Causal Discovery with Score Matching on Additive Models and Generic Noise

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Causal discovery methods are intrinsically constrained by the set of assumptions needed to ensure structure identifiability. Moreover additional restrictions are often imposed in order to simplify the inference task: this is the case for the Gaussian noise assumption on additive non-linear models, which is common to many causal discovery approaches. In this paper we show the shortcomings of inference under this hypothesis, analyzing the risk of edge inversion under violation of Gaussianity of the noise terms. Then, we propose a novel method for inferring the topological ordering of the variables in the causal graph, from data generated according to an additive non-linear model with a generic noise distribution. This leads to NoGAM (Not only Gaussian Additive noise Models), a causal discovery algorithm with a minimal set of assumptions and state of the art performance, experimentally benchmarked on synthetic data.

Keywords: Causal discovery; Generic noise distribution; Score matching

1. Introduction

Inferring cause-effect relationships from observational data is a central goal of causality research, as it enables formal reasoning about interventions on a system (Peters et al. (2017), Pearl (2009)) when these are expensive, unethical, or even impossible to perform. Structure identifiability results posit limits to what part of the causal graph can be inferred from pure observations from the joint distribution, and provides formal guidelines on which assumptions are needed to fully identify the causal graph underlying the data. Traditional causal discovery methods usually are limited to identify Markov equivalence classes (Glymour et al. (2019)) (directed acyclic graphs representing the same set of conditional independencies), which is the case for PC, FCI (Spirtes et al. (2000)) and GES (Chickering (2002)). More recently, methods based on properly defined Structural Causal Models (SCMs) have been proposed to distinguish the correct graph underlying the observed data, by mean of additional assumptions on the functional class of the SCM: Hoyer et al. (2009) and Zhang and Hyvärinen (2009) show that in case of non-linear additive noise models *almost any* non linearity typically yields an identifiable setting. This is the case for SCORE (Rolland et al. (2022)) and CAM (Bühlmann et al. (2014)), both of which output a unique and asymptotically consistent graph as result of the inference process. Other methods such as Lachapelle et al. (2020) and Zheng et al. (2018) formulate a continuous optimization problem which results in a unique directed acyclic graph (DAG).

In general, identifiability results require assumptions in order to infer the causal structure from observational data with theoretical guarantees. The shortcoming of this approach is that constraints in the form of assumptions reduce the scope of applicability of an algorithm. Instead it would be desirable to have methods working on a broad range of problems under different conditions, ideally showing a certain degree of robustness regarding violations of the model hypothesis. The

© 2023 .

strength of this viewpoint is manifest in deep learning practice, where the dominant approach is to apply algorithms that work on the task of interest, independently of the violation of the underlying assumptions. The motivation behind this paper is to provide a causal discovery tool in between these philosophies, by removing (from the identifiability perspective) unnecessary assumptions frequently made by some of the most prominent computational methods available. With this goal in mind, we design an algorithm for the inference of the causal graph underlying an additive non-linear model with generic noise terms, removing the common hypothesis of Gaussian distributions. This constraint removal broadens the scope of applicability of principled causal discovery, providing a state of the art method to practitioners interested in theoretical guarantees and operating in critical settings where the validity of the Gaussian noise assumption is hard to verify.

The rest of the paper is organized as follow: Section 2 provides an overview of the model under study, and a definition of the problem at hand; Section 3 analyzes the risk of inferring inversed edges under violation of the Gaussian noise assumption; Section 4 introduces a theoretically principled method to find the topological ordering of a causal graph by iteratively identifying its leaf nodes; Section 5 defines NoGAM, an algorithm for inference of the causal graph from the data; Section 6 is an overview of the experimental performance of such method with respect to classical and state of the art benchmarks.

2. Background knowledge

Model definition Let $\mathbf{V} = \{1, \dots, d\}$ be the vertices of a directed acyclic graph $\mathcal{G}$, and $\mathbf{X} \in \mathbb{R}^d$ be a set of random variables generated according to the Structural Causal Model (SCM)

$$X_i := f_i(\mathrm{PA}_i(\mathbf{X})) + N_i, \quad \forall\, i \in \mathbf{V}\,, \tag{1}$$

where $\mathrm{PA}_i(\mathbf{X})$ is the vector of parents of node i in the graph $\mathcal{G}$. We assume causal mechanisms f_i to be non-linear and continuously differentiable functions. Noise terms $N_i \in \mathbb{R}$ are continuous random variables with density $p_i(N_i)$, mean $\mu_i = 0$ and variance $\sigma_i^2 > 0$. We assume them to be jointly independent such that their joint distribution is $p_{\mathbf{N}}(\mathbf{N}) = \prod_i p_i(N_i)$. Under these assumptions the model (1) induces a joint distribution $p_{\mathbf{X}}(\mathbf{X})$ which is Markov with respect to $\mathcal{G}$, such that it admits the following factorization:

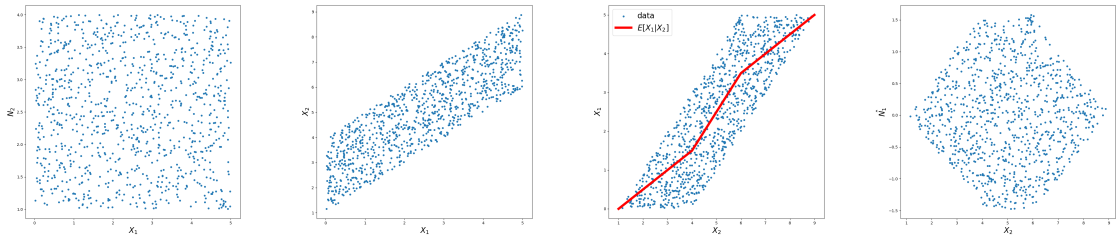
$$p_{\mathbf{X}}(\mathbf{X}) = \prod_i^d p_i(X_i \mid \mathrm{PA}_i(\mathbf{X}))\,, \tag{2}$$

where with an abuse of notation we distinguish the marginal $p_i(N_i)$ from $p_i(X_i \mid \mathrm{PA}_i(\mathbf{X}))$ by the argument.

In the reminder of this paper we will use X_i to denote both the random variable and the corresponding node $i \in \mathbf{V}$.

Topological ordering definition Let $\mathcal{G} = (\mathbf{X}, \mathcal{E})$ be a DAG. An ordering of the nodes $\mathbf{X}^\pi = X_{\pi_1}, \dots, X_{\pi_d}$ is a topological ordering relative to $\mathcal{G}$ if, whenever we have $X_{\pi_i} \to X_{\pi_j} \in \mathcal{E}$, then $i < j$ (Koller and Friedman (2009)).

Problem definition Given an i.i.d. sample from the joint distribution $p_{\mathbf{X}}$, we want to infer the true causal graph $\mathcal{G}$ underlying the model that generated the data. One approach to solve this problem is to divide the task in two steps: first we find a topological ordering for the vertices in the graph,



(a) Scatter plot of X_1 and ϵ_2

(b) Scatter plot of X_1 and X_2

(c) Scatter plot of X_2 and X_1 and regression curve

(d) Scatter plot of X_2 and estimated ϵ_1

Figure 1: Example to illustrate the limitations of the Gaussian noise assumption on the causal graph $X_1 \to X_2$.

and then we prune the fully connected graph obtained drawing an edge from each node to all its successors in the ordering. In this paper, we use the score function $\nabla \log p_{\mathbf{X}}(\mathbf{X})$ to propose a consistent method of inference of the topological ordering working without assuming any specific distribution of the noise terms N_i in Equation (1).

In practice, classical and state of the art causal discovery algorithms like SCORE (Rolland et al. (2022)), CAM (Bühlmann et al. (2014)), GES (Chickering (2002)) and GraN-DAG (Lachapelle et al. (2020)) assume the noise terms to be normally distributed. In the next section we show the limitations of this assumption, and how the topological ordering can be wrongly inferred when it is violated, leading to the estimation of a graph with inverted edges.

3. Limitations of the Gaussian noise assumption

It has been shown that for data generated by nonlinear models and additive noise, generally speaking, the causal direction between two variables is identifiable because in the reverse direction, one cannot find independent residual (Hoyer et al. (2009); Zhang and Hyvärinen (2009)), and hence the data likelihood given by the regression model (which assumes independent residuals) in the reverse direction is lower than that in the causal direction. The identifiability results (Zhang and Hyvärinen (2009)) imply that if the noise term in the causal model is Gaussian while the function is nonlinear, causal direction between two variables is identifiable. That is, the likelihood of the regression model in the correct direction is higher than that in the reverse direction, or equivalently, the total entropy of the estimated noise terms (including the hypothetical cause variable) is smaller in the causal direction (Zhang and Hyvärinen (2009)).

However, this result does not imply that when the noise distribution is assumed to be Gaussian, the correct causal direction can always give a higher likelihood or lower total entropy of the estimated noise terms.

Proposition 1 *Given a ground truth identifiable non-linear additive causal model, any inference algorithm based on observational data that wrongly assumes Gaussianity of the noise terms is not guaranteed to recover the correct direction of the edges in the underlying graph.*

In the reminder of the paper we consider Proposition 1 to be true, and justify the veridicity of the claim with an example.

Example 1 *We start defining an example for a linear additive noise model, such that closed form solutions for the regression problems at hand can be found. Then, we generalize the example to the non linear case.*

Let $X_1 \to X_2$ with $X_2 := X_1 + N_2$, where X_1 and N_2 are both uniformly distributed, as shown in Figure 1a. Figure 1b shows the scatter plot of X_1 and X_2. Figure 1c gives the scatter plot of X_2 and X_1, together with the regression curve, which, in this example, is piecewise linear. Figure 1d shows the scatter plot of X_2 and $\hat{N}_1$, the estimated regression noise in the reverse direction.
One can find the correct causal direction by comparing the total entropy of the estimated noise terms (note that the entropy of a variable is equivalent to the negative likelihood with the distribution estimated from data). Let $H(\cdot)$ denote the differential entropy. One can calculate that

$$H(X_1) + H(N_2) = 2.708 < H(X_2) + H(\hat{N}_1) = 2.954.$$

Let $H_{\mathcal{G}}(\cdot)$ denote the differential entropy under the Gaussianity assumption. One can then calculate that

$$H_{\mathcal{G}}(X_1) + H_{\mathcal{G}}(N_2) = 3.061 > H_{\mathcal{G}}(X_2) + H_{\mathcal{G}}(\hat{N}_1) = 3.043.$$

That is, under the Gaussianity assumption, the reverse direction gives a lower total entropy (or equivalently, a higher likelihood), and hence a wrong causal direction is inferred.
Why does it happen? Notice that for a variable with a fixed variance, the Gaussian distribution gives the highest differential entropy. So naturally, $H_{\mathcal{G}}(X_1) + H_{\mathcal{G}}(N_2) > H(X_1) + H(N_2)$ and $H_{\mathcal{G}}(X_2) + H_{\mathcal{G}}(\hat{N}_1) > H(X_2) + H(\hat{N}_1)$. Furthermore, it is totally possible (which is clearly the case in this example) that compared to the original independent variables X_1 and N_2, X_2 and $\hat{N}_1$ are closer to Gaussian, respectively. That is, the change in the total entropy of the estimated noise terms is smaller in the reverse direction. As a consequence, under the Gaussianity assumption, the reverse direction may give lower total entropy of the estimated noise terms, in contrast to the case using their true distributions.

Now consider the non linear model $X_2 := X_1^{1+\delta} + N_2$, with $\delta > 0$ and noise terms $X_1 := N_1$ and N_2 uniformly distributed. Clearly, for $\delta = 0$ this is equivalent to the linear example already discussed: here we set $\delta = 0.1$ to introduce a weak non-linearity in the generative process. Again, the ground truth causal direction can be identified by comparing the total entropy of the estimated noise terms in the correct and reversed direction. One can calculate that

$$H(X_1) + H(N_2) = 2.708 < H(X_2) + H(\hat{N}_1) = 2.926\,.$$

Similarly, under Gaussianity assumption one obtains

$$H_{\mathcal{G}}(X_1) + H_{\mathcal{G}}(N_2) = 3.061 > H_{\mathcal{G}}(X_2) + H_{\mathcal{G}}(\hat{N}_1) = 3.001\,,$$

showing that the inversion statement of Proposition 1 holds.

Given the shortcomings of the Gaussian assumption, we now propose a causal discovery method on additive non-linear models with generic noise terms, such that the inferred causal ordering (and hence, the edges direction) is guaranteed to be correct with respect to the causal graph.

4. Causal discovery via the score function

In this section we derive a principled approach to identify leaf nodes from the score function $s(\mathbf{X}) = \nabla \log p_{\mathbf{X}}(\mathbf{X})$, without assuming any distribution of the noise random variables in the generative model (1) of $\mathbf{X}$.

4.1. Score function of a data distribution

Given the distribution $p_{\mathbf{X}}(\mathbf{X})$ induced by model (1), we can define the vector of the score function as $s(\mathbf{X}) = \nabla \log p_{\mathbf{X}}(\mathbf{X})$. Exploiting the factorization of the joint distribution of Equation (2), we can write:

$$\log p_{\mathbf{X}}(\mathbf{X}) = \sum_{i=1}^{d} \log p_i(X_i \mid \mathrm{PA}_i) \,, \tag{3}$$

such that a single entry $s_i(\mathbf{X})$ of the score is equal to:

$$s_i(\mathbf{X}) = \partial_i \log p_i(X_i \mid \mathrm{PA}_i) + \sum_{k \in \mathrm{CH}_i} \partial_i f_k(\mathrm{PA}_k) \partial_i \log p_k(X_k \mid \mathrm{PA}_k) \,. \tag{4}$$

Under parents conditioning, the marginal of X_i is the same as the distribution of N_i shifted by the value of the mechanism $f_i(\mathrm{PA}_i)$, meaning that $p_i(X_i \mid \mathrm{PA}_i)$ can be replaced by $p_i(N_i = X_i - f_i(\mathrm{PA}_i) \mid \mathrm{PA}_i)$. This allows to rewrite the score as:

$$s_i(\mathbf{X}) = \partial_i \log p_i(N_i) + \sum_{k \in \mathrm{CH}_i} \partial_i f_k(\mathrm{PA}_k) \partial_i \log p_k(N_k) \,. \tag{5}$$

Then, for each $i = 1, \dots, d$ we define a function $g_i(N_i) \coloneqq \log p_i(N_i)$, such that the i-th score entry eventually becomes

$$s_i(\mathbf{X}) = \partial_i g_i(N_i) + \sum_{k \in \mathrm{CH}_i} \partial_i f_k(\mathrm{PA}_k) \partial_i g_k(N_k) \,. \tag{6}$$

For a leaf node X_l, Equation (6) of the score becomes

$$s_l(\mathbf{X}) = \partial_l g_l(N_l) \,, \tag{7}$$

telling us that, if g_l were known, we could predict the score of a leaf $s_l(\mathbf{X})$ from the noise N_l. In the next section, with this idea in mind, we define a regression problem for each variable X_i, where the input variables are all the remaining entries $\mathbf{X}_{\setminus\{i\}} \coloneqq \mathbf{X} \setminus \{X_i\}$: then we show that if the predicted variable is a leaf, the residuals of this learning problem are consistent estimators of the noise term in the corresponding structural equation of model (1).

4.2. Residuals estimation

Given an i.i.d. sample $X \in \mathbb{R}^{n \times d}$ from $p_{\mathbf{X}}$, for each $i = 1, \dots, d$ we define a regression problem predicting X_i from the remaining variables $\mathbf{X}_{\setminus\{i\}}$:

$$\begin{aligned} &\min_{q \in \mathcal{Q}} L(q), \quad L(q) = \int_{\mathbb{R}^d} (q(\mathbf{X}_{\setminus\{i\}}) - X_i)^2 dp_{\mathbf{X}}(\mathbf{X}) \\ &\text{given } \mathcal{D} = \left\{ \left(\mathbf{X}^k_{\setminus\{i\}}, X^k_i \right) \right\}_{k=1}^{n} \end{aligned} \,, \tag{8}$$

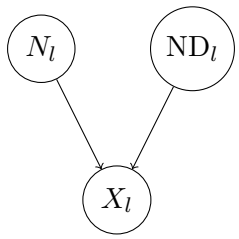


Figure 2: Consider a leaf node X_l generated according to the SCM defined in (1): by *d-separation* the noise term N_l is unconditionally independent from the set of non-descendants ND_l, as the only path between the two nodes contains a collider, namely X_l.

where $\mathcal{Q}$ is the space of measurable functions from the input space to $\mathbb{R}$. For all $\mathbf{X}_{\setminus\{i\}}$ in the input space, the minimizer $q^* \in \mathcal{Q}$ of L is the conditional expectation $\mathbf{E}[X_i \mid \mathbf{X}_{\setminus\{i\}}]$, which by linearity of the mean is equivalent to:

$$q^*(\mathbf{X}_{\setminus\{i\}}) = \mathbf{E}[f_i(\mathrm{PA}_i) \mid \mathbf{X}_{\setminus\{i\}}] + \mathbf{E}[N_i \mid \mathbf{X}_{\setminus\{i\}}] \,. \tag{9}$$

Given that $\mathrm{PA}_i \subset \mathbf{X}_{\setminus\{i\}}$, we can simply remove the expectation operator from the first term of the sum, obtaining:

$$q^*(\mathbf{X}_{\setminus\{i\}}) = f_i(\mathrm{PA}_i) + \mathbf{E}[N_i \mid \mathbf{X}_{\setminus\{i\}}] \,. \tag{10}$$

Now we define the residual of the learning problem in (8) as the difference between the response and the target function:

$$\begin{aligned} R_i &\coloneqq X_i - q^*(\mathbf{X}_{\setminus\{i\}}) \\ &= N_i - \mathbf{E}[N_i \mid \mathbf{X}_{\setminus\{i\}}] \,. \end{aligned} \tag{11}$$

We can further manipulate the residual expression by noticing that $\mathbf{X}_{\setminus\{i\}} = \mathrm{DE}_i \cup \mathrm{ND}_i$, with DE_i and ND_i respectively the set of descendants and non-descendants of a node X_i: Equation (11) then becomes

$$R_i = N_i - \mathbf{E}\left[N_i \mid \mathrm{DE}_i \cup \mathrm{ND}_i\right] \,. \tag{12}$$

For a leaf node X_l we can exploit the fact that $\mathrm{DE}_l = \emptyset$ in order to simplify the above expression of the residual in $R_l = N_l - \mathbf{E}\left[N_l \mid \mathrm{ND}_l\right]$. Additionally, we can use d-separation criterion to conclude that N_l is unconditionally independent of ND_l, as shown in Figure 2: the expectation on the error term of a leaf is then $\mathbf{E}\left[N_l \mid \mathrm{ND}_l\right] = \mathbf{E}\left[N_l\right]$. Finally, under the assumption of zero mean of the noise of model (1), the residual of Equation (12) is

$$R_l = N_l \,. \tag{13}$$

Exploiting this equivalence we can rewrite the score of a leaf in Equation (7) as a function of R_l, such that the score entry s_l satisfies:

$$\boxed{s_l(\mathbf{X}) = \partial_l g_l(R_l)} \,. \tag{14}$$

Now we want to use this result in order to show that it is possible to derive a consistent estimator of the score of a leaf from the associated residual.

By substituting the residual of Equation (11) in Equation (6), we can derive an analogous expression of the score entry s_i associated to a non leaf node X_i:

$$s_i(\mathbf{X}) = \partial_i g_i(R_i + \mathbf{E}\left[N_i \mid \mathbf{X}_{\setminus\{i\}}\right]) + \sum_{k \in \mathrm{CH}_i} \partial_i f_k(\mathrm{PA}_k) \partial_i g_k(R_k + \mathbf{E}\left[N_k \mid \mathbf{X}_{\setminus\{k\}}\right])) \,. \tag{15}$$

Discussion For a leaf node X_l, Equation (14) shows that the score is a function of a single variable, namely the residual R_l: this suggests that we can hope to consistently estimate the score $s_l(\mathbf{X})$ from such residual. On the other hand, if we consider a non-leaf node X_i, the form associated to its score $s_i(\mathbf{X})$ is more complicated (i.e. depending on a larger number of variables), as shown by Equation (15): intuitively we can see that R_i, as a predictor, is not sufficient to find a consistent estimator of $s_i(\mathbf{X})$. In the next section, we want to formalize these intuitions that will allow us to derive a theoretically principled method to identify leaf nodes by looking at the error of score entries predictions from their corresponding residual.

4.3. Identifying leaf nodes

Consider a leaf node X_l: given a set of i.i.d. observations $\left\{(R_l^k, s_l(\mathbf{X}^k))\right\}_{k=1}^n$, we want to find an estimator of the score using the residuals as input. Similarly to (8) we define a regression problem

$$\min_{h \in \mathcal{H}} L(h), \quad L(h) = \int_{\mathbb{R} \times \mathbb{R}^d} (h(R_l) - s_l(\mathbf{X}))^2 dp(R_l, \mathbf{X}) \,, \tag{16}$$

with $\mathcal{H}$ set of measurable functions from input to output space. For every $R_l \in \mathbb{R}$, the target function minimizing the expected risk is

$$\begin{aligned} h^*(R_l) &:= \mathbf{E}\left[s_l(\mathbf{X}) \mid R_l\right] \\ &= \mathbf{E}\left[\partial_l g_l(R_l) \mid R_l\right] \,. \end{aligned} \tag{17}$$

It is immediate to see that, since we are conditioning on R_l, we can remove the expectation operator, obtaining:

$$h^*(R_l) = \partial_l g_l(R_l), \; \forall R_l \in \mathbb{R} \,. \tag{18}$$

Then, given a sample $(R_l, s_l(\mathbf{X}))$, the difference between the prediction $h^*(R_l)$ and the ground truth $s_l(\mathbf{X}) = g_l(R_l)$ is

$$\boxed{h^*(R_l) - s_l(\mathbf{X}) = 0} \,. \tag{19}$$

Similarly, the regression problem of (16) can be defined for a non-leaf node X_i: the resulting regression function is $h^*(R_i) = \mathbf{E}\left[s_i(\mathbf{X}) \mid R_i\right]$, the conditional expectation of Equation (15). Now, we can exploit these results to define a criterion for identification of leaf nodes.

Lemma 1 *Let* $\mathbf{X}$ *be a random vector generated according to model* (1)*, and let* $X_j \in \mathbf{X}$*. Then*

$$\mathbf{E}\left[(h^*(R_j) - s_j(\mathbf{X}))^2\right] = 0 \Leftrightarrow X_j \text{ is a leaf.}$$

Discussion With Equation (19), we show that we can find a consistent estimator h^* that can exactly predict the score function associated to a leaf X_l, given that we observe the residual R_l. In general, this is not the case for a node X_i that is not a leaf. This intuition is formalized in Lemma 1, by considering the mean of the squared error of the prediction over all input and output realizations. The proof of the lemma can be found in Appendix A.

Based on the results of this section, we now introduce an algorithm for causal discovery that runs on non-linear additive models with generic distributions of the noise terms. Then, we compare its experimental performance against several existing methods on synthetic data.

5. Method

In Section 4.3 we show how to identify leaf nodes in a causal graph underlying observations generated according to model (1), consistently with the number of samples: the idea is that, given a set of observations, for each node $i = 1, \dots, d$ we can predict the score $s_i(\mathbf{X})$ from the corresponding residual R_i, choosing as leaf the node l index of the entry where the generalization error is minimized. Once a leaf is identified, it is removed from the graph, and the procedure is repeated iteratively up to the source node, allowing to infer a topological ordering $\mathbf{X}^\pi$ that is asymptotically consistent. In practice, given a finite set of observations $X \in \mathbb{R}^{n \times d}$, first we estimate $\nabla \log p(\mathbf{X})$ score function of the data using the Stein gradient estimator from SCORE (Rolland et al. (2022)), whose output is the vector $\hat{\mathbf{s}}(\mathbf{X})$. Then, we estimate the residuals $\hat{\mathbf{R}} = (\hat{R}_1, \dots, \hat{R}_d)$ by Kernel Ridge regression, solving the problem defined in (8). Next, we define the vector estimator $\tilde{\mathbf{s}}(\hat{\mathbf{R}})$ predicting $\hat{s}_i(\mathbf{X})$ from $\hat{R}_i$, for each variable $i = 1, \dots, d$: in order to avoid overfitting, we train K different models by K-fold cross validation (i.e. only on a subset of the observations), each one predicting on its corresponding test set, unseen during the training. Finally we compute the Mean Squared Error (MSE) between the predictions $\tilde{\mathbf{s}}(\hat{\mathbf{R}})$ and the ground truth $\hat{\mathbf{s}}(\mathbf{X})$ provided by SCORE's output: we select as leaf the node corresponding to the *argmin* of the vector of MSEs. This procedure is repeated such that at each iteration one leaf is identified and added to the topological ordering estimate. More details on the implementation of the algorithm described can be found in the box of Algorithm 1.

Given the order estimated by Algorithm 1, we use a pruning method, namely the pruning procedure of CAM (*CAM-pruning*), to remove superfluous edges from the fully connected graph admitted by the ordering.

6. Experiments

In this section we empirically study the performance of NoGAM (Algorithm 1). The experimental analysis is done on data synthetically generated using the Erdos and Renyi (1960) (ER) model. Mimicking the setting of Rolland et al. (2022), Lachapelle et al. (2020) and Zhu et al. (2020), we generate the mechanims f_i by sampling Gaussian processes with a unit bandwidth RBF kernel. Experiments are repeated with the number of nodes d equals 10 and 20, and expected number of edges equals to d (ER1) and $4d$ (ER4), to simulate inference on sparser and denser graphs. The size of the datasets is $N = 1000$. Performance is tested on datasets generated with noise terms under one of the following distributions: Beta, Exponential, Gamma, Gumbel, Laplace and Normal. Comparing the performance of NoGAM with state of the art methods working under gaussianity

Algorithm 1: NoGAM causal discovery

Input: data matrix $X \in \mathbb{R}^{n \times d}$

$X^{\pi} \leftarrow [\,]$

$nodes \leftarrow [1, \ldots, d]$

for $i = 1, \ldots, d$ **do**

  $\hat{\mathbf{s}} \leftarrow \text{SCORE}(X)$

  $\hat{\mathbf{R}} \leftarrow \{\hat{X}_j\}_{j=1}$ estimate from $\mathbf{X}_{\setminus\{i\}}$

  $\tilde{\mathbf{s}} \leftarrow \{\tilde{s}_j\}_{j=1}$ estimate from $\hat{R}_i$

  $MSE \leftarrow \text{Avg}\,[\hat{\mathbf{s}} - \tilde{\mathbf{s}}]^2$

  $l_{index} \leftarrow \text{argmin}\, MSE$

  $l \leftarrow nodes[l_{index}]$

  $X^{\pi} \leftarrow [l, X^{\pi}]$

  Remove l_{index}-th column from X; Remove l from $nodes$

end

$X^{\pi} \leftarrow reverse(X^{\pi})$ (first node is source, last node is leaf)

$\hat{\mathcal{G}} \leftarrow$ CAM-pruning(X^{π}) (CAM-pruning: pruning method of CAM algorithm)

return $\hat{\mathcal{G}}$

assumption, we are able to provide empirical evidence of the robustness of our algorithm with respect to changes in the distributions.

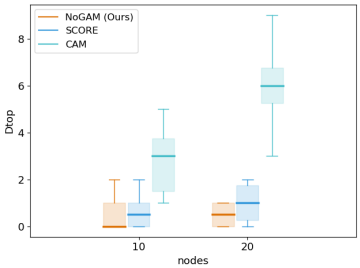
The residuals $\hat{\mathbf{R}}$ and the the score function entries $\tilde{\mathbf{s}}(\hat{\mathbf{R}})$ are estimated by Kernel Ridge regression, using the `scikit-learn` (Pedregosa et al. (2011)) implementation of the algorithm, with hyperparameters $\alpha = 0.01$, $\gamma = 0.1$: these values are tuned minimizing the generalization error on the estimated residuals, without using the performance on the causal graph ground truth. For the CAM pruning step the cutoff threshold is set at 0.001.

The metrics used to assess the quality of the inferred graph are the Structural Hamming Distance (SHD), which is the sum of false positive, false negative and reversed edges, and Structural Interventional Distance (SID, Peters and Bühlmann (2015)), accounting for the number of miscalculated interventional distributions from the inferred graph. We separately evaluate the quality of the topological ordering estimation as follow: given an ordering $\hat{\pi}$ and the ground truth adjacency matrix A, we use the topological ordering divergence defined in SCORE (Rolland et al. (2022)):

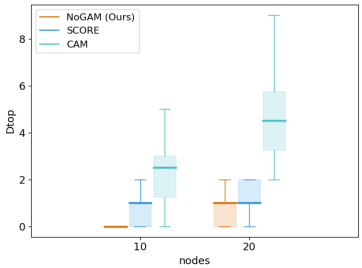
$$D_{top}(\hat{\pi}, A) = \sum_{i=1}^{d} \sum_{j:\hat{\pi}_i \succ \hat{\pi}_j} A_{ij}\,, \tag{20}$$

with $\hat{\pi}_i \succ \hat{\pi}_j$ meaning the node i is successive to j in the order, and $A_{ij} = 1$ if $X_i \in \text{PA}_j(\mathbf{X})$. In words this is the sum of the edges that can not be recovered due to the choice of the topological ordering. If $\hat{\pi}$ is correct with respect to A, then $D_{top}(\hat{\pi}, A) = 0$.

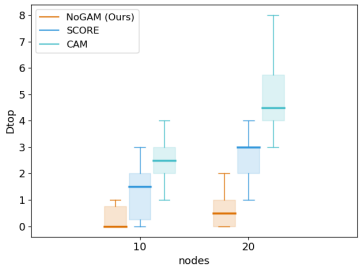
We compare the experimental performance of our algorithm against SCORE (Rolland et al. (2022)), CAM (Bühlmann et al. (2014)) and GES (Chickering (2002)), causal discovery methods working under the assumption of Gaussian noise. We exclude PC and FCI since in general they perform much worse (Bühlmann et al. (2014), Lachapelle et al. (2020)).



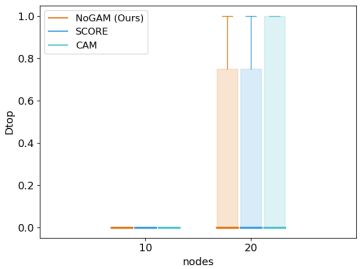
(a) Beta distribution



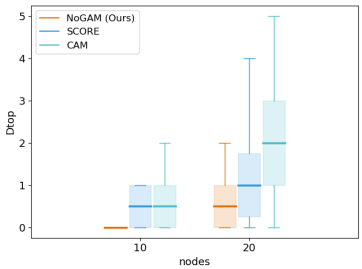
(b) Exponential distribution



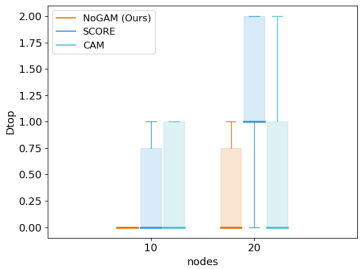
(c) Gamma distribution



(d) Normal distribution

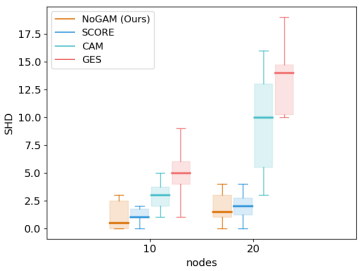


(e) Gumbel distribution

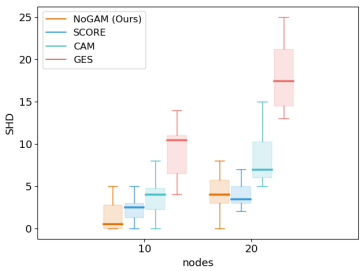


(f) Laplace distribution

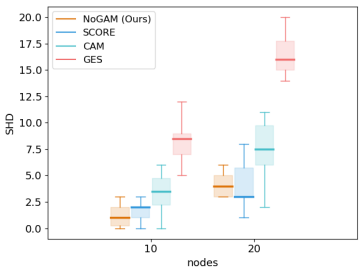
Figure 3: Boxplots over 10 runs showing topological ordering divergence D_{top} over sparse graphs ER1 (lower values correspond to better performance). GES algorithm does not appear since it does not require an explicit estimate of the topological ordering. Overall, NoGAM clearly outperforms CAM and SCORE.



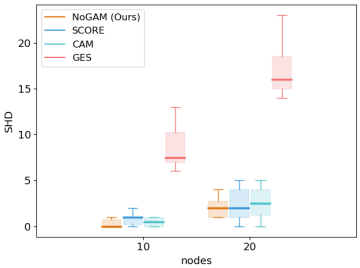
(a) Beta distribution



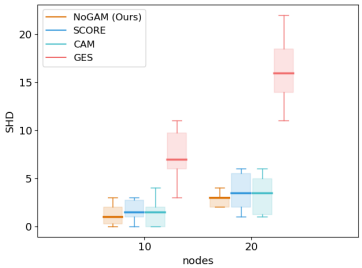
(b) Exponential distribution



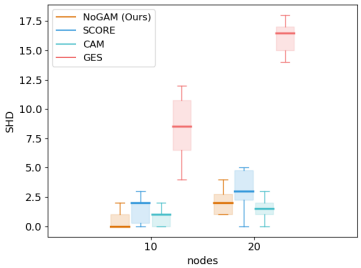
(c) Gamma distribution



(d) Normal distribution



(e) Gumbel distribution



(f) Laplace distribution

Figure 4: Boxplots over 10 runs showing SHD performance over sparse graphs ER1. NoGAM is in general better than all the other methods, with SCORE showing comparable performance.

Method	D_{top}	SHD	SID
NoGAM	8	12	45
SCORE	8	12	45
CAM	7	12	55
GES	–	17	62

Table 1: Experimental results on Sachs dataset (11 variables, 17 edges, 853 observables).

Figure 3 illustrate how in the sparse setting (ER1) topological ordering estimates of NoGAM systematically outperform the D_{top} results obtained by CAM and SCORE for every non-Gaussian noise distribution. The accuracy gap closes only for datasets generated under Gaussianity of the noise, coherently with our expectations. A similar performance gap in favor of our method is observed in the dense case (ER4), as shown in Figure 5 of the Appendix. Overall SCORE and NoGAM clearly show better SHD (Figure 4 and 6) with respect to the remaining methods. Note that the two algorithms differ only for the topological order inference step, while they share the same pruning method of CAM.
A complete overview of the experimental results is provided in the Appendix B.

Sachs real-world data We provide experimental results of the four algorithms under analysis on Sachs dataset (Sachs et al. (2005)), a common causal discovery benchmark made of real-world biological data. Results are provided in Table 1.

Additional experiments on Scale-free graphs (Barabasi and Albert (1999)) can be found in the Appendix C.

Discussion on SCORE robustness Despite being systematically outperformed by our method (Figure 3 and 5), SCORE algorithm shows significant robustness across different distributions of the noise terms. Here we want to provide a brief discussion on why this, in our opinion, is the case. In Lemma 1 of Rolland et al. (2022), authors propose to identify a leaf node from the Jacobian matrix of the score function $\mathbf{s}(\mathbf{X})$. In particular, computing the variance of the diagonal elements of the jacobian $\mathrm{Var}\left[\partial_i s_i(\mathbf{X})\right]$, $\forall i = 1, \ldots, d$, it can be shown the following: under gaussianity assumption, leaf nodes are associated to zero variance, such that $\mathrm{Var}\left[\partial_l s_l(\mathbf{X})\right] = 0 \iff \mathrm{CH}_l(\mathbf{X}) = \emptyset$, i.e. if and only if X_l is a leaf. In practice, due to statistical error introduced by the estimation of the score terms and of their variance, such expression in general is never exactly zero. To account for this, the SCORE algorithm iteratively selects leaf nodes as $l = \mathrm{argmin}_{i \in \{1,\ldots,d\}} \mathrm{Var}\left[\partial_i s_i(\mathbf{X})\right]$. We argue that such heuristic is key to determine the robustness of the algorithm as empirically observed outside of the Gaussian assumption. If the noise terms are not normally distributed, then the score function is equivalent to the one derived in Equation (4), namely:

$$s_i(\mathbf{X}) = \partial_i \log p_i(X_i \mid \mathrm{PA}_i) + \sum_{k \in \mathrm{CH}_i} \partial_i f_k(\mathrm{PA}_k) \partial_i \log p_k(X_k \mid \mathrm{PA}_k) \ .$$

It is clear that the variance of the partial derivative of such expression is proportional to the number of children in the summation term, as already pointed out in Lemma 2 of Rolland et al. (2022). In particular, the total variance of a diagonal element of the Jacobian of the score is the sum of the variances of the partial derivative of the two right-hand terms in Equation (4), plus their covariance:

	Method	Ordering time [s]	Total time [s]
d=10	NoGAM	6.98 ± 0.57	10.56 ± 0.84
	SCORE	4.28 ± 0.28	7.72 ± 01.07
	CAM	24.12 ± 1.58	27.98 ± 2.61
	GES	–	0.43 ± 0.11
d=20	NoGAM	25.75 ± 1.48	53.98 ± 2.01
	SCORE	9.51 ± 1.12	37.3 ± 4.21
	CAM	501.92 ± 11.92	529.61 ± 18.41
	GES	–	5.35 ± 1.67

Table 2: Experiments execution times. We report the average and standard deviation across 10 runs with fixed method and number of nodes. Noise terms distributions don't influence the execution time in our experimental setting, while significant differences under different sparsity conditions (ER1 and ER4) are observed only for GES, for which we report only ER1 running time in the table.

if the covariance happens to be negative and with magnitude large enough, the variance associated to a non-leaf node might be smaller than the one relative to all other nodes, leading to errors in the inferred topological ordering. Nevertheless, in general, this doesn't seem to be the case: despite the guarantee of vanishing variance for leaf nodes doesn't hold anymore, in practice the leaf selection criterion based on the argmin operator can still be expected to work.

6.1. Algorithmic complexity

We now provide an analysis of the algorithmic complexity of NoGAM topological ordering method. We denote with n the number of samples and d the number of nodes. Considering the implementation of Algorithm 1, each iteration of the for loop needs to solve a regression problem with $\mathcal{O}(n^3)$ cost, for each of the d residuals $\hat{R}_i$ estimated. The same analysis holds for the estimation of $\tilde{\mathbf{s}}$ from $\hat{\mathbf{R}}$. This provides an overall $\mathcal{O}(d^2 n^3)$ complexity. Similarly to NoGAM, SCORE iteratively identifies leaves one at the time: each iteration requires inverting the $n \times d$ dimensional matrix of the data, such that the topological ordering inference time scales with $\mathcal{O}(dn^3)$. An overview of the execution times of the experiments is provided in Table 2.

7. Conclusion

The assumption of Gaussian noise terms in an additive non-linear model, when violated, can lead causal discovery algorithms to infer graphs with inverted direction of the edges. In this work we prove such limitation, and in response to this problem we introduce NoGAM. Based on the interplay between score matching and causal discovery introduced by Rolland et al. (2022), our algorithm proposes a novel and consistent method of inference of the topological ordering that doesn't assume any distribution on the noise terms. We prove via systematic experiments that our approach outperforms traditional and state of the art causal discovery algorithms on almost any synthetic dataset generated with non-Gaussian distribution of the noise terms.

References

Albert-Laszlo Barabasi and Reka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. doi: 10.1126/science.286.5439.509. URL http://www.sciencemag.org/cgi/content/abstract/286/5439/509.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6), dec 2014. URL https://doi.org/10.1214%2F14-aos1260.

David Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 01 2002.

Paul Erdos and Alfred Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.*, 5:17–61, 1960.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00524. URL https://www.frontiersin.org/articles/10.3389/fgene.2019.00524.

Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklbKA4YDS.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jonas Peters and Peter Bühlmann. Structural Intervention Distance for Evaluating Causal Graphs. *Neural Computation*, 27(3):771–799, 03 2015. ISSN 0899-7667. doi: 10.1162/NECO_a_00708. URL https://doi.org/10.1162/NECO_a_00708.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russel, Bernhard Schölkopf, Dominik Janzing, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2203.04413.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005. URL https://www.science.org/doi/abs/10.1126/science.1105809.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

K. Zhang and A. Hyvärinen. Acyclic causality discovery with additive noise: An information-theoretical perspective. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2009*, Bled, Slovenia, 2009.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, page 647–655, Arlington, Virginia, USA, 2009. AUAI Press. ISBN 9780974903958.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g2skStPB.

Appendix A. Proof of Lemma 1

(i) X_j leaf node $\Rightarrow \mathbf{E}\left[(h^*(R_j) - s_j(\mathbf{X}))^2\right] = 0$: true by Equation (19).

(ii) $\mathbf{E}\left[(h^*(R_j) - s_j(\mathbf{X}))^2\right] = 0 \Rightarrow X_j$ leaf node: by contradiction, assume that X_j is not a leaf node, such that $h^*(R_j)$ is the expectation of Equation (15) conditional on R_j. The zero expectation can be rewritten as:

$$\int_{\mathbb{R}\times\mathbb{R}^d} (h^*(R_j) - s_j(\mathbf{X}))^2 \, dp(R_l, \mathbf{X}) = 0 \, .$$

Being the integral taken over a positive function, it is immediate that the equality with zero holds if and only if $h^*(R_j) = s_j(\mathbf{X})$ with probability 1, such that $\mathbb{P}(s_j(\mathbf{X}) = h^*(r_j) \mid R_j = r_j) = 1$, where r_j denotes a realized value of R_j. It follows that the conditional variance satisfies $\mathrm{Var}[s_j(\mathbf{X}) \mid R_j] = 0$. By inspection of Equation (15) we can see that this is a contradiction: being $s_j(\mathbf{X}) \mid R_j$ a function of random variables with strictly positive variance, then it must be that $\mathrm{Var}[s(\mathbf{X}) \mid R_j] > 0$.

Appendix B. Experiments on ER graphs

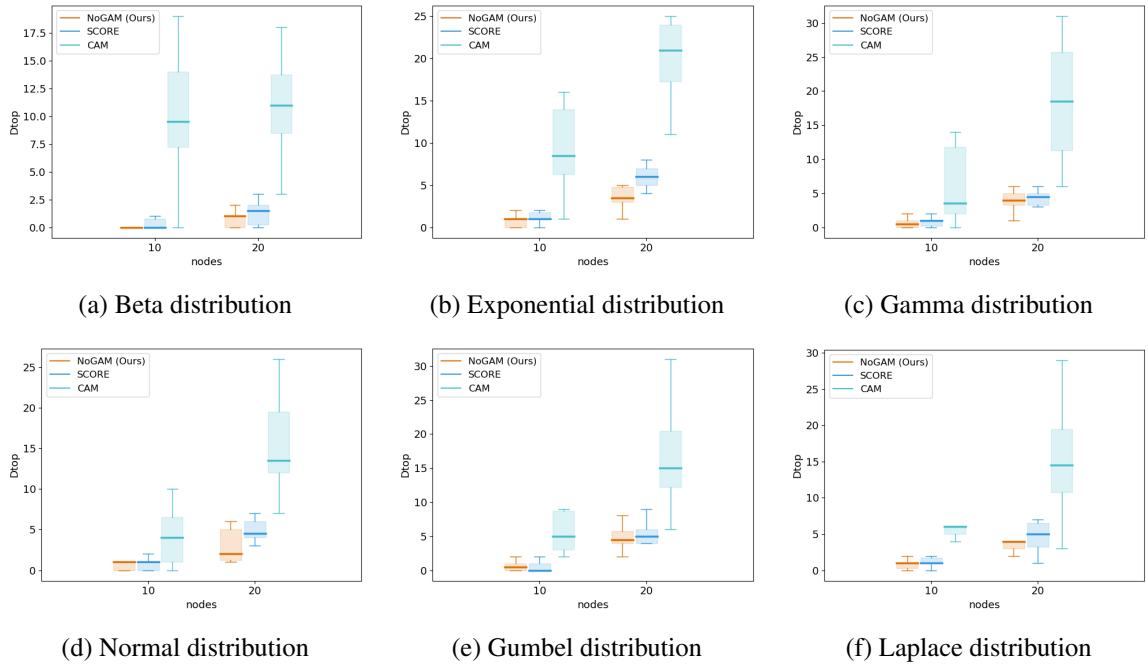
From Table 3 to 8 we provide the complete overview of the experiments described in section 6. Each metric is averaged over 10 runs, for which we record the empirical mean and deviation. Figure 5 shows how in the dense setting (ER4) NoGAM method of inference of the topological ordering outperforms CAM and SCORE, similarly to what observed in the sparse setting (ER1) illustrated in Figure 3. In Figure 6 we can see that SCORE and NoGAM have overall comparable SHD.

Table 3: Beta noise ER graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.1 ± 1.5	4.2 ± 5.4	0.6 ± 0.8	5.1 ± 1.3	11.4 ± 7.1	0.1 ± 0.3
	SCORE	1.6 ± 1.0	4.9 ± 3.5	0.6 ± 0.7	6.1 ± 2.0	9.6 ± 5.4	0.4 ± 0.7
	CAM	3.4 ± 1.9	13.8 ± 8.6	2.2 ± 1.4	14.4 ± 5.1	51.0 ± 14.3	7.7 ± 5.1
	GES	6.5 ± 2.0	22.0 ± 7.6	–	27.0 ± 5.2	72.0 ± 8.7	–
d=20	NoGAM	2.0 ± 1.6	8.0 ± 9.0	0.5 ± 0.5	30.0 ± 4.2	121.5 ± 22.6	0.9 ± 0.9
	SCORE	2.0 ± 1.5	8.0 ± 5.2	1.0 ± 0.3	33.1 ± 6.8	111.9 ± 29.1	1.4 ± 1.0
	CAM	9.5 ± 4.8	45.3 ± 16.0	5.7 ± 2.3	45.1 ± 7.2	220.8 ± 29.1	12.2 ± 7.0
	GES	12.9 ± 4.7	52.0 ± 14.4	–	78.5 ± 9.2	290.6 ± 25.9	–

Appendix C. Additional experiments on Scale Free graphs

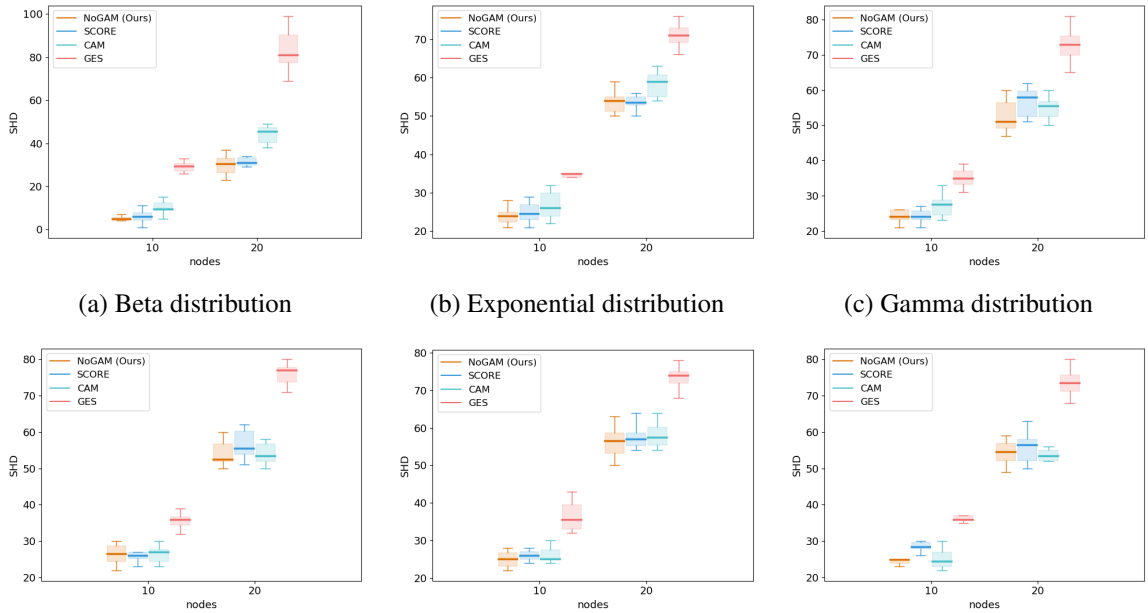
From Table 9 to 14 we provide additional experimental results on Scale-free (SF) graphs (Barabasi and Albert (1999)).



(a) Beta distribution (b) Exponential distribution (c) Gamma distribution

(d) Normal distribution (e) Gumbel distribution (f) Laplace distribution

Figure 5: Boxplots over 10 runs for topological ordering divergence D_{top} over dense graphs (ER4). GES algorithm doesn't appear since it does not require an explicit estimate of the topological ordering. From the illustrations we can see how NoGAM, in general, outperforms CAM and SCORE.



(a) Beta distribution (b) Exponential distribution (c) Gamma distribution

(d) Normal distribution (e) Gumbel distribution (f) Laplace distribution

Figure 6: Boxplots over 10 runs showing SHD performance over dense graphs (ER4).

Table 4: Exponential noise ER graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.2 ± 1.4	4.3 ± 5.2	0.3 ± 0.6	24.1 ± 2.0	43.3 ± 5.2	0.8 ± 0.8
	SCORE	2.3 ± 1.2	4.1 ± 4.2	0.8 ± 0.9	25.1 ± 3.0	40.9 ± 6.6	1.1 ± 0.6
	CAM	3.4 ± 1.4	14.8 ± 5.8	2.4 ± 1.0	24.6 ± 2.0	54.1 ± 11.9	4.4 ± 2.1
	GES	9.0 ± 3.0	33.1 ± 12.2	–	33.9 ± 3.7	80.6 ± 5.8	–
d=20	NoGAM	2.5 ± 1.4	11.3 ± 9.1	0.7 ± 0.6	54.2 ± 4.5	218.1 ± 28.7	3.6 ± 1.6
	SCORE	4.2 ± 2.8	19.7 ± 13.6	1.4 ± 1.4	54.1 ± 4.0	215.1 ± 40.4	6.1 ± 1.3
	CAM	8.2 ± 1.8	46.4 ± 24.4	5.5 ± 2.1	54.8 ± 4.3	258.4 ± 57.4	18.8 ± 7.5
	GES	19.8 ± 3.5	87.8 ± 24.4	–	72.9 ± 5.4	336.2 ± 24.0	–

Table 5: Gamma noise ER graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.2 ± 1.2	3.9 ± 5.4	0.3 ± 0.5	24.2 ± 1.8	41.4 ± 5.3	0.7 ± 0.8
	SCORE	1.6 ± 1.5	4.8 ± 6.3	1.4 ± 0.9	24.2 ± 0.9	41.6 ± 2.5	1.0 ± 1.1
	CAM	3.3 ± 1.6	17.3 ± 8.2	2.6 ± 1.7	26.5 ± 3.8	59.3 ± 15.1	8.0 ± 4.5
	GES	9.8 ± 2.2	34.6 ± 13.2	–	35.9 ± 3.6	80.8 ± 6.5	–
d=20	NoGAM	3.9 ± 1.7	13.5 ± 10.1	0.7 ± 0.8	52.7 ± 2.8	203.1 ± 24.9	3.7 ± 1.6
	SCORE	4.0 ± 1.8	14.8 ± 8.2	2.5 ± 1.3	56.5 ± 2.3	214.7 ± 21.7	4.7 ± 1.7
	CAM	8.3 ± 3.0	48.9 ± 23.3	5.1 ± 2.5	54.0 ± 3.5	264.5 ± 17.6	16.2 ± 4.6
	GES	18.8 ± 2.5	89.0 ± 19.4	–	73.9 ± 3.2	332.8 ± 19.2	–

Table 6: Gauss noise ER graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.4 ± 0.5	0.0 ± 0.0	0.0 ± 0.0	26.4 ± 2.5	43.9 ± 4.8	0.8 ± 0.9
	SCORE	0.9 ± 0.7	3.2 ± 2.4	0.1 ± 0.3	26.1 ± 3.0	43.7 ± 8.7	0.7 ± 0.6
	CAM	0.6 ± 0.7	1.3 ± 1.6	0.0 ± 0.0	26.9 ± 2.2	47.4 ± 5.2	5.1 ± 3.4
	GES	8.3 ± 1.8	31.9 ± 9.5	–	36.2 ± 2.4	85.3 ± 3.8	–
d=20	NoGAM	2.0 ± 1.0	10.1 ± 5.7	0.3 ± 0.5	54.0 ± 3.0	195.0 ± 17.0	3.1 ± 2.0
	SCORE	2.3 ± 1.3	11.6 ± 6.9	0.4 ± 0.7	56.5 ± 2.9	197.8 ± 22.3	5.0 ± 1.3
	CAM	2.1 ± 1.8	9.8 ± 9.9	0.8 ± 1.0	57.1 ± 1.9	204.1 ± 34.2	21.4 ± 8.4
	GES	17.1 ± 3.7	70.0 ± 22.1	–	71.9 ± 3.8	340.1 ± 20.9	–

Table 7: Gumbel noise ER graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.3 ± 1.4	3.3 ± 4.2	0.2 ± 0.4	24.9 ± 2.7	41.4 ± 4.8	0.6 ± 0.7
	SCORE	1.6 ± 1.4	6.3 ± 5.2	0.5 ± 0.5	26.1 ± 2.1	42.6 ± 5.7	0.5 ± 0.7
	CAM	1.4 ± 1.3	7.9 ± 8.8	0.8 ± 1.1	26.5 ± 1.9	55.9 ± 12.4	7.1 ± 4.3
	GES	9.7 ± 1.2	41.0 ± 11.1	–	38.2 ± 1.3	86.8 ± 2.5	–
d=20	NoGAM	2.7 ± 2.4	11.3 ± 6.8	0.6 ± 0.7	56.1 ± 3.9	207.9 ± 20.4	4.7 ± 2.4
	SCORE	3.5 ± 1.6	15.9 ± 17.4	1.3 ± 1.3	56.9 ± 2.2	205.6 ± 25.0	5.4 ± 1.6
	CAM	4.3 ± 1.4	29.7 ± 12.3	2.3 ± 1.4	57.0 ± 4.3	225.5 ± 24.7	17.6 ± 6.3
	GES	21.1 ± 1.9	95.4 ± 28.5	–	78.2 ± 2.4	348.5 ± 15.9	–

Table 8: Laplace noise ER graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.6 ± 0.7	2.2 ± 2.9	0.0 ± 0.3	24.5 ± 2.0	42.9 ± 6.5	0.8 ± 0.9
	SCORE	1.4 ± 1.0	3.0 ± 3.0	0.3 ± 0.4	27.9 ± 2.3	46.5 ± 6.1	1.1 ± 0.5
	CAM	1.1 ± 1.0	4.6 ± 4.9	0.4 ± 0.5	24.7 ± 1.7	43.8 ± 4.3	6.0 ± 3.1
	GES	10.0 ± 1.7	37.5 ± 10.1	–	35.6 ± 2.4	85.0 ± 2.3	–
d=20	NoGAM	2.0 ± 2.2	9.1 ± 6.4	0.4 ± 0.7	54.5 ± 3.4	204.2 ± 20.1	3.9 ± 1.5
	SCORE	3.1 ± 1.3	16.1 ± 6.6	1.3 ± 0.6	56.1 ± 1.9	209.6 ± 21.4	4.6 ± 2.0
	CAM	3.4 ± 1.2	11.1 ± 9.1	0.5 ± 0.7	54.8 ± 2.6	200.4 ± 16.1	18.5 ± 6.5
	GES	19.7 ± 2.4	90.4 ± 32.0	–	73.4 ± 3.1	329.4 ± 13.9	–

Table 9: Beta noise SF graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	2.6 ± 2.1	13.7 ± 15.9	1.7 ± 1.6	7.3 ± 2.3	12.9 ± 9.1	0.6 ± 0.7
	SCORE	2.3 ± 1.4	4.5 ± 4.0	0.7 ± 0.8	6.2 ± 1.4	16.0 ± 8.3	0.3 ± 0.5
	CAM	4.5 ± 1.2	29.9 ± 6.7	3.3 ± 0.8	14.3 ± 4.0	47.7 ± 18.6	6.4 ± 4.5
	GES	9.8 ± 2.9	56.3 ± 17.4	–	27.6 ± 5.8	73.7 ± 5.4	–
d=20	NoGAM	2.5 ± 1.7	24.8 ± 22.0	0.9 ± 1.0	31.1 ± 6.9	60.3 ± 20.0	0.8 ± 0.6
	SCORE	2.1 ± 1.4	5.8 ± 3.9	0.1 ± 0.3	31.6 ± 5.8	108.4 ± 18.7	1.2 ± 1.2
	CAM	6.3 ± 2.8	84.1 ± 40.8	4.6 ± 2.1	37.7 ± 7.4	178.9 ± 44.4	6.8 ± 4.2
	GES	21.3 ± 6.9	157.8 ± 48.5	–	93.3 ± 13.8	303.2 ± 13.1	–

Table 10: Exponential noise SF graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	1.5 ± 1.1	7.7 ± 6.7	0.9 ± 0.8	8.3 ± 2.1	31.6 ± 6.0	0.6 ± 0.8
	SCORE	0.7 ± 0.8	2.4 ± 3.6	0.3 ± 0.6	8.0 ± 2.6	39.4 ± 6.1	0.4 ± 0.5
	CAM	3.6 ± 2.0	26.6 ± 12.6	2.8 ± 1.3	11.7 ± 3.3	54.0 ± 9.1	4.8 ± 4.0
	GES	11.7 ± 2.6	59.0 ± 10.7	–	26.2 ± 3.0	81.0 ± 5.9	–
d=20	NoGAM	5.3 ± 2.5	61.7 ± 28.5	3.6 ± 1.7	24.5 ± 3.9	233.3 ± 24.4	4.8 ± 2.4
	SCORE	3.4 ± 1.7	14.8 ± 7.0	1.3 ± 0.9	24.7 ± 3.7	209.4 ± 29.3	4.2 ± 1.2
	CAM	7.7 ± 2.6	114.0 ± 37.8	5.8 ± 1.8	29.2 ± 3.2	271.5 ± 38.4	10.1 ± 5.2
	GES	26.8 ± 3.5	203.5 ± 48.5	–	59.0 ± 4.1	350.1 ± 10.6	–

Table 11: Gamma noise SF graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.7 ± 1.2	6.0 ± 12.0	0.6 ± 1.2	7.5 ± 2.6	36.4 ± 12.4	0.8 ± 1.2
	SCORE	0.7 ± 0.9	3.6 ± 4.7	0.5 ± 0.7	7.2 ± 2.5	41.3 ± 7.9	0.6 ± 1.0
	CAM	2.8 ± 2.4	19.3 ± 16.7	2.1 ± 1.8	10.4 ± 3.2	48.8 ± 14.8	4.2 ± 3.3
	GES	10.5 ± 2.5	58.5 ± 14.7	–	25.6 ± 2.3	80.4 ± 3.7	–
d=20	NoGAM	6.1 ± 2.3	71.5 ± 40.0	4.2 ± 1.9	26.6 ± 5.0	237.8 ± 21.4	4.3 ± 2.3
	SCORE	3.1 ± 2.1	13.2 ± 10.2	0.8 ± 0.6	27.1 ± 4.1	204.4 ± 23.3	3.7 ± 1.7
	CAM	8.6 ± 2.7	100.0 ± 38.3	5.6 ± 1.6	29.0 ± 6.6	270.7 ± 41.7	8.6 ± 4.3
	GES	27.1 ± 3.1	187.3 ± 31.1	–	59.8 ± 5.1	348.6 ± 11.1	–

Table 12: Gauss noise SF graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.3 ± 0.5	2.0 ± 4.0	0.2 ± 0.4	6.4 ± 1.9	34.1 ± 5.3	0.0 ± 0.0
	SCORE	0.3 ± 0.6	2.7 ± 5.8	0.1 ± 0.3	7.6 ± 2.9	31.6 ± 9.4	1.1 ± 0.7
	CAM	0.2 ± 0.4	1.5 ± 3.0	0.0 ± 0.0	9.8 ± 2.3	39.7 ± 9.9	1.0 ± 0.9
	GES	10.6 ± 3.0	49.7 ± 15.2	–	24.1 ± 3.3	80.7 ± 3.5	–
d=20	NoGAM	1.2 ± 0.9	14.2 ± 16.3	0.7 ± 0.6	15.8 ± 4.6	224.6 ± 22.4	2.3 ± 1.6
	SCORE	0.9 ± 0.9	13.8 ± 12.6	0.7 ± 0.8	17.5 ± 3.5	179.2 ± 23.8	4.9 ± 3.0
	CAM	0.3 ± 0.5	1.9 ± 5.7	0.0 ± 0.0	24.8 ± 3.3	240.7 ± 29.8	3.1 ± 2.4
	GES	28.1 ± 7.6	212.6 ± 51.0	–	58.5 ± 3.6	360.5 ± 2.7	0.0 ± 0.0

Table 13: Gumbel noise SF graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.9 ± 0.9	6.0 ± 5.2	0.6 ± 0.7	7.2 ± 2.3	28.1 ± 7.5	0.2 ± 0.4
	SCORE	1.2 ± 1.3	5.0 ± 5.1	0.7 ± 0.6	7.6 ± 3.2	44.0 ± 9.3	0.5 ± 0.5
	CAM	0.4 ± 0.5	4.0 ± 4.9	0.4 ± 0.5	8.0 ± 3.0	37.5 ± 13.4	1.5 ± 2.3
	GES	10.5 ± 2.6	55.1 ± 13.0	$-$	24.0 ± 3.4	80.3 ± 6.1	$-$
d=20	NoGAM	2.1 ± 1.4	26.5 ± 20.0	1.5 ± 0.9	25.3 ± 3.7	237.3 ± 21.0	2.7 ± 2.0
	SCORE	3.2 ± 1.9	20.0 ± 13.6	1.1 ± 1.4	28.1 ± 4.1	212.3 ± 19.7	5.6 ± 2.4
	CAM	1.5 ± 1.6	13.7 ± 16.6	0.8 ± 0.8	26.0 ± 4.5	247.0 ± 31.6	4.7 ± 2.2
	GES	28.0 ± 8.7	195.9 ± 34.7	$-$	58.9 ± 4.8	353.3 ± 12.7	$-$

Table 14: Laplace noise SF graphs

		ER1 (sparse)			ER4 (dense)		
	Method	SHD	SID	D_{top}	SHD	SID	D_{top}
d=10	NoGAM	0.1 ± 0.3	0.9 ± 2.7	0.1 ± 0.3	8.0 ± 2.1	30.4 ± 8.3	0.5 ± 0.9
	SCORE	0.5 ± 1.2	1.5 ± 3.4	0.0 ± 0.0	8.2 ± 1.7	43.6 ± 6.5	0.5 ± 0.7
	CAM	0.4 ± 0.5	1.9 ± 3.8	0.1 ± 0.3	7.0 ± 2.2	34.0 ± 12.1	1.0 ± 1.3
	GES	12.8 ± 4.7	60.4 ± 19.7	$-$	26.1 ± 1.9	82.4 ± 2.8	$-$
d=20	NoGAM	1.7 ± 0.9	12.6 ± 8.6	0.7 ± 0.6	25.9 ± 3.3	244.3 ± 22.9	2.5 ± 2.3
	SCORE	2.6 ± 1.1	16.7 ± 15.7	1.1 ± 1.0	27.4 ± 3.0	207.6 ± 18.4	4.8 ± 2.2
	CAM	1.1 ± 0.9	3.9 ± 7.8	0.1 ± 0.3	23.7 ± 3.9	227.2 ± 15.1	2.5 ± 1.8
	GES	29.0 ± 6.7	207.3 ± 34.7	$-$	58.0 ± 6.6	353.4 ± 7.7	$-$