Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Entropic Causal Inference

Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath
Department of Electrical and Computer Engineering,
The University of Texas at Austin, USA

Babak Hassibi
Department of Electrical Engineering
California Institute of Technology, USA

Abstract

We consider the problem of identifying the causal direction between two discrete random variables using observational data. Unlike previous work, we keep the most general functional model but make an assumption on the unobserved exogenous variable: Inspired by Occam's razor, we assume that the exogenous variable is *simple* in the true causal direction. We quantify simplicity using Rényi entropy. Our main result is that, under natural assumptions, if the exogenous variable has low H_0 entropy (cardinality) in the true direction, it must have high H_0 entropy in the wrong direction. We establish several algorithmic hardness results about estimating the minimum entropy exogenous variable. We show that the problem of finding the exogenous variable with minimum H_1 entropy (Shannon Entropy) is equivalent to the problem of finding minimum joint entropy given n marginal distributions, also known as minimum entropy coupling problem. We propose an efficient greedy algorithm for the minimum entropy coupling problem, that for $n = 2$ provably finds a local optimum. This gives a greedy algorithm for finding the exogenous variable with minimum Shannon entropy. Our greedy entropy-based causal inference algorithm has similar performance to the state of the art additive noise models in real datasets. One advantage of our approach is that we make no use of the values of random variables but only their distributions. Our method can therefore be used for causal inference for both ordinal and also categorical data, unlike additive noise models.

1 Introduction

Causality has been studied under several frameworks including potential outcomes (Rubin 1974) and structural equation modeling (Pearl 2009). Under the Pearlian framework (Pearl 2009) it is possible to discover some causal directions between variables using only observational data with conditional independence tests. The PC algorithm (Spirtes, Glymour, and Scheines 2001) and its variants fully characterize which causal directions can be learned in the general case. For large graphs, GES algorithm (Chickering 2002) provides a score-based test to greedily identify the highest scoring causal graph given the data. Unfortunately, these approaches do not guarantee the recovery of true causal direction between every pair of variables, since typically data could be generated by several statistically equivalent causal graphs.

A general solution to the causal inference problem is to conduct experiments, also called interventions. An intervention forces the value of a variable without affecting the other system variables. This removes the effect of its causes, effectively creating a new causal graph. These changes in the causal graph create a post-interventional distribution among variables, which can be used to identify some additional causal relations in the original graph. The procedure can be applied repeatedly to fully identify any causal graph (Hauser and Bühlmann 2012a), (Hauser and Bühlmann 2012b), (Hyttinen, Eberhardt, and Hoyer 2013), (Shanmugam et al. 2015).

For many problems, it can be very difficult to create interventions since they require additional experiments after the original data collection. Researchers would still like to discover causal relations between variables using only observational data, using so-called data-driven causality. Several recent works (Chen et al. 2014; Shajarisales et al. 2015) have developed such methods. To be able to make any conclusions on causal directions in this case, additional assumptions must be made about the mechanisms that generate the data.

In this paper we focus on the simplest causal discovery problem that involves only two variables. The two causal graphs $X \to Y$ and $X \leftarrow Y$ are statistically indistinguishable so conditional independence tests cannot make any causal inference from observational data without interventions. Statistical indistinguishability easily follows from the fact that any joint distribution on two variables $p(x, y)$ can be factorized both as $p(x)p(y/x)$ and $p(y)p(x/y)$.

The most popular assumption for two-variable data-driven causality is the additive noise model (ANM) (Shimizu et al. 2006). In ANM, any outside factor is assumed to affect the effect variable additively, which leads to the equation $Y = f(X) + E, E \perp\!\!\!\perp X$. Although restrictive, this assumption leads to strong theoretical guarantees in terms of identifiability, and provides the state of the art accuracy in real datasets. (Shimizu et al. 2006) showed that if f is linear and the noise is non-Gaussian the causal direction is identifiable. (Hoyer et al. 2008) showed that when f is non-linear, irrespective of the noise, identifiability holds in a non-adverserial setting of system parameters. (Peters, Janzing, and Schölkopf 2011) extended ANM to discrete variables.

Another approach is to exploit the postulate that the cause and mechanism are in general independently assigned by nature. The notion of *independence* here is vague and one needs

to assign maps, or conditional distributions to random variables to argue about independence of cause and mechanism. In this direction an information-geometry based approach is suggested (Janzing et al. 2012). Independence of cause and mechanism is captured by treating the log-slope of the function as a random variable, and assuming that it is independent from the cause. In the case of a deterministic relation $Y = f(X)$, there are theoretical guarantees on identifiability. However, this assumption is restrictive for real data.

Previous work exploited these two ideas, additive noise, and independence of cause and mechanism, to draw data-driven causal conclusions about problems in a diverse range of areas from astronomy to neuroscience (Shajarisales et al. 2015), (Schölkopf et al. 2015). (Shajarisales et al. 2015) uses the same idea that the cause and effect are independent in the time series of a linear filter. They suggest the spectral independence criterion, which is robust to time shifts. (Chen et al. 2014) uses kernel space embeddings with the assumption that the cause distribution $p(x)$ and mechanism $p(y|x)$ are selected independently to distinguish cause from effect. An exception to these frameworks is to use a binary classifier on the joint distribution on (X, Y) to classify distributions into those that come from the causal model $X \to Y$ and $Y \to X$ (Lopez-Paz et al. 2015; Lopez-Paz, Muandet, and Recht Dec 2015; Lopez-Paz and Oquab 2016). However, it is not clear what are the correct set of assumptions to show an identifiability result with this approach.

As noted by (Chen et al. 2014), although conceptually proposed before, using Kolmogorov complexity of the factorization of the joint distribution $p(y|x)p(x)$ and $p(x|y)p(y)$ as a criterion for deciding causal direction has not been used successfully until now.

The use of information theory as a tool for causal discovery is currently gaining increasing attention. This is through different appoaches, e.g., for time-series data, Granger causality and Directed Information can be used (Granger 1969; Etesami and Kiyavash 2016; Quinn, Kiyavash, and Coleman Dec 2015; Kontoyiannis and Skoularidou Aug 2016), see also (MCI 2016). However, researchers have not used entropy as a measure of simplicity in the causal discovery literature, probably because the entropies $H(Y|X)$ and $H(X|Y)$ do not give us any more information than $H(X)$ and $H(Y)$, due to the symmetry $H(Y) + H(X|Y) = H(X) + H(Y|X)$. In our work, as we will explain, we minimize $H(E)$ which initially sounds similar, *but is fundamentally different* from $H(Y|X)$. Entropy has found some additional uses in the causality literature recently: In (Gao et al. 2016), authors use maximum mutual information between X, Y in order to quantify the causal strength of a known causal graph.

The work that is most similar to ours in spirit is (Mooij et al. 2010), which also drops the additive noise assumption. Their approach and setup are different in many ways: Authors work with continuous data. To be able to handle this generic form, they have to make strong assumptions on the exogenous variable, function, and distribution of the cause: Mooij et al. assume that the exogenous variable is a standard Gaussian, a Gaussian mixture prior for the cause, and a Gaussian process as the prior of the function.

1.1 Our Contributions

In this paper, we propose a novel approach to the causal identifiability problem for discrete variables. Similar to (Mooij et al. 2010), we keep the most general functional model, but only put an assumption on the exogenous (background) variable. Based on Occam's razor, we employ a simplicity assumption on the unobserved exogenous variable. We use Rényi entropy, which is defined as $H_a(X) = \frac{1}{1-a} \log \left(\sum_i p_i^a\right)$, for a random variable X with state probabilities p_i. We focus on two special cases of Rényi entropy: H_0, which corresponds to the logarithm of the number of states, and H_1 which corresponds to Shannon entropy, but our framework can be extended.

Specifically, if the true causal direction is $X \to Y$, then the random variable Y is an arbitrary function of X and an exogenous variable E: $Y = f(X, E)$ where E is independent from the cause X. Our key assumption is that the exogenous variable E is *simple*, i.e., has low Rényi entropy. The postulate is that for any model in the wrong direction $X = f'(Y, \tilde{E})$, the exogenous variable $\tilde{E}$ has high Rényi entropy. We are able to prove this result for the H_0 special case of Rényi entropy, assuming generic distributions for X, Y. Furthermore, we empirically show that using H_1 Shannon entropy we obtain practical causality tests that work with high probability in synthetic datasets and that slightly outperforms the previous state of the art in real datasets.

Our assumption is an entropic interpretation of Occam's razor, motivated by what E represents in the causal model. The exogenous variable captures the combined effect of all the variables not included in the system model, which affect the distribution of Y. Our causal assumption can be stated as *"there should not be too much complexity not included in the causal model"*. For $a \to 1$, *i.e.*, Shannon entropy, $H(X) + H(E)$, $H(Y)+H(\tilde{E})$ are the number of random bits required to generate an input for the causal system $X \to Y$ and $X \leftarrow Y$, respectively. The simplest explanation of an observed joint distribution, i.e., the direction which requires nature to generate smaller number of random bits is selected as the true causal model. More precisely we have the following:

Assumption 1. *Entropy of the exogenous variable E is small in the true causal direction.*

The notions of simplicity that we consider are H_0, which is log-cardinality, and H_1, which is Shannon entropy. One significant advantage of using Shannon entropy as a simplicity metric is that it can be estimated more robustly in the presence of measurement errors, unlike cardinality H_0.

We prove an identifiability result for H_0 entropy, i.e., cardinality of E: If the probability values are not adversarially chosen, for most functions, the true causal direction is identifiable under Assumption 1. Based on experimental evidence, we conjecture that a similar identifiability result must hold for Shannon entropy H_1.

To use our framework we need algorithms that explain a dataset by finding an exogenous variable E with minimum cardinality H_0 and minimum Shannon entropy H_1. Since the entropies of X and Y can be very different, any metric to determine the true causal direction cannot only consider the entropy of the exogenous variable without incorporating

the entropy of the cause. We explain the exogenous variable in both directions and declare the causal direction to be the one with the smallest joint entropy $H_a(X) + H_a(E)$ versus $H_a(Y) + H_a(\tilde{E})$. Our method can be applied for any Rényi entropy H_a but in this paper we only use $a = 0$ and $a = 1$.

Unfortunately, minimizing $H_0(E)$ seems very hard for real datasets since it offers no noise robustness. For Shannon entropy we can do much better for real data. The first step in obtaining a practical algorithm is showing that the minimum H_1 explanation is equivalent to the following problem: For n random variables with given marginal distributions, find a joint distribution with the minimum Shannon entropy that is consistent with the given marginals. This problem is called the minimum Shannon entropy coupling and is known to be NP hard (Kovacevic, Stanojevic, and Senk 2012). We propose a greedy approximation algorithm for this problem that empirically performs very well. We also prove that, for $n = 2$, our algorithm always produces a local minimum.

In summary our contributions in this paper include[1]:

- We show identifiability for generic low-entropy causal models under Assumption 1 with H_0.
- We show that the problems of identifying the minimum cardinality (H_0) exogenous variable, and identifying the minimum Shannon entropy (H_1) exogenous variable given a joint distribution are both NP hard.
- We design a novel greedy algorithm for the minimum entropy coupling problem, which turns out to be equivalent to the problem of finding exogenous variable with minimum H_1 entropy.
- We empirically validate the conjecture that the causal direction is identifiable under Assumption 1 with H_1, using experiments on synthetic datasets.
- We empirically show that our causal inference algorithm based on Shannon entropy minimization has slightly better performance than the existing best algorithms on a real causal dataset. Interestingly, our algorithm uses only the probability distributions rather than the actual values of the random variables, and hence is applicable to categorical variables.

1.2 Background and Notation

A tuple $\mathcal{M} = (X, U, \mathcal{F}, D, p)$ is a causal model when, 1) $\mathcal{F} = \{f_i\}$ are deterministic functions, 2) $X = \{X_i\}$ are a set of endogenous (observed) variables $U = \{U_i\}$ are a set of exogenous (latent) variables with $X_i = f_i(Pa_i, U_i), \forall i$ where Pa_i are the endogenous parents and U_i is the exogenous parent of X_i in directed acyclic graph D, 3) U are mutually independent with respect to p. The observable variable set X has a joint distribution implied by the distributions of U, and the functional relations f_i. D is then a Bayesian network for the induced joint distribution of endogenous variables. A standard assumption employed in Pearl's model *causal sufficiency* is also used here: Every exogenous variable is a direct parent of at most one endogenous variable.

In this paper, we consider a simple two variable causal system which contains only two endogenous variables X, Y. Assume X causes Y, which is represented as $X \to Y$. The model is determined only by one exogenous variable E, and a function f, where $Y = f(X, E)$. The probability distribution of X and E, and f determines the distribution of Y. This model is shown by the tuple $\mathcal{M} = (\{X, Y\}, E, f, X \to Y, p_{X,E})$. Notice that we do not assign an exogenous variable to X, since it is the source node in the graph.

We denote the set $\{1, 2, \cdots, n\}$ by $[n]$. $\sum_i x_i$ is meant to run through every possible index. log refers to the logarithm base 2. For two variables X, Y, $\mathbf{Y}|\mathbf{X}$ and $\mathbf{X}|\mathbf{Y}$ denote the conditional probability distribution matrices, i.e., $\mathbf{Y}|\mathbf{X}(i, j) = p(y = i|x = j)$ and $\mathbf{X}|\mathbf{Y}(i, j) = p(x = i|y = j)$. The statistical independence of two random variables X and E are shown by $X \perp\!\!\!\perp E$. For notational convenience, probability distribution of random variable X is shown by $p(x)$ as well as $p_X(x)$. $\mathbf{x}$ shows the distribution of X in vector form , i.e., $x_i = \mathbf{x}(i) = \mathbb{P}(X = i)$. $n - 1$ simplex is the set of points x in n dimensional Euclidean space that satisfy $\sum_i x(i) = 1$. card is the cardinality of a set.

2 Causal Model with Minimum Cardinality Exogenous Variable

Consider the causal model $\mathcal{M} = (\{X, Y\}, E_0, f_0, X \to Y, p_{X,E})$. The task is to identify the underlying causal graph $X \to Y$ using independent identically distributed samples $\{(x_i, y_i)\}_i$. Assuming causal sufficiency, this task reduces to deciding whether X causes Y or Y causes X. To isolate the identifiability problem from estimation errors due to finite samples, we assume that the joint distribution of (X, Y) is available. Most proofs are deferred to the Appendix.

One way to identify that X causes Y is by showing that although there exists a function f and random variable E with $Y = f(X, E), X \perp\!\!\!\perp E$, there is no function, random variable pair $(g, \tilde{E})$ such that $X = g(Y, \tilde{E}), Y \perp\!\!\!\perp \tilde{E}$. However, without more assumptions, this is not possible: For any joint distribution one can find valid causal models for both $X \to Y, X \leftarrow Y$. This is widely known, although for completeness, we provide a proof (Lemma 4 in the Appendix).

Even when the true causal graph is known, one can create different constructions of f, E with $Y = f(X, E), X \perp\!\!\!\perp E$. There is no way to distinguish the true causal model. However, even though we cannot recover the actual function and the exogenous variable, we can still show identifiability.

First, we give an equivalent characterization of a causal model on two variables.

Definition 1 (Block Partition Matrices). *Consider a matrix $\mathbf{M} \in \{0, 1\}^{n^2 \times m}$. Let $\mathbf{m}_{i,j}$ represent the $i + (j - 1)n$ th row of $\mathbf{M}$. Let $S_{i,j} = \{k \in [m] : \mathbf{m}_{i,j}(k) \neq 0\}$. $\mathbf{M}$ is called a block partition matrix if it belongs to $\mathcal{C} := \{\mathbf{M} : \mathbf{M} \in \{0, 1\}^{n^2 \times m}, \bigcup_{i \in [n]} S_{i,j} = [m], S_{i,j} \cap S_{l,j} = \emptyset, \forall i \neq l\}$.*

$\mathcal{C}$ thus stands for $0, 1$ matrices with n^2 rows and m columns where each block of n rows correspond to a partitioning of the set $[m]$. We make the following key observation:

Lemma 1. *Given discrete random variables X, Y with distribution $p(x, y)$, $\exists$ a causal model $\mathcal{M} = (\{X, Y\}, E, f, X \to Y, p_{X,E})$, $E \in \mathcal{E}$ with card$(\mathcal{E}) = m$ if and only if $\exists \mathbf{M} \in \mathcal{C}, \mathbf{e} \in \mathbb{R}_+^m$ with $\sum_i \mathbf{e}(i) = 1$ that satisfy $vec(\mathbf{Y}|\mathbf{X}) = \mathbf{M}\mathbf{e}$.*

[1]For a version with proofs, see https://arxiv.org/abs/1611.04035

In other words, the existence of a causal pair $X \to Y$ is equivalent to the existence of a block partition matrix $\mathbf{M}$ and a vector $\mathbf{e}$ of proper dimensions with $vec(\mathbf{Y}|\mathbf{X}) = \mathbf{Me}$.

For simplicity, assume $|\mathcal{X}| = |\mathcal{Y}| = n$. We later remove this constraint. We first show that any joint distribution can be explained using a variable E with $n(n-1)+1$ states.

Lemma 2 (Upper Bound on Minimum Cardinality of E). *Let $X \in \mathcal{X}, Y \in \mathcal{Y}$ be two random variables with joint probability distribution $p_{X,Y}(x,y)$, where $|\mathcal{X}| = |\mathcal{Y}| = n$. Then $\exists$ a causal model $Y = f(X,E), X \perp\!\!\!\perp E$ that induces $p_{X,Y}$, where E has support size $n(n-1)+1$.*

We can show that, if the columns of $\mathbf{Y}|\mathbf{X}$ are uniformly sampled points in the $n-1$ dimensional simplex, then $n(n-1)$ states are also necessary for E (see Proposition 3in the Appendix). This shows, unless designed by nature through the causal mechanism, exogenous variable cannot have small cardinality. Based on this observation, the hope is to prove that in the wrong causal direction, say $X \to Y$ and we find an $\tilde{E} \perp\!\!\!\perp Y$ such that $X = g(Y, \tilde{E})$ for some g, the exogenous variable $\tilde{E}$ has to have large cardinality. In the next section, we show this is actually through, under mild conditions on f.

2.1 Identifiability for H_0 Entropy

In a causal system $Y = f(X,E)$, nature chooses the random variables X, E, and function f, and the conditional probability distributions are then determined by these. We are interested in the cardinality of variables $\tilde{E} \perp\!\!\!\perp X$ in the wrong causal direction $X = g(Y, \tilde{E})$. Considering $\mathbf{X}|\mathbf{Y}$, we can show that the same lower bound of $n(n-1)$ still holds despite nature now chooses E and X randomly, rather than choosing the columns of $\mathbf{X}|\mathbf{Y}$ directly. A mild assumption on f is needed to avoid degenerate cases (For counterexamples see the appendix).

Definition 2 (Generic Function). *Let $Y = f(X,E)$ where variables X, Y, E have supports $\mathcal{X}, \mathcal{Y}, \mathcal{E}$, respectively. Let $S_{y,x} = f_x^{-1}(y) \subset \mathcal{E}$ be the inverse map for x, e, i.e., $S_{y,x} = \{e \in \mathcal{E} : y = f(x,e)\}$. A function f is called "generic", if for each (x_1, x_2, y) triple $f_{x_1}^{-1}(y) \neq f_{x_2}^{-1}(y)$ and for every (x,y) pair $f_x^{-1}(y) \neq \emptyset$.*

In other words f is called generic if y^{th} row in the x_1^{th} block of matrix $\mathbf{M}$ in the decomposition $vec(\mathbf{Y}|\mathbf{X}) = \mathbf{Me}$ is different from y^{th} row in the x_2^{th} block, and both are nonzero. This is not a restrictive condition, for example if $p(y|x)$ are all different, no two rows of $\mathbf{M}$ can be the same. For any given conditional distribution, if the probabilities are perturbed by arbitrarily small continuous noise, the corresponding f will be generic almost surely. We have the following main identifiability result:

Theorem 1 (Identifiability). *Consider the causal model $\mathcal{M} = (\{X,Y\}, E_0, f_0, X \to Y, p_{X,E_0})$ where the random variables X, Y have n states, $E_0 \perp\!\!\!\perp X$ has θ states and f is a generic function .*

If the distributions of X and E are uniformly randomly selected from the $n-1$ and $\theta-1$ simplices, then with probability 1, any $\tilde{E} \perp\!\!\!\perp Y$ that satisfies $X = g(Y, \tilde{E})$ for some deterministic function g has cardinality at least $n(n-1)$.

Theorem 1 implies that the causal direction is identifiable, when the exogenous variable has cardinality $< n(n-1)$:

Corollary 1. *Assume that there exists an algorithm $\mathcal{A}$ that given n random variables $\{Z_i\}, i \in [n]$ with distributions $\{p_i\}, i \in [n]$ each with n states, outputs the distribution of the random variable E with minimum cardinality and functions $\{f_i, i \in [n]\}$ where $Z_i = f_i(E)$.*

Consider the causal pair $X \to Y$ where $Y = f(X, E_0)$. Assume that the cardinality of E_0 is less than $n(n-1)$, and f is generic. Then, $\mathcal{A}$ can be used to identify the true causal direction with probability 1, if X, E_0 are selected uniformly randomly from the proper dimensional simplices.

Proof. Feed the set of conditional distributions $\{\mathbb{P}(Y|X = i) : i \in [n]\}$ and $\{\mathbb{P}(X|Y = i) : i \in [n]\}$ to $\mathcal{A}$ to obtain $E, \tilde{E}$. From Theorem 1, with probability 1, $\mathcal{A}$ identifies $\tilde{E}$ with $\text{card}(\tilde{E}) \geq n(n-1)$. Then since $\text{card}(E) \leq \text{card}(E_0) < \text{card}(\tilde{E})$, comparing cardinalities give the true direction. □

Corollary 1 gives an algorithm for finding the true causal direction: Estimate $E, \tilde{E}$ with minimum H_0 entropy and declare $X \to Y$ if $|\tilde{E}| > |E|$ and declare $X \leftarrow Y$ if $|\tilde{E}| < |E|$. The result easily extends to the case where X and Y are allowed to have different number of states:

Proposition 1 (Inference algorithm). *Suppose $X \to Y$. Let $X \in \mathcal{X}, Y \in \mathcal{Y}$, $|\mathcal{X}| = n, |\mathcal{Y}| = m$. Assume that $\mathcal{A}$ is the algorithm that finds the exogenous variables $E, \tilde{E}$ with minimum cardinality. Then, if the underlying exogenous variable E_0 satisfies $|E_0| < n(m-1)$, with probability 1, we have $|X| + |E| < |Y| + |\tilde{E}|$.*

Proof follows from Corollary 1, and by extending the proof of Theorem 1 to different cardinalities for X, Y.

Unfortunately, it turns out there does not exist an efficient algorithm $\mathcal{A}$, unless P=NP:

Theorem 2. *Given a conditional distribution matrix $\mathbf{Y}|\mathbf{X}$, identifying $E \perp\!\!\!\perp X$ with minimum support size such that there exist a function f with $Y = f(X,E)$ is* NP *hard.*

The hardness of this problem sets us to search for alternative approaches.

3 Causal Model with Minimum H_1 Entropy

In this section, we propose a way to identify the causal model that explains the observational data with minimum Shannon entropy (entropy in short). Entropy of a causal model is measured by the number of random bits required to generate its input. In the causal graph $X \to Y$, where $Y = f(X,E)$, we identify the exogenous variable $E \perp\!\!\!\perp X$ with minimum entropy. We show that this corresponds to a known problem which has been shown to be NP hard. Later we propose a greedy algorithm.

Notice that $H(E)$ is different from the conditional entropy $H(Y|X)$. Certainly, since $Y = f(X,E)$, $H(Y|X) \leq H(E)$. The key is that since E is forced to be independent from X, $H(E)$ cannot be lowered to $H(Y|X)$. To see this, we can write $H(Y|X) = \sum_i p_X(i) H(Y|X = i)$, whereas since conditional probability distribution of $Y|X = i$ is the

same as the distribution of $f_i(E)$ for some function f_i, we have $H(E) \geq \max_i H(Y|X=i)$.

3.1 Finding *E* with Minimum Entropy

Consider the equation $Y = f(X,E), X \perp\!\!\!\perp E$. Let $f_x : \mathcal{E} \to \mathcal{Y}$ be the function mapping E to Y when $X = x$, i.e., $f_x(E) \coloneqq f(x,E)$. Then $\mathbb{P}(Y=y|X=x) = \mathbb{P}(f_x(E) = y|X=x) = \mathbb{P}(f_x(E)=y)$. The last equality follows from the fact that $X \perp\!\!\!\perp E$. Thus, we can treat the conditional distributions $\mathbb{P}(Y|X=x)$ as distributions that emerge by applying some function f_x to some unobserved variable E. Then the problem of identifying E with minimum entropy given the joint distribution $p(x,y)$ becomes equivalent to, given distributions of the variables $f_i(E)$, finding the distribution with minimum entropy (distribution of E), such that there exists functions f_i which map this distribution to the observed distributions of $Y|X=i$. It can be shown that $H(E) \geq H(f_1(E), f_2(E), \ldots, f_n(E))$. Regarding $f_i(E)$ as a random variable U_i, the best lower bound on $H(E)$ can be obtained by minimizing $H(U_1, U_2, \ldots, U_n)$. We can show that we can always construct an E that acheives this minimum. Thus the problem of finding the exogenous variable E with minimum entropy given the joint distribution $p(x,y)$ is equivalent to the problem of finding the minimum entropy joint distribution of the random variables $U_i = (Y|X=i)$, given the marginal distributions $p(Y|X=i)$:

Theorem 3 (**Minimum Entropy Causal Model**). *Assume that there exists an algorithm $\mathcal{A}$ that given n random variables $\{Z_i\}, i \in [n]$ with distributions $\{p_i\}, i \in [n]$ each with n states, outputs the joint distribution over Z_i consistent with the given marginals, with minimum entropy.*

Then, $\mathcal{A}$ can be used to find the causal model $\mathcal{M} = (\{X,Y\}, E, X \to Y, p_{X,E})$ with minimum input entropy, given any joint distribution $p_{X,Y}$.

The problem of minimizing entropy subject to marginal constraints is non-convex. In fact, it is shown in (Kovacevic, Stanojevic, and Senk 2012) that minimizing the joint entropy of a set of variables given their marginals is NP hard. Thus we have the following corollary:

Corollary 2. *Finding the causal model $\mathcal{M} = (\{X,Y\}, E, f, X \to Y, p_{E,X})$ with minimum $H(E)$ that induce a given distribution $p(x,y)$ is* NP *hard.*

For this, we propose a greedy algorithm. Using entropy to identify E instead of cardinality, despite both turning out to be NP hard, is useful since entropy is more robust to noise in data. In real data, we estimate the probability values from samples, and noise is unavoidable.

3.2 A Conjecture on Identifiability with H_1 Entropy

We have the following conjecture, supported by artificial and real data experiments in Section 4.

Conjecture 1. *Consider the causal model $\mathcal{M} = (\{X,Y\}, E, f, X \to Y, p_{X,E})$ where discrete random variables X, Y have n states, $E \perp\!\!\!\perp X$ has θ states. .*

If the distribution of X is uniformly randomly selected from the $n-1$ dimensional simplex and distribution of E is uniformly selected from the probability distributions that satisfy $H_1(E) \leq \log n + \mathcal{O}(1)$ and f is randomly selected from all functions $f : [n] \times [\theta] \to [n]$, then with high probability, any $\tilde{E} \perp\!\!\!\perp Y$ that satisfies $X = g(Y, \tilde{E})$ for some deterministic g entails $H(X) + H(E) < H(Y) + H(\tilde{E})$.

Proposition 2 (Assuming Conjecture 1)**.** *Assume there exists an algorithm $\mathcal{A}$ that given n random variables $\{Z_i\}, i \in [n]$ with distributions $\{p_i\}, i \in [n]$ each with n states, outputs the distribution of the random variable E with minimum entropy and functions $\{f_i\}, i \in [n]$ where $Z_i = f_i(E)$.*

Consider the causal pair $X \to Y$ where $Y = f(X, E_0)$, and cardinality of E_0 is cn for some constant c, and f is selected randomly. Then, $\mathcal{A}$ can be used to identify the true causal direction with high probability, if X, E_0 are uniformly random samples from the proper dimensional simplices.

3.3 Greedy Entropy Minimization Algorithm

Given m discrete random variables with n states, we provide a heuristic algorithm to minimize their joint entropy given their marginal distributions. The main idea is the following: Each marginal probability constraint must be satisfied. For example, for the case of two variables with distributions p_1, p_2, ith row of joint distribution matrix should sum to $p_1(i)$. The contribution of a probability mass to the joint entropy only increases when probability mass is divided into smaller chunks: $-p_1(i)\log p_1(i) \leq -a\log a - b\log b$, when $p_1(i) = a + b$, for $a, b \geq 0$. Thus, we try to keep large probability masses intact to assure that their contribution to the joint distribution is minimized.

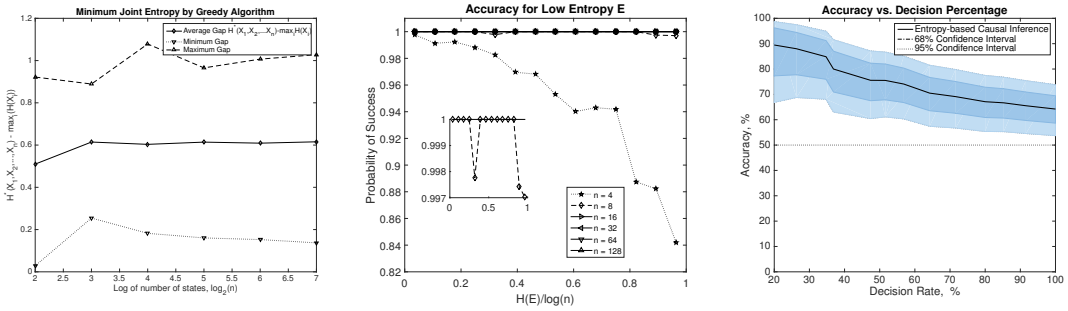
We propose Algorithm 1. The sorting step is only to simplify the presentation. Hence, although the given algorithm runs in time $\mathcal{O}(m^2 n^2 \log n)$, it can easily be reduced to $\mathcal{O}(\max(mn\log n, m^2 n))$ by dropping the sorting step. The algorithm simply proceeds by removing the most probability mass it can at each round. This makes sure the large probability masses remain intact.

Algorithm 1 Joint Entropy Minimization Algorithm

Input: Marginal distributions of m variables each with n states, in matrix form $\mathbf{M} = [p_1^T; p_2^T; ..., p_m^T]$.
$e = [\ ]$
Sort each row of M in decreasing order.
Find minimum of maximum of each row: $r \leftarrow \min_i(p_i(1))$
while $r > 0$ **do**
    $e \leftarrow [e, r]$
    Update maximum of each row: $p_i(1) \leftarrow p_i(1) - r, \forall i$
    Sort each row of M in decreasing order.
    $r \leftarrow \min_i(p_i(1))$
end while
return e.

One can easily construct the joint distribution using a variant: Instead of sorting, at each step, find $r = \min_i\{\max_j\{p_i(j)\}\}$ and assign r to the element with coordinates (a_i), where $a_i = \arg\max_j p_i(j)$.

Lemma 3. *Greedy entropy minimization outputs a point with entropy at most* $\log m + \log n$.



(a) Greedy Joint Entropy Min. (b) Identifiability, Artificial Data (c) Performance on Real Data

Figure 1: (a) Performance of greedy joint entropy minimization algorithm: n distributions each with n states are randomly generated for each value of n. As can be seen, the minimum joint entropy obtained by the greedy algorithm is at most 1 bit away from the largest marginal $\max_i H(X_i)$. (b) Identifiability with Entropy: We generate distributions of X, Y by randomly selecting f, X, E. Probability of success is the fraction of points where $H(X, E) < H(Y, \tilde{E})$. As observed, larger n drives probability of success to 1 when $H(E) \leq \log n$, supporting Conjecture 1. (c) Real Data Performance: Decision rate is the fraction of samples for which algorithm makes a decision for a causal direction. A decision is made when $|H(X, E) - H(Y, \tilde{E})| > t \log_2 n$, where t determines the decision rate. Confidence intervals are also provided.

Lemma 3 follows from the fact that the algorithm returns a support of size at most $m(n-1)+1$.

We also prove that, when there are two variables with n dimensions, the algorithm returns a point that satisfies the KKT conditions of the optimization problem, which implies that it is a local optimum (see Proposition 4 in the Appendix).

4 Experiments

In this section, we test the performance of our algorithms on real and artificial data. First, we test the greedy entropy minimization algorithm and show that it performs close to the trivial lower bound. Then, we test our conjecture of identifiability using entropy. Lastly, we test our entropy-minimization based causal identification technique on real data.

In order to test our algorithms, we sample points in proper dimensional simplices, which correspond to distributions for X and E. Distribution of points are uniform for selecting the distribution of X. It is well-known that a vector $[x_i/Z]_i$ is uniformly randomly distributed over the simplex, if x_i are i.i.d. exponential random variables with parameter 1, and $Z = \sum_i x_i$ (Onn and Weissman 2011). To sample low-entropy distributions for E, instead of exponential, we use a heavy tailed distribution for sampling each coordinate. Specifically, we use $[e_i/Z]_i$, where e_i are i.i.d. log-normal random variables with parameter σ. We observe that this allows us to sample a variety of distributions with small entropy.

Performance of Greedy Entropy Minimization: We sample distributions for n random variables $\{X_i\}, i \in [n]$ each with n states and apply Algorithm 1 to minimize their joint entropy. We compare our greedy joint entropy minimization algorithm with the simple lower bound of $\max_i H(X_i)$. Figure 1a shows average, maximum and minimum excess bits relative to this lower bound. Contrary to the pessimistic bound of $\log n$ bits, joint entropy is at most 1 bit away from $\max_i H(X_i)$ for the given range of n.

Verifying Entropy-Based Identifiability Conjecture: In this section, we empirically verify Conjecture 1. The distributions for X are uniformly randomly sampled from the simplex in n dimensions. We also select f randomly (see implementation details). For the log-normal parameter σ used for sampling the distribution of E from the $n(n-1)$ dimensional simplex, we sweep the integer values from 2 to 8. This allows us to get distribution samples from different regimes. We only consider the samples which satisfy $H(E) \leq \log n$.

After sampling E, X, f, we identify the corresponding $\mathbf{Y}|\mathbf{X}$ and $\mathbf{X}|\mathbf{Y}$ for $Y = f(X, E)$. We apply greedy entropy minimization on the columns of the induced distributions $\mathbf{Y}|\mathbf{X}, \mathbf{X}|\mathbf{Y}$ to get the estimates $E, \tilde{E}$ for both causal models $Y = f(X, E)$ and $X = g(Y, \tilde{E})$, respectively. Figure 1b shows the variation of success probability, i.e., the fraction of samples which satisfy $H(X) + H(E) < H(Y) + H(\tilde{E})$. As observed, as n is increased, probability of success converges to 1, when $H(E) \leq \log n$, which supports the conjecture.

Experiments on Real Cause Effect Pairs: We test our entropy-based causal inference algorithm on the CauseEffectPairs repository (Mooij et al. 2016a). ANM have been reported to achieve an accuracy of 63% with a confidence interval of $\pm 10\%$ (Mooij et al. 2016b). We also use the binomial confidence intervals as in (Clopper and Pearson 1934).

The cause effect pairs show very different characteristics. From the scatter plots, one can observe that they can be a mix of continuous and discrete variables. The challenge in applying our framework on this dataset is choosing the correct quantization. Small number of quantization levels may result in loss of information regarding the joint distribution, and a very large number of states might be computationally hard to work with. We pick the same number of states for both X and Y, and use a uniform quantization that assures each state of the variables has ≥ 10 samples on average. From the samples, we estimate the conditonal transition matrices $\mathbf{Y}|\mathbf{X}$ and $\mathbf{X}|\mathbf{Y}$ and feed the columns to the greedy entropy minimization algorithm (Algorithm 1), which outputs an approximate of

the smallest entropy exogenous variable. Later we compare $H(X, E)$ and $H(Y, \tilde{E})$ and declare the model with smallest input entropy to be the true model, based on Conjecture 1.

For a causal pair, we invoke the algorithm if $|H(X, E) - H(Y, \tilde{E})| \geq t \log(n)$ for threshold parameter t, which determines the decision rate. Accuracy becomes unstable for very small decision rates, since the number of evaluated pairs becomes too small. At 100% decision rate, algorithm achieves 64.21% which is slightly better than the 63% performance of ANM as reported in (Mooij et al. 2016b). In addition, our algorithm only uses probability values, and is applicable to categorical as well as ordinal variables.

Acknowledgements

This work has been supported by NSF Grants CCF 1344179, 1344364, 1407278, 1422549, ARO YIP W911NF-14-1-0258, NSF-1564167 and NSF-1559997. The work of Babak Hassibi has been supported in part by the National Science Foundation under grants CNS-0932428, CCF-1018927, CCF-1423663 and CCF-1409204, by a grant from Qualcomm Inc., by NASA's Jet Propulsion Laboratory through the President and Director's Fund, and by King Abdullah University of Science and Technology.

References

Chen, Z.; Zhang, K.; Chan, L.; and Schölkopf, B. 2014. Causal discovery via reproducing kernel hilbert space embeddings. *Neural Computation* 26:1484–1517.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3:507–554.

Clopper, C., and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26:404–413.

Etesami, J., and Kiyavash, N. 2016. Discovering influence structure. In *IEEE ISIT*.

Gao, W.; Kannan, S.; Oh, S.; and Viswanath, P. 2016. Conditional dependence via shannon capacity: Axioms, estimators and applications. In *ICML 2016*.

Granger, C. W. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* 424–438.

Hauser, A., and Bühlmann, P. 2012a. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* 13(1):2409–2464.

Hauser, A., and Bühlmann, P. 2012b. Two optimal strategies for active learning of causal networks from interventional data. In *Proceedings of Sixth European Workshop on Probabilistic Graphical Models*.

Hoyer, P. O.; Janzing, D.; Mooij, J.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. In *Proceedings of NIPS 2008*.

Hyttinen, A.; Eberhardt, F.; and Hoyer, P. 2013. Experiment selection for causal discovery. *Journal of Machine Learning Research* 14:3041–3071.

Janzing, D.; Mooij, J.; Zhang, K.; Lemeire, J.; Zscheischler, J.; Daniušis, P.; Steudel, B.; and Schölkopf, B. 2012. Information-geometric approach to inferring causal directions. *Artificial Intelligence* 182-183:1–31.

Kontoyiannis, I., and Skoularidou, M. Aug. 2016. Estimating the directed information and testing for causality. *IEEE Transactions on Information Theory* 62:6053–6067.

Kovacevic, M.; Stanojevic, I.; and Senk, V. 2012. On the hardness of entropy minimization and related problems. In *IEEE Information Theory Workshop*.

Lopez-Paz, D., and Oquab, M. 2016. Revisiting classifier two-sample tests. In *arXiv pre-print*.

Lopez-Paz, D.; Muandet, K.; Schölkopf, B.; and Tolstikhin, I. 2015. Towards a learning theory of cause-effect inference. In *Proceedings of ICML 2015*.

Lopez-Paz, D.; Muandet, K.; and Recht, B. Dec. 2015. The randomized causation coefficient. *Journal of Machine Learning Research* 16:2901–2907.

MCI. 2016. Munich workshop on causal inference and information theory 2016. https://www.lnt.ei.tum.de/events/munich-workshop-on-causal-inference-and-information-theory-2016/. Accessed: 2016 - 09 - 14.

Mooij, J. M.; Stegle, O.; Janzing, D.; Zhang, K.; and Schölkopf, B. 2010. Probabilistic latent variable models for distinguishing between cause and effect. In *Proceedings of NIPS 2010*.

Mooij, J. M.; Janzing, D.; Zscheischler, J.; and Schölkopf, B. 2016a. Cause effect pairs repository. [Online].

Mooij, M. J.; Peters, J.; Janzing, D.; Zscheischler, J.; and Scholkopf, B. 2016b. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research* 17(32):1–102.

Onn, S., and Weissman, I. 2011. Generating uniform random vectors over a simplex with implications to the volume of a certain polytope and to multivariate extremes. *Annals of Operations Research* 189:331–342.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Peters, J.; Janzing, D.; and Schölkopf, B. 2011. Causal inference on discrete data using additive noise models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33:2436–2450.

Quinn, C.; Kiyavash, N.; and Coleman, T. Dec. 2015. Directed information graphs. *IEEE Transactions on Information Theory* 61:6887–6909.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66 (5):688–701.

Schölkopf, B.; W. Hogg, D.; Wang, D.; Foreman-Mackey, D.; Janzing, D.; Simon-Gabriel, C.-J.; and Peters, J. 2015. Removing systematic errors for exoplanet search via latent causes. In *Proceedings of the 32 nd International Conference on Machine Learning*.

Shajarisales, N.; Janzing, D.; Schölkopf, B.; and Besserve, M. 2015. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of the 32 nd International Conference on Machine Learning*.

Shanmugam, K.; Kocaoglu, M.; Dimakis, A.; and Vishwanath, S. 2015. Learning causal graphs with small interventions. In *NIPS 2015*.

Shimizu, S.; Hoyer, P. O.; Hyvarinen, A.; and Kerminen, A. J. 2006. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7:2003—2030.

Spirtes, P.; Glymour, C.; and Scheines, R. 2001. *Causation, Prediction, and Search*. A Bradford Book.