Accurate 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*—Additive Noise Models (ANMs) provide a theoretically sound approach to inferring the most likely causal direction between pairs of random variables given only a sample from their joint distribution. The key assumption is that the effect is a function of the cause, with additive noise that is independent of the cause. In many cases ANMs are identifiable. Their performance, however, hinges on the chosen dependence measure, the assumption we make on the true distribution.**

In this paper we propose to use Shannon entropy to measure the dependence within an ANM, which gives us a general approach by which we do not have to assume a true distribution, nor have to perform explicit significance tests during optimization.

The information theoretic formulation gives us a general, efficient, identifiable, and, as the experiments show, highly accurate method for causal inference on pairs of discrete variables—achieving (near) 100% accuracy on both synthetic and real data.

***Index Terms*—Causal Inference, ANM, Shannon Entropy**

I. Introduction

Determining cause from effect is one of the fundamental problems in science. As it is often either very difficult, expensive, or simply impossible to perform a randomized study, the key question is whether we can accurately infer causal directions from observational data. Traditional constraint-based approaches, such as those based on conditional independence tests [1], [2] can do so up to Markov equivalence; they cannot distinguish between $X \to Y$ and $Y \to X$. In this paper we consider exactly this case. In particular, we give a highly accurate and identifiable method for inferring the causal direction between two discrete random variables X and Y given only a sample of their joint distribution.

As follows from the Structural Causal Model [1], we cannot infer the causal direction between a pair of random variables without making assumptions on the model generating the data. The Additive Noise Models (ANMs) are one of the most popular techniques in causal inference [3], [4], [5]. ANMs assume that *effect* is a function of *cause*, with additive noise that is independent of *cause*. Two random variables X and Y with a joint distribution $P(X, Y)$ are said to satisfy an ANM from X to Y if there exists a function f, and a random noise variable N_Y that is independent of X, i.e. $N_Y \perp\!\!\!\perp X$, such that $Y = f(X) + N_Y$. The model is *identifiable* if $P(X, Y)$ admits an ANM from X to Y, but not in the reverse direction. In that case, we say that, under ANM, X is likely the cause of Y. For discrete data, Peters et al. [5] showed that ANMs are generally identifiable.

In this work we propose ACID, an identifiable method for causal inference on discrete data based on information theory. In particular, we propose to use Shannon entropy to measure the dependence between X and N_Y within an ANM. That is, we infer X causes Y iff $H(X) + H(N_Y) < H(Y) + H(N_X)$. Additionally, our formulation allows us to define a natural significance test, which allows us to weed out insignificant inferences. Extensive evaluation shows that ACID achieves (near) 100% accuracy on both synthetic and real-world data, and outperforms the state of the art by a margin.

In the extended version of this paper [6], we additionally show the connection between information-theoretic and algorithmic-information-theoretic formulation of ANMs.

II. Causal Inference

Let X and Y be two discrete random variables with their domains $\mathcal{X}$ and $\mathcal{Y}$ respectively. Traditional statistical analysis of multivariate data involves inference from an *observed* sample from the joint distribution. Causal reasoning, however, requires *manipulation* (intervention, policy, treatment, etc.) on the variables of interest. To reason if X causes Y, we have to compare the distributions of Y under different manipulations of X. In particular, X causes Y if

$$P(Y \mid \text{manipulate } X \text{ to } x_1) \neq P(Y \mid \text{manipulate } X \text{ to } x_1).$$

Note that $P(Y \mid \text{manipulate } X \text{ to } x_1)$ is different from $P(Y \mid X = x)$; the former represents the distribution of Y post-manipulation on X, whereas the latter is merely an *observed* distribution of Y given that we observe $X = x$. The *do*-calculus [1, Chap. 3] provides a mathematical language for expressing such post-manipulation distributions. We represent $P(Y \mid \text{manipulate } X \text{ to } x)$ using do-calculus as $P(Y \mid do(X = x))$, shortly $P(Y \mid do(x))$.

In practice, manipulating variables (setting up an experiment) is often very expensive, unethical, or simply impossible; identifying if smoking causes lung cancer is one such example. Therefore it is desirable to identify causal relationships between variables from observational data. Roughly stated, $P(Y \mid do(x))$ is identifiable if we can estimate it from the observational data alone.

Pioneering work from Pearl [1, Thm. 3.2.5] shows that, under certain conditions, we can **identify** $P(Y \mid do(x))$ from observational data. Roughly, $P(Y \mid do(x))$ is **identifiable** from the observed sample drawn from the joint distribution $P(X, Y, Z)$ *given* a causal diagram $X \leftarrow Z \to Y$ of a *Markovian model* whenever all the common causes (confounders), Z, of both X and Y are measured.

In practice, we often do not know the true causal diagram. Instead, we seek to discover it from observational data. To this end, we can use the conditional independence test to *partially* identify the causal diagram of a Markovian model on the observed data. That is, if X and Y are conditionally independent given Z, i.e. $X \perp\!\!\!\perp Y \mid Z$, then we can infer X—Z—Y (notice that there is no direct link between X and Y in Z's presence). However, we cannot draw the directed edges, as $X \leftarrow Z \rightarrow Y$, $X \rightarrow Z \rightarrow Y$, and $X \leftarrow Z \leftarrow Y$ are Markov equivalent—they encode the same conditional independence.

Suppose that we partially identify X—Z—Y as our causal diagram from the observational data. Can we then draw the direct edges between the variables? Answering this question boils down to telling whether X causes Z, or the other way around, and whether Z causes Y, or vice versa. Inspired by this problem setting, in this work, we consider the general case of inferring the causal direction between two variables from a sample drawn from their joint distribution. For obvious reasons stated above, we assume that there is no common cause.

Slightly changing the notation, we formulate our **problem statement** as the following for two discrete random variables:

Problem 1 (Causal Inference from Two Discrete Variables)**.** *Suppose we are given* only *a sample drawn from the joint distribution $P(X, Y)$ of discrete random variables X and Y. We would like to infer whether X causes Y, or vice versa.*

Alternatively, can we **identify** $P(Y \mid do(x))$ given *only* an observed sample drawn from the joint distribution $P(X, Y)$? An algebraic counterpart to graphical modelling known as Structural Causal Modelling not only offers a powerful construct for formalising the general problem of causal inference [1, Chap. 1], but also has been instrumental in identifying causal relationships between two observed variables [3], [4], [5]. It also forms the basis of our causal inference framework.

III. Structural Causal Models

A Structural Causal Model (SCM) [1, Chap. 1] represents the data-generating process by a set of structural assignments. In case of two variables, given a causal diagram $X \rightarrow Y$, a SCM consists of two structural assignments:

$$\begin{aligned} X &\coloneqq N_X \ , \\ Y &\coloneqq f(X, N_Y) \ , \end{aligned} \tag{1}$$

where f is a function, and N_X and N_Y are statistically independent noises, i.e. $N_X \perp\!\!\!\perp N_Y$.

Note that we use an assignment operator instead of an equality to indicate a functional dependence in SCM since the assignment has a causal meaning; manipulating X leads to a change in the value of Y. To represent manipulations, such as $do(x)$, we simply replace the assignment in Eq. (1) by $X \coloneqq x$. The modified SCM then entails the distribution of Y post-manipulation on X, i.e. $P(Y \mid do(x))$.

For general SCMs without restrictions on the distribution of noise, or the functional form, however, we cannot tell if the sample drawn from the joint distribution $P(X, Y)$ is induced by an SCM from $X \rightarrow Y$, or $Y \rightarrow X$ as we can always construct a suitable function and noise in both directions [7, Prop. 4.1]. In other words, from SCMs in general, the causal structure of two variables is *not* identifiable from the joint distribution; we require additional assumptions to identify the causal direction. A special case of SCMs, known as Additive Noise Models, possess the identifiability that we seek.

The Additive Noise Models (ANMs) are a class of SCMs with the constraint that the noise is *additive*, and *independent* of the exogenous variable (variable whose value is independent from the state of other variables in the system). Given a causal diagram $X \rightarrow Y$, an ANM represents the data generating process as $Y = f(X) + N_Y$, where $N_Y \perp\!\!\!\perp X$.

With some restrictions on the functional form and the distribution of noise, the causal direction is **identifiable** under ANMs from observational data. Identifiability requires asymmetry of some sort in the data generating process. In ANMs, the asymmetry is observed in the independence of noise and the exogenous variable. A causal direction $X \rightarrow Y$ induced by an ANM is identifiable if $P(X, Y)$ admits an ANM from X to Y, but not vice versa.

Shimizu et al. [3] showed that an ANM is identifiable if the function is linear, and the noise is non-Gaussian. Hoyer et al. [4] showed that ANMs are generally identifiable even when the function is non-linear (without any restrictions on the distribution of noise). Of particular interest to us is the work by Peters et al. [5] which shows that ANMs are generally identifiable in the discrete case. As a result, we have following statement for the discrete case;

Definition 1 (Discrete Additive Noise Model)**.** *If a sample drawn from the joint distribution $P(X, Y)$ of two discrete random variables X and Y is induced by an ANM from X to Y, it generally holds that*

- *$X \sim P(X)$, and there exists at least one function f such that $Y = f(X) + N_Y$, where $N_Y \perp\!\!\!\perp X$, but*
- *for any function g, $N_X = X - g(Y)$ depends on Y.*

To identify the causal direction in practice, we fit ANMs in both directions, and choose the direction with independence as the causal direction. As a result the ANM approach hinges on the choice of dependence measure. Most dependence measures either assume the type of the sampling distribution of the test statistic, or require a kernel. Alternatively information-theory offers Shannon entropy as a very intuitive yet powerful tool to measure dependence.

In this work, we take an information-theoretic approach to ANMs, and use entropy as a dependence measure. As such, we avoid explicit null hypothesis testing with p-values. Moreover we can simply work with the empirical distribution. Note that although (differential) entropy has been studied in the context of ANMs on real-valued data [8], [9], seemingly straightforward application of Shannon entropy on ANMs for discrete data has been left out at large.

IV. Information-theoretic ANM

To arrive at the information-theoretic formulation of ANMs, we have to quantify the information contained in a sample drawn from the joint distribution $P(X, Y)$ under ANMs with graphical structures $X \rightarrow Y$ and $Y \rightarrow X$ using Shannon entropy. For a graphical structure $X \rightarrow Y$ modelled by an ANM, we have $P(Y \mid X) = P(N_Y \mid X)$ due to the discriminative nature of ANM modelling. Thus the total entropy of a sample assuming $X \rightarrow Y$ as an underlying graphical structure using an ANM is $H(X) + H(N_Y \mid X)$. Combining this observation with the property of joint Shannon entropy, we can prove the following result.

Theorem 1. *If a sample drawn from the joint distribution $P(X, Y)$ of two discrete random variables X and Y is induced by an ANM with $X \rightarrow Y$ graphical structure, it holds that*

$$H(X) + H(N_Y) < H(Y) + H(N_X),$$

where $N_Y = Y - f(X)$ such that $N_Y \perp\!\!\!\perp X$, and $N_X = X - g(Y)$ such that $N_X \not\perp\!\!\!\perp Y$.

Proof. The entropy of a sample under an ANM with $X \rightarrow Y$ as its underlying graphical structure is given by

$$\begin{aligned} H(X) + H(Y \mid X) &= H(X) + H(N_Y \mid X) \\ &= H(X) + H(N_Y) \qquad (N_Y \perp\!\!\!\perp X). \end{aligned}$$

In the other direction from $Y \rightarrow X$, we have

$$\begin{aligned} H(Y) + H(X \mid Y) &= H(Y) + H(N_X \mid Y) \\ &< H(Y) + H(N_X) \qquad (N_X \not\perp\!\!\!\perp Y). \end{aligned}$$

Combine the two right hand sides above. □

From here onwards, we use $H_{X \rightarrow Y}$ for $H(X) + H(N_Y)$, defining $H_{Y \rightarrow X}$ analogue. Based on Thm. 1, we can perform causal inference using a simple procedure:

- if $H_{X \rightarrow Y} < H_{Y \rightarrow X}$, we infer "$X$ causes Y",
- if $H_{X \rightarrow Y} > H_{Y \rightarrow X}$, we infer "$Y$ causes X",
- if $H_{X \rightarrow Y} = H_{Y \rightarrow X}$, we are undecided.

That is, we prefer the graphical structure with smaller entropy. The larger the absolute difference between the two indicators, i.e. $\Delta = |H_{X \rightarrow Y} - H_{Y \rightarrow X}|$, the stronger our *confidence* in the inference. In practice, we can always set a threshold τ on Δ and treat the results smaller than τ as undecided.

V. The ACID Algorithm

To use the causal inference rules for inferring the causal direction, we require noise variables on both directions. Therefore on each direction, we have to find a function that minimises entropy of the residual. In other words, we need a method for discrete regression.

Unlike for continuous regression, in the discrete case, there is no risk of overfitting; Y may take different values for each value of X, and hence there is no need for regularization. We can hence simply consider all possible functions, and take the one with the minimal value of the loss function.

As a loss function, we consider discrete Shannon entropy. Therefore we aim to find a function that minimises the entropy of the residual. However, even if range of the function lies within the domain of the target variable, we are left with exponentially many choices of functions, thereby making the problem intractable. Hence, we resort to heuristics.

We give the pseudocode for the ACID algorithm in Algorithm 1. To regress Y as a function of X, we start with a function that maps each x value to the most frequently co-occurring y value (line 2–3). Then we iteratively update the function for each x value. To ensure that the algorithm is deterministic, we do so in some canonical order (line 9). To update the function for a x value, we temporarily map x to other y values keeping all other mappings $f(\bar{x})$ with $\bar{x} \neq x$ fixed. We use $f_{j-1}^{x_i \rightarrow y}(X)$ to denote that f_{j-1} temporarily maps x_i to y. From all the mappings, we pick the best one as the one that results in the least entropy of the residual (line 10). If this residual complexity is better than the so-far best residual complexity, we update our function (line 11-14). We keep on iterating as long as the entropy of the residual reduces, or we arrive at the maximum number of iterations J (line 15).

In a nutshell, ACID performs coordinate descent in discrete space. Note that entropy is non-negative, and hence is bounded from below. Since the search space is finite and the entropy of the residual is strictly decreasing in every iteration, the algorithm will converge. It could, however, converge to a local optimum. We note that ACID bears similarity to DR [5]. Unlike DR however, ACID is deterministic, and minimises the entropy of the noise (residual).

The computational complexity of ACID is $\mathcal{O}(|\mathcal{Y}|^{|\mathcal{X}|})$. For early termination, we can set the maximum number of iterations J. In our experiments, we use $J = 10\,000$, of which ACID requires only a handful, finishing within seconds for pairs with reasonable domain size (roughly up till a hundred).

Algorithm 1: ACID

```
   Input: Two discrete i.i.d. sequences $X$ and $Y$, and the
          maximum number of iterations $J$
   Output: $H_{X \rightarrow Y}$
 1 $\mathcal{X}, \mathcal{Y} \leftarrow$ DOMAIN(X), DOMAIN(Y);
 2 for $x_i \leftarrow \mathcal{X}$ do
 3     $f_0(x_i) \leftarrow \arg\max_{y \in \mathcal{Y}} P(X = x_i, Y = y)$;
 4 $r \leftarrow H(Y - f_0(X))$;
 5 $j \leftarrow 0$;
 6 do
 7     $j \leftarrow j + 1$;
 8     $c \leftarrow false$;
 9     for $x_i \leftarrow$ CANONICALORDER($\mathcal{X}$) do
10         $t \leftarrow \min_{y \in \mathcal{Y}} H(Y - f_{j-1}^{x_i \rightarrow y}(X))$;
11         if $t < r$ then
12             $r \leftarrow t$;
13             $c \leftarrow true$;
14             $f_j(x_i) \leftarrow \arg\min_{y \in \mathcal{Y}} H(Y - f_{j-1}^{x_i \rightarrow y}(X))$;
15 while $j < J$ or $c = true$;
16 $H_{X \rightarrow Y} \leftarrow H(X) + r$;
17 return $H_{X \rightarrow Y}$
```

VI. Related Work

Existing methods for causal inference on a pair of discrete variables are roughly based on the following two frameworks:

Structural Causal Models The structural causal models express causal relationship in terms of a function of observed and unobserved variables. The ANMs assume that the unobserved variable (noise) is additive. Peters et al. [5] extend ANMs to discrete data, and propose the DR algorithm. DR uses chi-squared test of independence, which is more expensive to compute than Shannon entropy. Further ACID does not require p-value testing in every iteration. Moreover ACID is deterministic, whereas DR is non-deterministic.

Kocaoglu et al. [10] recently proposed a causal inference framework (ECI) for two discrete variables by postulating that the unobserved variable is simpler—in terms of the Rényi entropy—in the true direction. In particular, it is conjectured that if X causes Y, $H_\alpha(X)+H_\alpha(E) < H_\alpha(Y)+H_\alpha(\tilde{E})$ with H_α being the Rényi entropy, where $Y = f(X, E), X \perp\!\!\!\perp E$ and $X = f(Y, \tilde{E}), X \perp\!\!\!\perp \tilde{E}$. Unlike ANMs which assume the noise to be of additive type, the unobserved variable can be of arbitrary type in ECI.

Algorithmic Independence The algorithmic independence of Markov kernels postulates that if X causes Y, $P(X)$ and $P(Y \mid X)$ are *algorithmically* independent [11], [12]. As Kolmogorov complexity is not computable, causal inference methods based on algorithmic independence have to define a computable dependence measure.

CISC [13] employs *refined* MDL (an approximation from above to Kolmogorov complexity w.r.t. a model class) for causal inference from discrete data. Identifiability is a crucial aspect of causal inference as it distinguishes probabilistic conditioning $P(Y \mid X = x)$ from causal conditioning $P(Y|do(X = x))$. By the identifiability of ANM on discrete data, ACID is identifiable, whereas CISC is *not*.

Liu & Chan [14] (DC) propose to use distance correlation as a dependence measure. To infer the causal direction, DC computes the distance correlation between empirical marginal and conditional distributions in two directions. On account of the performance of DC against the state-of-the-art [13], we do not consider it for comparison.

VII. Experiments

We implemented ACID 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 MacBook Pro with 2.5 GHz Intel Core i7 processor and 16 GB memory. Unless specified otherwise, we use $\tau = 0$ for ECI, CISC and ACID. For DR, we use the level of significance of $\alpha = 0.05$ (also used by [5]) after examining its accuracy on 1000 cause-effect pairs, with 1000 outcomes each, sampled from a parameterised family of geometric distributions (explained in detail next) for various α values ranging from 0.05 to 0.001.

Synthetic Data We generate synthetic data with the ground truth X causes Y using ANM. Following the scheme of [5], we sample X from following model classes:

- uniform from $\{1, 2, \ldots, L\}$,
- binomial with parameters (n, p),
- geometric with parameter p,
- hypergeometric with parameters (M, K, N),
- poisson with parameter λ,
- negative binomial with parameter (n, p), and
- multinomial with parameter $\boldsymbol{\theta}$.

We randomly choose the parameters for each model class. In particular, we choose L uniformly between 2 and 10, p uniformly between 0.1 and 0.9, n, M and K uniformly between 1 and 40, N uniformly between 1 and $\min(40, M + K - 1)$, λ uniformly between 1 and 10, and $\boldsymbol{\theta}$ randomly s.t. $\sum_{\theta \in \boldsymbol{\theta}} \theta = 1.0$. We choose $f(x)$ uniformly between -7 and +7 for every x, and noise N uniformly, independent of X, between $-t$ and $+t$, where t is uniformly chosen between 1 and 7. To ensure identifiability, we use rejection sampling such that every cause-effect pair has a non-constant function f, and non-overlapping noise, i.e. $f(x) + \mathcal{N}$ are disjoint for $x \in \mathcal{X}$ [5], where $\mathcal{N}$ is the domain of the noise.

Accuracy We sample 1000 different models from each model class. For each model, we sample 1000 outcomes. In Fig. 1, for different model classes we compare the accuracy (percentage of correctly inferred cause-effect pairs) between ACID, DR, and CISC. The results show that unlike the other methods, ACID obtains 99 to 100% accuracy in all cases.

Whereas CISC performs consistently well across all model classes, it performs very poorly in the negative binomial model class. Note that CISC defines conditional stochastic complexity of Y given X as the *expected* stochastic complexity of Y conditional on the value of X. As a result, $\mathcal{S}(Y \mid X) \ll \mathcal{S}(Y)$, and $\mathcal{S}(X \mid Y) \ll \mathcal{S}(X)$. The inference process of CISC is hence largely dominated by the marginal stochastic complexities $\mathcal{S}(X)$, and $\mathcal{S}(Y)$, which in turn are dependent on the domain sizes $\mathcal{X}$ and $\mathcal{Y}$ respectively. Therefore CISC favours causal direction from the variable with smaller domain size towards the variable with larger domain size. The quintessential example of this bias is seen in the negative binomial model class. For this model class, unlike for the others, the domain size ($|\mathcal{X}|$) of the cause (X) is typically larger than the domain size ($|\mathcal{Y}|$) of the effect (Y), and hence CISC performs very poorly. Instead, by optimising the mapping between cause and effect, and evaluating the dependence measure over residual noise, both DR and ACID are able to avoid this bias.

The performance of ECI can be attributed to the difference in our data generating model, and the modelling assumption of ECI. Whereas ANMs assume that the noise is additive in nature, ECI assumes that the noise can be of arbitrary type.

Sample Size Next we study the effect of sample size on inference accuracy. For a fixed sample size, we compute the accuracy over 1000 models sampled from the geometric model

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

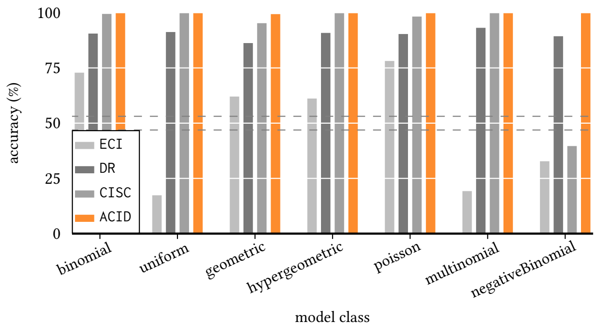

Figure 1: Accuracy on synthetic cause-effect pairs. The accuracy is reported over 1000 models from each model class. For each model, we sample 1000 data points. The dashed gray lines indicate the 95% confidence interval for a random coin flip in 1000 cause-effect pairs.

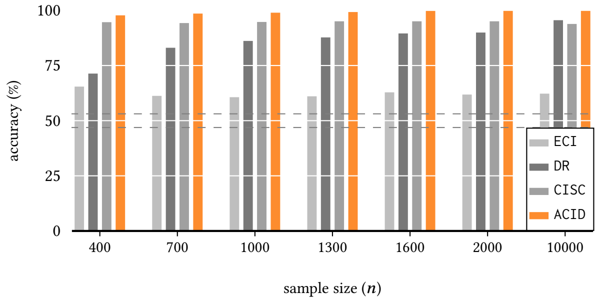

Figure 2: Effect of sample size on the accuracy on synthetic cause-effect pairs. For a fixed sample size, we sample 1000 models from the geometric model class. The dashed gray lines indicate the 95% confidence interval for a random coin flip in 1000 cause-effect pairs.

class. In Fig. 2, we compare for various sample sizes the accuracy of ACID against that of ECI, DR and CISC. We observe that ACID achieves 98% to 100% accuracy in all cases. DR performs poorly on small sample sizes, with its performance gradually improving for larger samples. CISC performs consistently around 94% accuracy, whereas ECI performs only slightly better than a random coin flip.

Significance Test The problem of differentiating between $X \to Y$ and $Y \to X$ can be cast as an *identity testing* problem. As our method is based on compression, we can use the compression based identity testing framework proposed by Ryabko & Astola [15] to assess the significance of inferred results. The framework can be roughly described as follows:

Definition 2 (Compression-based Identity Testing [15])**.** *Let x^n be a sequence over an alphabet Σ. Let $\mathcal{H}_0$ be the null hypothesis that the source of x^n has a distribution P, and h_1 be the alternative hypothesis that the source of x^n has a distribution Q. We reject the null hypothesis $\mathcal{H}_0$ if*

$$-\log P(x^n) - \{-\log Q(x^n)\} > -\log \alpha \ ,$$

where α is the level of significance.

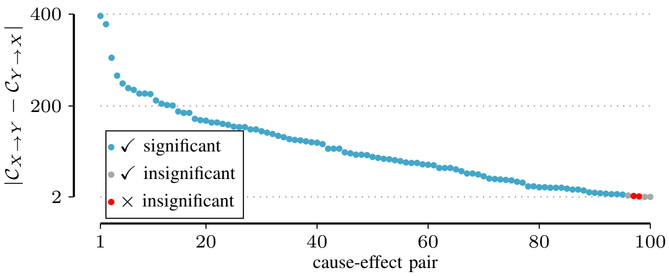

Figure 3: Synthetic cause-effect pairs sampled from a parameterized family of geometric distributions sorted by their corresponding difference in compression in two directions. We apply Benjamini-Hochberg correction to control the false discovery rate at a significance level of $\alpha = 0.01$.

The test statistic of the framework is given by $\delta = -\log P(x^n) + \log Q(x^n)$. The p-value of the test statistic is $2^{-\delta}$ due to the no-hypercompression-inequality [16, Chap 3.3] which gives an upper bound on the probability of an arbitrary distribution Q compressing the data better by δ bits than distribution P on the data.

From a cause-effect pair (X, Y) where ACID makes a decision (say $X \to Y$), we want to assess whether the decision is significant. To this end, our null hypothesis $\mathcal{H}_0$ will be the joint distribution under the alternative direction ($Y \to X$). Then the alternative hypothesis $\mathcal{H}_1$ will be that under the inferred direction. Since entropy gives the average number of bits per outcome in the sample, the compressed size of the sample using ANM from $X \to Y$ is given by $\mathcal{C}_{X\to Y} = nH_{X\to Y}$, and that from $Y \to X$ is $\mathcal{C}_{Y\to X} = nH_{Y\to X}$. Our test statistic would then be $\delta = \mathcal{C}_{Y\to X} - \mathcal{C}_{X\to Y}$. We reject $\mathcal{H}_0$ if $\delta > -\log \alpha$.

To control the false discovery rate for multiple hypothesis testing, we use the Benjamini-Hochberg procedure [17]. Let $\mathcal{H}_0^1, \mathcal{H}_0^2, \ldots, \mathcal{H}_0^m$ be the null hypotheses tested, and $p_1, p_2, \ldots, p_m$ their corresponding p-values. We sort the p-values in ascending order. For significance level of α, we find the largest k such that $p_k \leq \frac{k}{m}\alpha$. We reject the null hypothesis for all h_i, where $i = 1, \ldots, k$.

To evaluate, we sample 100 models from parameterized family of geometric distributions. For each model, we sample 350 outcomes. In Fig. 3, we sort the cause-effect pairs by their corresponding difference in compression in two directions (δ). That also corresponds to sorting the pairs by their p-values in ascending manner. At a significance threshold of $\alpha = 0.01$, after applying Benjamini-Hochberg correction, five inferences are insignificant, amongst which the two incorrect inferences. We observe similar behaviour with other model classes as well.

Real Data To investigate whether ACID discovers meaningful direction in real-world data, we consider three datasets.

Abalone This dataset is available from the UCI machine learning repository,[2] and contains physical measurements of 4 177 abalones (large, edible sea snails). We consider *sex* (X)

[2] http://archive.ics.uci.edu/ml/

Table I: Results on the *Abalone* dataset. For each pair with the ground truth, we report total compressed sizes using ACID in two directions, and the results of ACID, CISC, DR, and ECI respectively (✓ for correct, and $\approx$ for indecisive).

Truth	$\mathcal{C}_{X\to Y}$	$\mathcal{C}_{Y\to X}$	ACID	CISC	DR	ECI
sex → *length*	33713.74	34087.40	✓	✓	✓	✓
sex → *diameter*	32326.59	32886.51	✓	✓	✓	✓
sex → *height*	27046.16	27344.63	✓	✓	$\approx$	✓

of the abalone against length (Y_1), diameter (Y_2), and height (Y_3). The sex of the abalone is nominal (male, female, or infant), whereas length, diameter, and height are all measured in millimeters, and have 70, 57 and 28 unique values, respectively. Following [5], we treat the data as discrete. Since sex causes the size of the abalone and not the other way around, we regard $X \to Y_1$, $X \to Y_2$, and $X \to Y_3$ as the ground truth. We report the results in Table I. ACID infers correct direction in all three pairs with a large score difference between two directions. Both CISC and ECI also identify the correct directions in all three pairs. DR, on the other hand, remains indecisive in the third case.

Horse Colic This dataset is also available from the UCI machine learning repository, and contains the medical records of horses with 28 attributes, and 368 instances. Of particular interest to us are the two attributes: *abdomen status* (X) with 5 possible values, and *surgical lesion* (Y) with 2 possible values indicating whether the lesion (problem) was surgical. We remove the instances with missing values, ending up with a total of 225 instances. According to the domain experts, two abdomen statuses, namely *distended large intestine*, and *distended small intestine* indicate a surgical lesion. Therefore it is plausible to consider *abdomen status* as one of the causes of *surgical lesion*. Hence we regard $X \to Y$ as the ground truth. Both ACID and ECI recover the ground truth. Whereas DR remains indecisive, CISC infers the wrong direction with a very high confidence ($\delta = 85.73$ bits).

NLSchools This dataset is the 99-th pair in the Tübingen cause-effect benchmark pairs.[3] It contains the language test score (X), and socio-economic status of pupil's family (Y) of 2287 eighth-grade pupils from 132 classes in 131 schools in the Netherlands. The language test score has 47 unique values, and the socio-economic status of pupil's family has 21 unique values. We regard $Y \to X$ as the ground truth as the socio-economic status of pupil's family is one of the causes of the language test score. All methods recover the ground truth.

VIII. Conclusions

We proposed an information-theoretic framework for causal inference on discrete data using ANMs. The experiments show that the proposed algorithm, ACID, is highly accurate on synthetic data, obtaining at or near 100% accuracy for a wide range of source distributions and sample sizes, while qualitative case studies confirm that the results are sensible. ACID took few iterations to converge, and finished within seconds in our experiments. Moreover, the results of ACID can be assessed for statistical significance using the compression-based hypothesis testing framework.

The results suggest that Shannon entropy is a reasonably good choice as a dependence measure for causal inference using ANM from discrete data. First, marginal Shannon entropy is cheaper to compute. Further we do not have to explicitly test for the null hypothesis using p-values in every iteration unlike with the other statistical independence testing frameworks. If desired, one can always assess the significance of the final result using compression-based identity testing framework.

Acknowledgements

K. Budhathoki is supported by the International Max Planck Research School for Computer Science (IMPRS-CS). Both authors are supported by the Cluster of Excellence MMCI within the Excellence Initiative of the German Federal Government.

References

[1] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[2] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT press, 2000.

[3] 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.

[4] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf, "Nonlinear causal discovery with additive noise models," in *NIPS*, 2009, pp. 689–696.

[5] J. Peters, D. Janzing, and B. Schölkopf, "Identifying cause and effect on discrete data using additive noise models," in *AISTATS*, 2010, pp. 597–604.

[6] K. Budhathoki and J. Vreeken, "Accurate causal inference on discrete data," *CoRR*, vol. abs/1702.06776, 2018.

[7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference - Foundations and Learning Algorithms*, ser. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press, 2017.

[8] S. Kpotufe, E. Sgouritsa, D. Janzing, and B. Schölkopf, "Consistency of causal inference under the additive noise model," in *ICML*. JMLR, 2014, pp. 478–495.

[9] 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. 1, pp. 1103–1204, 2016.

[10] M. Kocaoglu, A. G. Dimakis, S. Vishwanath, and B. Hassibi, "Entropic causal inference," in *AAAI*, 2017, pp. 1156–1162.

[11] D. Janzing and B. Schölkopf, "Causal inference using the algorithmic markov condition," *IEEE TIT*, vol. 56, no. 10, pp. 5168–5194, 2010.

[12] J. Lemeire and E. Dirkx, "Causal models as minimal descriptions of multivariate systems. http://parallel.vub.ac.be/~jan," 2006.

[13] K. Budhathoki and J. Vreeken, "MDL for causal inference on discrete data," in *ICDM*, 2017, pp. 751–756.

[14] F. Liu and L. Chan, "Causal inference on discrete data via estimating distance correlations," *Neural Computation*, vol. 28, no. 5, pp. 801–814, 2016.

[15] B. Ryabko and J. Astola, "Application of data compression methods to hypothesis testing for ergodic and stationary processes," in *ICAA*, vol. AD. Discrete Mathematics and Theoretical Computer Science, 2005, pp. 399–408.

[16] P. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.

[17] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Jour. of the Royal Stat. Soc. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[3] https://webdav.tuebingen.mpg.de/cause-effect/