A Ladder of Causal Distances

Maxime Peyrard and **Robert West**

EPFL

{maxime.peyrard, robert.west}@epfl.ch

Abstract

Causal discovery, the task of automatically constructing a causal model from data, is of major significance across the sciences. Evaluating the performance of causal discovery algorithms should ideally involve comparing the inferred models to ground-truth models available for benchmark datasets, which in turn requires a notion of distance between causal models. While such distances have been proposed previously, they are limited by focusing on graphical properties of the causal models being compared. Here, we overcome this limitation by defining distances derived from the causal distributions induced by the models, rather than exclusively from their graphical structure. Pearl and Mackenzie [2018] have arranged the properties of causal models in a hierarchy called the "ladder of causation" spanning three rungs: observational, interventional, and counterfactual. Following this organization, we introduce a hierarchy of three distances, one for each rung of the ladder. Our definitions are intuitively appealing as well as efficient to compute approximately. We put our causal distances to use by benchmarking standard causal discovery systems on both synthetic and real-world datasets for which ground-truth causal models are available.

1 Introduction

Reasoning about the causes and effects driving physical and societal phenomena is an important goal of science. Causal reasoning facilitates the prediction of intervention outcomes and can ultimately lead to more principled policymaking [Spirtes *et al.*, 2000; Pearl and Mackenzie, 2018].

Given a causal model, reasoning about cause and effect corresponds to formulating causal queries, which have been organized by Pearl and Mackenzie [2018] in a three-level hierarchy termed the "ladder of causation": *observational* queries correspond to seeing and observing; *interventional* queries correspond to acting and intervening; and *counterfactual* queries correspond to imagining, reasoning, and understanding.

Asking such questions requires a causal model to begin with. Inferring a causal model from observational, interventional, or mixed data is the problem called *causal discovery*.

Much of science is concerned with causal discovery, and automating the task has been receiving increased attention in the machine learning community [Peters *et al.*, 2017], where causal models have become the tool of choice for tackling important problems such as transfer learning, generalization beyond spurious correlations [Rojas-Carulla *et al.*, 2018], and algorithmic fairness [Kusner *et al.*, 2017].

Evaluating causal discovery algorithms requires comparing the inferred causal models to ground-truth models on benchmark datasets, which in turn requires a notion of distance between causal models. When defining such a distance, it is not sufficient to rely on tools developed for comparing standard generative models, such as goodness of fit, as these tools only operate on the first, observational level of the ladder of causation. Remarkably little research has been done on the topic of evaluating and comparing arbitrary causal models higher up the ladder. Existing works focus on specific aspects, such as the outcome of a limited number of interventions manually selected in advance [Singh *et al.*, 2017] or the graph structure of the models being compared [Peters and Bühlmann, 2015]. The latter work, which proposed the *structural intervention distance* (SID), one of the most prominent causal distance measures, further assumes that the two models have an identical observational joint distribution. Unfortunately this assumption rarely holds in practice, and we show that even if the observational joint distributions are just slightly different, SID cannot be trusted. Furthermore, previous causal distances do not cover the counterfactual level.

To close this gap, we introduce three distances (Sec. 4), one for each rung of the ladder of causation. Our distances measure the difference between causal models for each type of causal query (observational, interventional, counterfactual). Each distance builds upon the distance one level below, thus mirroring the hierarchy of the ladder of causation. We highlight theoretical properties of the distances in relation to previously proposed distances (Sec. 5). Then, we study their behavior in a series of experiments and put them to use in evaluating existing causal discovery systems (Sec. 6). Code for reproducing our experiments[1] and an extended version of the paper (with added appendices)[2] are available online.

---

[1] https://github.com/epfl-dlab/causal-distances

[2] https://arxiv.org/abs/2005.02480

2 Preliminaries

2.1 Causal Graphs

We consider a finite ordered set of random variables $\mathbf{X} = \{X_1, \ldots, X_d\}$. A directed graph $\mathscr{G} = (\mathbf{X}, \mathbf{A})$ consists of the set of indexed nodes $\mathbf{X}$ together with a set of directed edges $\mathbf{A} \subseteq \mathbf{X} \times \mathbf{X}$. If $(X_i, X_j) \in \mathbf{A}$, we say that X_i is a *parent* of X_j and denote the set of all parents of X_j with $\mathbf{PA}_j$. If $\mathscr{G}$ contains no directed cycle it is called a *directed acyclic graph* (DAG).

DAGs are often used to encode causal assumptions by viewing an edge (X_i, X_j) as the statement "X_i is a direct cause of X_j" [Pearl, 2009]. A graph associated with such causal interpretation is called a *causal graph*.

2.2 Structural Causal Models (SCMs)

A *structural causal model* $\mathfrak{C}$ is a tuple $(\mathbf{X}, \mathbf{N}, \mathbf{F}, P_{\mathbf{N}})$, where $P_{\mathbf{N}}$ is a *noise distribution* over the (exogenous) noise variables $\mathbf{N}$ and $\mathbf{F} = \{f_1, \ldots, f_d\}$ is an ordered set of *structural equations* indicating, for each $X_i \in \mathbf{X}$, how its value is determined by its parents and noise:

$$X_i := f_i(\mathbf{PA}_i, N_i), \qquad (1)$$

where $\mathbf{PA}_i \subseteq \mathbf{X}$, and N_i is the noise variable associated with X_i. The noise models variations due to ignored variables or inherent randomness. We assume that the noise variables are independent [Pearl, 2009]. The associated causal graph $\mathscr{G}$ is obtained by viewing each variable in $\mathbf{X}$ as a vertex and drawing an arrow from each parent in $\mathbf{PA}_i$ to X_i.

Assumptions. Throughout the paper, we assume that all models satisfy four standard assumptions: the *Markov property*, *causal minimality*, *causal faithfulness*, and *positiveness* [Peters *et al.*, 2017].

2.3 Observational, Interventional, and Counterfactual Distributions

Fig. 1 is a graphical illustration of a three-variable SCM with queries about the observational, interventional, and counterfactual distributions. We define these distributions next.

Observational. A causal model $\mathfrak{C}$ entails a unique joint distribution of $\mathbf{X} = (X_1, \ldots, X_d)$ called the *observational distribution* and noted $P_{\mathbf{X}}^{\mathfrak{C}}$ [Peters *et al.*, 2017]. To sample from $\mathfrak{C}$, we can simply sample from the noise distribution $P_{\mathbf{N}}$ and use the structural assignments (cf. Eq. 1) following the topological order of $\mathbf{X}$ in the causal graph $\mathscr{G}$.

Interventional. An intervention on the set of variables $\mathbf{I} \subset \mathbf{X}$ of the causal model $\mathfrak{C}$ consists of replacing the structural assignments (cf. Eq. 1) of variables in $\mathbf{I}$ by forcing them to specific values $\mathbf{I} = \mathbf{i}$, a so-called *hard intervention*.[3] The new causal model obtained from $\mathfrak{C}$ via the intervention $\mathbf{I} = \mathbf{i}$ is denoted by $\mathfrak{C}; \mathrm{do}(\mathbf{I} = \mathbf{i})$ [Pearl, 2009]. Graphically, the *interventional model* is obtained by removing all incoming edges to the nodes in $\mathbf{I}$. After sampling the noise and following the new structural assignments, we obtain samples from the *interventional distribution* of $\mathbf{X}$, denoted by $P_{\mathbf{X}}^{\mathfrak{C}; \mathrm{do}(\mathbf{I} = \mathbf{i})}$.

---

[3]For simplicity, we focus on hard intervention. However, the approach can easily be extended to soft interventions.

Counterfactual. At the counterfactual level, we first (partially) observe the causal model in some state $\mathbf{E} = \mathbf{e}$, where $\mathbf{E} \subseteq \mathbf{X}$ is called the *evidence set.* Then we ask: "Given that $\mathbf{E} = \mathbf{e}$ actually happened, what would have happened had we done the intervention $\mathbf{I} = \mathbf{i}$?" This is different from the interventional level, where we only ask: "In general, what happens if we do $\mathbf{I} = \mathbf{i}$?" We now take into account the additional specific information provided by the evidence $\mathbf{E} = \mathbf{e}$.

Consider the causal model $\mathfrak{C}$ with noise distribution $P_{\mathbf{N}}$ for which we have some evidence $\mathbf{E} = \mathbf{e}$. The *counterfactual model* induced by $\mathfrak{C}$ and $\mathbf{E} = \mathbf{e}$ is denoted by $\mathfrak{C}|\mathbf{E} = \mathbf{e}$ and is identical to $\mathfrak{C}$ except for the noise distribution $P_{\mathbf{N}|\mathbf{E}=\mathbf{e}}$ which has been updated given the evidence using Bayes' rule [Pearl, 2009]:

$$P_{\mathbf{N}|\mathbf{E}=\mathbf{e}}(\mathbf{n}) = \frac{P_{\mathbf{E}|\mathbf{N}=\mathbf{n}}(\mathbf{e})}{P_{\mathbf{E}}(\mathbf{e})} P_{\mathbf{N}}(\mathbf{n}). \qquad (2)$$

The updated noise variables are not necessarily independent anymore. Note the difference in notation between the induced counterfactual model $\mathfrak{C}|\mathbf{E} = \mathbf{e}$ and the induced interventional model $\mathfrak{C}; \mathrm{do}(\mathbf{I} = \mathbf{i})$. The former corresponds to updating the noise distribution, whereas the latter corresponds to modifying the structural assignments of variables $\mathbf{I}$.

A counterfactual query corresponds to an intervention $\mathrm{do}(\mathbf{I} = \mathbf{i})$ in the counterfactual model $\mathfrak{C}|\mathbf{E} = \mathbf{e}$. Again, this intervention entails a distribution of $\mathbf{X}$, called the *counterfactual distribution* and denoted by $P_{\mathbf{X}}^{\mathfrak{C}|\mathbf{E}=\mathbf{e}; \mathrm{do}(\mathbf{I}=\mathbf{i})}$.

2.4 Metrics, Pseudometrics, and Premetrics

A *metric* d satisfies the four axioms of *non-negativity* ($d(x, y) \geq 0$), *identity of indiscernibles* ($x = y \iff d(x, y) = 0$), *symmetry* ($d(x, y) = d(y, x)$), and the *triangle inequality* ($d(x, z) \leq d(x, y) + d(y, z)$). The causal distances introduced in this paper are pseudometrics (i.e., they relax the identity of indiscernibles), whereas SID (Sec. 1 and 3) is a premetric (i.e., only non-negativity and $x = y \implies d(x, y) = 0$ hold).

3 Related Work

An important practical application of causal-model distances is the evaluation of causal discovery techniques. Ideally, one would like to compare an inferred causal model against a given ground-truth model for each type of causal query: observational, interventional, and counterfactual.

The comparison of observational distributions has been studied extensively in machine learning and statistics (cf. the overviews by Theis *et al.* [2015] and Sriperumbudur *et al.* [2010]), and distances between distributions have been used to evaluate causal discovery methods, typically by measuring the goodness of fit of the observational distribution induced by a model with respect to empirical samples from the true observational distribution (cf. Singh *et al.* [2017] for an overview). Importantly, such methods are inherently limited to the observational level and cannot measure how well the inferred causal model performs at the interventional and counterfactual levels.

These two levels have received relatively little attention, compared to the observational one. We are not aware of any previously proposed causal-model distance to consider the counterfactual level, and the few that consider the interventional level focus on a specific aspect of causal models: their
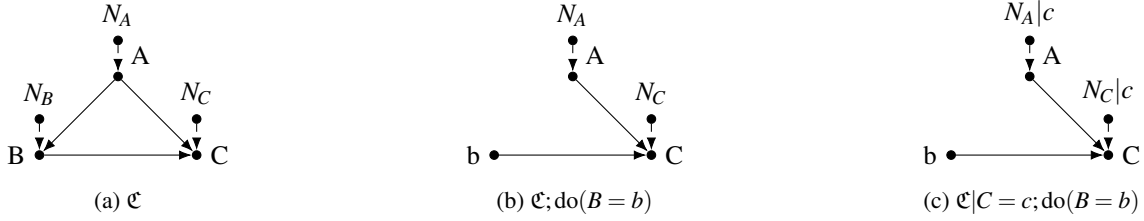
(a) $\mathfrak{C}$

(b) $\mathfrak{C}; \mathrm{do}(B = b)$

(c) $\mathfrak{C}|C = c; \mathrm{do}(B = b)$

Figure 1: Fig. 1a depicts an SCM over the variables $\mathbf{X} = \{A, B, C\}$. Sampling the noise variables $\mathbf{N}$ and following the structural assignments in topological order gives samples from the observational distribution. In Fig. 1b, the intervention $\mathrm{do}(B = b)$ replaces the structural assignment of B by the hard value b. Samples from this modified SCM are samples from the interventional distribution. Fig. 1c asks *what would have happened had we performed* $\mathrm{do}(B = b)$ *given that we actually observed* $C = c$? This counterfactual is obtained by updating the noise distribution and then performing $\mathrm{do}(B = b)$. Samples from this model are samples from the counterfactual distribution.

causal graphs [de Jongh and Druzdzel, 2009]. For instance, the popular *structural Hamming distance* (SHD) [Acid and de Campos, 2003] counts in how many edges the two input graphs differ. Peters and Bühlmann [2015] argue that previous graph comparison metrics, including SHD, are not in line with the end goal of causal discovery, namely, predicting the outcome of interventions, and propose the *structural intervention distance* (SID), a premetric (cf. Sec. 2.4) that counts the number of pairwise interventional distributions on which two causal models (with graphs $\mathscr{G}$ and $\mathscr{H}$, respectively) disagree:

$$\mathrm{SID}(\mathscr{G}, \mathscr{H}) \quad = |\{(X_i, X_j) \in \mathbf{X}^2 | \; P(X_i | \mathrm{do}(X_j)) \text{ is falsely} \quad (3)$$
$$\text{inferred in } \mathscr{H} \text{ with respect to } \mathscr{G}\}|. \quad (4)$$

Under the assumption that the causal models agree on an underlying observational distribution, Peters and Bühlmann [2015] show that the comparison of interventional distributions reduces to a purely graphical criterion. In particular, when the graphs are the same, both SHD and SID are 0, and $\mathrm{SHD}(\mathscr{G}, \mathscr{H}) = 0$ implies $\mathrm{SID}(\mathscr{G}, \mathscr{H}) = 0$.

Acharya *et al.* [2018] compare two causal models, but only for the purpose of testing identity. This constitutes a special case of our interventional distance. Recently, Gentzel *et al.* [2019] argued that causal discovery methods should be evaluated using interventional measures instead of structural ones such as SID and SHD. The causal distances we introduce are interventional and counterfactual measures.

3.1 Limitations of Related Work

Even though SID is focused on interventional distributions, it assumes that the underlying observational distribution has been estimated correctly. In practice, this is usually not the case, since the estimation is done using finitely many noisy samples. In general, SID cannot provide useful answers when the causal models disagree at the observational level. In fact, even when the observational distribution is just slightly off, SID may still produce highly inaccurate results.

To illustrate this problem, consider two causal models $\mathfrak{C}_1, \mathfrak{C}_2$, each with two nodes A, B. Both models have the graph $A \rightarrow B$, with $A \sim \mathscr{N}(0, \sigma_A)$ and B's noise $N_B \sim \mathscr{N}(0, \sigma_B)$:

$$\mathfrak{C}_1 : \; B := A + N_B, \; P_B^{\mathfrak{C}_1; \mathrm{do}(A=a)} = \mathscr{N}(a, \sigma_B) \quad (5)$$

$$\mathfrak{C}_2 : \; B := -A + N_B, \; P_B^{\mathfrak{C}_2; \mathrm{do}(A=a)} = \mathscr{N}(-a, \sigma_B). \quad (6)$$

The two models predict different values for the intervention $\mathrm{do}(A = a)$. In a toy interpretation, B could be the improvement in life expectancy, and A the daily intake of some drug. Then these two models would give rise to opposite policies given the goal of maximizing life expectancy. This should be reflected by a large distance between the models, but in fact the opposite happens: since $\mathfrak{C}_1$ and $\mathfrak{C}_2$ share the causal graph $\mathscr{G}$, we have $\mathrm{SHD}(\mathscr{G}, \mathscr{G}) = \mathrm{SID}(\mathscr{G}, \mathscr{G}) = 0$.

Strictly speaking, SID cannot even be applied in this case because the observational distributions are not identical. If, however, $\sigma_A \ll \sigma_B$, the observational distributions become almost indistinguishable, and one might be tempted to apply SID, obtaining a distance of 0 although the interventional distributions still give rise to opposite policies. We provide more details about this problem and its resolution in Appendix A (cf. footnote 2).

Another limitation of SID is that it is binary: either two pairwise interventional distributions are the same or not (cf. Eq. 4). It does not quantify the difference. In fact, for practical applications, two slightly wrongly inferred interventional distributions might be preferable to one completely wrongly inferred distribution. Also, if one has prior knowledge about which interventions are more critical, one might want to reflect this in the evaluation. Finally, SID and SHD cannot compare causal models at the counterfactual level as they ignore the structural equations and noise distributions (cf. Eq. 1).

4 Definition of Causal Distances

Let $\mathbf{X}$ be a set of random variables, and $\mathfrak{C}_1, \mathfrak{C}_2$ two causal models defined over them. We now introduce natural formulations of distances at the observational, interventional, and counterfactual level. Intuitively, they build upon an underlying distance between probability distributions and mirror the hierarchical aspect of Pearl and Mackenzie's ladder [2018].

4.1 Observational Distance (OD)

Let $P_{\mathbf{X}}^{\mathfrak{C}_1}, P_{\mathbf{X}}^{\mathfrak{C}_2}$ be the observational distributions induced by $\mathfrak{C}_1, \mathfrak{C}_2$. The *observational distance* (OD) is trivial and corresponds to choosing a distance D between probability distributions:

$$\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) = D\left(P_{\mathbf{X}}^{\mathfrak{C}_1}, P_{\mathbf{X}}^{\mathfrak{C}_2}\right). \quad (7)$$

Example choices for D include the Hellinger, total variation, or Wasserstein distance.

4.2 Interventional Distance (ID)

An intuitive way to compare two causal models $\mathfrak{C}_1, \mathfrak{C}_2$ at the interventional level is to compare all their interventional distributions. Let I denote the node on which the intervention is performed and μ a distribution over nodes that weighs the interventions on each node. In the absence of such information, μ may be chosen as the uniform distribution. Then, the *interventional distance* (ID) is defined as

$$\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = \mathbb{E}_{I \sim \mu} \mathbb{E}_{i \sim P_I} \left[\mathrm{OD}(\mathfrak{C}_1; \mathrm{do}(I = i), \mathfrak{C}_2; \mathrm{do}(I = i)) \right].$$

By convention, we include the empty intervention $I = \emptyset$, which corresponds to the observational distance OD.

In words, ID is the expected deviation in the interventional distributions if we sample a node I on which to intervene according to μ and sample its value according to P_I.

The expectation $\mathbb{E}_{i \sim P_I}$ indicates that I's values are drawn from the distribution P_I. For instance, P_I can be uniform for discrete models or standard Gaussian for continuous models. We only enforce P_I to be strictly positive for all possible values that I can take.

Variables in $\mathbf{X}$ may have different value scales, such that the averaged distances may be dominated by a few variables. This can be fixed by co-opting the weights μ in order to normalize each variable. In the tool we release, we give the option to z-normalize each variable and compute a scale-invariant ID.

Note that μ can give weights of 0 to some nodes, e.g., if they are unobservable to the causal discovery method.

4.3 Counterfactual Distance (CD)

The natural way to compare models at the counterfactual level is to consider their interventional distance on all counterfactual models, i.e., the counterfactual models induced by all possible evidences. Let $E = e$ denote the observed evidence and ν a distribution over nodes that weighs the counterfactuals induced by observing each node. Similar to μ, in the absence of further information, ν may be chosen to be uniform. Then, the *counterfactual distance* (CD) is defined as

$$\mathrm{CD}(\mathfrak{C}_1, \mathfrak{C}_2) = \mathbb{E}_{E \sim \nu} \mathbb{E}_{e \sim P_E} [\mathrm{ID}(\mathfrak{C}_1 | E = e, \mathfrak{C}_2 | E = e)],$$

By convention, we include the empty evidence $E = \emptyset$, which corresponds to the interventional distance ID.

The expectation $\mathbb{E}_{e \sim P_E}$ indicates that E's values are drawn from the distribution P_E. In the absence of information, P_E may be uniform for discrete models or standard Gaussian for continuous models.

5 Basic Properties of Causal Distances

In this section, we assume that both μ and ν are uniform over the set of nodes. The proofs are given in Appendix F.

Each distance builds on top of the distance defined at the level below (CD on ID; ID on OD), thus reflecting the hierarchical structure of the ladder of causation. Furthermore, one can verify the following connections.

Theorem 1. *For two causal models $\mathfrak{C}_1$ and $\mathfrak{C}_2$ over the variables $\mathbf{X}$, we have:*

$$\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) \leq (|\mathbf{X}| + 1) \, \mathrm{CD}(\mathfrak{C}_1, \mathfrak{C}_2) \tag{8}$$

$$\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) \leq (|\mathbf{X}| + 1) \, \mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) \tag{9}$$

Figure 2: $Z \sim \mathcal{N}(0, 1)$ is a hidden confounder in both graphs. The edges indicate a multiplicative factor, e.g., $X = \lambda Z$ in the left graph. The two models have the same joint distribution on (X, Y) and the same graph. Yet, $\mathrm{do}(Z = z)$ for $z \neq 0$ results in two different joint distributions. Their (Wasserstein) distance can be made arbitrarily large by increasing λ.

In particular, counterfactual equivalence implies interventional equivalence, which in turn implies observational equivalence.

5.1 Connection with Graph-based Metrics

The interventional distance (ID) is related to the graph-based SID and SHD via

Theorem 2. *For two causal models $\mathfrak{C}_1, \mathfrak{C}_2$ with causal graphs $\mathcal{G}_1, \mathcal{G}_2$,*

$$\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0 \implies \mathrm{SHD}(\mathcal{G}_1, \mathcal{G}_2) = 0 \implies \mathrm{SID}(\mathcal{G}_1, \mathcal{G}_2) = 0. \tag{10}$$

The reverse directions do not hold in general.

A further connection between SID and our causal distances is given by

Theorem 3. *Let $\mathfrak{C}_1, \mathfrak{C}_2$ be two causal models with causal graphs $\mathcal{G}_1, \mathcal{G}_2$. When $\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) = 0$,*

$$\mathrm{SID}(\mathcal{G}_1, \mathcal{G}_2) = 0 \iff \mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0. \tag{11}$$

When $\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) \neq 0$ the equivalence does not hold.

From Thm. 2, we know that ID being 0 guarantees that SID is 0. But SID being 0 only ensures that ID is 0 in the specific case where OD is also 0.

5.2 Hidden Variables

Until now, we considered the comparison of two Markovian causal models, i.e., with no hidden confounders. We may ask what happens in the non-Markovian case, where one or both models have hidden confounders.

If both models have hidden confounders that can be intervened on, we cannot bound the expected difference between two models, as the outcome of intervening on the hidden confounder can be made arbitrarily large, as shown by Fig. 2.

However, this constitutes a fairly peculiar scenario. Indeed, it is expected that comparing "incomplete" models can only give partial information. It is similar to trying to establish a distance between vectors where one dimension remains hidden. In practice, if we wish to compare two causal models either (i) one is fully known (e.g., the gold standard model to which we compare a model inferred by a causal discovery technique with variables unobservable during training) or (ii) the hidden variables cannot be intervened on. In this latter case, OD and ID computed on the observed subset preserve their interpretation.
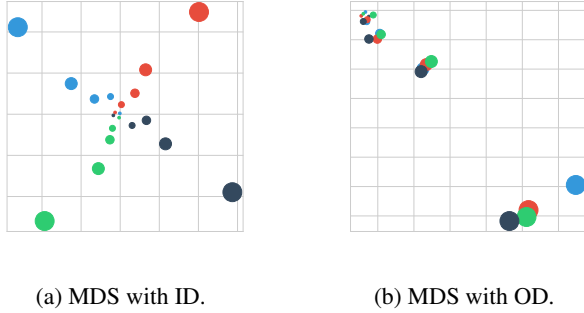
(a) MDS with ID.   (b) MDS with OD.

Figure 3: Comparison of multidimensional scaling (MDS) embeddings induced by OD and ID using the four causal models described in Sec. 6.1 (one color per model; size represents the strength of β).

5.3 Practical Implementation

In Appendix D, we describe the practical details related to the implementation and efficient estimation of the causal distances. We also discuss how to handle both continuous and discrete variables. This implementation results in a tool that we make publicly available to the community.

6 Experiments

We now conduct experiments with the causal distances, on both synthetic and real-world causal models. In all experiments, we use the sample Wasserstein distance [Villani, 2008] as the underlying distance D between probability distributions (cf. Eq. 7). In Appendix B, we validate the sample efficiency and sensitivity of the causal distances.

6.1 Geometry of Causal Distances

First, we illustrate the intuitive geometry induced by our causal distances. In particular, we consider simple models with two nodes A and B using linear structural equations and Gaussian noise. We let $\beta > 0$ denote the strength of the causal connection, let $N \sim \mathcal{N}(0,1)$ be B's noise, and consider four types of models:

$$A \sim \mathcal{N}(0,1) \text{ and } B \coloneqq \pm\beta A + N,$$
$$B \sim \mathcal{N}(0,1) \text{ and } A \coloneqq \pm\beta B + N,$$

We construct five models of each type with $\beta = 0.1, 0.5, 1, 2, 5$, respectively, resulting in 20 models overall. We then compute the pairwise distances between all models using ID and apply multidimensional scaling to obtain 2D embeddings of all models. The result is depicted in Fig. 3a and exhibits the geometrical structure induced by ID, where each type of model creates its own branch, and larger values of β push the different types further apart. When $\beta \to 0$, all models converge to a model where A and B are causally disconnected. When viewed in 3D, equal angles would separate all pairs of branches.

In contrast, SID induces a much poorer geometry where each model is projected on one of two points: one representing the graph $A \to B$, the other, the graph $B \to A$. With OD, shown in Fig. 3b, the models form one branch in the 2D embedding. They are only distinguished based on the amplitude of β; neither the sign nor the orientation of the graph are captured.
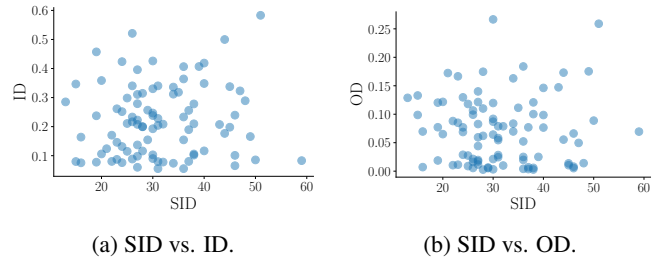

(a) SID vs. ID.   (b) SID vs. OD.

Figure 4: Comparisons between ID and OD against SID on 90 randomly sampled pairs of causal models.

6.2 Comparison of Causal Distances and SID

Whereas Thm. 3 connects ID, OD, and SID when two of these quantities are 0, we now empirically investigate their relationship when they deviate from 0. Fig. 4a shows a scatter plot comparing ID and SID, where each dot is a pair of random causal models. As we see, there is little correlation between ID and SID. It is possible to find pairs of causal models with low ID but high SID, and *vice versa*. These results highlight how the different distances capture different aspects of the models being compared.

6.3 Evaluation of Causal Discovery Systems

An important application of causal distances is the evaluation of causal discovery systems. In this section, we illustrate this by evaluating several causal discovery systems using both real-world and synthetic causal models.

We considered the following real-world Bayesian causal models from Elidan [2001]: *Cancer1*, *Cancer2*, *Earthquake*, *Survey*, *Protein*, *Child*, and *Insurance*. They range from 5 to 27 nodes and from 4 to 52 edges. For synthetic models, we sample random DAGs and random parametrizations using the CDT tool [Kalainathan and Goudet, 2019]. We consider the following parametrizations: *linear Gaussian (linGauss)*, *linear non-Gaussian (linNGauss)*, *Gaussian process with additive noise (GPAddit)*, and *Gaussian Process (GP)*.

For each model, we sample 2,000 observations from which causal discovery methods should recover the causal model. Since they are causal Bayesian networks, not structural causal models, we cannot compute counterfactuals [Pearl, 2019]. Thus, we restrict ourselves to ID, SHD, and SID.

Systems. We consider multiple causal discovery methods for recovering the causal graph: CCDr [Aragam and Zhou, 2015], PC [Spirtes *et al.*, 2000], GES [Chickering and Meek, 2002], GIES [Chickering, 2002], MMPC [Tsamardinos *et al.*, 2003a], IAMB [Tsamardinos *et al.*, 2003b], LiNGAM [Shimizu *et al.*, 2006], and CAM [Spirtes *et al.*, 2000]. Some of these techniques only output a partial DAG with undirected edges, and some only output a graph without parameters. To obtain a fair comparison of the full Bayesian networks, we perform all parameter estimation via maximum likelihood estimates based on the training data. When only a partial DAG is returned, we use the edge orientation that provides the best goodness of fit after the parameters have been estimated. Alternatively, one could report the mean performance over all possible directed graphs.
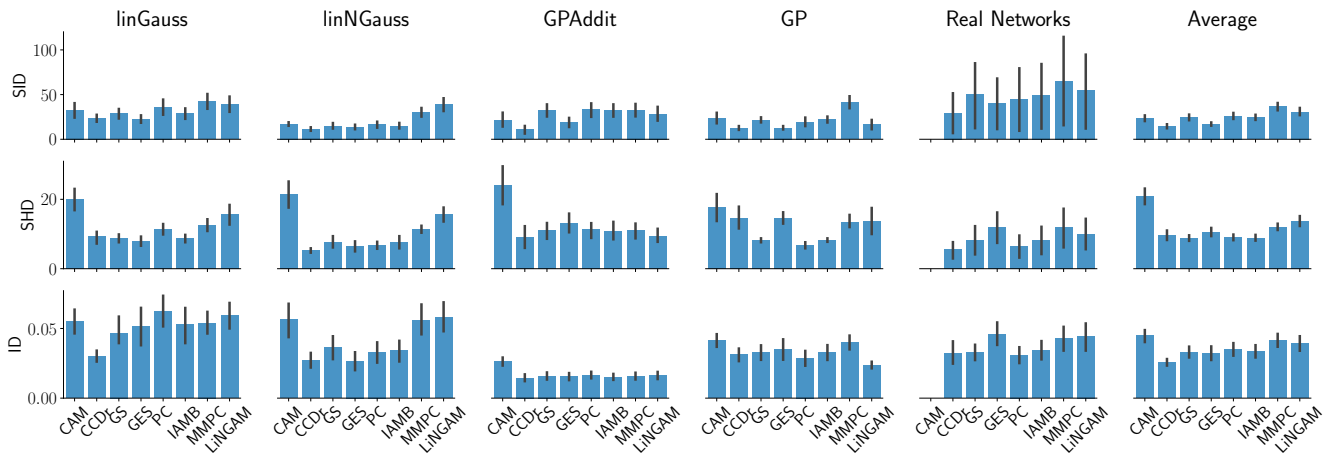
Figure 5: Evaluation of causal discovery techniques on synthetic networks and real-world networks ("Real Networks"). In the first row, models are evaluated by SID, in the second row, by SHD, and in the last row, by ID. The rightmost column shows average performance. Note that CAM yielded errors on some of the real-world networks and is thus not reported. Lower is better. Error bars are 95% confidence intervals.

Remark. Our distances compare full causal models, i.e., causal graphs after the parameters have been estimated. Here, by fixing the parameter estimation, we measure the impact of the causal graph on the intervention predictions. It also ensures that two methods that output the same graph (DAG or partial DAG) will obtain the same evaluation results. Yet, as shown below, it does not mean that these evaluation results agree with graph-based metrics such as SHD or SID. For the systems, we use the implementations available in CDT. We compute maximum likelihood estimates using the *Pomegranate* python framework.

Results. In Fig. 5, we report the evaluation results broken down by model parametrization. The *Real Networks* block corresponds to the performance of systems averaged over the real-world causal models described above. Results for each causal model are available in Appendix C. We see that different metrics yield different rankings of systems. Thus, the differences between metrics observed in Sec. 6.2 are also consequential in the task of causal discovery. These observations emphasize the importance of employing the evaluation metric that captures the desired behavior. If only the observational distribution matters, OD should be used, but then causal discovery may not be needed in the first place. If we care about the expected errors in predicting the outcome of interventions, ID should be used, and SID can be employed when we focus on the causal graph under the assumption that the underlying observational distribution has already been correctly estimated. Additionally, even a single metric produces different rankings of systems in different scenarios. Causal discovery requires assumptions about the underlying structure of the true causal model, and few guarantees are given when the respective assumptions are not met. Different networks fulfill different assumptions and are best handled by different causal discovery methods. An evaluation using causal distances such as OD, ID, and CD is indispensable for illuminating which causal discovery method is best suited for which kind of data. Interestingly, ID clearly reveals that systems struggle most for

the linear Gaussian case, which is known to be unidentifiable. While all other cases are identifiable, Gaussian processes with additive noise (GPAddit) seem to be the easiest for existing causal discovery systems. Overall, CCDr seems to perform fairly well in comparison to other systems. In Appendix E, we show that, contrary to machine learning, causal discovery systems do not benefit from more training data.

7 Conclusion

This paper introduces observational (OD), interventional (ID), and counterfactual (CD) distances between causal models, one for each rung of the ladder of causation (cf. Sec. 1). Each distance is defined based on the lower-level ones, reflecting the hierarchical structure of the ladder. We study the properties of our distances and propose practical approximations that are useful for evaluating causal discovery techniques. We release a Python implementation of our causal distances.

Our causal distances do not require the unrealistic assumptions of infinite samples and perfect statistical estimation that are currently common in the study of causality [Pearl, 2009]. Also, they quantify the difference between causal models on a continuous, rather than integer, scale and make use of the data at a finer granularity than the usual binary measurements of methods such as SHD and SID (cf. Sec. 3).

The proposed causal distances have both theoretical and empirical applications, and we hope the research community will use them to advance the study of causality.

Acknowledgments

References

[Acharya *et al.*, 2018] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9447–9460. Curran Associates, Inc., 2018.

[Acid and de Campos, 2003] Silvia Acid and Luis M. de Campos. Searching for Bayesian Network Structures in the Space of Restricted Acyclic Partially Directed Graphs. *Journal of Artificial Intelligence Research*, 18(1):445–490, may 2003.

[Aragam and Zhou, 2015] Bryon Aragam and Qing Zhou. Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. *Journal of Machine Learning Research*, 16(1):2273–2328, January 2015.

[Chickering and Meek, 2002] David Maxwell Chickering and Christopher Meek. Finding Optimal Bayesian Networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 94–102. Morgan Kaufmann Publishers Inc., 2002.

[Chickering, 2002] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[de Jongh and Druzdzel, 2009] Martijn de Jongh and Marek J Druzdzel. A Comparison of Structural Distance Measures for Causal Bayesian Network Models. *Recent Advances in Intelligent Information Systems, Challenging Problems of Science, Computer Science series*, pages 443–456, 2009.

[Elidan, 2001] Gal Elidan. Bayesian Network Repository. https://www.cse.huji.ac.il/~galel/Repository, 2001. Accessed: 2021-01-15.

[Gentzel *et al.*, 2019] Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11717–11727. Curran Associates, Inc., 2019.

[Kalainathan and Goudet, 2019] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python, 2019.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4066–4076. Curran Associates, Inc., 2017.

[Pearl and Mackenzie, 2018] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[Pearl, 2009] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[Pearl, 2019] Judea Pearl. The seven tools of inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, February 2019.

[Peters and Bühlmann, 2015] Jonas Martin Peters and Peter Bühlmann. Structural intervention distance for evaluating graphs. *Neural Computation*, 27(3):771–799, 3 2015.

[Peters *et al.*, 2017] Jonas Martin Peters, Dominik Janzing, and Bernard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

[Rojas-Carulla *et al.*, 2018] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

[Shimizu *et al.*, 2006] Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

[Singh *et al.*, 2017] Karamjit Singh, Garima Gupta, Vartika Tewari, and Gautam Shroff. Comparative benchmarking of causal discovery techniques. *arXiv preprint arXiv:1708.06246*, 2017.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N. Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, Prediction, and Search*. MIT press, 2000.

[Sriperumbudur *et al.*, 2010] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, August 2010.

[Theis *et al.*, 2015] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[Tsamardinos *et al.*, 2003a] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander Statnikov. Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 673–678, New York, NY, USA, 2003. Association for Computing Machinery.

[Tsamardinos *et al.*, 2003b] Ioannis Tsamardinos, Constantin F. Aliferis, and Alexander R. Statnikov. Algorithms for Large Scale Markov Blanket Discovery. In *FLAIRS Conference*, pages 376–381. AAAI Press, 2003.

[Verma and Pearl, 1991] Thomas Verma and Judea Pearl. Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, 1991.

[Villani, 2008] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.