Testing Conditional Independence on Discrete Data using Stochastic Complexity

Alexander Marx
Max Planck Institute for Informatics,
and Saarland University
Saarbrücken, Germany
amarx@mpi-inf.mpg.de

Jilles Vreeken
CISPA Helmholtz Center for Information Security,
and Max Planck Institute for Informatics
Saarbrücken, Germany
jv@cispa.saarland

Abstract

Testing for conditional independence is a core aspect of constraint-based causal discovery. Although commonly used tests are perfect in theory, they often fail to reject independence in practice, especially when conditioning on multiple variables.

We focus on discrete data and propose a new test based on the notion of algorithmic independence that we instantiate using stochastic complexity. Amongst others, we show that our proposed test, SCCI, is an asymptotically unbiased as well as a consistent estimator for conditional mutual information (CMI). Further, we show that SCCI can be reformulated to find a sensible threshold for CMI that works well on limited samples. Empirical evaluation shows that SCCI has a lower type II error than commonly used tests. As a result, we obtain a higher recall when we use SCCI in causal discovery algorithms, *without* compromising the precision.

1 Introduction

Testing for conditional independence plays a key role in causal discovery (Spirtes et al., 2000). If the true probability distribution of the observed data is faithful to the underlying causal graph, conditional independence tests can be used to recover the undirected causal network. In essence, under the faithfulness assumption (Spirtes et al., 2000) finding that two random variables X and Y are conditionally independent given a set of random variables Z, denoted as $X \perp\!\!\!\perp Y \mid Z$, implies that there is no direct causal link between X and Y.

As an example, consider Figure 1. Nodes F and T are d-separated given D, E. Based on the faithfulness assumption, we can identify this from i.i.d. samples of the joint distribution, as F will be independent of T given D, E. In contrast, $D \not\perp\!\!\!\perp T \mid E, F$, as well as $E \not\perp\!\!\!\perp T \mid D, F$.

Conditional independence testing is also important for recovering the Markov blanket of a target node—i.e. the minimal set of variables, conditioned on which all other variables are independent of the target (Pearl, 1988). There exist classic algorithms that find the correct Markov blanket with provable guarantees (Margaritis and Thrun, 2000; Peña et al., 2007). These guarantees, however, only hold under the faithfulness assumption and given a *perfect* independence test.

In this paper, we are not trying to improve these algorithms, but rather propose a new independence test to enhance their performance. Recently a lot of work focuses on tests for continuous data; methods ranging from approximating continuous conditional mutual information (Runge, 2018) to kernel based methods (Zhang et al., 2011), we focus on discrete data.

For discrete data, two tests are frequently used in practice, the G^2 test (Aliferis et al., 2010; Schlüter, 2014) and conditional mutual information (CMI) (Zhang et al., 2010). While the former is theoretically sound, it is very restrictive as it has a high sample complexity; especially when conditioning on

Preprint. Work in progress.

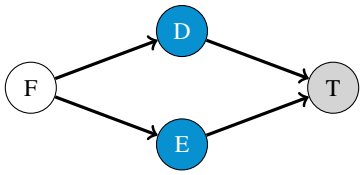


Figure 1: [d-Separation] Given the above causal DAG it holds that $F \perp\!\!\!\perp T \mid D, E$, or F is d-separated of T given D, E under the faithfulness assumption. Note that $D \not\perp\!\!\!\perp T \mid E, F$ and $E \not\perp\!\!\!\perp T \mid D, F$.

multiple random variables. When used in algorithms to find the Markov blanket, for example, this leads to low recall, as there it is necessary to condition on larger sets of variables.

If we had access to the true distributions, conditional mutual information would be the perfect criterium for conditional independence. Estimating CMI purely from limited observational data leads, however, to discovering spurious dependencies—in fact, it is likely to find no independence at all (Zhang et al., 2010). To use CMI in practice, it is therefore necessary to set a threshold. This is not an easy task, as the threshold should depend on both the domain sizes of the involved variables as well as the sample size (Goebel et al., 2005). Recently, Canonne et al. (2018) showed that instead of exponentially many samples, theoretically CMI has only a sub-linear sample complexity, although an algorithm is not provided. Closest to our approach is the work of Goebel et al. (2005) and Suzuki (2016). The former show that the empirical mutual information follows the gamma distribution, which allows them to define a threshold based on the domain sizes of the variables and the sample size. The latter employs an asymptotic formulation to determine the the threshold for CMI.

The main problem of existing tests is that these struggle to find the right balance for limited data: either they are too restrictive and declare everything as independent or not restrictive enough and do not find any independence. To tackle this problem, we build upon algorithmic conditional independence, which has the advantage that we not only consider the statistical dependence, but also the complexity of the distribution. Although algorithmic independence is not computable, we can instantiate this ideal formulation with stochastic complexity. In essence, we compute stochastic complexity using either factorized or quotient normalized maximum likelihood (fNML and qNML) (Silander et al., 2008, 2018), and formulate SCCI, the ***S**tochastic complexity based **C**onditional **I**ndependence criterium*.

Importantly, we show that we can reformulate SCCI to find a natural threshold for CMI that performs well given limited data and diminishes given enough data. In the limit, we prove that SCCI is an asymptotically unbiased and consistent estimator of CMI. For limited data, we find that the qNML threshold behaves similar to Goebel et al. (2005)—i.e. it considers the sample size as well as the dimensionality of the data. The fNML threshold, however, additionally considers the estimated probability mass functions of the conditioning variables. In practice, this reduces the type II error. Moreover, when applying SCCI based on fNML in constraint based causal discovery algorithms, we observe a higher precision and recall than related tests. In addition, in our empirical evaluation SCCI shows a sub-linear sample complexity.

In this work we build upon and extend the basic ideas we first presented as (Marx and Vreeken, 2018). Here we specifically focus on the theory and properties of using stochastic complexity for measuring conditional independence. Those readers that are interested in how SCI can be used in the discovery of directed Markov blankets we refer to (Marx and Vreeken, 2018).

For conciseness, we postpone some proofs and experiments to the supplemental material. For reproducibility of our experiments we make our code available online[1] and released an efficient version of SCCI in the R-package *SCCI*.

2 Conditional Independence Testing

In this section, we introduce the notation and give brief introductions to both standard statistical conditional independence testing, as well as to the notion of algorithmic conditional independence.

Given three possibly multivariate random variables X, Y and Z, our goal is to test the conditional independence hypothesis $H_0 \colon X \perp\!\!\!\perp Y \mid Z$ against the general alternative $H_1 \colon X \not\perp\!\!\!\perp Y \mid Z$, where

[1] `https://eda.mmci.uni-saarland.de/sci`

X, Y and Z denote possibly multivariate discrete random vectors with finite sample spaces (or domain sizes) $\mathcal{X}, \mathcal{Y}$ and $\mathcal{Z}$. In the context of independence testing in causal graphs, the random variables X and Y would be univariate in most cases, while Z could refer to a set of random variables. To not make the notation overly complicated, we will only use the set notation, e.g. $\boldsymbol{Z}$, when we want to clarify that we are referring to random variables that represent multiple nodes in a causal graph.

A perfect independence test minimizes both the type I error, that is, falsely rejecting the null hypothesis, as well as the type II error—i.e., falsely accepting the null hypothesis. A high type I error will lead to finding spurious edges in a causal discovery setup while having a high type II error means that we will miss out on true edges. A well-known measure for conditional independence is *conditional mutual information* (CMI) based on *Shannon entropy* (Cover and Thomas, 2006). The Shannon entropy of a possibly multivariate discrete random variable X with probability mass p is defined as

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \ ,$$

and the conditional Shannon entropy of a discrete random variable X given Y is defined as

$$H(X \mid Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} \ .$$

Using these definitions, we can define conditional mutual information as

$$I(X; Y \mid Z) = H(X \mid Z) - H(X \mid Z, Y) \ ,$$

where $X \perp\!\!\!\perp Y \mid Z$ iff $I(X; Y \mid Z) = 0$.

If we are given the true distribution of a random variable, CMI is ideal to test for conditional independencies on discrete data. In practice, we need to work with a limited sample size. On such a limited sample the plug-in estimator $\hat{H}$ tends to underestimate conditional entropies, and as a consequence, conditional mutual information is overestimated—even for completely independent data, as the following example shows.

Example 1 *Given three random variables X_1, X_2 and Y, with respective domain sizes $1\,000, 8$ and 4. Suppose that we are given 500 samples drawn from their joint distribution and find that $\hat{H}(Y \mid X_1) = \hat{H}(Y \mid X_2) = 0$. That is, Y is a deterministic function of X_1, as well as of X_2. However, as $|\mathcal{X}_1| = 1\,000$, and given only 500 samples, it is likely that a large fraction of values $v \in \mathcal{X}_1$ is only assigned to a single data point. Thus, finding that $\hat{H}(Y \mid X_1) = 0$ is likely due to the limited amount of samples, rather than that it depicts a true (functional) dependency, while $\hat{H}(Y \mid X_2) = 0$ is more likely to be due to a true dependency, since the number of samples $n \gg |\mathcal{Y}| \times |\mathcal{X}_2|$—i.e., the support for each value in the sample space is $\gg 1$.*

A possible solution is to set a threshold t such that $X \perp\!\!\!\perp Y \mid Z$ if $\hat{I}(X; Y \mid Z) \leq t$. Setting t is, however, not an easy task, as t is dependent on the quality of the entropy estimate, which by itself strongly depends on the complexity of the distribution and the given number of samples. Instead, to avoid this problem altogether, we will base our test on the notion of *algorithmic* independence.

2.1 Algorithmic Independence

To define algorithmic independence, we first need to briefly introduce to Kolmogorov complexity (Kolmogorov, 1965; Li and Vitányi, 1993).

Definition 1 ((Prefix) Kolmogorov Complexity) *The (prefix) Kolmogorov complexity of a finite binary string x is the length of the shortest self-delimiting binary program p^* for a universal prefix Turing machine $\mathcal{U}$ that generates x, and then halts. Formally, we have*

$$K(x) = \min\{|p| \mid p \in \{0,1\}^*, \mathcal{U}(p) = x\} \ .$$

In other words, program p^* is the most succinct *algorithmic* description of x, or the ultimate lossless compressor for that string. Importantly, the above definition defines prefix Kolmogorov complexity. There also exists plain Kolmogorov complexity, for which the program does not have to be self-delimiting, that is, it is assumed that the machine knows where codewords start and end. In this paper,

this distinction does not make a difference, however, for consistency we use prefix Kolmogorov complexity.

To define algorithmic independence, we will also need conditional (prefix) Kolmogorov complexity, that is,

$$K(x \mid y) = \min\{|q| \mid q \in \{0,1\}^*, \mathcal{U}(y,q) = x\} \ .$$

In essence, we still try to find the shortest program that outputs x, but we get y as an additional input. Similar to Shannon entropy, providing more information can only reduce the length of the program and hence $K(x \mid y) \leq K(x) + \mathcal{O}(1)$, where the constant does not depend on x or y. It is common to avoid writing the additional constant complexity term and instead write $\overset{+}{\leq}$ or $\overset{\pm}{=}$ to indicate that the inequality respectively the equality holds up to an additive constant.

By definition, Kolmogorov complexity makes maximal use of any effective structure in x; structure that can be expressed more succinctly algorithmically than by printing it verbatim. As such it is the theoretical optimal measure for complexity. In the context of our problem, instead of purely considering the statistical dependence between random variables, it also considers the complexity of the generating mechanism that induces the dependence.

Let us consider Example 1 again and let x_1, x_2, and y be the binary strings representing of the samples drawn from X_1, X_2 and Y. As Y can be expressed as a deterministic function of X_1 or X_2, $K(y \mid x_1)$ and $K(y \mid x_2)$ reduce to the programs describing the corresponding function. As the domain size of X_2 is 8 and $|\mathcal{Y}| = 4$, the program that describes the function from X_2 to Y only has to describe the mapping from 8 to 4 values, which will be shorter than describing a mapping from X_1 to Y, since $|\mathcal{X}_1| = 1\,000$—i.e., $K(y \mid x_2) \overset{+}{\leq} K(y \mid x_1)$ in contrast $\hat{H}(Y \mid X_1) = \hat{H}(Y \mid X_2)$.

To reject $X \perp\!\!\!\perp Y \mid Z$, we test whether providing information about Y and Z leads to a shorter program than only knowing Z (Chaitin, 1975).

Definition 2 (Algorithmic CMI) *Given the strings x, y and z, we write z^* to denote the shortest program for z, and analogously $(z,y)^*$ for the shortest program for the concatenation of z and y. Algorithmic conditional mutual information is defined as*

$$I_A(x; y \mid z) = K(x \mid z^*) - K(x \mid (z,y)^*) \ .$$

Similar to CMI, we say that x is algorithmically independent of y given z iff $I_A(x; y \mid z) \overset{\pm}{=} 0$. Due to the halting problem Kolmogorov complexity is, however, not computable nor approximable up to arbitrary precision (Li and Vitányi, 1993). The Minimum Description Length (MDL) principle (Grünwald, 2007) provides a statistically well-founded approach to approximate Kolmogorov complexity from above. For discrete data, this means we can use the stochastic complexity for multinomials (Kontkanen and Myllymäki, 2007), which belongs to the class of refined MDL codes.

3 Stochastic Complexity for Multinomials

In the following, we will define stochastic complexity for multinomials, which belongs to the class of refined MDL codes. On a high level, we will define stochastic complexity as the negative logarithm of the *normalized maximum likelihood* (NML), which has several nice properties.

Let X be a discrete random variable with $|\mathcal{X}| = k$, where we assume that X can be modelled by a parametric distribution P_θ, with parameter vector $\theta = (\theta_1, \ldots, \theta_k)$. Further, we denote all distributions that can be described with such a k-dimensional parameter θ by $\mathcal{M}_k$. Given a sample x^n of n data points drawn w.r.t. P_θ we denote the maximum likelihood (ML) estimate of θ w.r.t. to x^n by $\hat{\theta}(x^n)$. Shtarkov (1987) defined the NML density function as

$$f_{NML}(X \mid \mathcal{M}_k) = \frac{f_{\hat{\theta}(x^n)}(x^n)}{\mathcal{C}^n_{\mathcal{M}_k}} \ , \tag{1}$$

where $f_{\hat{\theta}(x^n)}$ is the empirical density function for X based on the maximum likelihood estimate $\hat{\theta}(x^n)$ under the model class $\mathcal{M}_k$. The normalizing factor, or regret $\mathcal{C}^n_{\mathcal{M}_k}$, relative to the model class $\mathcal{M}_k$ is defined as

$$\mathcal{C}^n_{\mathcal{M}_k} = \sum_{\tilde{x}^n \in \mathcal{X}^n} f_{\hat{\theta}(\tilde{x}^n)}(\tilde{x}^n) \, .$$

The sum iterates over every possible sample $\tilde{x}^n$ of length n and sample space $\mathcal{X}$, and for each considers the ML estimate for that data given model class $\mathcal{M}_k$. Whenever clear from context, we will drop $\mathcal{M}_k$ to simplify the notation—i.e., we write $P_{NML}(x^n)$ for $P_{NML}(x^n \mid \mathcal{M}_k)$ and let $\mathcal{C}_k^n$ to refer to $\mathcal{C}_{\mathcal{M}_k}^n$.

Notably, as shown by **?** the NML distribution incorporates all information in the data that can be extracted with the models in the model class $\mathcal{M}_k$. Moreover, the NML distribution is the optimal encoding w.r.t. the model class even if the data was generated by a model outside $\mathcal{M}_k$. The latter was formally shown by **?**, who proved that besides solving Shtarkov's minimax problem (Shtarkov, 1987), the NML distribution is also the solution to the minimax problem

$$\inf_q \sup_g E_g \log \frac{f_{\hat{\theta}(x^n)}(x^n)}{q(x^n)} \ ,$$

where the distributions q and g can range over virtually any distribution—i.e., g can be a distribution outside the model class.

Important for us is that for discrete data, we can assume the model class to be the class of multinomial distributions. Under this assumption, we can rewrite Equation (1) as (Kontkanen and Myllymäki, 2007)

$$f_{NML}(x^n) = \frac{\prod_{j=1}^k \left(\frac{c_j}{n}\right)^{c_j}}{\mathcal{C}_k^n} \ ,$$

where c_j is the empirical frequency of the j-th value in the sample space $\mathcal{X}$ in x^n. Respectively we can compute the regret as

$$\mathcal{C}_k^n = \sum_{c_1+\cdots+c_k=n} \frac{n!}{c_1! \cdots c_k!} \prod_{j=1}^k \left(\frac{c_j}{n}\right)^{c_j} \ .$$

Fortunately, Mononen and Myllymäki (2008) derived a formula to calculate the regret in sub-linear time, meaning that the whole formula can be computed in linear time w.r.t. n.

Building upon the definition of f_{NML}, we obtain the *stochastic complexity* of a discrete random variable X based on a sample x^n by simply taking the negative logarithm[2]—i.e.,

$$\begin{aligned} S(X) &= -\log f_{NML}(x^n) \ , \\ &= n\hat{H}(X) + \log \mathcal{C}_k^n \ . \end{aligned}$$

As a result, we see that the stochastic complexity decomposes into n times the empirical entropy and the log regret, which is also called *parametric complexity*.

3.1 Conditional Stochastic Complexity

Conditional stochastic complexity can be defined in different ways. We consider factorized normalized maximum likelihood (fNML) (Silander et al., 2008) and quotient normalized maximum likelihood (qNML) (Silander et al., 2018), which are equivalent except for the regret terms.

Given an empirical sample over two random vectors X and Y, conditional stochastic complexity using fNML is defined as

$$\begin{aligned} S_f(X \mid Y) &= -\sum_{y \in \mathcal{Y}} \log f_{NML}(x^n \mid y^n = y) \\ &= n\hat{H}(X \mid Y) + \sum_{y \in \mathcal{Y}} \log \mathcal{C}_{|\mathcal{X}|}^{c_y} \ , \end{aligned}$$

where c_y corresponds to the number of samples for which $Y = y$. Analogously, we define conditional stochastic complexity using qNML (Silander et al., 2018)

$$\begin{aligned} S_q(X \mid Y) &= -\log \frac{f_{NML}(x^n, y^n)}{f_{NML}(y^n)} \\ &= n\hat{H}(X \mid Y) + \log \frac{\mathcal{C}_{|\mathcal{X}|\cdot|\mathcal{Y}|}^n}{\mathcal{C}_{|\mathcal{Y}|}^n} \ . \end{aligned}$$

[2]As is common for MDL encodings, we want to obtain a code-length in terms of bits and hence compute the logarithm with respect to basis 2 and define $0 \log 0 = 0$.

In the following, we refer to conditional stochastic complexity as S and only use S_f or S_q whenever there is a conceptual difference. Further, we denote to the regret term of $S(X)$ as $\mathcal{R}(X) = \log C^n_{|\mathcal{X}|}$ and respectively refer to the regret of $S(X \mid Y)$ as $\mathcal{R}(X \mid Y)$, where

$$\mathcal{R}_f(X \mid Y) = \sum_{y \in \mathcal{Y}} \log \mathcal{C}^{c_y}_{|\mathcal{X}|} \; and$$

$$\mathcal{R}_q(X \mid Y) = \log \frac{\mathcal{C}^n_{|\mathcal{X}| \cdot |\mathcal{Y}|}}{\mathcal{C}^n_{|\mathcal{Y}|}} \; .$$

Next, we show that $\mathcal{C}^n_k$ is log-concave in n, which is a property we need to guarantee that our estimator is always smaller or equal than the empirical estimator $\hat{I}(X; Y \mid Z)$.

Lemma 1 *For $n \geq 1$, the regret $\mathcal{C}^n_k$ of the multinomial stochastic complexity of a random variable with a domain size of $k \geq 2$ is log-concave in n.*

For readability, we postpone the proof of Lemma 1 to Appendix A. In the following theorem, we present the first implication of this Lemma.

Theorem 1 *Given three discrete random variables X, Y and Z with domain sizes ≥ 2, it holds that $\mathcal{R}(X \mid Z) \leq \mathcal{R}(X \mid Z, Y)$.*

Proof: We start by proving the statement for $\mathcal{R}_f$. Consider that Z contains p distinct value combinations $\{r_1, \ldots, r_p\}$. If we add Y to Z, the number of distinct value combinations, $\{l_1, \ldots, l_q\}$, increases to q, where $p \leq q$. Consequently, to show the claim, it suffices to show that

$$\sum_{i=1}^{p} \log \mathcal{C}^{c_i}_k \leq \sum_{j=1}^{q} \log \mathcal{C}^{c_j}_k \tag{2}$$

where $\sum_{i=1}^{p} c_i = \sum_{j=1}^{q} c_j = n$. Next, consider w.l.o.g. that each value combination $\{r_i\}_{i=1,\ldots,p}$ is mapped to one or more value combinations in $\{l_1, \ldots, l_q\}$. Hence, Equation (2) holds, if $\log \mathcal{C}^n_k$ is sub-additive in n. Since we know from Lemma 1 that the regret term is log-concave in n (since both $p, q \geq 2$), sub-additivity follows by definition.

Next, consider $\mathcal{R}_q$. Let k, p and q be the domain sizes of X, Y and Z, we need to show that

$$\begin{aligned} &\mathcal{R}_q(X \mid Z) \leq \mathcal{R}_q(X \mid Z, Y) \\ \Leftrightarrow &\log \frac{\mathcal{C}^n_{kq}}{\mathcal{C}^n_q} \leq \log \frac{\mathcal{C}^n_{kpq}}{\mathcal{C}^n_{pq}} \; . \end{aligned}$$

We know from Silander et al. (2018) that for $p \in \mathbb{N}, p \geq 2$, the function $q \mapsto \frac{\mathcal{C}^n_{p \cdot q}}{\mathcal{C}^n_q}$ is increasing for every $q \geq 2$. This suffices to prove the statement above. □

4 Stochastic Complexity based Conditional Independence

With the above, we can formulate our new conditional independence test, which we will refer to as the ***S**tochastic **C**omplexity based **C**onditional **I**ndependence criterium*, or SCCI.

Definition 3 (SCCI) *Let X, Y and Z be discrete random vectors,* SCCI *is defined as*

$$\begin{aligned} \text{SCCI}(X; Y \mid Z) &= S(X \mid Z) - S(X \mid Z, Y) \\ &= n\hat{I}(X; Y \mid Z) + \mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z, Y) \; . \end{aligned} \tag{3}$$

Further, we say that $X \perp\!\!\!\perp Y \mid Z$ if $\text{SCCI}(X; Y \mid Z) \leq 0$.

From the second row in Equation 3, we see that the regret terms formulate a natural threshold t_S for the empirical estimate of CMI, where $t_S = \mathcal{R}(X \mid Z, Y) - \mathcal{R}(X \mid Z)$. From Theorem 1 we know that if we instantiate SCCI using fNML or qNML, we are guaranteed that $\mathcal{R}(X \mid Z, Y) - \mathcal{R}(X \mid Z) \geq 0$. Hence, Y has to provide a significant gain such that $\text{SCCI}(X; Y \mid Z) > 0$—i.e., we need $\hat{H}(X \mid Z) - \hat{H}(X \mid Z, Y) > t_S/n$. In other words, if

$$\hat{I}(X; Y \mid Z) \leq \frac{t_S}{n} \; ,$$

we would consider X and Y to be independent given Z. Thus, it is obvious that no matter what formulation of conditional stochastic complexity we choose, SCCI is more restrictive than the empirical estimator of CMI.

4.1 Factorized SCCI

To formulate our independence test based on factorized normalized maximum likelihood, we have to revisit the regret terms again. In particular, $\mathcal{R}_f(X \mid Z)$ is only equal to $\mathcal{R}_f(Y \mid Z)$, when the domain size of X is equal to the domain of Y. Further, $\mathcal{R}_f(X \mid Z) - \mathcal{R}_f(X \mid Z, Y)$ is not guaranteed to be equal to $\mathcal{R}_f(Y \mid Z) - \mathcal{R}_f(Y \mid Z, X)$. Consequently, our test would not be symmetric. Hence, we formulate SCCI using fNML as

$$\begin{aligned}\mathrm{SCCI}_f(X;Y \mid Z) = n\hat{I}(X;Y \mid Z) \\ + \max\{\mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z,Y), \mathcal{R}(Y \mid Z) - \mathcal{R}(Y \mid Z,X)\} \ .\end{aligned}$$

An alternative way to obtain a symmetric test using fNML would be to base the test on an equivalent formulation of (algorithmic) CMI, that is

$$\begin{aligned}I_A(x,y \mid z) &= K(x \mid z^*) - K(x \mid (z,y)^*) \\ &\stackrel{+}{=} K(x \mid z^*) + K(y \mid z^*) - K((x,y) \mid z^*) \ . \end{aligned} \tag{4}$$

If we approximate this alternative formulation using fNML, we get

$$\mathrm{SCCI}_{fs}(X;Y \mid Z) = S_f(X \mid Z) + S_f(Y \mid Z) - S_f(X,Y \mid Z) \ .$$

By writing down the regret terms, we see that SCCI_{fs} is symmetric. In particular, if we only consider the regret terms, we get

$$\sum_{z \in Z} \left(\mathcal{C}^{c_z}_{|\mathcal{X}|} + \mathcal{C}^{c_z}_{|\mathcal{Y}|} - \mathcal{C}^{c_z}_{|\mathcal{X}||\mathcal{Y}|} \right) \ .$$

All regret terms depend on the factorization given Z. For the previous formulation, however, we compare the factorizations of X given only Z to the one given Z and Y, or respectively the factorization of Y given only Z to the one given Z and X. Thus, for SCCI_f all regret terms correspond to the same domain, either to the domain of X or Y, whereas for SCCI_{fs} the regret terms are based on X, Y and the Cartesian product of them. Due to the latter, SCCI_{fs} is more conservative than SCCI_f, as we will show in our experiments. Apart from the fact that SCCI_f is more robust in the high-dimensional setup, both variants have a similar performance, which is why we mainly consider SCCI_f in the experiments.

4.2 Quotient SCCI

To formulate SCCI using quotient normalized maximum likelihood, we can straightforwardly replace S with S_q in the independence criterium—i.e.

$$\mathrm{SCCI}_q(X;Y \mid Z) = S_q(X \mid Z) - S_q(X \mid Z,Y) \ .$$

By writing down the regret terms for $\mathrm{SCCI}_q(X;Y \mid Z)$ and $\mathrm{SCCI}_q(Y;X \mid Z)$, we can see that they are equal and hence SCCI_q is symmetric. Another nice property of the qNML formulation is that we would get an equivalent formulation, if we were to base SCCI_q on the alternative formulation of algorithmic (CMI) that we showed in Equation 4. The only shortcoming of this formulation is that similar to SCCI_{fs}, SCCI_q is more restrictive than SCCI_f and thus does not perform as well on high-dimensional data.

Another way to instantiate SCCI, is to use the asymptotic approximation of stochastic complexity (Rissanen, 1996), which was done by Suzuki (2016) to approximate CMI. In practice, the corresponding test (JIC) is, however, very restrictive, which leads to a low recall.

Next, we will show that SCCI is a consistent estimator of CMI and hence in the sample limit able to reliably distinguish (conditional) independencies from dependencies. Thereafter, we compare SCCI to CMI using the threshold based on the gamma distribution (Goebel et al., 2005), and empirically evaluate the sample complexity of SCCI on a limited sample.

4.3 SCCI as a Consistent Estimator of CMI

In this part, our goal is to show that $\frac{1}{n}\text{SCCI}$ approaches the true conditional mutual information, as $n \to \infty$. That is, we need to show that

$$\frac{1}{n}\text{SCCI} = \hat{I} + \frac{t_S}{n}$$

approaches CMI. Since it is well-known that $\hat{I} \to I$ when $n \to \infty$ almost surely, it remains to prove that t_S/n approaches zero, which we will show below.

Theorem 2 *Given three discrete random variables X, Y and Z, we have that $\lim_{n\to\infty} \frac{1}{n}\text{SCCI}(X;Y \mid Z) = I(X;Y \mid Z)$, almost surely.*

Proof: To show the claim, we need to show that

$$\lim_{n\to\infty} \hat{I}(X;Y \mid Z) + \frac{1}{n}(\mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z,Y)) = I(X;Y \mid Z) \ .$$

The proof for any alternative formulation of SCCI follows equivalently. Since $\hat{I} \to I$ when $n \to \infty$ almost surely, we need to show that $\frac{1}{n}(\mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z,Y))$ goes to zero as $n \to \infty$. From Rissanen (1996) we know that $\log \mathcal{C}_k^n$ asymptotically behaves like $\frac{k-1}{2}\log n + \mathcal{O}(1)$, i.e., only grows logarithmically w.r.t. n. Hence, $\frac{1}{n}\mathcal{R}(X \mid Z)$ and $\frac{1}{n}\mathcal{R}(X \mid Z,Y)$ will approach zero if $n \to \infty$. $\square$

5 Link to Related Estimators

Goebel et al. (2005) estimate conditional mutual information through a second-order Taylor series and show that their estimator can be approximated with the gamma distribution. In particular, they state that

$$\hat{I}(X;Y \mid Z) \sim \Gamma\left(\frac{|\mathcal{Z}|}{2}(|\mathcal{X}|-1)(|\mathcal{Y}|-1), \frac{1}{n\ln 2}\right) \ ,$$

where $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ refer to the domains of X, Y and Z. This means by selecting a significance threshold α, we can derive a threshold for CMI based on the gamma distribution—for convenience we call this threshold t_Γ. In the following, we compare t_Γ against $t_S = \mathcal{R}(X \mid Z,Y) - \mathcal{R}(X \mid Z)$.

First of all, for qNML, like t_Γ, t_S depends purely on the sample size and the domain sizes. However, we consider the difference in complexity between only conditioning X on Z and the complexity of conditioning X on Z and Y. For fNML, we have the additional aspect that the regret terms for both $\mathcal{R}(X \mid Z)$ and $\mathcal{R}(X \mid Z,Y)$ also relate to the probability mass function of Z, and respectively the Cartesian product of Z and Y. Recall that for k being the size of the domain of X, we have that

$$\mathcal{R}_f(X \mid Z) = \sum_{z\in Z} \log \mathcal{C}_k^{c_z} \ .$$

As $\mathcal{C}_k^n$ is log-concave in n (Lemma 1), $\mathcal{R}_f(X \mid Z)$ is maximal if Z is uniformly distributed—i.e., it is maximal when $H(Z)$ is maximal. This is a favourable property, as the probability that Z is equal to X is minimal for uniform Z, as stated in the following Lemma.

Lemma 2 (Cover and Thomas (2006)) *If X and Y are i.i.d. with entropy $H(Y)$, then $P(Y = X) \geq 2^{-H(Y)}$ with equality if and only if Y has a uniform distribution.*

To elaborate on the link between t_Γ and t_S, we compare them empirically. In addition, we compare the results to the threshold provided from the JIC test. First, we compare t_Γ with $\alpha = 0.05$ and $\alpha = 0.001$ to t_S/n for fNML, qNML, and JIC on fixed domain sizes, with $|\mathcal{X}|$=$|\mathcal{Y}|$=$|\mathcal{Z}|$=4 and varying sample sizes (see Figure 2). For fNML we computed the worst case threshold by modelling Z as uniformly distributed. In general, the behaviour for each threshold is similar, whereas qNML, fNML and JIC are more restrictive than t_Γ.

Next, we keep the sample size fixed at 500 and increase the domain size of Z from 2 to 200, to simulate a multivariate random vector. Except to JIC, which seems to overpenalize in this case, we observe that fNML is most restrictive until we reach a plateau when $|\mathcal{Z}| = 125$. This is due to the fact that $|\mathcal{Z}||\mathcal{Y}| = 500$ and hence each data point is assigned to one value in the Cartesian product. We have that $\mathcal{R}_f(X \mid Z,Y) = |\mathcal{Z}||\mathcal{Y}|\mathcal{C}_k^1$.

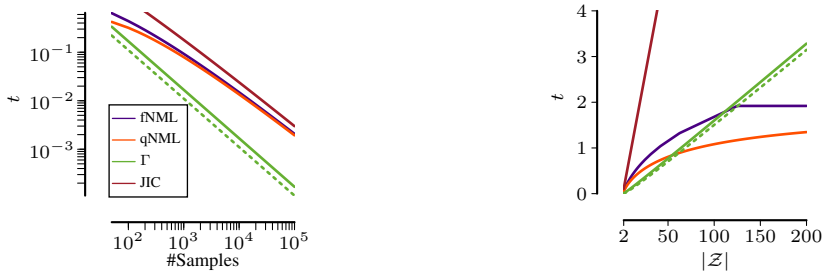


Figure 2: Threshold for CMI using fNML, qNML, JIC and the gamma distribution with $\alpha = 0.05$ (solid) and $\alpha = 0.001$ (dashed) for different sample sizes and fixed domain sizes equal to four (left) and fixed sample size of 500 and changing domain sizes (right).

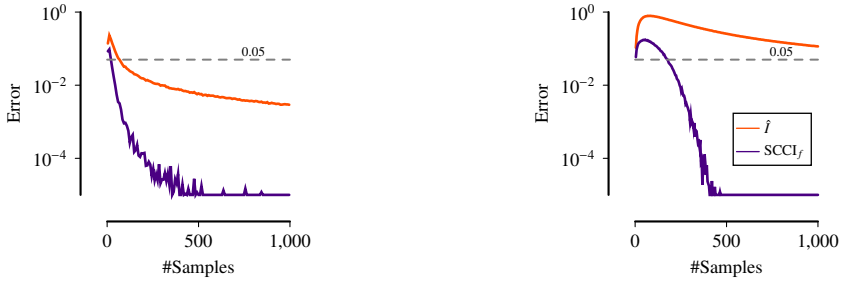


Figure 3: Error for SCCI_f and $\hat{I}$ compared to I, where $I(X;Y|Z) = 0$. Left: $|\mathcal{X}| = |\mathcal{Y}| = 4$ and $|\mathcal{Z}| = 4$. Right: $|\mathcal{X}| = |\mathcal{Y}| = 4$ and $|\mathcal{Z}| = 16$. Values smaller than 10^{-5} are truncated to 10^{-5}

Note that for the thresholds that we computed for fNML we pretend that Z and Y are divided equally over the joint domain $|\mathcal{Y}||\mathcal{Z}|$. In practice, this requirement may not be fulfilled, and hence the regret term for fNML can be smaller. In addition, it is possible that the number of distinct value combinations for Y and Z that we observe in the sample is smaller than their Cartesian product, which also reduces the regret for the fNML formulation.

5.1 Empirical Sample Complexity

In this section, we empirically evaluate the sample complexity of SCCI_f, where we focus on the type I error, i.e., $H_0 \colon X \perp\!\!\!\perp Y \mid Z$ is true and hence $I(X;Y \mid Z) = 0$. We generate data accordingly and draw samples from the joint distribution, where we set $P(x,y,z) = \frac{1}{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}$ for each value configuration $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. Per sample size we draw 1 000 data sets and report the average absolute error for SCCI_f and the empirical estimator of CMI. We show the results for two cases in Figure 3. We observe that in contrast to the empirical estimator $\hat{I}$, SCCI_f quickly approaches zero, and that the difference between both estimators is especially large if we increase the sample space.

If we take a second look at those plots, we see that SCCI_f only needs 180 samples to reach an average error smaller than 0.05 (in both), while the size of the domain space for the second experiment is $|\mathcal{X}||\mathcal{Y}||\mathcal{Z}| = 256$. Ideally, we would like to compute the number of samples that is needed for a reliable result as a function of the domain sizes of the involved variables. Formally, we would like to know the number of samples n that is required such that

$$P(|\text{SCCI}_f(X;Y \mid Z)/n - I(X;Y \mid Z)| \geq \epsilon) \leq \delta \ .$$

As a theoretical analysis is very challenging, we try to derive an empirical bound for $\epsilon = \delta = 0.05$.

We generate data according to the independence hypothesis like above and conduct empirical evaluations for varying domain sizes of X, Y and Z, where we define w.l.o.g. $|\mathcal{X}| \geq |\mathcal{Y}|$, as the test is symmetric. For each combination of domain sizes, we compute $P(|\text{SCCI}_f(X;Y \mid Z)/n - I(X;Y \mid Z)| \geq \epsilon) = P(\text{SCCI}_f(X;Y \mid Z)/n \geq 0.05) \leq 0.05$ as follows: We start with a small n, e.g. two, generate 1 000 data sets and check if over those data sets $P(\text{SCCI}_f(X;Y \mid Z)/n \geq 0.05) \leq 0.05$

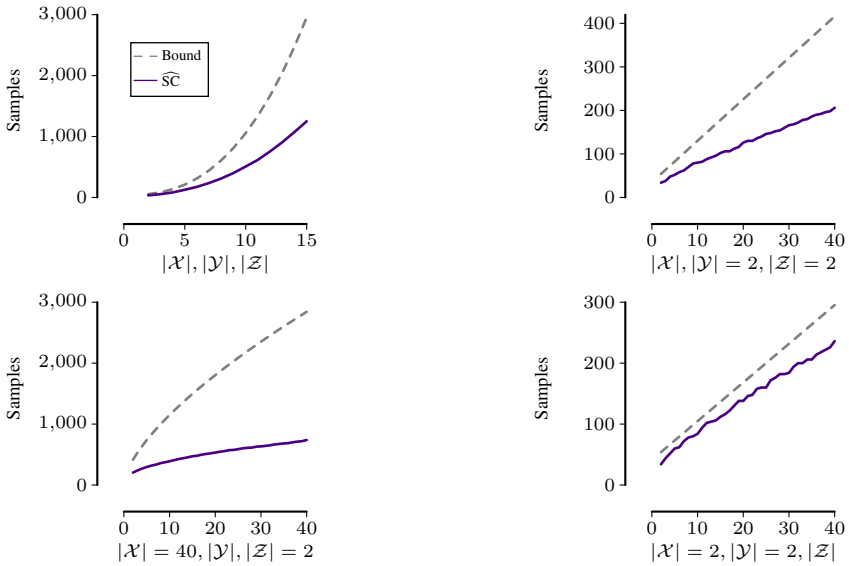


Figure 4: Estimated sample complexities for independently generated data s.t. $P(|\text{SCCI}_f/n - I| \geq 0.05) \leq 0.05$. For all settings, increasing the domain size of X, Y, Z jointly or independently, the suggested bound (calculated as $35 + 2|\mathcal{X}||\mathcal{Y}|^{\frac{2}{3}}(|\mathcal{Z}| + 1)$) is larger than the empirical value.

holds. If not, we increase n by the minimum domain size of X, Y and Z. We repeat this procedure until we reach an n, for which $P(\text{SCCI}_f(X; Y \mid Z)/n \geq 0.05) \leq 0.05$ holds and report this n.

In Figure 4 we plot those values for varying either the domain sizes of X, Y or Z independently or jointly. From these evaluations, we handcrafted a formula to show that it is possible to find an n that is sub-linear w.r.t. the domain sizes of X, Y and Z for which empirically $P(\text{SCCI}_f(X; Y \mid Z)/n \geq 0.05) \leq 0.05$ always holds. Hence, we additionally plot for each domain size the corresponding suggested bound for the sample complexity w.r.t. the formula $35 + 2|\mathcal{X}||\mathcal{Y}|^{\frac{2}{3}}(|\mathcal{Z}| + 1)$. We observe that the empirical values for n are always smaller than the values provided by this formula. Despite this positive result, we want to emphasize that this is only an example function to show the existence of a sub-linear bound for this data. From the plots we would expect that there exists an even tighter bound, however, we did not optimize for that. For future work we would like to theoretically validate a sub-linear bound function.

5.2 Discussion

The main idea for our independence test is to approximate conditional mutual information through algorithmic conditional independence. In particular, we estimate conditional entropy with stochastic complexity. We recommend SCCI_f, since the regret for the entropy term does not only depend on the sample size and the domain sizes of the corresponding random variables, but also on the probability mass function of the conditioning variables. In particular, when fixing the domain sizes and the sample size, higher thresholds are assigned to conditioning variables that are unlikely to be equal to the target variable.

By assuming a uniform distribution for the conditioning variables and hence eliminating this data dependence from SCCI_f, it behaves similar to SCCI_q and CMI where the threshold is derived from the gamma distribution (Goebel et al., 2005). SCCI_f is more restrictive and the penalty terms of all three decrease exponentially w.r.t. the sample size.

SCCI can also be extended for sparsification, as is possible to derive an analytical p-value for the significance of a decision using the no-hypercompression inequality (Grünwald, 2007; Marx and Vreeken, 2017).

Last, note that as we here instantiate SCCI using stochastic complexity for multinomials, we implicitly assume that the data follows a multinomial distribution. In this light, it is important to note

that stochastic complexity is a mini-max optimal refined MDL code (Grünwald, 2007). This means that for any data, we obtain a score that is within a constant term from the best score attainable given our model class. The experiments verify that indeed, SCCI performs very well, even when the data is sampled adversarially.

6 Experiments

In this section, we empirically evaluate SCCI based on fNML and compare it to the alternative formulation using qNML. To not overload the plots, we postpone most comparisons to SCCI_{fs} to Appendix B. In addition, we compare our results to the G^2 test from the *pcalg* R package (Kalisch et al., 2012), CMI_Γ (Goebel et al., 2005) and JIC (Suzuki, 2016).

6.1 Identifying Conditional (In)dependencies

To test whether SCCI can reliably identify (in)dependencies, we generate data according to the graph shown in Figure 1, where we assign the values of F uniformly w.r.t. to its domain space and model a dependency from X to Y by uniformly assigning a mapping form X to Y. We set the domain size for each variable to four and generate data under various samples sizes (100–2 500) and additive uniform noise settings (0%–95%). For each setup we generate 200 data sets and assess the accuracy. We report the correct identifications of $F \perp\!\!\!\perp T \mid \{D, E\}$ as the true positive rate and the false identifications $D \perp\!\!\!\perp T \mid \{E, F\}$ or $E \perp\!\!\!\perp T \mid \{D, F\}$ as false positive rate. For the G^2 test and CMI_Γ we select $\alpha = 0.05$, however, we found no significant differences for $\alpha = 0.01$. Note that for 0% noise all functions are deterministic, which leads to a faithfulness violation and thus $D \not\perp\!\!\!\perp T \mid \{E, F\}$ and $E \not\perp\!\!\!\perp T \mid \{D, F\}$ does not hold. Consequently, an accuracy of 50% is the best we can hope for in this setting.

We show the accuracy of the best performing competitors in Figure 5 and report the remaining results as well as the true and false positive rates for each approach in Appendix B. Overall, we observe that SCCI_f performs near perfect for less than 70% noise, while for $\geq 70\%$ additive noise, the type II error increases. Those results are even better than expected as from our empirical bound function we would suggest that at least 378 samples are required to have reliable results for this data set. SCCI_q has a similar but slightly worse performance. In contrast, CMI_Γ only performs well for less than 30% noise and fails to identify true independencies after more than 30% noise has been added, which leads to a high type I error. The G^2 test has problems with sample sizes up to 500 and performs inconsistently for more than 35% noise.

6.2 Changing the Domain Size

Using the same data generator as above, we now consider a different setup. We fix the sample size to 2 000 and use only 10% additive noise—a setup in which all tests performed well. What we change is the domain size of the source F from 2 to 20 while also restricting the domain sizes of the remaining variable to the same size. For each setup we generate 200 data sets.

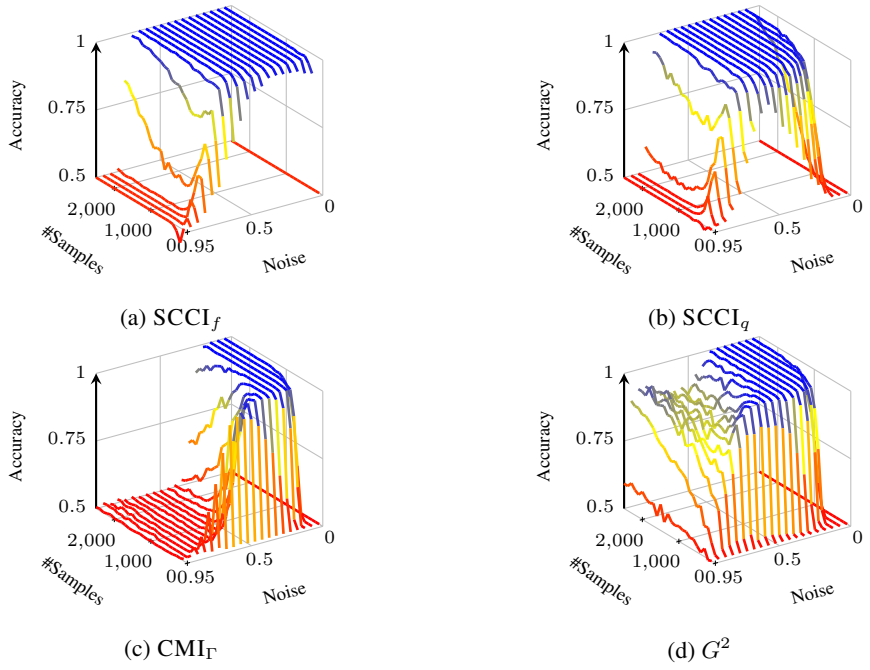
From the results in Figure 6 we see that only SCCI_f performs well for larger domain sizes, whereas all other test have a false positive rate of 100% for $|\mathcal{F}| > 10$, resulting in an accuracy of 50%.

6.3 Identifying Multiple Parents

In this experiment, we test the type II error. This we do by generating a certain number of parents Pa_T from which we generate a target node T. To generate the parents, we use either a

- uniform distribution with domain size $d \sim \text{Unif}(2, 5)$,
- geometric distribution with parameter $p \sim \text{Unif}(0.6, 0.8)$,
- hyper-geometric distribution with parameter $K \sim \text{Unif}(4, 6)$, or
- Poisson distribution with parameter $\lambda \sim \text{Unif}(1, 2)$.

Given Pa_T, we generate T as a mapping from the Cartesian product of the parents to T plus 10% additive uniform noise. Then we generate for each distribution 200 data sets with 2 000 samples, per



(a) SCCI_f (b) SCCI_q

(c) CMI_Γ (d) G^2

Figure 5: [Higher is better] Accuracy of SCCI_f, SCCI_q, CMI_Γ and G^2 for identifying d-separation using varying samples sizes and additive noise percentages, where a noise level of 0.95 refers to 95% additive noise.

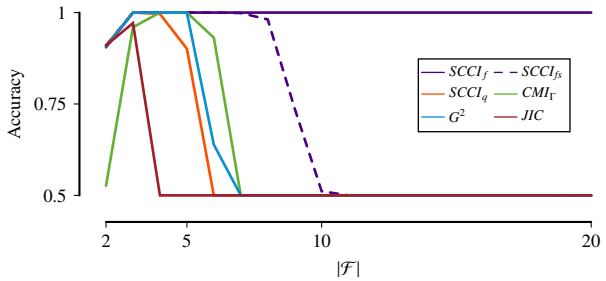


Figure 6: D-separation example with $2\,000$ samples and 10% noise. We gradually increase the domain size of the source node F, which propagates through the graph.

number of parents $k \in \{2, \ldots, 7\}$. We apply SCCI_f, SCCI_q, CMI_Γ and G^2 on each data set and check $\forall P \in \mathrm{Pa}_T$ if the corresponding test correctly identifies that $P \not\perp\!\!\!\perp T \mid \mathrm{Pa}_T \backslash \{P\}$.

We plot the averaged results for each k in Figure 7. It can clearly be observed that SCCI_f performs best and still has near to 100% accuracy for seven parents. Although not plotted here, we can add that the competitors struggled most with the data drawn from the Poisson distribution. We assume that this is due to the fact that the domain sizes for these data sets were on average larger than for the remaining distributions.

In the next experiment, we generate parents and target in the same way, whereas we now fix the number of parents to three. In addition, we generate $k \in \{1, \ldots, 7\}$ random variables N that are drawn jointly independent from T and Pa_T and are uniformly distributed. Then we assess whether the tests under consideration can still identify for each $P \in \mathrm{Pa}_T$ that $P \not\perp\!\!\!\perp T \mid N \cup \mathrm{Pa}_T \backslash \{P\}$.

The averaged results for G^2, JIC, SCCI_f, SCCI_q and CMI_Γ are plotted in Figure 7. Notice that the results for G^2 are barely visible, as they are close to zero for each setup. In general, the trend that we observe is similar to the previous experiment, except that the differences between SCCI_f and its competitors are even larger.

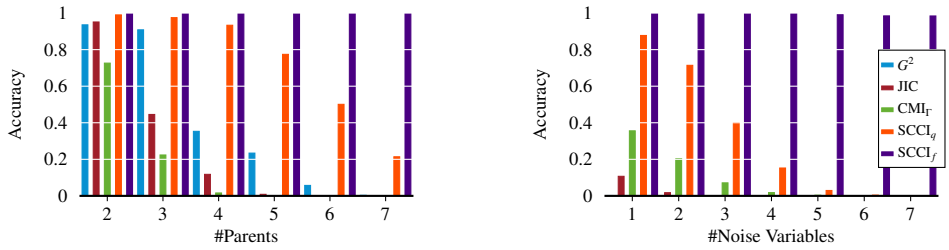


Figure 7: Left: Percentage of parents identified, where we start with only two parents and increase the number of parents to seven. Right: Percentage of parents identified, where we always use three parents, add independently drawn noise variables to the conditioning set.

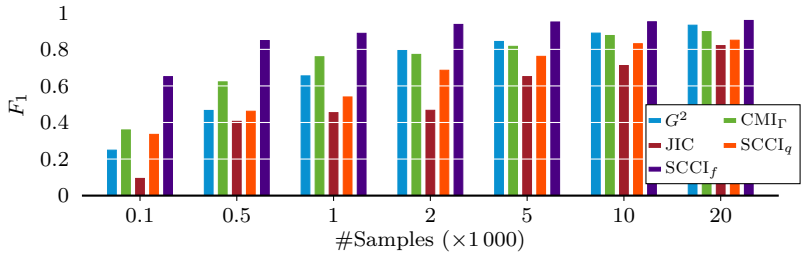


Figure 8: [Higher is better] F_1 score on undirected edges for stable PC using SCCI_f, SCCI_q, JIC, CMI_Γ and G^2 on the *Alarm* network for different sample sizes

6.4 Causal Discovery with SCCI

Last, we evaluate how SCCI performs in practice. Thus, we run the stable PC algorithm (Kalisch et al., 2012; Colombo and Maathuis, 2014) on the *Alarm* network (Scutari and Denis, 2014) from which we generate data with different sample sizes and average over the results of ten runs for each sample size. We equip the stable PC algorithm with SCCI_f, SCCI_q, JIC, CMI_Γ and the default, the G^2 test, and plot the average F_1 score over the undirected graphs in Figure 8. We observe that our proposed test, SCCI_f outperforms its competitors by a large margin, especially for $n \leq 1\,000$.

As a second practical test, we compute the Markov blanket for each node in the *Alarm* network and report the precision and recall. To find the Markov blankets, we run the PCMB algorithm (Peña et al., 2007) with the four independence tests. We plot the precision and recall for each variant in Figure 9. We observe that again SCCI_f performs best—especially regarding recall. As for Markov blankets of size m it is necessary to condition on at least $m - 1$ variables, the advantage in recall can be linked back to SCCI_f being able to correctly detect dependencies for larger domain sizes.

7 Conclusion

In this paper we introduced SCCI, a new conditional independence test for discrete data. We derive SCCI from algorithmic conditional independence and show that it is an unbiased asymptotic estimator for conditional mutual information (CMI). Further, we show how to use SCCI to find a threshold for CMI and compare it to thresholds drawn from the gamma distribution.

In particular, we propose to instantiate SCCI using fNML as in contrast to using qNML or thresholds drawn from the gamma distribution, fNML does not only make use of the sample size and domain sizes of the involved variables, but also utilizes the empirical probability mass function of the conditioning variable. Moreover, we observe that SCCI_f clearly outperforms its competitors on both synthetic and real world data. Last but not least, our empirical evaluations suggest that SCCI has a sub-linear sample complexity, which we would like to theoretically validate in future work.

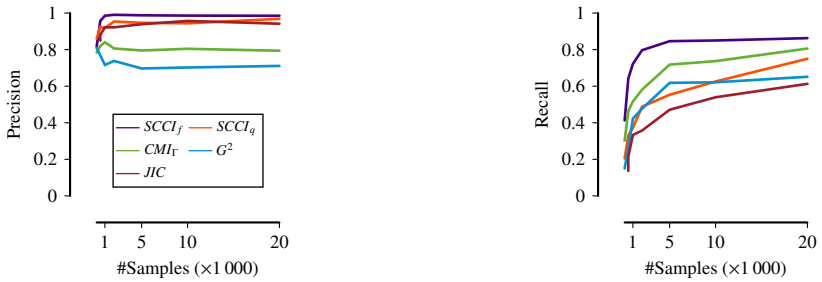


Figure 9: [Higher is better] Precision (left) and recall (right) for PCMB using SCCI_f, SCCI_q, JIC, CMI_Γ and G^2 to identify all Markov blankets in the *Alarm* network for different sample sizes.

Acknowledgements

The authors would like to thank David Kaltenpoth for insightful discussions. Alexander Marx is supported by the International Max Planck Research School for Computer Science (IMPRS-CS). Both authors are supported by the Cluster of Excellence on "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government.

References

Aliferis, C., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. (2010). Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation. *Journal of Machine Learning Research*, 11:171–234.

Canonne, C. L., Diakonikolas, I., Kane, D. M., and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *Proceedings of the Annual ACM Symposium on Theory of Computing*, pages 735–748. ACM.

Chaitin, G. J. (1975). A theory of program size formally identical to information theory. *Journal of the ACM*, 22(3):329–340.

Colombo, D. and Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley.

Goebel, B., Dawy, Z., Hagenauer, J., and Mueller, J. C. (2005). An approximation to the distribution of finite sample size mutual information estimates. In *IEEE International Conference on Communications*, volume 2, pages 1102–1106. IEEE.

Grünwald, P. (2007). *The Minimum Description Length Principle*. MIT Press.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Kolmogorov, A. (1965). Three approaches to the quantitative definition of information. *Problemy Peredachi Informatsii*, 1(1):3–11.

Kontkanen, P. and Myllymäki, P. (2007). MDL histogram density estimation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS), San Juan, Puerto Rico*, pages 219–226. JMLR.

Li, M. and Vitányi, P. (1993). *An Introduction to Kolmogorov Complexity and its Applications*. Springer.

Margaritis, D. and Thrun, S. (2000). Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems*, pages 505–511.

Marx, A. and Vreeken, J. (2017). Telling Cause from Effect using MDL-based Local and Global Regression. In *Proceedings of the 17th IEEE International Conference on Data Mining (ICDM), New Orleans, LA*, pages 307–316. IEEE.

Marx, A. and Vreeken, J. (2018). Causal discovery by telling apart parents and children. *arXiv preprint arXiv:1808.06356*.

Mononen, T. and Myllymäki, P. (2008). Computing the multinomial stochastic complexity in sub-linear time. In *Proceedings of the 4th European Workshop on Probabilistic Graphical Models*, pages 209–216.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Peña, J. M., Nilsson, R., Björkegren, J., and Tegnér, J. (2007). Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning*, 45(2):211–232.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Technology*, 42(1):40–47.

Runge, J. (2018). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 938–947. PMLR.

Schlüter, F. (2014). A survey on independence-based markov networks learning. *Artificial Intelligence Review*, pages 1–25.

Scutari, M. and Denis, J.-B. (2014). *Bayesian Networks with Examples in R*. Chapman and Hall, Boca Raton. ISBN 978-1-4822-2558-7, 978-1-4822-2560-0.

Shtarkov, Y. M. (1987). Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17.

Silander, T., Leppä-aho, J., Jääsaari, E., and Roos, T. (2018). Quotient normalized maximum likelihood criterion for learning bayesian network structures. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84, pages 948–957. PMLR.

Silander, T., Roos, T., Kontkanen, P., and Myllymäki, P. (2008). Factorized Normalized Maximum Likelihood Criterion for Learning Bayesian Network Structures. *Proceedings of the 4th European Workshop on Probabilistic Graphical Models*, pages 257–264.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

Suzuki, J. (2016). An estimator of mutual information and its application to independence testing. *Entropy*, 18(4):109.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813. AUAI Press.

Zhang, Y., Zhang, Z., Liu, K., and Qian, G. (2010). An improved IAMB algorithm for Markov blanket discovery. *Journal of Computers*, 5(11):1755–1761.

A Extended Theory

Lemma 1 *For $n \geq 1$, the regret $\mathcal{C}_k^n$ of the multinomial stochastic complexity of a random variable with a domain size of $k \geq 2$ is log-concave in n.*

Proof: To improve the readability of this proof, we write $\mathcal{C}_L^n$ as shorthand for $\mathcal{C}_{\mathcal{M}_L}^n$ of a random variable with a domain size of L. Since n is an integer, each $\mathcal{C}_L^n > 0$ and $\mathcal{C}_L^0 = 1$, we can prove Lemma 1, by showing that the fraction $\mathcal{C}_L^n/\mathcal{C}_L^{n-1}$ is decreasing for $n \geq 1$, when n increases. We know from Mononen and Myllymäki (2008) that $\mathcal{C}_L^n$ can be written as the sum

$$\mathcal{C}_L^n = \sum_{k=0}^{n} m(k,n) = \sum_{k=0}^{n} \frac{n^{\underline{k}}(L-1)^{\bar{k}}}{n^k k!} \ ,$$

where $x^{\underline{k}}$ represent falling factorials and $x^{\bar{k}}$ rising factorials. Further, they show that for fixed n we can write $m(k,n)$ as

$$m(k,n) = m(k-1,n)\frac{(n-k+1)(k+L-2)}{nk} \ , \tag{5}$$

where $m(0,n)$ is equal to 1. It is easy to see that from $n = 1$ to $n = 2$ the fraction $\mathcal{C}_L^n/\mathcal{C}_L^{n-1}$ decreases, as $\mathcal{C}_L^0 = 1, \mathcal{C}_L^1 = L$ and $\mathcal{C}_L^2 = L + L(L-1)/2$. In the following, we will show the general case. We rewrite the fraction as follows.

$$\begin{aligned}\frac{\mathcal{C}_L^n}{\mathcal{C}_L^{n-1}} &= \frac{\sum_{k=0}^{n} m(k,n)}{\sum_{k=0}^{n-1} m(k,n-1)} \\ &= \frac{\sum_{k=0}^{n-1} m(k,n)}{\sum_{k=0}^{n-1} m(k,n-1)} + \frac{m(n,n)}{\sum_{k=0}^{n-1} m(k,n-1)}\end{aligned} \tag{6}$$

Next, we will show that both parts of the sum in Equation 6 are decreasing when n increases. We start with the left part, which we rewrite to

$$\begin{aligned}\frac{\sum_{k=0}^{n-1} m(k,n)}{\sum_{k=0}^{n-1} m(k,n-1)} &= \frac{\sum_{k=0}^{n-1} m(k,n-1) + \sum_{k=0}^{n-1}\left(m(k,n) - m(k,n-1)\right)}{\sum_{k=0}^{n-1} m(k,n-1)} \\ &= 1 + \frac{\sum_{k=0}^{n-1} \frac{(L-1)^{\bar{k}}}{k!}\left(\frac{n^{\underline{k}}}{n^k} - \frac{(n-1)^{\underline{k}}}{(n-1)^k}\right)}{\sum_{k=0}^{n-1} m(k,n-1)} \ .\end{aligned} \tag{7}$$

When n increases, each term of the sum in the numerator in Equation 7 decreases, while each element of the sum in the denominator increases. Hence, the whole term is decreasing. In the next step, we show that the right term in Equation 6 also decreases when n increases. It holds that

$$\frac{m(n,n)}{\sum_{k=0}^{n-1} m(k,n-1)} \geq \frac{m(n,n)}{m(n-1,n-1)} \ .$$

Using Equation 5 we can reformulate the term as follows.

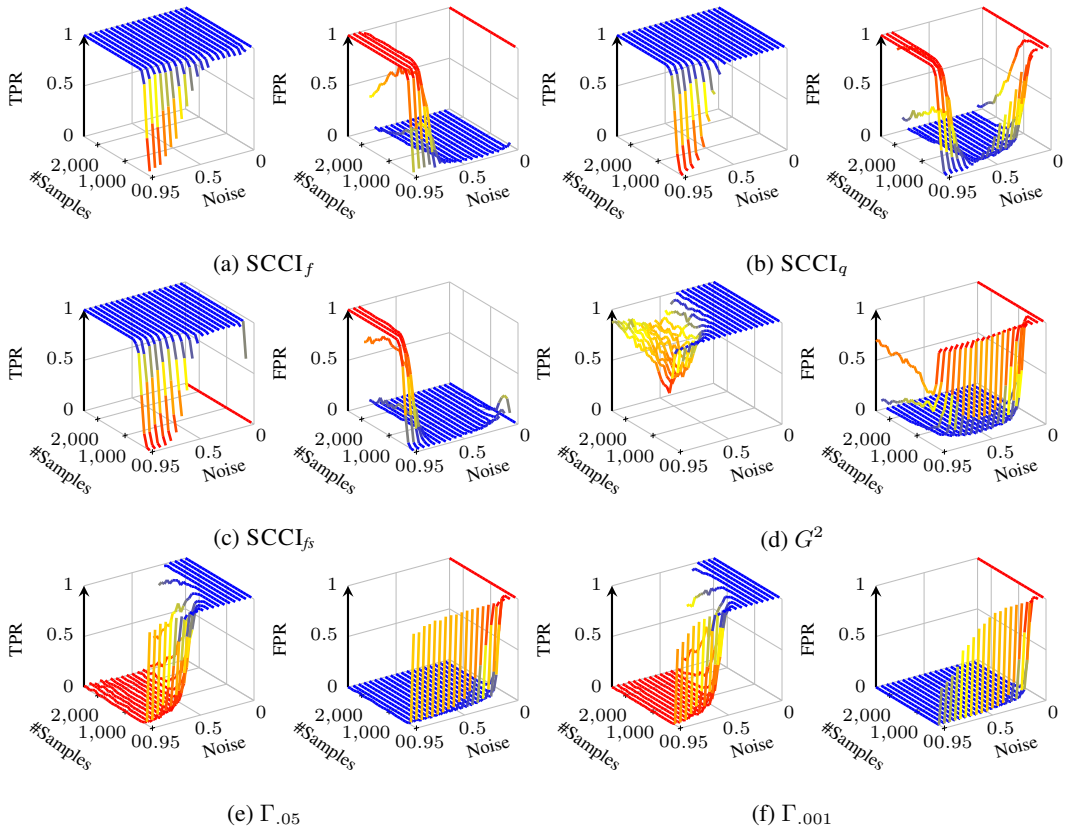
$$\frac{\frac{n+L-2}{n^2} m(n-1,n)}{m(n-1,n-1)} = \frac{n+L-2}{n^2}\left(1 + \frac{m(n-1,n) - m(n-1,n-1)}{m(n-1,n-1)}\right)$$

After rewriting, we have that $\frac{n+L-2}{n^2}$ is definitely decreasing with increasing n. For the right part of the product, we can argue the same way as for Equation 7. Hence the whole term is decreasing, which concludes the proof. □

B Experiments

In this section, we provide more details to the true positive and false positive rates w.r.t. the experiments in Section 6.1, which we show in Figure 10. In addition, we also provide the results for SCCI_{fs} and CMI_Γ with $\alpha = 0.001$. Since we did not provide the accuracy of JIC for this experiment in the main body of the paper, we plot the accuracy, true and false positive rates of JIC in Figure 12 and analyze those results at the end of this section.

From Figure 10, we see that SCCI_f and SCCI_{fs} perform best. Only for very high noise setups ($\geq 70\%$) they start to flag everything as independent. The G^2 test struggles with small sample sizes. It needs more than 500 samples and is inconsistent given more than 35% noise. Note that we forced



(a) SCCI_f (b) SCCI_q

(c) SCCI_{fs} (d) G^2

(e) $\Gamma_{.05}$ (f) $\Gamma_{.001}$

Figure 10: True positive (TPR) and false positive rates (FPR) of SCCI_f, SCCI_q, SCCI_{fs}, G^2 and CMI_Γ with $\alpha = 0.05$ ($\Gamma_{.05}$) and $\alpha = 0.001$ ($\Gamma_{.001}$) for identifying d-separation. We use varying samples sizes (x-axis) and additive noise percentages (y-axis) as in Figure 5, where a noise level of 0.95 refers to 95% additive noise.

G^2 to decide for every sample size, while the minimum number of samples recommended for G^2 on this data set would be $1\,440$, which corresponds to $10(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|$ (Kalisch et al., 2012). Further, we observe that there is barely any difference between CMI_Γ using $\alpha = 0.05$ or $\alpha = 0.001$ as a significance level. After more than 20% noise has been added, CMI_Γ starts to flag everything as dependent.

Next, we also show the accuracy for identifying d-separation for CMI with zero as threshold in Figure 11. Overall, it performs very poorly, which raises from the fact that it barely finds any independence. In addition to the accuracy of CMI, we also plot the average value that CMI reports for the true positive case ($F \perp\!\!\!\perp T \mid \{D, E\}$), where it should be equal to zero. It can be seen that it is dependent on the noise level as well as the sample size. This could explain, why SCCI_f performs best on the d-separation data. Since the noise is uniform, the threshold for SCCI_f is likely to be higher the more noise has been added.

The JIC test has the opposite problem. For the d-separation scenario that we picked it is too restrictive and falsely detects independencies where the ground truth is dependent, as shown in Figure 12. As the discrete version of JIC is calculated from the empirical entropies and a penalizing term based on the asymptotic formulation of stochastic complexity—i.e.,

$$\text{JIC}(X;Y \mid Z) := \max\{\hat{I}(X;Y \mid Z) - \frac{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}{2n} \log n, 0\}\,,$$

it penalizes quite strongly in our example since $|\mathcal{Z}| = 16$. As JIC is based on an asymptotic formulation of stochastic complexity, we expect it to perform better given more data.

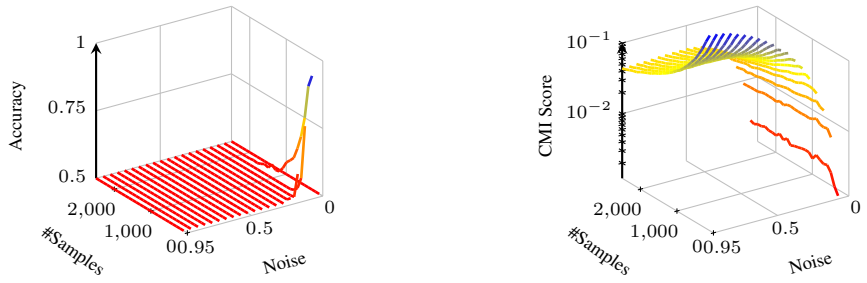


Figure 11: Accuracy of empirical CMI (left) and the average value of empirical CMI for the true independent case (right) for varying samples sizes and additive noise percentages. $\hat{I}(F;T \mid \{D,E\})$ is larger for small sample sizes and high noise settings.

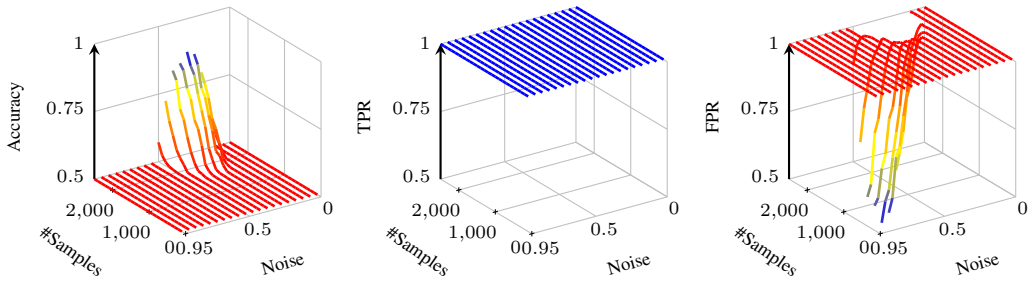


Figure 12: Accuracy, true positive (TPR) and false positive rates (FPR) of JIC for identifying d-separation. We use varying samples sizes and additive noise percentages, where a noise level of 0.95 refers to 95% additive noise.