LETTER **Communicated by Martijn de Jongh**

Structural Intervention Distance for Evaluating Causal Graphs

Jonas Peters
peters@stat.math.ethz.ch
Peter Bühlmann
buhlmann@stat.math.ethz.ch
Seminar for Statistics, Department of Mathematics, ETH Zürich 8092, Switzerland

Causal inference relies on the structure of a graph, often a directed acyclic graph (DAG). Different graphs may result in different causal inference statements and different intervention distributions. To quantify such differences, we propose a (pre-)metric between DAGs, the structural intervention distance (SID). The SID is based on a graphical criterion only and quantifies the closeness between two DAGs in terms of their corresponding causal inference statements. It is therefore well suited for evaluating graphs that are used for computing interventions. Instead of DAGs, it is also possible to compare CPDAGs, completed partially DAGs that represent Markov equivalence classes. The SID differs significantly from the widely used structural Hamming distance and therefore constitutes a valuable additional measure. We discuss properties of this distance and provide a (reasonably) efficient implementation with software code available on the first author's home page.

1 Introduction

Given a true causal directed acyclic graph (DAG) $\mathcal{G}$, we may want to assess the goodness of an estimate $\mathcal{H}$. The structural Hamming distance (SHD; see definition 1) counts the number of incorrect edges. Although this provides an intuitive distance between graphs, it does not reflect their capacity for causal inference. Instead, we propose to count the pairs of vertices (i, j), for which the estimate $\mathcal{H}$ correctly predicts intervention distributions within the class of distributions that are Markov with respect to $\mathcal{G}$. This results in a new (pre-)metric[1] between DAGs, the structural intervention distance, which adds valuable additional information to the established SHD.

Throughout this work, we consider a vector of random variables $\mathbf{X} = (X_1, \ldots, X_p)$ with index set $\mathbf{V} := \{1, \ldots, p\}$ (we use capital letters for random variables and bold letters for sets or vectors). We denote their joint

[1] We use the terms *distance* and *metric* interchangeably.

Neural Computation **27, 771–799** (2015)
doi:10.1162/NECO_a_00708

distribution by $\mathcal{L}(\mathbf{X})$ and denote corresponding densities of $\mathcal{L}(\mathbf{X})$ with respect to Lebesgue or the counting measure, by $p(\cdot)$ (implicitly assuming their existence). We also denote conditional densities and the density of $\mathcal{L}(\mathbf{Z})$ with $\mathbf{Z} \subset \mathbf{X}$ by $p(\cdot)$. A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of nodes $\mathbf{V}$ and edges $\mathcal{E} \subseteq \mathbf{V} \times \mathbf{V}$. With a slight abuse of notation, we sometimes identify the nodes (or vertices) $j \in \mathbf{V}$ with the variables X_j. In appendix A, we provide further terminology regarding DAGs (Lauritzen, 1996; Spirtes, Glymour, & Scheines, 2000; Koller & Friedman, 2009), which we require in our work.

This letter is organized as follows: Sections 1.1 and 1.2 review the structural Hamming distance and define intervention distributions (Pearl, 2009), respectively. In section 2 we introduce the new structural intervention distance, prove some of its properties and provide possible extensions. Section 3 contains experiments on synthetic data, and section 4 describes an efficient implementation of the SID.

1.1 Structural Hamming Distance. The structural Hamming distance (Acid & de Campos, 2003; Tsamardinos, Brown, & Aliferis, 2006) considers two partially directed acyclic graphs (PDAGs; see appendix A) and counts how many edges do not coincide.

Definition 1 *(structural Hamming distance). Let $\mathbb{P}$ be the space of PDAGs over p variables. The structural Hamming distance (SHD) is defined as*

$$\begin{aligned} SHD: \ & \mathbb{P} \times \mathbb{P} \to \mathbb{N} \\ & (\mathcal{G}, \mathcal{H}) \mapsto \#\{(i, j) \in \boldsymbol{V} \times \boldsymbol{V} \mid \mathcal{G} \text{ and } \mathcal{H} \text{ do not have the same type} \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{of edge between } i \text{ and } j\}, \end{aligned}$$

where edge types are defined in appendix A (no connection is also a type of an edge).

Equivalently, we count pairs (i, j), such that $\big((i, j) \in \mathcal{E}_{\mathcal{G}} \Delta \mathcal{E}_{\mathcal{H}}\big)$ or $\big((j, i) \in \mathcal{E}_{\mathcal{G}} \Delta \mathcal{E}_{\mathcal{H}}\big)$, where $A \Delta B := (A \setminus B) \cup (B \setminus A)$ is the symmetric difference. Definition 1 includes a distance between two DAGs since these are special cases of PDAGs. In this work, the SHD is primarily used as a measure of reference when comparing with our new structural intervention distance. A comparison to other similar structural distances (e.g., counting only missing edges) can be found in de Jongh and Druzdzel (2009); all distances they consider are of similar type as SHD.

1.2 Intervention Distributions. Assume that $\mathcal{L}(\mathbf{X})$ is absolutely continuous with respect to a product measure. Then $\mathcal{L}(\mathbf{X})$ is Markov with respect to $\mathcal{G}$ if and only if the joint density factorizes according to

$$p(x_1, \ldots, x_p) = \prod_{j=1}^{p} p(x_j \mid \mathbf{x}_{\mathbf{pa}_j})$$

(see Lauritzen, 1996, for example). The intervention distribution given $\mathrm{do}(X_i = \hat{x}_i)$ is then defined as

$$p_{\mathcal{G}}(x_1, \ldots, x_p \mid \mathrm{do}(X_i = \hat{x}_i)) = \prod_{j \neq i} p(x_j \mid \mathbf{x}_{\mathbf{pa}_j}) \delta(x_i = \hat{x}_i). \tag{1.1}$$

This, again, is a probability distribution. We can therefore take expectations or marginalize over some of the variables. A total effect from X to Y, for example,[2] is often defined as a difference between the distributions $p_{\mathcal{G}}(y \mid \mathrm{do}(X = \hat{x}))$ and $p(y)$. One can check (see the proof of proposition 2) that equation 1.1 implies $p_{\mathcal{G}}(y \mid \mathrm{do}(X = \hat{x})) = p(y)$ if Y is a nondescendant of X; intervening on X does not show any effect on the distribution of Y. If Y is not a parent of X, we can compute (marginalized) intervention distributions by taking into account only a subset of variables from the graph (Pearl, 2009).

Proposition 1 *(adjustment formula for parents). Let $X \neq Y$ be two different nodes in $\mathcal{G}$. If Y is a parent of X, then*

$$p_{\mathcal{G}}(y \mid do(X = \hat{x})) = p(y). \tag{1.2}$$

If Y is not a parent of X, then

$$p_{\mathcal{G}}(y \mid do(X = \hat{x})) = \sum_{\boldsymbol{pa}_X} p(y \mid \hat{x}, \boldsymbol{pa}_X)\, p(\boldsymbol{pa}_X), \tag{1.3}$$

where the sum is over all possible assignments of $\mathbf{PA}_X^{\mathcal{G}}$, the parents of node X in graph $\mathcal{G}$. Whenever we can compute the marginalized intervention distribution $p(y \mid \mathrm{do}(X = \hat{x}))$ by a summation $\sum_{\mathbf{z}} p(y \mid \hat{x}, \mathbf{z})\, p(\mathbf{z})$ as in equation 1.3, we call the set $\mathbf{Z}$ a valid adjustment set for the intervention $Y \mid \mathrm{do}(X)$. Proposition 1 states that $\mathbf{Z} = \mathbf{PA}_X^{\mathcal{G}}$ is a valid adjustment set for $Y \mid \mathrm{do}(X)$ if $Y \notin \mathbf{PA}_X^{\mathcal{G}}$. Section 2.2 shows that for a given graph, there may be other valid adjustment sets too.

2 Structural Intervention Distance

2.1 Motivation and Definition. We propose a new graph-based (pre-)metric, the structural intervention distance (SID). When comparing graphs (or DAGs in particular), there are many (pre-)metrics, one could consider that an appropriate choice should depend on the further use and purpose of the graphs. Often one is interested in a causal interpretation of a

[2] We sometimes denote variables by different letters rather than indices in order to avoid subscripts.

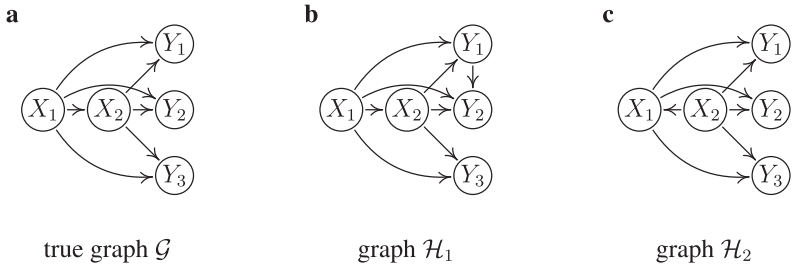

Figure 1: Panels b and c have the same SHD to the true graph, panel a, but differ in the SID.

graph that enables us to predict the result of interventions. We then require a distance that takes this goal into account. From now on, we implicitly assume that an intervention distribution is computed using adjustment for parents as in proposition 1; we discuss other choices of adjustment sets in section 2.4.5. Example 1 shows that the SHD (see definition 1) is not optimal in terms of capturing aspects of the graph that are related to intervention distributions.

Example 1. Figure 1 shows a true graph $\mathcal{G}$ (a) and two different graphs (e.g., estimates) $\mathcal{H}_1$ (b) and $\mathcal{H}_2$ (c). The only difference between $\mathcal{H}_1$ and $\mathcal{G}$ is the additional edge $Y_1 \to Y_2$, and the only difference between $\mathcal{H}_2$ and $\mathcal{G}$ is the reversed edge between X_1 and X_2. The SHD between the true DAG and the others is therefore one in both cases:

$$\mathrm{SHD}(\mathcal{G}, \mathcal{H}_1) = 1 = \mathrm{SHD}(\mathcal{G}, \mathcal{H}_2).$$

We now consider a distribution $p(.)$ that is Markov with respect to $\mathcal{G}$ and compute all intervention distributions using parent adjustment, equation 1.3. We will see that these two "mistakes" have different impacts on the correctness of those intervention distributions.

First, we consider the DAG $\mathcal{H}_1$. All nodes except Y_2 have the same parent sets in $\mathcal{G}$ and $\mathcal{H}_1$, and thus the parent adjustment implies exactly the same formula for interventions on nodes other than Y_2. Since X_1 and X_2 are parents of Y_2 in both graphs, the intervention distributions from Y_2 to X_1 and X_2 are correct. We will now argue why $\mathcal{G}$ and $\mathcal{H}_1$ agree on the intervention distribution from Y_2 to Y_3 and from Y_2 to Y_1. When computing the intervention distribution from Y_2 to Y_3 in $\mathcal{H}_1$, we adjust not only for $\{X_1, X_2\}$ as done in $\mathcal{G}$ but also for the additional parent Y_1. We thus have to check whether $\{X_1, X_2, Y_1\}$ is a valid adjustment set for $Y_3 \mid \mathrm{do}(Y_2)$. Indeed, since $Y_2 \perp\!\!\!\perp Y_1 \mid \{X_1, X_2\}$ (the distribution is Markov with respect to $\mathcal{G}$) we

have:

$$
\begin{aligned}
&p_{\mathcal{H}_1}(y_3 \mid \mathrm{do}(Y_2 = \hat{y}_2)) \\
&= \sum_{x_1,x_2,y_1} p(y_3 \mid x_1, x_2, y_1, \hat{y}_2) p(x_1, x_2, y_1) \\
&= \sum_{x_1,x_2,y_1} \frac{p(x_1, x_2, y_1, \hat{y}_2, y_3)}{p(\hat{y}_2 \mid x_1, x_2, y_1)} = \sum_{x_1,x_2,y_1} \frac{p(x_1, x_2, y_1, \hat{y}_2, y_3)}{p(\hat{y}_2 \mid x_1, x_2)} \\
&= \sum_{x_1,x_2} p(y_3 \mid x_1, x_2, \hat{y}_2) p(x_1, x_2) = p_{\mathcal{G}}(y_3 \mid \mathrm{do}(Y_2 = \hat{y}_2))
\end{aligned}
$$

It remains to show that $p_{\mathcal{G}}(y_1 \mid \mathrm{do}(Y_2 = \hat{y}_2)) = p(y_1) = p_{\mathcal{H}_1}(y_1 \mid \mathrm{do}(Y_2 = \hat{y}_2))$, where the last equality is given by equation 1.2. But since $Y_1 \perp\!\!\!\perp Y_2 \mid X_1, X_2$, it follows from the parent adjustment, equation 1.3, that $p_{\mathcal{G}}(y_1 \mid \mathrm{do}(Y_2 = \hat{y}_2)) = p(y_1)$. Thus, all intervention distributions computed in $\mathcal{H}_1$ agree with those computed in $\mathcal{G}$. Proposition 3 shows that this is not a coincidence. It proves that all estimates for which the true DAG is a subgraph correctly predict the intervention distributions.

The "mistake" in graph $\mathcal{H}_2$, namely, the reversed edge, is more severe. For computing the correct intervention distribution from X_2 to Y_1, for example, we need to adjust for the confounder X_1, as suggested by the parent adjustment, equation 1.3, applied to $\mathcal{G}$. In $\mathcal{H}_2$, however, X_2 does not have any parent, so there is no variable adjusted for. In general, $\mathcal{H}_2$ therefore leads to a wrong intervention distribution $p_{\mathcal{H}_2}(y_1 \mid \mathrm{do}(X_2 = \hat{x}_2)) \neq p_{\mathcal{G}}(y_1 \mid \mathrm{do}(X_2 = \hat{x}_2))$. Also, when computing the intervention distribution from X_1 to Y_i, $i = 1, 2, 3$, we are adjusting for X_2, which is now a parent of X_1 in $\mathcal{H}_2$. Again, this may lead to $p_{\mathcal{H}_2}(y_i \mid \mathrm{do}(X_1 = \hat{x}_1)) \neq p_{\mathcal{G}}(y_i \mid \mathrm{do}(X_1 = \hat{x}_1))$. Further, the intervention distributions from X_1 to X_2 and from X_2 to X_1 may not be correct either. In fact, $\mathcal{H}_2$ makes eight erroneous predictions for many observational distributions $p(.)$.

The preceding deliberations are reflected by the structural intervention distance we propose below (see definition 3). We will see that

$$\mathrm{SID}(\mathcal{G}, \mathcal{H}_1) = 0 \neq 8 = \mathrm{SID}(\mathcal{G}, \mathcal{H}_2).$$

Proposition 2 shows us how to read off the SID from the graph structures. It may come as a surprise that in this example, the "incorrect" graph $\mathcal{H}_1$ still obtains a distance of zero; it is due to the fact that $\mathcal{G}$ is a subgraph of $\mathcal{H}_1$ (see proposition 3). Section 2.4.3 shows how a small modification yields a measure that is zero only for the correct graph.

The following argumentation motivates the formal defintion of the SID. Given a true DAG $\mathcal{G}$ and an estimate $\mathcal{H}$, we would like see whether an intervention distribution, which is computed using the structure of $\mathcal{H}$, coincides with the "true" intervention distributions inferred from $\mathcal{G}$. This depends, however, on the observational distribution over all variables. Since we regard $\mathcal{G}$ as the ground truth, we assume that the observational distribution is Markov with respect to $\mathcal{G}$. Consider now a specific distribution that factorizes over all nodes: all variables are independent (this distribution is certainly Markov with respect to $\mathcal{G}$). Then $\mathcal{G}$ and $\mathcal{H}$ agree on all intervention distributions, even though their structure can be arbitrarily different. We therefore consider all distributions that are Markov with respect to $\mathcal{G}$ instead of only one: we count all pairs of nodes, for which the predicted interventions agree for all observational distributions that are Markov with respect to $\mathcal{G}$. Those pairs are said to "correctly infer" the intervention distribution.

Definition 2 *(correctly and falsely inferred intervention distributions). Let $\mathcal{G}$ and $\mathcal{H}$ be DAGs over variables $\boldsymbol{X} = (X_1, \ldots, X_p)$. For $i \neq j$ we say that the intervention distribution from i to j is correctly inferred by $\mathcal{H}$ with respect to $\mathcal{G}$ if*

$$p_{\mathcal{G}}(x_j \mid do(X_i = \hat{x}_i)) = p_{\mathcal{H}}(x_j \mid do(X_i = \hat{x}_i))$$
$$\forall \mathcal{L}(\boldsymbol{X}) \text{ Markov with regard to } \mathcal{G} \text{ and } \forall \hat{x}_i.$$

Otherwise, that is, if

$$\exists \mathcal{L}(\boldsymbol{X}) \text{ Markov with regard to } \mathcal{G} \text{ and } \hat{x}_i \text{ with}$$
$$p_{\mathcal{G}}(x_j \mid do(X_i = \hat{x}_i)) \neq p_{\mathcal{H}}(x_j \mid do(X_i = \hat{x}_i)),$$

we call the intervention distribution from i to j falsely inferred by $\mathcal{H}$ with respect to $\mathcal{G}$. Here, $p_{\mathcal{G}}$ and $p_{\mathcal{H}}$ are computed using parent adjustment as in proposition 1 (section 2.4.5 discusses an alternative to parent adjustment).

The SID counts the number of falsely inferred intervention distributions. The definition is independent of any distribution, which is crucial to allow for a purely graphical characterization.

Definition 3 *(Structural intervention distance). Let $\mathbb{G}$ be the space of DAGs over p variables. We then define*

$$\begin{aligned} SID : \ & \mathbb{G} \times \mathbb{G} \to \mathbb{N} \\ & (\mathcal{G}, \mathcal{H}) \mapsto \#\{(i, j), i \neq j \mid \text{the intervention distribution from } i \text{ to } j \\ & \qquad\qquad \text{is falsely inferred by } \mathcal{H} \text{ with respect to } \mathcal{G}\} \end{aligned} \tag{2.1}$$

as the structural intervention distance (SID).

The SID does not satisfy all properties of a metric; in particular it is not symmetric. However, section 2.3 shows that it is a (pre-)metric. In section 2.4.4, we introduce a symmetrized version of the SID.

Remark 1. Definition 2 remains exactly the same if we *also* allow for so-called soft interventions. For these we replace a variable X_i with another random variable that has density $\hat{p}$ and is independent of $X_1, \dots, X_p$. We define the corresponding intervention distribution as

$$p_{\mathcal{G}}\big(x_1, \dots, x_p \mid \mathrm{do}(X_i \sim \hat{p})\big) := \prod_{k \neq i} p(x_k \mid \mathbf{x}_{\mathbf{pa}_k}) \hat{p}(x_i). \tag{2.2}$$

If, for example, the intervention distribution from i to j is correctly inferred by $\mathcal{H}$ with respect to $\mathcal{G}$ using only "hard" interventions, equation 1.1, we can imply the same statement for soft interventions, equation 2.2, because

$$p_{\mathcal{G}}\big(x_j \mid \mathrm{do}(X_i \sim \hat{p})\big) = \int p_{\mathcal{G}}\big(x_j \mid \mathrm{do}(X_i = \hat{x}_i)\big)\, \hat{p}(\hat{x}_i)\, d\hat{x}_i.$$

2.2 An Equivalent Formulation. The SID as defined in equation 2.1 is difficult to compute. We now provide an equivalent formulation that is based on graphical criteria only. We will see that for each pair (i, j), the question becomes whether $\mathbf{PA}_{X_i}^{\mathcal{H}}$ is a valid adjustment set for the intervention $X_j \mid \mathrm{do}(X_i)$ in graph $\mathcal{G}$. Shpitser, der Weele, and Robins (2010) prove the following characterization of adjustment sets (note that we use a slightly simpler condition). The reader may think of $\mathbf{Z} = \mathbf{PA}_X^{\mathcal{G}}$, which is always a valid adjustment set, as stated in proposition 1.

Lemma 1 *(characterization of valid adjustment sets). Consider a DAG $\mathcal{G} = (\boldsymbol{V}, \mathcal{E})$, variables $X, Y \in \boldsymbol{V}$ and a subset $\boldsymbol{Z} \subset \boldsymbol{V} \setminus \{X, Y\}$. Consider the property of $\boldsymbol{Z}$ with regard to $(\mathcal{G}, X, Y)$:*

$$(*) \begin{cases} \text{In } \mathcal{G}\text{, no } Z \in \boldsymbol{Z} \text{ is a descendant of any } W \neq X \text{ which lies on a directed} \\ \quad \text{path from } X \text{ to } Y \text{ and } \boldsymbol{Z} \text{ blocks all nondirected paths from } X \text{ to } Y. \end{cases}$$

We then have the following two statements:

(i) Let $\mathcal{L}(\boldsymbol{X})$ be Markov with respect to $\mathcal{G}$. If $\boldsymbol{Z}$ satisfies $(*)$ with regard to $(\mathcal{G}, X, Y)$, then $\boldsymbol{Z}$ is a valid adjustment set for $Y \mid do(X)$.

(ii) If $\boldsymbol{Z}$ does not satisfy $(*)$ with regard to $(\mathcal{G}, X, Y)$, then there exists $\mathcal{L}(\boldsymbol{X})$ that is Markov with respect to $\mathcal{G}$ that leads to $p_{\mathcal{G}}(y \mid do(X = \hat{x})) \neq \sum_z p(y \mid \hat{x}, z)\, p(z)$, meaning $\boldsymbol{Z}$ is not a valid adjustment set.

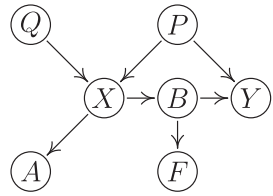

Figure 2: The sets $\mathbf{Z} = \{P, Q\}$ and $\mathbf{Z} = \{P, A\}$ are valid adjustment sets for $Y \mid \mathrm{do}(X)$; $\mathbf{Z} = \{P\}$ is the smallest adjustment set. Any set containing F, however, cannot be a valid adjustment set. See lemma 1.

If $Y \notin \mathbf{PA}_X^{\mathcal{G}}$, then $\mathbf{Z} = \mathbf{PA}_X^{\mathcal{G}}$ satisfies condition $(*)$ and statement i reduces to proposition 1. In fact, condition $(*)$ is a slight extension of the backdoor criterion (Pearl, 2009). It is not surprising that sets other than the parent set work too. We may adjust for children of X, for example, as long as they are not part of a directed path (see Figure 2). Similarly, we do not have to adjust for parents of X for which all unblocked paths to Y lead through X.

Using lemma 1, we obtain the following equivalent definition of the SID, which is entirely graph based and will later be exploited for computation:

Proposition 2. *The SID has the following equivalent definition:*

$$\begin{aligned} &SID(\mathcal{G}, \mathcal{H}) \\ &= \#\left\{ (i, j), i \neq j \;\middle|\; \begin{array}{ll} j \in \boldsymbol{DE}_i^{\mathcal{G}} & \text{if } j \in \boldsymbol{PA}_i^{\mathcal{H}} \\ \boldsymbol{PA}_i^{\mathcal{H}} \text{ does not satisfy } (*) \text{ for } (\mathcal{G}, i, j) & \text{if } j \notin \boldsymbol{PA}_i^{\mathcal{H}} \end{array} \right\} \end{aligned}$$

Here, $\mathbf{DE}_i^{\mathcal{G}}$ denotes the descendants of node i in graph $\mathcal{G}$ (see appendix A). The proof is provided in appendix B; it is based on lemma 1.

2.3 Properties. We first investigate metric properties of the SID. Let us denote the number of nodes in a graph by p (this is overloading notation but does not lead to any ambiguity). We then have that

$$0 \leq \mathrm{SID}(\mathcal{G}, \mathcal{H}) \leq p \cdot (p - 1)$$

and

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

The SID therefore satisfies the properties of what is sometimes called a premetric.[3]

The SID is not symmetric; for example, for a nonempty graph $\mathcal{G}$ and an empty graph $\mathcal{H}$, we have that $\mathrm{SID}(\mathcal{G}, \mathcal{H}) \neq 0 = \mathrm{SID}(\mathcal{H}, \mathcal{G})$ (if $\mathcal{G}$ is the empty DAG, all sets of nodes satisfy $(*)$ and are therefore valid adjustment sets).

If $\mathrm{SID}(\mathcal{G}, \mathcal{H}) = 0$, parent adjustment leads to the same intervention distributions in $\mathcal{G}$ and $\mathcal{H}$, but it does not necessarily hold that $\mathcal{G} = \mathcal{H}$. Example 1 shows graphs $\mathcal{G} \neq \mathcal{H}_1$ with $\mathrm{SID}(\mathcal{G}, \mathcal{H}_1) = 0$. We can characterize the set of DAGs that have structural intervention distance zero to a given true DAG $\mathcal{G}$:

Proposition 3. *Consider two DAGs $\mathcal{G}$ and $\mathcal{H}$. We then have*

$$SID(\mathcal{G}, \mathcal{H}) = 0 \quad \Leftrightarrow \quad \mathcal{G} \leq \mathcal{H}.$$

Here, $\mathcal{G} \leq \mathcal{H}$ means that $\mathcal{G}$ is a subgraph of $\mathcal{H}$ (see appendix A). The proof is provided in appendix C; it works for any type of adjustment set, not just the parent set (see section 2.4.5). Proposition 3 states that $\mathcal{H}$ can contain many more (additional) edges than $\mathcal{G}$ and still receive an SID of zero. Intuitively, the SID counts the number of pairs (i, j), such that the intervention distribution inferred from the graph $\mathcal{H}$ is wrong; the latter happens if the estimated set of parents $\mathbf{PA}_{X_i}^{\mathcal{H}}$ is not a valid adjustment set in $\mathcal{G}$. If an estimate $\mathcal{H}$ contains strictly too many edges, that is, $\mathcal{G} \leq \mathcal{H}$ and $\mathbf{pa}_{X_i}^{\mathcal{G}} \subseteq \mathbf{pa}_{X_i}^{\mathcal{H}}$ for all i, the intervention distributions are correct; this follows from $p(x_j \mid x_i, \mathbf{pa}_{X_i}^{\mathcal{H}}) = p(x_j \mid x_i, \mathbf{pa}_{X_i}^{\mathcal{G}})$ (see also lemma 1). For computing intervention distributions in practice, we have to estimate $p(x_j \mid x_i, \mathbf{pa}_{X_i}^{\mathcal{H}})$ based on finitely many samples. This can be seen as a regression task, a well-understood problem in statistics. It is therefore a question of the regression or feature selection technique, whether we see this equality (at least approximately) in practice as well. Section 2.4.3 shows a simple way to combine the SID with another measure in order to obtain zero distance if and only if the two graphs coincide.

The following proposition relates SID to the SHD by providing sharp bounds in some specific cases; these results underline the difference between these two measures. The proof is provided in appendix D.

Proposition 4 *(relating SID and SHD). Consider two DAGs $\mathcal{G}$ and $\mathcal{H}$.*

1a. When the SHD is zero, the SID is zero too:

$$SHD(\mathcal{G}, \mathcal{H}) = 0 \Rightarrow SID(\mathcal{G}, \mathcal{H}) = 0.$$

[3] A function $d : \mathbb{G} \times \mathbb{G} \to \mathbb{R}$ is called a premetric if $d(a, b) \geq 0$ and $d(a, a) = 0$.

1b. We have

$$SHD(\mathcal{G}, \mathcal{H}) = 1 \Rightarrow SID(\mathcal{G}, \mathcal{H}) \leq 2 \cdot (p-1).$$

This bound is sharp.

2. *There exists $\mathcal{G}$ and $\mathcal{H}$ such that $SID(\mathcal{G}, \mathcal{H}) = 0$ but $SHD(\mathcal{G}, \mathcal{H}) = p(p-1)/2$, which achieves the maximal possible value. Therefore, we cannot bound SHD from SID.*

2.4 Extensions

2.4.1 SID Between a DAG and a CPDAG. Let $\mathbb{C}$ denote the space of CPDAGs (completed partially directed acyclic graphs) over p variables. Some causal inference methods like the PC-algorithm (Spirtes et al., 2000) or greedy equivalence search (Chickering, 2002) do not output a single DAG, but rather a completed PDAG $\mathcal{C} \in \mathbb{C}$ representing a Markov equivalence class of DAGs. In order to compute the SID between a (true) DAG $\mathcal{G}$ and an (estimated) PDAG, we can in principle enumerate all DAGs in the Markov equivalence class and compute the SID for each single DAG. This way, we obtain a vector of distances instead of a single number, and we can compute lower and upper bounds for these distances.

Since the enumeration becomes computationally infeasible with growing graph size, we propose to extend the CPDAG locally. Especially for sparse graphs, this provides a considerable computational speed-up. We make use of the fact that the PDAG $\mathcal{C}$ represents a Markov equivalence class of DAGs only if each chain component is chordal (Andersson, Madigan, & Perlman, 1997); see appendix A for definitions. We extend each chordal chain component c locally to all possible DAGs $\mathcal{C}_{c,1}, \ldots, \mathcal{C}_{c,k}$, leaving the other chain components undirected (Meek, 1995). For each extension $\mathcal{C}_{c,h}$ $(1 \leq h \leq k)$ and for each vertex i within the chain component c, we consider

$$I(\mathcal{G}, \mathcal{C}_{c,h})_i := \#\left\{ j \neq i \;\middle|\; \begin{array}{ll} X_j \in \mathbf{DE}_{X_i}^{\mathcal{G}} & \text{if } X_j \in \mathbf{PA}_{X_i}^{\mathcal{C}_{c,h}} \\ \mathbf{PA}_{X_i}^{\mathcal{C}_{c,h}} \text{ does not satisfy } (*) \text{ for graph } \mathcal{G} & \text{if } X_j \notin \mathbf{PA}_{X_i}^{\mathcal{C}_{c,h}} \end{array} \right\}.$$

For each chain component c, we thus obtain k vectors $I(\mathcal{G}, \mathcal{C}_{c,1}), \ldots, I(\mathcal{G}, \mathcal{C}_{c,k})$, each having $\#c$ (= number of vertices in c) entries. We then represent each vector with its sum,

$$S(\mathcal{G}, \mathcal{C}_{c,h}) = \sum_{i \in c} I(\mathcal{G}, \mathcal{C}_{c,h})_i, \quad h = 1, \ldots, k,$$

and save the minimum and the maximum over the k values:

$$\min_h S(\mathcal{G}, \mathcal{C}_{c,h}), \quad \max_h S(\mathcal{G}, \mathcal{C}_{c,h}).$$

These values correspond to the "best" and "worst" DAG extensions. We then report the sum over all minima and the sum over all maxima as lower and upper bound, respectively:

$$\mathrm{SID}_{\mathrm{lower}}(\mathcal{G}, \mathcal{C}) = \sum_c \min_h S(\mathcal{G}, \mathcal{C}_{c,h}), \quad \mathrm{SID}_{\mathrm{upper}}(\mathcal{G}, \mathcal{C}) = \sum_c \max_h S(\mathcal{G}, \mathcal{C}_{c,h}).$$

This leads to the extended definition:

$$\begin{aligned} \mathrm{SID} : \ & \mathbb{G} \times \mathbb{C} \to \mathbb{N} \times \mathbb{N} \\ & (\mathcal{G}, \mathcal{C}) \mapsto \big(\mathrm{SID}_{\mathrm{lower}}(\mathcal{G}, \mathcal{C}), \mathrm{SID}_{\mathrm{upper}}(\mathcal{G}, \mathcal{C})\big). \end{aligned} \tag{2.3}$$

The definition guarantees that the neighborhood orientations of two nodes do not contradict each other. Both the lower and upper bounds are therefore met by a DAG member in the equivalence class of $\mathcal{C}$.

The differences between lower and upper bounds can be quite large. If the true DAG is a (Markov) chain $X_1 \to \ldots \to X_p$ of length p, the corresponding equivalence class contains the correct DAG resulting in a SID of zero (lower bound); it also includes the reversed chain $X_1 \leftarrow \ldots \leftarrow X_p$, resulting in a maximal SID of $p \cdot (p-1)$.

In order to provide a better intuition for these lower and upper bounds, we relate them to "identifiable" and "strictly identifiable" intervention distributions in the Markov equivalence class:

Definition 4. *Consider a CPDAG $\mathcal{C}$ and let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be the DAGs contained in the Markov equivalence class represented by $\mathcal{C}$. We say that the intervention distribution from i to j is*

- *strictly identifiable in $\mathcal{C}$ if $p_{\mathcal{C}_g}(x_j \mid do(X_i = x_i))$ is the same for all DAGs $\mathcal{C}_g \in \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ and for all distributions $p(.)$.*
- *identifiable in $\mathcal{C}$ if $p_{\mathcal{C}_g}(x_j \mid do(X_i = x_i))$ is the same for all $\mathcal{C}_g \in \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ and for all distributions $p(.)$ that are Markov with respect to $\mathcal{C}$.*
- *identifiable in $\mathcal{C}$ with regard to $\mathcal{G}$ if $p_{\mathcal{C}_g}(x_j \mid do(X_i = x_i))$ is the same for all DAGs $\mathcal{C}_g \in \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$ and for all distributions $p(.)$ that are Markov with regard to $\mathcal{G}$.*

Definition 2 further calls a (strictly) identifiable intervention distribution from i to j correctly inferred if $p_{\mathcal{G}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) = p_{\mathcal{C}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i))$ for all $\mathcal{L}(\mathbf{X})$ that are Markov with respect to $\mathcal{G}$. With this notation we have the following remark, which is visualized by Figure 3.

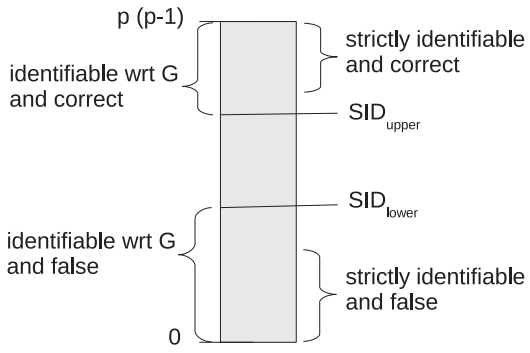

Figure 3: This is a visualization of remark 2. It describes the SID between a DAG $\mathcal{G}$ and a CPDAG $\mathcal{H}$.

Remark 2. Given a true DAG $\mathcal{G}$ and an estimated CPDAG $\mathcal{C}$. It then holds (see Figure 3) that

$$\#\left\{\text{intervention distributions that are } \begin{array}{c}\text{identifiable in } \mathcal{C} \text{ with regard to } \mathcal{G} \text{ and} \\ \text{inferred falsely by } \mathcal{C} \text{ with regard to } \mathcal{G}\end{array}\right\} = \mathrm{SID}_{\mathrm{lower}}(\mathcal{G}, \mathcal{C})$$

$$\#\left\{\text{intervention distributions that are } \begin{array}{c}\text{identifiable in } \mathcal{C} \text{ with regard to } \mathcal{G} \text{ and} \\ \text{inferred correctly by } \mathcal{C} \text{ with regard to } \mathcal{G}\end{array}\right\} = p \cdot (p-1) - \mathrm{SID}_{\mathrm{upper}}(\mathcal{G}, \mathcal{C})$$

$$\#\left\{\text{intervention distributions that are } \begin{array}{c}\text{strictly identifiable in } \mathcal{C} \text{ and} \\ \text{inferred falsely by } \mathcal{C} \text{ with regard to } \mathcal{G}\end{array}\right\} \leq \mathrm{SID}_{\mathrm{lower}}(\mathcal{G}, \mathcal{C})$$

$$\#\left\{\text{intervention distributions that are } \begin{array}{c}\text{strictly identifiable in } \mathcal{C} \text{ and} \\ \text{inferred correctly by } \mathcal{C} \text{ with regard to } \mathcal{G}\end{array}\right\} \leq p \cdot (p-1) - \mathrm{SID}_{\mathrm{upper}}(\mathcal{G}, \mathcal{C}).$$

As an example, consider the CPDAG $\mathcal{C}$ $X_1 - X_2 \rightarrow X_3 \leftarrow X_4$. If $\mathcal{G}$ is such that both $\emptyset$ and $\{X_1\}$ are valid adjustment sets for the effect from 2 to 3, then this intervention distribution is identifiable in $\mathcal{C}$ with regard to $\mathcal{G}$. That is, all DAGs in $\mathcal{C}$ "agree on" the intervention distribution from 2 to 3. If this distribution is inferred incorrectly, it contributes to $\mathrm{SID}_{\mathrm{lower}}(\mathcal{G}, \mathcal{C})$. However, the effect from 2 to 3 is not strictly identifiable in $\mathcal{C}$.

The procedure above fails if $\mathcal{C}$ is not a completed PDAG and therefore does not represent a Markov equivalence class. This may happen for some

versions of the PC algorithm if they are applied to finitely many data points or when hidden variables are present. For each node i, we can then consider all subsets of undirected neighbors as possible parent sets and again report lower and upper bounds. The same is done if the chain component is too large (with more than eight nodes). These modifications are implemented in our R-code that is available on the first author's home page.

2.4.2 SID Between a CPDAG and a DAG or CPDAG. If we simulate from a linear gaussian SEM with different error variances, for example, we cannot hope to recover the correct DAG from the joint distribution. If we assume faithfulness, however, it is possible to identify the correct Markov equivalence class. In such situations, one may want to compare the estimated structure with the correct Markov equivalence class (represented by a CPDAG) rather than with the correct DAG. Again, we denote the space of CPDAGs by $\mathbb{C}$. We have defined the SID on $\mathbb{G} \times \mathbb{G}$ (see definition 3) and on $\mathbb{G} \times \mathbb{C}$ (see section 2.4.1). We now want to extend the definition to $\mathbb{C} \times \mathbb{G}$ and $\mathbb{C} \times \mathbb{C}$, where we compare an estimated structure with a true CPDAG $\mathcal{C}$. The CPDAG $\mathcal{C}$ represents a Markov equivalence class that includes many different DAGs $\mathcal{G}_1, \ldots, \mathcal{G}_k$. These different DAGs lead to different intervention distributions. The main idea is therefore to consider only those (i, j) for which the intervention distribution from i to j is identifiable in $\mathcal{C}$ (see definition 4). Maathuis and Colombo (2013) introduce a generalized backdoor criterion that can be used to characterize the identifiability of intervention distributions. Lemma 2 is a direct implication of their corollary 4.2 and provides a graphical criterion in order to decide whether an intervention distribution is identifiable in a CPDAG. To formulate the result, we define that a path $X_{a_1}, \ldots, X_{a_s}$ in a partially directed graph is possibly directed if no edge between X_{a_f} and $X_{a_{f+1}}$, $f \in \{1, \ldots, s-1\}$, is pointing toward X_{a_f}.

Lemma 2. *Let X_i and X_j be two nodes in a CPDAG $\mathcal{C}$. The intervention distribution from i to j is not identifiable in $\mathcal{C}$ if and only if there is a possibly directed path from X_i to X_j starting with an undirected edge.*

We then define

$$\begin{aligned}
\text{SID}: \ & \mathbb{C} \times \mathbb{G} \to \mathbb{N} \\
& (\mathcal{C}, \mathcal{H}) \mapsto \#\{ (i, j), \\
& i \neq j \mid \\
& \text{the interv. distr. from } i \text{ to } j \text{ is identif. in } \mathcal{C} \text{ and} \\
& \exists \mathcal{C}_1 \in \mathcal{C}\, \exists \mathcal{L}(\mathbf{X}) \text{ such that } \mathcal{L}(\mathbf{X}) \text{ is Markov with regard to } \mathcal{C}_1 \text{ and} \\
& p_{\mathcal{C}_1}(x_j \mid \text{do}(X_i = \hat{x}_i)) \neq p_{\mathcal{H}}(x_j \mid \text{do}(X_i = \hat{x}_i))\}. \qquad (2.4)
\end{aligned}$$

In a DAG, all effects are identifiable. The definitions then reduce to the case of DAGs, equation 2.1. The extension to SID : $\mathbb{C} \times \mathbb{C} \to \mathbb{N} \times \mathbb{N}$ can be

done completely analogous to equation 2.3 with lower and upper bounds of the SID score, equation 2.4, between a true CPDAG and all DAGs in the estimated Markov equivalence class.

2.4.3 Penalizing Additional Edges. The estimated DAG may have strictly more edges than the true DAG and still receives an SID of zero (see Proposition 3). We argued in section 2.3 that for computing causal effects, this fact introduces statistical problems that can be dealt with only if the sample size increases. In some practical situations, however, it may nevertheless be seen as an unwanted side effect. This problem can be addressed by introducing an additional distance measuring the difference in number of edges (DNE) between $\mathcal{G}$ and $\mathcal{H}$:

$$\mathrm{DNE}(\mathcal{G}, \mathcal{H}) = |\#\text{edges in } \mathcal{G} - \#\text{edges in } \mathcal{H}|.$$

Here, a directed or undirected edge counts as one edge. For any DAG $\mathcal{G}$ and any DAG $\mathcal{H}$, it then follows directly from proposition 3 that

$$\mathcal{G} = \mathcal{H} \quad \Leftrightarrow \quad \big(\mathrm{SID}(\mathcal{G}, \mathcal{H}) = 0 \text{ and } \mathrm{DNE}(\mathcal{G}, \mathcal{H}) = 0\big).$$

Analogously, we have for any DAG $\mathcal{G}$ and any CPDAG $\mathcal{C}$,

$$\mathcal{G} \in \mathcal{C} \quad \Leftrightarrow \quad \big(\mathrm{SID}_{\mathrm{lower}}(\mathcal{G}, \mathcal{C}) = 0 \text{ and } \mathrm{DNE}(\mathcal{G}, \mathcal{C}) = 0\big).$$

2.4.4 Symmetrization. We may also want to compare two DAGs $\mathcal{G}$ and $\mathcal{H}$, where neither of them can be seen as an estimate of the other. For these situations, we suggest a symmetrized version of the SID:

$$\mathrm{SID}_{\mathrm{symm}}(\mathcal{G}, \mathcal{H}) = \frac{\mathrm{SID}(\mathcal{G}, \mathcal{H}) + \mathrm{SID}(\mathcal{H}, \mathcal{G})}{2}.$$

There are other possibilities to construct symmetric versions of SID, of course. Instead of the weighted average, one may want to consider the maximum of both values, for example (as suggested by an anonymous reviewer). Modifying definition 2, we could also count all pairs (i, j), such that the intervention distributions coincide for all distributions that are Markov with respect to both graphs. In our opinion, this option is less favorable since it would result in a distance that is always zero if one of its arguments is the empty graph.

2.4.5 Alternative Adjustment Sets. In this work we use the parent set for adjustment. Since it is easy to compute and depends only on the neighborhood of the intervened nodes, it is widely used in practice. Any other

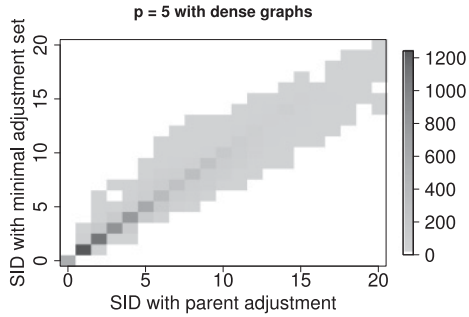

Figure 4: The SID between two DAGs is similar when it is computed with parent adjustment or the minimal adjustment set.

method to compute adjustment sets in graphs can be used too, of course. Choosing an adjustment set of minimal size (see Figure 2) is more difficult to compute but has the advantage of a small conditioning set. Textor and Liskiewicz (2011) discuss recent advances in efficient computation. In contrast to the parent set, it depends on the whole graph. Using the experimental setup from section 3.1, we compare the SID computed with parent adjustment with the SID computed with the minimal adjustment set for randomly generated dense graphs of size $p = 5$. Since the minimal adjustment set need not be unique, we decided to choose the smallest set that is found first by the algorithm. Figure 4 shows that the differences between the two values of SID, once computed with parent sets and once computed with minimal adjustment sets, are rather small (especially compared to the differences between SID and SHD; see section 3.1). In about 70% of the cases, they are exactly the same.

2.4.6 Hidden Variables (Future Work). If some of the variables are unobserved, not all of the intervention distributions are identifiable from the true DAG. We provide a road map on how this case can be included in the framework of the SID. As it was done for CPDAGs (see section 2.4.2), we can exclude the nonidentifiable pairs from the structural intervention distance. In the presence of hidden variables, the true structure can be represented by an acyclic directed mixed graph (ADMG), for which Shpitser & Pearl (2006) address the characterization of identifiable intervention distributions. Alternatively, we can regard a maximal ancestral graph (MAG) (Richardson & Spirtes, 2002) as the ground truth, for which the characterization becomes more difficult. There exists a generalized back-door criterion for MAGs and a way to construct a set satisfying the generalized back-door criterion; some identifiable effects, however, may not be identifiable via the generalized backdoor criterion (Maathuis & Colombo, 2013). Furthermore, methods like FCI (Spirtes et al., 2000) and its successors (Colombo,

Maathuis, Kalisch, & Richardson, 2012; Claassen, Mooij, & Heskes, 2013) output an equivalence class of MAGs that are called partial ancestral graphs (PAGs) (Richardson & Spirtes, 2002). To compare an estimated PAG to the true MAG, we would again go through all MAGs represented by the PAG (see section 2.4.1) and provide lower and upper bounds (as in section 2.4.1). Future work might show that this can be done efficiently.

2.4.7 Multiple Interventions (Future Work). The structural intervention distance compares the two graphs' predictions of intervention distributions. Until now, we have considered only interventions on single nodes. Instead, one may also consider multiple interventions. A direct modification of the adjustment idea does not work, however. The union of the parent sets, for example, is in general not a valid adjustment set; in some situations there might not even be a valid adjustment set (see Nandy, Maathuis, & Richardson, 2014, for discussion and possible alternatives). Furthermore, given a method that computes a valid adjustment set in the correct graph, one needs to handle the computational complexity that arises from the large number of possible interventions: for each number k of variables, there are $\binom{p}{k}$ possible intervention sets and $p-k$ possible target nodes. In total, we thus have $\sum_{k=1}^{p-1}\binom{p}{k}(p-k) = p(2^{p-1}-1)$ intervention distributions. In practice, one may first address the case of intervening on two nodes, where the number of possible intervention distributions is $p(p-1)(p-2)/2$.

3 Simulations

3.1 SID Versus SHD. For $p = 5$ and for $p = 20$, we sample 10,000 pairs of random DAGs and compute both the SID and the SHD between them. We consider two probabilities for i.i.d. sampling of edges, namely, $p_{\text{connect}} = 1.5/(p-1)$ (resulting in an expected number of $0.75p$ edges) for a sparse setting and $p_{\text{connect}} = 0.3$ for a dense setting. Furthermore, the order of the variables is chosen from a uniformly distributed permutation among the vertices. Figures 5a and 5c show two-dimensional histograms with SID and SHD. It is apparent that the SHD and SID constitute very different distance measures. For example, for SHD equal to a low number such as one or two (see $p = 5$ in the dense case), the SID can take on very different values. This indicates that, compared to the SHD, the SID provides additional information that is appropriate for causal inference. The observations are in par with the bounds provided in proposition 3.

For each pair $\mathcal{G}$ and $\mathcal{H}$ of graphs, we also generate a distribution by defining a linear structural equation model,

$$X_j = \sum_{k \in \mathbf{PA}_j^{\mathcal{G}}} \beta_{jk} X_k + N_j, \qquad j = 1, \ldots, p,$$

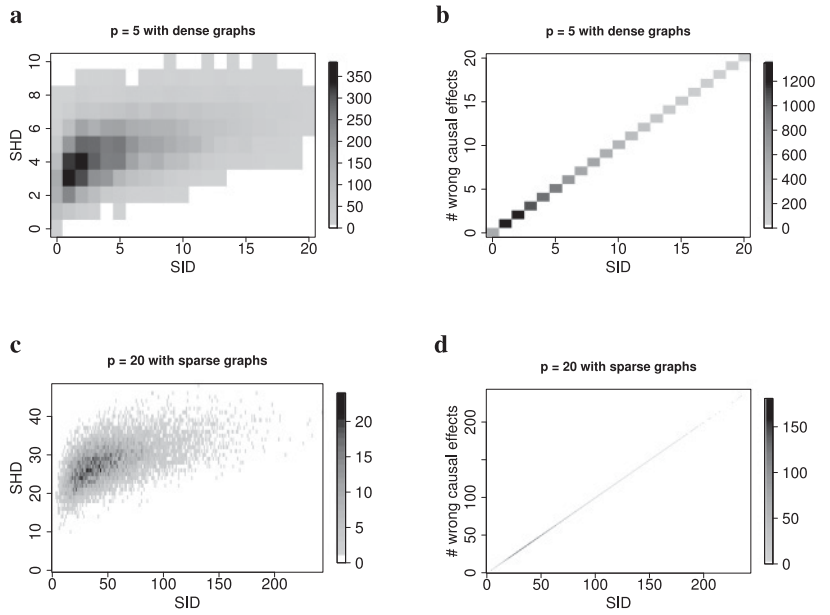

Figure 5: We generate 10, 000 pairs of random small dense graphs (a) and (b) and larger sparse graphs (c) and (d). For each pair of graphs $(\mathcal{G}, \mathcal{H})$ we also generate a distribution that is Markov with regard to $\mathcal{G}$. The two-dimensional histograms compare SID$(\mathcal{G}, \mathcal{H})$ with SHD$(\mathcal{G}, \mathcal{H})$ (a and c) and SID$(\mathcal{G}, \mathcal{H})$ with the number of pairs (i, j), for which the calculated causal effects differ (b and d). The SID measures exactly the number of wrongly estimated causal effects and thus provides additional and very different information as the SHD.

whose graph is identical to $\mathcal{G}$. We sample the coefficients β_{jk} uniformly from $[-1.0; -0.1] \cup [0.1; 1.0]$. The noise variables are normally distributed with mean zero and variance one. Due to the assumption of equal error variances for the error terms, the DAG is identifiable from the distribution (Peters & Bühlmann, 2014). With the linear gaussian choice, we can characterize the true intervention distribution $p_{\mathcal{G}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i))$ by one number, namely, the derivative of the expectation with respect to $\hat{x}_i$ (which is also called the total causal effect of X_i on X_j). Its derivation can be found in appendix E. We can then compare the intervention distributions from $\mathcal{G}$ and $\mathcal{H}$ and report the number of pairs (i, j), for which these two numbers differ. For numerical reasons, we regard two numbers as different if their absolute difference is larger than 10^{-8}. Figures 5b and 5d show the comparison to the SID. In all 20,000 cases, the SID counts exactly the number of those "wrong" causal effects. A priori this is not obvious since definition 3 requires only that there exists a distribution that discriminates between the intervention distributions. The result shown in Figure 5 suggests that the intervention

distributions differ for most distributions. Two possible reasons for inequality have indeed small probability: a nondetectable difference that is smaller than 10^{-8} and vanishing coefficients that would violate faithfulness (Spirtes et al., 2000). We are not aware of a characterization of the distributions that do not allow discriminating between the intervention distributions.

3.2 Comparing Causal Inference Methods. As in section 3.1 we simulate sparse random DAGs as ground truth (100 times for each value of p and n). We again sample n data points from the corresponding linear gaussian structural equation model with equal error variances (as above, coefficients are uniformly chosen from $[-1; -0.1] \cup [0.1; 1]$) and apply different inference methods. This setting allows us to use the PC algorithm (Spirtes et al., 2000), conservative PC (Ramsey, Zhang, & Spirtes, 2006), greedy equivalent search (GES) (Chickering, 2002) and greedy DAG search based on the assumption of equal error variances (GDS_{EEV}) (Peters & Bühlmann, 2014). For both PC versions and the GES we use the `R` packages `pcalg` (Kalisch, Mächler, Colombo, Maathuis, & Bühlmann, 2012) with the default choice of parameter values. Table 1 reports the average SID between the true DAG and the estimated ones. GDS_{EEV} is the only method that outputs a DAG. All other methods output a Markov equivalence class for which we apply the extension suggested in section 2.4.1. Additionally, we report the results for a random estimator RAND that does not take into account any of the data: we sample a DAG as in section 3.1 but with p_{connect} uniformly chosen between 0 and 1. Section 2.4.1 provides an example, for which the SID can be very different for two DAGs within the same Markov equivalence class. Table 1 shows that this difference can be quite significant even on average. While the lower bound often corresponds to a reasonably good estimate, the upper bound may not be better than random guessing for small sample sizes. In fact, for $p = 5$ and $n = 100$, the distance to the RAND estimate was less than the upper bound for PC in 77 of the 100 experiments (not directly readable from the aggregated numbers in the table). For the SHD, however, the PC algorithm outperforms random guessing; for example, for $p = 5$ and $n = 100$, RAND is better than PC in 8 out of 100 experiments. This supports the idea that the PC algorithm estimates the skeleton of a DAG more reliably than the directions of its edges. The results also show how much can be gained when additional assumptions are appropriate; all methods exploit that the data come from a linear gaussian SEM, while only GDS_{EEV} makes use of the additional constraint of equal error variances, which leads to identifiability of the DAG from the distribution (Peters & Bühlmann, 2014). We draw different conclusions if we consider the SHD (see Table 2). For $p = 40$ and $n = 100$, for example, PC performs best with respect to SHD while it is worst with respect to SID.

We use the SID for comparing causal inference methods. It is not straightforward to tune a method such that it performs well under the SID. Proposition 2 appears to suggest that removing edges can only increase the SID;

Table 1: Average SID to True DAG for 100 Simulation Experiments with Standard Deviation, for Different n and p.

	$n = 100$				
p	GDS_{EEV}	CPC	PC	GES	RAND
5	$\mathbf{1.7 \pm 2.2}$	2.9 ± 3.2	4.3 ± 4.7	3.3 ± 4.2	6.1 ± 4.0
		8.8 ± 5.2	7.7 ± 5.2	6.9 ± 4.6	
20	$\mathbf{14.1 \pm 10.5}$	22.8 ± 17.1	37.0 ± 26.8	24.4 ± 17.4	47.7 ± 28.8
		63.3 ± 38.0	52.8 ± 30.1	33.1 ± 19.1	
40	$\mathbf{37.2 \pm 27.2}$	56.7 ± 36.3	91.3 ± 58.3	58.9 ± 34.6	119.1 ± 63.8
		147.5 ± 78.6	124.2 ± 66.4	65.9 ± 36.2	
	$n = 1000$				
p	GDS_{EEV}	CPC	PC	GES	RAND
5	$\mathbf{0.6 \pm 1.6}$	1.7 ± 3.4	3.0 ± 4.7	1.9 ± 3.7	6.3 ± 5.0
		7.0 ± 4.8	6.7 ± 4.8	6.3 ± 4.4	
20	$\mathbf{3.0 \pm 6.7}$	7.4 ± 10.3	26.4 ± 28.7	8.3 ± 10.2	53.1 ± 36.6
		40.0 ± 28.4	40.2 ± 27.5	23.4 ± 13.1	
40	$\mathbf{7.8 \pm 10.2}$	13.8 ± 12.6	62.1 ± 45.5	19.7 ± 18.7	132.2 ± 79.8
		89.8 ± 49.5	91.9 ± 49.3	43.9 ± 22.7	

Notes: For the methods that output a Markov equivalence class (CPC, PC, and GES), two rows are shown: they represent DAGs from the equivalence class with the smallest and with the largest distance, that is, the lower and upper bounds in equation 4. The smallest averages are in bold.

Table 2: Same Experiment as in Table 1, This Time Reporting the Average SHD to the True DAG.

	$n = 100$				
p	GDS_{EEV}	CPC	PC	GES	RAND
5	$\mathbf{1.0 \pm 1.1}$	3.1 ± 1.4	2.6 ± 1.4	2.7 ± 1.5	6.2 ± 2.2
20	$\mathbf{11.3 \pm 3.1}$	13.4 ± 3.7	$\mathbf{11.3 \pm 3.1}$	15.0 ± 3.3	96.7 ± 47.8
40	43.7 ± 6.6	27.2 ± 4.9	$\mathbf{22.6 \pm 4.6}$	45.4 ± 6.1	377.9 ± 195.8
	$n = 1000$				
p	GDS_{EEV}	CPC	PC	GES	RAND
5	$\mathbf{0.3 \pm 0.6}$	2.6 ± 1.5	2.3 ± 1.4	2.5 ± 1.5	6.0 ± 2.0
20	$\mathbf{2.8 \pm 1.9}$	8.6 ± 2.7	7.7 ± 2.6	7.8 ± 2.7	98.4 ± 50.7
40	$\mathbf{10.6 \pm 3.6}$	17.0 ± 3.5	15.3 ± 3.4	17.8 ± 4.0	393.5 ± 189.8

Note: Smallest averages are in bold.

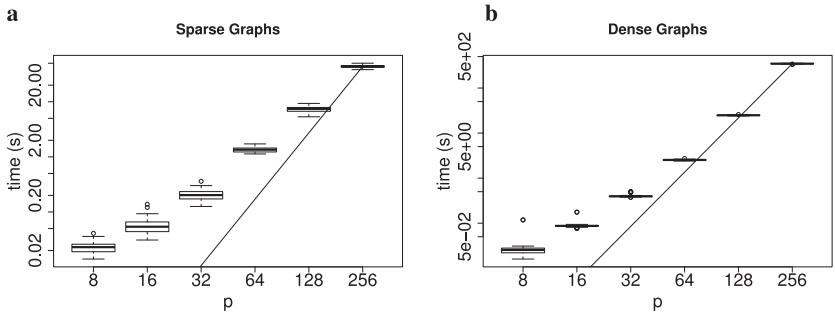

Figure 6: For varying p, the figure shows box plots for the processor time needed to compute the SID for one pair of random graphs (averaged over 100 pairs). (a) Results for sparse graphs. (b) Results for dense graphs. The line indicates the computational complexity of $\mathcal{O}(p^4)$.

thus, dense graph estimates would yield smaller SIDs than sparse estimates. This assumes, however, that the causal ordering is correct. In general, the benefit of estimating large graphs is either small or nonexistent. In the experimental setting of table 1, for example, random sparse DAGs ($p_{\text{connect}} = 1.5/(p-1)$) yield an average SID of 7.9 ± 4.6 and 134.3 ± 65.4 for $p = 5$ and $p = 40$, respectively (100 repetitions). Random dense DAGs ($p_{\text{connect}} = 1$) yield 5.9 ± 4.9 and 110.3 ± 61.9, respectively. For nonlinear additive noise models, Bühlmann, Peters, and Ernest (2014) estimate the causal ordering first and suggest pruning the graph later. Again, there is no significant drop of SID after the pruning (Bühlmann et al., 2014). Combining the SID with a measure penalizing additional edges as in section 2.4.3 removes any (possible) advantage of estimating dense graphs.

3.3 Scalability of the SID. For different values of p, we report here the processor time needed for computing the SID between two random graphs with p nodes. We choose the same setting for sparse and dense graphs as in section 3.1. Figure 6 shows box plots for 100 pairs of graphs for each value of p ranging between 2^3 and 2^8.[4] The figure does not provide a clean picture about the scaling of time complexity. It has not reached the worst-case analysis of $\mathcal{O}(p^4 \cdot \log_2(p))$ (which scales even faster than the shown p^4). Computing the SID between sparse graphs of size 1000, our code requires almost 7000 seconds (a bit less than 2 hours). The algorithm is written in `R`, and we expect significant speed-ups in other programming languages.

[4]The experiments were performed on a 64 bit Ubuntu machine using a single core of the Intel Core2 Duo CPU P8600 at 2.40 GHz.

4 Implementation

We sketch here the implementation of the structural intervention distance while details are presented in algorithms 1 and 2 in appendix F using pseudocode. The key idea of our algorithm is based on proposition 2. Condition $*$ contains two parts that need to be checked. Part 1 addressed the issue of whether any node from the conditioning set is a descendant of any node on a directed path (see line 28 in algorithm 1 in appendix F). Here, we make use of the $p \times p$ PathMatrix: its entry (i, j) is one if and only if there is a directed path from i to j. This can be computed efficiently by squaring the matrix $(\mathrm{Id} + \mathcal{G})$ $\lceil \log_2(p) \rceil$ times since $\mathcal{G}$ is idempotent; here we denote by $\mathcal{G}$ the adjacency matrix of the DAG $\mathcal{G}$. For part 2 of $(*)$, we check whether the conditioning set blocks all nondirected paths from i to j (see line 31 in algorithm 1). It is the purpose of the function rondp (line 9 in algorithm 1) to compute all nodes that can be reached on a nondirected path.

Algorithm 2, also presented in appendix F, describes the function rondp that computes all nodes reachable on nondirected paths. In a breadth-first search, we go through all node-orientation combinations and compute the $2p \times 2p$ reachabilityMatrix. Afterward we compute the corresponding PathMatrix (line 24 in algorithm 2). We then start with a vector reachableNodes (consisting of parents and children of node i) and read off all reachable nodes from the reachabilityPathMatrix. We then filter out the nodes that are reachable on a nondirected path.

Note that in the procedure, computing the PathMatrix is computationally the most expensive part. Making sure that this computation is done only once for all j is one reason that we do not use any existing implementation (e.g., for d-separation). The worst-case computational complexity for computing the SID between dense matrices is $\mathcal{O}(p \cdot \log_2(p) \cdot f(p))$, where squaring a matrix requires $\mathcal{O}(f(p))$; a naive implementation yields $f(p) = p^3$ while Coppersmith & Winograd (1987) report $f(p) = \mathcal{O}(p^{2.375477})$, for example. Sparse matrices lead to improved computational complexities, of course (see also section 3.3).

We also implemented the steps required for computing the SID between a DAG and a completed PDAG (both options from section 2.4.1) using a function that enumerates all DAGs from a partially directed graph. Those steps, however, are not shown in the pseudocode in order to ensure readability.

Our software code for SID is provided as `R`-code on the first author's home page. It will also be implemented in the `pcalg` package (Kalisch et al., 2012).

5 Conclusion

We have proposed a new (pre-)metric, the structural intervention distance (SID), between directed acyclic graphs and completed partially directed

acyclic graphs. Since the SID is a one-dimensional measure of distances between high-dimensional objects, it does not capture all aspects of the difference. The SID measures closeness between graphs in terms of their capacities for causal effects (intervention distributions). It is therefore well suited for evaluating different estimates of causal graphs. The distance can provide a useful complement to existing measures; for example, it differs significantly from the widely used structural Hamming distance (SHD). Based on known results for graphical characterization of adjustment sets, we have provided a representation of the SID that enabled us to develop an algorithm that scales up to a few thousand nodes.

In our simulation setting for causal inference, many methods outperform random guessing on small sample sizes in terms of SHD but not in terms of SID (e.g., some DAGs in the estimated Markov equivalence class are worse than randomly chosen DAGs). This means that if one uses simulation studies in order to determine how many samples are required to draw reliable causal conclusions from an estimated DAG (i.e., to obtain a small SID), the SHD will draw a picture that is too optimistic.

Appendix A: Terminology for Directed Acyclic Graphs

We summarize here some well-known facts about graphs, essentially taken from Peters (2012). Let $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ be a graph with $\mathbf{V} := \{1, \ldots, p\}$, $\mathcal{E} \subset \mathbf{V} \times \mathbf{V}$ and corresponding random variables $\mathbf{X} = (X_1, \ldots, X_p)$. A graph $\mathcal{G}_1 = (\mathbf{V}_1, \mathcal{E}_1)$ is called a *subgraph* of $\mathcal{G}$ if $\mathbf{V}_1 = \mathbf{V}$ and $\mathcal{E}_1 \subseteq \mathcal{E}$; we then write $\mathcal{G}_1 \leq \mathcal{G}$. If in addition, $\mathcal{E}_1 \neq \mathcal{E}$, we call $\mathcal{G}_1$ a *proper subgraph* of $\mathcal{G}$. A node i is called a *parent* of j if $(i, j) \in \mathcal{E}$ and a *child* if $(j, i) \in \mathcal{E}$. The set of parents of j is denoted by $\mathbf{PA}_j^{\mathcal{G}}$, and the set of its children by $\mathbf{CH}_j^{\mathcal{G}}$. Two nodes i and j are *adjacent* if either $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$. We call $\mathcal{G}$ *fully connected* if all pairs of nodes are adjacent. We say that there is an *undirected* edge between two adjacent nodes i and j if $(i, j) \in \mathcal{E}$ and $(j, i) \in \mathcal{E}$; we denote this edge by $i - j$. An edge between two adjacent nodes is directed if it is not undirected; if $(i, j) \in \mathcal{E}$, we denote it by $i \to j$. In the graphs we consider, there are four different *types of edges* between i and j: $i - j$, $i \to j$, $j \to i$, or i and j are not adjacent. The *skeleton* of $\mathcal{G}$ is the set of all edges without taking the direction into account; that is all (i, j), such that $(i, j) \in \mathcal{E}$ or $(j, i) \in \mathcal{E}$. The *number of edges* in a graph is the size of the skeleton; that is, undirected edges count as one.

A *path* $\langle i_1, \ldots, i_n \rangle$ in $\mathcal{G}$ is a sequence of (at least two) distinct vertices $i_1, \ldots, i_n$, such that there is an edge between i_k and i_{k+1} for all $k = 1, \ldots, n - 1$. If $(i_k, i_{k+1}) \in \mathcal{E}$ and $(i_{k+1}, i_k) \notin \mathcal{E}$ for all k, we speak of a *directed path* between i_1 and i_n and call i_n a *descendant* of i_1. We denote all descendants of i by $\mathbf{DE}_i^{\mathcal{G}}$ and all nondescendants of i by $\mathbf{ND}_i^{\mathcal{G}}$. We call a node j such that i is a descendant of j an *ancestor* of i and denote the set by $\mathbf{AN}_i^{\mathcal{G}}$. A path $\langle i_1, \ldots, i_n \rangle$ is called a *semidirected cycle* if $(i_j, i_{j+1}) \in \mathcal{E}$ for $j = 1, \ldots, n$ with $i_{n+1} = i_1$ and at least one of the edges is oriented as $i_j \to i_{j+1}$. If $(i_{k-1}, i_k) \in \mathcal{E}$

and $(i_{k+1}, i_k) \in \mathcal{E}$, as well as $(i_k, i_{k-1}) \notin \mathcal{E}$ and $(i_k, i_{k+1}) \notin \mathcal{E}$, i_k is called a *collider* on this path. $\mathcal{G}$ is called a *partially directed acyclic graph* (*PDAG*) if there is no directed cycle, that is, no pair (j, k), such that there are directed paths from j to k and from k to j. $\mathcal{G}$ is called a *chain graph* if there is no semidirected cycle between any pair of nodes. Two nodes j and k in a chain graph are called equivalent if there exists a path between j and k consisting only of undirected edges. A corresponding equivalence class of nodes (i.e., a (maximal) set of nodes that is connected by undirected edges) is called a *chain component*. We call an undirected graph (or a chain component) *chordal* if each of its cycles of four or more nodes has a chord; a chord is an edge that is not a part of the cycle and connects two nodes of the cycle. $\mathcal{G}$ is called a *directed acyclic graph* (*DAG*) if it is a PDAG and all edges are directed. A path in a DAG between i_1 and i_n is *blocked by a set* $\mathbf{S}$ (with neither i_1 nor i_n in this set) whenever there is a node i_k, such that one of the following two possibilities hold: (1) $i_k \in \mathbf{S}$ and $i_{k-1} \to i_k \to i_{k+1}$ or $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or $i_{k-1} \leftarrow i_k \to i_{k+1}$ or (2) $i_{k-1} \to i_k \leftarrow i_{k+1}$ and neither i_k nor any of its descendants is in $\mathbf{S}$. We say that two disjoint subsets of vertices $\mathbf{A}$ and $\mathbf{B}$ are *d-separated* by a third (also disjoint) subset $\mathbf{S}$ if every path between nodes in $\mathbf{A}$ and $\mathbf{B}$ is blocked by $\mathbf{S}$. The joint distribution $\mathcal{L}(\mathbf{X})$ is said to be *Markov with respect to the DAG* $\mathcal{G}$ if

$$\mathbf{A}, \mathbf{B}\ d\text{-sep. by } \mathbf{C} \ \Rightarrow\ \mathbf{X_A} \perp\!\!\!\perp \mathbf{X_B} \mid \mathbf{X_C}$$

for all disjoint sets $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. $\mathcal{L}(\mathbf{X})$ is said to be *faithful to the DAG* $\mathcal{G}$ if

$$\mathbf{A}, \mathbf{B}\ d\text{-sep. by } \mathbf{C} \ \Leftarrow\ \mathbf{X_A} \perp\!\!\!\perp \mathbf{X_B} \mid \mathbf{X_C}$$

for all disjoint sets $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$. Throughout this work, $\perp\!\!\!\perp$ denotes (conditional) independence.

We denote by $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markov with respect to $\mathcal{G}$:

$$\mathcal{M}(\mathcal{G}) := \{\mathcal{L}(\mathbf{X}) \,:\, \mathcal{L}(\mathbf{X}) \text{ is Markov with regard to } \mathcal{G}\}.$$

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent* if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This is the case if and only if $\mathcal{G}_1$ and $\mathcal{G}_2$ satisfy the same set of d-separations, which means the Markov-condition entails the same set of (conditional) independence conditions. A set of Markov equivalent DAGs (so-called Markov equivalence class) can be represented by a completed PDAG, which can be characterized in terms of a chain graph with undirected and directed edges (Andersson et al., 1997): this graph has a directed edge if all members of the Markov equivalence class have such a directed edge; it has an undirected edge if some members of the Markov equivalence class have an edge in the same direction and some members have an edge in the other direction, and it has no edge if all members in the Markov equivalence class have no corresponding edge.

Appendix B: Proof of Proposition 2

Let us denote by A the set of pairs (i, j) appearing in definition 3 and by B the corresponding set of pairs in proposition 2. We will show that $A = B$.

$A \subseteq B$: Consider $(i, j) \in A$. For case 1, if $X_j \in \mathbf{PA}_{X_i}^{\mathcal{H}}$, then $p_{\mathcal{H}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) = p(x_j)$. We will now prove that X_i not being an ancestor of X_j in $\mathcal{G}$ implies $p_{\mathcal{G}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) = p(x_j)$ (the latter statement contradicts $(i, j) \in A$ and therefore, X_i must be an ancestor of X_j in $\mathcal{G}$);

$$
\begin{aligned}
& p_{\mathcal{G}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) \\
&= \int_{\mathrm{anc}(j)} \int_{\text{non-anc}(j)} \prod_{k \in \substack{\mathrm{anc}(j) \cup \{j\} \cup \\ \text{non-anc}(j) \setminus \{i\}}} p(x_k \mid x_{\mathrm{pa}(k)}) \delta(x_i - \hat{x}_i)\, d\mathbf{x}_{\text{non-anc}(j)}\, d\mathbf{x}_{\mathrm{anc}(j)} \\
&\overset{(\dagger)}{=} \int_{\mathrm{anc}(j)} \prod_{k \in \mathrm{anc}(j) \cup \{j\}} p(x_k \mid x_{\mathrm{pa}(k)})\, d\mathbf{x}_{\mathrm{anc}(j)} \\
&= \int_{\mathrm{anc}(j)} \int_{\text{non-anc}(j)} p(x_1, \ldots, x_p)\, d\mathbf{x}_{\text{non-anc}(j)}\, d\mathbf{x}_{\mathrm{anc}(j)} = p(x_j)
\end{aligned}
$$

Equation (†) holds because we can integrate out all nonancestors of j (j is neither an ancestor nor a nonancestor of j): any sink node among nonancestors appears only as a left argument of $p(x_k \mid x_{\mathrm{pa}(k)})$ and can be integrated out. We then disregard that sink node from the graph and integrate out a new sink node among the nonancestors. Since a nonancestor of j cannot be a parent of an ancestor of j, we can integrate out *all* nonancestors and are left with the right-hand side of (†).

In case 2, if $X_j \notin \mathbf{PA}_{X_i}^{\mathcal{H}}$, then it follows by lemma 1i that $\mathbf{PA}_{X_i}^{\mathcal{H}}$ does not satisfy $(*)$. In both cases we have $(i, j) \in B$.

$A \supseteq B$. Now consider $(i, j) \in B$. In case 1, if $X_j \in \mathbf{PA}_{X_i}^{\mathcal{H}}$, then, again, $p_{\mathcal{H}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) = p(x_j)$ and $X_j \in \mathbf{DE}_{X_i}^{\mathcal{G}}$. Consider a linear gaussian structural equation model with error variances being one and equations $X_k = \sum_{\ell \in \mathbf{pa}_k^{\mathcal{G}}} 1 \cdot X_\ell + N_k$, corresponding to the graph structure $\mathcal{G}$. It then follows that $p_{\mathcal{G}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) \neq p(x_j)$. In case 2, if $X_j \notin \mathbf{PA}_{X_i}^{\mathcal{H}}$, then $\mathbf{PA}_{X_i}^{\mathcal{H}}$ does not satisfy $(*)$ and lemma 1ii implies $p_{\mathcal{G}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i)) \neq p_{\mathcal{H}}(x_j \mid \mathrm{do}(X_i = \hat{x}_i))$. In both cases we have $(i, j) \in A$.

Appendix C: Proof of Proposition 3

Assume that $\mathcal{G} \preceq \mathcal{H}$. We will use proposition 2 to show that the SID is zero. If $j \in \mathbf{DE}_i^{\mathcal{G}}$, then $j \in \mathbf{DE}_i^{\mathcal{H}}$, which implies that $j \notin \mathbf{PA}_i^{\mathcal{H}}$. It therefore remains to show that any set $\mathbf{Z}$ that satisfies $(*)$ in lemma 1 for $(\mathcal{H}, i, j)$ satisfies $(*)$ for $(\mathcal{G}, i, j)$ too. The first part of the condition is satisfied since any node

that lies on a directed path in $\mathcal{G}$ lies on a directed path in $\mathcal{H}$. The second part holds because any nondirected path in $\mathcal{G}$ is also a path in $\mathcal{H}$ and must therefore be blocked by $\mathbf{Z}$. If a path is blocked in a DAG, it is always blocked in the smaller DAG too.

Suppose now that $\mathcal{G}$ contains an edge $i \to j$ and that $i \notin \mathbf{PA}_j^{\mathcal{H}}$. We now construct an observational distribution $p(.)$ according to $X_k = N_k$ for all $k \neq j$, $X_j = X_i + N_j$ and $N_k \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$ for all k. This distribution is certainly Markov with respect to $\mathcal{G}$. We find for any $\hat{x}_j$ that $p_{\mathcal{G}}(x_i \mid \mathrm{do}(X_j = \hat{x}_j)) = p(x_i)$ and at the same time $p_{\mathcal{H}}(x_i \mid \mathrm{do}(X_j = \hat{x}_j)) = p(x_i \mid \hat{x}_j) \neq p(x_i)$. Therefore, the SID is different from zero.

Appendix D: Proof of Proposition 4

The different statements can be proved as follows:

1a. When the SHD is zero, each node has the same set of parents in $\mathcal{G}$ and $\mathcal{H}$. Therefore, all adjustment sets are valid and the SID is zero too.

1b. The bound clearly holds since an SHD of one can change the set of parents of at most two nodes. Extending the example shown in Figure 1 from example 3 to $p - 2$, different Y nodes proves that the bound is sharp.

2. Choosing $\mathcal{G}$ the empty graph and $\mathcal{H}$ (any) fully connected graph yields the result.

Appendix E: Computing Causal Effects for Linear Gaussian Structural Equation Models

Consider a linear gaussian structural equation model with known parameters. The covariance matrix $\Sigma_{\mathbf{X}}$ of the p random variables can then be computed from the structural coefficients and the noise variances. For a given graph, we are also able to compute the causal effects analytically. Since the intervention distribution $\mathcal{L}(X_j \mid \mathrm{do}(X_i = \hat{x}_i))$ is again gaussian with mean depending linearly on $\hat{x}_i$ and variance not depending on $\hat{x}_i$, we can summarize it by the so-called causal effect:

$$C_{ij} := \frac{\partial}{\partial \hat{x}_i} \mathbf{E}[X_j \mid \mathrm{do}(X_i = \hat{x}_i)].$$

Let us denote by Σ_2 the submatrix of $\Sigma_{\mathbf{X}}$ with rows and columns corresponding to X_i, $\mathbf{PA}_{X_i}$, and by Σ_1 the $(1 \times (\#\mathbf{PA}_{X_i} + 1))$-vector corresponding to the row from X_j and columns from X_i, $\mathbf{PA}_{X_i}$ of $\Sigma_{\mathbf{X}}$. Then,

$$C_{ij} = \Sigma_1 \cdot \Sigma_2^{-1} \cdot (1, 0 \ldots, 0)^T.$$

Appendix F: Algorithms

Here, we present pseudocode for computing the SID.

Algorithm 1: Computing Structural Intervention Distance.

input two adjacency matrices $\mathcal{G}$ and $\mathcal{H}$ of size $p \times p$.
incorrectCausalEffects $\leftarrow$ ZeroMatrix(p, p)
PathMatrix $\leftarrow$ computePathMatrix$(\mathcal{G})$
for $i = 1$ **to** p **do**
  $paG \leftarrow$ which$(\mathcal{G}[, i] == 1)$ # parents of i in $\mathcal{G}$
  $paH \leftarrow$ which$(\mathcal{H}[, i] == 1)$ # parents of i in $\mathcal{H}$
  $\tilde{\mathcal{G}} \leftarrow \mathcal{G}$ without edges leaving paH with a tail ($paH \rightarrow$)
  PathMatrix2 $\leftarrow$ computePathMatrix$(\tilde{\mathcal{G}})$
  $reachableOnNonDirectedPath \leftarrow$ rondp($\mathcal{G}$,i,paH,PathMatrix,PathMatrix2)
  for $j \neq i$ **from** 1 **to** p **do**
    $ijGNull, ijHNull, finished \leftarrow$ **false**
    if PathMatrix$[i, j] == 0$ **then**
      $ijGNull \leftarrow$ **true** # $\mathcal{G}$ predicts the causal effect to be zero
    end if
    if j is parent from i in $\mathcal{H}$ **then**
      $ijHNull \leftarrow$ **true** # $\mathcal{H}$ predicts the causal effect to be zero
    end if
    if !$ijGNull$ **and** $ijHNull$ **then**
      incorrectCausalEffects$[i, j] \leftarrow 1$
      $finished \leftarrow$ **true** # one mistake if only $\mathcal{H}$ predicts zero
    end if
    if $ijGNull$ **and** $ijHNull$ **or** $paG == paH$ **then**
      $finished \leftarrow$ **true** # no mistakes if both predictions coincide
    end if
    if !finished **then**
      $childrenOnDirectedPath \leftarrow$ children of i in $\mathcal{G}$ that have j as a descendant
      if sum(PathMatrix$[childrenOnDirectedPath, paH]) > 0$ **then**
        incorrectCausalEffects$[i, j] \leftarrow 1$ # part (1)
      end if
      if $reachableOnNonDirectedPath[j] == 1$ **then**
        incorrectCausalEffects$[i, j] \leftarrow 1$ # part (2)
      end if
    end if
  end for
end for
output sum(incorrectCausalEffects)

Algorithm 2: Finding All Reachable Nodes on Nondirected Paths (rondp).

input adjacency matrix $\mathcal{G}$ of size $p \times p$, node i, PaH, PathMatrix, PathMatrix2.

$Pai \leftarrow$ which($\mathcal{G}[, i] == 1$) # parents of i in $\mathcal{G}$
$Chi \leftarrow$ which($\mathcal{G}[i,] == 1$) # children of i in $\mathcal{G}$
$toCheck \leftarrow Pai + p$ and Chi # an index $> p$ indicates that this node is reached with an outgoing edge, $\leq p$ that it is reached with an incoming edge
$reachableNodes \leftarrow Pai$ and Chi
$reachableOnNonDirectedPath \leftarrow Pai + p \cdot 1_{\text{length}(Pai)}$
$\mathcal{G}[i, Chi] \leftarrow 0$, $\mathcal{G}[Pai, i] \leftarrow 0$

for all $currentNode$ **in** $toCheck$ **do**
  $PacN \leftarrow$ which($\mathcal{G}[, currentNode] == 1$)
  # If one of the Pa of $currentNode$ (cN) is reachable and is not included in PaH,
  # then cN is reachable, too (i.e. $\exists$ path from i that is not blocked by PaH).
  $PacN2 \leftarrow PacN$ setMinus PaH
  reachabilityMatrix$[PacN2, currentNode] \leftarrow 1$ # same index rule as before
  reachabilityMatrix$[PacN2 + p, currentNode] \leftarrow 1$
  # If $currentNode$ (cN) is reachable with $\rightarrow$ cN and cN is
  # an ancestor of PaH, then parents are reachable, too.
  if $currentNode$ is an ancestor of PaH **then**
    reachabilityMatrix$[currentNode, PacN + p] \leftarrow 1$
    add $PacN$ to $toCheck$ # $toCheck$ is a set; it contains each index only once
  end if
  # If $currentNode$ (cN) is reachable with $\leftarrow$ cN and cN is
  # not in PaH, then parents are reachable, too.
  if $currentNode$ is not in PaH **then**
    reachabilityMatrix$[currentNode + p, PacN + p] \leftarrow 1$
    add $PacN$ to $toCheck$ # $toCheck$ is a set; it contains each index only once
  end if
  ... # Apply analogous rules to the children ChcN of currentNode.
end for

reachabilityPathMatrix $\leftarrow$ computePathMatrix(reachabilityMatrix)
update $reachableNodes$ using reachabilityPathMatrix
update $reachableOnNonDirectedPath$ using reachabilityPathMatrix
We may have missed some nodes: if there is a directed (non-blocked) path
from i to k, then all parents of k are reachable from i on a non-directed path.
add more nodes to $reachableOnNonDirectedPath$: use PathMatrix2 to look for nodes j as in $i \rightarrow \ldots \rightarrow k \leftarrow j$ (k being a descendant of i with no node from PaH in between)
remove all direct connections between k and j in reachabilityPathMatrix
update $reachableOnNonDirectedPath$ using reachabilityPathMatrix
$reachableOnNonDir.Path \leftarrow \{j \mid j \text{ or } j + p \in reachableOnNonDir.Path\}$

output $reachableOnNonDirectedPath$

Acknowledgments

We thank Alain Hauser, Preetam Nandy, and Marloes Maathuis for helpful discussions. We especially thank the reviewers for their constructive and thoughtful comments. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no 326496.

References

Acid, S., & 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*, 445–490.

Andersson, S., Madigan, D., & Perlman, M. (1997). A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics*, *25*, 505–541.

Bühlmann, P., Peters, J., & Ernest, J. (2014). CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, *42*, 2526–2556.

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

Claassen, T., Mooij, J. M., & Heskes, T. (2013). Learning sparse causal models is not NP-hard. In *Proceedings of the 29th Annual Conference on Uncertainty in Artificial Intelligence*. Cambridge, MA: AUAI Press.

Colombo, D., Maathuis, M., Kalisch, M., & Richardson, T. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, *40*, 294–321.

Coppersmith, D., & Winograd, S. (1987). Matrix multiplication via arithmetic progressions. In *Proceedings of the 19th Annual ACM Symposium on Theory of Computing*. New York: ACM.

de Jongh, M., & Druzdzel, M. J. (2009). A comparison of structural distance measures for causal Bayesian network models. In M. Klopotek, A. Przepiorkowski, S. T. Wierzchon, & K. Trojanowski (Eds.), *Recent advances in intelligent information systems* (pp. 443–456). N.P.: Academic Publishing House.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*, 1–26.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. Cambridge, MA: MIT Press.

Lauritzen, S. (1996). *Graphical models*. New York: Oxford University Press.

Maathuis, M., & Colombo, D. (2013). A generalized backdoor criterion. *Annals of Statistics* (to appear).

Meek, C. (1995). Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence*. San Mateo, CA: Morgan Kaufmann.

Nandy, P., Maathuis, M. H., & Richardson, T. S. (2014). Estimating the effect of joint interventions from observational data in high-dimensional settings. ArXiv e-prints (1407.2451).

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge: Cambridge University Press.

Peters, J. (2012). *Restricted structural equation models for causal inference*. Doctoral dissertation, ETH Zurich and MPI for Intelligent Systems Tübingen. doi: 10.3929/ethz-a-007597940

Peters, J., & Bühlmann, P. (2014). Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, *101*, 219–228.

Ramsey, J., Zhang, J., & Spirtes, P. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*. N.p.: AUAI Press.

Richardson, T., & Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, *30*, 962–1030.

Shpitser, I., & Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)* (Vol. 2). Cambridge, MA: AAAI Press.

Shpitser, I., Van der Weele, T. J., & Robins, J. M. (2010). On the validity of covariate adjustment for estimating causal effects (corrected version). In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence*. N.p.: AUAI Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). Cambridge, MA: MIT Press.

Textor, J., & Liskiewicz, M. (2011). Adjustment criteria in causal diagrams: An algorithmic perspective. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence*. N.p.: AUAI Press.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, *65*, 31–78.

Received June 22, 2014; accepted October 1, 2014.