Simple Sorting Criteria Help Find the Causal Order in Additive Noise Models

Alexander Reisach

April 20, 2023

Alexander Reisach
Université Paris Cité

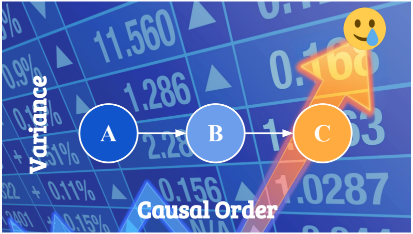
Myriam Tami
Université Paris Saclay

Christof Seiler
Maastricht University

Antoine Chambaz
Université Paris Cité

Sebastian Weichwald
Copenhagen University

Summary

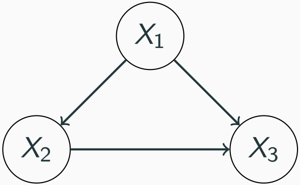


- Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. **“Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy To Game.”** In: Advances in Neural Information Processing Systems. Vol. 34. 2021
- Alexander G. Reisach, Myriam Tami, Christof Seiler, Antoine Chambaz, Sebastian Weichwald. **"Simple Sorting Criteria Help Find the Causal Order in Additive Noise Models."** arXiv preprint arXiv:2303.18211 (2023).

Setting

Structural Causal Models

A structural causal model consists of a graph $\mathcal{G}(V, E)$ encoding dependencies and a set of functions $\mathcal{F}_{\mathcal{G}}$ parametrizing them.



(a) Example graph

$$\begin{array}{cc} & \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \end{array} & \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \end{array}$$

(b) Adjacency matrix A

The Causal Discovery Set-Up

We parametrize the functions $\mathcal{F}_{\mathcal{G}}$ using a linear Additive Noise Model (ANM):

$X_t = f_t\left(\mathrm{Pa}_{\mathcal{G}}(X_t)\right) + \varepsilon_t$ with all ε_t iid and linear f_t.

Therefore, $P(X_1, \ldots, X_d) = \prod_{t=1}^{d} P(X_t \mid \mathrm{Pa}_{\mathcal{G}}(X_t))$.

$$\begin{array}{cc} & \begin{array}{ccc} X_1 & X_2 & X_3 \end{array} \\ \begin{array}{c} X_1 \\ X_2 \\ X_3 \end{array} & \begin{pmatrix} 0 & W_{12} & W_{13} \\ 0 & 0 & W_{23} \\ 0 & 0 & 0 \end{pmatrix} \end{array} \quad \underset{\text{Causal discovery}}{\overset{\text{Data generation}}{\begin{array}{c}\longrightarrow \\ \longleftarrow\end{array}}} \quad \begin{pmatrix} X_1^{(1)} & X_2^{(1)} & X_3^{(1)} \\ \vdots & \vdots & \vdots \\ X_1^{(n)} & X_2^{(n)} & X_3^{(n)} \end{pmatrix}$$

weighted Adjacency matrix W — Observations X

We have that $X = W^{\top} X + \varepsilon$.

Approaches to Causal Discovery

- **Constraint-based methods** Find graph structure by matching conditional independences to the data.
- **Score-based methods** Find graph structure with best goodness-of-fit criterion.
  $\rightarrow$ includes ordering-based search[a]
  1. Find a candidate causal order
  2. Perform sparse regression of each variables onto its predecessors in the order

  (The trendy[b] approach: Differentiable score-based causal discovery. This gave SOTA results on simulated data, even in non-identifiable settings!)

[a] Teyssier and Koller 2005.
[b] Zheng et al. 2018; Vowels, Camgoz, and Bowden 2022.

Methodology

The Elephant in the Room: Simulated Data

We have few high-quality real-world datasets. So when in doubt, we just simulate some data. What's the worst that could happen?

Nodes	30
Graph	Erdős–Rényi
Avg. in-degree	2
Noise	$\mathcal{N}(0, \sigma^2)$
Noise σ	Unif$(\pm(0.5, 2))$

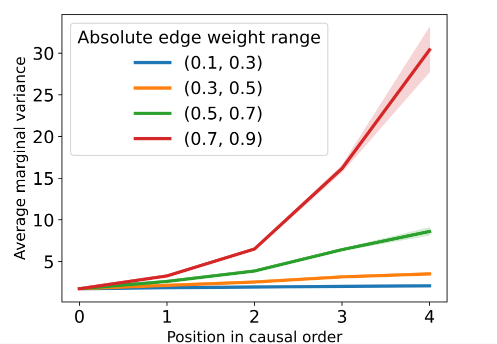


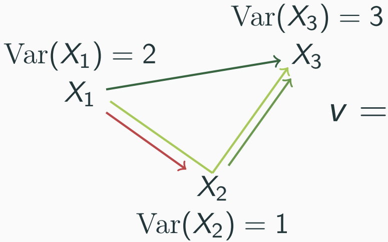
Figure 4: Variance tends to explode

Var-sortability: A Pattern in Causal Discovery Benchmarks

Var-sortability: The fraction of all cause-effect pairs for which the effect has higher variance than the cause.

Definition : Var-sortability

$$\mathsf{Vsb}(A_{\mathcal{G}}) = \frac{\sum_{i=1}^{d-1} \sum_{(s \to t) \in A^i_{\mathcal{G}}} \mathbb{1}_{(\mathrm{Var}(X_s) < \mathrm{Var}(X_t))}}{\sum_{i=1}^{d-1} \sum_{(s \to t) \in A^i_{\mathcal{G}}} 1}$$



$$v = \frac{1+1+1}{1+1+1+1} = \frac{3}{4}$$

Exploiting Var-sortability

We design two simple benchmark algorithms:

SortnRegress (a diagnostic tool for var-sortability)

1. Sort variables by increasing variance
2. Perform sparse regression of each node onto on all its predecessors

MSE-GDS ("Mean-Squared-Error Greedy DAG Search" - to show how MSE effectively sorts by variance)

1. Add the edge that reduces the total MSE the most
2. Stop when no more edges can be added, or no more improvement possible

Causal Discovery Benchmarks Are Easy to Game

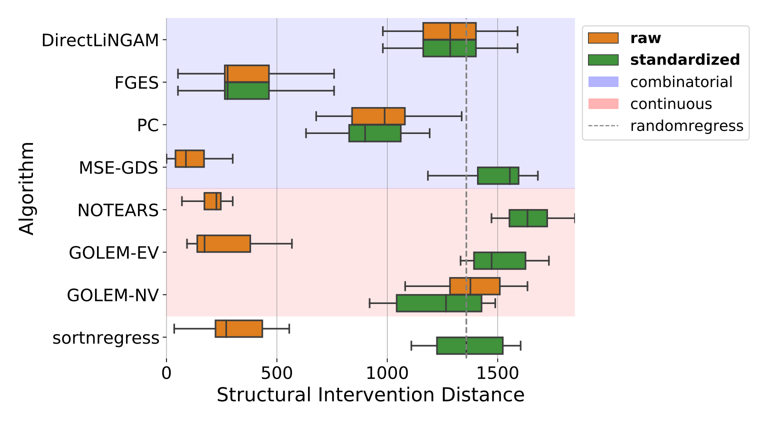


Figure 5: SID[a] on 30 Erdős–Rényi graphs with 50 nodes and Gaussian noise

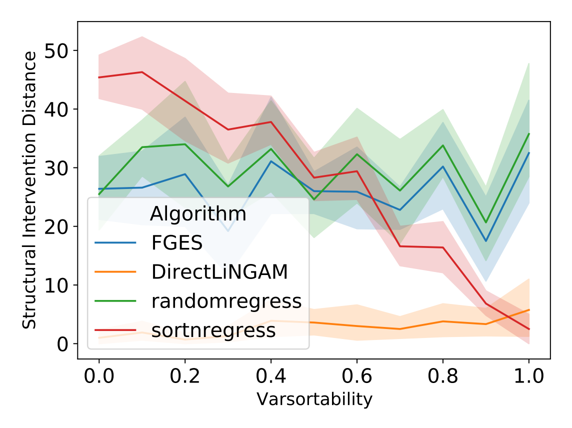
In common benchmarks, var-sortability is usually very high (0.9 to 1 for linear functions, 0.7 to 1 for non-linear)

[a]Peters and Bühlmann 2015.

What To Do With Var-sortability

High var-sortability make causal discovery very easy. Can this be realistic?

Standardization seems to offer a simple solution - but which var-sortability values should we expect in real-world data?



Sorting Our Way to Success

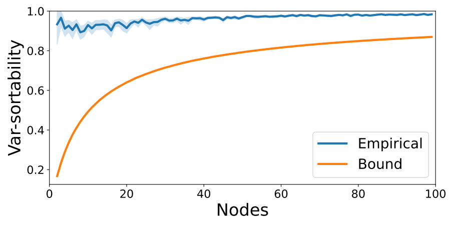
From Var-sortability to R^2-sortability

Drivers of Var-sortability (In Chain Graphs)

In a chain with weighted adjacency matrix W and noise standard deviations σ, the var-sortability between X_0 and X_p can be bounded as

$$P_{W,\sigma}\left(\mathrm{Var}(X_0) < \mathrm{Var}(X_p)\right) \geq \mathbb{P}\left(0 < \sum_{s=0}^{p-1} \ln|W_{s,s+1}|\right).$$

If $\mathbb{E}\left[\ln|P_W|\right] > 0$, this formulation can be transformed into a bound that only depends on the weight distribution:



Cause-explained Variance

An increase in total variance while noise variance are iid implies an increase (in expectation) in the fraction of inherited variance. We can not compute this quantity directly. But we can compute an upper bound given as

$$1 - \frac{\mathrm{Var}(X_t - E[X_t \mid X_S])}{\mathrm{Var}(X_t)}$$

where S is the set $\{1, \ldots, d\} \setminus \{t\}$.

In practice, we can simply compute the R^2 of a model $M^{\theta}_{t,S}(X_S)\colon \mathbb{R}^{|S|} \to \mathbb{R}, X_S \mapsto \theta^\top X_S$ which performs regression of X_t onto X_S with parameters $\theta \in \mathbb{R}^{|S|}$.

Generalizing Sortability and R^2-sortability

We propose a familiy of sortabilities for different criteria τ:

$$\mathrm{v}_\tau(X, \mathcal{G}) = \frac{\sum_{i=1}^{d} \sum_{(s \to t) \in A^i_{\mathcal{G}}} \mathbb{1}_{(\tau(X,s) < \tau(X,t))}}{\sum_{i=1}^{d} \sum_{(s \to t) \in A^i_{\mathcal{G}}} 1}.$$

We obtain the previously discussed var-sortability for $\tau(X, t) = \mathrm{Var}(X_t)$ and denote it as $\mathrm{v}_{\mathrm{Var}}$.

We newly introduce R^2-sortability for $\tau(X, t) = R^2(M^{\theta^*}_{t,S}, X)$ and denote it as v_{R^2}.

R^2-SortnRegress

1. Obtain a R^2 value for each variable given all others
2. Sort by increasing R^2
3. Perform sparse regression of each node onto all its predecessors

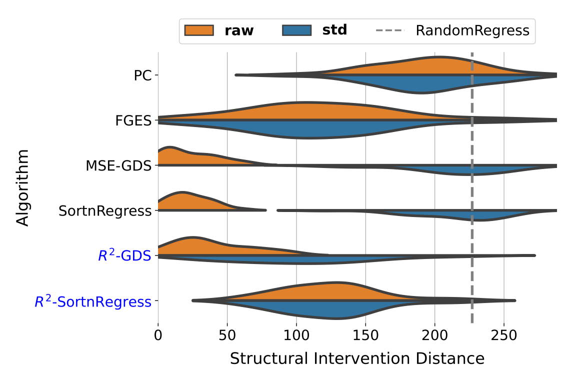
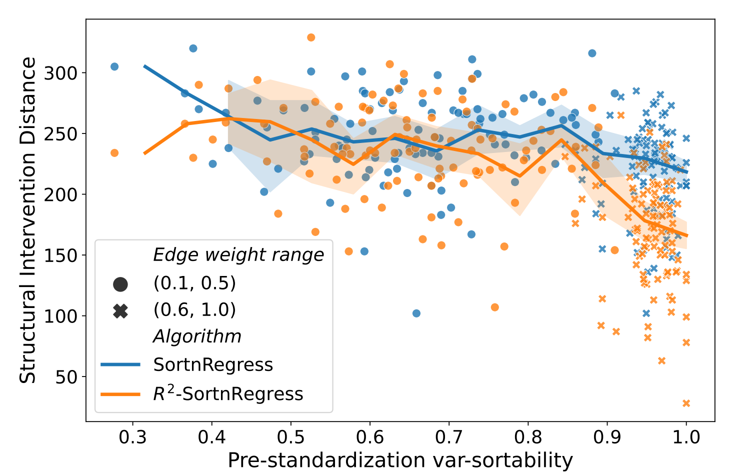


Figure 6: 30 Erdős–Rényi graphs, 20 nodes, Gaussian noise

Exploiting R^2-sortability on Standardized Data

Exploiting R^2-sortability is not as effective as exploiting var-sortability, but it does not require knowledge of the "true" data scale. An example on standardized data:

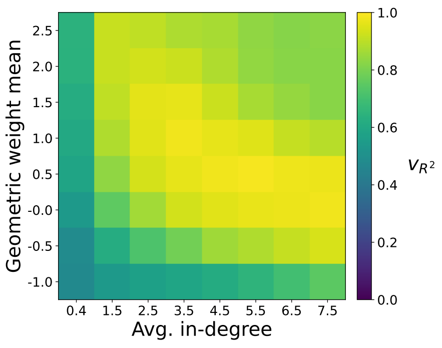


Why R^2-sortability?

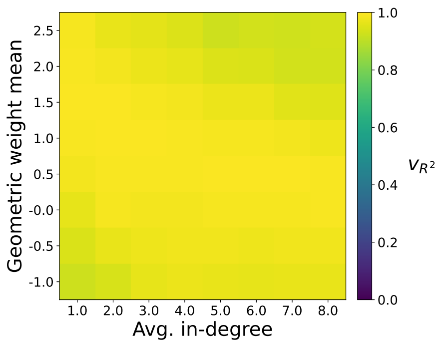
It is scale-invariant.

Realistic Values?

All parameters can affect R^2-sortability. The weight distribution and the connectivity (average in-degree) have a big effect. In Scale-free graphs, R^2-sortability is extremely high across a wide range of settings.



(a) Erdős–Rényi graphs



(b) Scale-free graphs

Takeaways

We make assumptions about *some* properties of ANM, but need to choose values for *all* properties of ANMs in simulations. In doing so, we may introduce patterns that are in effect additional assumptions.

R^2-sortability can help, because

- It provides a simple measure for one such simulation pattern.
- It is scale-invariant and can thus be assessed on real-world data, allowing us to match simulation values to real values.

References

[1] Jonas Peters and Peter Bühlmann. **"Structural Intervention Distance for Evaluating Causal Graphs".** In: *Neural computation* 27.3 (2015), pp. 771–799.

[2] Alexander G. Reisach, Christof Seiler, and Sebastian Weichwald. **"Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy To Game".** In: *Advances in Neural Information Processing Systems*. Vol. 34. 2021, pp. 27772–27784.

[3] Peter Spirtes and Clark Glymour. **"An Algorithm for Fast Recovery of Sparse Causal Graphs".** In: *Social Science Computer Review* 9.1 (1991), pp. 62–72.

[4] Marc Teyssier and Daphne Koller. **"Ordering-Based Search: A Simple and Effective Algorithm for Learning Bayesian Networks".** In: *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2005, pp. 584–590.

[5] Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. **"D'Ya Like DAGs? A Survey on Structure Learning and Causal Discovery".** In: *ACM Computing Surveys* 55.4 (2022).

[6] Xun Zheng et al. **"DAGs with NO TEARS: Continuous Optimization for Structure Learning".** In: *Advances in Neural Information Processing Systems*. Vol. 32. 2018, pp. 9472–9483.

Var-sortability and R^2-sortability

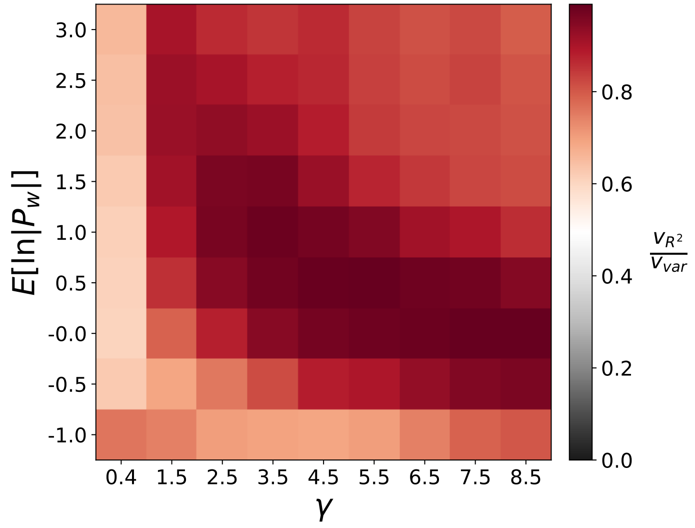


Figure 8: Alignment between var-sortability and R^2-sortability