Discovering Fully Oriented Causal Networks

Osman Mian, Alexander Marx and Jilles Vreeken
CISPA Helmholtz Center for Information Security
{osman.mian, alexander.marx, jv}@cispa.de

Abstract

We study the problem of inferring causal graphs from observational data. We are particularly interested in discovering graphs where *all* edges are oriented, as opposed to the partially directed graph that the state-of-the-art discover. To this end we base our approach on the *algorithmic* Markov condition. Unlike the *statistical* Markov condition, it uniquely identifies the true causal network as the one that provides the simplest—as measured in Kolmogorov complexity—factorization of the joint distribution. Although Kolmogorov complexity is not computable, we can approximate it from above via the Minimum Description Length principle, which allows us to define a consistent and computable score based on non-parametric multivariate regression. To efficiently discover causal networks in practice, we introduce the GLOBE algorithm, which greedily adds, removes, and orients edges such that it minimizes the overall cost. Through an extensive set of experiments we show GLOBE performs very well in practice, beating the state-of-the-art by a margin.

Introduction

Discovering causal dependencies from observational data is one of the most fundamental problems in science (Pearl 2009). We consider the problem of recovering the causal network over a set of continuous-valued random variables $\boldsymbol{X}$ based on an iid sample from their joint distribution. The state-of-the-art does so by first recovering an undirected causal skeleton—which identifies the variables that have a direct causal relation—and then uses conditional independence tests to orient as many edges as possible. By the nature of these tests this can only be done up to Markov equivalence classes, which means that these methods in practice return networks where only few edges are oriented. In contrast, we develop an approach that discovers fully directed causal graphs.

We base our approach on the algorithmic Markov condition (AMC), a recent postulate that states that the factorization of the joint distribution according to true causal network coincides with the one that achieves the lowest Kolmogorov complexity (Janzing and Schölkopf 2010). As an example, consider the case where X causes Y. Whereas the traditional *statistical* Markov condition cannot differentiate between $P(X)P(Y|X)$ and $P(Y)P(X|Y)$ as both are valid factorizations of joint distribution $P(X,Y)$, the *algorithmic* Markov condition takes the complexities of these distributions into account: in this case, the simplest factorization of $P(X,Y)$ is $K(P(X)) + K(P(Y|X))$ as only this factorization upholds the true independence between the marginal and conditional distribution—any competing factorization will be more complex because of inherent redundancy between the terms. As Kolmogorov complexity can capture any physical process (Li and Vitányi 2009) the AMC is a very general model for causality. However, Kolmogorov complexity is not computable, and hence we need a practical score to instantiate it. Here we do so through the Minimum Description Length principle (Grünwald 2007), which provides a statistically well-founded approach to approximate Kolmogorov complexity from above.

We develop an MDL-based score for directed acyclic graphs (DAGs), where we model the dependencies between variables through non-parametric multivariate regression. Simply put, the lower the regression error of the discovered model, the lower its cost, while more parameters mean higher complexity. We show this score is consistent: given sufficiently many samples from the joint distribution, we can uniquely identify the true causal graph if the causal relations are nearly deterministic. To efficiently discover causal networks directly from data we introduce the GLOBE algorithm, which much like the well-known GES (Chickering 2002) algorithm greedily adds and removes edges to optimize the score. Unlike GES, however, GLOBE traverses the space of DAGs rather than Markov equivalence classes—orienting edges during its search based on the AMC—and hence is guaranteed to result in a fully directed network.

Through extensive empirical evaluation we show that GLOBE performs well in practice and outperforms the state-of-the-art conditional independence and score based causal discovery algorithms. On synthetic data we confirm GLOBE does not discover spurious edges between independent variables, and overall achieves the best scores on both the structural hamming distance and the structural intervention distance. Last, but not least, on real-world data we show that GLOBE even works well when it is unlikely that our modelling assumptions are met.

For reproducibility we provide detailed pseudo-code in technical appendix, and make all code and data available.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Preliminaries

First, we introduce the notation for causal graphs and the main information theoretic concepts that we need later on.

Causal Graph We consider data over the joint distribution of m continuous valued random variables $\boldsymbol{X} = \{X_1, \ldots, X_m\}$. As is common, we assume *causal sufficiency*. That is, we assume that $\boldsymbol{X}$ contains all random variables that are relevant to the system, or in other words, that there exist no latent confounders. Under the assumptions of causal sufficiency and acyclicity, we can model causal relationships over $\boldsymbol{X}$ using a *directed acyclic graph* (DAG). A causal DAG G over $\boldsymbol{X}$ is a graph in which the random variables are the nodes and edges identify the causal relationship between a pair of nodes. In particular, a directed edge between two nodes $X_i \to X_j$ indicates that X_i is a *direct cause* or *parent* of X_j. We denote the set of all parents of X_i with $\text{Pa}(X_i)$.

When working on causal DAGs, we assume the common assumptions, the *causal Markov condition* and the *faithfulness condition*, to hold. Simply put, the combination of both assumptions implies that each separation present in the true graph G implies an independence in the joint distribution P over the random variables $\boldsymbol{X}$ and vice versa (Pearl 2009).

Identifiability of Causality A causal relationship is said to be *identifiable* if it is possible to unambiguously recover it from observational data alone. In general, causal dependencies are not identifiable without assumptions on the causal model. The common assumptions for discovering causal DAGs allow identification up to the Markov equivalence class (Pearl 2009). Given additional assumptions, such as that the relation between cause and effect is a non-linear function with additive Gaussian noise (Hoyer et al. 2009), it is possible to identify causal directions within a Markov equivalence class (Glymour, Zhang, and Spirtes 2019). This is the causal model we investigate.

Kolmogorov Complexity The Kolmogorov complexity of a finite binary string x is the length of the shortest binary program p^* for a universal Turing machine $\mathcal{U}$ that outputs x and then halts (Kolmogorov 1965; Li and Vitányi 2009). Simply put, p^* is the most succinct *algorithmic* description of x, and therewith Kolmogorov complexity of x is the length of its ultimate lossless compression. Conditional Kolmogorov complexity, $K(x \mid y) \leq K(x)$, is then the length of the shortest binary program p^* that generates x, and halts, given y as input.

The Kolmogorov complexity of a probability distribution P, $K(P)$, is the length of the shortest program that outputs $P(x)$ to precision q on input $\langle x, q \rangle$ (Li and Vitányi 2009). More formally, we have

$$K(P) = \min \left\{ |p| : p \in \{0,1\}^*, |\mathcal{U}(p, x, q) - P(x)| \leq \frac{1}{q} \right\} .$$

The conditional, $K(P \mid Q)$, is defined similarly except that the universal Turing machine $\mathcal{U}$ now gets the additional information Q. For more details on Kolmogorov complexity see Li and Vitányi (2009).

Minimum Description Length Principle Although Kolmogorov complexity is not computable, we can approximate it from above through lossless compression (Li and Vitányi 2009). The Minimum Description Length (MDL) principle (Rissanen 1978; Grünwald 2007) provides a statistically well-founded and computable framework to do so. Conceptually, instead of all programs, *ideal MDL* considers only those programs for which we know that they output x and halt, i.e., lossless compressors. Formally, given a model class $\mathcal{M}$, MDL identifies the best model $M \in \mathcal{M}$ for data D as the one minimizing $L(D, M) = L(M) + L(D \mid M)$, where $L(M)$ is the length in bits of the description of M, and $L(D \mid M)$ is the length in bits of the description of data D given M. This is known as two-part, or *crude* MDL. There also exists one-part, or *refined* MDL. Although refined MDL has theoretically appealing properties, it is efficiently computable for a small number of model classes. Asymptotically, there is no difference between the two (Grünwald 2007).

To use MDL in practice we need to define a model class, and how to encode a model, resp. the data given a model, into bits. Note that we are only concerned with optimal code *lengths*, not actual codes—our goal is to measure the *complexity* of a dataset under a model class, after all (Grünwald 2007). Hence, all logarithms are to base 2, and we use the common convention that $0 \log 0 = 0$.

Theory

In this section, we will first introduce the algorithmic model of causality which is based on Kolmogorov complexity. To put it into practice, we need to introduce a set of modelling assumptions that allow us to approximate it using MDL. We conclude this section by providing consistency guarantees.

Algorithmic Model of Causality

Here we introduce the main concepts of algorithmic causal inference as introduced by Janzing and Schölkopf (Janzing and Schölkopf 2010), starting with the causal model.

Postulate 1 (Algorithmic Model of Causality). *Let G be a DAG formalizing the causal structure among the strings $x_1, \ldots, x_m$. Then, every x_j is computed by a program q_j with constant length from its parents $Pa(x_j)$ and an additional input n_j. That is*

$$x_j = q_j(Pa(x_j), n_j) \ ,$$

where the inputs n_j are jointly independent.

As any mathematical object x can be described as a binary string, and a program q_j can model any physical process (Deutsch 1985) or possible function h_j (Li and Vitányi 2009), this is a particularly general model of causality. Equivalent to the statistical model, we can derive that the algorithmic model of causality fulfils the *algorithmic* Markov property (Janzing and Schölkopf 2010), that is

$$K(x_1, \ldots, x_m) \overset{+}{=} \sum_{j=1}^{m} K(x_j \mid \text{Pa}^*(x_j)) \ ,$$

where $\overset{+}{=}$ denotes equality up to an additive constant. Meaning, to most succinctly describe all strings, it suffices to know what are the parents and additional inputs n_j for each string x_j. Unlike its statistical counterpart which can only identify the causal network up to Markov equivalence, the *algorithmic* Markov property can identify a single DAG as the most succinct description of all strings. As any mathematical object, including distributions, can be described by a binary string, Janzing and Schölkopf (2010) define the following postulate.

Postulate 2 (Algorithmic Markov Condition)**.** *A causal DAG G over random variables $\boldsymbol{X}$ with joint density P is only acceptable if the shortest description of P factorizes as*

$$K(P(X_1, \ldots, X_m)) \overset{+}{=} \sum_{j=1}^{m} K(P(X_j \mid \mathit{Pa}(X_j))) \ . \quad (1)$$

Hence, under the assumption that the true causal graph can be modelled by a DAG, it has to be the one minimizing Eq. (1). As K is not computable we cannot directly compute this score. What we can however, restrict our model class from allowing all possible functions to a subset of these and then approximate K using MDL.

Causal Model

As causal model we consider a rich class of structural equation models (Pearl 2009) (SEMs) where the value of each node is determined by a linear combination of functions over all possible subsets of parents and additional independent noise. Formally, for all $X_i \in \boldsymbol{X}$ we have

$$X_i := \sum_{\mathcal{S}_j \in \mathcal{P}(\text{Pa}(X_i))} h_j(\mathcal{S}_j) + N_i \ , \quad (2)$$

where h_j is a non-linear function of the j-th subset over the power set, $\mathcal{P}(\text{Pa}(X_i))$, of parents of X_i, and N_i is an independent noise term. We assume that all noise variables are jointly independent, Gaussian distributed and that $N_i \perp\!\!\!\perp \text{Pa}(X_i)$.

MDL Encoding of the Causal Model

Next, we specify our MDL score for DAGs. Given an iid sample $\boldsymbol{X}^n$ drawn from the joint distribution P over $\boldsymbol{X}$, our goal is to approximate Eq. (1) using two-part MDL, which means we need to define a model class $\mathcal{M}$ for which we can compute the optimal code length. Here, we define $\mathcal{M}$ to include all possible DAGs over $\boldsymbol{X}$ and their corresponding parametrization according to our causal model. That is, for each node X_i a model $M \in \mathcal{M}$ contains an index indicating the parents of X_i, which is equivalent to storing the DAG structure, and the corresponding functional dependencies.

Building upon Eq. (1), we want to find that model $M^* \in \mathcal{M}$ such that

$$\begin{aligned} M^* &= \operatorname*{argmin}_{M \in \mathcal{M}} L(\boldsymbol{X}^n, M) \\ &= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{m} L(X_i^n \mid \text{Pa}(X_i), M) \right) \\ &= \operatorname*{argmin}_{M \in \mathcal{M}} \left(L(M) + \sum_{i=1}^{m} L(\epsilon_i) \right) \end{aligned}$$

where $\text{Pa}(X_i)$ are the parents of X_i according to the model M. In the last line, we replace $L(X_i^n \mid \text{Pa}(X_i), M)$ with $L(\epsilon_i)$ to clarify that encoding a node given M and its parents comes down to encoding the residuals ϵ_i.

Encoding the Model The model complexity $L(M)$ for a model $M \in \mathcal{M}$, comprises of the parameters of the functional dependencies and the graph structure. The total cost is simply the sum of the code lengths of the individual nodes

$$L(M) = \sum_{i=1}^{m} L(M_i) \ .$$

To encode the individual nodes X_i, we need to transmit its parents, the form of the functional dependency, and the bias or mean shift μ_i. We encode the model M_i for a node X_i as

$$L(M_i) = L_{\mathbb{N}}(k) + k \log m + L_F(f_i) + L_p(\mu_i) \ ,$$

where we first encode the number of parents using $L_{\mathbb{N}}$, the MDL-optimal encoding for integers $z \geq 0$ (Rissanen 1983). It is defined as $L_{\mathbb{N}}(z) = \log^* z + \log c_0$, where $\log^* z = \log z + \log \log z + \ldots$ and we consider only the positive terms, and c_0 is a normalization constant to ensure the Krafft-inequality holds (Kraft 1949). Next, we identify which out of the m random variables these are, and then proceed to encode the function f_i over these parents, where f_i represents the summation term on the right hand side of Eq. (2). Last, we encode the bias term using L_p, defined later in Eq.(3).

Encoding the Functions We will instantiate the framework using non-parametric functions h_i that also allow for non-linear transformations of the parent variables. To this end, we fit non-parametric Multivariate Adaptive Regression Splines (Friedman 1991). In essence, we estimate X_i as

$$\hat{X}_i := \sum_{j=1}^{|H|} h_j(\mathcal{S}_j) \ ,$$

where h_j is called a hinge function that is applied to a subset of the parents, $\mathcal{S}_j$, with size $|\mathcal{S}_j|$, that is associated with the j-th hinge. A hinge takes the form

$$h(\mathcal{S}) = \prod_{i=1}^{T} a_i \cdot \max(0, g_i(s_i) - b_i) \ ,$$

where T denotes the number of multiplicative terms in h, $s_i \in \mathcal{S}$ is the parent associated with the i-th term, g_i is a non-linear transformation applied to s_i where g_i belongs to the function class $\mathcal{F}$, e.g. the class of all polynomials up to a certain degree. We specify $\mathcal{F}$ in more detail in the supplementary section, but the encoding can be very general and can include any regression function as long as we can describe the parameters and $|\mathcal{F}| < \infty$. If $T = 1$ for all hinges, the above definition simplifies to an additive model over individual parents. We encode a hinge function as follows

$$\begin{aligned} L_F(h) = \ & L_{\mathbb{N}}(|H|) + \sum_{h_j \in H} \Big[\, L_{\mathbb{N}}(T_j) + \log \binom{|\mathcal{S}| + T_j - 1}{T_j} \\ & + T_j \log(|\mathcal{F}|) + L_p(\theta(h_j)) \, \Big] \end{aligned}$$

First, we use $L_{\mathbb{N}}$ to encode the number of hinges and the number of terms per hinge. We then transmit the correct assignment of terms T_j to parents in $\mathcal{S}$, and finally need $\log(|\mathcal{F}|)$ bits to identify the specific non-linear transformation that is used for each of the T_j terms in the hinge.

Encoding Parameters To encode the bias we use the proposal of Marx and Vreeken (2017) for encoding parameters up to a user specified precision p. We have

$$L_p(\theta) = |\theta| + \sum_{i=1}^{|\theta|} L_{\mathbb{N}}(s_i) + L_{\mathbb{N}}(\lceil \theta_i \cdot 10^{s_i} \rceil) \ , \qquad (3)$$

where s_i is the smallest integer such that $|\theta_i| \cdot 10^{s_i} \geq 10^p$. Simply put, $p = 2$ implies that we consider two digits of the parameter. We need one bit to store the sign of the parameter, then we encode the shift s_i and the shifted parameter θ_i.

Encoding Residuals Last, we need to encode the residual term, $L(\epsilon_i)$. Since we use regression functions, we aim to minimize variance of the residual—and hence should encode the residual ϵ as Gaussian distributed with zero-mean (Marx and Vreeken 2017; Grünwald 2007)

$$L(\epsilon) = \frac{n}{2} \left(\frac{1}{\ln 2} + \log 2\pi\hat{\sigma}^2 \right) \ ,$$

where we can compute the empirical variance $\hat{\sigma}^2$ from ϵ.

Combining the above, we now have a lossless MDL score for a causal DAG.

Consistency

Since MDL can only upper bound Kolmogorov complexity, but not compute it, it is not possible to directly derive strict guarantees from the AMC. We can, however, derive consistency results. We first show that our score allows for identifying the Markov equivalence class of the true DAG i.e. the partially directed network for which each collider is correctly identified. Then, we show that under slightly stricter assumptions, we can orient the remaining edges correctly.

The main idea for the first part is to show that our score is consistent—simply put, *the likelihood term dominates in the limit*. For a score with such properties e.g. BIC (Haughton 1988), Chickering (2002) showed that it is possible to identify the Markov equivalence class of the true DAG. To show that our score behaves in the same way, we need to make two light weight assumptions for $n \to \infty$:

1. the number of hinges of $|H|$ is bounded by $\mathcal{O}(\log n)$, and
2. the precision of the parameters θ is constant w.r.t. to n and hence $L_p(\theta) \in \mathcal{O}(1)$.

Based on these assumptions, we can show that our score is consistent as it asymptotically behaves like BIC, meaning that the penalty term for the parameters only grows with $\mathcal{O}(\log n)$ complexity, while the likelihood term grows linearly with n and hence is the dominating term as $n \to \infty$.

Theorem 1. *Given a causal model as defined in Eq.* (2) *and corresponding data $\boldsymbol{X}^n$ drawn iid from joint distribution P. Under Assumptions (1) and (2), $L(\boldsymbol{X}^n, M)$ asymptotically behaves like BIC.*

Algorithm 1: The GLOBE Algorithm

```
Data: Data $\boldsymbol{X}^n$ over $\boldsymbol{X}$
Result: Causal DAG $G$
$Q \leftarrow$ EDGESCORING($\boldsymbol{X}^n$)
$G \leftarrow$ FORWARDSEARCH($Q, \boldsymbol{X}^n$)
$G \leftarrow$ BACKWARDSEARCH($G$)
return $G$
```

With the above, we know that given sufficient data our score will identify the correct Markov equivalence class.

To infer the complete DAG, we need to be able to infer the direction for those edges that cannot be inferred using collider structures—i.e. single edges like $X - Y$. Closest to our approach is the work of Marx and Vreeken (2019) who showed that it is possible to distinguish between $X \to Y$ and $Y \to X$ using any L_0 regularized score—e.g. BIC, if we assume that the underlying causal function is near deterministic i.e. $Y := f(X) + \alpha N$, where f is a non-linear function and N is an unbiased, unit-variance noise regulated by a small constant $\alpha > 0$. Since our score in the limit behaves like an L_0-based score (ref. Theorem 1), we can distinguish between Markov equivalent DAGs under these stricter assumptions. For a detailed discussion, readers are directed to the proof of Theorem 1 in technical appendix.

Although our score is consistent and can be used to distinguish Markov equivalent DAGs, these guarantees only hold if we were to score all DAGs over $\boldsymbol{X}$. Since this is infeasible for large graphs, we propose a modified greedy DAG search algorithm to minimize $L(\boldsymbol{X}^n, M)$.

The GLOBE Algorithm

We now present GLOBE, a score-based method for discovering directed acyclic causal graphs from multivariate continuous valued data. GLOBE consists of three steps: edge scoring, forward and backward search, as shown in Algorithm 1.[1]

Edge Scoring To improve the forward search where we greedily add the edge that provides the highest gain, we first order all potential edges in a priority queue by their causal strength. We measure the causal strength of an edge, using the absolute gain in bits for orienting an edge in either direction in our model. Formally, let $e = (X_i, X_j)$ be an undirected edge between X_i and X_j, and further let $\vec{e}$ refer to the directed edge $X_i \to X_j$ and $\overleftarrow{e}$ the directed edge in the reverse direction. Now, let M be the current model. We write $M \oplus \overleftarrow{e}$ to refer to the model where we add edge $\overleftarrow{e}$, and $M \oplus \vec{e}$ for the model where we add $\vec{e}$. We define the gain in bits, δ, associated with edge $\overleftarrow{e}$ as

$$\delta(\overleftarrow{e}) = \max\{0, L(\boldsymbol{X}^n, M) - L(\boldsymbol{X}^n, M \oplus \overleftarrow{e})\}$$

where $L(\boldsymbol{X}^n, M)$ is defined according to the causal model specified in the theory section, and define $\delta(\vec{e})$ analogously. Based on $\delta(\overleftarrow{e})$ and $\delta(\vec{e})$, we define the directed gain $\Psi(\overleftarrow{e})$ for a given edge as

$$\Psi(\overleftarrow{e}) = \delta(\overleftarrow{e}) - \delta(\vec{e}) \ ,$$

[1] We provide detailed pseudocodes in the technical appendix

where $\Psi(\overleftarrow{e}) = -\Psi(\vec{e})$. The higher the value of $\Psi(\overleftarrow{e})$, the higher edge $\overleftarrow{e}$ is ranked. Intuitively, the larger the difference between the edge direction, the more certain we are that we inferred the correct direction. The algorithm for this step is straightforward, we pick each undirected edge e, calculate δ and Ψ for $\overleftarrow{e}$ and $\vec{e}$, and add the edges to a priority queue.

Forward Search For forward search phase, we use the priority queue obtained from the edge ranking step to build the causal graph by iteratively adding the highest ranked edge. We reject edges that would introduce a cycle. After adding an edge $X_i \rightarrow X_j$ we need to update the score of all edges pointing towards X_j and re-rank them in the priority queue. Due to the greedy nature of the algorithm, we may add edges in the wrong direction when we do not yet know all the parents of a node. Hence, after adding edge $X_i \rightarrow X_j$ to the current model—i.e. discovering a new parent for X_j—we check for all children of X_j, whether flipping the direction of the edge improves the overall score. If so, we delete that edge $\vec{e}$ from our model, re-calculate δ and Ψ for $\vec{e}$ and $\overleftarrow{e}$, and push them again to the priority queue (see Fig. 1). The forward search stops when the priority queue is empty.

To avoid spurious edges, we check for significance of the gain. Let $k = \delta(\overleftarrow{e})$, based on the no-hypercompression inequality (Grünwald 2007), the probability to gain k bits over the null model is smaller or equal to 2^{-k}. If for an edge the gain k is not significant—i.e. $2^{-k} > \alpha$, where α is a user defined significance threshold, we disregard the edge.

Backward Search To further refine the graph discovered in the forward search, we iteratively remove superfluous edges. In particular, for each node X_j with $|\mathrm{Pa}(X_j)| = k \geq 2$ we score all graphs for which we only use a subset of the parents of size $k-1$. If any of these graphs provides a gain in compression, we select the one that provides the largest gain and update the model accordingly. We continue this process until we cannot find such a subset for any node and output the current graph as our predicted causal DAG.

Complexity Analysis

The edge ranking does one pass over the edges, it has a runtime of $\mathcal{O}(|V|^2)$. In the forward search, each edge can lead to at most $(|V|-1)$ ranking updates due to edge flips. Resulting in a total complexity in $\mathcal{O}(|V|^3)$. The backwards search has a loose upper bound of $\mathcal{O}(|V|^3)$, that results when the forward search returns a fully connected graph and we delete each of those edges in the backwards search. Hence, the overall complexity of Globe is in $\mathcal{O}(|V|^3)$. In practice, Globe is fast enough for networks as large as 500 nodes.

Instantiation

We instantiate Globe[2] using the open-source implementation in R of Multivariate Adaptive Regression Splines framework (Friedman 1991). Since we could face issues like multicollinearity (Farrar and Glauber 1967) and unrealistic run times if we allow for arbitrary many interactions between parents, we restrict the maximum number of interaction terms to 2 for experiments.

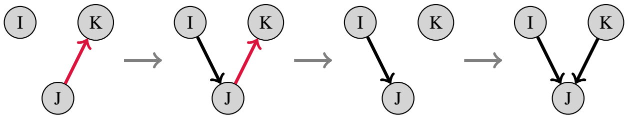


Figure 1: Edge reversal in the forward search: We start with the graph where we wrongly added edge $X_j \rightarrow X_k$, then we add the correct edge $X_i \rightarrow X_j$. Revisiting the children of X_j we see that flipping $X_j \rightarrow X_k$ improves our score and hence delete the edge. In the next step we add the correct edge.

Related Work

Causal discovery on observational data has drawn more attention in recent years (Bühlmann et al. 2014; Huang et al. 2018; Hu et al. 2018; Margaritis and Thrun 2000) and is still an open problem. To give a succinct overview, we focus on the most related methods, ones that aim to recover a DAG or its Markov equivalence class from continuous valued data. We exclude methods that aim at weakening assumptions such as causal sufficiency or acyclicity (Spirtes et al. 2000), or methods for discrete data (Budhathoki and Vreeken 2017).

Most approaches can be classified as constraint based or score based. Both rely on the Markov and faithfulness conditions to recover Markov equivalence classes of the true DAG. Constraint based methods such as the PC and FCI algorithm (Spirtes et al. 2000), their extensions (Colombo and Maathuis 2014; Pearl, Verma et al. 1991) as well as the Grow-Shrink algorithm (Margaritis and Thrun 2000) rely on conditional independence (CI) tests to first recover the undirected causal graph and then infer edge directions only up to the Markov equivalence class using additional edge orientation rules (Meek 1995). The main bottleneck for those approaches is the CI test. The standard choice is the Gaussian CI test (Kalisch and Bühlmann 2007). However, it cannot capture non-linear correlations. The current state-of-the-art uses kernel based tests such as HSIC (Gretton et al. 2005), which can capture non-linear dependencies.

Score based methods define a scoring function, $S(G, \boldsymbol{X}^n)$, that evaluates how well a causal DAG G fits the provided data $\boldsymbol{X}^n$. If the true causal graph G^* is a DAG, then given infinite data the highest scoring DAG is part of the equivalence class of G^* (Chickering 2002). Score based approaches start with an empty graph and greedily traverse to the highest scoring Markov equivalence class that is reachable by adding, deleting or reversing an edge. Well-known algorithms in this category include the greedy equivalence search (GES) (Chickering 2002; Hauser and Bühlmann 2012), its extensions (Ramsey et al. 2017), and the current state-of-the-art, generalized-GES (GGES) (Huang et al. 2018) which uses kernel regression to capture complex dependencies.

In contrast, additive noise models (ANMs) aim to discover the fully directed graph (Hoyer et al. 2009). The primary as-

[2] Globe stems from discovering fully, rather than locally, oriented networks, as well as from it being based on Multivariate Adaptive Regression Splines (Mars), of which the public implementation is known as Earth.

sumption is that the effect can be written as a function of the cause plus additive noise that is independent of the cause. Under this assumption, the function is only admissible in causal direction and not vice-versa (Hoyer et al. 2009). Methods range from linear non-Gaussian (LINGAM) (Shimizu et al. 2006), non-linear functions (RESIT) (Peters et al. 2014) to mixtures of non-linear additive noise models (Hu et al. 2018). The main caveat of ANMs is also the CI test. Fitting a non-linear function that maximizes the independence between the cause and noise is a slow process which restricts ANMs application to small networks (Hoyer et al. 2009).

Most related to our work are methods based on regression error. Those methods have been shown to successfully decide between Markov equivalent DAGs under the assumption of having a non-linear function and low noise (Marx and Vreeken 2017; Blöbaum et al. 2018; Marx and Vreeken 2019) or proven to correctly identify the causal ordering of all nodes (CAM) (Bühlmann et al. 2014). Directly comparing a causal ordering to a DAG is, however, not straightforward.

In this paper, we combine the advantages of score based methods and methods based on regression error by discovering the fully oriented graph and allowing for complex non-linear dependencies, while being fast in practice.

Experiments

We evaluate GLOBE on both synthetic and real-world data with known ground truth. GLOBE is implemented in Python and both the source code, as well as the synthetic data are made available for reproducibility.[3] We compare GLOBE to the state-of-the-art from different classes of algorithms. We compare to RESIT (Peters et al. 2014) and LINGAM (Shimizu et al. 2006) as representative ANM-based methods, to GGES as the best score-based method (Huang et al. 2018), and to PC with the Hilbert Schmidt Independence Criteria, short PC_{HSIC} (Colombo and Maathuis 2014; Gretton et al. 2005), as the state-of-the-art constraint-based method for causal discovery. Comparison with FASTGES (Ramsey et al. 2017) is ommitted since its performance was significantly worse than the other methods. We provide details on experimental setup as well as additional experiments, involving a case-study in the technical appendix. GLOBE finished within ten minutes for each experimental instance except one real-world dataset with 500 nodes, on which it took 3 days. While the competitors could not handle this data.

Evaluation Metrics We evaluate the predicted and the ground truth graphs on the basis of their structural, as well as their *causal* similarity. We justify using our proposed evaluation metrics in the technical appendix.

The Structural Hamming Distance (SHD) (Kalisch and Bühlmann 2007), between two partially directed acyclic graphs (PDAGs) G and $\hat{G}$ is the the total number of edges where the two graphs differ. Denoting the edge adjacency matrix of G and $\hat{G}$ with X resp. $\hat{X}$ we have

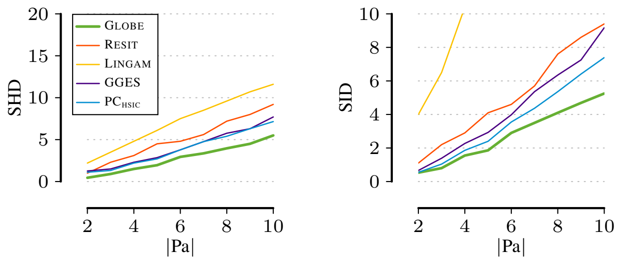


Figure 2: [Lower is better] SHD (left) and SID (right) for increasing number of parents.

$$SHD(G, \hat{G}) \coloneqq \sum_{1 \le i < j \le m} \mathbf{I}((X_{ij} \oplus \hat{X}_{ij}) \vee (X_{ji} \oplus \hat{X}_{ji})) \; ,$$

where $\oplus$ denotes an XOR operation and $\mathbf{I}(x)$ is 1 when the expression x is $true$ and 0 otherwise.

However, SHD tells us nothing about the causal similarity between two graphs. Hence, we use the Structural Intervention Distance (SID) (Peters and Bühlmann 2015) pre-metric. SID counts the pairs of nodes u and v such that the effect of intervention from u to v is falsely estimated by $\hat{G}$ with respect to G. In case of a PDAG, SID is an interval, with smallest and largest scores indicating the best resp. worst scores for the DAGs in a given Markov equivalence class.

Synthetic Data

We start with a sanity check to ensure that GLOBE can reliably avoid false positives and build up to the case of varying sample sizes over a more complex network. We generated 100 instances each with 1 000 observations for the discussed structures, unless stated otherwise. We standardized the data to have zero mean and unit variance.

Independent Data As a sanity check, we test the methods on instances of a graph containing 10 independent nodes where the value of each node is sampled independently from a Gaussian distribution. We expect all the methods to report empty sets of edges for the instances in this experiment. GLOBE did not report a single spurious edge on *any* of the instances. On the other hand, LINGAM reported at least one spurious edge for 38%, RESIT for 42% and PC_{HSIC} and GGES for half resp. 10% of the instances.

Effect of Multiple Parents Next we test GLOBE on a simple case of a collider where we vary the number of parents from 2 up to 10. The collider node is calculated as a linear combination of non-linear parent functions given as

$$X_j = \sum_{X_i \in \text{Pa}(X_j)} a_i \cdot (X_i + b_i)^{c_i} \quad . \tag{4}$$

Since it is possible to identify a collider structure using conditional independence tests, we expect GGES and PC_{HSIC} to discover a fully directed network. The results for both

[3] http://eda.mmci.uni-saarland.de/globe/

Table 1: [Lower is Better] Averaged normalized *SID* for the methods. Interval for GGES and PC$_{\text{HSIC}}$ indicates the best, resp. worst possible intervention distance for the DAGs in the discovered Markov equivalence class.

n	GLOBE	RESIT	LINGAM	GGES	PC$_{\text{HSIC}}$
100	**0.28**	0.45	0.47	[0.18 , 0.48]	[0.28 , 0.54]
500	**0.26**	0.43	0.43	[0.17 , 0.48]	[0.21 , 0.55]
1000	**0.26**	0.42	0.42	[0.17 , 0.48]	[0.20 , 0.54]
1500	**0.27**	0.40	0.43	[0.17 , 0.48]	[0.19 , 0.53]
2000	**0.26**	0.40	0.40	[0.18 , 0.49]	[0.19 , 0.54]

SHD and *SID* are shown in Figure 2. In case of *SID*, we compare favorably to both GGES and PC$_{\text{HSIC}}$ by only reporting the *best possible* achievable score for their predicted graphs' Markov equivalence class. Even with this favorable comparison, GLOBE outperforms the competition.

Data Sampled from a Causal Network Next, we show GLOBE's effectiveness in finding the causal relationships in a more general setting. We consider multiple instances of the graph that contains all possible connections that could exist in a DAG. In this setting, each child node, X_j can alternatively be calculated using more complex multiplicative interactions between the parents given by

$$X_j = a_j \cdot \prod_{X_i \in \text{Pa}(X_j)} {X_i}^{c_i} + b_j \ . \tag{5}$$

We generate data where we choose between Eq. (4) and (5) with probability 0.7 resp. 0.3 and report results over varying sample sizes. We report the values for *SID* in Table 1. Overall we see that GLOBE outperforms RESIT and LINGAM by a margin. The causal networks predicted by GLOBE have *SID* closer to the better end of the range of scores possible for PC$_{\text{HSIC}}$ and GGES. In terms of *SHD*, all the methods were found to be consistent over varying sample sizes, with GLOBE slightly outperforming the competition.

Real World Data

For real world data with known ground truth, we consider three distinct networks of sizes 5, 15 and 500 nodes from the reged dataset (Statnikov et al. 2015), each containing 1 000 rows. Looking at the results shown in Figure 3, we see that GLOBE is closest to the true causal network for both the 5 node (REGED5) and the 15 node (REGED15) network. For REGED15, GLOBE reports a better *SID* than all the competitors. We see that for the REGED15 network, GGES fails to orient most of the edges, which results in a graph where both extremes of the *SID* are possible.

For the 500 node network, GLOBE was the *only* algorithm to produce any kind of result in reasonable time (3 days), with a reported normalized *SID* and *SHD* of 0.1 resp. 0.01. While GGES failed to terminate within one month, all other methods could not process the data.

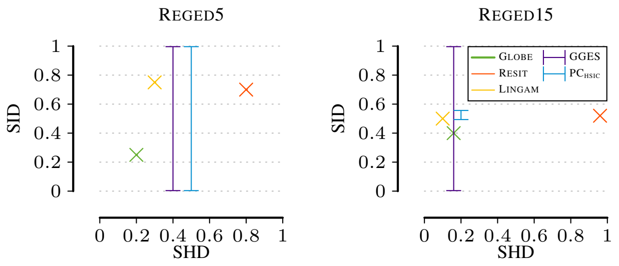


Figure 3: [Closer to Origin is Better] Comparison of Normalized *SHD* and Normalized *SID* for real world networks.

Discussion and Future Work

Instantiating GLOBE using the MARS framework is just one of the many realizations of the algorithm. Other regression approaches, as long as we define a consistent lossless encoding for them, can also be incorporated into GLOBE and may give better results based on the application domain. For proof of concept, we show how to instantiate GLOBE using parametric regression in the technical appendix.

Due to computational reasons, we only traverse the space of DAGs and not the Markov equivalence classes, which could result in a locally optimal solution. We try to mitigate this using the edge flipping step during the forward search. However, by incorporating a more complex search strategy, like the beam search, we could both expand our search space, and eliminate the need for the edge flip.

Our score is specifically defined for continuous valued data. An extension of GLOBE would be to discover causal relationships over discrete and mixed type data. As MDL-based scores have been proposed for inference on discrete (Budhathoki and Vreeken 2017) and mixed (Marx and Vreeken 2018) data, but only for pairs of variables, it would be interesting to extend GLOBE to handle both cases.

Conclusion

We considered discovering fully directed causal graphs from observational data. To tackle this problem, we built upon the algorithmic Markov condition that is based on Kolmogorov complexity. Since the latter cannot be computed directly, we proposed a score based on MDL to approximate it from above. We showed that for non-linear mixture models with additive noise, our score allows for discovering the Markov equivalence class of the true DAG and if the noise term is assumed to have a low variance, we can discover the fully directed causal graph. To minimize our score, we proposed GLOBE, a greedy DAG search algorithm that iteratively builds a DAG to find a locally optimal solution. We modeled functional dependencies using non-parametric regression functions.

Through an extensive set of experiments, we showed that GLOBE beats the state-of-the-art by a margin, reliably orients the edges in the presence of multiple parents, discovers graphs that are *structurally* and *causally* similar to the ground truth and is fast enough to infer networks up to 500 nodes.

References

Blöbaum, P.; Janzing, D.; Washio, T.; Shimizu, S.; and Schölkopf, B. 2018. Cause-Effect Inference by Comparing Regression Errors. In *AISTATS*, 900–909.

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

Bühlmann, P.; Peters, J.; Ernest, J.; et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals Stat.* 42(6): 2526–2556.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *JMLR* 3(Nov): 507–554.

Colombo, D.; and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. *JMLR* 15(1): 3741–3782.

Deutsch, D. 1985. Quantum Theory, the Church-Turing Principle and the Universal Quantum Computer. *R. Statist. Soc. A* 400(1818): 97–117.

Farrar, D. E.; and Glauber, R. R. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics* .

Friedman, J. H. 1991. Multivariate adaptive regression splines. *The annals of statistics* 1–67.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* .

Gretton, A.; Bousquet, O.; Smola, A.; and Schölkopf, B. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT*. Springer.

Grünwald, P. 2007. *The Minimum Description Length Principle*. MIT Press.

Haughton, D. M. 1988. On the choice of a model to fit data from an exponential family. *Annals Math. Stat.* 16(1): 342–355.

Hauser, A.; and Bühlmann, P. 2012. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *JMLR* 13(Aug): 2409–2464.

Hoyer, P. O.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. In *NIPS*, 689–696.

Hu, S.; Chen, Z.; Partovi Nia, V.; CHAN, L.; and Geng, Y. 2018. Causal Inference and Mechanism Clustering of A Mixture of Additive Noise Models. In *NeurIPS*.

Huang, B.; Zhang, K.; Lin, Y.; Schölkopf, B.; and Glymour, C. 2018. Generalized Score Functions for Causal Discovery. In *KDD*. ACM.

Janzing, D.; and Schölkopf, B. 2010. Causal Inference Using the Algorithmic Markov Condition. *IEEE TIT* 56(10): 5168–5194.

Kalisch, M.; and Bühlmann, P. 2007. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *JMLR* 8(Mar): 613–636.

Kolmogorov, A. 1965. Three Approaches to the Quantitative Definition of Information. *Problemy Peredachi Informatsii* 1(1): 3–11.

Kraft, L. G. 1949. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. Ph.D. thesis, Massachusetts Institute of Technology.

Li, M.; and Vitányi, P. 2009. *An Introduction to Kolmogorov Complexity and its Applications*. Springer.

Margaritis, D.; and Thrun, S. 2000. Bayesian network induction via local neighborhoods. In *NIPS*, 505–511.

Marx, A.; and Vreeken, J. 2017. Telling Cause from Effect using MDL-based Local and Global Regression. In *ICDM*, 307–316. IEEE.

Marx, A.; and Vreeken, J. 2018. Causal inference on multivariate and mixed-type data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 655–671. Springer.

Marx, A.; and Vreeken, J. 2019. Identifiability of Cause and Effect using Regularized Regression. In *KDD*. ACM.

Meek, C. 1995. Causal Inference and Causal Explanation with Background Knowledge. In *UAI*, 403–410. Morgan Kaufmann Publishers Inc.

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

Pearl, J.; Verma, T.; et al. 1991. A theory of inferred causation. *KR* 91: 441–452.

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

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal Discovery with Continuous Additive Noise Models. *JMLR* 15.

Ramsey, J.; Glymour, M.; Sanchez-Romero, R.; and Glymour, C. 2017. A million variables and more: the Fast Greedy Equivalence Search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics* .

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14(1): 465–471.

Rissanen, J. 1983. A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals Stat.* 11(2): 416–431.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *JMLR* 7.

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

Statnikov, A.; Ma, S.; Henaff, M.; Lytkin, N.; Efstathiadis, E.; Peskin, E. R.; and Aliferis, C. F. 2015. Ultra-Scalable and Efficient Methods for Hybrid Observational and Experimental Local Causal Pathway Discovery. *JMLR* 16: 3219–3267.

Discovering Fully Oriented Causal Networks

Technical Appendix

A. Consistency and Proofs

Theorem 1. *Given a causal model as defined in Eq.* (2) *(in the main text) and corresponding data $\boldsymbol{X}^n$ drawn iid from joint distribution P. Under Assumptions (1) and (2), $L(\boldsymbol{X}^n, M)$ asymptotically behaves like BIC.*

Proof. First note that we can rewrite the encoding of the residuals $L(\epsilon)$ as $c_1 n \log \hat{\sigma}^2 + \mathcal{O}(1)$, where the additive constant is independent of the model. Next, we upper bound $L(M)$. From Assumption (1) we get that $|H| \in \mathcal{O}(\log n)$. Per hinge we need to encode the number of multiplicative terms $L_{\mathbb{N}}(T_j)$, the function type per term $T_j \log |\mathcal{F}|$, the number of possible assignments from terms to parents $\log \binom{|\mathcal{S}|+T_j-1}{T_j}$ and the parameter vector per hinge $L_p(\theta(h_j))$. Each parameter vector is constant, by Assumption (2). Since the number of parents are independent of n as they are fixed for a certain network, the number of possible interacting terms T_j is also constant w.r.t. n, which means that for large n $L_{\mathbb{N}}(T_j)$, $T_j \log |\mathcal{F}|$ (for a finite function class) and $\log \binom{|\mathcal{S}|+T_j-1}{T_j}$ are also constants. Since we encode for each non-source node a function where we need to encode each hinge, we get an asymptotic complexity of $c_2 \log n + \mathcal{O}(1)$. In addition, we need to encode the parents and number of hinges for each node, which adds to the constant term. Combining the above statements, we arrive at

$$c_1 n \log \hat{\sigma}^2 + c_2 \log n + \mathcal{O}(1) \ .$$

If we set $c_1 = 1$ and $c_2 = \frac{d}{2}$, where d is the number of degrees of freedom of the model, we arrive at the BIC score. □

Based on Theorem 1 and the knowledge the BIC is consistent (Haughton 1988), we can conclude that $L(\boldsymbol{X}^n, M)$ is also consistent. Further, Chickering (Chickering 2002) showed that it is possible to identify the Markov equivalence class of the true DAG with a consistent score.

With the above, we know that given sufficient data our score will identify the correct Markov equivalence class. To infer the complete DAG, we need to be able to decide between Markov equivalent DAGs. That is, we need to be able to infer the direction for those edges that cannot be inferred using collider structures—i.e. single edges like $X - Y$ where either X is the only parent of Y or vice versa Y is the only parent of X. We can exclude that the child node of both nodes has multiple parents, as otherwise we could have inferred them by detecting a collider structure. Hence, we need to be able to distinguish between the two DAGs $X \to Y$ and $Y \to X$, which is a well studied problem (Peters, Janzing, and Schölkopf 2017), that, however, can only be solved under stronger assumptions (Pearl 2009). Closest to our approach is the work of Marx and Vreeken (2019). They show that it is possible to distinguish between $X \to Y$ and $Y \to X$ using any L_0 regularized score—e.g. BIC. To use their result, we need to make the following assumptions: If in the true generating process, Y is the effect of X, where X and Y are continuous random variables with compact supports, Y is generated as

$$Y := f(X) + \alpha N,$$

where f is a non-linear function and N is an unbiased noise variable with unit variance that is regulated by a small constant $\alpha > 0$. Further, let ϕ be the function the minimizes the expected error if we fit Y as a function of X and ψ be defined equivalently for fitting X as a function of Y. From the algorithmic Markov condition, Marx and Vreeken (2019) derived that the function ϕ in the true causal direction is at most as complex—in terms of Kolmogorov complexity—as the function ψ in the anti-causal direction. Since Kolmogorov complexity is not computable, they formulate a practical version of this statement as an assumption. Simply put, the causal function needs at most as many parameters θ_ψ as the function for the anti causal direction. Formally, they assume that $\|\theta_\phi\|_0 \leq \|\theta_\psi\|_0$, which holds for many generating functions such as invertible functions, linear combinations of polynomial functions and non-invertible functions. Then it is possible to distinguish the Markov equivalent DAG $X \to Y$ from $Y \to X$ using an L_0-based score–e.g. BIC or AIC. Since our score in the limit behaves like an L_0-based score (see Theorem 1), we can distinguish between Markov equivalent DAGs.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

B. Psuedocode

Algorithm 1: The GLOBE Algorithm

```
  Data: Data $X^n$ over $X$
  Result: Causal DAG $G$
1 $Q \leftarrow$ EDGESCORING($X^n$)
2 $G \leftarrow$ FORWARDSEARCH($Q, X^n$)
3 $G \leftarrow$ BACKWARDSEARCH($G$)
4 return $G$
```

Algorithm 2: Edge Scoring

```
  Data: n samples over $X$
  Result: priority queue of edges $Q$
1 $Q \leftarrow \emptyset$
2 foreach pair $(u, v) \in X$ do
3     $\psi \leftarrow \delta(e_{uv}) - \delta(e_{vu})$
4     $Q \leftarrow Q \oplus (e_{uv}, \psi)$
5     $Q \leftarrow Q \oplus (e_{vu}, -\psi)$
6 return $Q$
```

Algorithm 3: Forward Search

```
   Data: priority queue of edges $Q$, n samples over $X$
   Result: graph $G$
1  $E \leftarrow \emptyset$
2  $G \leftarrow (X, E)$
3  while $Q$ not empty do
4      $e_{uv} \leftarrow$ take top most entry from $Q$
5      if $E \oplus e_{uv}$ is not cyclic and $e_{uv}$ is significant
        then
6          $E \leftarrow E \oplus e_{uv}$
7          foreach incoming edge to $v$, $e_{xv} \in Q$ do
8              $\psi \leftarrow \delta(e_{xv}) - \delta(e_{vx})$
9              update the value of $e_{xv}$ in $Q$ to $\psi$
10             update the value of $e_{vx}$ in $Q$ to $-\psi$
11         foreach outgoing edge from $v$, $e_{vy} \in E$ do
12             $E' \leftarrow (E \ominus e_{vy}) \oplus e_{yv}$
13             $G' \leftarrow (X, E')$
14             if $Cost(G') < Cost(G)$ then
15                 $E \leftarrow E \ominus e_{vy}$
16                 $\psi \leftarrow \delta(e_{vy}) - \delta(e_{yv})$
17                 $Q \leftarrow Q \oplus (e_{vy}, \psi)$
18                 update the value of $e_{yv}$ in $Q$ to $-\psi$
19                 foreach
                    incoming edge to $y$, $e_{ky} \in Q$ do
20                     $\psi \leftarrow \delta(e_{ky}) - \delta(e_{yk})$
21                     update value of $e_{ky}$ in $Q$ to $\psi$
22                     update value of $e_{yk}$ in $Q$ to $-\psi$
23 return $G$
```

Algorithm 4: Backward Search

```
  Data: graph $G$
  Result: pruned graph $G$
1 foreach node $v \in G$ do
2     while node updated and $(|$Pa$(v)| \geq 2)$ do
3         $(p, c) \leftarrow ($Pa$(v)$ , Cost$(v \mid$ Pa$(v)$ $))$
4         foreach $p' \subset$ Pa$(v)$ $s.t$ $|p'| =$ $|$Pa$(x)| - 1$
           do
5             $c' \leftarrow Cost(v \mid p')$
6             if $c' < c$ then
7                 $(p, c) \leftarrow (p', c')$
8         Pa$(v) \leftarrow p$
9 return $G$
```

Edge Scoring The pseudocode for the edge scoring phase is given in Algorithm 2. For each of the possible node pairs (u, v) in our set of variables, X, we consider directed edges in both directions, e_{uv} and e_{vu} and calculate in line 3 the gain (δ) resp. score (ψ) as defined in the Algorithm section. The pairs (e_{uv}, ψ) and $(e_{vu}, -\psi)$ are then added to the priority queue.

Forward Search We give the pseudocode for forward search phase in Algorithm 3. We iteratively take the best edge from the top of the priority queue (line 4). The validity checks are carried out (line 5) before adding the edge to the graph (line 6). Next, we update the rankings of all the edges in Q which are incident on the same child node as the the currently added edge (lines 7-10). Last, we check for possible edge flips (lines 11-22). For each possible edge flip, we create another instance of our graph, G' with the reversed edge instead of the actual edge (lines 12-13) and compare the costs of both the graphs (line 14). If G' has a lower cost than G, we calculate the score for this edge and add it to the priority queue again (lines 15-18). Suppose the edge e_{vy} was removed, then we also need to update the cost of incoming edges e_{ky}, for y which are still present in Q. (lines 19-22).

Backward Search The pseudocode for the backward search is given in Algorithm 4. As input we get the graph, G, from the forward search phase. The following process is repeated for all of the nodes (line 1): denoting the number of parents for the node by k, we check the cost of storing the node given each of the $k - 1$ sized subsets of its parents (line 4). If such a subset is found, we update the cost of the node (lines 5-7) and we repeat the same procedure, using the newly found subset of parents (line 8). This is done until no subset of parents is better at encoding the current node than the current set of parents. The graph at the end of the backward search phase is the causal DAG predicted by GLOBE.

C. Experimental Setup

We execute all the experiments on an 4-core Intel i7 Laptop running Linux with 8GB of RAM. For all of the individual instances in each of the experiments GLOBE finished within ten minutes. The only exception was the 500 node network from the REGED dataset on which GLOBE took 3 days, whereas the competitor methods could not handle this data. GGES did not produce any result even after a month, while RESIT, LINGAM and PC$_{\text{HSIC}}$ crashed and could not handle this network.

Competitor Methods The code for the competitor methods were taken from the following sources.

- GGES: Implementation provided by the authors of the paper on Github (Huang et al. 2018).
- PC$_{\text{HSIC}}$ and LINGAM: Causal Discovery Toolbox (Kalainathan and Goudet 2019).
- RESIT: Source code downloaded from the homepage of the author (Peters et al. 2014).

Real world datasets All the data sets used in our real-world experiments are publicly available. In the following we provide the sources for our experimental data

- REGED: Causality Workbench (2014).
- SACHS: Bayesian Network (*bnlearn*) repository (Scutari 2009).
- REALESTATE: UCI Machine Learning Repository (Dua and Graff 2017).

Instantiation

We instantiate GLOBE using the open-source implementation in R of Multivariate Adaptive Regression Splines framework (Friedman 1991), called EARTH. For non-linear transformation function types described in our the Theory section, we use polynomials up to degree four, exponential functions as well as reciprocal functions up to degree two. Since we could face issues like multi-collinearity (Farrar and Glauber 1967) and unrealistic run times if we allow for arbitrary many interactions between parents, we restrict the maximum number of interaction terms to 2 for experiments. In theory, we could allow for arbitrary many interactions. In practice, we are limited by the maximum number of interactions supported by the implementation of EARTH, which was 10 in our case.

Evaluation Metrics

As our evaluation metric, we choose the *SHD* (Kalisch and Bühlmann 2007) to measure structural similarity. Measures such as the False Discovery Rate (FDR) (Benjamini and Hochberg 1995) or the Area under the ROC curve (Fawcett 2006) can also be used to evaluate GLOBE. However, they require us to rank our predictions in order of decreasing confidence. While this is not possible for both LINGAM and RESIT, both GGES and PC$_{\text{HSIC}}$ only provide us with p-values over undirected edges and hence it is not straightforward to compare the algorithms using such a metric. While alternate scores such as the F1 score (Sasaki et al. 2007) are also possible, our performance evaluation using the F1 score was found to follow a similar trend as the *SHD*. However, *SHD* provides for a more intuitive and easy-to-interpret measure when considered together with *SID*, our measure of *causal* similarity. Evaluating our results using both *SHD* as well as the F1 score only gives us somewhat redundant information about our performance.

We stress, while structural similarity between predicted and the ground truth graphs is important, it is not as important as measuring the causal similarity between the two. A single reversed edge between a cause and its effect could have drastic *causal* implications as shown by Peters and Bühlmann (2015). Therefore, any evaluation of causal graphs without assessing the quality of causal statements made by the predicted graph is incomplete and may result in wrong conclusions. Hence we evaluate all our methods using the *SID* (pre)metric. Using *SID* together with the *SHD* gives us a comprehensive picture of the performance of a causal discovery algorithm.

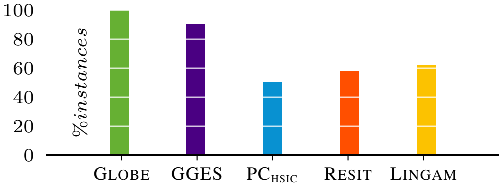


Figure 1: [Higher is better] Percentage of instances with no spurious edges reported for the independent data. GLOBE achieves a perfect score.

D. Extended Main Results

In this section we provide additional details about the experiments presented in the main text.

Independent Data

As a sanity check, we test the methods on instances of a graph containing 10 independent nodes where the value of each node is sampled independently from a Gaussian distribution. The ground truth for this network is an empty graph. We expect the methods to report empty set of edges for the instances in this experiment. GLOBE did not report a single spurious edge on *any* of the instances. Whereas, LINGAM reported at least one spurious edge for 38%, RESIT for 42% and PC$_{\text{HSIC}}$ and GGES for half resp. 10% of the instances. This shows that GLOBE is robust against false positives.

Structures used for Synthetic Data

In this section, we provide details on the graph structures used for our main experiments.

Effect of Multiple Parents We generate our data according the graph structure depicted in Figure 2. The collider node is calculated as a linear combination of non-linear par-

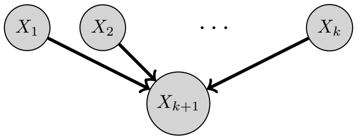


Figure 2: Common effect node with k parents—i.e. X_{k+1} is generated as a function over X_1 to X_k plus additive noise.

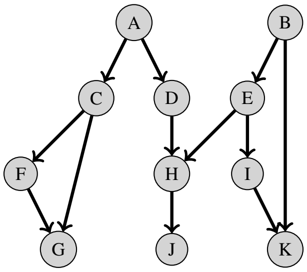


Figure 3: DAG used to generate non-linear data. This graph contains all connections possible in a DAG: a collider ($D \rightarrow H \leftarrow E$), a fork ($C \leftarrow A \rightarrow D$), a chain ($A \rightarrow C \rightarrow F$) and a feed forward loop ($C \rightarrow F \rightarrow G$ and $C \rightarrow G$).

ent functions given as

$$X_j = \sum_{X_i \in \text{Pa}(X_j)} a_i \cdot (X_i + b_i)^{c_i} \quad . \tag{1}$$

The range of values are chosen such that we avoid taking roots of negative numbers. Since it is possible to identify a collider structure using conditional independence tests, we expect GGES and PC$_{\text{HSIC}}$ to discover a fully directed network.

Data Sampled from a Causal Network Data for these experiments was generated using the graph structure that is shown in Figure 3. In this setting, each child node, X_j can alternatively be calculated using more complex multiplicative interactions between the parents given by

$$X_j = a_j \cdot \prod_{X_i \in \text{Pa}(X_j)} {X_i}^{c_i} + b_j \quad . \tag{2}$$

We generate data where we choose between Eq. (1) and (2) with pre-specified probability of 0.7 and 0.3 respectively. Previously Ghanbari, Lasserre, and Vingron (2018) have used a similar graph structure as well. Our usage differs from the latter in the sense that we define a *different* causal function for each edge in each of our experimental instances. On the other hand, Ghanbari, Lasserre, and Vingron (2018) always generate their synthetic data using the *same* function for each of the experiments.

E. Additional Experiments

In this section, we provide results over two additional experiments. The first experiment was to test the methods over adversarial settings for which we used the SACHS network (Sachs et al. 2005) dataset which contains discretized interventional data. Secondly we perform a case-study on real-world data without known ground truth to show that GLOBE discovers meaningful causal relations from the data.

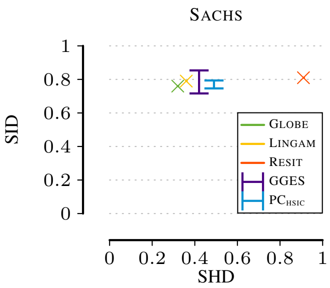
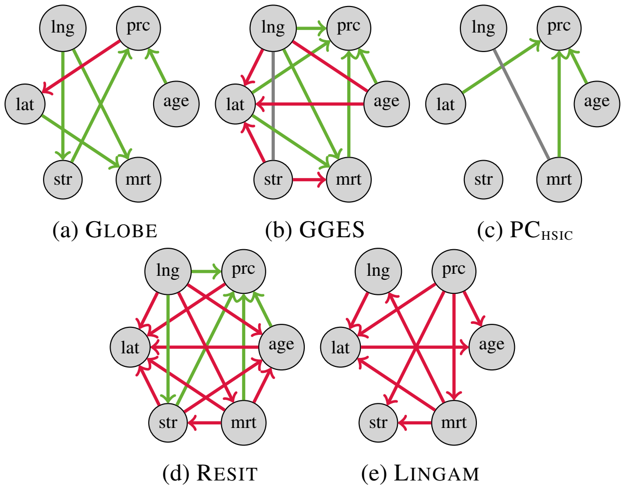


Figure 4: [Closer to Origin is Better] Normalized *SHD* and Normalized *SID* for real world SACHS data.



(a) GLOBE (b) GGES (c) PC$_{\text{HSIC}}$

(d) RESIT (e) LINGAM

Figure 5: Discovered DAGs on the real estate data set. Green edges are causal directions that agree with our domain knowledge, directed red edges are wrongly oriented causal dependency. Gray edges are associations that agree with domain knowledge but are left unoriented. Undirected red edges are associations that disagree with domain knowledge.

SACHS Protein Signalling Network

The SACHS network (Sachs et al. 2005) consists of interventional rather than observational data over discretize variables—both of which break our assumptions. The data consists of 5400 readings over 12 variables (Scutari 2009). We evaluate the methods on this data and show the results in Fig. 4. All methods struggle with the SACHS network and result in networks of relatively poor quality. Overall, GLOBE has a *SHD* of 0.32, lowest among all the competitors, and *SID* of 0.76 which is close to the best possible score over the equivalence class that GGES reports, $[0.71, 0.86]$.

Case Study: REALESTATE Data

To conclude our evaluation, we perform a case study on a real estate dataset (Yeh and Hsu 2018; Dua and Graff 2017) of market valuation of properties in the Sindian district, Taiwan. The data contains six continuous valued attributes: the age of the property (*age*), the distance to the nearest MRT station (*mrt*), the number of convenience stores reachable by foot from the house (*str*), the geographical coordinates (*lat*,*lng*) and the price of the property (*prc*). After additionally normalizing the data between zero and one, we run all the methods on this data and report the results in Figure 5. Overall, we see that the causal dependencies GLOBE discovers are in accordance with our domain knowledge: it finds that by changing the coordinates of the property (*lat*,*lng*), we can alter the distance to the train station (*mrt*) and that the latitude of the property determines the number of nearby stores. GLOBE also discovers three possible causal relations to the price of the property namely *age*, *str* and *lat*. However it wrongly orients the direction of price to latitude.

The other methods perform less well. We see that they either orient most edges against domain knowledge (RESIT, LINGAM), discover spurious causal relations (GGES), or report meaningful edges but only for a single variable (PC$_{\text{HSIC}}$).

F. GLOBE for Parametric Regression

In this section we will show how to instantiate GLOBE if we are given a set of pre-specified regression functions, $\mathcal{F}$, of size $|\mathcal{F}|$, over a simplified version of our causal model.

Causal Model

To model dependencies between nodes in our causal DAG G, consider a simplified causal model where we assign each node to be a linear combination of functions over its parents plus additional independent noise. We have $\forall X_i \in X$:

$$X_i := \sum_{X_j \in \text{Pa}(X_i)} f_j(X_j) + N_i \ , \qquad (3)$$

where f_j is a non-linear function over the $j-th$ parent of X_i and N_i is an independent noise term. We assume that all noise variables are jointly independent, Gaussian distributed and that N_i is independent of $\text{Pa}(X_i)$. Likewise, our model reduces only to the additive noise term, if we model a source node that has no parents.

Encoding the Model

The model complexity $L(M)$ for a model $M \in \mathcal{M}$, is defined exactly as in the theory section of the main text. It comprises of the parameters of the functional dependencies and the graph structure. The total cost is simply the sum of the code lengths of the individual nodes

$$L(M) = \sum_{i=1}^{m} L(M_i) \ .$$

To encode the individual nodes X_i, we need to transmit its parents, the form of the functional dependency, and the bias or mean shift μ_i. We encode the model M_i for a node X_i as

$$L(M_i) = L_{\mathbb{N}}(k) + k \log m + L_F(f_i) + L_p(\mu_i) \ ,$$

where we first encode the number of parents using $L_{\mathbb{N}}$, the MDL-optimal encoding for integers $z \geq 0$ (Rissanen 1983). It is defined as $L_{\mathbb{N}}(z) = \log^* z + \log c_0$, where $\log^* z = \log z + \log \log z + \ldots$ and we consider only the positive terms, and c_0 is a normalization constant to ensure the Krafft-inequality holds (Kraft 1949). Given the number of parents, we then identify which out of the m random variables these are, and then proceed to encode the function f_i over these parents, where f_i represents the summation term on the right hand side of Eq. (3). Last, we encode the bias term using L_p, from Eq. (4).

Encoding the Functions The parametric model follows our causal model in Eq. (3). one-to-one. That is, we can estimate X_i given its parents as:

$$\hat{X}_i := \sum_{X_j \in \text{Pa}(X_i)} f_j(X_j) \ ,$$

where f_j belongs to the function class $\mathcal{F}$ that we are already provided, e.g. the class of all polynomials up to a certain degree. To encode a function f_j , we specify the functional form of f_j and its parameters. We have:

$$L_F(f) = k \log |\mathcal{F}| + \sum_{j=1}^{k} L_p(\theta(f_j)) \ ,$$

where $\theta(f_j)$ is the parameter vector of function f_j. First, we use $\log |\mathcal{F}|$ bits to identify the regression function used for each of the k parents. Next we encode the parameters of each of the functions over those k parents.

Encoding Parameters To encode the bias (or mean shift) we use the proposal of Marx and Vreeken (2017) for encoding parameters up to a user specified precision p as as follows.

$$L_p(\theta) = |\theta| + \sum_{i=1}^{|\theta|} L_{\mathbb{N}}(s_i) + L_{\mathbb{N}}(\lceil \theta_i \cdot 10^{s_i} \rceil) \ , \qquad (4)$$

where s_i is the smallest integer such that $|\theta_i| \cdot 10^{s_i} \geq 10^p$. Simply put, $p = 2$ implies that we consider two digits of the parameter. We need one bit to store the sign of the parameter, then we encode the shift s_i and the shifted parameter θ_i.

Consistency

Comparing our current score for parametric functions to the non-parametric hinge functions that we actually use to initialize GLOBE, it is straightforward to see that $\forall X_i \in X$: $|H| \geq |\text{Pa}(X_i)|$. This is because each parent can be assigned to more than one hinge in the non-parametric formulation. Therefore, the costs for the non-parametric function will asymptotically be larger than for the parametric function under Assumption (2). This additionally implies that assumptions (1) holds trivially for this new formulation because $|\text{Pa}(X_i)| \leq |H| \in \mathcal{O}(\log n)$. Lastly, the likelihood term $c_1 n \log \hat{\sigma}^2$ defined below in Eq. (5) depends only on the encoding of our residuals and hence is not affected by our choice of the model.

Using the above results and following similar line of reasoning as we did for our original score, we arrive at

$$c_1 n \log \hat{\sigma}^2 + c_2 \log n + \mathcal{O}(1) \ . \qquad (5)$$

If we set $c_1 = 1$ and $c_2 = \frac{d}{2}$, where d is the number of degrees of freedom of the model, we arrive at the BIC score.

References

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57.

Causality Workbench, E. Z. 2014. Causality workbench.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *JMLR* 3(Nov):507–554.

Dua, D., and Graff, C. 2017. Uci machine learning repository.

Farrar, D. E., and Glauber, R. R. 1967. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*.

Fawcett, T. 2006. An introduction to roc analysis. *Pattern recognition letters* 27(8):861–874.

Friedman, J. H. 1991. Multivariate adaptive regression splines. *The annals of statistics* 1–67.

Ghanbari, M.; Lasserre, J.; and Vingron, M. 2018. The distance precision matrix: computing networks from nonlinear relationships. *Bioinformatics* 35(6):1009–1017.

Haughton, D. M. 1988. On the choice of a model to fit data from an exponential family. *Annals Math. Stat.* 16(1):342–355.

Huang, B.; Zhang, K.; Lin, Y.; Schölkopf, B.; and Glymour, C. 2018. Generalized score functions for causal discovery.

Kalainathan, D., and Goudet, O. 2019. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*.

Kalisch, M., and Bühlmann, P. 2007. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *JMLR* 8(Mar):613–636.

Kraft, L. G. 1949. *A device for quantizing, grouping, and coding amplitude-modulated pulses*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Marx, A., and Vreeken, J. 2017. Telling Cause from Effect using MDL-based Local and Global Regression. In *ICDM*, 307–316. IEEE.

Marx, A., and Vreeken, J. 2019. Identifiability of cause and effect using regularized regression. In *KDD*. ACM.

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

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

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal discovery with continuous additive noise models.

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

Rissanen, J. 1983. A universal prior for integers and estimation by minimum description length. *Annals Stat.* 11(2):416–431.

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.

Sasaki et al., Y. 2007. The truth of the f-measure. 2007.

Scutari, M. 2009. Learning bayesian networks with the bnlearn r package. *arXiv preprint arXiv:0908.3817*.

Yeh, I.-C., and Hsu, T.-K. 2018. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing* 65:260–271.