Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy To Game

Anonymous Author(s)
Affiliation
Address
email

Abstract

Simulated DAGs may exhibit properties that inadvertently render their structure identifiable and unexpectedly affect structure learning algorithms. Here, we show that marginal variance tends to increase along the causal order for generically sampled additive noise models. We introduce *varsortability* as a measure of the agreement between the order of increasing marginal variance and the causal order. For common sampling schemes, we show that the remarkable performance of some continuous structure learning algorithms can be explained by high varsortability and matched by a simple baseline method. Yet, this performance may not transfer to real-world data where varsortability may be absent or dependent on the choice of measurement scales. On standardized data, the same algorithms no longer identify the ground-truth DAG or its Markov equivalence class. While standardization removes the pattern in marginal variance, we show that data generating processes that incur high varsortability also leave a distinct covariance pattern that may be exploited even after standardization. Our findings challenge the significance of generic benchmarks with independently drawn parameters.

1 Introduction

Causal structure learning aims to infer a causal model from data. Academic disciplines anywhere from biology, medicine, finance, to machine learning are interested in causal models [Rothman et al., 2008, Imbens and Rubin, 2015, Sanford and Moosa, 2012, Schölkopf, 2019]. Causal models not only describe the observational joint distribution of variables but also formalize how to predict the effects of interventions and counterfactuals [Spirtes et al., 2000, Pearl, 2009, Peters et al., 2017]. Causal structure is commonly represented by a directed acyclic graph (DAG). The nodes represent variables and the directed edges pointing from cause to effect represent the causal relationships. This graphical representation rests on strong assumptions which have been critically questioned, for example by Dawid [2010]. Inferring causal structure from observational data is difficult: Often we can only identify the DAG up to its Markov equivalence class (MEC) and identifying high-scoring DAGs is NP-hard [Chickering, 1996, Chickering et al., 2004]. Here, we focus on DAGs encoding linear additive noise models (ANM).

Data scale and marginal variance can affect causal discovery and may carry information about the data generating process. It can dominate benchmarking results, such as the outcome of the NeurIPS Causality 4 Climate competition [Runge et al., 2020]. Here, the magnitude of regression coefficients was informative about the existence of causal links and enabled ordinary regression-based methods on raw data to outperform causal discovery algorithms [Weichwald et al., 2020]. Multiple prior works implicitly or explicitly state the importance of data scale. The structure identified by *ICA-LiNGAM* [Shimizu et al., 2006], for example, is susceptible to rescaling of the variables. This motivated the development of *DirectLiNGAM* [Shimizu et al., 2011], a scale-invariant causal

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

discovery algorithm for linear non-Gaussian models. The causal structure of additive noise models is proven to be identifiable given the noise scale (cf. Section 2.2). Yet, such identifiability results require knowledge about the true data scale.

Simulated DAGs may be identifiable from marginal variances for generic choices of parameter distributions. An instructive example is the causal graph $A \to B$ with structural equations $A = N_A$ and $B = wA + N_B$ where w is non-zero and N_A, N_B are independent noise variables with zero-centered distributions. It holds that $\mathrm{MSE}\,(A \to B) < \mathrm{MSE}\,(A \leftarrow B) \iff \mathrm{Var}(A) < \mathrm{Var}(B) \iff (1 - w^2)\,\mathrm{Var}(N_A) < \mathrm{Var}(N_B)$ (see Appendix A). Deciding the directionality of the edge between A and B based on the MSE amounts to inferring an edge from the lower-variance variable to the higher-variance variable. For error variances $\mathrm{Var}(N_A) \leq \mathrm{Var}(N_B)$ and any non-zero edge weight w, the MSE-based inference is correct. This resembles known scale-based identifiability results based on equal or monotonously increasing error variances [Peters and Bühlmann, 2014, Park, 2020]. However, if the values of A were multiplied by a sufficiently large constant, the MSE-based inference would wrongfully conclude that $A \leftarrow B$. This is problematic as we may be changing the variance just by choosing our units of measurement differently. Arguably, this is often the case for observations from real-world systems: There is no canonical choice as to whether we should pick meters or yards for distances, gram or kilogram for weights, yuan or dollar for currency. Presently, a researcher applying any method that leverages the data scale (such as, for example, Peters and Bühlmann [2014], Zheng et al. [2018], Park [2020]) cannot rely on obtaining the same results for different measurement scales or after re-scaling the data.

Continuous causal structure learning algorithms optimize model fit under a differentiable acyclicity constraint [Zheng et al., 2018]. This allows for the use of continuous optimization and avoids the explicit traversal of possible causal structures. This idea has found numerous applications and extensions such as those by Lachapelle et al. [2019], Lee et al. [2019], Ng et al. [2020], Yu et al. [2019], Brouillard et al. [2020], Pamfil et al. [2020], Wei et al. [2020], Zheng et al. [2020], Bhattacharya et al. [2021]; see Vowels et al. [2021] for a review. *NOTEARS* [Zheng et al., 2018] use the MSE with reference to Loh and Bühlmann [2014], while *GOLEM* [Ng et al., 2020] assesses model fit by the penalized likelihood assuming a jointly Gaussian model. On simulated data and across noise types, both methods recover graphs that are remarkably close to the ground truth causal graph in structural intervention distance (SID) and structural hamming distance (SHD). We agree with the original authors that these empirical findings, especially under model misspecification and given the non-convex loss landscape, may seem surprising at first. Here, we investigate the performance under data standardization and explain how the causal order is (partially) identifiable from the data scale alone in common generically simulated benchmarking data.

Contribution. We show that causal structure drives the marginal variances of nodes in an additive noise model and can lead to (partial) identifiability. This pattern in marginal variances is dominant in ANM benchmark simulations with edge coefficients drawn identically and independently. We introduce varsortability as a measure of the information the data scale carries about the causal structure. We argue that high varsortability affects the identification of stationary points in the optimization procedures of continuous structure learning algorithms. Our experiments demonstrate that this effect dominates their behavior and helps them achieve state-of-the-art performance. Data standardization or an unknown data scale remove this information and the same algorithms no longer recover the DAG or its Markov equivalence class. We observe this behavior even for the score-equivalent criterion used in *GOLEM* and the effect extends to the identification of the MEC. To illustrate that recent benchmark results depend heavily on high varsortability, we provide a simple baseline method that exploits increasing marginal variances to achieve state-of-the-art results on these benchmarks. We thereby provide an explanation of the unexpected performance of recent continuous structure learning algorithms in identifying the true DAG. We further compare the algorithms on real-world data and do not find a dominating algorithm. We show how, even if data is standardized and even in non-linear additive noise models, a causal discovery benchmark may be gamed due to covariance patterns. Consequently, recent benchmark results may not transfer to real-world settings in which no such knowledge is available or where causal edge weights are not drawn iid. We conclude that benchmarks on generically sampled additive noise models may be unexpectedly distorted due to inadvertent patterns in the data.

2 Background

2.1 Model Class

We consider acyclic linear additive noise models. Individual observations are denoted by $x^{(i)} \in \mathbb{R}^d$ where $x_j^{(i)}$ denotes the ith i.i.d. observation of the jth dimension of $X = [X_1, ..., X_d]^\top$. All observations are stacked as $\mathbf{X} = [x^{(1)}, ..., x^{(n)}]^\top \in \mathbb{R}^{n \times d}$ and $x_j \in \mathbb{R}^n$ refers to the jth column of $\mathbf{X}$. Analogously, $n^{(i)}$ denotes the corresponding i.i.d. observation of the noise vector $N = [N_1, ..., N_d]^\top$, which has zero-centred independent components. The linear effect of variable X_k on X_j is denoted by $w_{k \to j} = w_{kj}$. The causal structure corresponding to the adjacency matrix $W = [w_{kj}]_{k,j=1,...,d}$ with columns $w_j = [w_{k \to j}]_{j=1,...,d} \in \mathbb{R}^d$ can be represented by a directed acyclic graph $G = (V_G, E_G)$ with vertices $V_G = \{1, ..., d\}$ and edges $E_G = \{(k, j) : w_{k \to j} \neq 0\}$. Edges can be represented by an adjacency matrix E such that the $(k, j)^{\text{th}}$ entry of E^l is non-zero if and only if a directed path of length l from k to j exists in G. For a given graph, the parents of j are denoted by $\mathrm{PA}(j)$. The structural causal model is $X = W^\top X + N$.

2.2 Identifiability of Additive Noise Models

Identifiability of the causal structure or its MEC requires causal assumptions. Under causal faithfulness and Markov assumptions, the causal graph can be recovered up to its MEC [Chickering, 1995, Spirtes et al., 2000]. Faithfulness, however, is untestable [Zhang and Spirtes, 2008]. Shimizu et al. [2006] show that under the assumption of no unobserved confounders, faithfulness, linearity, and non-Gaussian noise, the causal graph can be recovered from data. Hoyer et al. [2009] show that this holds for any noise distribution under the assumption of strict non-linearity. This finding is generalized to post-nonlinear functions by Zhang and Hyvarinen [2012]. Peters and Bühlmann [2014] prove that the causal structure of a linear causal model with Gaussian noise is identifiable if the error variances are equal or known. Any unknown re-scaling of the data breaks this condition. For the case of linear structural causal models, Loh and Bühlmann [2014] provide a framework for DAG estimation based on a noise variance-weighted least squares score function. For additive noise models, they give conditions under which the general Gaussian case can be identified via approximating it by the equal noise-variance case given knowledge of the (approximate) noise scale. Finally, subsuming further prior results on (linear) additive noise models [Hoyer et al., 2009, Ghoshal and Honorio, 2017, 2018, Chen et al., 2019], Park [2020] show that the causal structure is identifiable under regularity conditions on the conditional variances along the causal order. In particular, identifiability holds if the error variances of nodes are weakly monotone increasing along the causal order.

2.3 Structure Learning Algorithms

Combinatorial structure learning algorithms (such as *PC, FGES, DirectLiNGAM*) solve the combinatorial problem of finding the true DAG or its MEC and separately find the optimal parameters for each structure. To remain computationally feasible, the search space of potential structures is often restricted or traversed according to a heuristic. This can, for example, be accomplished by carefully choosing which conditional independence statements to evaluate in constraint-based algorithms, or by greedy (equivalence class) search in score-based algorithms. In our experiments, we consider *PC* [Spirtes and Glymour, 1991], *FGES* [Meek, 1997, Chickering, 2002], and *DirectLiNGAM* [Shimizu et al., 2011]. For details see Appendix I.

Continuous structure learning algorithms (such as *NOTEARS* and *GOLEM*) employ continuous optimization to simultaneously optimize over structures and parameters. As a first step towards expressing causal structure learning as a continuous optimization problem, Aragam and Zhou [2015] propose l^1-regularization instead of the conventional l^0-penalty for model selection. Zheng et al. [2018] propose a differentiable acyclicity constraint, allowing for end-to-end optimization of score functions over graph adjacency matrices. We examine and compare the continuous structure learning algorithms *NOTEARS* [Zheng et al., 2018] and *GOLEM* [Ng et al., 2020]. For details see Appendix I.

NOTEARS [Zheng et al., 2018] minimizes the MSE between observations and model predictions subject to a hard acyclicity constraint. The MSE with respect to W on observations $\mathbf{X}$ is defined as

$$\mathrm{MSE}_{\mathbf{X}}(W) = \|\mathbf{X} - \mathbf{X}W\|_2^2.$$

GOLEM [Ng et al., 2020] performs maximum likelihood estimation (MLE) under the assumption of a Gaussian distribution with equal (EV) or non-equal (NV) noise variances. There are soft acyclicity and sparsity penalties. The unnormalized likelihood-parts of the objective function are

$$\mathcal{L}_{EV}(W, \mathbf{X}) = \log(\mathrm{MSE}_{\mathbf{X}}(W)), \quad \text{or} \quad \mathcal{L}_{NV}(W, \mathbf{X}) = \sum_{j=1}^{d} \log\left(\|x_j - \mathbf{X}w_j\|_2^2\right).$$

To ease notation, we sometimes drop the explicit reference to $\mathbf{X}$ when referring to MSE, $\mathcal{L}_{EV}, \mathcal{L}_{NV}$.

3 Varsortability

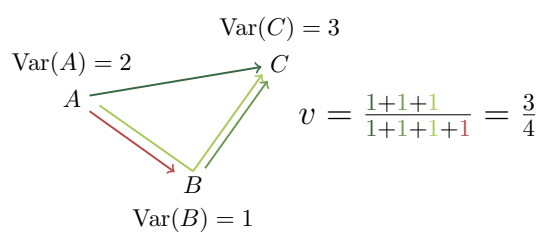
In this section we present how the data generating process leaves information about the causal order in the data scale. We introduce varsortability as a measure of such information. When varsortability is maximal, the causal order is identifiable. We show that varsortability is high in common simulation schemes used for benchmarking causal structure learning algorithms. We describe the process by which even score-equivalent continuous structure learning algorithms are affected by marginal variances and how they may leverage this information to infer the causal order. This explains the results of continuous methods reported by Zheng et al. [2018], Ng et al. [2020], and others. We reveal the extent to which increasing marginal variances drive benchmarking results by matching the performance of continuous structure learning algorithms using a simple baseline method *sortnregress* (cf. Appendix F) which sorts by marginal variance followed by parent selection.

3.1 Definition of Varsortability

We propose varsortability as a measure of agreement between the order of increasing marginal variance and the causal order. For any causal graph $G = (V, E)$ over variables $V = \{X_1, ..., X_d\}$, we define varsortability as the fraction of directed paths of any length which start from a node with strictly lower variance than the node they end in, that is,

$$v := \frac{\sum_{k=1}^{d-1} \sum_{i \to j \in E^k} \mathbf{1}[\mathrm{Var}(X_i) < \mathrm{Var}(X_j)]}{\sum_{k=1}^{d-1} \sum_{i \to j \in E^k} 1} \in [0, 1].$$

Consider the following illustration:



Varsortability equals one if the marginal variance of each node is strictly greater than that of its causal ancestors. Varsortability equals zero if the marginal variance of each node is greater or equal than that of its causal descendants. Varsortability does not depend on choosing one of the possibly multiple causal orders and captures the overall agreement between the partial order induced by the marginal variances and all pathwise descendant relations implied by the causal structure. In the two-node introductory example (cf. Section 1), varsortability $v = 1$ implies that

$$\mathrm{Var}(A) < \mathrm{Var}(B) \iff \mathrm{Var}(N_A) < w^2\,\mathrm{Var}(N_A) + \mathrm{Var}(N_B)$$

where A and B are nodes in the causal graph $A \xrightarrow{w} B$ with noise variances $\mathrm{Var}(N_A)$ and $\mathrm{Var}(N_B)$.

Varsortability can also be understood as a property of the distribution from which graphs and parameters are sampled. As such, varsortability is of concern for synthetic benchmarking settings. If the weights and noise variances are drawn at random, the distributions they are drawn from determine the probability of obtaining varsortability for each individual connected node pair in the graph. Even for modest probabilities of varsortability between any two neighboring nodes, the variance of nodes

along the causal order increases quickly. The further apart a pair of nodes along the chain, the greater the chance their ordering by variance corresponds to the causal order. In practice, for many instantiations of linear and non-linear additive noise models, there are high chances an ordering of nodes by marginal variance is closer to the causal ordering than a random ordering (cf. Appendix E).

3.2 Varsortability and Identifiability

If varsortability $v = 1$, the causal structure is identifiable. It can be recovered by ordering the nodes by increasing marginal variance and regressing each node onto its predecessors using conventional sparse regression approaches. While closely related, the identifiability conditions in Park [2020] are not equivalent to varsortability, which we prove in Appendix C. Decoupling the causal structure learning problem into causal order estimation and parent selection is not uncommon in the literature [Shojaie and Michailidis, 2010]. It is warranted since we only need the causal ordering to consistently estimate interventional distributions (cf. e.g. [Bühlmann et al., 2014, Section 2.6]). Parent selection prunes the fully-connected DAG for a previously identified order to identify the causal structure.

While identifiability of the causal order is immediate if varsortability $v = 1$, it shares severe drawbacks with other identifiability conditions that rely on data scale [Peters and Bühlmann, 2014, Loh and Bühlmann, 2014, Park, 2020]. First, it is difficult to verify or assess the plausibility of such assumptions for any given dataset, even if variables are on the same measurement scale. Second, any unknown rescaling may break previously met identifiability conditions.

3.3 Varsortability in Benchmarking Scenarios

For real-world data we cannot readily assess the extent to which varsortability holds as we do not know the true data scale of the data generating process. When benchmarking causal structure learning algorithms, however, we can evaluate varsortability for the simulated DAGs and parameter settings. The probability that any two nodes in a graph which are connected by a directed path have increasing variance along that path depends on the distributions of edge weights and noise variances. We may acquire an intuition about how probable it is to encounter varsortable cause-effect pairs in our simulation settings by considering two nodes A and B in the sampled graph with direct edge $A \overset{w}{\to} B$ and without any common ancestors and no other directed path from A to B. The probability that this variable pair is varsortable in a simulated ANM with edge coefficients w and noise variances $\sigma^2_{N_A}, \sigma^2_{N_B}$ drawn independently is bounded from below by

$$\mathrm{P}[(1 - w^2)\sigma^2_{N_A} < \sigma^2_{N_B}] \quad \text{(cf. Appendix B).}$$

In particular, a variable pair is varsortable whenever the direct edge weight turned out to be $w \geq 1$, assuming all noise variances are non-zero. In the common experimental settings used by, for example, Zheng et al. [2018, 2020], Lachapelle et al. [2019], Ng et al. [2020], weights $w \in [a, b]$, and standard deviations of the noise terms $\sigma_{N_A}, \sigma_{N_B} \in [c, d]$ are bounded and uniformly, identically, and independently distributed. For $w \overset{\text{iid}}{\sim} \mathrm{Unif}((-2, -0.5) \cup (0.5, 2))$ and $\sigma_{N_A}, \sigma_{N_B} \overset{\text{iid}}{\sim} \mathrm{Unif}((0.5, 2))$ the variable pair is varsortable with probability ≈ 0.93. Empirically, we find that varsortability averages above 0.94 in our simulated graphs and above 0.71 in commonly considered non-linear additive noise models (cf. Appendix E). This result is indicative of how often varsortability may hold in benchmark simulations and to what extent they may be gamed by sorting variables by their marginal variance. If A and B have a common ancestor or mediator C, which is improbable in commonly simulated sparse graphs, the effect of C on B may either compound or partially cancel out the effect of A on B. In practice, the effect commonly increases the variance of the effect node B, which may be attributed to the independent sampling of path coefficients which also renders faithfulness violations improbable [Meek, 1995]. We find varsortability to increase with graph density and the lower bound presented above to be loose. In general, when simulating additive noise models we advocate to evaluate the varsortability of the simulated graphs empirically (cf. Appendix E.3 for the implementation). We emphasize that even for varsortability < 1, where identifiability of the causal order is no longer guaranteed, experimental results may still be largely driven by the overall agreement between increasing marginal variance and causal order. The extent to which varsortability may distort experimental comparisons for linear ANMs is demonstrated in Section 4 and indicative results for non-linear ANMs are shown in Section 5.2.

3.4 Marginal Variance yields Asymmetric Gradients for Causal and Anti-Causal Edges

Here, we explain how varsortability may dominate the performance of continuous structure learning algorithms. We do not expect combinatorial structure learning algorithms such as *PC*, as local constraint-based algorithm, nor *FGES*, as locally greedy score-based search using a score-equivalent criterion, nor *DirectLiNGAM*, a procedure minimising residual dependence, to be dependent on the data scale. Recent continuous structure learning algorithms, however, may depend on data scale irrespective of whether the employed score is score-equivalent (e.g. *GOLEM* for Gaussian models) or not (e.g. *GOLEM* under likelihood misspecification or *NOTEARS*).

We first establish how varsortability affects the gradients of score functions that assume equal noise variances and initialise with the empty graph $0_{d\times d}$. Since $\nabla \mathrm{MSE}\,(W) \propto \mathbf{X}^\top(\mathbf{X} - \mathbf{X}W)$ we have that $\nabla \mathrm{MSE}\,(0_{d\times d}) \propto \mathbf{X}^\top\mathbf{X}$ and $\nabla\mathcal{L}_{EV}(0_{d\times d}) \propto \log(\mathbf{X}^\top\mathbf{X})$ where the logarithm is applied element-wise. The initial gradient step of both *NOTEARS* and *GOLEM-EV* is symmetric. From the next step onward, the symmetry is broken. We have $\nabla \mathrm{MSE}(W) \propto [\mathbf{X}^\top(x_1 - \mathbf{X}w_1), ..., \mathbf{X}^\top(x_d - \mathbf{X}w_d)]$ where the jth column $\mathbf{X}^\top(x_j - \mathbf{X}w_j)$ reflects the vector of empirical covariances of the jth residual vector $x_j - \mathbf{X}w_j$ with each x_i. Provided a small identical learning rate was used across all entries of W in the first step, we expect the residuals after the first gradient step to be larger in those components that have higher marginal variance. As a result, the entries in the gradient column $\mathbf{X}^\top(x_j - \mathbf{X}w_j)$ are larger for those x_j with higher marginal variance. This leads to bigger gradient steps for edges pointing in the direction of high-variance nodes. Given high varsortability, this corresponds predominantly to the causal direction. This way, the global information about the causal order due to varsortability is effectively exploited.

Once we allow for unequal noise variances as in *GOLEM-NV*, the marginal variances lead the gradients differently. Letting $\mathrm{MSE}_j\,(w_j) = \|x_j - \mathbf{X}w_j\|_2^2$, we have

$$\nabla\mathcal{L}_{NV}(W) \propto \left[\frac{1}{\mathrm{MSE}_j\,(w_j)} \log\left(\mathbf{X}^\top(x_j - \mathbf{X}w_j)\right)\right]_{j=1,\ldots,d}$$

such that the logarithmic derivative breaks the symmetry of the first step for the non-equal variance formulation of *GOLEM*. We have $\nabla\mathcal{L}_{NV}(0_{d\times d}) \propto \log(\mathbf{X}^\top\mathbf{X})\,\mathrm{diag}(\|\mathrm{x}_1\|_2^{-2}\ldots, \|\mathrm{x}_\mathrm{d}\|_2^{-2})$. Although mitigated by the $\log$-term, the column-wise inverse MSE scaling leads to bigger gradient steps for edges pointing in the direction of low-variance nodes. Given high varsortability, this corresponds predominantly to the anti-causal direction.

We focus on the first optimization steps to explain a) why continuous structure learning algorithms that assume equal noise variance work remarkably well in the presence of high varsortability and b) why performance changes once data is standardized and the marginal variances no longer hold information about the causal order. We conjecture that the first gradient steps have a dominant role in determining the causal structure, even though afterwards the optimization is governed by a non-trivial interplay of optimizer, model fit, constraints, and penalties. Because of the acyclicity constraint, it may be enough for a weight $w_{i\to j}$ to be greater in magnitude than its counterpart $w_{j\to i}$ early on in the optimization for the smaller edge to be pruned from there on. For a discussion of the interplay between sparsity penalties and data scale, we defer the reader to Appendix H, which indicates that the nodes need to be on a comparable data scale for l^1-penalization to be well calibrated. We refer to Ng et al. [2020] for a discussion of sparsity and acyclicity constraints in continuous DAG learning.

4 Simulations

We compare the performance of the algorithms introduced in Section 2.3 on raw and standardized synthetic data. In our comparison, we distinguish between settings with different noise distributions, graph types, and graph sizes. Our experimental set-up follows those in Zheng et al. [2018], Ng et al. [2020] and we contribute results obtained repeating their experiments in Appendix J. We complement our and previous DAG-recovery results by additionally evaluating how well the DAG output by continous structure learning algorithms identifies the MEC of the ground-truth DAG.

4.1 Data Generation

We sample Erdös-Rényi (ER) [Erdős and Rényi, 1960] and Scale-Free (SF) [Barabási and Albert, 1999] graphs and the parameters for additive noise models according to the simulation details in

Table 1. For a graph specified as ER-k or SF-k with d nodes, we simulate dk edges. For every combination of parameters, we create a raw data instance and a standardized version that is de-meaned and re-scaled to unit variance. On standardized data, we have varsortability $v = 0$ and the marginal variances hold no information about the causal ordering of the nodes. In all our experimental settings, varsortability averages above 0.94 on the raw data scale (cf. Appendix E.1).

Table 1: Parameters for synthetic data generation.

Repetitions	10
Graphs	ER-2, SF-2, SF-4
Edge weights	i.i.d. $\mathrm{Unif}((-2, -.5) \cup (.5, 2))$
Noise distributions	Exponential, Gaussian, Gumbel
Noise standard deviations	1 (Gaussian-EV); i.i.d. $\mathrm{Unif}(.5, 2)$ (others)
Nodes	$d \in \{10, 30, 50\}$
Samples	$n = 1000$

4.2 Evaluation

We evaluate performance using structural intervention distance (SID) [Peters and Bühlmann, 2015] and structural Hamming distance (SHD) between the recovered graph and ground truth. Additionally, we contribute a performance assessments of continuous structure learning algorithms in terms of the SID and SHD between the ground truth MEC and the recovered MEC (PC, FGES) or the MEC identified by the recovered graph (NOTEARS, GOLEM). SID assesses in how far the recovered structure enables to correctly predict the effect of interventions. SHD measures the distance between the true and recovered graph by counting how many edge insertions, deletions, and reversals are required to turn the former into the latter. Since interventional distributions can consistently be estimated given only the causal ordering [Bühlmann et al., 2014], SID is less susceptible to arbitrary choices of edge pruning procedures and thresholds than SHD. Intuitively, SID prioritizes the causal order, while SHD prioritizes the correctness of individual edges. We follow common practices for edge thresholding and scoring (see Appendix J).

4.3 Performance on Raw Versus Standardized Data

We group our algorithms into combinatorial and continuous algorithms. *sortnregress* serves as reference marking the performance achievable by exploiting marginal variances. We indicate its performance on standardized data as *randomregress*, since it amounts to choosing a random order and regressing each node on its predecessors. We use boxplots aggregating the performance achieved in the 10 repetitions on raw and standardized data by each algorithm and create separate plots per noise type to account for identifiability and MLE specification differences. We show the results obtained for ER-2 graphs and Gaussian noise with non-equal variances in Figure 1. These results are representative of the results obtained for different graphs and noise distributions (cf. Appendix J).

We observe that combinatorial algorithms are less scale-sensitive than continuous algorithms or *sortnregress*, which perform vastly different on raw and standardized data. For the simulated settings, varsortability is high (> 0.94) on the raw data scale (cf. Appendix E.1). *sortnregress* achieves competitive performance on raw, and baseline performance on standardized data. It thus qualifies as diagnostic tool to highlight how much of a given causal structure learning task can be resolved by exploiting data scale and sorting nodes by their marginal variance. The performance of *NOTEARS* and *GOLEM-EV*, which both assume equal noise variances, follows the same pattern. The performance of *GOLEM-NV* is also scale-sensitive but improves upon standardization, as predicted by our gradient analysis in Section 3.4. This evidence corroborates our claim that the remarkable performance on raw data and the overall behavior upon standardization of the continuous structure learning algorithms are driven by high varsortability. On a real-world causal protein signaling data set [Sachs et al., 2005] we find no indication for high varsortability (cf. Appendix G).

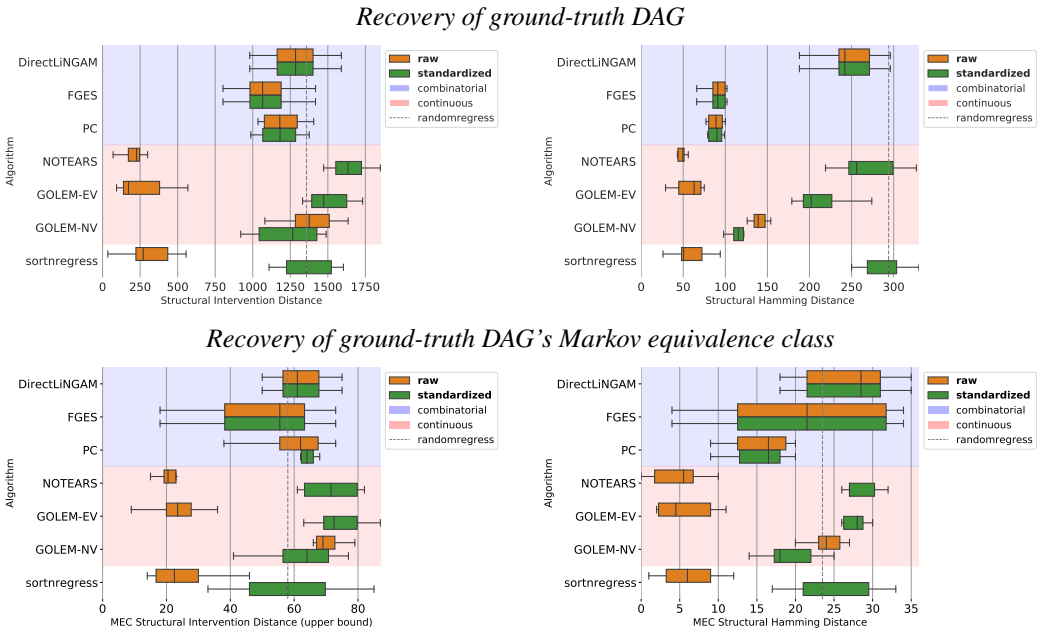


Figure 1: SID (left, lower is better) and SHD (right, lower is better) between recovered and ground-truth graphs (top) or Markov equivalence classes (bottom) for ER-2 graphs with 50 (top) or 10 (bottom) nodes and Gaussian-NV noise. The performance of *sortnregress*, which only exploits varsortability, matches that of the continuous methods *NOTEARS* and *GOLEM*.

5 Gaming Further Benchmarks

5.1 Orienting Causal Chains on Standardized Data

In order to design a causal discovery benchmark that does not favour methods that explicitly exploit marginal variances we may standardize the data or employ coefficient re-scaling schemes. Mooij et al. [2020], for example, propose a scale-harmonization by dividing each column $w_j = [w_{k\to j}]_{j=1,\dots,d} \in \mathbb{R}^d$ of the drawn adjacency matrices by $\sqrt{\|w_j\|^2 + 1}$ such that each variable would have comparable scale if all its direct parents were independently standard-normally distributed. However, this does not avoid the problem of inadvertent patterns in simulated ANMs. Even after standardization or scale-harmonization, DAGs with previously high varsortability generate data with distinct covariance patterns that may be exploited.

In Appendix D we present an instructive example of a decision rule that can infer the orientation of a causal chain from raw, standardized, and scale-harmonized data with accuracy strictly greater than 50%. For a causal chain $X_1 \to X_2 \to \dots \to X_d$ where edge weights and noise terms are drawn iid we can decide between the two Markov-equivalent graphs $X_1 \to X_2 \to \dots \to X_d$ and $X_1 \leftarrow X_2 \leftarrow \dots \leftarrow X_d$ whith above-chance accuracy. The empirical results for varying chain-lengths and various edge-weight distribution are deferred to the appendix where we discuss the 3-variable chain in detail and illustrate that the phenomenon extends from finite-sample to the population setting.

The intuition is as follows. Consider data generated by $X_1 \to X_2 \to \dots \to X_d$ and the aim is to infer from standardized data whether $X_1 \to \dots \to X_d$ or $X_1 \leftarrow \dots \leftarrow X_d$. For data with high varsortability and comparable noise variance on the raw data scale it holds that the further downstream a node X_i is in the causal chain, the stronger the variance of its parent $\mathrm{Var}(X_{i-1})$ contributes to its marginal variance $\mathrm{Var}(X_i) = \mathrm{Var}(X_{i-1}) + \mathrm{Var}(N_i)$ relative to its noise variance $\mathrm{Var}(N_i)$, and the stronger is it correlated with its parent. Thus, the sequence of regression coefficients, which in the standardized case amounts to $(\mathrm{Corr}(X_i, X_{i+1}))_{i=1,\dots,d-1}$, tends to increase in magnitude along the causal order and decrease in the anti-causal direction. The proposed decision rule predicts the causal direction as the one in which the absolute values of the regression coefficients tend to increase. This chain orientation rule achieves above-chance performance on raw, standardized, and scale-harmonized data (cf. Appendix D).

5.2 Sorting by Variance in Non-Linear Settings

Table 2 shows the results of a sorting by variance and filling in all possible edges in a non-linear setting. This is more naive than *sortnregress*. We do not show performance in terms of SHD, as this varsortability-baseline is only considering the fully connected graph. While the comparisons are not on the original dataset, we still consider it instructive to observe how well even a crude variance sorting compares to more involved methods. This sorting by variance performs substantially better than a random graph, and plain variance-sorting or *sortnregress* may therefore be much more informative baselines than random graphs.

This begs the question, how much GraN-DAG, or other successful methods, exploit varsortability, and how they perform under data standardization. We encourage such an exploration in future work and suggest that varsortability and (a nonlinear version of) *sortnregress* should always be included in future benchmarks.

Table 2: Naive baseline in non-linear setting. 1000 observations of ER-4 and SF-4 Gauss-ANM 20-node graphs. Comparison to GraN-DAG and random baseline as in [Lachapelle et al., 2019, Table 8, A.6].

graph-type	varsortability	SID variance sorting	SID GraN-DAG	SID random graph
ER-4	0.90	180.20 ± 12.27	165.1 ± 21.0	298.5 ± 28.7
SF-4	0.96	74.00 ± 8.96	62.5 ± 18.8	204.8 ± 38.5

6 Discussion and Conclusion

We find that continuous structure learning methods are highly susceptible to data rescaling and some do not perform well without access to the true data scale. Therefore, scale-variant causal structure learning methods should be applied and benchmarked with caution, especially if the variables do not share a measurement scale or when the true scale of the data is unattainable. It is important to declare whether data is being standardized prior to being fed to various structure learning algorithms.

The drop in performance of *NOTEARS* upon standardizing the data was independently and later reported by Kaiser and Sipos [2021] who present a low-dimensional exemplary case. Beyond a reporting of impaired performance, we provide an extension to score-equivalend methods, exhaustive simulation experiments, and a general explanation for the phenomenon.

Our aim is to raise awareness of the severity with which scaling properties in data from simulated DAGs and causal additive models may distort algorithm performance. Increasing marginal variances can render scenarios identifiable, which may commonly not be expected to be so—for example the Gaussian case with non-equal variances. We therefore argue that varsortability should be taken into account for future benchmarking. Yet, with any synthetic benchmark there remains a risk that the results are not indicative of algorithm performance on real data. Our results indicate that current structure learning algorithms may perform within the range of naive baselines on real-world datasets.

The theoretical results of our paper are limited to the setting of linear ANMs. Our empirical findings indicate that causal discovery benchmarks can be similarly gamed on standardized data and in non-linear settings, but further research is needed to . We focus on a specific subset of algrorithms. The impact of patterns in benchmarking data on a wider class of algorithms and score functions remains to be explored.

Varsortability arises in many additive noise models and the marginal variances increase drastically along the causal order, at least in common simulation settings. This begs the question whether this phenomenon can also be observed in nature, and if not, how applicable the additive noise assumption is in practice. The critical examination of the validity or plausibility of the additive noise assumption for real-world data might help guide future research in causal structure learning.

References

Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse Gaussian Bayesian networks. *Journal of Machine Learning Research*, 16(1):2273–2328, 2015.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439): 509–512, 1999.

Rohit Bhattacharya, Tushar Nagarajan, Daniel Malinsky, and Ilya Shpitser. Differentiable causal discovery under unmeasured confounding. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, 2021.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, pages 21865–21877, 2020.

Peter Bühlmann, Jonas Peters, and Jan Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Wenyu Chen, Mathias Drton, and Y Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 09 2019.

David M. Chickering. A Transformational Characterization of Equivalent Bayesian Network Structures. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995)*, page 87–98, 1995.

David M. Chickering. Learning Bayesian networks is NP-complete. In *Learning from data*, pages 121–130. Springer, 1996.

David M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(11):507–554, 2002.

David M. Chickering, David Heckerman, and Chris Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5, 2004.

Philip A. Dawid. Beware of the DAG! In *Causality: objectives and assessment*, pages 59–86. PMLR, 2010.

Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

Asish Ghoshal and Jean Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems*, volume 30, pages 6457–6466, 2017.

Asish Ghoshal and Jean Honorio. Learning linear structural equation models in polynomial time and sample complexity. In *International Conference on Artificial Intelligence and Statistics*, pages 1466–1475. PMLR, 2018.

Patrik 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*, volume 21, pages 689–696, 2009.

Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

Marcus Kaiser and Maksim Sipos. Unsuitability of NOTEARS for Causal Graph Discovery. *arXiv preprint arXiv:2104.05441*, 2021.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. *arXiv preprint arXiv:1906.02226*, 2019.

Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T. Cherng, and Joel T. Dudley. Scaling structural learning with no-bears to infer causal transcriptome networks. In *Pac Symp Biocomput*. World Scientific, 2019.

Po-Ling Loh and Peter Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1):3065–3105, 2014.

Christopher Meek. Strong Completeness And Faithfulness In Bayesian Networks. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence (UAI1995)*, pages 411–418, August 1995.

Christopher Meek. *Graphical Models: Selecting causal and statistical models*. PhD thesis, Carnegie Mellon University, 1997.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint Causal Inference from Multiple Contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. *Advances in Neural Information Processing Systems*, 33, 2020.

Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pages 1595–1605, 2020.

Gunwoong Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

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

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. The MIT Press, 2017.

Joseph D. Ramsey, Kun Zhang, Madelyn Glymour, Ruben S. Romero, Biwei Huang, Imme Ebert-Uphoff, Savini Samarasinghe, Elizabeth A. Barnes, and Clark Glymour. TETRAD—A toolbox for causal discovery. In *8th international workshop on Climate Informatics (CI 2018)*, 2018.

Kenneth J. Rothman, Sander Greenland, and Timothy L. Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

Jakob Runge, Xavier-Andoni Tibau, Matthias Bruhns, Jordi Muñoz-Marí, and Gustau Camps-Valls. The causality for climate competition. In *NeurIPS 2019 Competition and Demonstration Track*, pages 110–120. PMLR, 2020.

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.

Andrew D. Sanford and Imad A. Moosa. A bayesian network structure for operational risk modelling in structured finance operations. *Journal of the Operational Research Society*, 63(4):431–444, 2012.

Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

Ali Shojaie and George Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 07 2010.

Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*, 2021.

Dennis Wei, Tian Gao, and Yue Yu. DAGs with No Fears: A Closer Look at Continuous Optimization for Learning Bayesian Networks. *arXiv preprint arXiv:2010.09133*, 2020.

Sebastian Weichwald, Martin E. Jakobsen, Phillip B. Mogensen, Lasse Petersen, Nikolaj Thams, and Gherardo Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *NeurIPS 2019 Competition and Demonstration Track*, pages 27–36. PMLR, 2020.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. *arXiv preprint arXiv:1904.10098*, 2019.

Jiji Zhang and Peter Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18 (2):239–271, 2008.

Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Xun Zheng, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep K. Ravikumar, and Eric P. Xing. Learning sparse nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The paper warns against benchmarks that do not transfer to real-world settings and thereby aims to prevent their potentially harmful impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Besides the experimental suite, which will be available on github, we provide relevant code snippets in Appendix E.3 and Appendix F.1 for varsortability and sortnregress, and attach code to reproduce Appendix D.2 in the supplementary zip file.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Table 1 and Appendix I.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] For all results we display variation across several repetitions (box plots, standard error envelopes, min/max spreads.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] Experiments do not depend on HPC clusters.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See e.g. Appendix I.
   (b) Did you mention the license of the assets? [Yes] See Appendix I.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See Appendix E.3 and Appendix F.1 for varsortability and sortnregress.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]