Testing Conditional Independence on Discrete Data using Stochastic Complexity

Alexander Marx
Max Planck Institute for Informatics,
and Saarland University

Jilles Vreeken
Helmholtz Center for Information Security,
and Max Planck Institute for Informatics

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, *SCI*, is an asymptotically unbiased as well as L_2 consistent estimator for conditional mutual information (*CMI*). Further, we show that *SCI* can be reformulated to find a sensible threshold for *CMI* that works well on limited samples. Empirical evaluation shows that *SCI* has a lower type II error than commonly used tests. As a result, we obtain a higher recall when we use *SCI* 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.

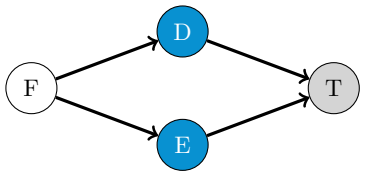


Figure 1: [d-Separation] Given the above causal DAG it holds that $F \perp\!\!\!\perp T \mid D, E$, or F is d-separated of T given D, E under the faithfulness assumption. Note that $D \not\perp\!\!\!\perp T \mid E, F$ and $E \not\perp\!\!\!\perp T \mid D, F$.

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,

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

it is very restrictive as it has a high sample complexity; especially when conditioning on 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 *SCI*, the ***S**tochastic complexity based **C**onditional **I**ndependence criterium*.

Importantly, we show that we can reformulate *SCI* to find a natural threshold for *CMI* that works very well given limited data and diminishes given enough data. In the limit, we prove that *SCI* is an asymptotically unbiased and L_2 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 *SCI* 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 *SCI* shows a sub-linear sample complexity.

For conciseness, we postpone some proofs and experiments to the supplemental material. For reproducibility, we make our code available online.[1]

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 $H_1 \colon X \not\perp\!\!\!\perp Y \mid Z$. The main goal of a good independence test is to minimize the type I and type II error. The type I error is defined as falsely rejecting the null hypothesis and the type II error is defined as falsely accepting the null hypothesis.

A well known theoretical measure for conditional independence is conditional mutual information based on Shannon entropy (Cover and Thomas, 2006).

Definition 1 *Given random variables X, Y and Z. If*

$$I(X;Y \mid Z) := H(X \mid Z) - H(X \mid Z,Y) = 0$$

then X and Y are called statistically independent given Z.

In theory, conditional mutual information (*CMI*) works perfectly as an independence test for discrete data. However, this only holds if we are given the true distributions of the random variables. In practice, those are not given. On a limited sample the plug-in estimator tends to underestimates conditional entropies, and as a consequence, the conditional mutual information is overestimated—even for completely independent data, as in the following Example.

Example 1 *Given three random variables X_1, X_2 and Y, with resp. domain sizes $1\,000, 8$ and 4. Suppose that we are given $1\,000$ samples over 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 $1\,000$ samples, it is likely that we will have only a single sample for each $v \in \mathcal{X}_1$. That is, 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 $n \gg |\mathcal{X}_2|$—i.e. we have more evidence.*

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

A possible solution is to set a threshold t such that $X \perp\!\!\!\perp Y \mid Z$ if $I(X;Y \mid Z) \leq t$. 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 need to give a brief introduction to Kolmogorov complexity. The Kolmogorov complexity of a finite binary string x is the length of the shortest binary program p^* for a universal Turing machine $\mathcal{U}$ that generates x, and then halts (Kolmogorov, 1965; Li and Vitányi, 1993). Formally, we have

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

That is, program p^* is the most succinct *algorithmic* description of x, or in other words, the ultimate lossless compressor for that string. To define algorithmic independence, we will also need conditional Kolmogorov complexity, $K(x \mid y) \leq K(x)$, which is again the length of the shortest binary program p^* that generates x, and halts, but now given y as input for free.

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 this point, algorithmic independence differs from statistical independence. In contrast to purely considering the dependency between random variables, it also considers the complexity of the process behind the dependency.

Let us consider Example 1 again and let x_1, x_2, and y be the binary strings representing 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 to describe Y from X_2 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. $K(y \mid x_2) \leq K(y \mid x_1)$ in contrast $\hat{H}(Y \mid X_1) = \hat{H}(Y \mid X_2)$. To reject $Y \perp\!\!\!\perp X \mid Z$, we test whether providing the information of X leads to a shorter program than only knowing Z. Formally, we define algorithmic conditional independence as follows (Chaitin, 1975).

Definition 2 *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. If*

$$I_A(x; y \mid z) := K(x \mid z^*) - K(x \mid (z,y)^*) \stackrel{+}{=} 0$$

holds up to an additive constant that is independent of the data, then x and y are called algorithmically independent given z.

Due to the halting problem Kolmogorov complexity is not computable, however, 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 it 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

Given n samples of a discrete univariate random variable X with a domain $\mathcal{X}$ of $|\mathcal{X}| = k$ distinct values, $x^n \in \mathcal{X}^n$, let $\hat{\theta}(x^n)$ denote the maximum likelihood estimator for x^n. Shtarkov (1987) defined the mini-max optimal *normalized maximum likelihood (NML)*

$$P_{NML}(x^n \mid \mathcal{M}_k) = \frac{P(x^n \mid \hat{\theta}(x^n), \mathcal{M}_k)}{\mathcal{C}^n_{\mathcal{M}_k}} \; , \qquad (1)$$

where the normalizing factor, or regret $\mathcal{C}^n_{\mathcal{M}_k}$, relative to the model class $\mathcal{M}_k$ is defined as

$$\mathcal{C}^n_{\mathcal{M}_k} = \sum_{\tilde{x}^n \in \mathcal{X}^n} P(\tilde{x}^n \mid \hat{\theta}(\tilde{x}^n), \mathcal{M}_k) \, . \qquad (2)$$

The sum goes over every possible $\tilde{x}^n$ over the domain of X, and for each considers the maximum likelihood for that data given model class $\mathcal{M}_k$. Whenever clear from context, we will drop the model class to simplify the notation—i.e. we write $P_{NML}(x^n)$ for $P_{NML}(x^n \mid \mathcal{M}_k)$ and $\mathcal{C}^n_k$ to refer to $\mathcal{C}^n_{\mathcal{M}_k}$.

For discrete data, assuming a multinomial distribution, we can rewrite Eq. (1) as (Kontkanen and Myllymäki, 2007)

$$P_{NML}(x^n) = \frac{\prod_{j=1}^{k} \left(\frac{|v_j|}{n}\right)^{|v_j|}}{\mathcal{C}^n_k} \; ,$$

writing $|v_j|$ for the frequency of value v_j in x^n, resp. Eq. (2) as

$$\mathcal{C}^n_k = \sum_{|v_1|+\cdots+|v_k|=n} \frac{n!}{|v_1|! \cdots |v_k|!} \prod_{j=1}^{k} \left(\frac{|v_j|}{n}\right)^{|v_j|} \; .$$

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.

We obtain the stochastic complexity for x^n by simply taking the negative logarithm of P_{NML}, which decomposes into a Shannon-entropy and the log regret

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

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 x^n and y^n drawn from the joint distribution of two random variables X and Y, where k is the size of the domain of X. Conditional stochastic complexity using the fNML formulation is defined as

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

where $y^n = v$ denotes the set of samples for which $Y = v$, $\mathcal{Y}$ the domain of Y with domain size l, and $|v|$ the frequency of a value v in y^n.

Analogously, we can define conditional stochastic complexity S_q using qNML (Silander et al., 2018). We prove all important properties of our independence test for both fNML and qNML definitions, but for conciseness, and because S_f performs superior in our experiments, we postpone the definition of S_q and the related proofs to the supplemental material.

In the following, we always consider the sample size n and slightly abuse the notation by replacing $S(x^n)$ by $S(X)$, similar so for the conditional case. We refer to conditional stochastic complexity as S and only use S_f or S_q whenever there is a conceptual difference. In addition, we refer to the regret terms of the conditional $S(X \mid Z)$ as $\mathcal{R}(X \mid Z)$, where

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

Next, we show that the multinomial regret term is log-concave in n, which is a property we need later on.

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

For conciseness, we postpone the proof of Lemma 1 to the supplementary material. Based on Lemma 1 we can now introduce our main theorem that is necessary for our proposed independence test.

Theorem 1 *Given three random variables X, Y and Z, it holds that $\mathcal{R}_f(X \mid Z) \leq \mathcal{R}_f(X \mid Z, Y)$.*

Proof: 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 that Theorem 1 is true, it suffices to show that

$$\sum_{i=1}^{p} \log \mathcal{C}_k^{|r_i|} \leq \sum_{j=1}^{q} \log \mathcal{C}_k^{|l_j|} \tag{3}$$

where $\sum_{i=1}^{p} |r_i| = \sum_{j=1}^{q} |l_j| = n$. 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, Eq. (3) holds, if the $\log \mathcal{C}_k^n$ is sub-additive in n. Since we know from Lemma 1 that the regret term is log-concave in n, sub-additivity follows by definition. □

Now that we have all the necessary tools, we can define our independence test in the next section.

4 Stochastic Complexity based Conditional Independence

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

Definition 3 *Let X, Y and Z be random variables. We say that $X \perp\!\!\!\perp Y \mid Z$, if*

$$SCI(X; Y \mid Z) := S(X \mid Z) - S(X \mid Z, Y) \leq 0 \ . \tag{4}$$

In particular, Eq. 4 can be rewritten as

$$\begin{aligned} SCI(X; Y \mid Z) = n \cdot I(X; Y \mid Z) \\ + \mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z, Y) \ . \end{aligned}$$

From this formulation, we see that the regret terms formulate a threshold t_S for conditional mutual information, where $t_S = \mathcal{R}(X \mid Z, Y) - \mathcal{R}(X \mid Z)$. From Theorem 1 we know that if we instantiate *SCI* using fNML that $\mathcal{R}(X \mid Z, Y) - \mathcal{R}(X \mid Z) \geq 0$. Hence, Y has to provide a significant gain such that $X \not\perp\!\!\!\perp Y \mid Z$—i.e. we need $\hat{H}(X \mid Z) - \hat{H}(X \mid Z, Y) > t_S/n$.

Next, we show how we can use *SCI* in practice by formulating it using fNML.

4.1 Factorized SCI

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 $\mathcal{R}_f(Y \mid Z) - \mathcal{R}_f(Y \mid Z,X)$. As a consequence,

$$I_S^f(X;Y \mid Z) := S_f(X \mid Z) - S_f(X \mid Z,Y)$$

is not always equal to

$$I_S^f(Y;X \mid Z) := S_f(Y \mid Z) - S_f(Y \mid Z,X) \ .$$

To achieve symmetry, we formulate SCI_f as

$$SCI_f(X;Y \mid Z) := \max\{I_S^f(X;Y \mid Z), I_S^f(Y;X \mid Z)\}$$

and say that $X \perp\!\!\!\perp Y \mid Z$, if $SCI_f(X;Y \mid Z) \leq 0$.

There are other ways to achieve such symmetry, such as via an alternative definition of conditional mutual information. However, as we show in detail in the supplementary, there exist serious issues with these alternatives when instantiated with fNML.

Instead of the exact fNML formulation, it is also possible 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 low recall.

In the next section, we show the main properties for *SCI* using fNML. Thereafter, we compare *SCI* to *CMI* using the threshold based on the gamma distribution (Goebel et al., 2005), and empirically evaluate the sample complexity of *SCI*.

4.2 Properties of SCI

In the following, for readability, we write *SCI* to refer to properties that hold for both SCI_f and SCI_q.

First, we show that if $X \perp\!\!\!\perp Y \mid Z$, we have that $SCI(X;Y \mid Z) \leq 0$. Then, we prove that $\frac{1}{n}SCI$ is an asymptotically unbiased estimator of conditional mutual information and is L_2 consistent. Note that by dividing *SCI* by n we do not change the decisions we make as long as $n < \infty$. Since we only accept H_0 if $SCI \leq 0$, any positive output will still be > 0 after dividing it by n.

Theorem 2 *If $X \perp\!\!\!\perp Y \mid Z$, $SCI(X;Y \mid Z) \leq 0$.*

Proof: W.l.o.g. we can assume that $I_S^f(X;Y \mid Z) >= I_S^f(Y;X \mid Z)$. Based on this, it suffices to show that $I_S^f(X;Y \mid Z) \leq 0$ if $X \perp\!\!\!\perp Y \mid Z$. As the first part of I_S^f consists of $n \cdot I(X;Y \mid Z)$, it will be zero by definition. We know that $\mathcal{R}_f(X \mid Z) - \mathcal{R}_f(X \mid Z,Y) \leq 0$ (Theorem 1), which concludes the proof. □

Next, we show that $\frac{1}{n}SCI$ converges against conditional mutual information and hence is an asymptotically unbiased estimator of conditional mutual information and is L_2 consistent to it.

Lemma 2 *Given three random variables X, Y and Z, it holds that $\lim_{n\to\infty} \frac{1}{n}SCI(X;Y \mid Z) = I(X;Y \mid Z)$.*

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

$$\lim_{n\to\infty} I(X;Y \mid Z) + \frac{1}{n}(\mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z,Y)) = 0 \ .$$

The proof for $I_S^f(Y;X \mid Z)$ follows analogously. In essence, we need to show that $\frac{1}{n}(\mathcal{R}(X \mid Z) - \mathcal{R}(X \mid Z,Y))$ goes to zero as n goes to infinity. From Rissanen (1996) we know that $\log \mathcal{C}_k^n$ asymptotically behaves like $\frac{k-1}{2}\log n + \mathcal{O}(1)$. Hence, $\frac{1}{n}\mathcal{R}(X \mid Z)$ and $\frac{1}{n}\mathcal{R}(X \mid Z,Y)$ will approach zero if $n \to \infty$. □

As a corollary to Lemma 2 we find that $\frac{1}{n}SCI$ is an asymptotically unbiased estimator of conditional mutual information and is L_2 consistent to it.

Theorem 3 *Let X, Y and Z be discrete random variables. Then $\lim_{n\to\infty} \mathbb{E}[\frac{1}{n}SCI(X;Y|Z)] = I(X;Y|Z)$, i.e. $\frac{1}{n}SCI$ is an asymptotically unbiased estimator for conditional mutual information.*

Theorem 4 *Let X, Y and Z be discrete random variables. Then $\lim_{n\to\infty} \mathbb{E}[(\frac{1}{n}SCI(X;Y|Z) - I(X;Y|Z))^2] = 0$ i.e. $\frac{1}{n}SCI$ is an L_2 consistent estimator for conditional mutual information.*

Next, we compare both of our tests to the findings of Goebel et al. (2005).

4.3 Link to Gamma Distribution

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

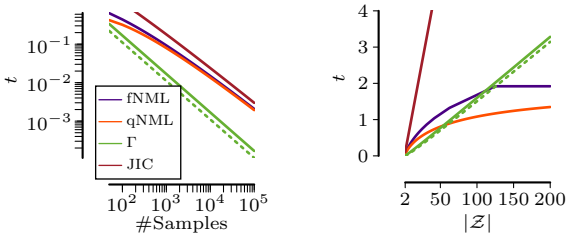


Figure 2: Threshold for *CMI* using fNML, qNML, *JIC* and the gamma distribution with $\alpha = 0.05$ (solid) and $\alpha = 0.001$ (dashed) for different sample sizes and fixed domain sizes equal to four (left) and fixed sample size of 500 and changing domain sizes (right).

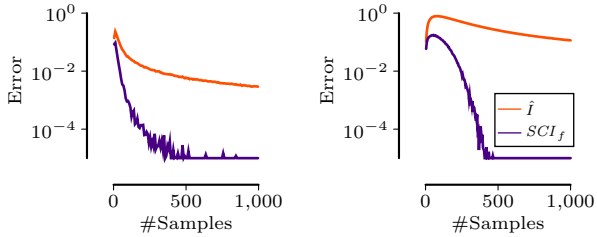


Figure 3: Error for SCI_f and $\hat{I}$ compared to I, where $I(X;Y|Z)=0$. Left: $|\mathcal{X}|=|\mathcal{Y}|=2$ and $|\mathcal{Z}|=4$. Right: $|\mathcal{X}|=|\mathcal{Y}|=4$ and $|\mathcal{Z}|=16$. Values smaller than 10^{-5} are truncated to 10^{-5}.

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 functions 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

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

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 (Cover and Thomas, 2006).

Lemma 3 *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 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 and qNML, and *JIC* on fixed domain sizes, with $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}| = 4$ and varying the sample sizes (see Figure 2). For fNML we computed the worst case threshold under the assumption that Z is 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 fix at 500 and increase the domain sizes of Z from 2 to 200, to simulate multiple variables in the conditioning set. 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 $\mathcal{R}_f(X \mid Z,Y) = |\mathcal{Z}||\mathcal{Y}|\mathcal{C}_k^1$.

It is important to note, however, that the thresholds that we computed for fNML assume that Z and Y are uniformly distributed and $Y \perp\!\!\!\perp Z$. In practice, when this requirement is not fulfilled, the regret term of fNML can be smaller than this value, since it is data dependent. In addition, it is possible that the number of distinct values that we observe from the joint distribution of Z and Y is smaller than their Cartesian product, which also reduces the difference in the regret terms for fNML.

4.4 Empirical Sample Complexity

In this section, we empirically evaluate the sample complexity of SCI_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 SCI_f and the empirical estimator of CMI, $\hat{I}$. We show the results for two cases in Fig. 3. We observe that in contrast to the empirical plug-in estimator $\hat{I}$, SCI_f quickly approaches zero, and that the difference is especially large for larger domain sizes.

In the supplemental material we give a more in depth analysis alltogether. Our evaluation suggest that the sample complexity is sub-linear. In particular, we find that the number of samples n required such that $P(|SCI_f^n(X;Y \mid Z)/n - I(X;Y \mid Z)| \geq \epsilon) \leq \delta$, with $\epsilon = \delta = 0.05$ is smaller than $35 + 2|\mathcal{X}||\mathcal{Y}|^{2/3}(|\mathcal{Z}| + 1)$.

To illustrate this, consider the left example in Figure 3 again. We observe that for $\epsilon = \delta = 0.05$, n needs to be at least 52, which is smaller than the value from our empirical bound function, that is equal to 67. If we require $\epsilon = 0.01$ and $\delta = 0.05$, we observe that n must

be at least 72. In comparison, for $\hat{I}$, n needs to be at least 140 for $\epsilon = 0.05$ and 684 for $\epsilon = 0.01$.

4.5 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 SCI_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 SCI_f, it behaves similar to SCI_q and CMI where the threshold is derived from the gamma distribution (Goebel et al., 2005). SCI_f is more restrictive and the penalty terms of all three decrease exponentially w.r.t. the sample size.

SCI 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).

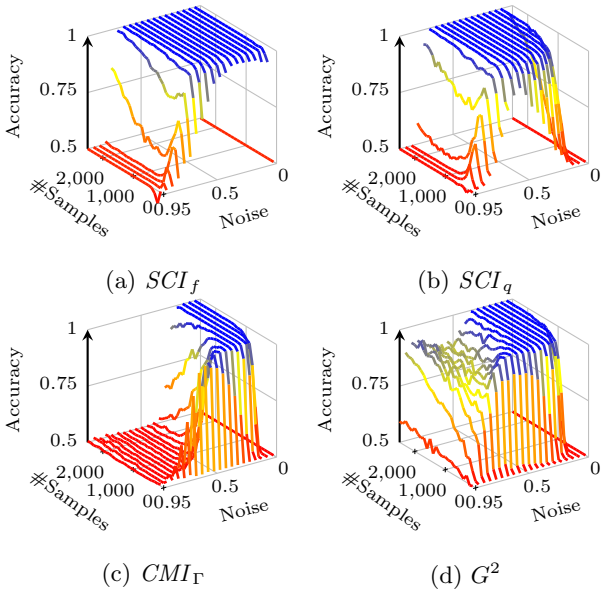
Last, note that as we here instantiate SCI using stochastic complexity for multinomials, we implicitly assume that the data follows a multinomial distribution. In this light, it is important to note that stochastic complexity is a mini-max optimal refined MDL code (Grünwald, 2007). This means that for 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, SCI performs very well, even when the data is sampled adversarially.

5 Experiments

In this section, we empirically evaluate SCI based on fNML and compare it to the alternative formulation using qNML. In addition, we compare it to the G^2 test from the *pcalg* R package (Kalisch et al., 2012), CMI_Γ (Goebel et al., 2005) and JIC (Suzuki, 2016).

5.1 Identifying d-Separation

To test whether SCI can reliably distinguish between independence and dependence, we generate data as depicted in Figure 1, where we draw F from a uniform distribution and model a dependency from X to Y by simply assigning uniformly at random each $x \in \mathcal{X}$ to a $y \in \mathcal{Y}$. We set the domain size for each variable to 4 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. In particular, 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.[2] For the G^2 test and CMI_Γ we select $\alpha = 0.05$, however, we found no significant differences for $\alpha = 0.01$.



(a) SCI_f (b) SCI_q (c) CMI_Γ (d) G^2

Figure 4: [Higher is better] Accuracy of SCI_f, SCI_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.

In the interest of space we only plot the accuracy of the best performing competitors in Figure 4 and report the remaining results as well as the true and false positive rates for each approach in the supplemental material. Overall, we observe that SCI_f performs near perfect for less than 70% additive noise. When adding 70% or more 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. SCI_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 given more than 35% noise.

[2] For 0% noise, F has all information about D and E, hence $D \not\perp\!\!\!\perp T \mid E, F$ and $E \not\perp\!\!\!\perp T \mid D, F$ does not hold.

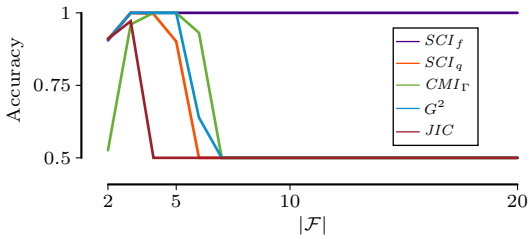


Figure 5: d-Separation with 2 000 samples and 10% noise on different domain sizes of the source node F.

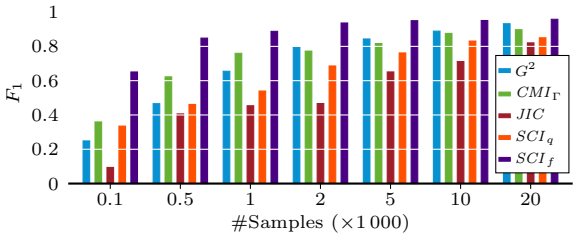


Figure 6: [Higher is better] F_1 score on undirected edges for stable PC using SCI_f, SCI_q, JIC, CMI_Γ and G^2 on the *Alarm* network for different sample sizes.

5.2 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 where 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 5 we can clearly see that only SCI_f is able to deal with larger domain sizes as for all other test, the false positive rate is at 100% for larger domain sizes, resulting in an accuracy of 50%.

5.3 Plug and Play with SCI

Last, we want to show how *SCI* performs in practice. To do this, 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 10 runs for each sample size. We equip the stable PC algorithm with SCI_f, SCI_q, JIC, CMI_Γ and the default, the G^2 test, and plot the average $F1$ score over the undirected graphs in Figure 6. We observe that our proposed test, SCI_f outperforms the other tests for each sample size with a large margin and especially for small sample sizes.

As a second practical test, we compute the Markov blanket for each node in the *Alarm* network and report

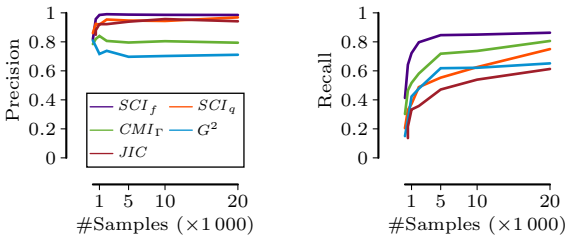


Figure 7: [Higher is better] Precision (left) and recall (right) for PCMB using SCI_f, SCI_q, JIC, CMI_Γ and G^2 to identify all Markov blankets in the *Alarm* network for different sample sizes.

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 7. We observe that again SCI_f performs best—especially with regard to recall. As for Markov blankets of size k it is necessary to condition on at least $k-1$ variables, this advantage in recall can be linked back to SCI_f being able to correctly detect dependencies for larger domain sizes.

6 Conclusion

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

In particular, we propose to instantiate *SCI* 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 SCI_f clearly outperforms its competitors on both synthetic and real world data. Last but not least, our empirical evaluations suggest that *SCI* has a sub-linear sample complexity, which we would like to theoretically validate in future work.

Acknowledgment

The authors would like to thank David Kaltenpoth for insightful discussions. AM is supported by the International Max Planck Research School for Computer Science (IMPRS-CS). Both authors are supported by the DFG Cluster of Excellence MMCI.

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.

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

A.1 Proof of Lemma 1

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 Eq. 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 Eq. 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 Eq. 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 Eq. 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 Eq. 7. Hence the whole term is decreasing, which concludes the proof. □

A.2 Quotient SCI

Conditional stochastic complexity can also be defined via quotient normalized maximum likelihood (qNML), which is defined as follows

$$S_q(x^n \mid y^n) = \sum_{v \in \mathcal{Y}} |v| \hat{H}(x^n \mid y^n = v) + \log \frac{\mathcal{C}_{|\mathcal{X}|\cdot|\mathcal{Y}|}^n}{\mathcal{C}_{|\mathcal{Y}|}^n} \ .$$

We refer to the regret term of $S_q(X \mid Z)$ with

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

Analogously to Theorem 1 for fNML, we can define the following theorem for qNML.

Theorem 5 *Given three random variables X, Y and Z, it holds that $\mathcal{R}_q(X \mid Z) \leq \mathcal{R}_q(X \mid Z, Y)$.*

Proof: Consider n samples of three random variables X, Y and Z, with corresponding domain sizes k, p and q. It should hold 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 proof the statement above. □

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

$$SCI_q(X; Y \mid Z) := S_q(X \mid Z) - S_q(X \mid Z, Y)$$

and say that $X \perp\!\!\!\perp Y \mid Z$, if $SCI_q(X; Y \mid Z) \leq 0$. By writing down the regret terms for $SCI_q(X; Y \mid Z)$ and $SCI_q(Y; X \mid Z)$, we can see that they are equal and hence SCI_q is symmetric, that is, $SCI_q(X; Y \mid Z) = SCI_q(Y; X \mid Z)$.

Since we showed that Theorem 5 holds for qNML, Theorems 2-4 can also be proven for qNML using the same arguments as for fNML.

A.3 Alternative Symmetry Correction for Factorized SCI

To instantiate *SCI* using fNML, we take the maximum between $I^f_S(X; Y \mid Z)$ and $I^f_S(Y; X \mid Z)$ to achieve symmetry. We could also achieve symmetry when we base our formulation on an alternative formulation of conditional mutual information, that is

$$CMI(X; Y \mid Z) = H(X \mid Z) + H(Y \mid Z) - H(X, Y \mid Z) \ . \tag{8}$$

In particular, we formulate our alternative test by replacing the conditional entropies in Eq. 8 with stochastic complexity based on fNML

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

By writing down the regret terms, we see that $SCI_{fs}(X; Y \mid Z) = SCI_{fs}(Y; X \mid Z)$. In particular, if we only consider the regret terms, we get

$$\sum_{z \in Z} \left(\mathcal{C}^{|z|}_{|\mathcal{X}|} + \mathcal{C}^{|z|}_{|\mathcal{Y}|} - \mathcal{C}^{|z|}_{|\mathcal{X}||\mathcal{Y}|} \right) \ . \tag{9}$$

From Eq. 9 we see that all regret terms depend on the factorization given Z. For $I^f_S(X; Y \mid Z)$, however, we compare the factorizations of X given only Z to the one given Z and Y, and similarly so for $I^f_S(Y; X \mid Z)$. In addition, for SCI_f all regret terms correspond to the same domain, either to the domain of X or Y, whereas for SCI_{fs} the regret terms are based on X, Y and the Cartesian product of them. Since the last regret term of SCI_{fs} is based on the Cartesian product of X and Y it performs worse than SCI_f for large domain sizes. This can also be seen in Figure 8, for which we conducted the same experiment as in Section 5.2, but also applied SCI_{fs}. SCI_q exhibits similar behaviour like SCI_{fs}, as it also considers products of domain sizes.

There also exists a third way to formulate *CMI*, i.e.

$$CMI(X; Y \mid Z) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z) \ . \tag{10}$$

When we replace all entropy terms with the stochastic complexity in Eq. 10, we would get an equivalent formulation to SCI_q, as the regret terms would sum up to exactly the same values. Hence, we do not elaborate further on this alternative.

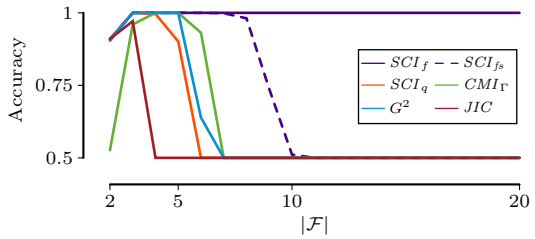


Figure 8: d-Separation with 2 000 samples and 10% noise on different domain sizes of the source node F.

B Experiments

In this section, we provide more details to the true positive and false positive rates w.r.t. d-separation. Further, we show how well SCI and its competitors can recover multiple parents with and without additional noise variables in the conditioning set.

B.1 TPR and FPR for d-Separation

In Section 5.1 we analyzed the accuracy of SCI_f, SCI_q, CMI_Γ and G^2 for identifying d-separation. In Figure 9, we plot the true and false positive rates to the corresponding experiment. In addition, we also provide the results for SCI_{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 10 and analyze those results at the end of this section.

From Figure 9, we see that SCI_f and SCI_{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 and is inconsistent given more than 35% noise. Note that we forced 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 SCI_f performs best on the d-separation data. Since the noise is uniform, the threshold for SCI_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 10. 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.

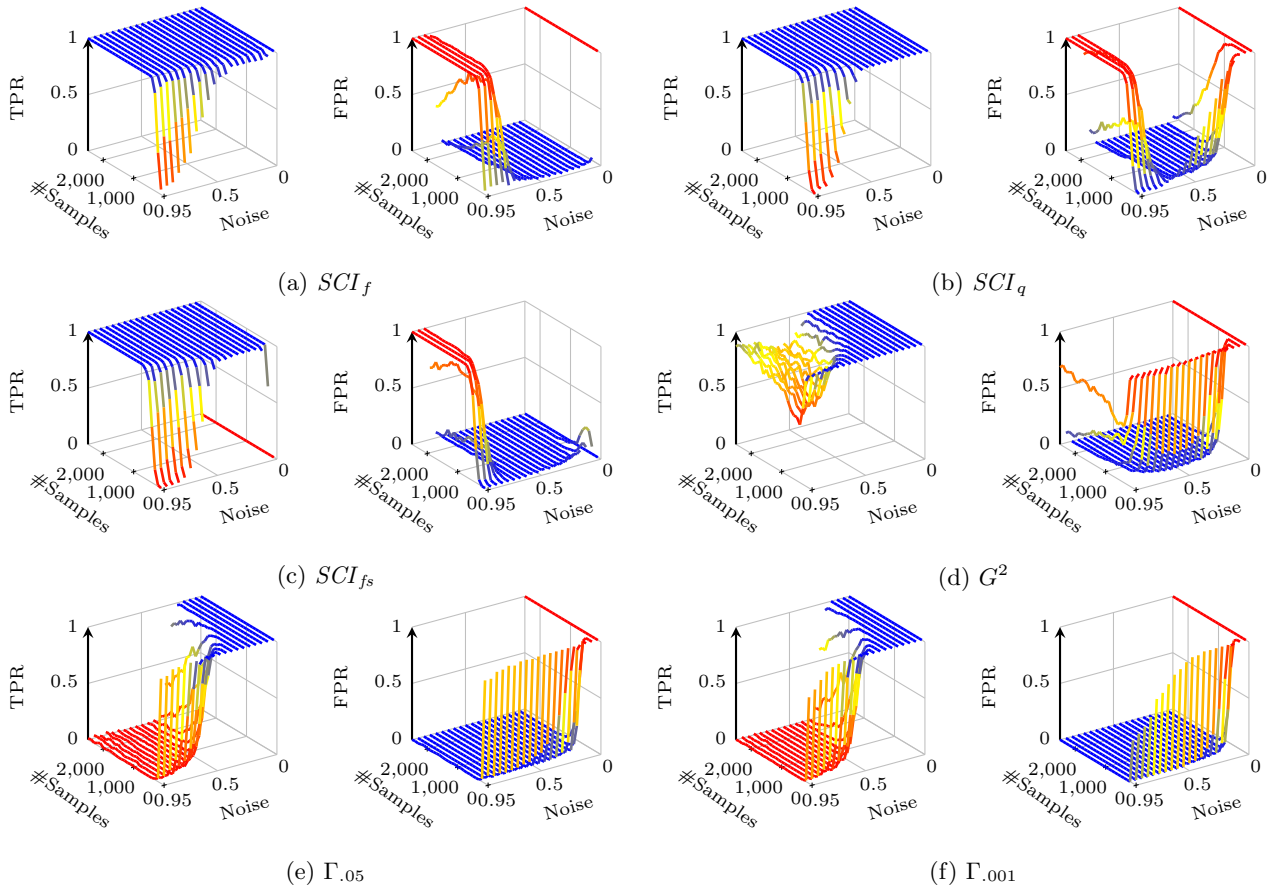
$$JIC(X;Y \mid Z) := \max\{\hat{I}(X;Y \mid Z) - \frac{(|\mathcal{X}|-1)(|\mathcal{Y}|-1)|\mathcal{Z}|}{2n} \log n, 0\}\,,$$

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.

B.2 Identifying the 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 a domain size d drawn uniformly with $d \sim \text{unif}(2, 5)$,
- geometric distribution with parameter $p \sim \text{unif}(0.6, 0.8)$,



(a) SCI_f (b) SCI_q

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

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

Figure 9: True positive (TPR) and false positive rates (FPR) of SCI_f, SCI_q, SCI_{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 and additive noise percentages, where a noise level of 0.95 refers to 95% additive noise.

- hyper-geometric distribution with parameter $K \sim \text{unif}(4, 6)$, or
- poisson distribution with parameter $\lambda \sim \text{unif}(1, 2)$.

Given the parents, 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 number of parents $k \in \{2, \ldots, 7\}$. We apply SCI_f, SCI_q, CMI_Γ and G^2 on each data set and we check for each $p \in PA_T$ whether they output the correct result, that is, $p \not\perp\!\!\!\perp T \mid PA_T \backslash \{p\}$.

We plot the averaged results for each k in Figure 12. It can clearly be observed that SCI_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 all other tested distributions.

In the next experiment, we generate parents and target in the same way as mentioned above, 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 as described above. Then we test whether the conditional independence tests under consideration can still identify for each $p \in PA_T$ that $p \not\perp\!\!\!\perp T \mid N \cup PA_T \backslash \{p\}$.

The averaged results for G^2, JIC, SCI_f, SCI_q and CMI_Γ are plotted in Figure 12. 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 SCI_f and its competitors are even larger.

Figure 10: Accuracy of *CMI* (left) and the average value returned by *CMI* for the true independent case (right) for varying samples sizes and additive noise percentages. $I(F;T \mid D,E)$ is larger for small sample sizes.

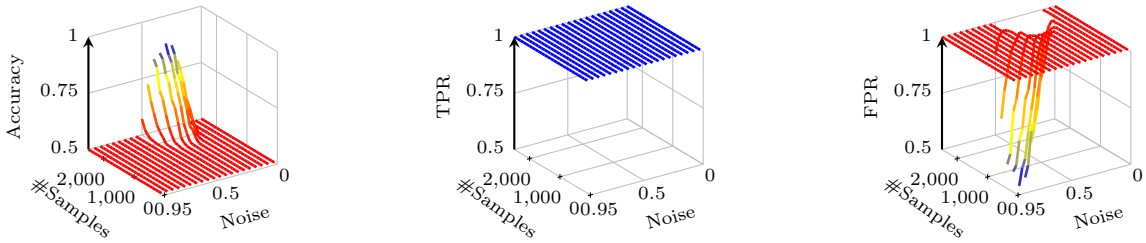


Figure 11: Accuracy, true positive (TPR) and false positive rates (FPR) of *JIC* for identifying d-separation. We use varying samples sizes and additive noise percentages, where a noise level of 0.95 refers to 95% additive noise.

B.3 Empirical Sample Complexity

To give an intuition to the sample complexity of *SCI*, we provide an empirical evaluation. The goal of this section is to show that there exists a bound for the sample complexity of *SCI*, that is sub-linear w.r.t the size Cartesian product of the domain sizes and always larger than the bounds calculated from synthetic data. However, we do not argue that this is the minimal bound that can be found, nor that it is impossible to pass the bound, as we can only evaluate a subset of all possible data sets. What makes us optimistic is that it has been shown that there exists an algorithm with sub-linear sample complexity to estimate *CMI* (Canonne et al., 2018).

The problem that we would like to solve is to provide a formula that calculates the number of samples n such that $P(|SCI_f^n(X;Y \mid Z) - I(X;Y \mid Z)| \geq \epsilon) \leq \delta$, for small ϵ and δ. Thereby, we focus on an n such that the probability of making a type I error, i.e. rejecting independence when $H_0\colon X \perp\!\!\!\perp Y \mid Z$ is true, is low. In our empirical evaluation, we set $\epsilon = \delta = 0.05$ and draw samples from data with the ground truth $I(X;Y \mid Z) = 0$ by assigning equal probabilities to each value combination of X, Y and Z—i.e. we set $P(x,y,z) = \frac{1}{|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|}$ for each value configuration $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. We 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 calculate $P(|SCI_f^n(X;Y \mid Z) - I(X;Y \mid Z)| \geq \epsilon) = P(SCI_f^n(X;Y \mid Z) \geq 0.05) \leq 0.05$ as follows: We start with a small n, e.g. 2, generate 1 000 data sets and check if over those data sets $P(SCI_f^n(X;Y \mid Z) \geq 0.05) \leq 0.05$ 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 $P(SCI_f^n(X;Y \mid Z) \geq 0.05) \leq 0.05$ holds and report this n.

In Figure 13 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 that shows 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 $P(SCI_f^n(X;Y \mid Z) \geq 0.05) \leq 0.05$ 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. 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 a tighter bound, however, we did not optimize for that. For future work we would like to theoretically validate a sub-linear bound function.

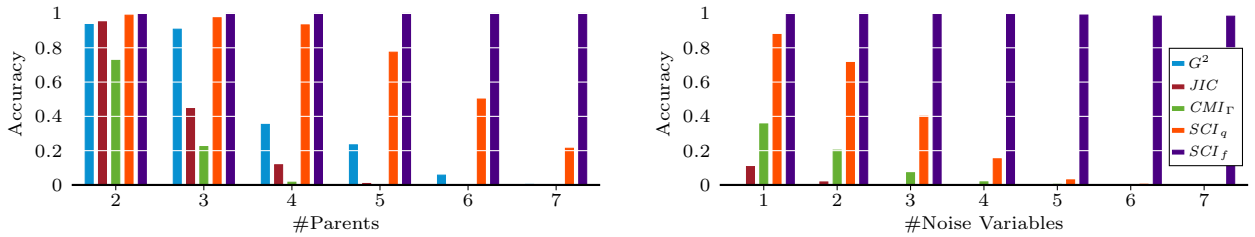


Figure 12: Left: Percentage of parents identified, where we start with only two parents and increase the number of parents to seven. Right: Percentage of parents identified, where we always use three parents, add independently drawn noise variables to the conditioning set.

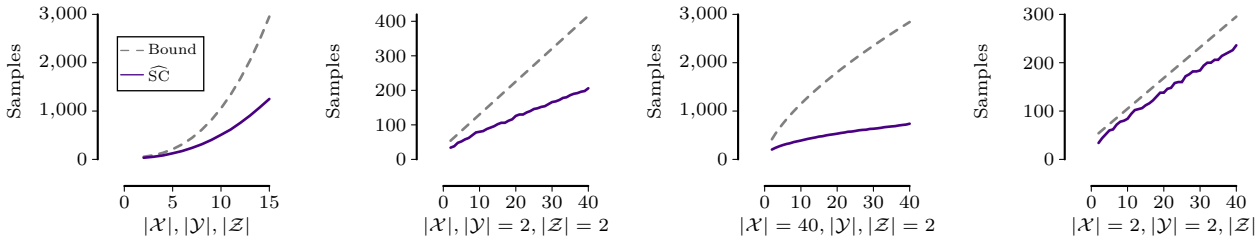


Figure 13: Estimated sample complexities for independently generated data s.t. $P(|SCI_f^n/n - I| \geq 0.05) \leq 0.05$. The suggested bound is calculated as $35 + 2|\mathcal{X}||\mathcal{Y}|^{\frac{2}{3}}(|\mathcal{Z}| + 1)$. For all setups, increasing the domain size of X, Y, Z together or independently, the bound function is larger than the empirical value.