MDL for Causal Inference on Discrete Data

Kailash Budhathoki and Jilles Vreeken
Max Planck Institute for Informatics and Saarland University
Saarland Informatics Campus, Saarbrücken, Germany
{kbudhath,jilles}@mpi-inf.mpg.de

***Abstract*—The algorithmic Markov condition states that the most likely causal direction between two random variables X and Y can be identified as the direction with the lowest Kolmogorov complexity. This notion is very powerful as it can detect *any* causal dependency that can be explained by a physical process. However, due to the halting problem, it is also not computable.**

In this paper we propose an computable instantiation that provably maintains the key aspects of the ideal. We propose to approximate Kolmogorov complexity via the Minimum Description Length (MDL) principle, using a score that is mini-max optimal with regard to the model class under consideration. This means that even in an adversarial setting, the score degrades gracefully, and we are still maximally able to detect dependencies between the marginal and the conditional distribution.

As a proof of concept, we propose CISC, a linear-time algorithm for causal inference by stochastic complexity, for pairs of univariate discrete variables. Experiments show that CISC is highly accurate on synthetic, benchmark, as well as real-world data, outperforming the state of the art by a margin, and scales extremely well with regard to sample and domain sizes.

***Index Terms*—causal inference; MDL; discrete data**

I. Introduction

Causal inference from data that was not collected through carefully controlled randomised trials is a fundamental problem in both business and science [23], [15]. A particularly interesting setting is to tell cause from effect between a pair of random variables X and Y, given data over the joint distribution. That is, to identify which of the Markov equivalent cases $X \to Y$ or $Y \to X$ is the most likely.

In recent years, a number of ideas have been proposed for causal inference based on properties of the joint distribution. These ideas include the Additive Noise Model (ANM), where we assume the effect is a function of the cause with additive noise independent of the cause [21], [16], [17], and that of the algorithmic Markov condition [7], [1], which is based on Kolmogorov Complexity. Loosely speaking, the key idea of the latter is that if X causes Y, the shortest description of the joint distribution $P(X, Y)$ is given by the separate descriptions of $P(X)$ and $P(Y \mid X)$. Kolmogorov complexity, however, is not computable, and hence any method that builds on this observation requires a computable approximation, which in general involves arbitrary choices [20], [24], [11], [8].

We define a causal inference rule that, while based on the algorithmic Markov condition, is computable and guaranteed to maintain the key properties of the ideal score. That is, we propose to approximate Kolmogorov complexity via the Minimum Description Length (MDL) principle using Stochastic Complexity, which is a score that is mini-max optimal with regard to the model class under consideration. This means that even in an adversarial setting where the true data generating distribution does not reside in our model class $\mathcal{M}$, we still obtain the optimal encoding for the data relative to $\mathcal{M}$ [3].

We show the strength of this approach by instantiating it for pairs of univariate discrete data, using multinomial stochastic complexity. For this setting, stochastic complexity is remarkably efficiently to compute, by which our score has only a linear-time computational complexity. Through experiments on synthetic and benchmark data we show that our method performs very well in practice and outperforms the state of the art by a large margin. Last, but not least, we perform two case studies that show CISC indeed infers sensible causal directions from real-world data.

In sum, the main contributions of this paper are as follows.

(a) we propose the first computable framework for causal inference by the algorithmic Markov condition with provable mini-max optimality guarantees,
(b) define a causal indicator for pairs of discrete variables based on stochastic complexity,
(c) show how to efficiently compute it,
(d) provide extensive experimental results on synthetic, benchmark, and real-world data, and
(e) make our implementation and all used data available

II. Preliminaries

In this section, we introduce notations and background definitions we will use in subsequent sections.

A. 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 [9], [10]. Thus, $K(x) = |p^*|$. The conditional Kolmogorov complexity $K(x \mid y)$ of x relative to y is defined similarly as the length of the shortest program that generates x, and halts, given y as input. The amount of algorithmic information that y contains about x is defined as $I(y : x) = K(y) - K(y \mid x)$. Up to an additive constant term, $I(x : y) = I(y : x)$.

The Kolmogorov complexity of a probability distribution P, $K(P)$, is the length of the shortest program that outputs $P(x)$ to precision q on input $\langle x, q \rangle$ [4]. The conditional variant $K(P \mid Q)$ is defined similarly but with the additional information Q. Finally the algorithmic mutual information between distributions P and Q is $I(P : Q) = K(P) - K(P \mid Q^*)$, where Q^* is the shortest program for Q.

III. Causal Inference by AIT

Given two statistically dependent variables X and Y, we want to infer their causal relationship. In particular, we want to infer whether X causes Y, whether Y causes X, or they are only correlated. In doing so, we take the usual assumption of causal sufficiency [16], [17], [8]. That is, we assume there is no confounding variable, i.e. hidden common cause, Z of X and Y. We use $X \to Y$ to indicate X causes Y.

We base our inference method on the following postulate:

Postulate 1 (independence of input and mechanism [20])**.** *If $X \to Y$, the marginal distribution of the cause $P(X)$, and the conditional distribution of the effect given the cause, $P(Y \mid X)$ are* independent *— $P(X)$ contains no information about $P(Y \mid X)$ — and vice versa since they correspond to independent mechanisms of nature.*

The notion of *independence*, however, is abstract, and requires formalization. A rather general, yet theoretical sound formalization is that using the algorithmic information theory [7] (AIT). The following theorem is hence a consequence of the *algorithmic* independence of input and mechanism.

Theorem 1 (Th. 1 in [13])**.** *If X is a cause of Y,*

$$K(P(X)) + K(P(Y \mid X)) \leq K(P(Y)) + K(P(X \mid Y)) \; .$$

holds up to an additive constant.

In other words, we can perform causal inference simply by identifying that direction between X and Y for which the factorization of the joint distribution has the lowest Kolmogorov complexity. Although this inference rule has sound theoretical foundations, Kolmogorov complexity is not computable due to the *halting problem*. However, the Minimum Description Length (MDL) principle [18], [3] provides a statistically sound and computable means for approximating Kolmogorov complexity [3].

IV. Causal Inference by MDL

In this section, we discuss MDL for causal inference.

A. Minimum Description Length Principle

The Minimum Description Length (MDL) [18] principle is a practical version of Kolmogorov complexity. Instead of all possible programs, it considers only programs for which we know they generate x and halt. That is, lossless compressors.

In MDL theory, programs are often referred to as *models*. By MDL principle, the best model is the one that describes the data best when the complexity of the model is also accounted for [3]. The shortest description of the data relative to a model class is called the *stochastic complexity*. Typically code length is used as a measure of description of the data.

B. Stochastic Complexity

Let $X = (x_1, x_2, \ldots, x_n)$ be an i.i.d. sample of n observed outcomes, where each outcome x_i is an element from domain $\mathcal{X}$. Let $\Theta \in \mathrm{R}^d$, where $d \in \mathbb{Z}^+$, be the parameter space. As model class $\mathcal{M}$ we consider a family of probability distributions consisting of all the different distributions $P(. \mid \boldsymbol{\theta})$ that can be produced by varying the parameters $\boldsymbol{\theta}$. Formally, a model class $\mathcal{M}$ is defined as $\mathcal{M} = \{P(\cdot \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$.

Let $P(\cdot \mid \hat{\boldsymbol{\theta}}(X, \mathcal{M}))$ be distribution induced by the maximum likelihood estimate $\hat{\boldsymbol{\theta}}(X, \mathcal{M})$ of X relative to $\mathcal{M}$. The Normalized Maximum Likelihood (NML) distribution is then defined as

$$P_{\mathrm{NML}}(X; \mathcal{M}) = \frac{P(X \mid \hat{\boldsymbol{\theta}}(X, \mathcal{M}))}{R(\mathcal{M}, n)} \; ,$$

where the normalizing term $R(\mathcal{M}, n)$ is the sum over maximum likelihoods of all possible datasets of size n relative to $\mathcal{M}$. For discrete data, $R(\mathcal{M}, n)$ is defined as

$$R(\mathcal{M}, n) = \sum_{X' \in \mathcal{X}^n} P(X' \mid \hat{\boldsymbol{\theta}}(X'; \mathcal{M})) \; , \tag{1}$$

where $\mathcal{X}^n$ is the n-fold Cartesian product $\mathcal{X} \times \cdots \times \mathcal{X}$ indicating the set of all possible datasets of size n with domain $\mathcal{X}$. If data X is defined over a continuous sample space, the summation symbol in Eq. (1) is replaced by an integral.

The NML distribution has a number of important theoretical properties. First, it gives a unique solution to the minimax problem posed by Shtarkov [22],

$$\min_{\bar{P}} \max_{X} \log \frac{P(X \mid \hat{\boldsymbol{\theta}}(X, \mathcal{M}))}{\bar{P}(X \mid \mathcal{M})} \; .$$

That is, for *any* data X, $P_{\mathrm{NML}}(X; \mathcal{M})$ assigns a probability, which differs from the highest achievable probability within the model class — the maximum likelihood $P(X \mid \hat{\boldsymbol{\theta}}(X; \mathcal{M}))$ — by a constant factor $R(\mathcal{M}, n)$. In other words, the NML distribution is the *mini-max optimal universal model* with respect to the model class.

Second, it also provides solution to another mini-max problem formulated by Rissanen [19], which is given by

$$\min_{\bar{P}} \max_{Q} E_Q \left(\log \frac{P(X \mid \hat{\boldsymbol{\theta}}(X; \mathcal{M}))}{\bar{P}(X; \mathcal{M})} \right) \; ,$$

where Q is the worst-case data generating distribution (outside the model class $\mathcal{M}$), and E_Q is the expectation over X. That is, even if the true data generating distribution does *not* reside in the model class $\mathcal{M}$ under consideration, $P_{\mathrm{NML}}(X \mid \mathcal{M})$ *still* gives the optimal encoding for the data X relative to $\mathcal{M}$.

These properties are very important and relevant when modelling real-world problems. In most cases, we do not know the true data generating distribution. In such cases, ideally we would want to encode our data as best as possible — close to the optimal under the true distribution. The NML distribution provides a theoretically sound means for that.

The *stochastic complexity* of data X relative to a model class $\mathcal{M}$ using the NML distribution is defined as

$$\begin{aligned} \mathcal{S}(X; \mathcal{M}) &= -\log P_{\mathrm{NML}}(X; \mathcal{M}) \\ &= -\log P(X \mid \hat{\boldsymbol{\theta}}(X; \mathcal{M})) + \log R(\mathcal{M}, n) \; . \end{aligned}$$

The term $\log R(\mathcal{M}, n)$ is the *parametric* complexity of the model class $\mathcal{M}$. It indicates how well $\mathcal{M}$ can fit random data.

The stochastic complexity of data under a model class $\mathcal{M}$ gives the shortest description of the data relative to $\mathcal{M}$. Hence the richer the $\mathcal{M}$, the closer we are to Kolmogorov complexity. Intuitively, it is also the amount of information, in bits, in the data relative to the model class.

C. Multinomial Stochastic Complexity

We consider discrete random variable X with m values. Furthermore we assume that data $X = (x_1, \ldots, x_n)$ is multinomially distributed, and the space of observations $\mathcal{X}$ is $\{1, 2, \ldots, m\}$. The multinomial model class $\mathcal{M}_m$ then is

$$\mathcal{M}_m = \{P(X \mid \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_m\} \ ,$$

where Θ is the simplex-shaped parameter space given by

$$\Theta_m = \{\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m) : \theta_j \geq 0, \theta_1 + \cdots + \theta_m = 1\} \ ,$$

with $\theta_j = P(X = j \mid \boldsymbol{\theta}), j = 1, \ldots, m$. The maximum likelihood parameters for a multinomial distribution are given by $\hat{\boldsymbol{\theta}}(X, \mathcal{M}_m) = (h_1/n, \ldots, h_m/n)$, where h_j is the number of times an outcome j is seen in X. Then the distribution induced by the maximum likelihood parameters for X under the model class $\mathcal{M}_m$ is given by

$$P(X \mid \hat{\boldsymbol{\theta}}(X; \mathcal{M}_m)) = \prod_{j=1}^{m} \left(\frac{h_j}{n}\right)^{h_j} \ .$$

And the normalizing term $R(\mathcal{M}_m, n)$ is given by

$$R(\mathcal{M}_m, n) = \sum_{h_1 + \cdots + h_m = n} \frac{n!}{h_1! \cdots h_m!} \prod_{j=1}^{m} \left(\frac{h_j}{n}\right)^{h_j} \ .$$

Then the NML distribution for X relative to $\mathcal{M}_m$ is

$$P_{\text{NML}}(X; \mathcal{M}_m) = \frac{\prod_{j=1}^{m} (h_j/n)^{h_j}}{R(\mathcal{M}_m, n)} \ .$$

At last, the stochastic complexity of X relative to $\mathcal{M}_m$ is

$$\mathcal{S}(X; \mathcal{M}_m) = n \log n - \sum_{j=1}^{m} h_j \log h_j + \log R(\mathcal{M}_m, n) \ .$$

We can compute the counts h_j in $\mathcal{O}(n)$ with a single pass over the data. Although the normalizing sum is exponential in m, we can approximate it up to a finite floating-point precision of d digits in *sub-linear* time with respect to the data size n given precomputed counts h_i [12]. Altogether we can compute the multinomial stochastic complexity in linear time, $\mathcal{O}(n)$. In the experiments we use $d = 10$.

D. Conditional Stochastic Complexity

For our purpose, we also need the *conditional* stochastic complexity $\mathcal{S}(Y \mid X; \mathcal{M}_m)$. Let $\mathcal{S}(Y \mid X = x; \mathcal{M}_m)$ be the stochastic complexity of Y conditioned on $X = x$. To obtain the complexity of the conditional distribution, we define the conditional stochastic complexity $\mathcal{S}(Y \mid X; \mathcal{M}_m)$ as a weighted sum of $\mathcal{S}(Y \mid X = x; \mathcal{M}_m)$ over all possible values of X, using the relative frequencies $w_x = h_x/n$ as weights. More formally,

$$\mathcal{S}(Y \mid X; \mathcal{M}_m) \coloneqq \sum_{x \in \mathcal{X}} w_x \mathcal{S}(Y \mid X = x; \mathcal{M}_m) \ .$$

That is, first we form the sub-populations of Y by grouping those outcomes that share the same $x \in \mathcal{X}$ value. We then compute the stochastic complexities of each group. Finally, we aggregate the locally computed stochastic complexities using the frequency of the corresponding x value as a weight.

If there exists a bijective function between X and Y, then $P_{\text{NML}}(Y \mid X = x; \mathcal{M}) = 1/K$, where K is the normalizing term which is a constant for a fixed domain $\mathcal{Y}$, is maximal and hence $\mathcal{S}(Y \mid X; \mathcal{M}_m)$ is minimal. If there is no bijection, $\mathcal{S}(Y \mid X = x; \mathcal{M}_m)$ gives the additional bits that we need compared to the bijective case. This comes very close in spirit to the Additive Noise Models (ANM). Instead of assuming noise as an additive variable that represents the shortcoming of the bijective function in explaining Y as in ANM, we consider noise as the randomness introduced by different mappings for a specific x value directly.

We can compute $\mathcal{S}(Y \mid X = x; \mathcal{M}_m)$ in $\mathcal{O}(n)$. To compute the conditional stochastic complexity $\mathcal{S}(Y \mid X; \mathcal{M}_m)$, we have to compute $\mathcal{S}(Y \mid X = x; \mathcal{M}_m)$ over all $x \in \mathcal{X}$. Hence the computational complexity of conditional stochastic complexity is $\mathcal{O}(n|\mathcal{X}|)$. Now that we have defined both stochastic complexity, and its conditional variant, next we discuss how they can be used for causal inference.

E. Causal Inference by Stochastic Complexity

The stochastic complexity of data X relative to model class $\mathcal{M}$ corresponds to the complexity of the NML distribution of the data relative to $\mathcal{M}$. This means we can use the stochastic complexity of X as an approximation of the Kolmogorov complexity of $P(X)$. As such, it provides a general, yet computable, theoretically sound approach for causal inference based on the algorithmic Markov condition.

To infer the causal direction, we look over total stochastic complexity in two directions — X to Y and vice versa. The total stochastic complexity from X to Y, approximating $K(P(X)) + K(P(Y \mid X))$ is given by

$$\mathcal{S}_{X \to Y} = \mathcal{S}(X; \mathcal{M}_m) + \mathcal{S}(Y \mid X; \mathcal{M}_m) \ ,$$

and that from Y to X is given by

$$\mathcal{S}_{Y \to X} = \mathcal{S}(Y; \mathcal{M}_m) + \mathcal{S}(X \mid Y; \mathcal{M}_m) \ .$$

Following Theorem 1, using the above indicators we arrive at the following causal inference rules.

- If $\mathcal{S}_{X \to Y} < \mathcal{S}_{Y \to X}$, we infer $X \to Y$.
- If $\mathcal{S}_{X \to Y} > \mathcal{S}_{Y \to X}$, we infer $Y \to X$.
- If $\mathcal{S}_{X \to Y} = \mathcal{S}_{Y \to X}$, we are undecided.

That is, if describing X and then describing Y given X is easier — in terms of stochastic complexity — than vice versa, we infer X is likely the cause of Y. If it is the other way around, we infer Y is likely the cause of X. If both

ways of describing are equally complex, or within a user-specific threshold, we remain undecided. We refer to this framework as CISC. The computational complexity of CISC is $\mathcal{O}(n \max(|\mathcal{X}|, |\mathcal{Y}|))$.

Causal inference using stochastic complexity has a number of powerful properties. First, unlike Kolmogorov complexity, stochastic complexity is computable. Second, the inference rule is generic in the sense that we are not restricted to one data type or distribution—we are only constrained by the model class $\mathcal{M}$ under consideration, yet by the mini-max property of NML we know that even if the data generating distribution is adversarial, we still identify the best encoding w.r.t. $\mathcal{M}$.

V. Related Work

Constraint-based approaches like conditional independence test [15] are one of the widely used causal inference frameworks. However, they require at least three observed random variables. Therefore they cannot distinguish between $X \to Y$ and $Y \to X$ as the factorization of the joint distribution $P(X, Y)$ is the same in both direction, i.e. $P(X)P(Y \mid X) = P(Y)P(X \mid Y)$.

In recent years, several methods have been proposed that exploit sophisticated properties of the joint distribution. The linear trace method [6] infers linear causal relations of the form $Y = AX$, where A is the structure matrix that maps the cause to the effect, using the linear trace condition. The kernelized trace method [2] infers non-linear causal relations by mapping the observations to high dimensional reproducing kernel Hilbert space.

The Additive Noise Models (ANM) [21] assume that the effect is a function of the cause and the additive noise that is independent of the cause. Causal inference is then done by finding the direction that admits such a model. Over the years, many frameworks for causal inference from real-valued data have been proposed using ANMs [21], [5], [25], [17].

Algorithmic information theory (AIT) also provides a theoretically sound foundation for causal inference [7]. However, as Kolmogorov complexity is not computable, practical instantiations require computable notions of independence. The information-geometric approach [8] defines independence via orthogonality in information space. Vreeken [24] proposes a causal framework based on relative conditional complexity and instantiates it with cumulative entropy to infer the causal direction in continuous real-valued data. Budhathoki & Vreeken [1] propose a decision tree based approach for causal inference on univariate and multivariate binary data.

All above methods consider either continuous real-valued or binary data. Causal inference from discrete data has received much less attention. Peters et al. [16] extend additive noise models to discrete data, and propose the DR algorithm. Liu & Chan [11] (DC) define independence in terms of the distance correlation between empirical distributions $P(X)$ and $P(Y \mid X)$ to infer the causal direction from categorical data.

VI. Experiments

We implemented CISC in Python and provide the source code for research purposes, along with the used datasets, and synthetic dataset generator.[1] All experiments were executed single-threaded on Intel Xeon E5-2643 v3 machine with 256GB memory running Linux. We compare CISC against Discrete Regression (DR) [16], and DC [11]. In particular, we use significance level of $\alpha = 0.05$ for the independence test in DR, and threshold of $\epsilon = 0.0$ for DC.

A. Synthetic Data

To evaluate CISC on the data with known ground truth, we consider synthetic data. We generate synthetic cause-effect pairs with the ground truth $X \to Y$ using an additive noise model (ANM). That is, first we generate the cause X, and then generate the effect Y using the model given by $Y = f(X) + N, N \perp\!\!\!\perp X$, where f is a function, and N is additive noise that is independent of X.

Following [16], we sample X from following distributions, using independently generated uniform noise.

- uniform from $\{1, \ldots, L\}$,
- binomial with parameters (n, p),
- geometric with parameter p,
- hypergeometric with parameters (M, K, N),
- poisson with parameter λ,
- multinomial with parameters $\boldsymbol{\theta}$, and
- negative binomial with parameters (n, p).

We note that even though we generate data following ANM from X to Y, the joint distribution $P(X, Y)$ might admit an additive noise model in the reverse direction. Therefore in some cases where we say that $X \to Y$ is the true direction, $Y \to X$ might also be equally plausible, and hence full accuracy might not be achievable in some cases. However, this happens in only few trivial instances [16].

We choose parameters of the distributions randomly for each model class. We choose L uniformly between 1 and 10, M, K uniformly between 1 and 40, N uniformly between 1 and $\min(40, M + K)$, p uniformly between 0.1 and 0.9, λ uniformly between 1 and 10, $\boldsymbol{\theta}$ randomly s.t. $\sum_{\theta \in \boldsymbol{\theta}} \theta = 1.0$, for every x choose $f(x)$ uniformly between -7 to $+7$, and noise N uniformly between $-t$ to $+t$, where t is uniformly randomly chosen between 1 and 7.

Accuracy — From each model class, we sample 1000 different models, and hence 1000 different cause-effect pairs. For each model, we sample 1000 points, i.e. $n = 1000$. In Figure 1, we compare the *accuracy* (percentage of correct decisions) of CISC against DC and DR for various model classes. We see that CISC either outperforms or is as good as the other methods in all but one case. This certainly proves the generality of CISC.

Although we compute the stochastic complexity under multinomial model class, we are still able to perform as good with other model classes. This is due to the mini-max optimality property of the NML distribution.

Decision Rate — Next we investigate the accuracy of CISC against the fraction of decisions CISC is forced to make. To this end, for each model class, we sample 1000 new different

[1] http://eda.mmci.uni-saarland.de/cisc/

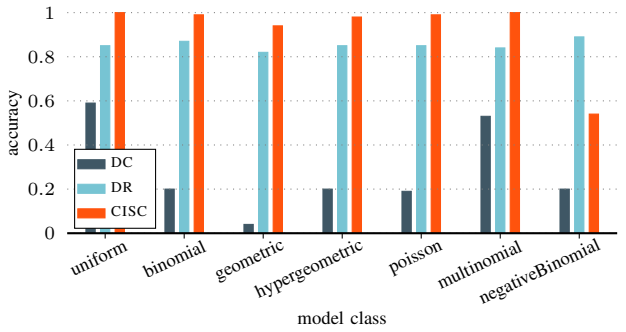


Figure 1: Accuracy on synthetic cause-effect pairs sampled from various model classes.

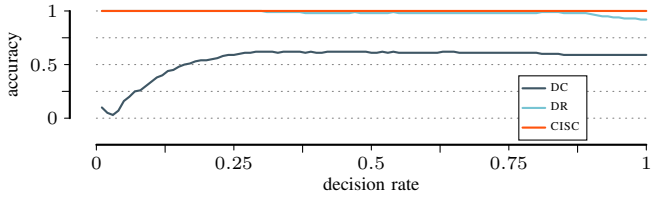


Figure 2: Accuracy versus decision rate on synthetic cause-effect pairs sampled from uniform model class.

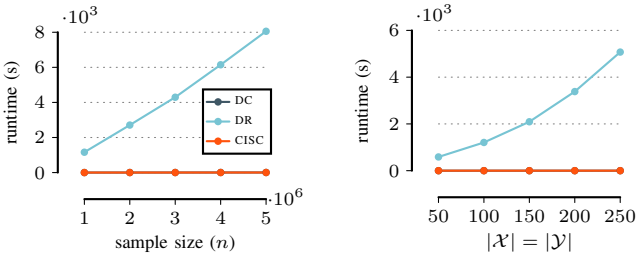
cause-effect pairs. For each cause-effect pair, we sample 1000 points. We sort the pairs by their absolute score difference in two directions ($X \to Y$ vs. $Y \to X$), i.e. $|\mathcal{S}_{X\to Y} - \mathcal{S}_{Y\to X}|$ in descending order. Then we compute the accuracy over top-$k\%$ pairs. The decision rate is the fraction of *top* cause-effect pairs that we consider. Alternatively, it is also the fraction of cause-effect pairs whose $|\mathcal{S}_{X\to Y} - \mathcal{S}_{Y\to X}|$ is greater than some threshold δ. For undecided pairs, we flip a coin. For other methods, we follow the similar procedure with their respective absolute score difference.

In Figure 2, we show the decision rate versus accuracy for uniform model class on these samples.[2] We see that both CISC and DR are highly accurate up to a very high decision rate in all cases. Both CISC and DR are highly accurate on the cause-effect pairs where the absolute score difference is very high — where the methods are most decisive. DC, on the other hand, performs poorly. We also observe the similar behaviour with other model classes.

Scalability — Next we empirically investigate the scalability of CISC. First, we examine runtime with regard to the sample size. To this end, we fix the domain size of the cause-effect pairs to 20, i.e. $|\mathcal{X}| = |\mathcal{Y}| = 20$. Then for a given sample size, we sample X uniformly randomly between 1 and $|\mathcal{X}|$. Likewise for Y.

In Figure 3a, we show the runtime of CISC, DC, and DR for various sample sizes. We observe that both CISC and DC (overlapping line) finish within seconds. DR, on the other hand, takes in the order of hours.

[2]As in this experiment we force the algorithms to decide, as well as consider a fresh sample of randomly generated data, the accuracies at 100% decision rate may differ a little in comparison to Fig. 1



(a) $|\mathcal{X}| = |\mathcal{Y}| = 20$. (b) $n = 100\,000$.

Figure 3: Runtime versus (a) sample size, and (b) domain size.

Next we fix the sample size to $n = 100\,000$ and vary the domain size $|\mathcal{X}| = |\mathcal{Y}|$. We observe that both CISC and DC again finish within seconds over the whole range. As DR iteratively searches over the entire domain, it shows a non-linear runtime behaviour with respect to the domain size.

Overall, these results indicate that DR is fairly accurate, but relatively slow. DC is fast, but (highly) inaccurate. CISC is both highly accurate, and fast.

B. Benchmark Data

To evaluate how well CISC fares on real data that is highly unlikely drawn from a multinomial we evaluate it on 95 real-world benchmark cause-effect pairs with known ground truth [14]. Most of these pairs are continuous valued. As there does not exist a discretization strategy that provably preserves the causal relationship between variables, we not know the underlying domains of the data, following Peters et al. [16] for all pairs we simply consider the unique values as discrete.

In Figure 4, we compare the accuracy of CISC against DC and DR at various decision rate together with the 95% confidence interval for a random coin flip. If we look over all the pairs, we find that CISC infers correct direction in roughly 67% of all the pairs. When we consider only those pairs where CISC is most decisive—with a very high value of $|\mathcal{S}_{X\to Y} - \mathcal{S}_{Y\to X}|$, it is 100% accurate on top 22% of the pairs, 80% accurate on top 45% of the pairs. Also in the same figure, we compare the accuracy of CISC against various state-of-the-art frameworks for continuous real-valued data namely IGCI [8], CURE [20], and pHSIC [14], [5]. Overall we see that CISC is even on-par with the top-performing causal inference frameworks for continuous real-valued data. DC and DR, on the other hand, are insignificant at almost every decision rate. Especially DR performs notably less well here than on synthetic data.

Further we perform permutation testing on the benchmark pairs using CISC at a significance level of 0.05. In particular, we observe that results of CISC on 55 pairs are statistically significant. Out of those 55 pairs, CISC identified correct direction in 38 cases (69.09%). On the insignificant pairs, CISC performs almost with a coin flip (22 correct, 18 incorrect).

C. Real Data

Abalone — First we consider the *Abalone* dataset from the UCI machine learning repository. The dataset contains the

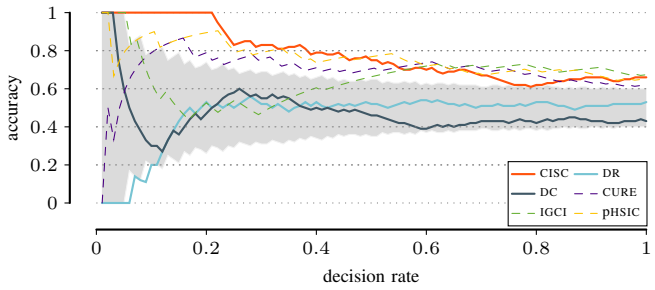


Figure 4: Accuracy versus decision rate for univariate Tübingen cause-effect pairs. Gray area indicates the 95% confidence interval for a random coin flip. Frameworks for discrete data are indicated by solid lines, and that for continuous real-valued data are indicated by dashed lines.

physical measurements of 4 177 abalones, which are large, edible sea snails. We examine *sex* (X) against *length* (Y_1), *diameter* (Y_2), and *height* (Y_3). Following Peters et al. [16], we treat the data as being discrete, and consider $X \to Y_1$, $X \to Y_2$, and $X \to Y_3$ as the ground truth as sex causes the size of the abalone and not the other way around.

We observe that both CISC, and DC infers correct direction in all three cases with a large score difference between two directions in all cases. DR, on the other hand, remains indecisive in the third case.

NLSchools — The *NLSchools* dataset is the 99-th pair in the Tübingen cause-effect benchmark pairs. It contains the language test score (X), and socio-economic status of pupil's family (Y) of 2287 eighth-grade pupils (aged about 11) from 132 classes in 131 schools in the Netherlands.

We regard $Y \to X$ as the ground truth as the socio-economic status of the family is one of the causes of the language test score. With CISC, we get $\mathcal{S}_{X \to Y} = 12168.68$ bits, and $\mathcal{S}_{Y \to X} = 10208.60$ bits. Therefore CISC infers $Y \to X$, which is also the true direction. We note that both DC and DR also identify the correct direction.

Overall, these results illustrate that CISC finds sensible causal directions from real-world data.

VII. Conclusion

We proposed a general, yet *computable* framework for information-theoretic causal inference with optimality guarantees. As a proof of concept, we proposed the linear-time CISC algorithm for causal inference between pairs of univariate discrete variables, using stochastic complexity over the class of multinomial distributions. Extensive evaluation on synthetic, benchmark, and real-world data showed that CISC is highly accurate, outperforming the state of the art by a margin, and scales extremely well to both sample and domain sizes. Future work includes considering richer model classes, as well as structure learning for causal discovery.

Acknowledgements

Kailash Budhathoki is supported by the International Max Planck Research School for Computer Science. Both authors are supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government.

References

[1] K. Budhathoki and J. Vreeken, "Causal inference by compression," in *ICDM*. IEEE, 2016, pp. 41–50.
[2] Z. Chen, K. Zhang, and L. Chan, "Nonlinear causal discovery for high dimensional data: A kernelized trace method," in *ICDM*. IEEE, 2013, pp. 1003–1008.
[3] P. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.
[4] P. D. Grünwald and P. M. B. Vitányi, "Algorithmic information theory," *CoRR*, vol. abs/0809.2754, 2008.
[5] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *NIPS*, 2009, pp. 689–696.
[6] D. Janzing, P. Hoyer, and B. Schölkopf, "Telling cause from effect based on high-dimensional observations," in *ICML*. JMLR, 2010, pp. 479–486.
[7] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic markov condition," *IEEE TIT*, vol. 56, no. 10, pp. 5168–5194, 2010.
[8] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, "Information-geometric approach to inferring causal directions," *AIJ*, vol. 182-183, pp. 1–31, 2012.
[9] A. Kolmogorov, "Three approaches to the quantitative definition of information," *Problemy Peredachi Informatsii*, vol. 1, no. 1, pp. 3–11, 1965.
[10] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 1993.
[11] F. Liu and L. Chan, "Causal inference on discrete data via estimating distance correlations," *Neur. Comp.*, vol. 28, no. 5, pp. 801–814, 2016.
[12] T. Mononen and P. Myllymäki, "Computing the multinomial stochastic complexity in sub-linear time," in *PGM*, 2008, pp. 209–216.
[13] J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Schölkopf, "Probabilistic latent variable models for distinguishing between cause and effect," in *NIPS*. Curran, 2010, pp. 1687–1695.
[14] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf, "Distinguishing cause from effect using observational data: Methods and benchmarks," *JMLR*, vol. 17, no. 32, pp. 1–102, 2016.
[15] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.
[16] J. Peters, D. Janzing, and B. Schölkopf, "Identifying cause and effect on discrete data using additive noise models," in *AISTATS*. JMLR, 2010, pp. 597–604.
[17] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf, "Causal discovery with continuous additive noise models," *JMLR*, vol. 15, pp. 2009–2053, 2014.
[18] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, no. 1, pp. 465–471, 1978.
[19] ——, "Strong optimality of the normalized ML models as universal codes and information in data," *IEEE TIT*, vol. 47, no. 5, pp. 1712–1717, 2001.
[20] E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf, "Inference of cause and effect with unsupervised inverse regression," in *AISTATS*. JMLR, 2015, pp. 847–855.
[21] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, "A linear non-gaussian acyclic model for causal discovery," *JMLR*, vol. 7, pp. 2003–2030, 2006.
[22] Y. M. Shtarkov, "Universal sequential coding of single messages," *Problems of Information Transmission*, vol. 23, no. 3, pp. 175–186, 1987.
[23] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT press, 2000.
[24] J. Vreeken, "Causal inference by direction of information," in *SDM*. SIAM, 2015, pp. 909–917.
[25] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *UAI*. AUAU Press, 2009, pp. 647–655.