A LADDER OF CAUSAL DISTANCES

Maxime Peyrard
EPFL
maxime.peyrard@epfl.ch

Robert West
EPFL
robert.west@epfl.ch

May 7, 2020

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. Finally, we highlight the usefulness of our causal distances by briefly discussing further applications beyond the evaluation of causal discovery techniques.

1 Introduction

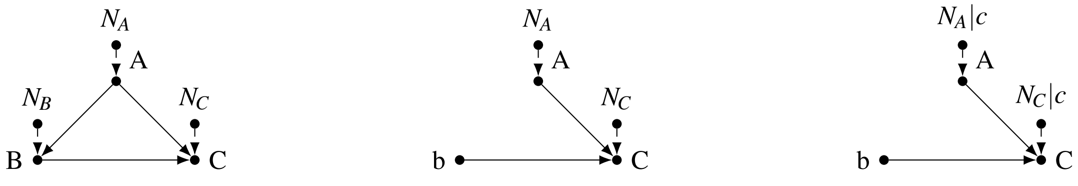
Reasoning about the causes and effects driving physical and societal phenomena is an important goal of science. Causal reasoning facilitates the prediction of intervention outcomes and can ultimately lead to more principled policymaking (Spirtes et al., 2000; Pearl and Mackenzie, 2018).

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:*

1. *Observational* queries: seeing and observing. What can we tell about Y if we observe $X = x$?
2. *Interventional* queries: acting and intervening. What can we tell about Y if we do $X = x$?
3. *Counterfactual* queries: imagining, reasoning, and understanding. Given that $E = e$ actually happened, what would have happened to Y had we done $X = x$?

Asking such questions requires a causal model to begin with. The problem of inferring such a model from observational, interventional, or mixed data is 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, Janzing, and Schölkopf, 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), algorithmic fairness (Kusner et al., 2017), and interpretability (Lipton, 2018).

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



(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.

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 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. We highlight theoretical properties of the distances in relation to previously proposed distances (Sec. 5) and discuss how to efficiently approximate them in practice (Sec. 6). Then, we study their behavior in a series of experiments and put them to use to evaluate existing causal discovery systems (Sec. 7). We conclude with a discussion of implications and further applications (Sec. 8). The implementation of causal distances and our experiments are available at: `https://github.com/epfl-dlab/causal-distances`.

2 Preliminaries

2.1 Causal graphs

We consider a finite ordered set of random variables $\mathbf{X} = (X_1, \dots, X_d)$. A directed graph $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ consists of the set of indexed nodes $\mathbf{X}$ together with a set of directed edges $\mathbf{E} \subseteq \mathbf{X} \times \mathbf{X}$. If $(X_i, X_j) \in \mathbf{E}$, 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 $\mathcal{G}$ 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 (exogeneous) noise variables $\mathbf{N}$ and $\mathbf{F} = (f_1, \dots, f_d)$ is a set of *structural equations* indicating, for each $X_i \in \mathbf{X}$, how its value is determined by its parents and noise:

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

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), i.e.,

$$P_{\mathbf{N}}(n_1, \dots, n_d) = \prod_{i=1}^{d} P_{N_i}(n_i). \tag{2}$$

The associated causal graph $\mathcal{G}$ 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 these assumptions:

- *Markov property*: Every conditional independence statement entailed by the causal graph is satisfied by the joint distribution.
- *Causal minimality*: The joint distribution satisfies the Markov property for $\mathscr{G}$ but not for any proper subgraph of $\mathscr{G}$.
- *Causal faithfulness*: Every conditional independence within the joint distribution is entailed by the causal graph.
- *Positiveness*: The entailed marginal and conditional distributions are strictly positive.

2.3 Observational, interventional, and counterfactual distributions

Fig. 1 provides a graphical illustration of an SCM over three variables 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, Janzing, and Schölkopf, 2017). To sample from $\mathfrak{C}$, we can simply sample from the noise distribution $P_{\mathbf{N}}$ and use the structural assignments (cf. (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. (1)) of variables in $\mathbf{I}$ by forcing them to specific values $\mathbf{I} = \mathbf{i}$, so-called *hard intervention*.[1] 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 the *interventional distribution* of $\mathbf{X}$, denoted by $P_{\mathbf{X}}^{\mathfrak{C};\mathrm{do}(\mathbf{I}=\mathbf{i})}$.

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*. 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):

$$P_{\mathbf{N}|\mathbf{E}=\mathbf{e}}(\mathbf{n}) = \frac{P_{\mathbf{E}|\mathbf{N}=\mathbf{n}}(\mathbf{e})}{P_{\mathbf{E}}(\mathbf{e})} P_{\mathbf{N}}(\mathbf{n}). \tag{3}$$

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, while 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* ($d(x,z) \leq d(x,y) + d(y,z)$). A *pseudometric* relaxes the identity of indiscernibles such that $x = y \implies d(x,y) = 0$, but the implication does not necessarily hold in the opposite direction. A *premetric* only satisfies non-negativity and $x = y \implies d(x,y) = 0$.

The causal distances introduced in this paper are pseudometrics, whereas SID (Sec. 1 and 3) is a premetric.

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.

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

The comparison of observational distributions has been studied extensively in machine learning and statistics (cf. the overviews by Theis, Oord, and Bethge (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 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:

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

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 $\mathrm{SID}(\mathscr{G},\mathscr{H}) = 0$.

Singh et al. (2017) also discuss evaluation methods that measure the performance of an inferred causal model with respect to some predefined causal effect: for fixed $X,Y \in \mathbf{X}$, is $P_Y^{\mathfrak{C};\mathrm{do}(X=x)}$ estimated correctly? Acharya et al. (2018) compare two causal models, but only for the purpose of testing identity. Both constitute special cases of our interventional distance.

Recently, Gentzel, Garant, and Jensen (2019) argued that causal discovery methods should be evaluated using interventional measures instead of structural ones like SID and SHD. The causal distances introduced here form a general class of such interventional and counterfactual measures.

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 $A \to B$, with $A \sim \mathscr{N}(0, \sigma_A)$ and B's noise $N_B \sim \mathscr{N}(0, \sigma_B)$:

$$\mathfrak{C}_1 : \ B := A + N_B, \tag{5}$$

$$\mathfrak{C}_2 : \ B := -A + N_B. \tag{6}$$

Note that the two models predict different values for the intervention $\mathrm{do}(A = a)$ for $a \neq 0$:

$$P_B^{\mathfrak{C}_1;\mathrm{do}(A=a)} = \mathscr{N}(a, \sigma_B), \tag{7}$$

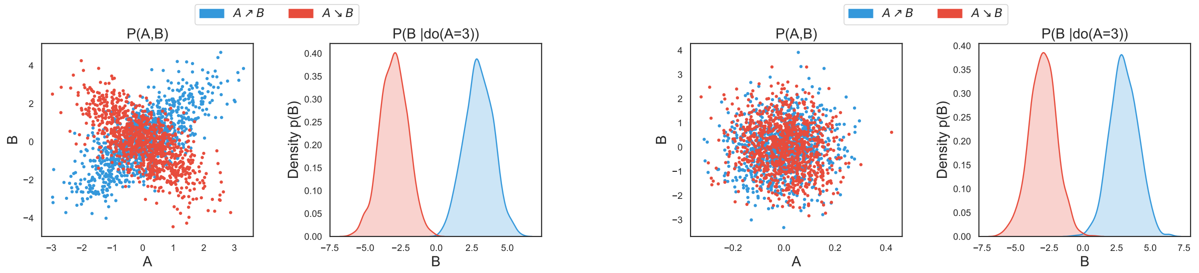
$$P_B^{\mathfrak{C}_2;\mathrm{do}(A=a)} = \mathscr{N}(-a, \sigma_B). \tag{8}$$

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.

Fig. 2a depicts the observational and interventional distributions of both models under the action $\mathrm{do}(A = 3)$ with $\sigma_A = \sigma_B = 1$. Similarly, Fig. 2b shows the same distributions but with $\sigma_A = 0.1 \cdot \sigma_B$. We observe that the interventional distributions remain the same and different (ID is constant) even if the observational distributions become almost indistinguishable when $\frac{\sigma_A}{\sigma_B}$ becomes smaller.

This problem is resolved by considering the intervention distance ID instead of SID which recognizes the intervention distribution as strictly different. Indeed, $\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) \approx 1.4$ for both noise ratios.



(a) Observational and interventional distributions of both models with $\sigma_A = \sigma_B = 1$

(b) Observational and interventional distribution of both models with $\sigma_A = 0.1 \cdot \sigma_B$

Figure 2: Example of two causal model with the same graph and *similar* observational distribution (Fig. 2b) but different interventional distributions.

Another limitation of SID is that it is based on binary decisions: either two pairwise interventional distributions are the same or not (cf. (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 measure.

Finally, SID and SHD cannot compare causal models at the counterfactual level because they ignore the structural equations and noise distributions (cf. (1)). In contrast, we now propose distances for comparing causal models at all rungs of the ladder of causation.

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 discuss natural formulation 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 between probability distributions:

$$\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) = D\left(P_{\mathbf{X}}^{\mathfrak{C}_1}, P_{\mathbf{X}}^{\mathfrak{C}_2}\right). \tag{9}$$

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, 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 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 and standard Gaussian for continuous ones. We only enforce P_I to be strictly positive for all possible values that I can take.

Note that μ can give weights of 0 to some nodes that, for example, were unobservable to the causal discovery method. In such cases, the computation of ID is effectively performed on a subset of nodes.

Note that computing the effect of one variable X on another variable Y, as discussed by Singh et al. (2017) (cf. Sec. 3), is a special case where $\mu(X) = 1$ and the comparison of the resulting distributions is restricted to the marginals of Y.

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 observation and ν a distribution over nodes that weighs the counterfactual induced by observing each node. Similar to μ, ν may be chosen to be uniform in the absence of further information.

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 distribution 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 and standard Gaussian for continuous ones.

5 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 B.

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, for all $\epsilon \geq 0$:*

$$\mathrm{CD}(\mathfrak{C}_1, \mathfrak{C}_2) \leq \epsilon \implies \mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) \leq (|\mathbf{X}| + 1)\epsilon \tag{10}$$

$$\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) \leq \epsilon \implies \mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) \leq (|\mathbf{X}| + 1)\epsilon \tag{11}$$

In particular, counterfactual equivalence implies interventional equivalence which implies observational equivalence (corresponding to the case $\epsilon = 0$).

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 $\mathscr{G}_1, \mathscr{G}_2$,*

$$\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0 \implies \mathscr{G}_1 = \mathscr{G}_2 \implies \mathrm{SHD}(\mathscr{G}_1, \mathscr{G}_2) = 0 \implies \mathrm{SID}(\mathscr{G}_1, \mathscr{G}_2) = 0. \tag{12}$$

The reverse direction of (12) does not hold in general.

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

Theorem 3. *For two causal models $\mathfrak{C}_1, \mathfrak{C}_2$ with causal graphs $\mathscr{G}_1, \mathscr{G}_2$. When $\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) = 0$, we have:*

$$\mathrm{SID}(\mathscr{G}_1, \mathscr{G}_2) = 0 \iff \mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0. \tag{13}$$

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 might wonder 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. 3.

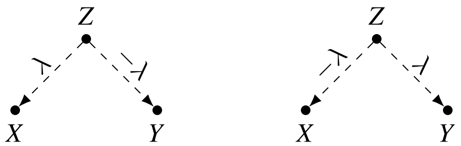


Figure 3: $Z \sim \mathcal{N}(0,1)$ is a hidden confounder in both graphs. The edges indicate a multiplicative factor, e.g., $X = \lambda Z$ on the left graph. The two models have the same joint distribution on (X,Y) and the same graph. Yet, $\text{do}(Z = z), z \neq 0$ results in two different joint distributions. Their (Wasserstein) distance can be made arbitrarily large by increasing λ.

However, this constitutes a fairly peculiar scenario. Indeed, it is expected that comparing "incomplete" models can only give partial information.[2] 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.

5.3 Pseudo-metrics

Technically, OD, ID, and CD are not metrics, since the identity of indiscernibles does not hold: there can be two distinct causal models with an OD, ID, or CD of 0. If D satisfies the other metric axioms (but not otherwise, e.g., if D is the asymmetric KL-divergence; Sec. 2.4), our distances are pseudometrics. Like all pseudometrics, however, they can be turned into proper metrics by considering equivalence classes as the objects of comparison, where the equivalence class of a model is the set of all models to which it has distance 0. Interestingly, these equivalence classes are tightly connected to problems of identifiability (Pearl, 2009).

6 Estimating causal distances in practice

We now discuss the practical computation of OD, ID, and CD. For general causal models, they cannot be computed analytically. Instead, we must draw finitely many samples and use empirical distances $\tilde{D}$ instead of the theoretical D. This results in estimated distances denoted by $\tilde{\text{OD}}$, $\tilde{\text{ID}}$, and $\tilde{\text{CD}}$.

Observational. In order to estimate OD, we draw k samples from the joint observational distribution of each model and use a sample distance $\tilde{D}$. Consequently, the estimated $\tilde{\text{OD}}$ directly inherits the statistical properties of the chosen estimator $\tilde{D}$ and has sampling complexity $\mathcal{O}(k)$.

Interventional. The computation of ID involves the application of $\tilde{\text{OD}}$ to compare dl pairs of interventional distributions: for each node I from the set of all d nodes, l intervention values i are sampled from P_I, and the corresponding interventional distribution $P_{\mathbf{X}}^{\mathfrak{C};\text{do}(I=i)}$ is estimated by drawing k samples. Thus, the sampling complexity is $\mathcal{O}(dlk)$. Algorithm 1 is the pseudo-code for the computation of $\tilde{\text{ID}}$.

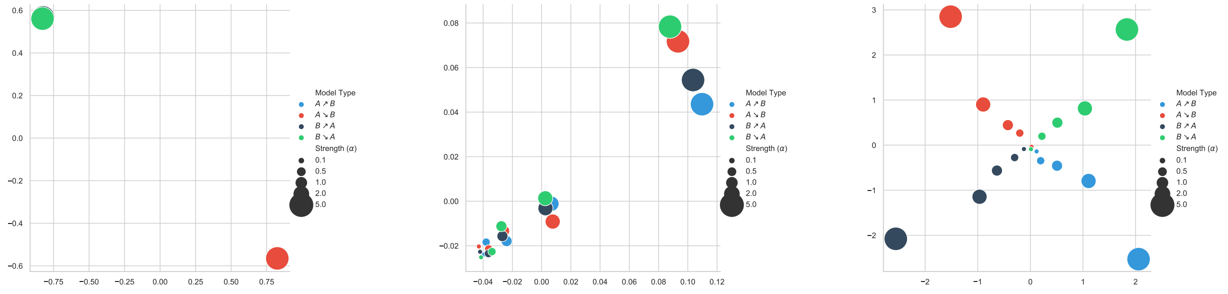
Algorithm 1: Computing ID in practice

```
Function $\tilde{\text{ID}}(\mathfrak{C}_1, \mathfrak{C}_2, k, l, \mu)$:
    $id := 0$
    for $i \in \{1, \dots, d\}$ do
        $id \mathrel{+}=$
        $\sum_{\substack{x \sim P_I \\ j=1}}^{j=l} \mu(I)\tilde{\text{OD}}(P_{\mathbf{X}}^{\mathfrak{C}_1;do(X_i=x)}, P_{\mathbf{X}}^{\mathfrak{C}_2;do(X_i=x)}, k)$
    end
    return $\frac{id}{l \cdot d}$
```

Algorithm 2: Computing CD in practice

```
Function $\tilde{\text{CD}}(\mathfrak{C}_1, \mathfrak{C}_2, m, l, k, \nu)$:
    $cd := 0$
    for $E \in \mathbf{X}$ do
        for $i \in \{1, \dots, m\}$ do
            $e \sim P_E$
            $(\mathfrak{C}_1 | E = e) := (\mathbf{X}, \mathbf{N}, \mathbf{F}, P_{\mathbf{N}_1 | E=e})$
            $(\mathfrak{C}_2 | E = e) := (\mathbf{X}, \mathbf{N}, \mathbf{F}, P_{\mathbf{N}_2 | E=e})$
            $cd \mathrel{+}= \nu(E)\tilde{\text{ID}}(P_{\mathbf{X}}^{\mathfrak{C}_1|E=e}, P_{\mathbf{X}}^{\mathfrak{C}_1|E=e}, l, k)$
        end
    end
    return $\frac{cd}{d \cdot m}$
```

[2] It is similar to trying to establish a distance between vectors where one dimension remains hidden.



(a) MDS with SID.

(b) MDS with OD.

(c) MDS with ID.

Figure 4: Comparison of multidimensional scaling embeddings induced by OD, SID and ID using various causal models relating the two variables A and B. For SID, all models are collapsed onto the two points.

Counterfactual. The estimation of CD involves the computation of $\widetilde{\mathrm{ID}}$ on several modified causal models. For each node E, m evidence values e are sampled from P_E. For each evidence $E = e$, the noise distributions of both models are updated using Bayes' rule (cf. (3)) and $\widetilde{\mathrm{ID}}$ is computed on these modified causal models. The sampling complexity of $\widetilde{\mathrm{CD}}$ is therefore $\mathscr{O}(d^2mlk)$. Algorithm 2 is the pseudo-code for the computation of $\widetilde{\mathrm{CD}}$.

The Bayesian update can be computationally demanding. To address this, we first observe that we only need to sample from $P(\mathbf{N}|E=e)$ (or $P_{\mathbf{N}|E=e}$ in the notation of (3)) to estimate ID in the induced counterfactual model. Using a general Gibbs sampler, it suffices to compute the likelihood term $P(E=e|\mathbf{N}=\mathbf{n})$, and thanks to the Markov factorization property, this reduces to $P(E=e|\mathbf{PA}_E)$. The value of E is set according to the structural equation $E = f_E(\mathbf{PA}_E, N_E)$. When $\mathbf{PA}_E$ is given but not the noise N_E, we obtain a probability distribution for E. Each likelihood could then be estimated at runtime by sampling the noise $\mathbf{N}$ and empirically estimating $P(E=e|\mathbf{PA}_E)$ with techniques such as density estimation. To speed up the computation, we instead propose a faster alternative, as follows.

If E takes on values from the discrete set $\{e_1, \ldots, e_n\}$, we can speed up the Gibbs sampler by simply precomputing the likelihood estimates $\{P(E=e_i|\mathbf{PA}_E)\}$. Otherwise, if E is a continuous random variable, we first discretize it and then pick n evenly spaced values $\{e_1, \ldots, e_n\}$ for which we precompute the likelihood estimates. At runtime, when the observation $E = e$ is given, we retrieve the nearest neighbor of e and use its precomputed value. In practice, the computation of $\widetilde{\mathrm{CD}}$ is orders of magnitude faster than the naïve algorithm where the likelihood terms are estimated at runtime.

Handling of Continuous Input. We require that the intervention and evidence values i are drawn from a distribution with full support over Ω_I (Peters, Janzing, and Schölkopf, 2017). In the discrete case, it is straightforward to assign a uniform distribution over the elements of Ω_I. However, in the continuous case, we use the standard Gaussian distribution.

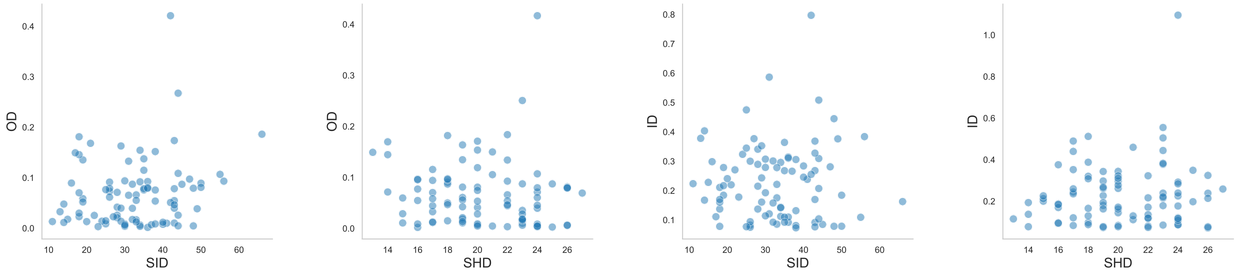
7 Experiments

We now conduct a wide range of experiments. First, using synthetic causal models, we highlight how our causal distances differ among each other as well as from the popular SID (Sec. 7.1, 7.2). Then, still using synthetic models, we evaluate computational aspects, namely sample efficiency and sensitivity to perturbation (Sec. 7.3, 7.4) Finally, we include real-world causal models and use our distances, as well as SID and SHD, to evaluate 8 causal discovery methods from the literature (Sec. 7.5).

In all experiments, we use the sample Wasserstein distance as the underlying distance D between probability distributions (cf. (9)) (Villani, 2008).

7.1 Geometry of causal distances

First, we illustrate the intuitive geometry induced by our causal distances. In particular, we focus on ID and look at 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, $N \sim \mathcal{N}(0,1)$ be B's noise, and consider 4 types of models:



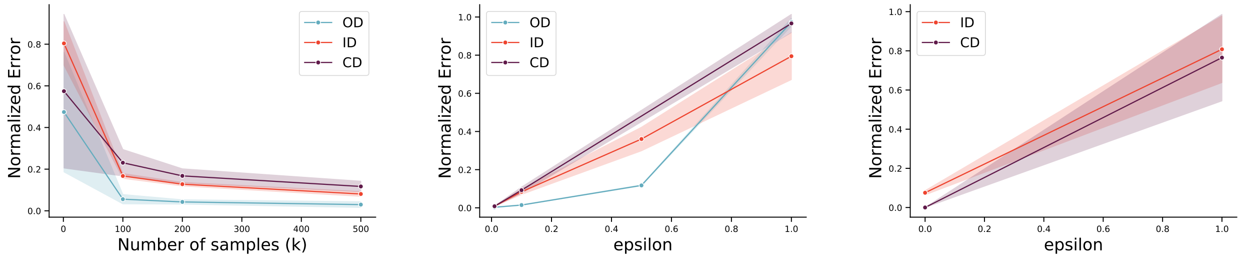
(a) Scatter-plot comparing SID and OD.

(b) Scatter-plot comparing SHD and OD.

(c) Scatter-plot comparing SID and ID.

(d) Scatter-plot comparing SHD and OD.

Figure 5: Comparisons between OD and ID against SID and SHD on 90 randomly sampled pairs of causal models



(a) Sample efficiency of $\widetilde{OD}$, $\widetilde{ID}$, and $\widetilde{CD}$.

(b) Sensitivity of $\widetilde{OD}$, $\widetilde{ID}$, and $\widetilde{CD}$ to perturbations.

(c) Sensitivity of $\widetilde{ID}$ (resp. $\widetilde{CD}$) to perturbations that leave OD (resp. ID) unchanged.

Figure 6: Sample efficiency and sentitivity of the proposed approximations of OD, ID, and CD.

$$\begin{array}{rcl} A \nearrow B & : & A \sim \mathcal{N}(0,1) \text{ and } B := \beta A + N, \\ A \searrow B & : & A \sim \mathcal{N}(0,1) \text{ and } B := -\beta A + N, \\ B \nearrow A & : & B \sim \mathcal{N}(0,1) \text{ and } A := \beta B + N, \\ B \searrow A & : & B \sim \mathcal{N}(0,1) \text{ and } A := -\beta B + N. \end{array}$$

We make 5 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. 4c 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 $\beta \rightarrow 0$, all models converge to a model where A and B are causally disconnected. Note that in 3D, equal angles separate all pairs of branches.[3]

In contrast, SID depicted in Fig. 4a induces a much poorer geometry where each model is projected on one of two points: one representing the graph $A \rightarrow B$, the other, the graph $B \rightarrow A$. With OD, shown in Fig. 4b, 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.

7.2 Comparison of causal distances and SID

While Thm. 3 connects ID, OD and SID when two of these quantities are 0, we empirically investigate their relationship when they deviate from 0. Fig. 5c shows a scatter-plot comparing the correlations between ID and SID, where each dot is a pair of randomly sampled causal models (between 5 and 20 nodes and expected degree of 3). 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*. Similarly, there is little correlation between SID and OD as shown by Fig. 5a.

The same behaviour is observed when SID is replaced by SHD, as depicted by Fig. 5b and Fig. 5d.

[3] Visualization will be available in the accompanying Jupyter notebook.

	Cancer1			Cancer2			Child			Earthquake			Insurance			Protein			Survey		
	SID	SHD	ID	SID	SHD	ID	SID	SHD	ID	SID	SHD	ID	SID	SHD	ID	SID	SHD	ID	SID	SHD	ID
LinGAM	38	14	5.4	12	6	3.44	282	45	4.21	16	7	10.56	528	91	5.56	58	32	4.44	26	8	1.45
CCDr	6	11	1.75	1	1	1.32	57	13	3.65	0	5	8.	456	51	5.53	18	27	3.25	12	9	1.7
GS	18	6	1.82	16	7	3.7	273	44	3.93	0	1	4.82	542	64	5.18	51	22	4.31	27	11	1.88
GES	44	21	6.15	20	8	3.35	189	78	3.64	20	11	9.26	545	98	5.6	50	48	5.09	27	15	1.8
PC	11	4	1.54	12	6	2.64	182	27	3.84	0	1	4.79	488	49	5.17	40	20	4.40	27	9	1.81
IAMB	18	6	1.63	16	7	3.62	253	43	3.81	0	1	6.17	588	67	5.26	51	22	4.31	27	11	1.82
MMPC	51	16	5.63	16	7	3.19	367	61	4.31	20	9	9.13	682	97	5.07	59	34	4.35	27	11	1.97

Table 1: Evaluation of various causal discovery techniques with SID, OD and ID.

These results highlight how the different distances capture different aspects of the models being compared. They should be considered complementary to one another.

7.3 Sample efficiency

Next, we validate the sample efficiency of approximating OD, ID, and CD (Sec. 6) by observing how quickly the estimates converge to 0 when comparing a causal model to itself.

In Fig. 6a, we fix $l = 100$ and $m = 10$, sample graphs with $d = 6$ nodes and an expected degree of 3, and vary the number k of samples used to estimate the distributions. The shaded area represents the standard deviation after repeating the experiment 10 times with different random seeds. We observe a quick decrease towards 0, especially for $\widetilde{\mathrm{OD}}$ and $\widetilde{\mathrm{ID}}$. When going up the ladder toward $\widetilde{\mathrm{CD}}$, errors can accumulate due to the finite sample size.

7.4 Sensitivity to perturbation

In the next experiment, we verify that $\widetilde{\mathrm{OD}}$, $\widetilde{\mathrm{ID}}$, and $\widetilde{\mathrm{CD}}$ can capture perturbations of causal models despite the imperfect approximations due to the finite sample size.

We randomly draw a causal model $\mathfrak{C}$ of $d = 10$ nodes and 34 edges and perturb one of its mechanisms f_i by adding another random mechanism g_i according to a perturbation parameter $\epsilon \in [0, 1]$. This results in a new causal model $\mathfrak{C}_{\epsilon,g_i}$ which is identical to $\mathfrak{C}$ except for the i-th mechanism, which is replaced by $(1 - \epsilon) f_i + \epsilon g_i$. When $\epsilon = 0$, $\mathfrak{C}_{\epsilon,g_i} = \mathfrak{C}$ and we expect the distance to be 0. As ϵ increases, the distance should grow.

Fig. 6b plots the growth of $\widetilde{\mathrm{OD}}$, $\widetilde{\mathrm{ID}}$, and $\widetilde{\mathrm{CD}}$ as functions of the perturbation ϵ using $k = 1000$, $l = 100$ and $m = 10$. Each distance increases with ϵ, with the effect being more visible higher up the ladder of causation. Intuitively, slight perturbations have the potential to induce large deviations when going up the ladder. Indeed, a perturbation of the mechanism modifies the likelihood terms, which also modifies the Bayesian update of the noise variables. Note that graph-based metrics such as SHD and SID cannot capture such nuances because the causal graph remains unchanged.

Next, we perturb $\mathfrak{C}$ in a way that OD (resp. ID) remains unchanged and observe the variations in $\widetilde{\mathrm{ID}}$ (resp. $\widetilde{\mathrm{CD}}$). We detail how we proceed to create such perturbations in Appendix A. We denote with ϵ the parameter that quantifies these perturbations and report the results in Fig. 6c, which shows that both $\widetilde{\mathrm{ID}}$ and $\widetilde{\mathrm{CD}}$ detect their level-specific perturbations.

7.5 Evaluation of causal discovery systems

An important application of causal distances is the evaluation of causal discovery systems. OD, ID, and CD can precisely evaluate the inferred causal model $\mathfrak{C}$ in comparison to a ground-truth causal model $\mathfrak{D}$ at each rung of the ladder of causation. 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: **Cancer1** (Lauritzen and Spiegelhalter, 1990), a model of lung cancer (8 nodes, 8 edges); **Cancer2** (Korb and Nicholson, 2010), a toy model connecting pollution and smoking to lung cancer (5 nodes, 4 edges); **Earthquake** (Korb and Nicholson, 2010), a model of alarm triggering (5 nodes, 4 edges); **Survey** (Scutari and Denis, 2014), a model of survey outcomes (6 nodes, 6 edges); **Protein** (Sachs et al., 2005), a model of protein signaling (11 nodes, 17 edges); **Child** (Elidan, 2001), a model of the diseases that birth asphyxia may cause (20 nodes, 25 edges); **Insurance** (Binder et al., 1997), a model of car insurance policies (27 nodes, 52 edges).

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.

Furthermore, we sample causal models for each of the following parametrization:

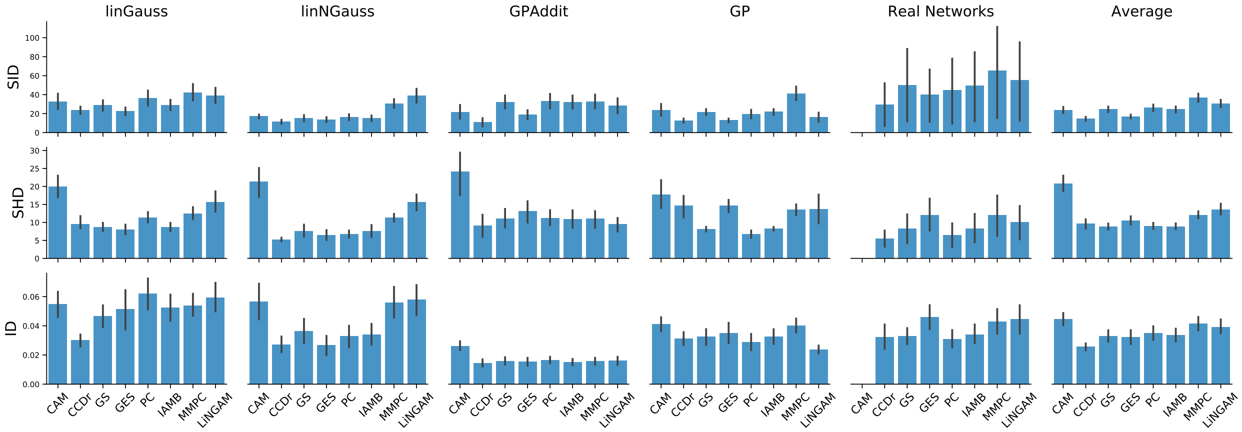


Figure 7: Evaluation of causal discovery techniques on synthetic and real-world networks. In the first row models are evaluated by SID, in the second row by SHD and the last row is for ID. The second column depicts the average performance. Note that CAM yields errors on some of the *Real Networks* and is thus not reported.

- **Linear Gaussian (linGauss)**: $X_i = \sum_{X_j \in \mathbf{PA}_i} \alpha_i X_j + N_i$ with a Gaussian noise: $N_i \sim \mathcal{N}(0,1)$ and $\alpha_i \in \mathbb{R}$.
- **Linear Non-Gaussian (linNGauss)**: Same as above but the noise is distributed according to a Γ distribution: $N_i \sim \Gamma(a,b),\ a \sim \mathcal{N}(0,1),\ b \sim \mathcal{N}(0,1)$.
- **Gaussian Process Additive (GPAddit)**: $X_i = GP(\mathbf{PA}_i) + N_i$. The mechanism is a multivariate Gaussian process, the noise is additive and Gaussian.
- **Gaussian Process (GP)**: $X_i = GP(\mathbf{PA}_i, N_i)$. The noise is not additive, it is one dimension of the Gaussian process. The noise follows a standard normal distribution.

To sample a causal model, we first sample a causal graph using the Erdős-Rény model and remove cycles to obtain a DAG. For each parametrization, we sample both a 5-node and a 10-node causal model. For each causal model, we sample several training datasets with different number of samples: 250, 500, 1000 and 2000. This results in $4 \cdot 2 \cdot 4 = 16$ training datasets.

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, Aliferis, and Statnikov, 2003a), IAMB (Tsamardinos, Aliferis, and Statnikov, 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 fix the parameter estimation as an MLE estimates based on the training data. When only a partial DAG is returned, we use the edge orientation which provides the best goodness of fit after the parameters have been estimated.

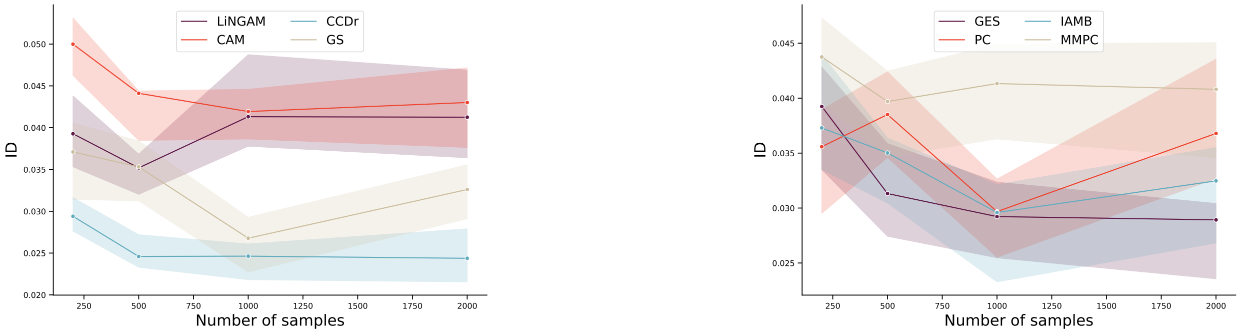
Our distances like ID compare full causal models, thus causal graph after the parameters have been estimated. Here, by fixing the parameter estimation, we measure the impact of the causal graph on the intervention predictions. It also ensures that two methods which output the same graph (DAG or partial DAG) will obtain the evaluation results.

For the systems, we use the implementations available in CDT[4]. We compute MLE estimates using the Pomegranate framework.[5]

Results. The results on real networks are reported in Table 1. We observe that different metrics produce different rankings of systems. This shows that the differences between metrics observed in Sec. 7.2 are also visible in the causal discovery evaluation setup. In particular, we observe low agreement between SID and ID on the Earthquake and Insurance networks. Also, IAMB and MMPC have the same SID (16) and SHD (7) on Cancer2 but different graphs which is distinguished by ID. On the contrary, on the Protein dataset, GS and IAMB have the same graph and, with fixed parameter estimation, the same SHD, SID and ID.

[4] https://github.com/Diviyan-Kalainathan/CausalDiscoveryToolbox

[5] https://pomegranate.readthedocs.io



(a) LiNGAM, CAM, CCDr, and GS.

(b) GES, PC, IAMB, and MMPC.

Figure 8: Variation in performance of causal discovery systems measured by ID when more training data is available.

These observations emphasize the importance of employing the evaluation metric which 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.

A peculiarity of SID is that it outputs integers only, which can result in ties, whereas ID produce continuous values and do not have this problem. Furthermore, ID is normalized by default whereas SID and SHD greatly vary depending on the number of nodes.

Additionally, even a single metric produces different rankings of systems for different networks. 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.

Thus, we also perform an evaluation of causal discovery system broken down by model parametrization, shown in Fig. 7. The *Real Networks* block corresponds to results of Table 1 averaged across networks for comparison.

Interestingly, ID clearly reveals that systems struggle most for the linear Gaussian case, which is known to be unidentifiable without further assumptions. While the other cases are identifiable, the non-linear additive case seems to be the easiest for existing systems. Overall, CCDr seems to perform fairly well in comparison to other systems.

Finally, since we generated several datasets with varying number of training samples, we can measure how well systems benefit from more training data. This is reported in Fig. 8a and Fig. 8b, where the systems are arbitrarily split into two groups to avoid overcrowding one figure. Interestingly, systems seem to not clearly benefit from accessing more training data. In fact, it is a particularity of causal inference that even infinite observational data does not necessarily help to infer the causal model.

8 Applications and future work

A straightforward application of causal distances concerns the evaluation of causal discovery methods in comparison to a known gold standard as demonstrated by Sec. 7.5. Furthermore, ID can be used to train and fine-tune hyper-parameters of causal discovery algorithms using a validation set of data with ground-truth causal models.

The causal distances also have many other applications. For instance, one can cluster causal models with similar answers to causal queries (interventions or counterfactuals). Alternatively, by considering OD and ID together, one can quantitatively understand the important cases where two models have similar joint distributions but deviate largely in their responses to interventions. These are the cases where a causal understanding is critical because the joint distributions are greatly warped under interventions.

Furthermore, one can embed causal models into vector spaces based on ID or CD as demonstrated by Sec. 7.1. This could be beneficial for performing structure and parameter search in continuous spaces without discarding the interventional properties of the causal model.

The causal distances can also be extended to time series. For example, an expected interventional distance between two causal models at time step t, ID^t, can easily be defined. Then, one can study the expected differences $\{\mathrm{ID}^t\}_{t\in[T,T+\tau]}$ within a time interval $[T, T+\tau]$ aggregated through time or visualized itself as a time series. This could have important practical applications in comparing candidate models of dynamical systems upon which we might perform policy changes, e.g., economic models, climate models, etc..

Interestingly, ID can also be used as a measure of the resilience of a system by measuring the expected deviation between the system and itself under small random external modifications. Intuitively, a system is resilient if it does not change much under small unexpected perturbations, i.e., ID should remain small when P_I has small variance.

Finally, most causal models currently used as benchmarks (cf. Sec. 7.5) have a fairly low number of nodes, and in such cases, our proposed estimates are sample-efficient. When, however, the number of nodes is large, one needs to compare joint distributions of potentially high dimensionality. While in principle any specialized technique to compare high-dimensional distributions may be chosen as D, scalability becomes an issue that deserves further investigation.

9 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.[6]

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 required by 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.

References

Acharya, J.; Bhattacharyya, A.; Daskalakis, C.; and Kandasamy, S. 2018. Learning and testing causal models with interventions. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc. 9447–9460.

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

Aragam, B., and Zhou, Q. 2015. Concave Penalized Estimation of Sparse Gaussian Bayesian Networks. *Journal of Machine Learning Research* 16(1):2273–2328.

Binder, J.; Koller, D.; Russell, S.; and Kanazawa, K. 1997. Adaptive Probabilistic Networks with Hidden Variables. *Machine Learning* 29(2–3):213–244.

Chickering, D. M., and Meek, C. 2002. Finding Optimal Bayesian Networks. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, 94–102. Morgan Kaufmann Publishers Inc.

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

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

Elidan, G. 2001. Bayesian Network Repository. http://www.cs.huji.ac.il/site/labs/compbio/Repository.

Gentzel, A.; Garant, D.; and Jensen, D. 2019. The case for evaluating causal models using interventional measures and empirical data. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. 11717–11727.

Korb, K., and Nicholson, A. 2010. *Bayesian Artificial Intelligence*. Chapman and Hall, 2nd edition.

[6] `available-after-reviewing.`

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 4066–4076.

Lauritzen, S. L., and Spiegelhalter, D. J. 1990. Local computations with probabilities on graphical structures and their application to expert systems. In Shafer, G., and Pearl, J., eds., *Readings in Uncertain Reasoning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 415–448.

Lipton, Z. C. 2018. The mythos of model interpretability. *Commun. ACM* 61(10):36–43.

Pearl, J., and Mackenzie, D. 2018. *The Book of Why*. New York: Basic Books.

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

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

Peters, J. M., and Bühlmann, P. 2015. Structural intervention distance for evaluating graphs. *Neural Computation* 27(3):771–799.

Peters, J. M.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press.

Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; and Peters, J. 2018. Invariant models for transfer learning. *Journal of Machine Learning Research* 19(36):1–34.

Sachs, K.; Perez, O.; Pe'er, D.; Lauffenburger, D. A.; and Nolan, G. P. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721):523–529.

Scutari, M., and Denis, J.-B. 2014. *Bayesian networks: with examples in R*. Chapman and Hall/CRC.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research* 7(Oct):2003–2030.

Singh, K.; Gupta, G.; Tewari, V.; and Shroff, G. 2017. Comparative benchmarking of causal discovery techniques. *arXiv preprint arXiv:1708.06246*.

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, Prediction, and Search*. MIT press.

Sriperumbudur, B. K.; Gretton, A.; Fukumizu, K.; Schölkopf, B.; and Lanckriet, G. R. 2010. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research* 11:1517–1561.

Theis, L.; Oord, A. v. d.; and Bethge, M. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*.

Tsamardinos, I.; Aliferis, C. F.; and Statnikov, A. 2003a. 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*, 673678. New York, NY, USA: Association for Computing Machinery.

Tsamardinos, I.; Aliferis, C. F.; and Statnikov, A. R. 2003b. Algorithms for Large Scale Markov Blanket Discovery. In *FLAIRS Conference*, 376–381. AAAI Press.

Verma, T., and Pearl, J. 1991. Equivalence and Synthesis of Causal Models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, 255–270.

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

A Details about the Sensitivity Experiment

A.1 Perturbating ID while leaving OD constant

We know that if two causal graphs are within the same Markov class they can support the same observational distribution (Verma and Pearl, 1991). Thus, we take a causal model $\mathfrak{C}$ with graph $\mathscr{G}$ and compute its Markov equivalence class $\mathscr{M}(\mathscr{G})$.

We then consider a causal graph $\mathscr{H} \in \mathscr{M}(\mathscr{G})$ from the Markov equivalence class and select the perturbation quantification ϵ as:

$$\epsilon = \frac{\mathrm{SID}(\mathscr{H},\mathscr{G})}{\max\{\mathrm{SID}(\mathscr{H},\mathscr{G}) | \mathscr{H} \in \mathscr{M}(\mathscr{G})\}}. \tag{14}$$

Then, we train an MLE parameter estimator using $\mathscr{H}$ to find the parameters that yield (almost) the same observational distribution as $\mathfrak{C}$. Thus, OD is expected to be (almost) constant while ID is perturbated. In particular, when $\epsilon = 0$, ID is expected to be 0.

A.2 Perturbating CD while leaving ID constant

The interventional distributions remain unchanged if all the conditional distributions $P(X|\mathbf{PA}_X)$ do not change. To preserve the interventional distributions, we can perturbate the structural equations like described in Sec. 7.4 to generate Fig. 6b, but also adjust the noise to precisely cancel the perturbation and keep the conditional distribution constant.

In practice, at one node X, we perturbate the noise distribution by adding a random Gaussian Mixture $GMM(k,\mu,\sigma)$. Here k is the number of Gaussians, μ is an k-dimensional vector of means and Σ the covariance matrix.

$$P_{N_X}^{(\epsilon)} := (1-\epsilon)P_{N_X} + \epsilon GMM(k,\mu,\Sigma) \tag{15}$$

Here, ϵ quantifies the perturbation. In order to preserve the conditional probability distribution $P(X|\mathbf{PA}_X)$ we fit a Gaussian process g_X^{ϵ} such that:

$$g_X^{(\epsilon)}(\mathbf{PA}_X, P_{N_X}^{(\epsilon)}) \approx f_X(\mathbf{PA}_X, P_{N_X}) \tag{16}$$

Thus, ID is expected to stay (almost) fixed while CD is expected to be affected because the noise is changed. In particular, when ϵ is 0 the causal model is not modified and when ϵ is 1 the noise is fully replaced by the random Gaussian Mixture.

B Proofs

B.1 Preliminaries

Assumptions Throughout the proofs, we assume that all models satisfy:

- *Markov property*: Every conditional independence statement entailed by the causal graph is satisfied by the joint distribution. :$\forall\ \mathbf{W},\mathbf{Y},\mathbf{Z} \in 2^{\mathbf{X}},\ \mathbf{W} \perp\!\!\!\perp_{\mathscr{G}} \mathbf{Y}|\mathbf{Z} \implies \mathbf{W} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$, where $\perp\!\!\!\perp_{\mathscr{G}}$ stands for d-separated in $\mathscr{G}$ (Pearl, 2009).
- *Causal minimality*: The joint distribution satisfies the Markov property for $\mathscr{G}$ but not for any proper subgraph of $\mathscr{G}$.
- *Causal faithfulness*: Every conditional independence within the joint distribution is entailed by the causal graph. :$\forall\ \mathbf{W},\mathbf{Y},\mathbf{Z} \in 2^{\mathbf{X}},\ \mathbf{W} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z} \implies \mathbf{W} \perp\!\!\!\perp_{\mathscr{G}} \mathbf{Y}|\mathbf{Z}$
- *Positiveness*: The entailed marginal and conditional distributions are strictly positive.

Then, we say that a node $X \in \mathbf{X}$ *has an effect on* another node $Y \in \mathbf{X}$ when:

$$\exists x \neq x',\ \ P_Y^{\mathfrak{C};\mathrm{do}(X=x)} \neq P_Y^{\mathfrak{C};\mathrm{do}(X=x')} \tag{17}$$

- If X has an effect on Y, then X is an ancestor of Y.
- If X is a parent of Y, then X has an effect on Y except if there exists a canceling path, i.e., $\exists Z_1 \dots Z_k,\ \ X \to Z_1 \to \dots \to Z_k \to Y$ which precisely cancels the effect $X \to Y$.

We order the proofs out of convenience instead of following the order in which they appear in the paper.

B.2 Proof of Theorem 1

Theorem 1. *For two causal models $\mathfrak{C}_1$ and $\mathfrak{C}_2$ over the variables $\mathbf{X}$, we have, for all $\epsilon \geq 0$:*

$$\mathrm{CD}(\mathfrak{C}_1,\mathfrak{C}_2) \leq \epsilon \implies \mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) \leq (|\mathbf{X}|+1)\epsilon \tag{10}$$

$$\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) \leq \epsilon \implies \mathrm{OD}(\mathfrak{C}_1,\mathfrak{C}_2) \leq (|\mathbf{X}|+1)\epsilon \tag{11}$$

Proof. Let $\mathfrak{C}_1$ and $\mathfrak{C}_2$ be two models such that, for some $\epsilon \geq 0$:

$$\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) \leq \epsilon \tag{18}$$

We note $\mathrm{OD}_i = \mathrm{OD}(\mathfrak{C}_1;\mathrm{do}(I=i),\mathfrak{C}_2;\mathrm{do}(I=i))$, the distance between the interventional distributions resulting from $\mathrm{do}(I=i)$. Then, ID can be decomposed as:

$$\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) = \tag{19}$$

$$\frac{1}{|\mathbf{X}|+1}\left(\mathrm{OD}(\mathfrak{C}_1,\mathfrak{C}_2) + \sum_{I\in\mathbf{X}} \mathbb{E}_{i\sim P_I}\,\mathrm{OD}_i\right) \tag{20}$$

$$\tag{21}$$

Since OD is a distance between distributions, OD_i is positive, and the expectations inside the sum are positive. Finally:

$$\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) \leq \epsilon \tag{22}$$

$$(|\mathbf{X}|+1)\,\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) \leq (|\mathbf{X}|+1)\epsilon \tag{23}$$

$$\mathrm{OD}(\mathfrak{C}_1,\mathfrak{C}_2) \leq (|\mathbf{X}|+1)\epsilon \tag{24}$$

The same reasoning gives:

$$\mathrm{CD}(\mathfrak{C}_1,\mathfrak{C}_2) \leq \epsilon \implies \mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) \leq (|\mathbf{X}|+1)\epsilon \tag{25}$$

□

B.3 Proof of Theorem 2

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

$$\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) = 0 \implies \mathscr{G}_1 = \mathscr{G}_2 \implies \mathrm{SHD}(\mathscr{G}_1,\mathscr{G}_2) = 0 \implies \mathrm{SID}(\mathscr{G}_1,\mathscr{G}_2) = 0. \tag{12}$$

The reverse direction of (12) does not hold in general.

Proof. Suppose $\mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) = 0$. From Thm. 1, we know that $\mathrm{OD}(\mathfrak{C}_1,\mathfrak{C}_2) = 0$.

The two models belong to the same Markov equivalence class. Thus, they have the same skeleton and v-structures (Verma and Pearl, 1991). Furthermore, orienting new edges cannot create new v-structures.

Some edges may still be oriented differently. For example, consider the edge between X and Y left unoriented in the Markov equivalence class. Without loss of generality, suppose $X \to Y$ in $\mathscr{G}_1$.

Now, X has an effect on Y in $\mathscr{G}_1$ because there cannot be a cancelling path. If there were a cancelling path, the orientation $X \to Y$ would create new v-structure. Since X has an effect on Y in $\mathscr{G}_1$, X also has an effect on Y in $\mathscr{G}_2$ because the models agree on any interventions. Finally, the edge goes from X to Y in $\mathscr{G}_2$.

Thus, we conclude $\mathscr{G}_1 = \mathscr{G}_2$. Finally, SID and SHD only consider the adjacency matrices and therefore they are also 0. Peters and Bühlmann (2015) proved that $\mathrm{SHD}(\mathfrak{C}_1,\mathfrak{C}_2) = 0 \implies \mathrm{SID}(\mathfrak{C}_1,\mathfrak{C}_2) = 0$. A counterexample to the converse implication of (12) is given by the models of the case study presented in the paper. □

B.4 Proof of Theorem 3

Theorem 3. *For two causal models $\mathfrak{C}_1,\mathfrak{C}_2$ with causal graphs $\mathscr{G}_1,\mathscr{G}_2$. When $\mathrm{OD}(\mathfrak{C}_1,\mathfrak{C}_2) = 0$, we have:*

$$\mathrm{SID}(\mathscr{G}_1,\mathscr{G}_2) = 0 \iff \mathrm{ID}(\mathfrak{C}_1,\mathfrak{C}_2) = 0. \tag{13}$$

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

Proof. Suppose $\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) = 0$. Then, the two models have graphs belonging to the same Markov equivalence class, i.e., same skeleton and v-structures (Verma and Pearl, 1991).

We already know from Thm. 2 that $\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0 \implies \mathrm{SID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0$.

Suppose $\mathrm{SID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0$. Then, no edge between any two nodes can be oriented differently in the two graphs. To see that, consider the edge between X and Y left unoriented in the Markov equivalence class. Without loss of generality, suppose $X \to Y$ in $\mathscr{G}_1$. Now, X has an effect on Y in $\mathscr{G}_1$ because there cannot be a cancelling path. If there were a cancelling path, the orientation $X \to Y$ would create new v-structure. Since X has an effect on Y in $\mathscr{G}_1$, X also has an effect on Y in $\mathscr{G}_2$ because the models agree on any interventions. Finally, the edge goes from X to Y in $\mathscr{G}_2$. Therefore, the two graphs are the same.

Since the two graphs are the same and the observation distributions are the same, we can conclude that $\mathrm{ID}(\mathfrak{C}_1, \mathfrak{C}_2) = 0$.

A counter-example when $\mathrm{OD}(\mathfrak{C}_1, \mathfrak{C}_2) \neq 0$ is given by the case study presented in the paper. □