Information-Theoretic Approximation to Causal Models

Peter Gmeiner
Global Data Science
GfK SE, Germany
peter.gmeiner@gfk.com

Abstract

Inferring the causal direction and causal effect between two discrete random variables X and Y from a finite sample is often a crucial problem and a challenging task. However, if we have access to observational and interventional data, it is possible to solve that task. If X is causing Y, then it does not matter if we observe an effect in Y by observing changes in X or by intervening actively on X. This invariance principle creates a link between observational and interventional distributions in a higher dimensional probability space. We embed distributions that originate from samples of X and Y into that higher dimensional space such that the embedded distribution is closest to the distributions that follow the invariance principle, with respect to the relative entropy. This allows us to calculate the best information-theoretic approximation for a given empirical distribution, that follows an assumed underlying causal model. We show that this information-theoretic approximation to causal models (IACM) can be done by solving a linear optimization problem. In particular, by approximating the empirical distribution to a monotonic causal model, we can calculate probabilities of causation. It turns out that this approximation approach can be used to successfully solve causal discovery problems in the bivariate, discrete case. Experimental results on both labeled synthetic and real-world data demonstrate that our approach outperforms other state-of-the-art approaches in the discrete case with low cardinality.

1 Introduction

Detecting causal relationships from data is a significant issue in many disciplines. The understanding of causal relations between variables can help to understand how a system behaves under intervention, stabilize future predictions, and has many other important implications. Identifying causal links (causal discovery) from observed data alone is only possible with further assumptions and/or additional data. Despite the various causal discovery methods available, the problem of finding the causal structure between two random variables remains notoriously hard. In this paper, we use additional data and assume a very natural principle to solve that task. Our work is based on a mathematical framework proposed by Pearl [Pea09], that formalizes causality and causal relations. It introduces *causal models* that represent an (unknown) underlying data generation mechanism responsible for the distribution of the sampled data [PJS17]. We include sampled data from situations (environments) where interventions took place together with samples from pure observations. Recent developments in that direction revealed promising results [PBM16, HDPM18], but often these methods are conservative, leading to situations where no direction is preferred. This paper focuses on the bivariate discrete case and is based on a natural and weak principle. The *principle of independent mechanism* assumes that the data generating mechanism is independent of the data that is feed into such a mechanism. From this principle, we derive an invariance relation that states that it does not matter if we observe an effect due to an observation of its cause or due to an intervention on its cause. Distributions

Preprint. Under review.

that are generated by an underlying causal model fulfill these invariance relations. If X and Y are discrete random variables, then we can characterize the support set of joint distributions that fulfill these relations by embedding the distributions from observational and interventional samples into a higher dimensional space and creating a link between them. That means we first embed the empirical distributions into a higher dimensional space and then find the best approximation of this embedding to the probability distributions that are compatible with the invariance principle such that the relative entropy between them minimizes. We call this approach an information-theoretic approximation to causal models (IACM) since the relative entropy can be interpreted as an error telling us how much a finite sample deviates from a sample that comes from an assumed underlying causal model. It turns out that solving this optimization problem is equivalent to solving a linear optimization problem, which ends up in an efficient algorithm. We use IACM to formulate a causal discovery algorithm that infers the causal direction of two random variables. For this, we approximate to a causal model were X causes Y and to a model were Y causes X. We prefer the direction that has lower relative entropy. With respective preprocessing, this can also be applied to continuous data.

If we additionally assume that the underlying causal model is monotonic w.r.t. X or Y, then we can include this assumption into the support set characterization used by our approach. In the case of binary random variables, an approximation to a monotonic causal model enables us to calculate probabilities about how necessary, sufficient, or necessary and sufficient a cause is for an effect as defined in [Pea09]. We will use this as a strength of a causal link and include it in our causal discovery algorithm.

For the rest of this paper, we assume that we have two random variables X and Y that attain values in finite ranges $\mathcal{X}_X$ and $\mathcal{X}_Y$, respectively. The contribution of this paper is twofold. The first contribution is an approximation of (empirical) distributions to a set of distributions that is compatible with an invariance condition induced by an assumed causal model. The second contribution is a method for causal discovery based on this approximation procedure. This method can also be applied if we have observed data from X and Y that are heterogeneous and continuous. In experiments, we were able to verify the strength of our causal discovery approach, especially in the case that we have discrete ranges with low cardinality, we outperformed alternative state-of-the-art methods.

The paper is organized as follows. Section 2 introduces causal models and formulates the invariance statement. In Section 3 we present an information-theoretic approximation of distributions to one that is generated by causal models. We derive the theoretic foundation, illustrate the results for the binary case, and formulate the approximation algorithm. Section 4 shows applications of the approximation algorithm. In particular, the calculation of probabilities for causes and the application to causal discovery. Section 5 describes experiments to verify our approach and we conclude in Section 6.

2 Causal Models

We describe causal relations in the form of a *directed graph* $G = (V, E)$ with a finite vertex set V and a set of directed edges $E \subset V \times V$. A *directed edge* from $u \in V$ to $v \in V$ is an ordered pair (u, v) and often represented as an arrow between vertices, e.g. $u \to v$. For a directed edge (u, v) the vertex u is a *parent* of v and v is a *child* of u. The set of parents of a vertex u is denoted by PA_u. We only consider directed graphs that have no cycles and call them *directed acyclic graphs* (DAGs). In a DAG we interpret the vertices as random variables $V = \{X_1, \ldots, X_n\}$ and a directed edge (X_i, X_j) as a causal link between X_i and X_j. We say that X_i is a *direct cause* of X_j and X_j is a *direct effect* of X_i. We further specify causal links by introducing functional relations between them.

Definition 1 *A* **structural causal model** *(SCM) is a tuple $\mathcal{C} := (S, P_N)$ where S is a collection of d structural assignments*

$$X_j := f_j(\mathrm{PA}_j, N_j), \qquad j = 1, \ldots, d,$$

where $\mathrm{PA}_j \subseteq \{X_1, \ldots, X_d\} \backslash \{X_j\}$ are the parents of X_j and $P_N = P_{N_1,\ldots,N_d}$ is a joint distribution over the noise variables N_j that are assumed to be jointly independent.

We consider an SCM as a model for a data generating process [PJS17]. This enables us to model a system in an observational state and under perturbations at the same time. An SCM defines a unique distribution $P_X^{\mathcal{C}}$ over the variables $X = (X_1, \ldots, X_d)$. Perfect *interventions* are done by replacing an assignment in an SCM. Given an SCM $\mathcal{C}$ we can replace the assignment for X_k by $X_k := \tilde{f}(\tilde{\mathrm{PA}}_k, \tilde{N}_k)$. The distribution of that new SCM $\tilde{\mathcal{C}}$ is denoted by $P_X^{\tilde{\mathcal{C}}} =: P_X^{\mathcal{C};\mathrm{do}(X_k := \tilde{f}(\tilde{\mathrm{PA}}_k, \tilde{N}_k))}$

and called *intervention distribution* [PJS17, Pea09]. When modeling causality, we assume the *principle of independent mechanism*. Roughly speaking, this principle states that a change in a variable does not change the underlying causal mechanism, see [PJS17]. Formally for an SCM, this means that a change in a child variable X will not change the mechanism f that is responsible to obtain an effect from X. From this principle the following invariance statement follows:

$$p^{\mathcal{C}}(x_j|x_{\mathrm{PA}_j}) = p^{\mathcal{C};\mathrm{do}(X_k:=x)}(x_j|x_{\mathrm{PA}_j}), \tag{1}$$

where $p^{\mathcal{C}}(x_j|x_{\mathrm{PA}_j})$ is the conditional density of $P^{\mathcal{C}}_{X_j|X_{\mathrm{PA}_j}=x_{\mathrm{PA}_j}}$ evaluated at x_j for some $k \neq j$. Informally, this means that, if X_k is a cause of X_j, then it doesn't matter if we observe x_j when x is present due to an observation of X_k or when x is present due to an intervention on X_k.

3 Approximation to Causal Models

3.1 The General Case

Given two random variables X, Y with finite ranges $\mathcal{X}_X, \mathcal{X}_Y$, and data from observations of X, Y as well as from interventions on X or Y.[1] We further assume that the data from different interventions are independent of each other, and that there is no confounding variable. In practical applications, the interventional data can be obtained from experiments or more implicitly from heterogeneous data. Condition (1), is in general, not fulfilled by empirical distributions obtained from such data. We derive a method that enables us to find a joint probability distribution of X and Y that fulfill (1) and is closest to a given empirical distribution in an information-theoretic sense.

Without loss of generality we assume that the intervention took place on X with values in $\mathcal{X}_X$ and $X \to Y$. In the following we assume that the elements of $\mathcal{X}_X$ are in a fixed order. We summarize $\mathbf{X} := (X, (X_a)_{a\in\mathcal{X}_X})$, $\mathbf{Y} := (Y, (Y_a)_{a\in\mathcal{X}_X})$, where X, Y are the observed data and $(X_a)_{a\in\mathcal{X}_X}, (Y_a)_{a\in\mathcal{X}_X}$ the interventional data. We define $V := \{X, Y\} \cup \bigcup_{a\in\mathcal{X}_X} Y_a$ that takes values in $\mathcal{X}_V := \mathcal{X}_X \times \mathcal{X}_Y \times \bigtimes_{a\in\mathcal{X}_X} \mathcal{X}_{Y_a}$ and with P_V we denote the joint distribution over V. The space of probability distributions on $\mathcal{X}_V$ is denoted by $\mathcal{P}(\mathcal{X}_V)$ and for $A \subset V$ the marginalization of $P \in \mathcal{P}(\mathcal{X}_V)$ is defined by $\pi_A : \mathcal{P}(\mathcal{X}_V) \to \mathcal{P}(\mathcal{X}_A)$ with $\pi_A(P)(x) := \sum_{y\in\mathcal{X}_{V\setminus A}} P(y, x)$, where $x \in \mathcal{X}_A$ and $\mathcal{X}_A := \bigtimes_{a\in A} \mathcal{X}_a$. The next Lemma gives us a characterization of distributions that fulfill (1).

Lemma 1 *The set of joint probability distributions for X, Y, Y_a, for all $a \in \mathcal{X}_X$ which fulfill the consistency condition (1) is called $\mathcal{M}_C$ and given as*

$$\begin{aligned}\mathcal{M}_C \quad = \quad & \{P \in \mathcal{P}(\mathcal{X}_V) \mid \pi_{X,Y,Y_a} P(a, y_a, \overline{y}_a) = \pi_{X,Y,Y_a} P(a, \overline{y}_a, y_a) = 0 \\ & \qquad \forall\, y_a \in \mathcal{X}_{Y_a}, a \in \mathcal{X}_X\}\,,\end{aligned}$$

where $\overline{y}_a \in \mathcal{X}_{Y_a}\setminus\{y_a\}$.

The proof is given in the Supplementary Material. The support of $\mathcal{M}_C$ is therefore given by

$$\mathrm{supp}(\mathcal{M}_C) = \mathcal{X}_V \setminus \bigcup_{\substack{y\in\mathcal{X}_Y, y_{x_i}\in\mathcal{X}_{Y_{x_i}}, y=y_{x_i}, \\ x_i\in\mathcal{X}_X, i\in\{1,\ldots,|\mathcal{X}_X|\}}} x_i \times y \times \mathcal{X}_{Y_{x_1}} \times \ldots \times \mathcal{X}_{Y_{x_{i-1}}} \times y_{x_i} \times \mathcal{X}_{Y_{x_{i+1}}} \times \ldots \times \mathcal{X}_{Y_{x_{|\mathcal{X}_X|}}}.$$

Given observational and interventional samples of X and Y and its corresponding empirical distributions $P_{X,Y}$, P_{Y_a} for $a \in \mathcal{X}_X$ we try to find a distribution $\hat{P} \in \mathcal{M}_C$ such that

$$\pi_{X,Y}\hat{P} = P_{X,Y}, \quad \text{and } \pi_{Y_a}\hat{P} = P_{Y_a} \text{ for } a \in \mathcal{X}_X. \tag{2}$$

We can always find a joint distribution $\hat{P} \in \mathcal{P}(\mathcal{X}_V)$ such that (2) holds, since the distributions P_{XY}, P_{Y_a} for all $a \in \mathcal{X}_X$ are independent to each other. Although this does not guarantee $\hat{P} \in \mathcal{M}_C$,

[1] We can also relax the assumption of having interventional data and assume that the data are heterogeneous and show a rich diversity. Alternatively, we can say that we have data of X and Y from different environments, where each environment belongs to a different intervention on X or Y.

we can try to find a distribution in $\mathcal{M}_C$ that has minimal relative entropy to $\hat{P}$. This minimal relative entropy can be interpreted as an approximation error to an assumed causal model. The *relative entropy* or *Kullback-Leibler divergence* (KL-divergence) between two distributions $P, Q \in \mathcal{P}(\mathcal{X}_V)$ is defined as follows

$$D(P||Q) := \begin{cases} \sum_{x \in \mathcal{X}_V} P(x) \log\left(\frac{P(x)}{Q(x)}\right), & \text{if } \operatorname{supp}(Q) \supseteq \operatorname{supp}(P), \\ \infty, & \text{else.} \end{cases}$$

We use the convention that $0 \log \frac{0}{q} = 0$ for $q > 0$, see also [CT91, Kak99]. This leads to:

$$\min_{\substack{\hat{P} \in \mathcal{P}(\mathcal{X}_V), \\ \pi_{X,Y}\hat{P} = P_{XY}, \pi_{Y_a}\hat{P} = P_{Y_a}, a \in \mathcal{X}_X}} \quad \min_{\tilde{P} \in \mathcal{M}_C} D(\tilde{P}||\hat{P}). \tag{3}$$

That is a nonlinear min-min optimization problem with linear constraints. It turns out that in our situation, the problem simplifies to a linear optimization problem.

Proposition 1 *The optimization problem (3) simplifies to the following linear optimization problem*

$$\max_{\substack{\hat{P} \in \mathcal{P}(\mathcal{X}_V), \\ \pi_{X,Y}\hat{P} = P_{XY}, \pi_{Y_a}\hat{P} = P_{Y_a}, a \in \mathcal{X}_X}} \quad S(\hat{P}),$$

with $S(\hat{P}) := \sum_{z \in \operatorname{supp}(\mathcal{M}_C)} \hat{P}(z)$.

The proof is given in the Supplementary Material and an application of the Lagrangian multiplier method. The statements of Proposition 1 holds also for any other support set characterization rather than $\mathcal{M}_C$. The global approximation error is given by $D(\tilde{P}||\hat{P}) = -\log(S(\hat{P}))$. Inspired by [JMZ$^+$12, DJM$^+$10, JBGWS13] and by the intuition that the information loss from $\pi_X \hat{P}$ to $\pi_Y \hat{P}$ should be smaller than the other way round (due to an assumed mechanism from X to Y) the quantity $D(\pi_X \hat{P}||\pi_Y \hat{P})$ could also be seen as a kind of approximation error to the causal model $X \to Y$.

3.2 The Binary Case

To illustrate our approach, we consider the binary case. That means $\mathcal{X}_X = \mathcal{X}_Y = \{0, 1\}$, and $V = \{X, Y, Y_0, Y_1\}$. The set of consistent probability distributions is characterized by

$$\mathcal{M}_C = \{P \in \mathcal{P}(\mathcal{X}_V) \mid P_{0010} = P_{0011} = P_{0100} = P_{0101} = P_{1001} = P_{1011} = P_{1100} = P_{1110} = 0\}$$

and therefore $\operatorname{supp}(\mathcal{M}_C) = \{0000, 0001, 0110, 0111, 1000, 1010, 1101, 1111\}$. A probability distribution $\hat{P} \in \mathcal{P}(\mathcal{X}_V)$ is a non-negative vector with 16 elements that sums up to 1. We encode the conditions (2) into a contraint matrix $\mathcal{C}$ that takes the following form

$$\mathcal{C} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and into a corresponding right-hand side

$$\begin{aligned} c \quad = \quad & (1, P_{Y_0}(Y_0 = 1), P_{Y_1}(Y_1 = 1), P_{X,Y}(X = 1, Y = 1), P_{X,Y}(X = 1, Y = 0), \\ & P_{X,Y}(X = 0, Y = 1)). \end{aligned}$$

The non-negativity can be encoded in an identity matrix $\mathbb{1}_{16}$ of length 16 and a zero vector 0_{16} of length 16 as the right-hand side. A probability distribution $\hat{P}$ that solves (2) is then a solution to the following linear optimization problem

$$\min S(\hat{P}) \ s.t. \ \mathcal{C} \cdot \hat{P} = c \text{ and } \mathbb{1}_{16} \cdot \hat{P} \geq 0_{16}.$$

The proof of Proposition 1 tells us that a distribution $\tilde{P}$ that fulfill condition (1) and is as close as possible to $\hat{P}$ in an information-theoretic sense can be obtained by the following re-weighting of $\hat{P}$

$$\tilde{P}(x) := \frac{\hat{P}(x)}{S(\hat{P})}, \quad \text{if } x \in \operatorname{supp}(\mathcal{M}_C) \quad \text{and} \quad \tilde{P}(x) := 0, \quad \text{if } x \notin \operatorname{supp}(\mathcal{M}_C).$$

3.3 Implementation

The procedure in Subsection 3.2 can be generalized for arbitrary finite ranges of X and Y. The pseudo-code of the algorithm is shown in Algorithm 1. The size of the finite ranges is denoted by $b_x := |\mathcal{X}_X|$ and $b_y := |\mathcal{X}_Y|$. We further assume that we have for every $x \in \mathcal{X}_X$ interventional data available. Therefore, the constraint matrix $\mathcal{C}$ has dimension $b_x(2b_y - 1) \times b_x b_y^{b_x+1}$. The first row of $\mathcal{C}$ contains 1 at each column, the following $b_x(b_y - 1)$ rows contain the support patterns of P_{Y_a} and the final $b_x b_y - 1$ rows contain the support pattern of P_{XY}. The function `getConstraintDistribution` prepares the right hand side of $\mathcal{C}$ accordingly. Since we assumed that the intervention took place on X the underlying assumed causal model is $X \to Y$. Note that $S(.)$ is depending on $\mathcal{M}_C$.

Algorithm 1 IACM(P, b_x, b_y, $\mathcal{M}_C$)

```
$\mathcal{C} \leftarrow$ createConstraintMatrix$(b_x, b_y)$
$c \leftarrow$ getConstraintDistribution$(P, b_x, b_y)$
Solve LP problem: $\min S(\hat{P})$ s.t. $\mathcal{C}\hat{P} = c$ and $\mathbb{1}_{b_x b_y b_y^{b_x}} \hat{P} \geq 0_{b_x b_y b_y^{b_x}}$
$\tilde{P}(x) \leftarrow \begin{cases} \frac{\hat{P}(x)}{S(\hat{P})}, & \text{for } x \in \text{supp}(\mathcal{M}_C), \\ 0, & \text{for } x \notin \text{supp}(\mathcal{M}_C). \end{cases}$
$D_{X \to Y} \leftarrow -\log(S(\hat{P}))$ or $D_{X \to Y} \leftarrow D(\pi_X \hat{P} || \pi_Y \hat{P})$ depending on the setting
return $\tilde{P}, D_{X \to Y}$
```

We implemented this procedure in Python and used the `cvxpy` package to solve the linear program. The dimension of $\mathcal{C}$ will grow exponentially with the size of ranges for X and Y. However, it turns out that it is enough to consider low range sizes $b_x \leq 4, b_y \leq 4$. For possibly preprocessed sample data of X and Y with much higher or continuous range sizes, we apply an equal-frequency discretization based on quantiles, that is done before calculating and feeding P into Algorithm 1.

4 Applications

The approximation approach described in Section 3 has several applications. We describe two of them in the following subsections.

4.1 Probabilities for Causes

To measure the effect of a cause-effect relation Pearl proposed in [Pea09] *counterfactual statements* that give information about the necessity, the sufficiency, and the necessity and sufficiency of cause-effect relations. A *counterfactual statement* is a do-statement in a hypothetical situation that can, in general, not observed or simulated. Formally this means we condition an SCM to an observed situation and apply a do-operator. The corresponding intervention distribution reads for example $P^{\mathcal{C}|(X,Y)=(1,1);\text{do}(X=0)}(Y = 0)$ which means the probability that Y equals 0 if X would have been 0 where indeed we observed that X is 1 and Y is 1.

Definition 2 *Let X, Y be random variables in an SCM $\mathcal{C}$ such that X is a (hypothetical) cause of Y and $x \in \mathcal{X}_X, y \in \mathcal{X}_Y$.*

- *The probability that $X = x$ is necessary as a cause for an effect $Y = y$ is defined as* $\text{PN}_{x \to y} := P^{\mathcal{C}|(X,Y)=(x,y);\text{do}(X \in \overline{x})}(Y \in \overline{y})$*, where* $\overline{x} = \mathcal{X}_X \setminus \{x\}$*.*
- *The probability that $X = x$ is sufficient as a cause for an effect $Y = y$ is defined as* $\text{PS}_{x \to y} := P^{\mathcal{C}|(X,Y) \in (\overline{x},\overline{y});\text{do}(X=x)}(Y = y)$*.*
- *The probability that $X = x$ is necessary and sufficient as a cause for an effect $Y = y$ is defined as* $\text{PNS}_{x \to y} := P(X = x, Y = y)\text{PN}_{x \to y} + P(X \in \overline{x}, Y \in \overline{y})\text{PS}_{x \to y}$.

In general, counterfactual statements cannot be calculated from observational data and without knowing the true underlying SCM. However, Pearl identified situations in which we can exploit the presence of observational and interventional data to calculate the probabilities defined above. One such situation is when the underlying SCM is monotonic.

Definition 3 *An SCM $\mathcal{C}$ with $Y := f(X, N_Y)$ for two random variables X and Y is called* monotonic *relative to X, if and only if f is monotonic in X independent of N_Y.*

If X and Y are binary and if Y is increasing monotonic relative to X, then Theorem 9.2.15 in [Pea09] give us

$$\mathrm{PN}_{1\to 1} = \frac{P(Y=1) - P^{\mathcal{C};\mathrm{do}(x=0)}(Y=1)}{P(Y=1, X=1)}, \tag{4}$$

$$\mathrm{PS}_{1\to 1} = \frac{P^{\mathcal{C};\mathrm{do}(x=1)}(Y=1) - P(Y=1)}{P(Y=0, X=0)}, \tag{5}$$

$$\mathrm{PNS}_{1\to 1} = P^{\mathcal{C};\mathrm{do}(x=1)}(Y=1) - P^{\mathcal{C};\mathrm{do}(x=0)}(Y=1). \tag{6}$$

Similar if Y is decreasing monotonic relative to X, then we could also derive in the same fashion as Pearl did it the following formulas

$$\mathrm{PN}_{0\to 1} = \frac{P^{\mathcal{C};\mathrm{do}(x=1)}(Y=0) - P(Y=0)}{P(Y=1, X=0)}, \tag{7}$$

$$\mathrm{PS}_{0\to 1} = \frac{P(Y=0) - P^{\mathcal{C};\mathrm{do}(x=0)}(Y=0)}{P(Y=0, X=1)}, \tag{8}$$

$$\mathrm{PNS}_{0\to 1} = P^{\mathcal{C};\mathrm{do}(x=0)}(Y=1) - P^{\mathcal{C};\mathrm{do}(x=1)}(Y=1). \tag{9}$$

By approximating empirical observational and interventional distributions to a monotonic causal model we can calculate $\mathrm{PN}_{x\to y}$, $\mathrm{PS}_{x\to y}$, and $\mathrm{PNS}_{x\to y}$. To do this, we need to further restrict the set $\mathcal{M}_C$ and note that the monotonicity of f implies that either $Y_0 = 1$ and $Y_1 = 0$ has zero probability or that $Y_0 = 0$ and $Y_1 = 1$ has zero probability. This means that either $P_{0110} = P_{1010} = 0$ or $P_{0001} = P_{1101} = 0$ has to hold in addition to the conditions given in $\mathcal{M}_C$. We define $\mathcal{M}_{M_i} := \{P \in \mathcal{M}_C | P_{0110} = P_{1010} = 0\}$ as the set of probability conditions with an underlying monotonic increasing data generation process and $\mathcal{M}_{M_d} := \{P \in \mathcal{M}_C | P_{0001} = P_{1101} = 0\}$ as the set of probability conditions with an underlying monotonic decreasing data generation process. An approximation in the sense of Subsection 3.1 to $\mathcal{M}_{M_d}$ or $\mathcal{M}_{M_i}$ instead of $\mathcal{M}_C$ will only change the definition of $S(P)$, the rest will remain the same. In order to calculate $\mathrm{PN}_{x\to y}$, $\mathrm{PS}_{x\to y}$, and $\mathrm{PNS}_{x\to y}$ we approximate to $\mathcal{M}_{M_d}$ and $\mathcal{M}_{M_d}$, choose the one with the least approximation error and use the formulas given above. We state the pseudo-code in Algorithm 2.

Algorithm 2 CalcCausalProbabilities(P)

```
$\tilde{P}_i, D_i \leftarrow \mathrm{IACM}(P, 2, 2, \mathcal{M}_{M_i})$
$\tilde{P}_d, D_d \leftarrow \mathrm{IACM}(P, 2, 2, \mathcal{M}_{M_d})$
if $D_i < D_d$ then
    Calculate PN, PS, PNS using $\tilde{P}_i$ and formulas (4) - (6)
else
    Calculate PN, PS, PNS using $\tilde{P}_d$ and formulas (7) - (9)
end if
return PN, PS, PNS
```

4.2 Causal Discovery

When we assume that $X \to Y$ we can test how well the given data fit that assumption and obtain an approximation error $D_{X\to Y}$. Switching the roles of X and Y we get $D_{Y\to X}$. The direction with the smallest error is the one we infer as the causal direction. If the error difference is below a small tolerance $\epsilon > 0$, we consider both directions as equal and return "no decision". If X and Y are binary and the error to the monotone models is smaller than to the non-monotone models, then we apply Algorithm 2 to determine PNS for both directions and use this as a decision criterion for the preferred direction (the direction with the higher PNS determines the direction). In general, some kind of data preprocessing and discretization before applying the causal discovery method is of advantage. In our implementation, we included several different preprocessing steps that treat X and Y different depending on the assumed causal direction, see Supplementary Material for more details.

Algorithm 3 IACMDiscovery(X, Y)

```
$\mathrm{data}_X, b_x \leftarrow$ preprocessing of $X, Y$ w.r.t. $X$
$\mathrm{data}_Y, b_y \leftarrow$ preprocessing of $X, Y$ w.r.t. $Y$
if $b_x = b_y = 2$ AND monotone model is preferred then
    use CalcCausalProbabilities to get PNS, $D_{X\to Y}, D_{Y\to X}$ for $X \to Y$ and $Y \to X$
    If $(D_{X\to Y} - D_{Y\to X}) < \epsilon$ then return direction with highest PNS
else
    $D_{X\to Y} \leftarrow \mathrm{IACM}(P_{\mathrm{data}_X}, b_x, b_y, \mathcal{M}_C)$
    $D_{Y\to X} \leftarrow \mathrm{IACM}(P_{\mathrm{data}_Y}, b_x, b_y, \mathcal{M}_C)$
    If $(D_{X\to Y} - D_{Y\to X}) < \epsilon$ then return no decision
end if
If $D_{X\to Y} < D_{Y\to X}$ then return $X \to Y$ else return $Y \to X$
```

5 Experiments

We test Algorithm 3 with synthetic and real-world benchmark data against alternative causal discovery methods. It runs with $b_x = b_y = 2$, and with a heuristic preprocessing procedure described in the Supplementary Material. Depending on the input data structure, we used $D(\pi_X \hat{P} || \pi_Y \hat{P})$ or $-\log(S(\hat{P}))$ as an approximation error, see also Supplementary Material for more details.

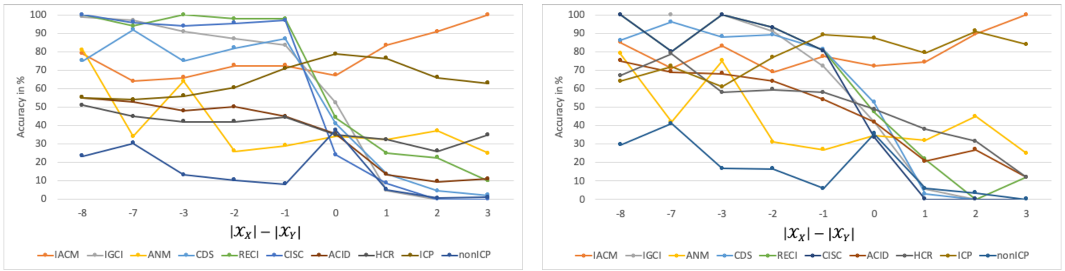
5.1 Pairwise Causal Discovery Methods

Among the various causal discovery approaches for continuous, discrete, nonlinear bivariate data, we select those that do not include any training of labeled cause-effect pairs to have a fair comparison. One well-known method uses additive noise models (ANM) that assume SCMs with additive noise and applies for continuous and discrete data [HJM$^+$09, PJS11]. Furthermore, we select an information-geometric approach (IGCI) [JMZ$^+$12] designed for continuous data and some recent methods for discrete data that are using minimal description length (CISC) [BV17], Shannon entropy (ACID) [BV18], and a compact representation of the causal mechanism (HCR) [CQZ$^+$18]. We further select conditional distribution similarity (CDS) [Fon19], regression error based causal inference (RECI) [BJW$^+$18], and (nonlinear) invariant causal prediction (nonICP, ICP) [HDPM18, PBM16] as baseline methods. For all methods we used the default parameter settings.[2]

5.2 Synthetic Data

We generate a set of synthetic data that are different in its structure (linear, nonlinear, discrete, non-discrete) and its range size. These synthetic data consists of observed data and data from perfect interventions. We use SCMs with additive or multiplicative noise to generate these data $X := N_X, Y := f(X) + N_Y$ or $Y := f(X) * N_Y$ where N_X, N_Y are sampled independently from a t-distribution for which the degrees of freedom are chosen randomly from $\{2, \ldots, 10\}$ and we randomly decide if we use an additive or multiplicative model. The nonlinear function f is randomly selected between the following functions $f_1(x) := \max(0, x)$, $f_2(x) := \sin(2\pi x)$, and $f_3(x) := \mathrm{sign}(x)\sqrt{|x|}$. The linear function is given by $f(x) := \alpha x$, where α is randomly selected from the interval $[-10, 10]$. The discrete data are generated using a k-bins discretization. We simulate perfect interventions on X by setting them to every value in the range if the range is discrete and to some randomly selected value if the range is continuous. The sample size is chosen randomly from $\{100, 500, 1000\}$ and we run 100 simulations for each configuration. Figure 1a shows the averaged accuracy of correct inferred causal direction for linear synthetic data relative to the difference in range size $|\mathcal{X}_X| - |\mathcal{X}_Y|$ for small range sizes. Our method performs substantially better for positive differences than all alternative approaches. A similar picture can be seen in Figure 1b for nonlinear synthetic data. Therefore, it seems that our method is well suited for situations where the range size of the cause is greater than the range size of the effect. In Figure 2a we see that our method performs also well for synthetic linear and nonlinear continuous data.

[2] For HCR, nonICP, and ICP we use the R-packages from the references, for CISC, ACID the corresponding Python code and for ANM, IGCI, CDS, RECI the Python package `causal discovery toolbox` [KG19].



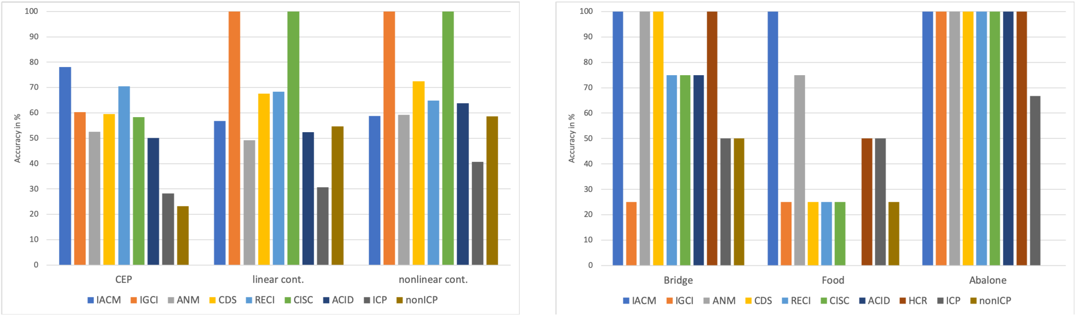
(a)

(b)

Figure 1: Averaged accuracies of correct inferred causal direction for linear (a) and nonlinear (b) synthetic data relative to the difference in range size $|\mathcal{X}_X| - |\mathcal{X}_Y|$ for small range sizes.

5.3 Real-World Data

As a benchmark set consisting of real-world data, we use 102 manually labeled continuous cause-effect pairs (CEP) from different contexts [MPJ+16, DG19]. As real-world discrete data sets, we use 4 anonymous discrete cause-effect pairs where food intolerances cause health issues (Food)[3], the Pittsburgh bridges dataset (Bridge, 4 pairs) from [DG19] as it has been used in [CQZ+18], and the Abalone dataset (Abalone, 3 pairs) from the UCI Machine Learning Repository [DG19]. For the CEP data set IACM outperforms all other methods. Also for discrete real-world data we can see in Figure 2b that IACM successfully recovers all causal directions and can keep pace with state-of-the-art methods.



(a) Accuracies of synthetic linear and nonlinear continuous data and of continuous real-world data (CEP).

(b) Accuracies of different discrete real-world data sets.

Figure 2: Accuracies of correct inferred causal directions for continuous and discrete data.

6 Discussion

In this paper, we proposed a way how empirical distributions coming from observations and experiments can be approximated to ones that follow the restrictions enforced by an assumed causal model. This can be used to calculate probabilities of causation and leads to a new causal discovery method. In our experiments, we could confirm that our approach can compete with the current state-of-the-art methods, also on real-world datasets (continuous and discrete) and without the explicit knowledge of experimental data. Especially for the discrete setting in which the range size of the cause is greater than the range size of the effect, our method has advantages compared to other approaches. Since IACM ran with small range sizes b_x and b_y, it seems that in many cases the essential cause-effect information can be encoded with much less information than we might have in the data. This is interesting by itself and could serve as a base for future research.

[3] This dataset, given as discrete timeseries data, has been provided by the author and the causal direction has been independently confirmed by medical tests.

Broader Impact

As all methods for causal discovery that use observational and/or data from implicit interventions the work in this paper could help to avoid unethical experiments. Furthermore, it contributes to a more relieable detection of causal relations, since it fills a gap in the existing causal discovery landscape. Therefore, our research can help during the evaluation and design of studies with few discrete features and contribute to more solid conclusions of those studies. On the other hand there is a potential risk that our method is used for data that are not following the assumptions of this paper. This may lead to wrong causal directions and to wrong conclusions, but can be avoided by checking the assumptions on the data before applying our method. Finally, it should be noted that this article may inspire future research projects in the field.

References

[BJW+18] P. Bloebaum, D. Janzing, T. Washio, S. Shimizu, and B. Schölkopf, *Cause-effect inference by comparing regression errors*, International Conference on Artificial Intelligence and Statistics (2018), 900–909.

[BV17] K. Budhathoki and J. Vreeken, *MDL for causal inference on discrete data*, 2017 IEEE International Conference on Data Mining (ICDM) (2017), 751–756.

[BV18] ______, *Accurate causal inference on discrete data*, 2018 IEEE International Conference on Data Mining (ICDM) (2018), 881–886.

[CLRS01] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, MIT Press, Cambridge, 2001.

[CQZ+18] R. Cai, J. Qiao, K. Zhang, Z. Zhang, and Z. Hao, *Causal discovery from discrete data using hidden compact representation*, Adv Neural Inf Process Syst (2018), 2666-2674.

[CT91] T. Cover and J. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.

[DG19] D. Dua and C. Graff, *UCI machine learning repository*, 2019.

[DJM+10] P. Daniušis, D. Janzing, J. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf, *Inferring deterministic causal relations*, Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (2010), 143–150.

[Fon19] J. Fonollosa, *Conditional distribution variability measures for causality detection*, Cause Effect Pairs in Machine Learning (I. Guyon, A. Statnikov, and B. Batu, eds.), Springer, 2019.

[HDPM18] C. Heinze-Deml, J. Peters, and N. Meinshausen, *Invariant causal prediction for nonlinear models*, Journal of Causal Inference **6** (2018), no. 2.

[HJM+09] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf, *Nonlinear causal discovery with additive noise models*, In Neural Information Processing Systems (NIPS) (2009), 689–696.

[JBGWS13] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, and B. Schölkopf, *Quantifying causal influences*, The Annals of Statistics **41** (2013), no. 5, 2324–2358.

[JMZ+12] D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf, *Information-geometric approach to inferring causal directions*, Artificial Intelligence **182-183** (2012), 1–31.

[Kak99] Y. Kakihara, *Abstract methods in information theory*, World Scientific Publishing Co. Pte. Ltd., 1999.

[KG19] D. Kalainathan and O. Goudet, *Causal discovery toolbox: Uncover causal relationships in python*.

[MPJ+16] J. Mooij, J. Peters, D. Janzing, J. Zscheischler, and S. B., *Distinguishing cause from effect using observational data: methods and benchmarks*, Journal of Machine Learning Research 17 (2016), 1–102.

[PBM16] J. Peters, P. Bühlmann, and N. Meinshausen, *Causal inference by using invariant prediction: identification and confidence intervals*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **78** (2016), no. 5, 947–1012.

[Pea09] J. Pearl, *Causality, models, reasoning, and inference*, Cambridge University Press, 2009.

[PJS11] J. Peters, D. Janzing, and B. Schölkopf, *Causal inference on discrete data using additive noise models*, IEEE Transactions on Pattern Analysis and Machine Intelligence **33** (2011), 2436–2450.

[PJS17] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference*, MIT Press, 2017.

Supplementary Material for Information-Theoretic Approximation to Causal Models

A Proofs

A.1 Proof of Lemma 1

Lemma 1

The set of joint probability distributions for X, Y, Y_a, for all $a \in \mathcal{X}_X$ which fulfill the consistency condition (1) is called $\mathcal{M}_C$ and given as

$$\begin{aligned}\mathcal{M}_C \quad = \quad & \{P \in \mathcal{P}(\mathcal{X}_V) \mid \pi_{X,Y,Y_a} P(a, y_a, \overline{y}_a) = \pi_{X,Y,Y_a} P(a, \overline{y}_a, y_a) = 0 \\ & \qquad \forall\, y_a \in \mathcal{X}_{Y_a}, a \in \mathcal{X}_X\},\end{aligned}$$

where $\overline{y}_a \in \mathcal{X}_{Y_a} \backslash \{y_a\}$.

Proof.

The consistency condition (1) implies the following relation for some $P \in \mathcal{P}(\mathcal{X}_V)$ and $a \in \mathcal{X}_X$

$$\pi_{X,Y} P(a, y_a) \quad = \quad \pi_{X,Y,Y_a} P(a, y_a, y_a).$$

These relation implies

$$\pi_{X,Y,Y_a} P(a, y_a, \overline{y}_a) = \pi_{X,Y,Y_a} P(a, \overline{y}_a, y_a) = 0, \text{ for } a \in \mathcal{X}_X,$$

which characterizes the joint distributions that satisfy (1). □

A.2 Proof of Proposition 1

Proposition 1

The optimization problem (3) simplifies to the following linear optimization problem

$$\max_{\substack{\hat{P} \in \mathcal{P}(\mathcal{X}_V), \\ \pi_{X,Y}\hat{P} = P_{XY}, \pi_{Y_a}\hat{P} = P_{Y_a}, a \in \mathcal{X}_X}} \qquad S(\hat{P}),$$

with $S(\hat{P}) := \sum_{z \in \mathrm{supp}(\mathcal{M}_C)} \hat{P}(z)$.

Proof.

We first consider the inner minimization problem of (3) for a given joint distribution $\hat{P} \in \mathcal{P}(\mathcal{X}_V)$. This is a constrained optimization problem where the constraints in $\mathcal{M}_C$ are equivalent to the equation

$$S(\tilde{P}) = 1,$$

since $\tilde{P}$ is a probability distribution. Therefore, the Lagrange functional of this minimization problem reads

$$\Lambda(\tilde{P}) := D(\tilde{P}||\hat{P}) + \lambda \left(S(\tilde{P}) - 1\right),$$

with λ as Lagrange multiplier. Using the Lagrange multiplier method we obtain explicit expressions for the approximating distribution $\tilde{P} \in \mathcal{M}_C$

$$\tilde{P}(z) = \frac{\hat{P}(z)}{S(\hat{P})}$$

for $z \in \operatorname{supp}(\mathcal{M}_C)$ and $\tilde{P}(z) = 0$ for all $z \notin \operatorname{supp}(\mathcal{M}_C)$. Thus we have solved the inner minimization problem explicitly and the relative entropy simplifies to

$$D(\tilde{P}||\hat{P}) = -\log(S(\hat{P})).$$

Therefore, we can now optimize on the space of possible joint distributions and (3) simplifies to

$$\max_{\substack{\hat{P}\in\mathcal{P}(\mathcal{X}_V),\\ \pi_{X,Y}\hat{P}=P_{XY},\pi_{Y_a}\hat{P}=P_{Y_a},a\in\mathcal{X}_X}} \log(S(\hat{P})).$$

Since $\log$ is a monotone function it suffices to maximize $S(\hat{P})$ given the constraints. But this is nothing than a linear optimization problem which can be solved by linear programming using the simplex algorithm, see, for example, [CLRS01]. □

B Application to Timeseries Data

Algorithm 1 can also be applied when we assume that the underlying causal model has a time lag of T, which is $X_{t-T} \to Y_t$, and the observational and interventional data have a time order. We only have to shift the incoming data for X_t and Y_t so that Algorithm 1 applies to X_t, Y_{t+T}, and has to take care that we preserve the data order during preprocessing steps. If we do not know the exact time lag we can run the approximation several times with different time lags to find the approximation with the lowest error.

C Experiments

C.1 Data Preprocessing

Algorithm 3 includes several preprocessing steps for sample data from X and Y that can be parametrized. The preprocessing steps split the data into obersvational data and interventional data for X and Y accordingly. These steps are:

- `discrete-cluster`: discretize data from X and Y using `KBinsDiscretizer` to a number of bins. Apply a k-means clustering to the discretized data with fixed number of clusters. We split the data according to the variance in the identified cluster. If X is the assumed cause, then the cluster with the lowest variance in X determines the observed data and the union of the other clusters the interventional data. If Y is the assumed cause, then we apply the same procedure by exchanging X with Y.
- `cluster-discrete`: the steps of `discrete-cluster` are interchanged. The data are first clustered, then discretized, and split according the variance in the clusters.

C.2 Parameter Setting in Experiments

In the experimental runs we used the following parameter configuration to run IACM. If $|\mathcal{X}_X| = 2, |\mathcal{X}_Y| < 5$ or $|\mathcal{X}_X| < 5, |\mathcal{X}_Y| = 2$ we used no preprocessing, since the binary discretization build into IACM itself seems sufficient enough. If $|\mathcal{X}_X| - |\mathcal{X}_Y| < 0$ we used `discrete-cluster` as preprocessing and used $D(\pi_X\hat{P}||\pi_Y\hat{P})$ as a decision criterion for the causal direction. In the other case when $|\mathcal{X}_X| - |\mathcal{X}_Y| \geq 0$ we used `discrete-cluster` but $-\log(S(\hat{P}))$ as decision criterion. For data with continuous or large range sizes we used `cluster-discrete` as preprocessing and the approximation error as decision criterion. The number of bins used in the discretization part is chosen such that it is a little above the range size of the data. The number of clusters is determined by a simple search procedure. We apply KMeans clustering for every number of clusters between 2 and the number of bins to the data and determine the relative entropies between the observational data and between the interventional data resulting from the cluster split as described above. The number of clusters with the lowest sum of those relative entropies is chosen for preprocessing. Furthermore, before feeding data into preprocessing we applied a robust scaling to them that is robust w.r.t. outliers. Our experiments also show that these heuristics defining the meta parameters for IACM is by far not optimal. We left that for future research.