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 of two discrete random variables X and Y from a finite sample is often a crucial problem and a challenging task in many disciplines. 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 can be formalized using causal models and creates a link between observational and interventional distributions in a higher dimensional probability space. Given finite samples of X and Y coming from observations and interventions on either X or Y, we embed the corresponding empirical distributions into that higher dimensional space such that the embedded distribution is closest to the set of distributions fulfilling the invariance principle with respect to the relative entropy. This allows us to calculate the best information-theoretic approximation for a given empirical distribution such that it follows an assumed underlying causal model. We show that this information-theoretic approximation to causal models (IACM) can be done by just solving a linear optimization problem. In particular, we can calculate certain probabilities of causation by approximating the empirical distribution to a monotonic causal model. It turns out that this simple approximation approach can also be applied to timeseries data and can be used to solve causal discovery problems in the bivariate, discrete setting. We run several experiments with labeled synthetic and real-world data and compare the performance with alternative methods designed for continuous and discrete data. In the discrete setting with low cardinality, our approach seems to be well-suited and outperform alternative state of the art approaches.

1 Introduction

Detecting causal relationships from data is a major 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 relations (causal discovery) from observed data alone is only possible with further assumptions and/or additional data. Pearl proposed in [Pea09] a mathematical framework that formalizes causality and causal relations. This framework allows a systematic study of causal models and their identification. That are models that represent an (unknown) underlying data generation mechanism that is responsible for the distribution of the sampled data [PJS17]. Despite the various causal discovery methods available the problem of finding the causal structure of random variables is still not sufficiently solved. In particular, the case with two random variables X and Y remains challenging and some methods fail to resolve the causal direction in that case. However, under certain assumptions, we can infer the correct causal direction. One example is to include sampled data from situations (environments) where interventions took place together with samples from observations. Recent developments in that

Preprint. Under review.

direction revealed promising results [PBM16, HDPM18], but often these methods are conservative leading to situations were 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 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 set of joint distributions that fulfill these relations by embedding them into a higher-dimensional space. This allows us to link distributions from observational and interventional samples and embed them into that higher-dimensional space. 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 is minimized. We call this approach an information-theoretic approximation to causal models (IACM). The relative entropy can be interpreted as an error that tells 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.

With respective preprocessing we can apply IACM also to continuous data. This allows us to formulate a causal discovery algorithm that infers the causal direction of two random variables. For this, we apply the approximation 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.

If we additionally assume that the underlying causal model is monotonic w.r.t. X or Y, then we can include this assumption into our approach. For binary random variables, Pearl defined counterfactual statements that give information about how necessary, sufficient, or necessary and sufficient a cause is for an effect [Pea09]. If the variables are generated by a monotonic causal model he also provides closed formulas for these measures. We can use this as a strength of a causal link by using the approximated distribution to a monotonic causal model. We also include the approximation to that more specific model into our causal discovery algorithm.

The contribution of this paper is twofold. Let us assume that we have two random variables X and Y that attain values in finite ranges $\mathcal{X}_X$ and $\mathcal{X}_Y$, respectively. If we know (or assume) for some reason the causal direction and have data from X, Y available that come from observations and experiments, then our first contribution is a new method for approximating empirical probability distributions such that they fulfill an invariance condition of the assumed causal model. With sophisticated preprocessing, this method can even then be applied if we only know that the data are heterogeneous. If we further assume monotonicity, then we can calculate probability measures for the assumed causal relation. The second contribution is a method for causal discovery that is based on this approximation procedure. By measuring the distance between the empirical probability distribution and its causal model approximation we can formulate a simple causal discovery algorithm. This causal discovery algorithm 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 in the case that we have discrete ranges with low cardinality. For that case, 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 for the general case, illustrate the results for the binary case, and formulate the approximation algorithm. Section 4 shows some applications of the approximation algorithm. In particular, we discuss the calculation of probabilities for causes, the application to timeseries data, and the application to causal discovery. Section 5 describes experiments to verify the approach and Section 6 concludes with a discussion.

2 Causal Models

In this section we introduce models to formally describe causality and causal relations. 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. In this paper 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 can further specify those causal links by introducing functional relations between parent and child vertices.

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

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

where $\mathrm{PA}_j \subseteq \{X_1, \ldots, X_L\} \backslash \{X_j\}$ are the parents of X_j and $P_N = P_{N_1,\ldots,N_L}$ 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* can be formalized 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 would mean 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 $p(x_j | x_{\mathrm{PA}_j})$ is the conditional density of $P_{X_j | X_{\mathrm{PA}_j} = x_{\mathrm{PA}_j}}$ evaluated at x_j for some $k \neq j$. Informally, this condition means that if there is a cause X_k which has X_j as an effect, then it doesn't matter if we observe x_j when x is present or if we observe x_j 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. We can also relax that assumption and assume instead that the data are heterogeneous and show a rich diversity. Formally we have observed data of X and Y and data from perfect interventions on X or Y denoted as X_a, Y_a when the intervention on X or Y was some $a \in \mathcal{X}_X$ or $a \in \mathcal{X}_Y$, respectively.[1] We further assume that the different interventional data are independent of each other. In practical applications, these interventional data can be obtained from experiments or more implicit from heterogeneous data.

Condition (1) is in general not fulfilled by empirical distributions that are obtained from such data. In this section, we derive a method that enables us to find a joint probability distribution of X and Y that fulfill the consistency condition (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 $\mathcal{X}_X = \{x_1, \ldots, x_d\}$, where $d := |\mathcal{X}_X|$. 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 $V := \{X, Y, Y_{x_1}, \ldots, Y_{x_d}\}$ that takes values in $\mathcal{X}_V := \mathcal{X}_X \times \mathcal{X}_Y \times \mathcal{X}_{Y_{x_1}} \times \ldots \times \mathcal{X}_{Y_{x_d}}$ 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 a probability distribution $P \in \mathcal{P}(\mathcal{X}_V)$ is defined by

[1] 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.

$\pi_A : \mathcal{P}(\mathcal{X}_V) \to \mathcal{P}(\mathcal{X}_A)$ with $\pi_A(P)(x) := \sum_{y \in \mathcal{X}_{V \setminus A}} P(y, x)$, where $x \in \mathcal{X}_A$ and $\mathcal{X}_A := \times_{a \in A} \mathcal{X}_a$. The next Lemma give us a characterization of distributions that fulfill condition (1).

Lemma 1 *The set of joint probability distributions for $X, Y, Y_{x_1}, \ldots, Y_{x_d}$ which fulfill the consistency condition (1) is called $\mathcal{M}_C$ and given as*

$$\begin{aligned} \mathcal{M}_C \quad = \quad & \big\{ P \in \mathcal{P}(\mathcal{X}_V) \mid \pi_{X,Y,Y_{x_i}} P(x_i, y, \overline{y}_{x_i}) = \pi_{X,Y,Y_{x_i}} P(x_i, \overline{y}, y_{x_i}) = 0 \\ & \qquad\qquad \forall\, y \in \mathcal{X}_Y, y_{x_i} \in \mathcal{X}_{Y_{x_i}}, i \in \{1, \ldots, d\} \big\}, \end{aligned}$$

where $\overline{y}_{x_i} \in \mathcal{X}_{Y_{x_i}} \setminus \{y_{x_i}\}$ for $i \in \{1, \ldots, d\}$ and $\overline{y} \in \mathcal{X}_Y \setminus \{y\}$.

Proof.

The consistency condition (1) implies the following relation for some $P \in \mathcal{P}(\mathcal{X}_V)$ and $i \in \{1, \ldots, d\}$

$$\pi_{X,Y} P(x_i, y) \quad = \quad \pi_{X,Y,Y_{x_i}} P(x_i, y, y_{x_i}), \qquad \text{with } y = y_{x_i}.$$

These relation implies

$$\pi_{X,Y,Y_{x_i}} P(x_i, y, \overline{y}_{x_i}) = \pi_{X,Y,Y_{x_i}} P(x_i, \overline{y}, y_{x_i}) = 0, \text{ for } i \in \{1, \ldots, d\},$$

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

The support of $\mathcal{M}_C$ is therefore given by

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

Given observation and intervention data of X and Y and its corresponding empirical distributions $P_{X,Y}$, $P_{Y_{x_i}}$ for $i \in \{1, \ldots, d\}$ we try to find a distribution $\hat{P} \in \mathcal{M}_C$ such that

$$\pi_{X,Y} \hat{P} = P_{X,Y}, \quad \text{ and } \pi_{Y_{x_i}} \hat{P} = P_{Y_{x_i}} \text{ for } x_i \in \mathcal{X}_X, i \in \{1, \ldots, d\}. \tag{2}$$

We can always find a joint distribution $\hat{P} \in \mathcal{P}(\mathcal{X}_V)$ such that (2) holds, since the distributions $P_{XY}, P_{Y_{x_i}}$ for all $x_i \in \mathcal{X}_X$ are independent to each other. But this does not guarantee that $\hat{P} \in \mathcal{M}_C$. Nevertheless, we can try to find a distribution in $\mathcal{M}_C$ that has minimal relative entropy to $\hat{P}$.[2] This minimal relative entropy can be interpreted as an approximation error to the 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

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

We use the convention that $0 \log \frac{0}{q} = 0$ for $q > 0$. The relative entropy is not symmetric, but it holds that $D(P||Q) \geq 0$, with equality if and only if $P = Q$. Although, it is not a distance measure it is an important quantity that appears in many contexts concerning information and probability theory, see [CT91, Kak99].

This lead us to the following optimization problem:

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

This is a nonlinear min-min optimization problem with linear constraints. But it turns out that in our situation the problem simplifies to a linear optimization problem.

[2] In other words we try to find the information-geometric projection of $\hat{P}$ to $\mathcal{M}_C$.

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

$$\max_{\substack{\hat{P}\in\mathcal{P}(\mathcal{X}_V),\\ \pi_{X,Y}\hat{P}=P_{XY},\pi_{Y_{x_i}}\hat{P}=P_{Y_{x_i}}}} S(\hat{P}),$$

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

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

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

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

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

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

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

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]. □

The global approximation error is given by $D_{X\to Y} := D(\hat{P}||\tilde{P}) = -\log(S(\hat{P}))$ and the local approximation errors are given by $D(P_{XY}||\pi_{X,Y}\hat{P})$, $D(P_{Y_{x_i}}||\pi_{Y_{x_i}}\hat{P})$ for all $i \in \{1, \ldots, d\}$.[3]

In practical applications, we can now test if an empirical joint distribution of X, Y is in $\mathcal{M}_C$ by solving the linear problem above. In the next subsection, we consider the simplest case where the ranges of X and Y are binary.

3.2 The Binary Case

To illustrate our approximation approach we consider the binary case. That means $d = 2$, $\mathcal{X}_X = \mathcal{X}_Y = \{0, 1\}$, 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

[3] Henceforth, we use the global approximation error since it bounds the local errors from above.

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

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 $\hat{P}$

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

The global approximation error is then quantified by $-\log(S(\hat{P}))$.

3.3 Implementation

The procedure explained in the previous subsection 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 $P_{Y_{x_i}}$ and the final $b_x b_y - 1$ rows contain the support pattern of P_{XY}. This is done by the function `createConstraintMatrix`. 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 $\mathbb{1}_{b_x b_y b_y^{b_x}} \hat{P} \geq 0_{b_x b_y b_y^{b_x}}$
$\tilde{P}(x) \leftarrow \begin{cases} \frac{\hat{P}(x)}{S(\hat{P})}, & \text{for } x \in \operatorname{supp}(\mathcal{M}_C), \\ 0, & \text{for } x \notin \operatorname{supp}(\mathcal{M}_C). \end{cases}$
$D_{X \to Y} \leftarrow -\log(S(\hat{P}))$
return $\tilde{P}, D_{X \to Y}$
```

We implemented this procedure in Python and used the `cvxpy` package to solve the linear program.[4] The dimension of $\mathcal{C}$ will grow exponentially in the size of ranges for X and Y. However, we will see in Section 5 that already for small ranges we get satisfactory results.

4 Applications

The approximation approach described in the previous section has several applications. We describe three of them in the following subsections.

[4] We provide the code for the presented algorithms at `www.github.com/???`.

4.1 Probabilities for Causes

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 you can in general not observe or simulate. Formally this means we condition an SCM to an observed situation and execute a do-operation in the conditioned SCM. The corresponding intervention distribution reads for example $P^{\mathcal{C}|(X,Y)=(1,1);\mathrm{do}(X=0)}(Y=0)$ 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*

$$\mathrm{PN}_{x\to y} := P^{\mathcal{C}|(X,Y)=(x,y);\mathrm{do}(X\in\overline{x})}(Y \in \overline{y}),$$

where $\overline{x} = \mathcal{X}_X \backslash \{x\}$.

- *The probability that $X = x$ is sufficient as a cause for an effect $Y = y$ is defined as*

$$\mathrm{PS}_{x\to y} := P^{\mathcal{C}|(X,Y)\in(\overline{x},\overline{y});\mathrm{do}(X=x)}(Y = y).$$

- *The probability that $X = x$ is necessary and sufficient as a cause for an effect $Y = y$ is defined as*

$$\mathrm{PNS}_{x\to y} := P(X = x, Y = y)PN_{x\to y} + P(X \in \overline{x}, Y \in \overline{y})PS_{x\to y}.$$

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

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

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

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

With data from an observational and interventional setting we can extend the approximation approach in subsection 3.1 such that we can calculate $\mathrm{PN}_{x\to y}$, $\mathrm{PS}_{x\to y}$, and $\mathrm{PNS}_{x\to y}$. For this we have to further restrict the set $\mathcal{M}_C$. We first note that the monotonicity of f implies that either the probability that $Y_0 = 1$ and $Y_1 = 1$ is zero or the probability that $Y_0 = 0$ and $Y_1 = 1$ is zero. These two cases translate that either $P_{0110} = P_{1010} = 0$ or $P_{0001} = P_{1101} = 0$ in addition to the conditions given in $\mathcal{M}_C$. We therefore get two sets of probability distributions that fulfill the consistency condition

(1) and follow a data generation process that is monotone in one of two directions. 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 $\text{PN}_{x\to y}$, $\text{PS}_{x\to y}$, and $\text{PNS}_{x\to y}$ we approximate to $\mathcal{M}_{M_d}$ and $\mathcal{M}_{M_d}$ choose the one with the least approximation error and calculate it with the formulas given above. We state the pseudo-code of this in Algorithm 2.

Algorithm 2 CalcCausalProbabilities(P)

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

4.2 Timeseries Data

The approximation method can also be applied when we assume that the underlying causal model has a time lag T, that is $X_{t-T} \to Y_t$, and the observational and interventional data are ordered by time. 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 have to take care that we preserve the order in the data 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. This is a first step in the direction of identifying the actual underlying causal model using the approximation method.

4.3 Causal Discovery

The approximation method described in subsection 3.1 can be used to formulate a causal discovery algorithm. When we assume that $X \to Y$ we can test how well the given data fit that assumption and obtain an approximation error $D_{X\to Y}$. When we switch the roles of X and Y we get $D_{Y\to X}$ and can compare them. The direction with the smallest error is the one that is assumed to be the causal direction. If the difference between these errors is below a small tolerance $\epsilon > 0$ we consider both directions as equal and return that there is no decision. If we are in a binary situation and the error to the monotone models is smaller that to the non-monotone models, then we can 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). However, depending on the data it could be necessary to first discretize these datasets to a fixed alphabet size. In general, some kind of data preprocessing before applying the causal discovery method is of advantage. In our implementation, we included several different data preprocessing steps suitable for different situations.[5] Depending on the causal direction we want to test, the preprocessing treats the variables X and Y different and apply for example sorting on either X or Y.

5 Experiments

We test Algorithm 3 with synthetic and real-world benchmark data against alternative causal discovery methods for continuous and discrete data.

[5] For example we use KMeans clustering in order to further discretize the data and split the data according to the variance in the identified cluster.

Algorithm 3 IACMDiscovery(X, Y, b_x, b_y)

```
$\mathrm{data}_X \leftarrow$ preprocessing of $X, Y$ w.r.t. $X$
$\mathrm{data}_Y \leftarrow$ 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 $Y \to X$
```

5.1 Pairwise Causal Discovery Methods

There are many causal discovery approaches for the continuous, discrete, nonlinear bivariate case. We select those that do not include any training of labeled cause-effect pairs to have a fair comparison. A well-known method uses additive noise models (ANM) that assume SCMs with additive noise and can be applied for continuous and discrete data [HJM$^+$09, PJS11]. Furthermore, we select an information-geometric approach (IGCI) [JMZ$^+$12] designed for continuous data and some recent methods designed for discrete data that make use of minimal description length (CISC) [BV17], Shannon entropy (ACID) [BV18], and of a compact representation of the causal mechanism (HCR) [CQZ$^+$18]. We also use some more continuous methods namely conditional distribution similarity (CDS) [Fon19], regression error based causal inference (RECI) [BJW$^+$18], and (nonlinear) invariant causal prediction (nonICP, ICP) [HDPM18, PBM16] as baseline methods.[6]

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 that come from perfect interventions. We use the following SCM with additive noise to generate these data $X := N_X, Y := f(X) + N_Y$ where N_X, N_Y are independently sampled from a t-distribution for which the degrees of freedom are chosen at random from $\{2, \ldots, 10\}$. We vary the nonlinear functions f randomly 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 $\{100, 500, 1000\}$.

Figure 1 shows the averaged accuracy of correct inferred causal direction for each method we took into account relative to the difference in alphabet size $|\mathcal{X}| - |\mathcal{Y}|$. Our method performs substantially better for non-negative differences than all alternative approaches except for ICP. A similar picture can be seen in Figure 2 where the averaged accuracy is shown for nonlinear synthetic data. Therefore, it seems that our method is better suited for situations were the alphabet size of the cause is greater or equal to the alphabet size of the effect.

5.3 Real-World Data

As a benchmark set that consists of real-world data, we use a database of continuous cause-effect pairs that consists of manual labeled cause-effect pairs (CEP) from different contexts [MPJ$^+$16, DG19]. For all these pairs the ground truth is known and it is assumed that the pairs are not influenced by a confounder. As real-world discrete data sets we use anonymous discrete cause-effect pairs were food

[6] 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].

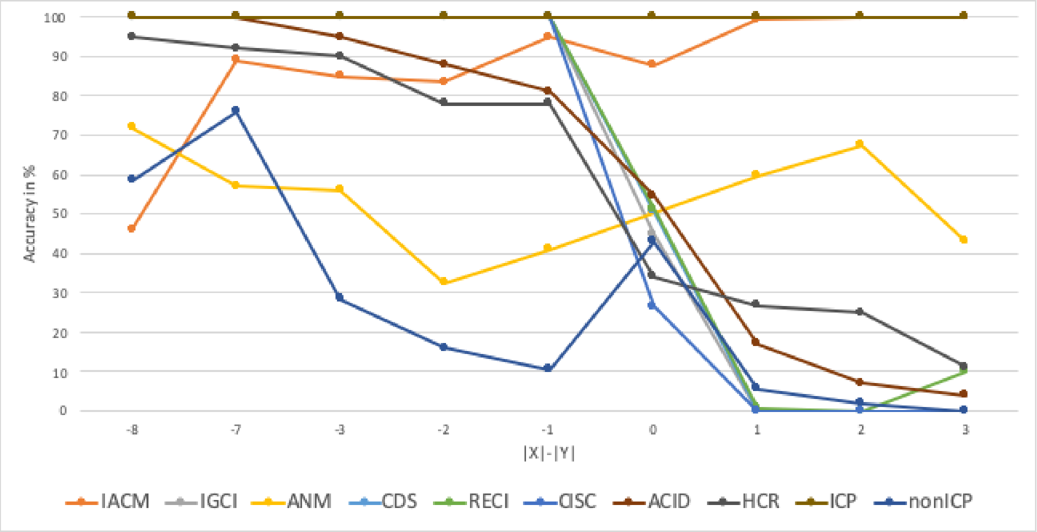


Figure 1: Averaged accuracy of correct inferred causal direction for linear synthetic data relative to the difference in alphabet size $|\mathcal{X}| - |\mathcal{Y}|$ for small range sizes.

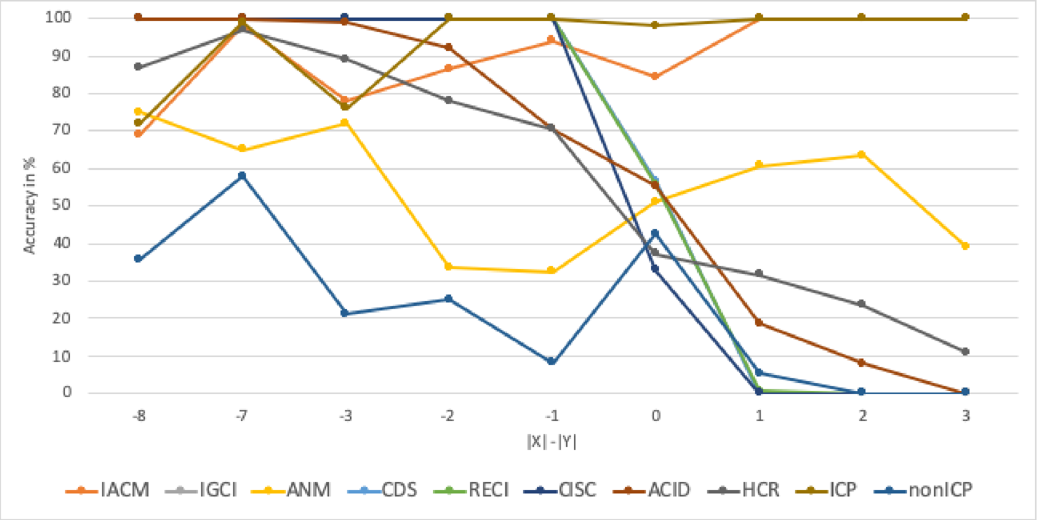
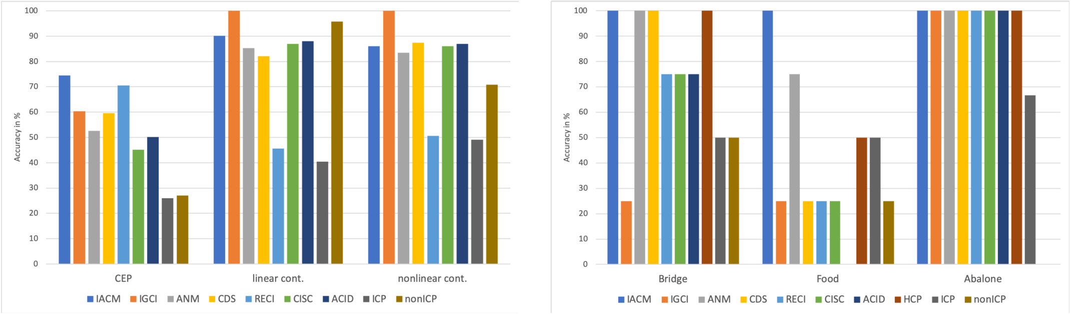


Figure 2: Averaged accuracy of correct inferred causal direction for nonlinear synthetic data relative to the difference in alphabet size $|\mathcal{X}| - |\mathcal{Y}|$ for small range sizes.

intolerances causes health issues (Food)[7], the Pittsburgh bridges dataset (Bridge) from [DG19] as it has been used in [CQZ$^+$18], and the Abalone dataset (Abalone) from the UCI Machine Learning Repository [DG19].

In Figure 3 (a) we see that our method performs well for synthetic linear and nonlinear continuous data. For the CEP data set it outperforms all other methods. Also for discrete real-world data we can see in Figure 3 (b) that our methods successfully can recover all causal directions and can keep pace with state of the art methods.

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



(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 3: Accuracies of correct inferred causal directions for continuous and discrete data.

6 Discussion

In this paper we proposed a way how empirical distributions that come from observations and experiments can be approximated to ones that follow the restrictions enforced by an assumed causal model. This approximated distribution can then be used to calculate probabilities of causation and together with the approximation error this 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 even on real-world datasets (continuous and discrete) and without the explicit knowledge of experimental data. Especially, for the discrete setting in which the alphabet size of the cause is greater or equal than the alphabet size of the effect our method has advantages compared to other approaches. ICP was the only method that performed better for synthetic data sets. However, for real-world data sets ICP seems too conservative and IACM was more reliable there. With a sophisticated preprocessing we also succeeded to apply that to the continuous setting with small b_x and b_y. Therefore, it seems that in many cases we can encode the essential cause-effect information with much less information than we might have available in the data. This is interesting by itself and could serve as a base for future research.

References

[BJW$^+$18] P. Bloebaum, D. Janzing, T. Washio, S. Shimizu, and B. Schoelkopf, *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.

[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. Scholkopf, *Nonlinear causal discovery with additive noise models*, In Neural Information Processing Systems (NIPS) (2009), 689–696.

[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. Scholkopf, *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.