x_{i+1}$.

Theorem 2 (Equivalence of Markov conditions and information decomposition) *Let the nodes $x_1, \dots, x_k$ of a DAG G be elements of some lattice Ω and R be an information measure on Ω. Then the following three properties are equivalent*

(1) $x_1, \dots, x_k$ fulfill the (local) causal Markov condition.

(2) For every ancestral set[2] $A \subseteq \{x_1, \dots, x_k\}$, R decomposes according to G:

$$R(A) = \sum_{x_i \in A} R(x_i|pa_i).$$

(3) The global Markov condition holds, i.e., if two sets of nodes A and B are d-separated in G given a set C disjoint from A and B, then

$$\left(\bigvee_{a \in A} a\right) \perp\!\!\!\perp \left(\bigvee_{b \in B} b\right) \mid \left(\bigvee_{c \in C} c\right).$$

We omit the proof due to space constraints. The second condition shows that the joint information of observations can be recursively computed according to the causal structure. The third condition describes explicitly which sets of independences are implications of the causal Markov condition.

Our next Theorem will show that the CMC follows from a general notion of a functional model. At its basis is the following Lemma describing that the CMC on a given set of observations can be derived from the causal Markov condition with respect to an extended causal graph (see Figure 1).

[2] A set A of nodes of a DAG G is called ancestral, if for every $v \in A$ the parents of v are in A too.

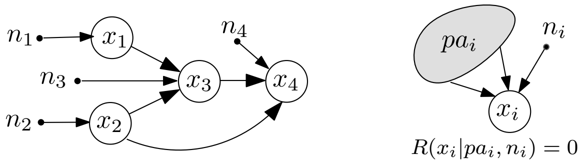


Figure 1: On the left a causal model of four observations $x_1,\dots,x_4$ is shown together with the 'noise' for each node. In Lemma 4 it is shown that the causal Markov condition on this extended graph implies the CMC for $x_1,\dots,x_4$. On the right hand side the functional model assumption is illustrated: The generation of x_i from its parents pa_i and the 'noise' does not produce additional information.

Lemma 4 (causal Markov condition from extended graph) *Let the nodes $x_1,\dots,x_k$ of a DAG G be elements of a lattice Ω with an independence relation I that is monotone and satisfies the chain rule. If there exist additional elements $n_1,\dots,n_k\in\Omega$ such that for all j*

$$I(x_j\,:\,nd_j,n_{-j}\,|\,pa_j,n_j)=0\,,\qquad \textit{where}\quad n_{-j}=\bigvee_{i\neq j} n_i, \tag{4}$$

and the n_j are jointly independent in the sense that

$$I(n_j:n_{-j})=0\,, \tag{5}$$

then the $x_1,\dots,x_k$ fulfill the causal Markov condition with respect to G.

Proof: Based on G we construct a new graph G' with node set $\{n_1,\dots,n_k\}\cup\{x_1,\dots,x_k\}$ and an additional edge $n_j\to x_j$ for every j, $(1\leq j\leq k)$. We first show that the causal Markov condition holds for the nodes of G': By construction, the join of non-descendants nd'_j of x_j with respect to G' is equal to $n_{-j}\vee nd_j$. Since the join of the parents pa'_j of x_j in G' are $pa_j\vee n_j$, assumption (4) just states $I(x_j:nd'_j|pa'_j)=0$ which is the local CMC with respect to x_j. To see that CMC also holds for n_j, observe that the non-descendants of n_j are equal to the non-descendants of x_j in G' and since n_j does not have any parents, we have to show

$$I(n_j:nd'_j)=0. \tag{6}$$

Using $nd'_j=n_{-j}\vee nd_j$ together with the chain rule for mutual information we get

$$I(n_j:nd_j,n_{-j})=I(n_j:n_{-j})+I(n_j:nd_j|n_{-j})=I(n_j:nd_j|n_{-j}),$$

where the last equality follows from (5). Let $ND_j=\{x_{j_1},\dots,x_{j_{k_j}}\}$ be the set of non-descendants of x_j in G. Note that ND_j is ancestral, that is if $x\in ND_j$, then so are the ancestors of x. We introduce a topological order on ND_j, such that if there is an edge $x_{j_a}\to x_{j_b}$ in G, then $x_{j_a}<x_{j_b}$. Using the chain rule for mutual information iteratively we get

$$I(n_j:nd_j|n_{-j})=\sum_{a=1}^{k_j} I\big(n_j:x_{j_a}|x^{(<)}_{j_a},n_{-j}\big),$$

where $x^{(<)}_{j_a}$ denotes the join of elements of ND_j smaller than x_{j_a}. By choice of our ordering the mutual information of n_j and x_{j_a} is conditioned at least on its parents and we can write $x^{(<a)}_{j_a}=pa_{j_a}\vee pa^c_{j_a}$, where $pa^c_{j_a}$ is the join of elements smaller than x_{j_a} in ND_j that are not its parents. Therefore, again by the chain rule, each summand on the right hand side can be bounded from above by writing

$$\begin{aligned} I\big(n_j:x_{j_a}|x^{(<a)}_{j_a},n_{-j}\big) &\leq I\big(n_{-j_a},pa^c_{j_a}:x_{j_a}|pa_{j_a},n_{j_a}\big)\\ &\leq I\big(n_{-j_a},nd_{j_a}:x_{j_a}|pa_{j_a},n_{j_a}\big)=0, \end{aligned}$$

where the second inequality is true because by construction $pa^c_{j_a}$ is the join of non-descents of x_{j_a}. The right hand side vanishes because of assumption (4). This proves (6) and therefore the causal Markov condition with respect to G'.
By Theorem 2, d-separation on G' implies independence. Due to the special structure of G' one can check that d-separation in G implies d-separation in the extended graph G'. Again by Theorem 2, d-separation implies the causal Markov condition for G, which proves the lemma. □

Now we formalize the intuition that in a generalized functional model a node only contains information that is already contained in the direct causes and the noise together (see Figure 1):

Definition 4 (functional model) *Let G be a DAG with nodes $x_1, \ldots, x_k$ in the lattice Ω. If there exists an additional node $n_j \in \Omega$ for each x_j, such that the n_j are jointly independent and*

$$R(x_j, pa_j, n_j) = R(pa_j, n_j) \qquad \text{for all } j, (1 \leq j \leq k) \tag{7}$$

then G together with $n_1, \ldots, n_k$ is called a functional model *of the $x_1, \ldots, x_k$.*

If we restrict our attention to causal mechanism of the above form, the CMC is justified:

Theorem 3 (functional model implies CMC) *If there exists a functional model for the nodes $x_1, \ldots, x_k$ of a DAG G then they fulfill the causal Markov condition with respect to G.*

Proof: In the functional model with noise nodes n_i it holds $R(x_j, pa_j, n_j) = R(pa_j, n_j)$ for all j. This implies $I(nd_j : x_j | pa_j, n_j) = 0$. Since the n_j in a functional model are assumed to be jointly independent, Lemma 4 can be applied and proves the theorem. □

The following section describes examples of causal mechanisms that can be seen as functional models with respect to various information measures.

5 Examples of information measures and their functional models

Let $S = \{x_1, \ldots, x_k\}$ be a finite set of observations which are in a canonical way elements of the lattice of subsets $(2^S, \cup, \cap)$. Let the causal structure be a DAG with $x_1, \ldots, x_k$ as nodes.

5.1 Shannon entropy of random variables

Let the x_i be discrete random variables with joint probability mass function $p(x_1, \ldots, x_k)$. For a subset $A \subseteq \{x_1, \ldots, x_k\}$ denote by $x_A := \times_{x_i \in A} x_i$ the random variable with distribution $p_A := p((x_i)_{x_i \in A})$. The Shannon entropy for the subset A is defined as $H(A) := -\mathbb{E}_p \log p_A$. Monotonicity as well as submodularity are well-known properties (Cover & Thomas, 2006). The corresponding notion of independence is the familiar (conditional) stochastic independence, its information-theoretic quantification I being mutual information. Then $H(x_i, pa_i, n_i) = H(pa_i, n_i)$ is equivalent to the existence of some function f_i with

$$x_i = f_i(pa_i, n_i) \, .$$

This restricts the set of mechanisms to those which were deterministic if one could take all latent factors into account. Note that continuous Shannon entropy is not monotone under restriction to subsets. Nevertheless, in this case the chain rule and non-negativity is true and therefore the CMC can be motivated by independences with respect to an extended causal model (Lemma 4 of the previous section).

5.2 Kolmogorov complexity of binary strings

Let the x_i be binary strings and the information measure be the Kolmogorov complexity as information measure. More explicitly, for a subset of strings $A \subseteq S$ denote by x_A a concatenation of the strings in a prefix free manner (which guarantees that the concatenation can be uniquely decoded into its components). The Kolmogorov complexity $K(x_A)$ is then defined as the length of the shortest program that generates the concatenated string x_A on a universal prefix-free Turing machine. It is submodular up to a logarithmic constant (Hammer et al., 2000). For two strings s, t the conditional Kolmogorov complexity $K(s|t)$ of s, given t is defined as the length of the shortest program that computes s from the input t. It must be distinguished from $K(s|t^*)$, the length of the shortest program that computes s from the shortest compression of t. Note that defining $R(s) := K(s)$ implies that the conditional information reads $R(s|t) = K(s|t^*)$ due to (Chaitin, 1975)

$$K(s, t) \overset{+}{=} K(t) + K(s|t^*) \, ,$$

see also (Gács et al., 2001). Then

$$K(x_i, pa_i, n_i) \overset{+}{=} K(pa_i, n_i) \quad \text{is equivalent to} \quad K(x_i | (pa_i, n_i)^*) \overset{+}{=} 0 \, ,$$

which, in turn, is equivalent to the existence of a program of length $O(1)$ that computes x_i from the shortest compression of (pa_i, n_i). Here we have considered the number k of nodes as a constant, which ensures that the order of the strings does not matter. Such an “algorithmic model of causality”(Janzing & Schölkopf, 2007) restricts causal influences to *computable* ones. Uncomputable mechanisms can easily be defined (as in the halting problem). However, in the spirit of the Church-Turing thesis, we will assume that they don’t exist in nature and conjecture that the algorithmic model of causality is the most general model of a causal mechanism as long as we restrict the attention to the non-quantum world (where the model could be replaced with a quantum Turing machine).

5.3 Period length of time series

We now present an example of an information measure on a lattice of observations different from the lattice of subsets. Let every observation be a natural number $x_i \in \mathbb{N}$ and consider them elements of the lattice of natural numbers where $\vee$ denotes the least common multiple and $\wedge$ the greatest common divisor, hence for $S \subseteq \{x_1, \ldots, x_k\}$

$$x_S := \vee_{x_i \in S} x_i := \mathrm{lcm}(S) \qquad .$$

We define an information measure by

$$R(x_S) := \log x_S \,.$$

Non-negativity and monotonicity of R are clear and submodularity even holds with equality: For $a, b \in \mathbb{N}$

$$\begin{aligned} R(a \vee b) + R(a \wedge b) &= \log \mathrm{lcm}(a,b) + \log \gcd(a,b) = \log \frac{ab}{\gcd(a,b)} + \log \gcd(a,b) \\ &= R(a) + R(b). \end{aligned}$$

The corresponding conditional dependence measure reads

$$I(a : b|c) = R\big(\gcd(a,b)/\gcd(a,b,c)\big) = \log \gcd(a,b) - \log \gcd(a,b,c),$$

so a and b are independent given c if c contains all prime factors that are shared by a and b (with at least the same multiplicity).

We define a functional model where every node x_i contains only prime factors that are already contained in its parents and its noise node (with at least the same multiplicity) and the noise terms are assumed to be relatively prime.

Such a lattice of observations can occur in real-life if x_i denotes the period length of a periodic time series over $\mathbb{Z}$. Then the period length of the joint time series defined by a set of nodes is obviously the least common multiple. If every time series at node i is a function F_i of its parents and noise node (each being a time series) and F_i is *time-covariant*, x_i divides their period lengths.

Assuming that the period lengths of the noise time series are relatively prime is indeed a strong restriction, but if we assume that the periods are large numbers and interpret independence in the approximate sense

$$\log \mathrm{lcm}(x_1, \ldots, x_k) \approx \sum_{i=1}^{k} \log x_i \,,$$

we obtain the condition that their periods have no *large* factors in common. This seems to be a reasonable assumption if the noise time series have no common cause.

One can easily think of generalizations where every observation x_i is characterized by a symmetry group and the join of nodes by the group intersection describing the joint symmetry. One may then define functional models where every node inherits all those symmetries that are shared by all its parents and the noise node.

5.4 Size of vocabulary in a text

Let every observation x_i be a text and for every collection of texts $S \subseteq \{x_1, \ldots, x_k\}$ let $R(S)$ be the number of different meaningful words in S. Here, meaningful means that we ignore words like articles and prepositions. To see that R is submodular we observe that it is just the number of elements of a set.

We can use R to explore which author has copied parts of the texts written by other authors: Let every x_i be written by another author and a causal arrow from x_i to x_j means that the author of x_i was influenced by x_i when writing x_j.

The noise n_i can be interpreted as the set of words the author usually uses and the condition $R(x_i, pa_i, n_i) = R(pa_i, n_i)$ then means that he/she combines only words from the texts he/she has seen with the own vocabulary.

To conclude this section we want to emphasize that the above example refers to a dependence measure that is non-increasing under conditioning, that is for collections S, T, U and V of texts $I(S : T|U) \geq I(S : T|V)$ whenever $U \subseteq V$. This is because $I(S : T|U)$ is equal to the number of meaningful words contained in S and T, but not in U.[3] We will elaborate on this point in the next section because it imposes special challenges for causal inference.

[3] In general, the above information measure can be viewed as rank or height function on the lattice of sets of meaningful words and it can be shown that dependence measures originating from information functions that are rank functions on distributive lattices are always non-increasing under conditioning.

6 Faithfulness for monotone dependence measures

Apart from the CMC, the essential postulate of independence based causal inference is usually *causal faithfulness*. It states that all observed independence relations are structural, that is, they are induced by the true causal DAG through d-separation. This postulate allows the identification of causal DAGs up to "Markov equivalence classes" imposing the same independences.

Faithfulness has already been defined for abstract conditional independence statements and we start by rephrasing the definition following (Spirtes et al. (2001), p.81).

Definition 5 (faithfulness) *A DAG G is said to represent a set of conditional independence relations $\mathcal{L}$ on a set of observations X* faithfully, *if $\mathcal{L}$ consists* exactly *of the independence relations implied by G through d-separation. Further, a set of observations X is said to be faithful (w.r.t. a given dependence measure), if there exists a causal DAG that represents X faithfully.*

The above definition of faithfulness makes sense for the probabilistic and algorithmic notions of dependence, but there is a problem with respect to dependence measures on which conditioning can only decrease information. As mentioned above, rank functions of distributive lattices lead to this kind of dependence measures, that we will call *monotone* in the following. To see the problem, consider for three observations a, b, c a causal DAG G of the form $a \rightarrow b \leftarrow c$. By d-separation, a is independent of c and for a monotone dependence measure this implies $a \perp\!\!\!\perp c \,|b$, which is not an independence induced by d-separation. Hence, G does not faithfully represent the objects and one can easily check that a faithful representation does not exist (e.g. using the theorem below). However, we can modify faithfulness such that it also accounts for those independences that follow from monotonicity under conditioning:

Definition 6 (monotone faithfulness) *A DAG G is said to represent a set $\mathcal{L}$ of conditional independences of observations X monotonically faithful, if the following condition is true for all disjoint subsets $S, T, U \subseteq X$ whose join is denoted by s, t and u: Whenever $s \perp\!\!\!\perp t \,|u$ is in $\mathcal{L}$ and u is* minimal *among all the sets that render s and t independent, then s and t are d-separated by u in G. Further, a set of observations X is said to be monotonically faithful (w.r.t. a given dependence measure), if there exists a causal DAG that represents X monotonically faithful.*

Note that, trivially, every faithful representation is a monotonically faithful representation, hence faithful observations are monotonically faithful observations. Faithful representations have already been characterized (Theorem 3.4 in (Spirtes et al., 2001)) and we prove an equivalent characterization that holds simultaneously for monotonically faithful and for faithful observations.

Theorem 4 (characterization of monotonically faithful representations) *A set of (monotonically) faithful observations X is represented (monotonically) faithfully by a DAG G if and only if (1) and (2) holds, where:*

(1) two observations a and b are adjacent in G if and only if they can not be made independent by conditioning on any join of observations in $X \backslash \{a, b\}$.

(2) for three observations a, b, c, such that a is adjacent to b, b is adjacent to c and a is not adjacent to c, it holds that $a \rightarrow b \leftarrow c$ in G if and only if there exists a set $U \subseteq X \backslash \{a, b, c\}$ such that a is independent of c given the join of the observations in U.

We omit the proof due to space constraints. The theorem implies in particular, that every monotonically faithful representation of faithful objects is already a faithful representation.
The PC algorithm (Spirtes & Glymour, 1991; Spirtes et al., 2001) for causal inference takes a set of conditional independences on faithful objects and returns the equivalence class of faithful representations. Since the above theorem is used to prove the correctness of the algorithm in the faithful case, we conclude that the algorithm correctly returns monotonically faithful representations given monotonically faithful observations. We apply the PC-algorithm with respect to compression based information functions in the following section. Also they are not monotone in a strict theoretical sense, empirical observations indicate that it is unlikely for the mutual information to increase.

7 Compression based information

In this section we demonstrate that our framework enables us to do causal inference on single *objects* (coded as binary strings) without relying on the uncomputable measure of Kolmogorov complexity. To this end, instead of defining complexity with respect to a universal Turing machine we explicitly limit ourselves to specific production processes of strings. The underlying measure of information is motivated by universal compression algorithms like LZ77 (Ziv & Lempel, 1977) and grammar based compression (Yang & Kieffer, 2000) that detect repeated occurrences of identical substrings within a given input string and encode them more efficiently. The choice of a compression scheme can be seen as a prior analogously to the choice of a universal Turing machine in the case of algorithmic information. The measures considered in this section quantifiy the information of an observation (string) in terms of the diversity of its substrings and entail the following assumption on causal processes: A mechanism that produces a string y from a string x is considered as simple, if it constructs y by concatenating a small number of substrings from x (see Lemma 6 below for a formal statement). Further, the amount of dependence of observations is approximately given by the number of substrings that they share.

We are going to describe two specific measures of information that are closely related to the total length of the compressed string, but have better formal properties than the latter. This way our conclusions will be independent of the actual implementation of the compression scheme and proving theoretical results gets easier.

In the last part of this section we describe experiments on real data in which the PC algorithm is applied to infer the causal structure using either of the two introduced measures of information.

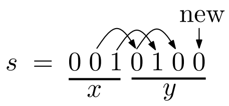
Note that *distance metrics* based on compression length have already been used to cluster various kinds of data (see (Cilibrasi & Vitányi, 2005) for computable distance metrics motivated by algorithmic mutual information or (Hanus et al., 2007) for an application to molecular biology). These metrics can be used to reconstruct trees (hierarchical clustering) but if two nodes are linked by more than one path a measure of conditional mutual information is needed to reconstruct the data-generation process. To the best of our knowledge, methods that rely on compression based *conditional* mutual information have not been used before to infer non-tree-like DAGs.

7.1 Lempel-Ziv information (LZ-information)

LZ-information has been introduced as a complexity measure for strings in (Ziv & Lempel, 1976). It has been applied to quantify the complexity of time series in biomedical signal analysis (Aboy et al., 2006) and distance measures based on versions of LZ-information have been used to analyze neural spike train data (Blanc et al., 2008) and to reconstruct phylogenetic trees (Zhen et al., 2009). We start by defining

Definition 7 (production and reproduction from prefix) *Let $s = xy$ be a string. We say s is* reproducible *from its prefix x and write $x \rightarrow s$ if y is a substring of $x\overline{y}$, where $\overline{y}$ is equal to y without its last symbol. We say s is* producible *from x and write $x \Rightarrow s$ if $x \rightarrow \overline{s}$, where $\overline{s}$ is equal to s without its last symbol.*

Contrary to reproducibility, producibility allows for the generation of new substrings, for if $x \Rightarrow s$, the last symbol of s can be arbitrary.



Example: For a given string $s = xy$ let $\overline{s}$ be the string without its last symbol. The figure on the left shows that $\overline{s}$ is producible from its prefix x by copying the second symbol of x to the first of y and so on. The string s itself is not producible from x, but reproducible.

Informally, LZ-information counts the minimal number of times during the process of parsing the input string from left to right, in which the string can not be reproduced from its prefix and a production step is needed.

Definition 8 (LZ-information, (Ziv & Lempel, 1976)) *Let s be a string of length n. Denote by s_i the i-th symbol of s and by $s(i,j)$ the substring $s_i s_{i+1} \cdots s_j$. A* production history *H_s of s is a partition of s into substrings $s = s(h_0,h_1)s(h_1+1,h_2)\cdots s(h_k+1,h_{k+1})$ with $h_0 = 1$ and $h_{k+1} = n$, such that*

$$s(1,h_i) \Rightarrow s(1,h_{i+1}) \quad \textit{for all } i \in \{1,\ldots,k\}.$$

A history H_s is called exhaustive *if additionally*

$$s(1,h_i) \not\rightarrow s(1,h_{i+1}) \quad \textit{for all } i \in \{1,\ldots,k-1\}.$$

The substrings $s(h_i+1,h_{i+1})$, $(0 \le i \le k)$ will be called components of H_s and the length $|H_s|$ of H_s is defined as the number of its components.

The LZ-information *of s, denoted by $c(s)$, is defined as the length of its (unique) exhaustive history.*

In an exhaustive history, each h_i is chosen maximal such that $s(1,h_i-1)$ is reproducible from its prefix $s(1,h_{i-1})$. As an example, for $s = 000100101100110$ the exhaustive history partitions s into

$$s = (0)(001)(00101)(10011)(0),$$

hence $c(s) = 5$.

In the original paper of Ziv and Lempel (1976) it was shown that c is subadditive: for two strings x and y the information of the concatenated string xy is at most the information of x plus the information of y. This already suggests to define the non-negative unconditional dependency measure $i(x:y) = c(x)+c(y)-c(xy)$. As it turns out, non-negativity of conditional information holds up to a negligible constant independent of the involved string lengths:

Lemma 5 (non-negativity of conditional LZ-information, asymmetric version) *Let x, y, z be finite strings over some alphabet $\mathcal{A}$. Further let α and β be symbols not contained in $\mathcal{A}$ that will be used as separators. Then*

$$i(x:y|z) := c(z\alpha x) + c(z\alpha y) - c(z\alpha x\beta y) - c(z) \geq -1. \qquad (8)$$

Proof: Let $E_{z\alpha}$ be the exhaustive history of $z\alpha$. The exhaustive history of $z\alpha x$ is of the form $E_{z\alpha x} = [E_{z\alpha}, E_{x|z}]$, where $E_{x|z}$ describes the partition of x induced by $E_{z\alpha x}$. This is because α is not part of the alphabet, hence the component in $E_{z\alpha x}$ containing α must be of the form $(t\alpha)$ for some substring t. Analogously $E_{z\alpha y} = [E_{z\alpha}, E_{y|z}]$. It is not difficult to see that

$$H_{z\alpha x\beta y} = [E_{z\alpha}, E_{x|z}, \beta, E_{y|z}].$$

is a production history of $z\alpha x\beta y$. Theorem 1 in (Ziv & Lempel, 1976) states that a production history is at least as long as the exhaustive history, hence

$$\big|[E_{z\alpha}, E_{x|z}, \beta, E_{y|z}]\big| \geq |E_{z\alpha x\beta y}| = c(z\alpha x\beta y),$$

Further, $c(z) \leq |E_{z\alpha}|$ and so (8) can be bounded from below by

$$\begin{aligned} c(z\alpha x) + c(z\alpha y) - c(z\alpha x\beta y) - c(z) &\geq \big|[E_{z\alpha}, E_{x|z}]\big| + \big|[E_{z\alpha}, E_{y|z}]\big| - \big|[E_{z\alpha}, E_{x|z}, \beta, E_{y|z}]\big| - |E_z| \\ &= -1. \end{aligned}$$

□

The above Lemma shows, that for two sets of strings $A = \{z, x\}$ and $B = \{z, y\}$ the LZ-information of $A \cup B$ and $A \cap B$ (represented by the information of strings $z\alpha x\beta y$ and z) exceeds the LZ-information of A and B (represented by the information of the strings $z\alpha x$ and $z\alpha y$) by at most one. This can be interpreted as approximate 'submodularity' with respect to A and B.

Within the functional models introduced before a node x_i was assumed to contain at most as much information as its parents pa_i and an independent noise n_i. The following Lemma states that if x_i is produced by concatenating complex substrings of pa_i and n_i, this is approximately the case with respect to LZ-information.

Lemma 6 (functional model for LZ-information, asymmetric version) *Let pa_i and n_i be two strings over an alphabet $\mathcal{A}$ and construct a third string string x_i by concatenating k substrings of pa_i and n_i. Then*

$$c(pa_i\,\alpha\,n_i\,\beta\,x_i) \leq c(pa_i\,\alpha\,n_i\beta) + k,$$

where α and β are symbols not in $\mathcal{A}$ used as separators.

Proof: A production history of $pa_i\alpha n_i\beta x_i$ can be generated by concatenating the exhaustive history of $pa_i\alpha n_i\beta$ with the list of the at most k substrings out of which x_i is constructed. The length of this history is $c(pa_i\alpha n_i) + k + 1$ and bounds $c(pa_i\alpha n_i\beta x_i)$ from above by Theorem 1 in (Ziv & Lempel, 1976). □

In particular, if xy is producible from x, by appending y, the information is at most increased by one. Hence, if we restrict the mechanisms that generate a node to consist of a limited number of concatenations of substrings from its parents and the independent noise (compared to the amounts of information involved) the causal Markov condition would follow if c were an information function. This is not the case since c is not defined on sets of strings (in particular it is not symmetric ($c(xy) \neq c(yx)$), therefore we define the LZ-information of a set of strings to be the LZ-information of their concatenation with respect to a given order (e.g. lexicographic).

Definition 9 (LZ-information, set version) *Let $\{x_1, \ldots, x_k\}$ be a set of strings over some alphabet $\mathcal{A}$. Choose k distinct symbols $\alpha_1, \ldots, \alpha_k$ not contained in $\mathcal{A}$ that will be used as separators.*
Let $X = \{x_{i_1}, \ldots, x_{i_m}\}$ be a subset and assume $x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_m}$ with respect to a given order on the set of strings over $\mathcal{A}$. We define the LZ-information of X as

$$LZ(X) = c\big(x_{i_1}\,\alpha_{i_1}\cdots x_{i_m}\,\alpha_{i_m}\big),$$

where the argument of c is understood as the concatenation of the strings.

LZ is not monotone and submodular in a strict sense. However, empirical observations suggest that for sufficiently large strings the violations of submodularity induced by the asymmetries like $c(x\alpha y) \neq c(y\alpha x)$ are negligible compared to the amounts of information.

Hypothesis: For practical purposes $LZ(\cdot)$ is an information measure up to constants that are negligible compared to the amounts of information of the strings involved. The associated independence measure I is monotonically decreasing (through conditioning).

We close by mentioning that the calculation of the LZ-information is very inefficient for large strings since one has to search over all substrings of the part of the string already parsed. In our implementation we therefore considered only substrings of length limited by a constant (we chose 30 for strings of English text, since it is unlikely that a substring of length 30 is repeated exactly).

7.2 Grammar based information

In the grammar based approach to compression an input string x is transformed into a context-free grammar that generates x. This grammar is then compressed for example using arithmetic codes. We discuss this approach because it has been successfully applied to compress RNA data (e.g. (Liu et al., 2008)). Further the LZ-based compression discussed in the previous section can be rephrased into this framework. As there are many grammars that produce a given string, it is essential that the transformation of strings to grammars produces economic representations of x (for an overview see (Lehman & Shelat, 2002)) We implemented the so called greedy grammar transform from Yang and Kieffer (2000). It constructs the grammar iteratively by parsing the input string x. Due to space restrictions we just give an example of a string and its generated grammar.
Example: The binary string $x = 1001110001000$ is transformed using the greedy grammar transform from Yang and Kieffer (2000) to the grammar $G(x)$:

$$\begin{aligned} s_0 &\rightarrow s_1 11 s_2 s_2 \\ s_1 &\rightarrow 100 \\ s_2 &\rightarrow s_1 0, \end{aligned}$$

where s_0, s_1 and s_2 are variables of the grammar and x can be reconstructed by starting from s_0 and then iteratively substituting s_i by the right hand side of each production rule above. The *length of a grammar* $|G(x)|$ is defined as the sum of all symbols on the right of every production rule, so for the above example $|G(x)| = 10$. We view the length of the constructed grammar as information measure of the string that it produces and define analog to the LZ-information

Definition 10 (grammar based information) *Let $\{x_1, \ldots, x_k\}$ be a set of strings over some alphabet $\mathcal{A}$. Choose k distinct symbols $\alpha_1, \ldots, \alpha_k$ not contained in $\mathcal{A}$ that will be used as separators.*
Let $X = \{x_{i_1}, \ldots, x_{i_m}\}$ be a subset and assume $x_{i_1} \leq x_{i_2} \leq \ldots \leq x_{i_m}$ with respect to a given order on the set of strings over $\mathcal{A}$. We define the grammar based information of X as

$$GR(X) = \left|G\left(x_{i_1}\, \alpha_{i_1} \cdots x_{i_m}\, \alpha_{i_m}\right)\right|,$$

where the input of the grammar construction G is understood as the concatenation of the strings.

By definition GR is non-negative. However, experiments show that submodularity is violated, but the amount of violation still allows to draw causal conclusions for sufficiently large strings.

7.3 Experiments

This section reports the results on causal inference using the introduced LZ-information and grammar based information measures. Matlab code of the algorithms used in the experiments can be downloaded from the homepage of the first author.

Experiment 1: Markov chains of English texts

We start with a string of English text s_0 from which we construct further strings $s_1, \ldots, s_k$ as follows: To generate s_{i+1} we translate s_i using an automatic translator from Google[4] to a randomly chosen European language. Then s_{i+1} is defined as the string that we obtain when we translate s_i back to English using the same translator. Since s_{i+1} is *determined* by s_i, the process can be modeled by a 'Markov' chain $s_0 \rightarrow \cdots \rightarrow s_k$. We then apply the PC algorithm[5] to infer the corresponding equivalence class of (monotonically) faithful causal models consisting of the DAGs:

$$s_0 \leftarrow \cdots \leftarrow s_i \rightarrow \cdots \rightarrow s_k \qquad \text{for} \quad 0 \leq i \leq k.$$

In our experiments we chose several starting texts of 1000 to 5000 symbols (e.g. news articles and the abstract of this paper) and generated three strings ($k = 3$) using the described procedure. In every string we transformed all non-space characters to numbers $0, \ldots, 8$ using a modulo operation on the ASCII value to reduce the alphabet size. Repeated spaces were deleted and the space character has been encoded separately by the number 9 to ensure that words of the string remain separated.
Results: Based on the two information measures, the PC algorithm returned the correct class of DAGs in every case. For LZ-information the chosen threshold used to determine independence did not even have to depend on the starting texts s_0. Grammar based information seems to be more sensitive to the string lengths involved and we had to choose a different threshold for every chosen text s_0. Further, we successfully tried the method on the chain of preliminary versions of the abstract of this paper.

[4] accessible at http://translate.google.de/

[5] Our implementation of the PC algorithm for causal inference was based on the BNT-Toolbox for Matlab written by Kevin Murphy and available at http://code.google.com/p/bnt/.

Finally note that methods based on compression distance could also be applied to recover the correct equivalence class. The crucial difference to our approach consists in the fact that we did not have to assume that the underlying graph is a tree.

Experiment 2: Four-node networks

We want to infer the equivalence classes of (monotonically) faithful causal models depicted in Figures (a) and (b) below. To this end we randomly choose segments of a large English text and then construct the strings corresponding to the nodes a, b, c and d in a way that ensures the resulting observation $\{a, b, c, d\}$ to be (monotonically) faithful. Explicitely, we choose segments s_x and s_{xy} for each node x and for each edge between nodes x and y respectively. Further, for every ordered triple of nodes (x, y, z) whose subgraph is not equal to $x \rightarrow y \leftarrow z$, we pick a segment s_{xyz}. This way we obtain the following segments with respect to the graph in Figure (a):

$$s_a, s_b, s_c, s_d, s_{ab}, s_{ac}, s_{bd}, s_{cd}, s_{bac}, s_{abd}, s_{acd}$$

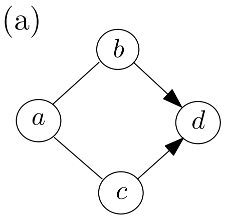
and with respect to the graph in Figure (b) we get segments

$$s_a, s_b, s_c, s_d, s_{ac}, s_{ad}, s_{bc}, s_{bd}, s_{cad}, s_{cbd}.$$

Finally, the string at a node is constructed as the concatenation of all segments that contain the name of the node in its index (the order is arbitrary), e.g. in the case of Figure (a)

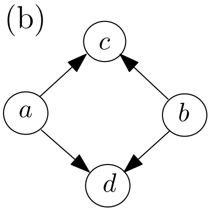
$$b = s_b s_{ab} s_{bd} s_{bac} s_{abd}.$$

As text source we chose an English version of Anna Karenina by Lev Tolstoi [6]. We then transformed all non-space characters to numbers from $0, \ldots, 8$ using a modulo operation on the ASCII value to reduce the size of the alphabet. Repeated spaces were deleted and the space character has been encoded separately by the number 9 to ensure that words of the string remain separated. The resulting string consisted of a total of approximately two million symbols. Using the above construction, we generated 100 observations $\{a, b, c, d\}$ with respect to each graph and applied the PC algorithm. The length N of the randomly chosen segments was chosen uniformly between 100 and 200 in the first run and between 300 and 500 in the second run. The choice of the threshold to determine independence depended only on the information measure and on the two possible ranges of N, but not on the individual observations. Further, the graph of Figure (b) implies an unconditional independence of a and b. Since two disjoint segments of English text can not be expected to be independent, we conditioned all informations that we calculate on background knowledge in terms of fixed segment of length 5000.



Correct answers of PC:

$N \in [100, 200]$	
$LZ:$	98%
$GR:$	53%
$N \in [300, 500]$	
$LZ:$	100%
$GR:$	56%



Correct answers of PC:

$N \in [100, 200]$	
$LZ:$	95%
$GR:$	97%
$N \in [300, 500]$	
$LZ:$	100%
$GR:$	99%

Results: Above, the percentages of correct results from the PC-algorithm are shown. Note that using LZ-information we were able to recover the correct equivalence class in almost all runs independently of the graph structure and segment length. Grammar based inference did not perform quite as well, but in the majority of cases in which it did not return the correct Markov equivalence class most of the independences still were detected correctly.

8 Conclusions

We have introduced conditional dependence measures that originate from submodular measures of information. We argued that these notions of conditional dependence (generalizing statistical dependence) can be used to infer the causal structure among observations even if the latter are not generated by i.i.d. sampling. To this end, we formulated a generalized causal Markov condition (with significant formal analogies to the statistical one) and proved that the condition is justified provided that the attention is restricted to a class of causal mechanisms that depends on the underlying measure of information. We demonstrated that existing compression schemes like Lempel-Ziv define interesting notions of information and described the class of mechanisms that justify the causal Markov condition in this case. Accordingly, we showed that the PC-algorithm successfully infers causal relations among texts when based on a notion of dependence that is induced by compression schemes.

[6] The text is available at http://www.gutenberg.org/etext/1399.

References

Aboy, M., Hornero, R., Abasolo, D., & Alvarez, D. (2006). Interpretation of the Lempel-Ziv Complexity Measure in the Context of Biomedical Signal Analysis. *IEEE Transactions on Biomedical Engineering, 53, issue 11*, 2282–2288.

Birkhoff, G. (1995). *Lattice theory*. American Mathematical Society. 3rd edition.

Blanc, J.-L., Schmidt, N., Bonnier, L., Pezard, L., & Lesne, A. (2008). Quantifying neural correlations using Lempel-Ziv complexity. *Deuxime confrence franaise de Neurosciences Computationnelles*.

Chaitin, G. J. (1975). A theory of program size formally identical to information theory. *J. ACM*, *22*, 329–340.

Cilibrasi, R., & Vitányi, P. M. B. (2005). Clustering by compression. *IEEE Transactions on Information Theory*, *51*, 1523–1545.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience. 2nd edition.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*, 1–31.

Gács, P., Tromp, J. T., & Vitányi, P. M. (2001). Algorithmic statistics. *IEEE Transactions on Information Theory*, *47*, 2443–2463.

Hammer, D., Romashchenko, A., Shen, A., & Vereshchagin, N. (2000). Inequalities for Shannon entropy and Kolmogorov complexity. *Journal of Computer and System Sciences*, *60*, 442 – 464.

Hanus, P., Dingel, J., Zech, J., Hagenauer, J., & Müller, J. C. (2007). Information theoretic distance measures in phylogenomics. *Proc. International Workshop on Information Theory and Applications (ITA 2007)*.

Janzing, D., & Schölkopf, B. (2007). Causal inference using the algorithmic Markov condition. `http://arxiv.org/abs/0804.3678`, to appear *in IEEE Transactions on Information Theory*.

Lauritzen, S. L. (1996). *Graphical models*. Oxford Statistical Science Series. Oxford University Press, USA.

Lehman, E., & Shelat, A. (2002). Approximation algorithms for grammar-based compression. *In Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms* (pp. 205–212). ACM/SIAM.

Liu, Q., Yang, Y., Chen, C., Bu, J., Zhang, Y., & Ye, X. (2008). RNACompress: Grammar-based compression and informational complexity measurement of RNA secondary structure. *BMC Bioinformatics*, *9*, 176.

Lovász, L. (1983). Submodular functions and convexity. *Mathematical Programming–The State of the Art*, *22*, 235–257.

Madiman, M., & Tetali, P. (2008). Information inequalities for joint distributions, with interpretations and applications.

Matus, F. (1994). Probabilistic conditional independence structures and matroid theory: Background. *Int. J. of General Systems*, *22*, 185–196.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Reichenbach, H. (1956). *The direction of time*. University of Califonia Press.

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, *9*, 62–72.

Spirtes, P., Glymour, C., & Scheines, R. (2001). *Causation, prediction, and search, second edition (adaptive computation and machine learning)*. The MIT Press.

Yang, E.-H., & Kieffer, J. C. (2000). Efficient universal lossless data compression algorithms based on a greedy sequential grammar transform. *IEEE Transactions on Information Theory*, *46*, 755–777.

Zhen, X., Li, C., & Wang, J. (2009). A complexity-based measure and its application to phylogenetic analysis. *Journal of Mathematical Chemistry*, *4*, 1149–1157.

Ziv, J., & Lempel, A. (1976). On the complexity of finite sequences. *IEEE Transactions on Information Theory*, *22*, 75–81.

Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, *23*, 337–343.