Stochastic Complexity for Testing Conditional Independence on Discrete Data

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 part of constraint-based causal discovery. We specifically focus on discrete data. Although commonly used tests are perfect in theory, due to data sparsity they often fail to reject independence in practice—especially when conditioning on multiple variables.

We propose a new conditional independence test based on the notion of algorithmic independence. To instantiate this ideal formulation in practice, we use stochastic complexity. We show that our proposed test SCI is an asymptotically unbiased estimator for conditional mutual information (CMI) as well as L_2 consistent. Further, we show that SCI can be reformulated to find a sensible threshold for CMI that works well given only limited data.

Empirical evaluation shows that SCI has a lower type II error than commonly used tests, which leads to a higher recall when we use it 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 probability distribution from which the observed data was generated is faithful to the true 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 d-separation from an i.i.d. sample of the joint distribution $P(D, E, F, T)$, as F will be independent of T given D, E. In contrast, $D \not\perp\!\!\!\perp T \mid E, F$, as well as $E \not\perp\!\!\!\perp T \mid D, F$. Identifying these cases correctly depends on the quality of the test.

Conditional independence testing is also important for recovering the Markov blanket of a target node: 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 when given a *perfect* independence test.

In this paper, we are not trying to improve those algorithms, but rather propose a new independence test to enhance their performance. While recently a lot of work focuses on continuous data, as methods range from approximating the continuous conditional mutual information (Runge, 2018) to kernel based methods (Zhang et al., 2011), we focus on discrete data.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

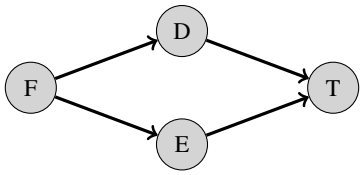


Figure 1: [d-Separation] Given the above causal DAG it holds that $F \perp\!\!\!\perp T \mid D, E$, or F is d-separated of T given D, E under the faithfulness assumption. Note that $D \not\perp\!\!\!\perp T \mid E, F$ and $E \not\perp\!\!\!\perp T \mid D, F$.

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 G^2 test is theoretically sound, it is very restrictive and needs large samples sizes to detect dependencies, especially when conditioned 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, CMI 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 it in practice, it is therefore necessary to set a threshold. This is not an easy task, as the threshold will depend on both the domain sizes of the involved variables and the sample size (Goebel et al., 2005). Often it is assumed that an exponential number of samples is required, but Canonne et al. (2018) show the existence of a sample-efficient algorithm to distinguish dependence from independence using CMI that has sub-linear sample complexity. Perhaps closest to our approach is the work of Goebel et al. (2005). They use a second-order Taylor series to approximate the conditional mutual information and show that this estimate follows the gamma distribution, which allows them to define a threshold based on the domain sizes of the variables and the sample size.

The main problem that previous tests have is that they 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 approach this problem, we build upon algorithmic conditional independence, which has the advantage that we not only consider the statistical dependence, but also its complexity. We can instantiate this ideal formulation with the stochastic complexity. In essence, we compute the stochastic complexity using the factorized normalized maximum likelihood (fNML) (Silander et al., 2008), 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 proof that SCI is an asymptotically unbiased estimator of CMI and L_2 consistent. For limited data, we show that this threshold behaves similar to the one defined by Goebel et al. (2005), but has additional properties. That is, the threshold derived from SCI does not only consider the sample size and the dimensionality of the data, but also the estimated probability mass functions of the conditioning variables. In practice, this reduces the type II error as we show in our experiments. Further, when applying SCI in constraint based causal discovery algorithms, we observe a higher precision and recall than related tests.

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 $H_0 : X \perp\!\!\!\perp Y \mid Z$ against the general alternative $H_1 : X \not\perp\!\!\!\perp Y \mid Z$. The main goal of an independence test is to minimize the type I and II error. The type I error is defined as falsely rejecting the null hypothesis and the type II error is defined as falsely accepting H_0.

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, only if we are given the true distributions of the random variables. In practice, those are not given and we have to estimate the entropy from limited data. As a consequence, conditional mutual information often overestimates dependencies, even if the involved variables are independent of each other, i.e. consider Example 1.

Example 1 *Given three random variables X, Y_1 and Y_2, with corresponding domain sizes $4, 8$ and $1\,000$. Suppose that given $1\,000$ samples X is a deterministic function of Y_1, as well as of Y_2. That is, $\hat{H}(X \mid Y_1) = \hat{H}(X \mid Y_2) = 0$. However, as only a single sample exists for each $v \in Y_2$, it is very likely that $\hat{H}(X \mid Y_2) = 0$ only due to the limited amount of samples and will be > 0 given more samples. It is more likely that $\hat{H}(X \mid Y_1) = 0$ due to a true dependency, since the number of samples $n \gg |\mathcal{Y}_1|$—i.e. we have more evidence.*

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 sample size and the domain sizes of X, Y and Z. 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, y_1 and y_2 be the binary strings representing X, Y_1 and Y_2. As X can be expressed as a deterministic function of Y_1 or Y_2, $K(x \mid y_1)$ and $K(x \mid y_2)$ reduces to the program describing this function. As the domain size of Y_1 is 8 and $|\mathcal{X}| = 4$, the program to describe X from Y_1 only has to describe the mapping from 8 to 4 values, which will be shorter than describing a mapping from Y_2 to X, since $|\mathcal{Y}_2| = 1\,000$—i.e. $K(x \mid y_1) \leq K(x \mid y_2)$ in contrast $\hat{H}(X \mid Y_1) = \hat{H}(X \mid Y_2)$. To reject $X \perp\!\!\!\perp Y \mid Z$, we test whether prividing the information of Y 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.

Sadly, Kolmogorov complexity is not computable, nor approximable up to arbitrary precision due to the halting problem (Li and Vitányi, 1993). We can approximate it, however, through the Minimum

Description Length (MDL) principle (Grünwald, 2007). 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)*

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

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

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

The sum goes over every possible 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 $\mathcal{C}^n_k$ for $\mathcal{C}^n_{\mathcal{M}_k}$.

For discrete data, assuming a multinomial, we 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

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

Next, we show how to compute the conditional stochastic complexity.

3.1 Conditional Stochastic Complexity

To compute the conditional stochastic complexity, we use the factorized normalized maximum likelihood (fNML) (Silander et al., 2008).

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

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

with $\mathcal{Y}$ the domain of Y with domain size l, and $|v|$ is the frequency of a value v in y^n.

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 the ideal conditional stochastic

complexity as S and to the instantiation based on fNML with S_f. In addition, we refer to the regret terms of the conditional $S(X \mid Z)$ as $\mathcal{R}(X \mid Z)$, where for fNML

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

Next, we introduce two important properties of the regret term.

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.*

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

For conciseness, we postpone both proofs to the appendix. 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*

$$SCI(X; Y \mid Z) := S(X \mid Z) - S(X \mid Z, Y) \leq 0 \ . \tag{3}$$

In particular, Eq. 3 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 can see that the regret terms essentially formulate a threshold t_S for the 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 $\mathcal{R}_f(X \mid Z, Y) - \mathcal{R}_f(X \mid Z) \geq 0$. Hence, Y has to provide a significant gain such that $X \not\perp\!\!\!\perp Y \mid Z$—i.e. $n \cdot (H(X \mid Z) - H(X \mid Z, Y))$ must be $> t_S$.

Next, we show how we can use SCI in practice by formulating it based on fNML.

4.1 Factorized SCI

To formulate our independence test based on the 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 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 $SCI_f(X; Y \mid Z) \leq 0$.

Note that we could also define alternative formulations of SCI, such as defining it via the recently proposed quotient normalized maximum likelihood (qNML) (Silander et al., 2018), however, preliminary results showed that these formulations lead to worse results. A more in depth analysis of alternative formulations of SCI is part of future work.

In the next section, we show the main properties for SCI_f. Thereafter, we compare SCI_f to CMI using the threshold based on the gamma distribution, as proposed by Goebel et al. (2005).

4.2 Properties of SCI

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

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

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 this formulation consists of $n \cdot I(X;Y \mid Z)$, it will be zero by definition. From Theorem 1, we know that $\mathcal{R}_f(X \mid Z) - \mathcal{R}_f(X \mid Z,Y) \leq 0$, which concludes the proof. □

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

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

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

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

The proof for $I_S^f(Y;X \mid Z)$ follows analogously. In essence, we need to show that $\frac{1}{n}(\mathcal{R}_f(X \mid Z) - \mathcal{R}_f(X \mid Z,Y))$ 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, $\frac{1}{n}\mathcal{R}_f(X \mid Z)$ and $\frac{1}{n}\mathcal{R}_f(X \mid Z,Y)$ will approach zero if $n \to \infty$. □

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

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

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

Next, we compare SCI_f to the findings of Goebel et al. (2005).

4.3 Link to Gamma Distribution

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

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

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 the conditional mutual information based on the gamma distribution—for convenience we call this threshold t_Γ.

In contrast to t_Γ, $t_S = \mathcal{R}_f(X \mid Z,Y) - \mathcal{R}_f(X \mid Z)$ the regret terms for both $\mathcal{R}_f(X \mid Z)$ and $\mathcal{R}_f(X \mid Z,Y)$ 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 (see 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 (see (Cover and Thomas, 2006)).

Figure 2: Threshold for CMI using fNML and the gamma distribution with $\alpha = 0.05$ ($\Gamma_{.05}$) and $\alpha = 0.001$ ($\Gamma_{.001}$) for different sample sizes and fixed domain sizes equal to four (left) and fixed sample size of 500 and changing domain sizes (right).

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 empirically compare them. First, we compare t_Γ with $\alpha = 0.05$ and $\alpha = 0.001$ to t_S/n on fixed domain sizes, with $|\mathcal{X}| = |\mathcal{Y}| = |\mathcal{Z}| = 4$ and varying the sample sizes (see Figure 2). For t_S we computed the worst case threshold under the assumption that Z is uniformly distributed. In general, both thresholds behave similar, whereas t_S is more restrictive.

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. Again we observe that t_S is more restrictive than t_Γ 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 t_S 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.

5 Experiments

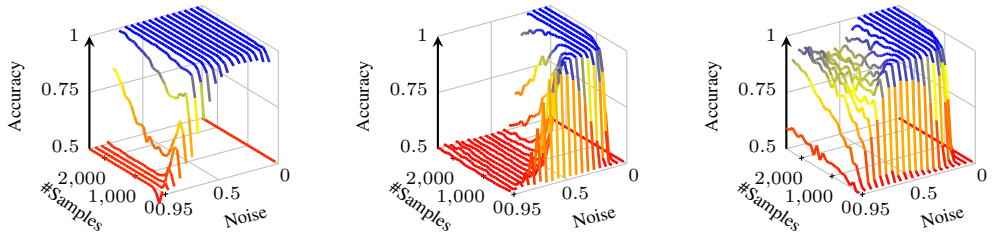
In this section, we empirically evaluate our proposed independence test based on fNML and compare it to the G^2 test from the *pcalg* R package (Kalisch et al., 2012) and CMI_Γ (Goebel et al., 2005).

5.1 Identifying d-Separation

To test whether SCI_f 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 $x \in X$ to a $y \in Y$. 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.[1] For the G^2 test and CMI_Γ we select $\alpha = 0.05$, however, we found no significant differences for $\alpha = 0.01$.

We plot the accuracy for SCI_f, G^2 and CMI_Γ in Figure 3. 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. In contrast, CMI_Γ only performs well for less than 30% noise and then fails to identify the true independencies, 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. 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 $10(|\mathcal{X} - 1|)(|\mathcal{Y} - 1|)(|\mathcal{Z}|)$.

[1] Note that for 0% noise, F has all information about D and E therefore $D \not\perp\!\!\!\perp T \mid E, F$ and $E \not\perp\!\!\!\perp T \mid D, F$ cannot be identified.



(a) SCI_f (b) CMI_Γ (c) G^2

Figure 3: Accuracy of SCI_f, 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.

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 and also restrict the domain sizes of the remaining variable to the same size. For each setup we generate 200 data sets.

From the results in Figure 4 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_f 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 generated data with different sample sizes and averaged over the results of 10 runs for each sample size. We equipped the stable PC algorithm with SCI_f, CMI_Γ and the default, the G^2 test, and plot the average $F1$ score over the undirected graphs in Figure 5. 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.

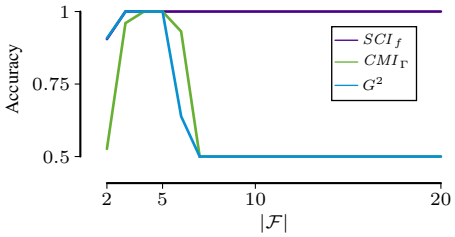


Figure 4: Accuracy of SCI_f, CMI_Γ and G^2 for d-separation with $2\,000$ samples and 10% noise for increasing domain size of the F.

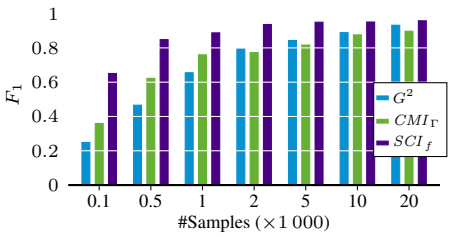


Figure 5: F_1 score on undirected edges for stable PC using SCI_f, CMI_Γ and G^2 on the *Alarm* network given different sample sizes.

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 different independence tests. We plot the precision and recall for each variant in Figure 6. We observe that 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

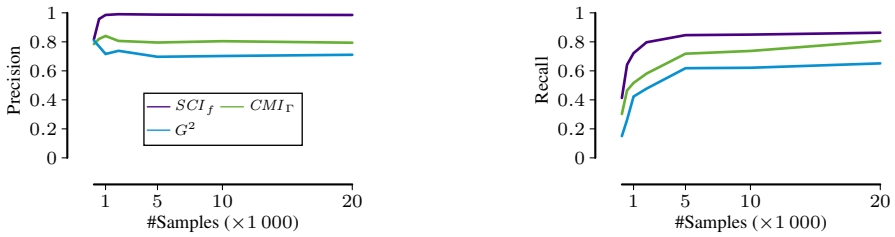


Figure 6: [Higher is better] Precision (left) and recall (right) for PCMB using SCI_f, CMI_Γ and G^2 to identify all Markov blankets in the *Alarm* network for different sample sizes.

for the 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 thresholds drawn from the gamma distribution, SCI_f 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 SCI_f clearly outperforms its competitors on synthetic and real world data.

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.

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., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT press.

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 Appendix

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

$$m(k,n) = m(k-1,n)\frac{(n-k+1)(k+L-2)}{nk} \ , \tag{4}$$

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.

$$\begin{aligned} \frac{\mathcal{C}_L^n}{\mathcal{C}_L^{n-1}} &= \frac{\sum_{k=0}^{n} m(k,n)}{\sum_{k=0}^{n-1} m(k,n-1)} \\ &= \frac{\sum_{k=0}^{n-1} m(k,n)}{\sum_{k=0}^{n-1} m(k,n-1)} + \frac{m(n,n)}{\sum_{k=0}^{n-1} m(k,n-1)} \end{aligned} \tag{5}$$

Next, we will show that both parts of the sum in Eq. 5 are decreasing when n increases. We start with the left part, which we rewrite to

$$\begin{aligned} \frac{\sum_{k=0}^{n-1} m(k,n)}{\sum_{k=0}^{n-1} m(k,n-1)} &= \frac{\sum_{k=0}^{n-1} m(k,n-1) + \sum_{k=0}^{n-1} \left(m(k,n) - m(k,n-1)\right)}{\sum_{k=0}^{n-1} m(k,n-1)} \\ &= 1 + \frac{\sum_{k=0}^{n-1} \frac{(L-1)^{\bar{k}}}{k!} \left(\frac{n^{\underline{k}}}{n^k} - \frac{(n-1)^{\underline{k}}}{(n-1)^k}\right)}{\sum_{k=0}^{n-1} m(k,n-1)} \ . \end{aligned} \tag{6}$$

When n increases, each term of the sum in the numerator in Eq. 6 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. 5 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. 4 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. 6. Hence the whole term is decreasing, which concludes the proof. □

A.2 Proof of Theorem 1

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, $\{l_1, \ldots, _q\}$, increases to q, where $p \leq q$. Consequently, to show that Theorem 1 is true, it suffices to show that

$$\sum_{i=1}^{p} \log \mathcal{C}_k^{|r_i|} \leq \sum_{j=1}^{q} \log \mathcal{C}_k^{|l_j|} \tag{7}$$

whereas $\sum_{i=1}^{p} |r_i| = \sum_{j=1}^{q} |l_j| = n$. Next, consider w.l.o.g. that each value combination $\{r_i\}_{i=1,\dots,p}$ is mapped to one or more value combinations in $\{l_1, \dots, l_q\}$. Hence, Eq. (7) 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. $\square$