Causal Inference by Direction of Information

Jilles Vreeken°

Abstract

We focus on data-driven causal inference. In particular, we propose a new principle for causal inference based on algorithmic information theory, i.e. Kolmogorov complexity. In a nutshell, we determine how much information one data object gives about the other, and vice versa, and identify the most likely causal direction by the strongest *direction* of information.

To apply this principle in practice, we propose ERGO, an efficient instantiation for inferring the causal direction between multivariate real-valued data pairs. ERGO is based on cumulative and Shannon entropy. Therewith, we do not have to assume distributions, nor have to restrict the type of correlation. Extensive empirical evaluation on synthetic, benchmark, and real-world data shows that ERGO is robust against both noise and dimensionality, efficient, and outperforms the state of the art by a wide margin.

1 Introduction

Causal inference is concerned with identifying *causality*. Loosely speaking, the goal in causal inference is to determine from empirical data whether X causes Y, or the other way around, or, whether they are only correlated. Clearly, causal inference has a broad area of application – science is all about discovering causes and effects, after all. In biology and medicine, for example, key questions include *"do pills X cure disease Y"*, or, *"what are the genes X that cause phenotype Y"*. Being able to automatically determine cause and effect is hence one of the holy grails in data mining.

Toward this goal, we propose a causal inference rule based on Kolmogorov complexity, or, algorithmic information theory [8]. In a nutshell, we consider the amount of information an object $\mathbf{X}$ gives about object $\mathbf{Y}$ – and vice versa – and infer causality based on the strongest *direction of information* between these two. Our rule hence closely embraces the common postulate of causal inference: it is simpler to explain *effect* through *cause* than the other way around [11]. While this has been adopted by a number of recent proposals [5,6,18], our approach allows for causal inference regardless of correlation type, noise model, and without having to assume anything about the distribution of the data.

Kolmogorov complexity has very nice theoretical properties, but due to the halting problem it is sadly not computable [8]. To put our principle to practice, we hence introduce ERGO, an efficient instantiation for inferring the causal direction between pairs of real-valued data – multivariate or univariate – measuring the direction of information by a combination of cumulative and Shannon entropy.

Most research effort on causal analysis considers pairs of univariate variables, and is aimed at inferring whether X causes Y, or vice versa, from the joint observations of (X, Y) under the assumption that there are no hidden confounders [6, 19, 24]. Recently, there has been more focus on inferring the causal direction between multivariate random variables [1, 4, 26]. These methods only consider *linear* and *invertible functional* correlations. Further, they require both X and Y to be strictly multivariate. Last, but not least, non-linear functional correlation analysis techniques are computationally expensive, and hence not suited for large data. ERGO alleviates each and every of these points.

Extensive empirical evaluation on synthetic, benchmark, and real-world data shows that by considering the complexity of both the data and the model, ERGO outperforms the state of the art by a wide margin. It is highly resilient to noise and dimensionality, and can handle *both* univariate and multivariate variables, as well as non-deterministic, complex, and non-invertible correlations. Moreover, it is very efficient, permitting usage on large data.

In summary, our contributions include a new algorithmic information theoretic principle for causal inference between arbitrary objects, a practical instantiation based on cumulative and Shannon entropy, and a method for efficiently inferring the causal direction between two real-valued random variables X and Y without having to make any assumption on their relationship or distribution.

The paper is organised as usual. We introduce our theory for causal inference in Sec. 2, and its entropy-based instantiation in Sec. 3. The details for ERGO, its efficient implementation are in Sec. 4. Sec. 5 discusses related work. We empirically evaluate in Sec. 6, and round up with discussion and conclusions in Sec. 7 and 8. For conciseness and readability we postpone the proofs to Appendix A.

2 Causal Inference by Algorithmic Information Theory

Consider two *objects* $\mathbf{X}$ and $\mathbf{Y}$ that we know to be correlated – as identified by e.g. a domain expert or an appropriate test. Our goal is to infer the causal relationship between $\mathbf{X}$ and $\mathbf{Y}$. We assume there are no hidden confounders, there is no hidden $\mathbf{Z}$ causing both $\mathbf{X}$ and $\mathbf{Y}$. That is, we assume causal sufficiency. Hence, our decision comes down to determining which of $\mathbf{X} \to \mathbf{Y}$ and $\mathbf{Y} \to \mathbf{X}$ is most plausible [11].

°Max Planck Institute for Informatics and Saarland University, Saarbrücken, Germany. Email: jilles@mpi-inf.mpg.de

Loosely speaking, we will derive a rule for causal inference that calls $\mathbf{X}$ to be more likely to cause $\mathbf{Y}$ than vice versa, when the data $\mathbf{Y}$ is more easily described knowing $\mathbf{X}$ than vice versa – that is, when $\mathbf{X}$ provides relatively more information about $\mathbf{Y}$ than the other way around.

2.1 Kolmogorov Complexity We base our inference rule on algorithmic information theoretic principles, using Kolmogorov complexity as the main foundation. The key aspect of Kolmogorov complexity [8], as well as that of its practical implementations Minimum Description Length (MDL) [14] and Minimum Message Length (MML) [23], is perhaps best captured by the slogan *Induction by Compression*.

Given a finite binary string s, its Kolmogorov complexity $K(s)$ is defined as the length of the *shortest* program s^* for a universal Turing machine $\mathcal{U}$ that generates s and halts [8]. That is, $K(s) = l(s^*)$. Intuitively, s^* is the most succinct algorithmic description of s. To derive our inference rule, we will need conditional Kolmogorov complexity, $K(s \mid t)$, the length of the shortest program s^* that given the information in t 'for free' generates s, and then halts.

In fact, s^* can be split such that the true structure of the data is separated from meaningless noise [21]. The main idea is as follows. Let S be a set of strings containing s. Foundational to information theory is that given a set of objects, and without any further information, all elements are equally likely to be chosen. The most efficient way to identify s from S is hence by an index, i.e. $K(s \mid S) = \log |S| + O(1)$ bits. Let $K(S)$ be the length of the shortest program that generates S, and then halts. Now,

$$K(s) = K(S) + \log |S| + O(1) \quad ,$$

identifies the best model S for s as the most easily described set of strings S for which s is a typical element. This means that *all* structure in s that can algorithmically described succinctly is captured by $K(S)$ – including the *form* of the noise, e.g. Gaussian. Written more intuitively, with equality up to a constant, we have [21]

$$K(s) \stackrel{+}{=} K(s') + K(s \mid s') \quad , \tag{2.1}$$

where $K(s')$ is the length in bits of the structure in s – the part of s that can be described succinctly algorithmically – and $K(s \mid s')$ is the randomness of s – the length in bits we need to reach s given the modelled data s'. This two-part definition of Kolmogorov complexity is the foundation of two-part MDL [21]. Here we use if for causal inference.

2.2 Causal Inference by Direction of Information We will now develop our causal inference rule using Kolmogorov complexity. For readability, in the remainder we will refer to objects $\mathbf{X}$ and $\mathbf{Y}$ as the input data, rather than string s. Note that the complexity of $\mathbf{X}$ is equivalent to that of string s up to the constant cost of serialising $\mathbf{X}$ into s.

For completeness, let us define the two-part decomposition of $K(\mathbf{X})$ using Eq. 2.1 as

$$K(\mathbf{X}) \stackrel{+}{=} K(\mathbf{X}') + K(\mathbf{X} \mid \mathbf{X}') \tag{2.2}$$

such that $K(\mathbf{X}')$ is the cost for $\mathbf{X}'$, the compressible part of $\mathbf{X}$, and $K(\mathbf{X} \mid \mathbf{X}')$ the cost of the incompressible part of $\mathbf{X}$. Similarly, $\mathbf{Y}'$ is the compressible part of $\mathbf{Y}$.

A cornerstone postulate in causal inference states that, if X causes Y, describing $\mathbf{Y}$ using $\mathbf{X}$ will be simpler than the other way around [11]. This makes sense from an algorithmic information theoretic perspective. That is, if X causes Y, $\mathbf{X}$ will provide more information about $\mathbf{Y}$ than vice versa. To use this for causal inference we need to measure the amount of information that $\mathbf{X}$ provides towards most succinctly describing $\mathbf{Y}$, and vice versa. That is, we need to be able to infer the strongest *direction of information* between the two objects.

In terms of Kolmogorov complexity, when X causes Y we expect $K(\mathbf{Y} \mid \mathbf{X}) < K(\mathbf{X} \mid \mathbf{Y})$ as intuitively it will require a much simpler algorithm to generate $\mathbf{Y}$ knowing $\mathbf{X}$, than vice versa. This, however, assumes that the process generating $\mathbf{X}$ directly caused $\mathbf{Y}$, and that no noise was added to either after this causal influence. Clearly, it is more general to assume that the process generating $\mathbf{X}'$ causes $\mathbf{Y}'$ and that we only observe the noise-distorted objects $\mathbf{X}$ and $\mathbf{Y}$. For this subtly different, and more general process, we expect $K(\mathbf{Y} \mid \mathbf{X}') < K(\mathbf{X} \mid \mathbf{Y}')$ as then we measure the information provided by the models $\mathbf{X}'$ and $\mathbf{Y}'$, the best *generalisations* of $\mathbf{X}$ and $\mathbf{Y}$, and ignore noise (randomness) that may be present in the observed data.

When conditioning Eq. 2.2, we have

$$K(\mathbf{Y} \mid \mathbf{X}') \stackrel{+}{=} K(\mathbf{Y}' \mid \mathbf{X}') + K(\mathbf{Y} \mid \mathbf{X}', \mathbf{Y}') \quad , \tag{2.3}$$

where the first term is the complexity of $\mathbf{Y}'$, the optimal model for $\mathbf{Y}$, given the optimal model for $\mathbf{X}$. The second term measures the complexity of $\mathbf{Y}$ given both models.

In practice, $\mathbf{X}$ and $\mathbf{Y}$ may be of different complexities. Inferring causal direction based on the absolute difference between $K(\mathbf{Y} \mid \mathbf{X}')$ and $K(\mathbf{X} \mid \mathbf{Y}')$ would hence be biased towards the simplest object. To more reliably identify the direction of information we therefore normalise, and instead consider the difference in *relative* conditional complexity. We define the relative amount of directed information as

$$\Delta_{\mathbf{X} \to \mathbf{Y}} = \frac{K(\mathbf{Y}' \mid \mathbf{X}') + K(\mathbf{Y} \mid \mathbf{X}', \mathbf{Y}')}{K(\mathbf{Y})} \quad , \tag{2.4}$$

defining $\Delta_{\mathbf{Y} \to \mathbf{X}}$ analogously. Eq. 2.4 takes a value of 1 when $\mathbf{X}'$ does not provide any information about $\mathbf{Y}$ and (close to) 0 when $\mathbf{X}'$ identifies $\mathbf{Y}$ non-deterministically. If $\Delta_{\mathbf{X} \to \mathbf{Y}} < \Delta_{\mathbf{Y} \to \mathbf{X}}$, $\mathbf{X}'$ provides more information than $\mathbf{Y}'$, and by the direction of information we infer that it is more plausible that $\mathbf{X}$ caused $\mathbf{Y}$ than vice versa. Alternatively, when $\Delta_{\mathbf{Y} \to \mathbf{X}} < \Delta_{\mathbf{X} \to \mathbf{Y}}$, we infer $\mathbf{Y} \to \mathbf{X}$.

By using algorithmic information theory as our foundation we only need to consider data objects – there is not a distribution in sight. This also means we can determine the most likely causal direction between *arbitrary* objects and not restricted to series of observations. The most important observation to make, however, is that to make reliable inferences we have to take both the complexity of the model and that of the data under the model into account.

3 Causal Inference by Entropy over Data and Model

Kolmogorov complexity provides strong theoretical foundations, but, it is not computable and hence not practical. We can, however, approximate it from above by compression [8]. Here, we will do so by cumulative and Shannon entropy.

More in particular, instead of objects in general we consider real-valued random-variables X and Y with the goal to determine whether $X \to Y$, or $Y \to X$. To calculate the direction of information, we will approximate the complexity of the data we have over these variables, $\mathbf{X}$ and $\mathbf{Y}$, using cumulative entropy [3]. To calculate conditional entropy we have to perform estimation, i.e. we have to model. For calculating the complexity of these models, $\mathbf{X}'$ and $\mathbf{Y}'$, we will use Shannon entropy [2].

3.1 Notation Throughout the remainder, we consider a k-dimensional random variables $X = \{X_1, \ldots, X_k\}$ and l-dimensional random variables $Y = \{Y_1, \ldots, Y_l\}$ where each X_i and each Y_j are real-valued. We write $\mathbf{X}$ to denote the data we have for X. If we have collected n observations $\mathbf{X}$ represents the n-by-k data matrix of X. Similarly, we use $\mathbf{X}_i$ for the data over variable X_i, and analogue, slightly abusing notation, we say $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_k\}$.

We write $p(X)$ for the probability density function (pdf) of X. We write $p(x)$ as a short form for $p(X = x)$. We define $p(Y)$ similarly. We assume that the domain of $X_i \in X$ is $[min(X_i), max(X_i)]$. Considering a univariate random variable X, we write the cumulative distribution function (cdf) of X as $P(X)$, with $P(x)$ as the short form of $P(X \leq x)$. All logarithms are to base 2, and we adopt the usual convention of $0 \log 0 = 0$.

3.2 Cumulative Entropy Loosely speaking, cumulative entropy [3] captures the information content, i.e. complexity, of a probability distribution. However, different from Shannon entropy, it works with (conditional) cdfs, and can be regarded as a substitute for Shannon entropy for real-valued data. The cumulative entropy of a real-valued univariate random variable X, denoted as $h(X)$, is given as

$$h(X) = -\int P(x) \log P(x) dx \quad .$$

The conditional cumulative entropy of a real-valued univariate random variable X given $Z \in \mathbb{R}^s$ is defined as [3]

$$E_Z[h(X \mid Z)] = \int h(X \mid z) p(z) dz \quad . \tag{3.5}$$

It has two properties that are of particular importance to us.

THEOREM 3.1. *$h(X \mid Z) \geq 0$ with equality iff X is a function of Z. $h(X \mid Z) \leq h(X)$ with equality iff X is statistically independent of Z.*

Proof. We postpone the proof to Appendix A.

Even more importantly, unconditional cumulative entropy can be computed in closed-form for empirical data. Let $x_1 \leq \ldots \leq x_n$ be the ordered records of X. We have

$$h(X) = -\sum_{i=1}^{n-1} (x_{i+1} - x_i) \tfrac{i}{n} \log \tfrac{i}{n} \quad .$$

When Z is discrete valued, computing the conditional cumulative entropy is equally straightforward. It imply is the weighted sum of cumulative entropies over X for every $z \in Z$. Let $(x_1, z) \leq \ldots \leq (x_m, z)$ be the m records for which $Z = z$ ordered by value of X. We then have

$$h(X \mid Z) = \sum_{z \in Z} h(X \mid z) p(z) \quad .$$

When Z is continuous real-valued and the probability density function $p(Z)$ is unavailable, we need estimation. We will return to this in Section 4.

3.3 Entropy-based Direction of Information We will now proceed to define a practical version of Eq. 2.4 for real-valued data. First we propose how to approximate $K(\mathbf{Y} \mid \mathbf{X}', \mathbf{Y}')$ resp. $K(\mathbf{Y}' \mid \mathbf{X}')$, and then the discuss the normalisation term. Finally, we construct the practical inference rule and discuss its properties.

Complexity of the Data We will use cumulative entropy to approximate $K(\mathbf{Y} \mid \mathbf{X}', \mathbf{Y}')$, the conditional complexity of the data. Cumulative entropy, however, is only defined for single univariate random variables whereas Y may be multivariate. We therefore apply a chain rule. That is, we factorise $p(Y \mid X)$ into $p(Y_1 \mid X)$, $p(Y_2 \mid X, Y_1)$, $\ldots$, $p(Y_l \mid X, Y_1, \ldots, Y_{l-1})$. We can now quantify the complexity of $p(Y_i \mid \cdot)$ using $h(Y_i \mid \cdot)$.

To compute $h(Y_i \mid \cdot)$ we require the pdf over the conditioning terms, e.g. $p(Y)$ and $p(X)$. These are unavailable in practice and will need to be estimated. Postponing the details to Sec. 4, we will do so using density estimation. This provides discrete models $\mathbf{X}'$ and $\mathbf{Y}'$. For readability we slightly abuse notation, and write $h(\mathbf{Y})$ and $h(\mathbf{Y} \mid \mathbf{X}')$ for resp. $h(Y)$ and $h(Y \mid X)$ over real-valued $\mathbf{Y}$ and discrete $\mathbf{X}'$.

There exist $l!$ factorisations of $p(Y \mid X)$, raising the question which one to use. To approximate $K(\cdot)$ we define $h(\mathbf{Y} \mid \cdot)$ as the *minimum* entropy over all factorisations. Let

σ_Y denote a permutation of the attributes of Y. We then have

$$h(\mathbf{Y} \mid \mathbf{X}') = \min_{\sigma_Y} h(\mathbf{Y}_{\sigma_Y(1)} \mid \mathbf{X}') + h(\mathbf{Y}_{\sigma_Y(2)} \mid \mathbf{X}', \mathbf{Y}'_{\sigma_Y(1)}) + \cdots + h(\mathbf{Y}_{\sigma_Y(l)} \mid \mathbf{X}', \mathbf{Y}'_{\sigma_Y(1)}, \ldots, \mathbf{Y}'_{\sigma_Y(l-1)}) \quad . \tag{3.6}$$

We postpone the details for how to compute, or rather, approximate the σ_Y^* that minimises Eq. 3.6 to Sec. 4. For normalisation, it is good to know we have an upper bound.

LEMMA 3.1. *It holds that* $h(\mathbf{Y} \mid \mathbf{X}') \leq h(\mathbf{Y})$.

Proof. We postpone the proof to Appendix A.

Complexity of the Model Approximating the conditional complexity of the model is more straightforward. Under the assumption that X causes Y, estimating $p(X)$ and $p(Y)$ yields an estimation $\widehat{p}_{X \to Y}(X, Y)$ of $p(X, Y)$. Let $\mathbf{X}'$ and $\mathbf{Y}'$ be the discrete models of data $\mathbf{X}$ and $\mathbf{Y}$ as induced by this estimation. Given that $\mathbf{X}'$ and $\mathbf{Y}'$ are discrete, their joint Shannon entropy $H(\mathbf{X}', \mathbf{Y}')$ corresponds to the complexity of $\widehat{p}_{X \to Y}(X, Y)$, which here therefore provides a natural approximation of $K(\mathbf{Y}' \mid \mathbf{X}')$.

Normalisation So far we have $h(\mathbf{Y} \mid \mathbf{X}') + H(\mathbf{X}', \mathbf{Y}')$ as a practical approximation of Eq. 2.3. It can be interpreted as the average number of bits needed to describe an observation under the assumption $X \to Y$. In Eq. 2.4, we can use $K(\mathbf{Y})$ to normalise, as it is a natural, single, and tight upper bound for $K(\mathbf{Y} \mid \mathbf{X}')$. In our practical setting we are not so lucky and will need to define upper bounds h^{u} and H^{u} ourselves. There are many options to this end. A straightforward solution follows from $h(\mathbf{Y} \mid \mathbf{Y}', \mathbf{X}') \leq h(\mathbf{Y} \mid \mathbf{Y}')$ and $H(\mathbf{X}', \mathbf{Y}') \leq H(\mathbf{X}') + H(\mathbf{Y}')$. With our estimation scheme, these do not work well for univariate X or Y. Instead, we therefore define H^{u} and h^{u} alternatively. First, we observe

$$h(\mathbf{Y} \mid \mathbf{X}') \leq h(\mathbf{Y}) \leq \sum_{i=1}^{l} h(\mathbf{Y}_i) \quad ,$$

where $h(\mathbf{Y}_i)$ can be calculated directly. This gives us a natural upper bound for $h(\mathbf{Y} \mid \mathbf{X}')$ in the form of

$$h^{\mathrm{u}}(\mathbf{Y}) = \sum_{i=1}^{l} h(\mathbf{Y}_i) \quad .$$

To obtain a practical upper bound for $H(\mathbf{X}', \mathbf{Y}')$, we first factorise it into the independence model, $H^{\mathrm{u}}(\mathbf{X}', \mathbf{Y}') = H^{\mathrm{u}}(\mathbf{X}') + H^{\mathrm{u}}(\mathbf{Y}')$. Analogue to above, we could use $H(\cdot)$ to instantiate the two terms. Preliminary experiments showed, however, that this does not work well in practice. Hence, instead we use that the uniform distribution has the largest Shannon entropy. That is, for discrete data $\mathbf{X}' = \{\mathbf{X}'_1, \ldots, \mathbf{X}'_k\}$, with $|\mathbf{X}'_i|$ the number of bins of $\mathbf{X}'_i$,

$$H(\mathbf{X}') \leq \sum_{i=1}^{l} \log(|\mathbf{X}'_i|) \quad .$$

We use this to define an upper bound, $H^{\mathrm{u}}(\mathbf{X}')$, for the complexity of the discrete (modelled) data $\mathbf{X}'$ as

$$H^{\mathrm{u}}(\mathbf{X}') = \sum_{i=1}^{l} \log(|\mathbf{X}_i|) \quad ,$$

and similar for $H^{\mathrm{u}}(\mathbf{Y}')$. If all components of $\mathbf{X}$ and $\mathbf{Y}$ have only one bin, by convention we say $\frac{0}{0} = 0$.

It is easy to see that h^{u} and H^{u} are not equally tight upper bounds. This would mean that when we normalise $h(\mathbf{Y} \mid \mathbf{X}') + H(\mathbf{X}', \mathbf{Y}')$ by $h^{\mathrm{u}}(\mathbf{Y} \mid \mathbf{X}') + H^{\mathrm{u}}(\mathbf{X}', \mathbf{Y}')$ we introduce a bias that depends on this relative tightness. To avoid this, we will normalise both terms separately.

Determining the Direction of Information With the above, we can now define our practical entropy-based approximation of Eq. 2.4. We define the relative amount of directed information of real-valued $X \to Y$ as

$$\Delta_{X \to Y} = \frac{1}{2} \left(\frac{h(\mathbf{Y} \mid \mathbf{X}', \mathbf{Y}')}{h^{\mathrm{u}}(\mathbf{Y} \mid \mathbf{X}', \mathbf{Y}')} + \frac{H(\mathbf{X}', \mathbf{Y}')}{H^{\mathrm{u}}(\mathbf{X}', \mathbf{Y}')} \right) \tag{3.7}$$

and analogue for $\Delta_{Y \to X}$. As above, if $\Delta_{X \to Y} < \Delta_{Y \to X}$, we infer by direction of information that $X \to Y$ is more plausible. If $\Delta_{Y \to X} < \Delta_{X \to Y}$, we infer that $Y \to X$.

It is important to note that we do not make any assumption on the distribution of, or the type of correlation between, X and Y, and neither on the presence, or form, of noise.

A natural interpretation of $\Delta_{X \to Y}$ is that it measures the divergence between causal determinacy $X \to Y$ and independence $X \perp\!\!\!\perp Y$. That is, analogue to Eq. 2.4, the lower the value of Eq. 3.7, the stronger the causal relationship, and the closer to its maximum, 1, the weaker.

4 ERGO – Causal Inference by Direction of Information

Next we give the implementation details of ERGO,[1] for efficiently calculating the relative amount of directed information. In particular, we explain our design choices on how to approximate the minimal entropy factorisation, how to estimate conditional cumulative entropy, and how to increase scalability. Finally, we discuss its time complexity.

4.1 Minimal Entropy Factorisations

As detailed above, for computing $h(\mathbf{Y} \mid \mathbf{X}')$ we need the permutation σ_Y of Y that minimises the conditional cumulative entropy. First, however, we need $\mathbf{X}'$ – the modelled, discretised version of

[1] ERGO is Latin for therefore, consequently.

$\mathbf{X}$ such that we can calculate the conditional entropy of $\mathbf{Y}$ per bin of $\mathbf{X}'$. Computing $\mathbf{X}'$ happens to be equivalent to computing $h(\mathbf{X})$, i.e., searching for the permutation σ_X of X with minimal cost.

To identify the minimal entropy factorisation of X, we would have to exhaustively consider all $k!$ permutations. This will be infeasible for high-dimensional data. We therefore propose a greedy solution that approximates σ_X^*. Let $\sigma_X(i)$ denote the i^{th} element of permutation σ_X of X. For readability, wherever clear from context we do not identify the attribute and simply write $\sigma(i)$. For σ_X, we first select $\sigma(1)$ such that $\mathbf{X}_{\sigma(1)}$ has minimal cumulative entropy, i.e. we have $\sigma(1) = \arg\min_i h(\mathbf{X}_i)$. We then choose $\sigma(2)$ such that $h(\mathbf{X}_{\sigma(2)} \mid \mathbf{X}'_{\sigma(1)})$ is minimal, and proceed until every dimension of $\mathbf{X}$ has been selected. We consider the permutation σ_X where dimensions are picked to be the approximate optimal permutation of X.

We compute $h(\mathbf{Y} \mid \mathbf{X}')$ analogue. That is, we choose σ_Y such that $h(\mathbf{Y}_{\sigma(1)} \mid \mathbf{X}')$ is minimal, then choose $\sigma(2)$ such that $h(\mathbf{Y}_{\sigma(2)} \mid \mathbf{X}', \mathbf{Y}'_{\sigma(1)})$ is minimal, and again proceed until every dimension of Y have been considered. We denote this permutation by σ_Y and consider it the approximate minimum entropy permutation of $\mathbf{Y}$.

Note that $\mathbf{Y}_{\sigma(l)}$, the last chosen dimension of Y, does not have to be discretised. Its model complexity is therefore minimal, i.e. $H(\mathbf{Y}_{\sigma(l)}) = 0$, by which we have $H(\mathbf{X}', \mathbf{Y}') = H(\mathbf{X}', \mathbf{Y}' \setminus \{\mathbf{Y}'_{\sigma(l)}\})$.

4.2 Estimating Conditional Cumulative Entropy We combine the estimation of conditional cumulative entropy with our algorithm for selecting the permutations of X and Y. For illustration purposes, we first consider $h(\mathbf{X})$. That is, after selecting $\mathbf{X}_{\sigma(1)}$, we calculate $h(\mathbf{X}_i \mid \mathbf{X}'_{\sigma(1)})$ for every dimension $\mathbf{X}_i$ not yet discretised such that $h(\mathbf{X}_i \mid \mathbf{X}'_{\sigma(1)})$ is minimal; we select dimension $\mathbf{X}_i$ with minimum entropy.

At every subsequent step, we only discretise the dimension picked in the previous step. That is, we do not re-discretise any earlier chosen dimensions. First and foremost, this increases the efficiency of the algorithm. Second, and more importantly, it allows us to measure the model complexity in a straightforward manner – we only have to consider one discretisation per dimension.

Next, we show that the discretisation at a step can be done efficiently and optimally by dynamic programming. For exposition, let us consider X. Let $X' \subset X$ be the set of dimensions from X already picked *and* discretised. We denote X_p as the dimension picked in the previous step but not yet discretised. Consider $X_c \in X \setminus (X' \cup \{X_p\})$ as a candidate dimension to be picked next, i.e. for which we will have to discretise X_p into X'_p such that $h(X_c \mid X' \cup \{X'_p\})$ is minimal. Further, let $x_1 \leq \ldots \leq x_n$ be realisations of X_p. We write $x_{j,u}$ for $\{x_j, x_{j+1}, \ldots, x_u\}$ where $j \leq u$. Slightly abusing notation, whenever we consider all data points, i.e. $x_{1,n}$ we simply write X_p. We use $h(X_c \mid X', \langle x_{j,u} \rangle)$ to denote $h(X_c \mid X')$ computed using the $(u - j + 1)$ points of X corresponding to x_j to x_u, projected onto X_p. For $1 \leq l \leq u \leq n$, we write

$$f(u, l) = \min_{g:|g|=l} h(X_c \mid X', x^g_{1,u})$$

where g is a discretisation of $x_{1,u}$, $|g|$ is its number of bins, and $x^g_{1,u}$ is the discretised version of $x_{1,u}$ by g. For $1 < l \leq u \leq n$, we have

THEOREM 4.1. $f(u, l) = \min_{j \in [l-1, u)} \mathcal{A}_j$ *where* $\mathcal{A}_j = \frac{j}{u} f(j, l-1) + \frac{u-j}{u} h(X_c \mid X', \langle x_{j+1,u} \rangle)$.

Proof. We postpone the proof to Appendix A.

Theorem 4.1 shows that the optimal discretisation of $x_{1,u}$ can be derived from that of $x_{1,j}$ with $j < u$. This allows us to find the discretisation of X_p that minimises $h(X_c \mid X', X_p)$ by dynamic programming. We note that we have to impose a maximum number of bins on all discretisation g considered. This is because in the extreme case when all realisations of X_p are distinct and $|g| = n$, $h(X_c \mid X', X_p)$ will be zero. Therefore, following [13], we impose the restriction that $|g| < n^\epsilon$ where $\epsilon \in (0, 1)$.

The computation of $h(\mathbf{Y} \mid \mathbf{X}')$ is done analogously. A small difference is that when searching for $\mathbf{Y}_{\sigma(1)}$ we concurrently seek the discretisation of $\mathbf{X}'_{\sigma(k)}$. Note that, as mentioned above, after processing all dimensions of $\mathbf{Y}$, all but $\mathbf{Y}_{\sigma(l)}$ are discretised.

Alternatively, kernel methods can be used to estimate densities [15]. Our strategy to approximate minimal conditional entropy automatically provides a good dimension permutation – plus, we do not have to choose a kernel.

4.3 Increasing Scalability To identify the optimal discretisation of a dimension using dynamic programming and the data points as cut points would result in a time-complexity of $O(n^3)$. Most cut-points, however will not be used in the optimal discretisation. To gain efficiency, we can hence impose a maximum grid size $max\_grid = n^\epsilon$ and limit the number of cut points to $c \times max\_grid$ with $c > 1$. To find these candidate cut points, we follow Reshef et al. [13] and apply equal-frequency binning per dimension with the number of bins equal to $(c \times max\_grid + 1)$.

4.4 Complexity Analysis The larger we choose ϵ and c, the more candidate discretisations we consider, and hence at the expense of additional computation the better its result. Preliminary empirical analysis shows that $\epsilon = 0.3$ and $c = 10$ offers a good balance between quality and efficiency.

This makes the cost of discretising a single dimension $O(n)$. Therewith, the overall complexity of computing $\Delta_{X \to Y}$ and $\Delta_{Y \to X}$ is $O((k^2 + l^2) \cdot n)$.

5 Related Work

Traditional causal inference methods [11] rely on conditional independence tests and hence require at least three observed random variables; they are not designed to infer the causal direction for just two observed random variables.

The existing algorithmic information-theoretic approach to causal inference [6, 7] postulates that $X \to Y$ is only acceptable if $p(X)$ and $p(Y \mid X)$ are algorithmically independent, i.e. when the shortest description of $p(X, Y)$ is given by separate descriptions of $p(X)$ and $p(Y \mid X)$. Kolmogorov complexity, however, is defined over *data*, not over distributions [8]. One can argue that $K(p(X))$ corresponds to model complexity $K(\mathbf{X}')$, and that by considering only model complexity, the problem of information symmetry, $K(\mathbf{X}) + K(\mathbf{Y} \mid \mathbf{X}) = K(\mathbf{Y}) + K(\mathbf{X} \mid \mathbf{Y})$ [8], is avoided. This assumes, however, that the relative complexity of the data under the model is neglible. That is, access to the true distributions. When the amount of empirical data approaches infinity, $p(\cdot)$ can be estimated with arbitrarily high accuracy [17]. That is, there will be no difference in the complexities of the estimated joint probabilities for $X \to Y$ and $Y \to X$. In practice, however, sample sizes are finite, and generic (conditional) pdfs are hard to estimate. Consequently, the complexity of the estimated joint probability $\widehat{p}_{X \to Y}(X, Y)$ will often be non-negligible, and we need to take this into account for reliable inference of causal direction. Our inference principle does take this complexity explicitly into account as we consider the complexities of both the model and of the data under the model.

As Kolmogorov complexity is not computable, frameworks as these require practical implementations. Janzing et al. [5] consider information-geometry to detect (in)dependencies between $p(X)$ and $p(Y \mid X)$ for deterministically related univariate X and Y. Earlier [4, 26], they considered covariance matrices to detect linear relations $Y = \mathbf{A} \times X + \mathbf{E}$ between multivariate variables. Chen et al. [1] allow non-linear correlations, yet require correlations to be deterministic, functional, and invertible.

Causal inference based on the additive noise model (ANM) [12,16] postulates that if $Y = f(X) + E$ with X the cause, Y the effect, and E an additive error term statistically independent of X, that there typically does not exist an additive noise model for the opposite direction. [24] generalised this to a post-nonlinear model. The intuition of both can be justified by the above algorithmic information-theoretic approach, in particular the algorithmic independence postulate.

Cumulative entropy was proposed in [3]. Earlier, we used it for non-parametric (non-)linear correlation analysis [9, 10]. To the best of our knowledge, we are the first to use it for causal inference.

It is important to point out that unlike any of the above, our framework only considers data, and does not require assumptions on either distribution, noise, or correlation.

Method	*Univariate*	*Multivariate*
ERGO	✓	✓
GPI [18]	✓	−
IGCI [5]	✓	−
LTR [4]	−	✓
KTR [1]	−	✓

Table 1: Characteristics of casual inference methods. (✓) means it can consider data of that type, (−) means it cannot.

6 Experiments

Next, we empirically evaluate ERGO with respect to inferring correct causal directions. We will compare the performance of ERGO to LTR [4], KTR [1], GPI [18], and IGCI [5]. LTR and KTR are state of the art for causal inference for multivariate pairs, while GPI and IGCI are state of the art for univariate pairs. Table 1 summarises their characteristics. We implemented ERGO in Java. We use $\epsilon = 0.3$ and $c = 10$ for all experiments. All experiments were conducted on an Intel i5-2500K Windows machine with 16GB RAM.

6.1 Causal Inference for Univariate Pairs We first evaluate ERGO on a benchmark set of cause-effect pairs with known ground truth [25]. We compare to GPI [18] and IGCI [5], two state of the art methods for univariate pairs. We consider 75 pairs from various domains. All of them are noisy, i.e. the relationship between each pair is non-deterministic. We find that ERGO infers the correct causal direction with an accuracy of 74.7%, outperforming both IGCI (69.3%) and GPI (61.3%) with a margin. It is reassuring to note that for correctly inferred pairs the difference between $\Delta_{X \to Y}$ and $\Delta_{Y \to X}$ differ more (0.1 on average) than when ERGO draws the wrong conclusion (0.05).

6.2 Causal Inference for Multivariate Pairs Second, we evaluate ERGO on real-world benchmark data where X and/or Y are multivariate. We consider nine non-deterministic data pairs for which the causal direction is known. We give the base statistics in Table 2. The first five are drawn from [25], the others from [4]. In the interest of space we refer to the original papers for their descriptions.

We compare against LTR [4] and KTR [1], two causal inference methods for the multivariate setting. LTR assumes the correlation between X and Y to be linear. KTR relaxes this requirement but explicitly requires the relationship to be deterministic (no noise), functional, and invertible.

We summarise the results in Table 2. As LTR and KTR require both X and Y to be multivariate they are hence inapplicable on the ozone concentration problem. Inspecting the results, we find that LTR and KTR obtain an accuracy of 50%. In comparison, ERGO is accurate in 78% of the cases.

Data	n	k	l	ERGO	LTR	KTR
Symptoms	120	6	2	✓	✓	✓
Climate forecast	10 266	4	4	✓	✓	–
Radiation	72	16	16	✓	–	✓
Ozone	989	1	3	✓	(n/a)	(n/a)
Car efficiency	392	3	2	–	–	✓
Precipitation	4 748	3	12	✓	✓	–
7 Stock indices	2 394	4	3	✓	–	✓
9 Stock indices	2 394	6	3	✓	–	–
Pollution	1 440	3	6	–	✓	–

Table 2: **ERGO is accurate** Results on multivariate cause-effect pairs with known ground truth. (✓) means the correct causal direction is inferred, and (−) otherwise. (n/a) means the respective method is inapplicable to the given pair.

6.3 Robustness and Scalability Finally, we evaluate ERGO with regard to robustness and scalability. To this end we consider synthetic data with known complex non-deterministic causal relations as ground truth. To this end, we first generate multivariate $X_{k\times 1} = \mathbf{A}_{k\times k} \times Z_{k\times 1}$ where $z_i \sim \textit{Gaussian}(0,1)$ and $a_{ij} \sim \textit{Uniform}[0,1]$. Then, using a function f that describes the relation between X and Y we generate $Y_{n\times 1}$ with $y_i = f(u_i) + e_i$ where $\mathbf{U}_{l\times 1} = \mathbf{B}_{l\times k} \times X_{k\times 1}$ with $b_{ij} \sim \textit{Uniform}[0, 0.5]$, and $e_i \sim \textit{Gaussian}(0,\sigma)$ with σ a free parameter. Through σ we can control the level of noise. For $\sigma = 0$ the relationship is deterministic, while larger values correspond to more noise. We use three non-linear, complex, and non-invertible instantiations of f, i.e.

$$
\begin{aligned}
f_1(x) &= \tanh(2x) + \tanh(3x+1) + \tanh(4x+2) \\
f_2(x) &= \sin(2x) + \sin(3x+1) \\
f_3(x) &= \sin(2x) + \sin(3x+1) + \\
&\quad \tfrac{1}{3}\big(\tanh(2x) + \tanh(3x+1) + \tanh(4x+2)\big)
\end{aligned}
$$

As competitors we again consider LTR [4] and KTR [1]. Per experiment, we generate 100 data sets per function, and for every data set we infer the causal direction per method.

Robustness to Complexity We first evaluate robustness against functional complexity and data dimensionality. We set $k = l$ and vary it between 5 to 120. We fix $n = 1\,000$ and use $\sigma = 0.5$. We show the average accuracy, the relative number of correct inferences, in Figure 1. We see that ERGO performs very well, obtaining 100% accuracy for every setting. In comparison, LTR and KTR show almost as good scores for f_1. For f_2, however, their performance decays to approx. 60%, while for the most complex function (f_3) these methods most often indicate the wrong direction.

Robustness to Noise Next, we evaluate robustness against noise. Following Janzing et al. [4], we vary σ from 0 to 2. We fix $n = 1\,000$ and $k = l = 5$. For brevity we only discuss the results on f_1. We show the results in Figure 2(a). We see that for (near) deterministic relations ($\sigma = 0$ and 0.5) all three methods make perfect inferences. For the higher levels of noise we see that ERGO is clearly the most robust, outperforming both LTR and KTR at a fair margin.

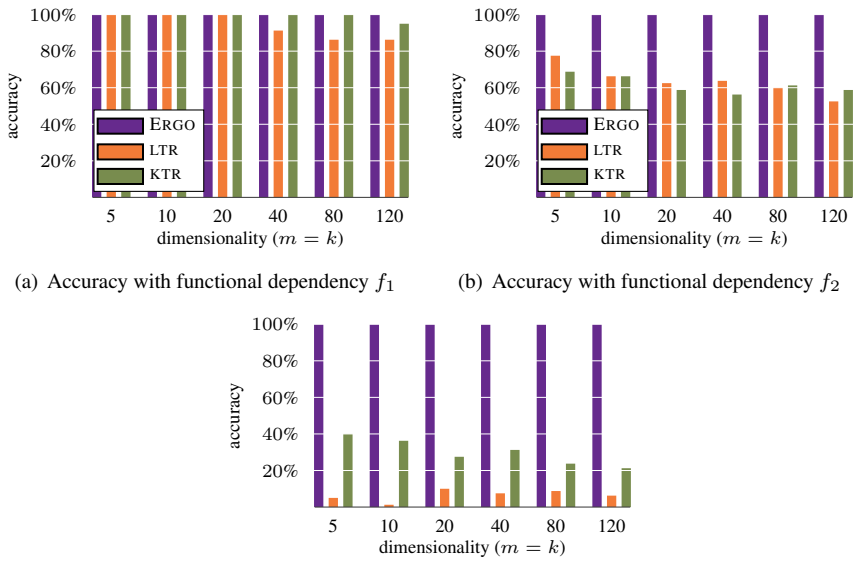
Scalability As last experiment on this data we investigate scalability. We consider two scenarios. First, we set $\sigma = 0.5$ and $k = l = 5$, varying n from $1\,000$ to $15\,000$. Second, we keep $n = 1\,000$ and $\sigma = 0.5$, while varying $k = l$ from 5 to 120. We give the results in Fig. 2(b) resp. Fig. 2(c). We find that ERGO scales linearly to data size and quadratically to dimensionality. This agrees with our analysis in Section 4.4. We observe that ERGO scales better than KTR but worse than LTR. Taking into account performance, we find that ERGO yields a good balance between quality and efficiency. As it scales linearly to data size it is applicable to large data sets.

6.4 Causal Discovery in Real-World Data Last, we consider the discovery of causal relations in non-benchmark data. To this end, we consider a noisy real-world data set on insurances [20]. It consists of 9000 user profiles over 86 dimensions, viz. income, education, social class, number and average price per insurance policy type, etc. The data is known to be rather noisy, and hence difficult to analyse [20].

To mine correlated dimensions we use MAC [9], a non-parametric method for discovering (non-)linear correlations. Next, we apply ERGO and IGCI [5] on the 120 discovered pairs to determine their most likely causal direction.

Inspecting the results, we find that both methods identify sensible causal relations. Overall, the inferences by ERGO correspond to intuition more often – probably as, unlike IGCI assumes, the relations between X and Y are not deterministic. Examples of causal directions correctly identified by ERGO, but not by IGCI, include the following

- *# of Roman Catholic family members* → *# of married couples in the family* as $\Delta_{X\to Y} = \mathbf{0.82} < \Delta_{Y\to X} = 0.89$
- *# of family members with high education* → *# of family members with high status* as $\Delta_{X\to Y} = \mathbf{0.85} < \Delta_{Y\to X} = 0.88$
- *average income of the whole family* → *# of home owners* as $\Delta_{X\to Y} = \mathbf{0.87} < \Delta_{Y\to X} = 0.92$
- *# of family members with low education* → *# of unskilled workers in the family* as $\Delta_{X\to Y} = \mathbf{0.88} < \Delta_{Y\to X} = 0.92$



(a) Accuracy with functional dependency f_1

(b) Accuracy with functional dependency f_2

(c) Accuracy with functional dependency f_3

Figure 1: **ERGO is accurate** Accuracy vs. number of dimensions (varying $k = l$), resp. for non-linear, complex and non-invertible functional dependencies f_1, f_2, and f_3. (See Section 6.3 for definition of these dependencies).

- *# of unskilled workers in the family* $\rightarrow$ *# of members with low income in the family* as $\Delta_{X \to Y} = \mathbf{0.87} < \Delta_{Y \to X} = 0.93$

Although far from a comprehensive analysis, these results show that ERGO can indeed be used to discover meaningful causal relations from noisy real-world data. Developing algorithms that can efficiently discover the partitioning of a correlated multivariate Z into X and Y such that $\Delta_{X \to Y}$ is minimal will make for engaging future work.

7 Discussion

The experiments clearly show that ERGO performs well in practice. It yields high accuracies for both univariate and multivariate data, is robust against dimensionality, and performance is particularly promising with regard to noise. The results of ERGO corroborate that the relative complexity of the data under the model is indeed important to reliably determine the direction of information.

Despite these encouraging results, causal inference is not solved with ERGO. For example, we currently compute the complexity of pdfs by means of cumulative entropy. Our estimation scheme clearly allows room for improvement; we currently greedily construct a Markov-chain – allowing more degrees of freedom in choosing the 'parent' node will likely improve estimation. Fursther, instead of cumulative entropy, one could use Shannon entropy as long as there is a reliable estimation schemes with good justifications.

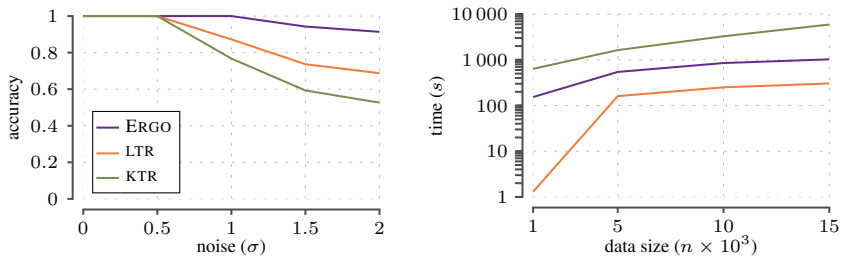
In addition, while in the Kolmogorov case $K(\mathbf{Y})$ provides the ideal single normalisation term, in ERGO we have to employ two tailored normalisation terms. More detailed characterisation of the complexity scores will likely identify better a normalisation scheme. Alternatively, one could perceive the normalisation as a weighting scheme and adjust the weights accordingly to fit the underlying application domain.

Further, given real-valued univariate random variables $X_1, \ldots, X_k$, we can use ERGO to efficiently derive their causal ordering, which in turn can be used to assess the plausibility of a given causal DAG. Moreover, applying ERGO in a framework for structure learning seems a particularly promising avenue of future work. That is, we can use ERGO to estimate the amount of directed information over an edge, and so iteratively construct a causal graph.

Whereas ERGO is currently restricted to real-valued data, our inference principle is defined over data in general. We see two main lines for future work in this regard. First of all, time series are a standard set-up for causal analysis. We are exploring to what extend our framework provides a theoretical foundation for Granger causality. Second, we aim to instantiate our framework for discrete categorical data. Recent results in pattern-based modelling provide promising results to this end [22].

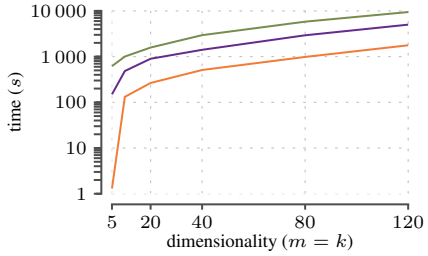
8 Conclusion

We proposed a new information theoretic principle for causal inference based on Kolmogorov complexity. In a nutshell, we measure the relative amount of information that one data object gives about the other, and vice versa, in order to determine the most likely causal direction by the strongest



(a) Accuracy vs. noise (varying σ) for function f_1 with $n = 1\,000$ and $m = k = 5$

(b) Runtime against n with $\sigma = 0.5$ and $k = l = 5$



(c) Runtime against $k = l$ with $n = 1\,000$ and $\sigma = 0.5$

Figure 2: **ERGO is noise-resistant and scalable** Accuracy vs. noise (by varying σ), runtime vs. data size, and runtime vs. dimensionality. (vertical axes in log-scale).

direction of information between the objects.

To apply this in practice we presented ERGO, an efficient instantiation for inferring the causal direction between pairs of univariate or multivariate random variables. ERGO is based on cumulative and Shannon entropy, and allows reliable causal inference without having to make assumptions on the distributions or correlation relationship of the data.

Empirical evaluation showed that ERGO is highly accurate, very resilient to noise, and outperforms the state of the art by a wide margin. As future work, we plan to refine the measures of complexity that ERGO uses, extend it toward inferring causal networks, as well as studying instantiations of our framework for time-series and discrete data.

Acknowledgements

The author would like to thank Hoang-Vu Nguyen for extensive discussions on the topic and contributions to an early draft of this work. The author is supported by the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government.

References

[1] Z. Chen, K. Zhang, and L. Chan. Nonlinear causal discovery for high dimensional data: A kernelized trace method. In *ICDM*, pages 1003–1008, 2013.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience New York, 2006.

[3] A. D. Crescenzo and M. Longobardi. On cumulative entropies. *J. Stat. Plan. and Inf.*, 139(2009):4072–4087, 2009.

[4] D. Janzing, P. O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *ICML*, pages 479–486, 2010.

[5] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artif. Intell.*, 182-183:1–31, 2012.

[6] D. Janzing and B. Schölkopf. Causal inference using the algorithmic markov condition. *IEEE TIT*, 56(10):5168–5194, 2010.

[7] J. Lemeire and D. Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Mind Mach*, 23(2):227–249, 2013.

[8] M. Li and P. Vitányi. *An Introduction to Kolmogorov Complexity and its Applications*. Springer, 1993.

[9] H.-V. Nguyen, E. Müller, J. Vreeken, and K. Böhm. Multivariate maximal correlation analysis. In *ICML*, pages 775–783. JMLR, 2014.

[10] H.-V. Nguyen, E. Müller, J. Vreeken, F. Keller, and K. Böhm. CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In *SDM*, pages 198–206. SIAM, 2013.

[11] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[12] J. Peters, D. Janzing, and B. Schölkopf. Causal inference on discrete data using additive noise models. *IEEE TPAMI*, 33:2436–2450, 2011.

[13] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman,

G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

[14] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(1):465–471, 1978.

[15] S. Seth and J. C. Príncipe. Assessing Granger non-causality using nonparametric measure of conditional independence. *IEEE TNN*, 23(1):47–59, 2012.

[16] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. J. Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, 2006.

[17] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[18] O. Stegle, D. Janzing, K. Zhang, J. M. Mooij, and B. Schölkopf. Probabilistic latent variable models for distinguishing between cause and effect. In *NIPS*, pages 1687–1695, 2010.

[19] X. Sun, D. Janzing, and B. Schölkopf. Causal reasoning by evaluating the complexity of conditional densities with kernel methods. *Neurocomp.*, 71(7-9):1248–1256, 2008.

[20] P. van der Putten and M. van Someren. A bias-variance analysis of a real world learning problem: The coil challenge 2000. *Mach. Learn.*, 57(1-2):177–195, 2004.

[21] N. Vereshchagin and P. Vitanyi. Kolmogorov's structure functions and model selection. *IEEE TIT*, 50(12):3265–3290, 2004.

[22] J. Vreeken, M. van Leeuwen, and A. Siebes. KRIMP: Mining itemsets that compress. *Data Min. Knowl. Disc.*, 23(1):169–214, 2011.

[23] C. S. Wallace. *Statistical and inductive inference by minimum message length*. Springer, 2005.

[24] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *UAI*, pages 647–655, 2009.

[25] J. Zscheischler. Database with cause-effect pairs. `http://webdav.tuebingen.mpg.de/cause-effect/` accessed Sep 2014.

[26] J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *UAI*, pages 839–846, 2011.

A Proofs

Proof. [Theorem 3.1a] Intuitively, since $h(X \mid z) \geq 0$, its expectation is also non-negative. Consequently, $E_Z[h(X \mid Z)]$ is zero if and only if $h(X \mid z) = 0$ for all $z \in dom(Z)$. This implies: $\forall z \in dom(Z), \forall x \in dom(X)$:

$$P(X \leq x \mid z) \log P(X \leq x \mid z) = 0 \quad .$$

From $x \log x = 0$ if and only if $x = 0$ or $x = 1$, we arrive at

$$\begin{aligned} &\forall z \in dom(Z), \forall x \in dom(X) : \\ &\quad P(X \leq x \mid z) = 0 \vee P(X \leq x \mid z) = 1 \quad . \end{aligned}$$

Let $A_z = \{x : P(X \leq x \mid v) = 1\}$ then $A_z \neq \emptyset$ and $\min A_z$ exists since

$$\begin{aligned} \lim_{x \to +\infty} P(X \leq x \mid z) = 1 \quad , \\ \lim_{x \to -\infty} P(X \leq x \mid z) = 0 \quad . \end{aligned}$$

Hence, for every z, there exists a unique $x_z = \min A_z$ such that $p_{X|Z}(X = x_z \mid z) = 1$, i.e., X is a function of Z. □

Proof. [Theorem 3.1b] Given a fixed $x_0 \in dom(X)$, $P(X \leq x_0 \mid z)$ is a number depending on z. So if we let $V = P(X \leq x_0 \mid Z)$, V is then a random variable. According to the Jensen's inequality, it holds that

$$E_V[V \log V] \geq E_V(V) \log E_V(V) \quad .$$

Thus,

$$\begin{aligned} E_Z[P(X \leq x_0 \mid Z) \log P(X \leq x_0 \mid Z)] \geq \\ E_Z[P(X \leq x_0 \mid Z)] \log E_Z[P(X \leq x_0 \mid Z)] \quad . \end{aligned}$$

From

$$P(X \leq x_0 \mid z) = \int_{-\infty}^{x_0} p_{X|Z}(x \mid z) dx \quad ,$$

we arrive at

$$E_Z[P(X \leq x_0 \mid Z)] = \int_{dom(Z)} \int_{-\infty}^{x_0} p_{X,Z}(x, z) dx dz \quad ,$$

which leads to

$$E_Z[P(X \leq x_0 \mid Z)] = P(X \leq x_0) \quad .$$

Therefore

$$\begin{aligned} E_Z[P(X \leq x_0 \mid Z) \log P(X \leq x_0 \mid Z)] \geq \\ P(X \leq x_0) \log P(X \leq x_0) \quad . \end{aligned}$$

Replacing x_0 by x, integrating and negating both sides w.r.t. x, we obtain

$$E_Z[h(X \mid Z)] \leq h(X) \quad .$$

Since $g(w) = w \log w$ is strictly convex, equality holds if and only if $V = E_V[V]$. This implies $P(X \leq x \mid Z) = P(X \leq x)$, i.e. X is independent of Z. □

Proof. [Lemma 3.1] Let σ and σ' be the permutations of Y that yield $h(\mathbf{Y})$ and $h(\mathbf{Y} \mid \mathbf{X}')$, respectively. We have

$$\begin{aligned} &\sum_{i=1}^{l} h(\mathbf{Y}_{\sigma'(i)} \mid \mathbf{X}, \mathbf{Y}_{\sigma'(1)}, \ldots, \mathbf{Y}_{\sigma'(i-1)}) \\ &\leq \sum_{i=1}^{l} h(\mathbf{Y}_{\sigma(i)} \mid \mathbf{X}, \mathbf{Y}_{\sigma(1)}, \ldots, \mathbf{Y}_{\sigma(i-1)}) \\ &\leq \sum_{i=1}^{l} h(\mathbf{Y}_{\sigma(i)} \mid \mathbf{Y}_{\sigma(1)}, \ldots, \mathbf{Y}_{\sigma(i-1)}) \quad . \end{aligned}$$

Therewith, we have $h(\mathbf{Y} \mid \mathbf{X}') \leq h(\mathbf{Y})$. □

Proof. [Theorem 4.1] Let $g^* = \arg \min_{g:|g|=l} h(X_c | B, x_{1,u}^g)$. We denote l bins that g^* generates on X_p as $b(X_p)_1, \ldots, b(X_p)_l$. We write $|b(X_p)_t|$ as the number of values of X_p in $b(X_p)_t$. For each $X_i' \in X'$, we denote its bins as $b(X_i')_1, \ldots, b(X_i')_{n_i'}$.

Further, let $c_z = \sum_{i=1}^{z} |b(X_p)_t|$. Note that each bin of X_p is non-empty, i.e., $c_z \geq z$. We use $h(X_c \mid X', b_t)$ to denote $h(X_c \mid X')$ computed using the points of $\mathbf{X}$ corresponding to the realizations of X_p in $b(X_p)_t$, projected onto X_c and X'.

We write $|(t, t_1, \ldots, t_k)|$ as the number of points in the cell made up by bins $b(X_p)_t, b(X_1')_{t_1}, \ldots, b(X'_{|X'|})_{t_{|X'|}}$.

We have

$$\begin{aligned} f(u,l) &= \sum_{t=1}^{l} \sum_{t_1=1}^{n_1'} \cdots \sum_{t_{|X'|}=1}^{n'_{|X'|}} \frac{|(t, t_1, \ldots, t_k)|}{u} \times \\ &\quad h(X_c \mid b(X_p)_t, b(X_1')_{t_1}, \ldots, b(X'_{|X'|})_{t_{|X'|}}) \\ &= \sum_{t=1}^{l-1} \sum_{t_1=1}^{n_1'} \cdots \sum_{t_{|X'|}=1}^{n'_{|X'|}} \frac{|(t, t_1, \ldots, t_k)|}{u} \times \\ &\quad h(X_c \mid b(X_p)_t, b(X_1')_{t_1}, \ldots, b(X'_{|X'|})_{t_{|X'|}}) \\ &\quad + \frac{|b_l|}{u} h(X_c \mid X', b_l) \\ &= \frac{c_{l-1}}{u} \sum_{t=1}^{l-1} \sum_{t_1=1}^{n_1'} \cdots \sum_{t_{|X'|}=1}^{n'_{|X'|}} \frac{|(t, t_1, \ldots, t_k)|}{u} \times \\ &\quad h(X_c \mid b(X_p)_t, b(X_1')_{t_1}, \ldots, b(X'_{|X'|})_{t_{|X'|}}) \\ &\quad + \frac{u - c_{l-1}}{u} h(X_c \mid X', \langle e_{c_{l-1}+1,u} \rangle) \\ &= \frac{c_{l-1}}{u} f(c_{l-1}, l-1) \\ &\quad + \frac{u - c_{l-1}}{u} h(X_c \mid X', \langle e_{c_{l-1}+1,m} \rangle) \quad . \end{aligned}$$

In the last line,

$$\sum_{t=1}^{l-1} \sum_{t_1=1}^{n'_1} \cdots \sum_{t_{|X'|}=1}^{n'_{|X'|}} \frac{|(t, t_1, \ldots, t_k)|}{u} \times$$
$$h(X_c \mid b(X_p)_t, b(X'_1)_{t_1}, \ldots, b(X'_{|X'|})_{t_{|X'|}})$$

is equal to $f(c_{l-1}, l-1)$ because otherwise, we could decrease $f(u, l)$ by choosing a different discretization of $e_{1,c_{l-1}}$ into $l-1$ bins. This in turn contradicts our definition of $f(u, l)$. Since $c_{l-1} \in [l-1, u)$ and $f(u, l)$ is minimal over all $j \in [l-1, u)$, we arrive at the final result. □

B Extension of Section 6.2 – Description of Data Pairs

The first data set is from the medical domain and has 120 records [25]. X contains 6 symptoms of patients and Y contains diagnosis results of two diseases. We note that Y is essentially categorical. It is expected that $X \to Y$.

The second contains climate forecast data [25]. Both X and Y contain four components $\{$*air temperature, pressure at surface, sea level pressure, relative humidity*$\}$. They were measured at the same location grid at different times. $\mathbf{X}$ contains 10 266 observations on day 50 of year 2000. $\mathbf{Y}$ contains respectively 10 266 observations on day 51 of year 2000. The ground truth is assumed to be $X \to Y$.

The third data set contains daily mean values of ozone values and sun radiation in the last 83 days of 2009 at 16 different places in Switzerland [25]. We set X to contain 16 measures of ozone values and Y to contain 16 measures of sun radiation. The causal direction is assumed to be $Y \to X$.

Fourth, we test on data capturing the relationship between ozone concentration and different meteorological parameters [25]. The data contains 989 observations and was collected in Neckarsulm and Heilbronn, Germany. We define $X = \{$*ozone concentration*$\}$ and $Y = \{$*wind speed, global radiation, temperature*$\}$. According to Janzing et al. [4] the production of ozone is impacted by wind, sun radiation, and air temperature, by which the ground truth causal direction is $Y \to X$. We note that X is univariate and Y is multivariate.

The fifth data set is on car efficiency in terms of miles per gallon, and has 392 records [25]. We define $X = \{$*displacement, horsepower, weight*$\}$ and $Y = \{mpg, acceleration\}$. It is natural to assume that $X \to Y$.

The fifth to ninth data sets are taken from [4]. The next three are resp. on precipitation data, and two on stock indices. The ninth data set is on weather and air pollution in Chemnitz, Germany [4]. It contains 1440 observations. X has 3 dimensions on the wind direction and air temperature. Y has 6 dimensions on air quality indices, e.g. sulfur dioxid and dust concentration. The ground truth is that $X \to Y$.