Sortability in Structural Causal Models

Alexander Reisach
Université Paris Cité

Myriam Tami
Université Paris Saclay

Christof Seiler
Maastricht University

Antoine Chambaz
Université Paris Cité

Sebastian Weichwald
Copenhagen University

Workshop on Graphical Models and Clustering
Montpellier, May 17, 2024

Published Works

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

Alexander G. Reisach[1,2] **Christof Seiler**[2,3] **Sebastian Weichwald**[1]

[1]Department of Mathematical Sciences, University of Copenhagen, Denmark
[2]Department of Data Science and Knowledge Engineering, Maastricht University, The Netherlands
[3]Mathematics Centre Maastricht, Maastricht University, The Netherlands

Abstract

Simulated DAG models may exhibit properties that, perhaps 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 commonly sampled graphs and model parameters, 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 moderate or dependent on the choice of measurement scales. On standardized data, the same algorithms fail to 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. The code is available at `https://github.com/Scriddie/Varsortability`.

A Scale-Invariant Sorting Criterion to Find a Causal Order in Additive Noise Models

Alexander G. Reisach*
CNRS, MAP5
Université Paris Cité
F-75006 Paris, France

Myriam Tami
CentraleSupélec
Université Paris-Saclay
F-91190 Gif-sur-Yvette, France

Christof Seiler
Department of Advanced
Computing Sciences, Maastricht
University, The Netherlands

Antoine Chambaz
CNRS, MAP5
Université Paris Cité
F-75006 Paris, France

Sebastian Weichwald
Department of Mathematical Sciences
and Pioneer Centre for AI,
University of Copenhagen, Denmark

Abstract

Additive Noise Models (ANMs) are a common model class for causal discovery from observational data. Due to a lack of real-world data for which an underlying ANM is known, ANMs with randomly sampled parameters are commonly used to simulate data for the evaluation of causal discovery algorithms. While some parameters may be fixed by explicit assumptions, fully specifying an ANM requires choosing all parameters. Reisach et al. (2021) show that, for many ANM parameter choices, sorting the variables by increasing variance yields an ordering close to a causal order and introduce 'var-sortability' to quantify this alignment. Since increasing variances may be unrealistic and cannot be exploited when data scales are arbitrary, ANM data are often rescaled to unit variance in causal discovery benchmarking.

We show that synthetic ANM data are characterized by another pattern that is scale-invariant and thus persists even after standardization: the explainable fraction of a variable's variance, as captured by the coefficient of determination R^2, tends to increase along the causal order. The result is high 'R^2-sortability', meaning that sorting the variables by increasing R^2 yields an ordering close to a causal order. We propose a computationally efficient baseline algorithm termed 'R^2-SortnRegress' that exploits high R^2-sortability and that can match and exceed the performance of established causal discovery algorithms. We show analytically that sufficiently high edge weights lead to a relative decrease of the noise contributions along causal chains, resulting in increasingly deterministic relationships and high R^2. We characterize R^2-sortability on synthetic data with different simulation parameters and find high values in common settings. Our findings reveal high R^2-sortability as an assumption about the data generating process relevant to causal discovery and implicit in many ANM sampling schemes. It should be made explicit, as its prevalence in real-world data is an open question. For causal discovery benchmarking, we provide implementations of R^2-sortability, the R^2-SortnRegress algorithm, and ANM simulation procedures in our library CausalDisco.

Section Overview

Causality

Take two continuous random variables X and Y with joint distribution $P(X, Y)$. The joint can be factorized into

$$P(Y|X)P(X) \quad \text{or} \quad P(X|Y)(Y). \tag{1}$$

Causality

Take two continuous random variables X and Y with joint distribution $P(X, Y)$. The joint can be factorized into

$$P(Y|X)P(X) \quad \text{or} \quad P(X|Y)(Y). \tag{1}$$

We may want to predict one of the variables given the other. Depending on which is our target, we may be interested in the conditional expectation

$$\mathbb{E}[Y \mid X] \quad \text{or} \quad \mathbb{E}[X \mid Y]. \tag{2}$$

Causality

Take two continuous random variables X and Y with joint distribution $P(X, Y)$. The joint can be factorized into

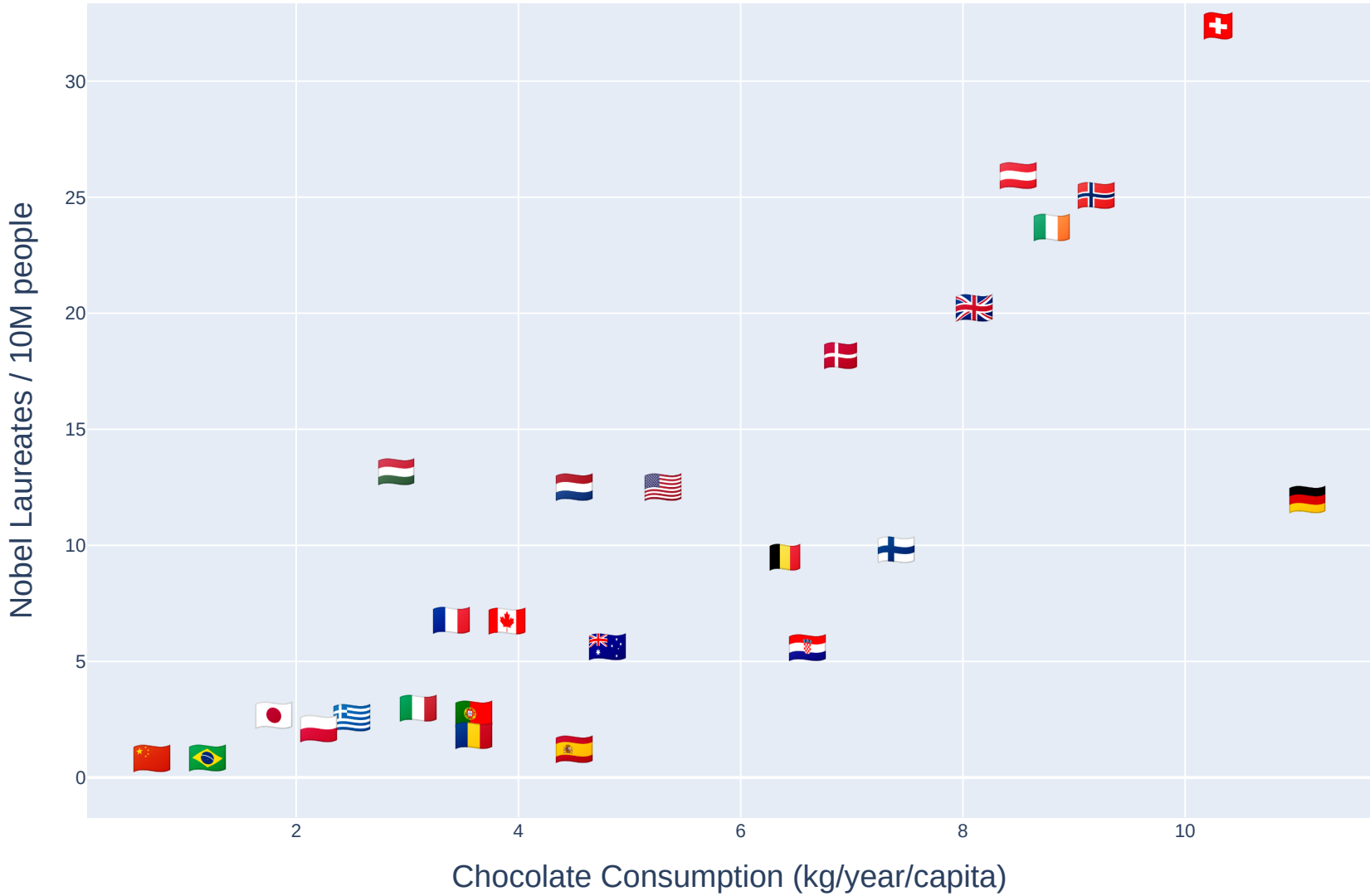
$$P(Y|X)P(X) \quad \text{or} \quad P(X|Y)(Y). \tag{1}$$

We may want to predict one of the variables given the other. Depending on which is our target, we may be interested in the conditional expectation

$$\mathbb{E}[Y \mid X] \quad \text{or} \quad \mathbb{E}[X \mid Y]. \tag{2}$$

For prediction, either option can be useful. However, this may not be the case when we are interested in the effect of changing the variables.

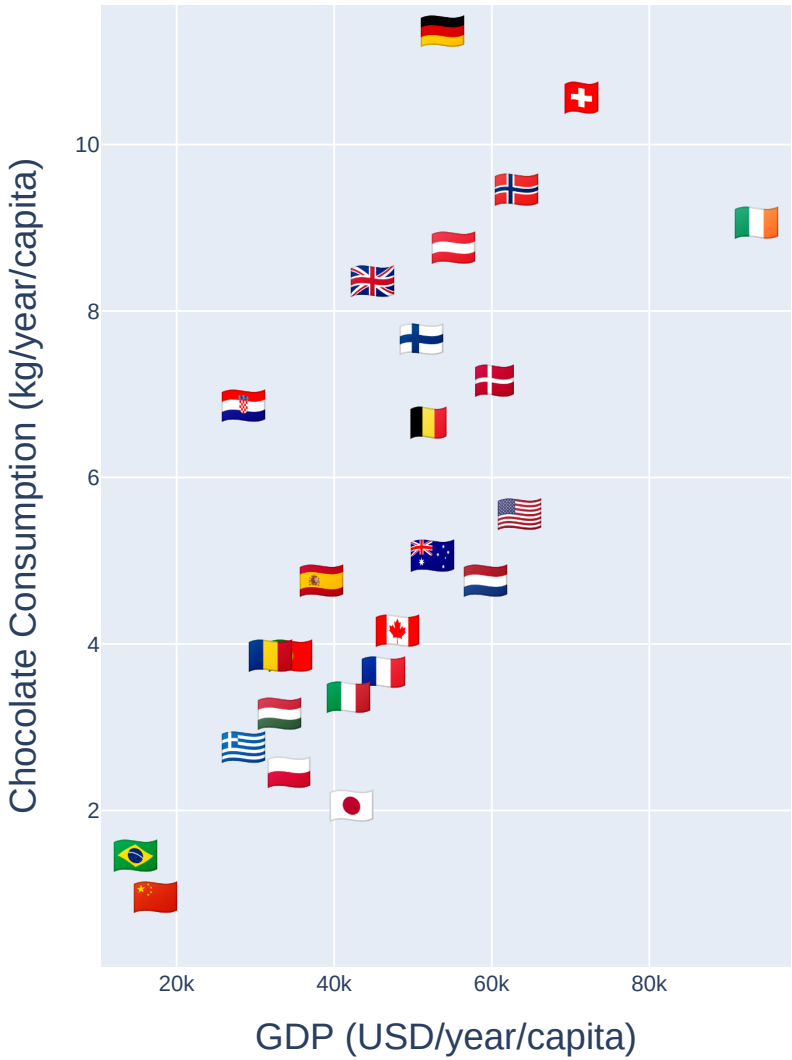
Correlation vs. Causation

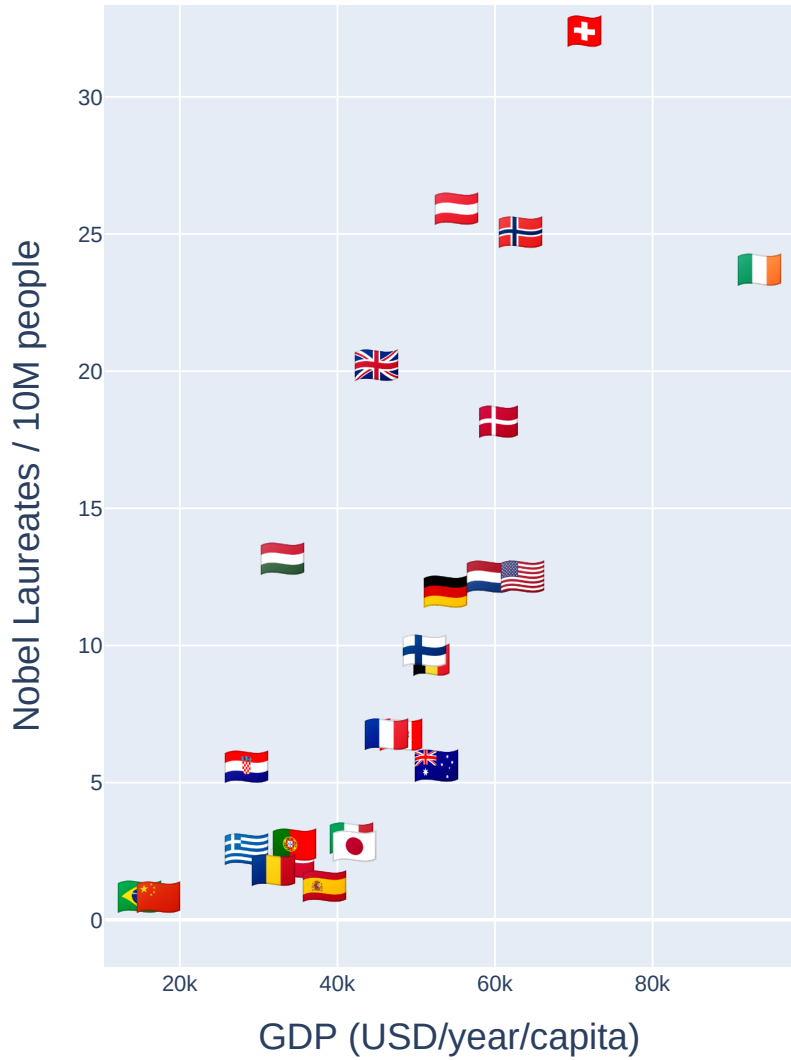


1

[1] Messerli 2012.

Correlation vs. Causation





Structural Equation Models

The data generating process may be given by the structural equations

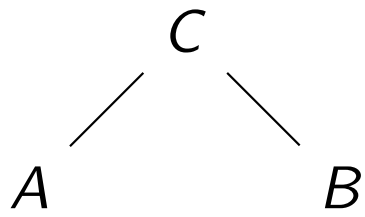
$$\begin{aligned} C &:= N_C && \text{(GDP)} \\ A &:= f(C) + N_A && \text{(chocolate)} \\ B &:= f(C) + N_B && \text{(nobel laureates)}, \end{aligned}$$

where N_A, N_B, N_C are mutually independent noise random variables.

Changing the value of a variable (replacing its structural equation with a constant) is called an intervention. In this case, changing A has no effect on B, because A is not a cause of B.

Structural Equations as Graphs

We can draw the relationships between variables in a graph - note that C separates[2] A and B.



We write the independence between A and B given C as

$$A \perp B | C.$$

[2] Removing C and all adjacent edges removes all paths between X and Y.

Graphical Models

Graphical models are independence models (adhering to the 5 graphoid axioms[3]). Let $\mathcal{G} = (V, E)$ be a graph with nodes V and edges[4] E.

[3] Lauritzen 1996, Section 2.5.1.

[4] undirected, no self-loops, and no multiple edges.

Graphical Models

Graphical models are independence models (adhering to the 5 graphoid axioms[3]). Let $\mathcal{G} = (V, E)$ be a graph with nodes V and edges[4] E.

Let A, B, C, D be disjoint subsets of V.

$$(S1) \quad A \perp B | C \implies B \perp A | C \qquad \text{(symmetry)}$$

$$(S2) \quad A \perp B \cup D | C \implies A \perp B | C \text{ and } A \perp D | C \qquad \text{(decomposition)}$$

$$\vdots$$

[3] Lauritzen 1996, Section 2.5.1.

[4] undirected, no self-loops, and no multiple edges.

Graphical Models

Graphical models are independence models (adhering to the 5 graphoid axioms[3]). Let $\mathcal{G} = (V, E)$ be a graph with nodes V and edges[4] E.

Let A, B, C, D be disjoint subsets of V.

$$(S1) \quad A \perp B | C \implies B \perp A | C \qquad \text{(symmetry)}$$

$$(S2) \quad A \perp B \cup D | C \implies A \perp B | C \text{ and } A \perp D | C \qquad \text{(decomposition)}$$

$$\vdots$$

These are very general axioms, which allow us to use graphical models to represent a variety of independence relationships.

[3] Lauritzen 1996, Section 2.5.1.

[4] undirected, no self-loops, and no multiple edges.

Probabilistic Graphical Models

Let $\mathcal{G} = (V, E)$ be our graph. Let X be a set of random variables[5] $\{X_v\}_{v \in V}$ with joint distribution $P(X)$.

[5] think of the vertices $V = \{V_1, \ldots, V_d\}$ corresponding to $\{X_1, \ldots, X_d\}$

Probabilistic Graphical Models

Let $\mathcal{G} = (V, E)$ be our graph. Let X be a set of random variables[5] $\{X_v\}_{v \in V}$ with joint distribution $P(X)$.

$P(X)$ is Markov w.r.t. $\mathcal{G}$ if for all $X_1, X_2, X_3 \subset X$ disjoint,

$$X_1 \perp X_3 \mid X_2 \implies X_1 \perp\!\!\!\perp X_3 \mid X_2. \tag{3}$$

[5] think of the vertices $V = \{V_1, \ldots, V_d\}$ corresponding to $\{X_1, \ldots, X_d\}$

Probabilistic Graphical Models

Let $\mathcal{G} = (V, E)$ be our graph. Let X be a set of random variables[5] $\{X_v\}_{v \in V}$ with joint distribution $P(X)$.

$P(X)$ is Markov w.r.t. $\mathcal{G}$ if for all $X_1, X_2, X_3 \subset X$ disjoint,

$$X_1 \perp X_3 \mid X_2 \implies X_1 \perp\!\!\!\perp X_3 \mid X_2. \tag{3}$$

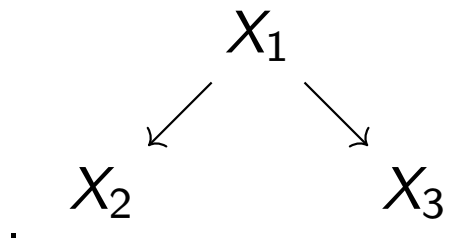
Note: the other direction of eq. (3) does not hold in general.[6]

[5] think of the vertices $V = \{V_1, \ldots, V_d\}$ corresponding to $\{X_1, \ldots, X_d\}$

[6] It would be very convenient, so it is often assumed that it does.

Causal Graphical Models

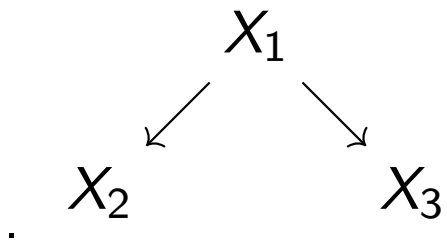
Causal graphical models use directed acyclic graphs (DAGs) to reason about interventions[7].



[7] recall our example about GDP, chocolate consumption, and nobel prizes.

Causal Graphical Models

Causal graphical models use directed acyclic graphs (DAGs) to reason about interventions[7].



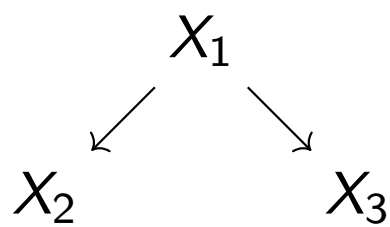
If $P(X)$ is Markov w.r.t. $\mathcal{G}$, we have that

$$P(X) = \prod_{i=1}^{d} P(X_i | \text{Pa}_{\mathcal{G}}(Xi)).$$

[7] recall our example about GDP, chocolate consumption, and nobel prizes.

Structural Causal Models (SCMs)

Structural causal models combine a causal graphical model with a set of structural equations, and a corresponding.
Given $X = \{X_1, X_2, X_3\}$ and independent noise variables N_1, N_2, N_3, consider for example the following model:



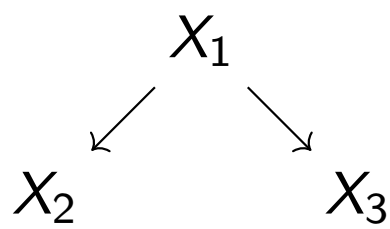
(a) Corresponding DAG G

$$X_1 := f_1(N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(X_1, N_3)$$

(b) Structural equations

Structural Causal Models (SCMs)

Structural causal models combine a causal graphical model with a set of structural equations, and a corresponding.
Given $X = \{X_1, X_2, X_3\}$ and independent noise variables N_1, N_2, N_3, consider for example the following model:



(a) Corresponding DAG G

$$X_1 := f_1(N_1)$$
$$X_2 := f_2(X_1, N_2)$$
$$X_3 := f_3(X_1, N_3)$$

(b) Structural equations

In SCMs, we can compute the effect of interventions using the structural equations, and we can compute graphical properties using the DAG.

Section Overview

Causal Discovery

The goal of causal discovery is to learn a structural causal model from data. This requires either

- interventional data (e.g. randomized control trials) and strong assumptions, or
- observational data and extremely strong assumptions.

Causal Discovery on Observational data

Two main approaches:

- **Constraint-based methods:** perform (conditional) independence testing to narrow down the graph structure. (Yields partially directed graphs.)
- **Score-based methods:** optimize a score criterion (e.g. the likelihood) to find the best an estimated graph. (Yields DAGs!)

Additive Noise Models (ANMs)

Additive noise models encode a popular functional assumption that allows learning causal structure from observational data using score-based methods.[8]

[8] Peters et al. 2011.

Additive Noise Models (ANMs)

Additive noise models encode a popular functional assumption that allows learning causal structure from observational data using score-based methods.[8]

We consider **linear ANMs** of the form

$$X = W^\top X + N, \text{ where}$$

- $X = (X_1, \ldots, X_d)^\top$ are random variables
- $W \in \mathbb{R}^{d \times d}$ is a weighted adjacency matrix
- $N = (N_1, \ldots, N_d)^\top$ is a vector of independent noise variables.

[8] Peters et al. 2011.

Additive Noise Models (ANMs)

Additive noise models encode a popular functional assumption that allows learning causal structure from observational data using score-based methods.[8]

We consider **linear ANMs** of the form

$$X = W^\top X + N, \text{ where}$$

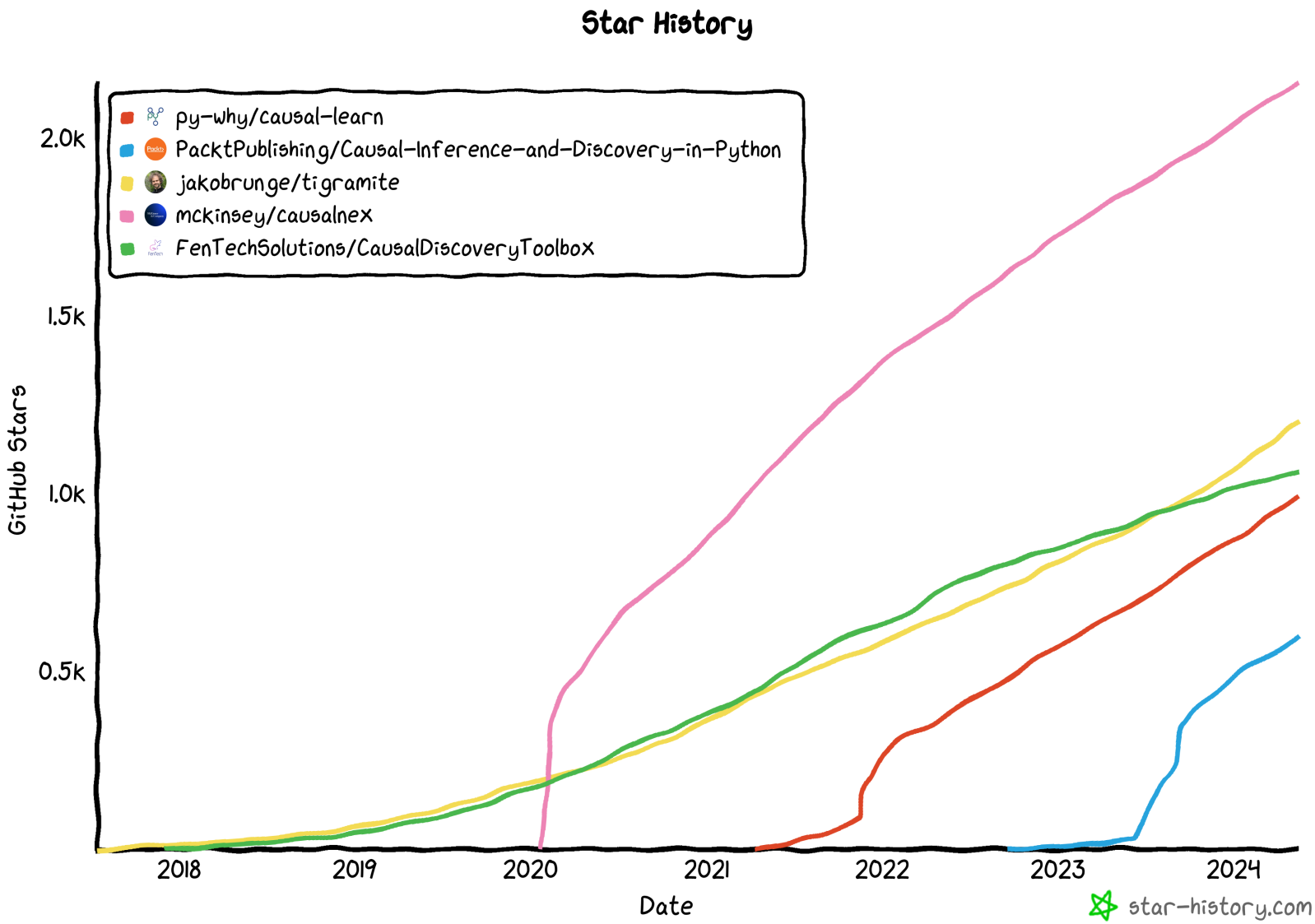
- $X = (X_1, \ldots, X_d)^\top$ are random variables
- $W \in \mathbb{R}^{d \times d}$ is a weighted adjacency matrix
- $N = (N_1, \ldots, N_d)^\top$ is a vector of independent noise variables.

The goal is to learn W (the causal DAG and weights) from observations of X.

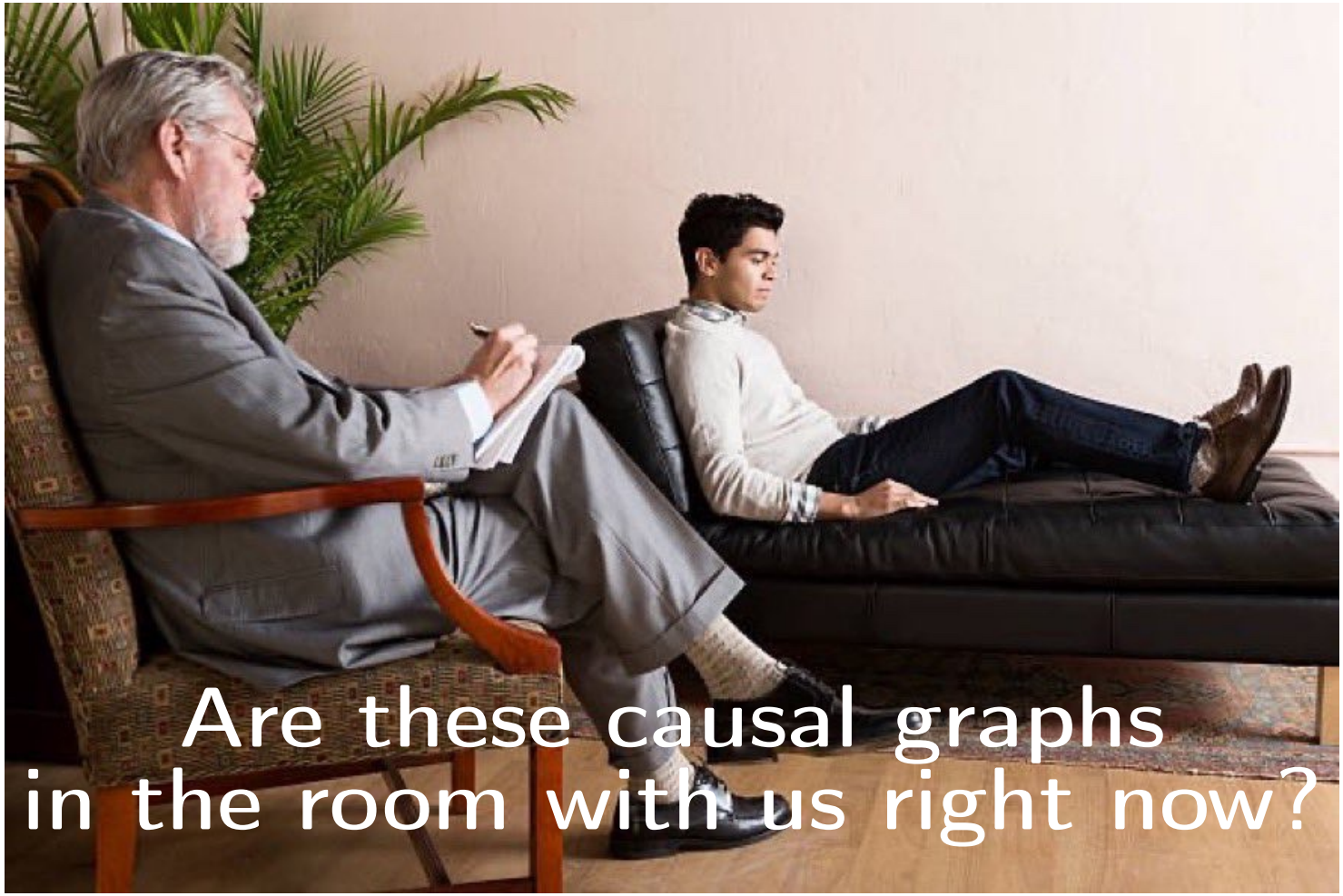
[8] Peters et al. 2011.

Section Overview

A Glimpse at the Causal Discovery Ecosystem



Lack of Ground Truth



How (Not?) to Evaluate Causal Discovery Algorithms

If we have no data, let us simulate some! In the case of ANMs:

1. Choose the number of variables d.
2. Determine a connectivity parameter γ.
3. Draw random graphs from a distribution $P_{\mathcal{G}}$.
4. Draw edge weights from a distribution P_W.
5. Draw noise standard deviations from P_σ.
6. Draw noise from a distribution $\mathcal{P}_N(\sigma)$.

How (Not?) to Evaluate Causal Discovery Algorithms

If we have no data, let us simulate some! In the case of ANMs:

1. Choose the number of variables d.
2. Determine a connectivity parameter γ.
3. Draw random graphs from a distribution $P_{\mathcal{G}}$.
4. Draw edge weights from a distribution P_W.
5. Draw noise standard deviations from P_σ.
6. Draw noise from a distribution $\mathcal{P}_N(\sigma)$.

But what parameters should we choose?

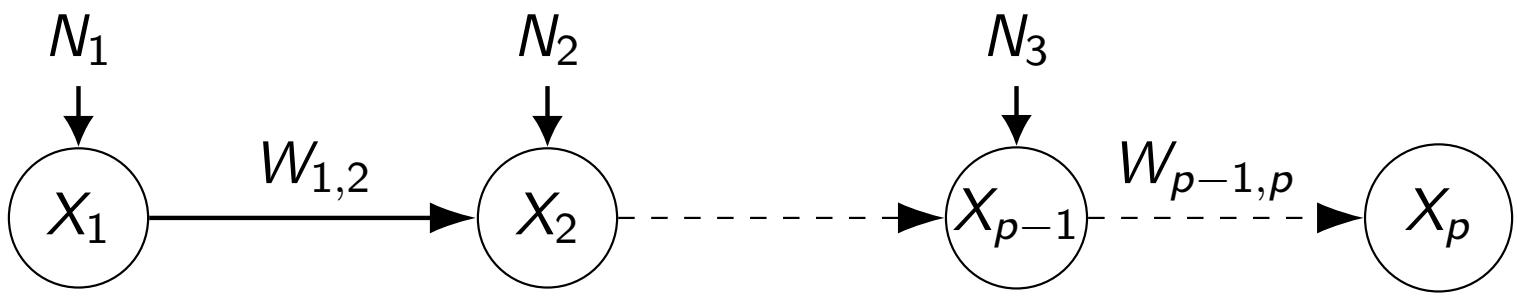
Section Overview

Parameter Choices in Causal Chains

Consider a causal chain of the form

$$
\begin{aligned}
X_1 &:= N_1 \\
X_2 &:= W_{1,2} X_1 + N_2 \\
&\dots \\
X_p &:= W_{p-1,p} X_{p-1} + N_p,
\end{aligned}
$$

corresponding to the chain DAG



What Happens Along the Chain?

Consider a causal chain $(X_0 \xrightarrow{w_{0,1}} X_1 \xrightarrow{w_{1,2}} X_2 \xrightarrow{w_{2,3}} \cdots \xrightarrow{w_{p-1,p}} X_p)$ of length $p > 0$ with *fixed* edge weights $w_{j,j+1}$ for $j = 0, ..., p-1$ and *fixed* standard deviations σ_j for $j = 0, ..., p$ of the independent noise variables $N_0, ..., N_p$.

What Happens Along the Chain?

Consider a causal chain $(X_0 \xrightarrow{w_{0,1}} X_1 \xrightarrow{w_{1,2}} X_2 \xrightarrow{w_{2,3}} \cdots \xrightarrow{w_{p-1,p}} X_p)$ of length $p > 0$ with *fixed* edge weights $w_{j,j+1}$ for $j = 0, ..., p-1$ and *fixed* standard deviations σ_j for $j = 0, ..., p$ of the independent noise variables $N_0, ..., N_p$.

$$\mathrm{Var}(X_p) = \mathrm{Var}(w_{p-1,p}\, X_{p-1}) + \mathrm{Var}(N_p)$$

What Happens Along the Chain?

Consider a causal chain $(X_0 \xrightarrow{w_{0,1}} X_1 \xrightarrow{w_{1,2}} X_2 \xrightarrow{w_{2,3}} \cdots \xrightarrow{w_{p-1,p}} X_p)$ of length $p > 0$ with *fixed* edge weights $w_{j,j+1}$ for $j = 0, ..., p-1$ and *fixed* standard deviations σ_j for $j = 0, ..., p$ of the independent noise variables $N_0, ..., N_p$.

$$\begin{aligned} \mathrm{Var}(X_p) &= \mathrm{Var}(w_{p-1,p}\, X_{p-1}) + \mathrm{Var}(N_p) \\ &= \cdots (\text{unfolding the recursion}) \cdots \end{aligned}$$

What Happens Along the Chain?

Consider a causal chain $(X_0 \xrightarrow{w_{0,1}} X_1 \xrightarrow{w_{1,2}} X_2 \xrightarrow{w_{2,3}} \cdots \xrightarrow{w_{p-1,p}} X_p)$ of length $p > 0$ with *fixed* edge weights $w_{j,j+1}$ for $j = 0, ..., p-1$ and *fixed* standard deviations σ_j for $j = 0, ..., p$ of the independent noise variables $N_0, ..., N_p$.

$$\begin{aligned}
\mathrm{Var}(X_p) &= \mathrm{Var}(w_{p-1,p}\, X_{p-1}) + \mathrm{Var}(N_p) \\
&= \cdots (\text{unfolding the recursion}) \cdots \\
&= \sum_{i=0}^{p-1} \sigma_i^2 \left(\prod_{j=i}^{p-1} w_{j,j+1} \right)^2 + \sigma_p^2
\end{aligned}$$

What Happens Along the Chain?

Consider a causal chain $(X_0 \xrightarrow{w_{0,1}} X_1 \xrightarrow{w_{1,2}} X_2 \xrightarrow{w_{2,3}} \cdots \xrightarrow{w_{p-1,p}} X_p)$ of length $p > 0$ with *fixed* edge weights $w_{j,j+1}$ for $j = 0, ..., p-1$ and *fixed* standard deviations σ_j for $j = 0, ..., p$ of the independent noise variables $N_0, ..., N_p$.

$$\begin{aligned}
\mathrm{Var}(X_p) &= \mathrm{Var}(w_{p-1,p}\, X_{p-1}) + \mathrm{Var}(N_p) \\
&= \cdots (\text{unfolding the recursion}) \cdots \\
&= \sum_{i=0}^{p-1} \sigma_i^2 \left(\prod_{j=i}^{p-1} w_{j,j+1} \right)^2 + \sigma_p^2 \\
&\geq \sigma_0^2 \prod_{j=0}^{p-1} w_{j,j+1}^2
\end{aligned}$$

What Happens Along the Chain?

Consider a causal chain $(X_0 \xrightarrow{w_{0,1}} X_1 \xrightarrow{w_{1,2}} X_2 \xrightarrow{w_{2,3}} \cdots \xrightarrow{w_{p-1,p}} X_p)$ of length $p > 0$ with *fixed* edge weights $w_{j,j+1}$ for $j = 0, ..., p-1$ and *fixed* standard deviations σ_j for $j = 0, ..., p$ of the independent noise variables $N_0, ..., N_p$.

$$
\begin{aligned}
\mathrm{Var}(X_p) &= \mathrm{Var}(w_{p-1,p}\, X_{p-1}) + \mathrm{Var}(N_p) \\
&= \cdots (\text{unfolding the recursion}) \cdots \\
&= \sum_{i=0}^{p-1} \sigma_i^2 \left(\prod_{j=i}^{p-1} w_{j,j+1} \right)^2 + \sigma_p^2 \\
&\geq \sigma_0^2 \prod_{j=0}^{p-1} w_{j,j+1}^2 \\
&\geq \sigma_0^2 \sum_{j=0}^{p-1} \log |w_{j,j+1}|.
\end{aligned}
$$

A Sufficient Criterion for Diverging Variances in Chains

Let now $w_{j,j+1}$ be drawn from the distribution of edge weights P_W and let σ_0 have bounded positive support.

We have by the strong law of large numbers that

$$\sigma_0^2 \sum_{j=0}^{p-1} \log |w_{j,j+1}| \xrightarrow[p\to\infty]{\text{a.s.}} +\infty$$

given

$$0 < \mathbb{E}\left[\log |V|\right] < +\infty, \text{ with } V \sim P_W, \tag{4}$$

since the $w_{j,j+1}$ are sampled iid from P_W.

Illustration

Given sufficiently large weights in W, the variance tends to increase along causal chains.

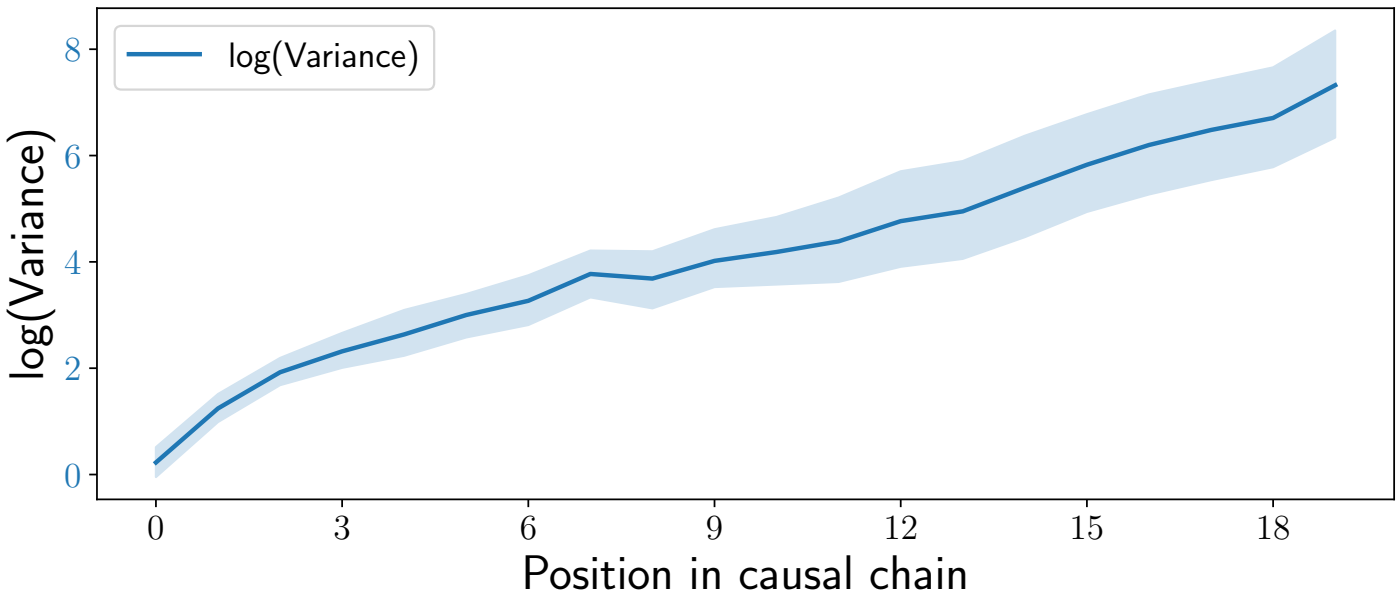


Figure: Causal chains with weights drawn from Unif$(0.5, 2)$ and Gaussian noise with standard deviations drawn from Unif$(0.5, 2)$; 30 chains simulated independently.

(For a $V \sim \text{Unif}(0.5, 2)$, we have $\mathbb{E}[\log|V|] \approx 0.16$.)

Exploiting The Variance Pattern For Causal Discovery

Var-SortnRegress – a simple causal discovery algorithm

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

Exploiting The Variance Pattern For Causal Discovery

Var-SortnRegress – a simple causal discovery algorithm

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

$\rightarrow$ performs extremely well.[9] In fact, it performs too well. Consider:

- causal discovery is a hard problem across many sciences.
- the variance depends on arbitrary measurement units.

[9] Reisach, Seiler, and Weichwald 2021

Exploiting The Variance Pattern For Causal Discovery

Var-SortnRegress – a simple causal discovery algorithm

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

$\rightarrow$ performs extremely well.[9] In fact, it performs too well. Consider:

- causal discovery is a hard problem across many sciences.
- the variance depends on arbitrary measurement units.

[9] Reisach, Seiler, and Weichwald 2021

But is Standardization Enough?

As before, we note that in a chain we have that

$$\mathrm{Var}(X_j) = \mathrm{Var}(w_{j-1,j}\, X_{j-1}) + \sigma_j^2.$$

But is Standardization Enough?

As before, we note that in a chain we have that

$$\mathrm{Var}(X_j) = \mathrm{Var}(w_{j-1,j}\, X_{j-1}) + \sigma_j^2.$$

The fraction of the variance due to the cause, which we call the fraction of cause-explained variance (CEV), is given as

$$\frac{\mathrm{Var}(w_{j-1,j}\, X_{j-1})}{\mathrm{Var}(w_{j-1,j}\, X_{j-1}) + \sigma_j^2}. \tag{5}$$

But is Standardization Enough?

As before, we note that in a chain we have that

$$\mathrm{Var}(X_j) = \mathrm{Var}(w_{j-1,j}\, X_{j-1}) + \sigma_j^2.$$

The fraction of the variance due to the cause, which we call the fraction of cause-explained variance (CEV), is given as

$$\frac{\mathrm{Var}(w_{j-1,j}\, X_{j-1})}{\mathrm{Var}(w_{j-1,j}\, X_{j-1}) + \sigma_j^2}. \tag{5}$$

Idea: for diverging variances and iid noise variances with bounded support, eq. (5) must converge to 1.
Problem: we cannot estimate eq. (5) without knowing the causal structure.

R^2 as a Scale-Invariant Sorting Criterion

We can use the fraction of explainable variance given all other variables (not just the cause) as an upper bound.[10]

It is captured by the coefficient of determination

$$R^2(X_j) = 1 - \frac{\mathrm{Var}(X_j - \mathbb{E}(X_j \mid X_{\{1\ldots d\}\setminus\{j\}}))}{\mathrm{Var}(X_j)}.$$

[10] Reisach, Tami, et al. 2023.

R^2 as a Scale-Invariant Sorting Criterion

We can use the fraction of explainable variance given all other variables (not just the cause) as an upper bound.[10]

It is captured by the coefficient of determination

$$R^2(X_j) = 1 - \frac{\mathrm{Var}(X_j - \mathbb{E}(X_j \mid X_{\{1\ldots d\}\setminus\{j\}}))}{\mathrm{Var}(X_j)}.$$

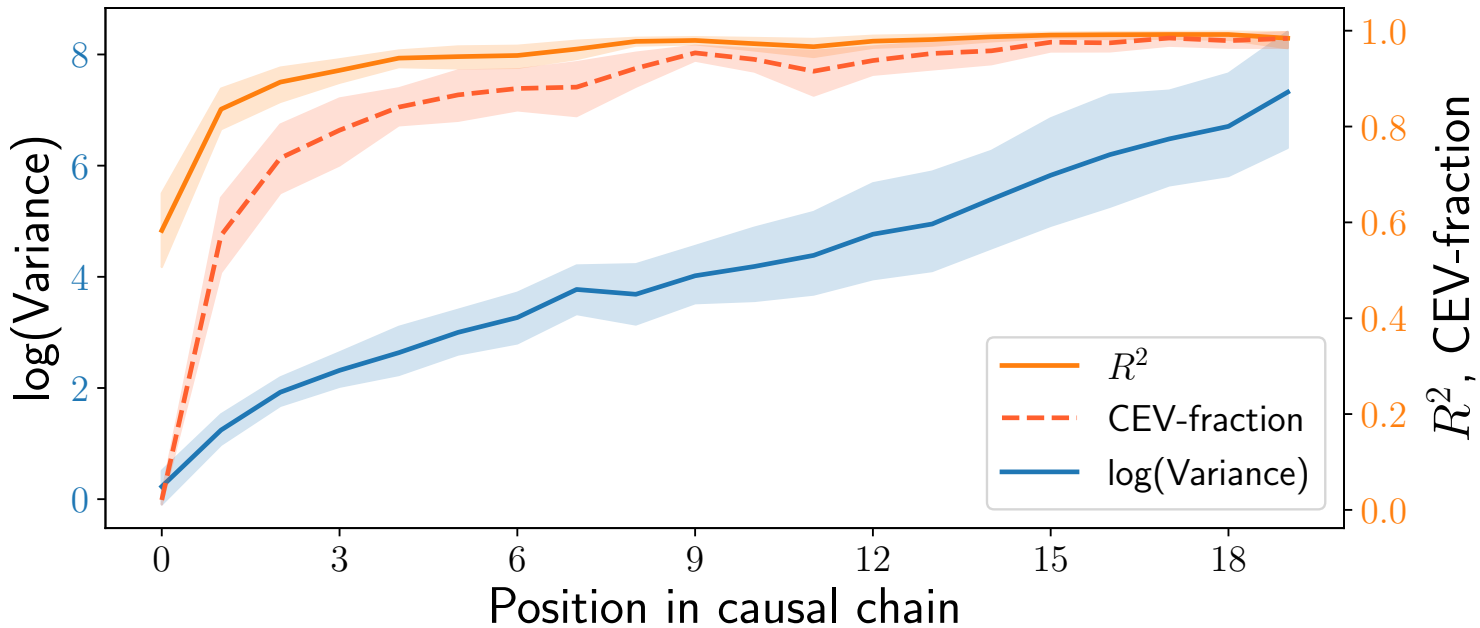
In causal chains, if the condition for the divergence of the variances to infinity is fulfilled, R^2 converges to 1 since

$$\begin{aligned} R^2(X_j) &\geq 1 - \frac{\mathrm{Var}(X_j - \mathbb{E}\left[X_j \mid X_{j-1}\right])}{\mathrm{Var}(X_j)} \\ &= 1 - \frac{\sigma_j^2}{\mathrm{Var}(X_j)} \xrightarrow[j\to\infty]{\text{a.s.}} 1. \end{aligned}$$

[10] Reisach, Tami, et al. 2023.

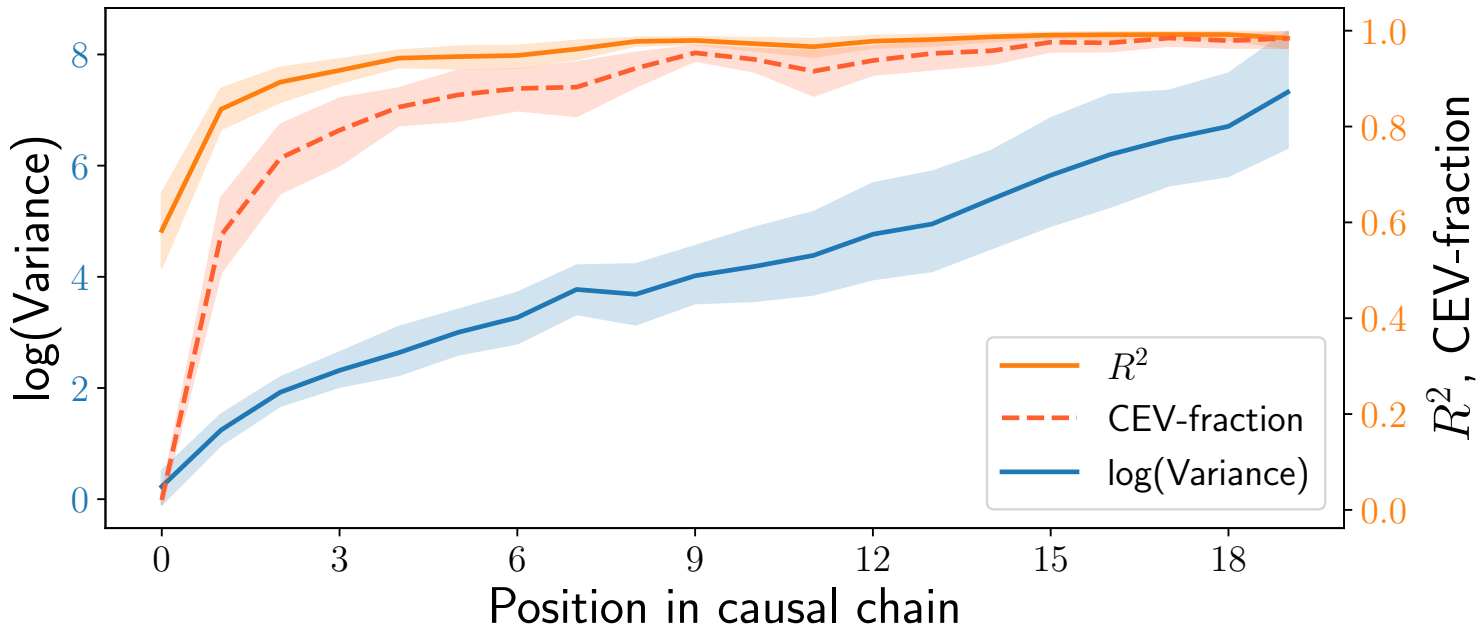
Illustration

Given sufficiently large weights in W, the total variance, cause-explained variance, and R^2 tend to increase along causal chains.



Illustration

Given sufficiently large weights in W, the total variance, cause-explained variance, and R^2 tend to increase along causal chains.



We can use R^2 as a sorting criterion to obtain a candidate causal order! But how well does an ordering by R^2 approximate a causal ordering?

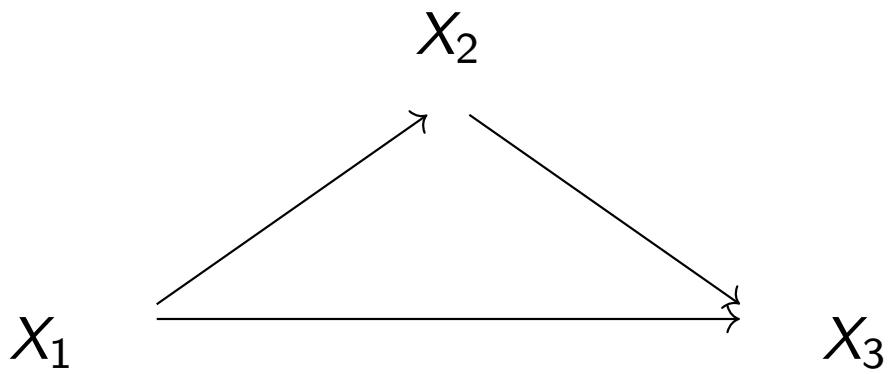
Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

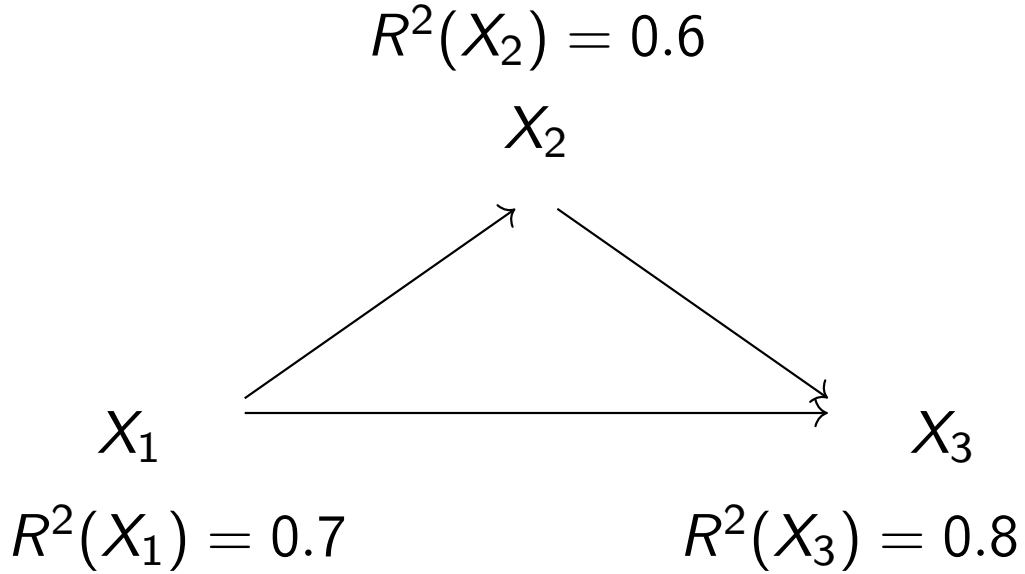
Example R^2-sortability: $\tau(X, i) = R^2(X_i)$



Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

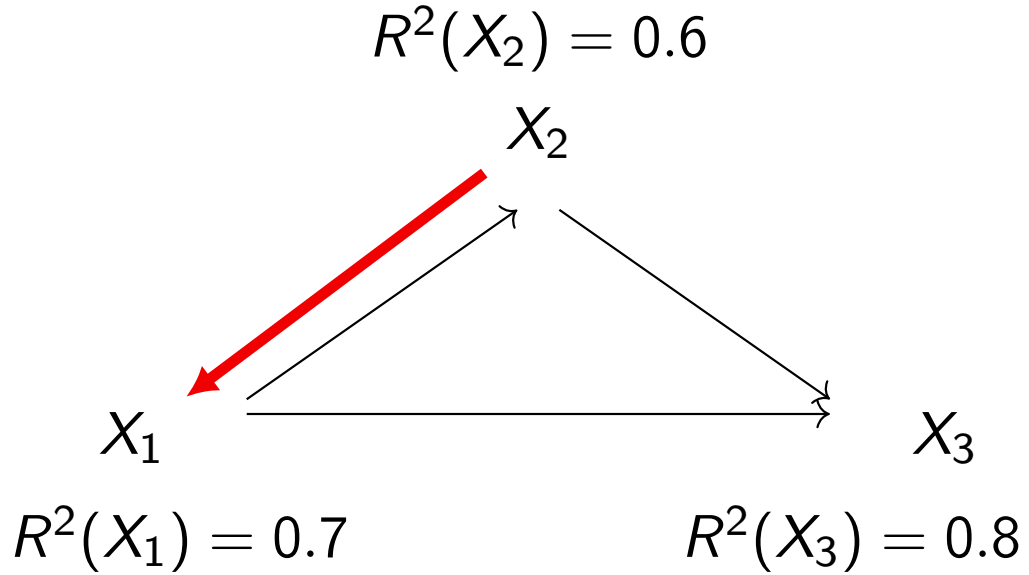
Example R^2-sortability: $\tau(X, i) = R^2(X_i)$



Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

Example R^2-sortability: $\tau(X, i) = R^2(X_i)$

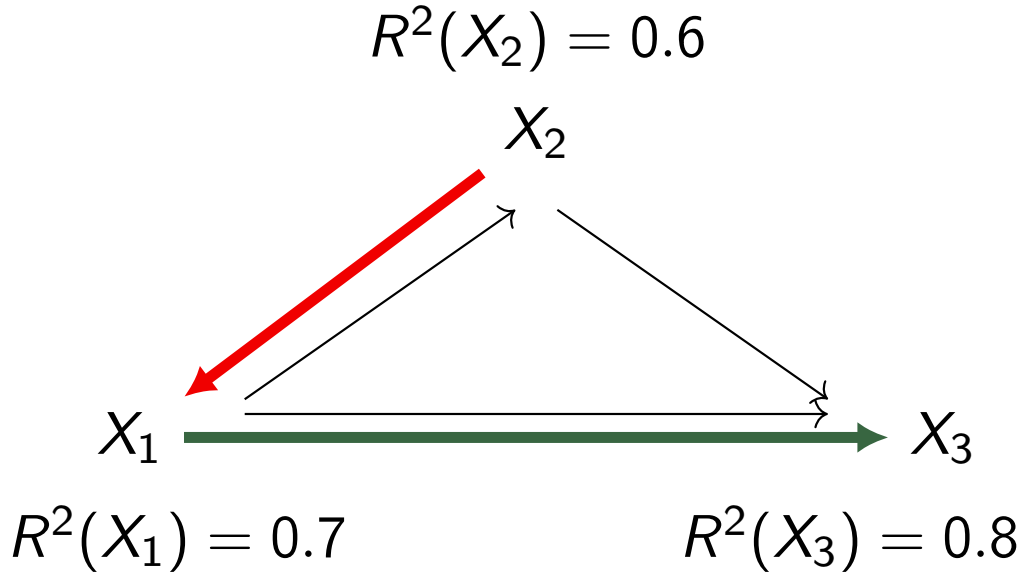


$$v = \frac{}{1}$$

Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

Example R^2-sortability: $\tau(X, i) = R^2(X_i)$

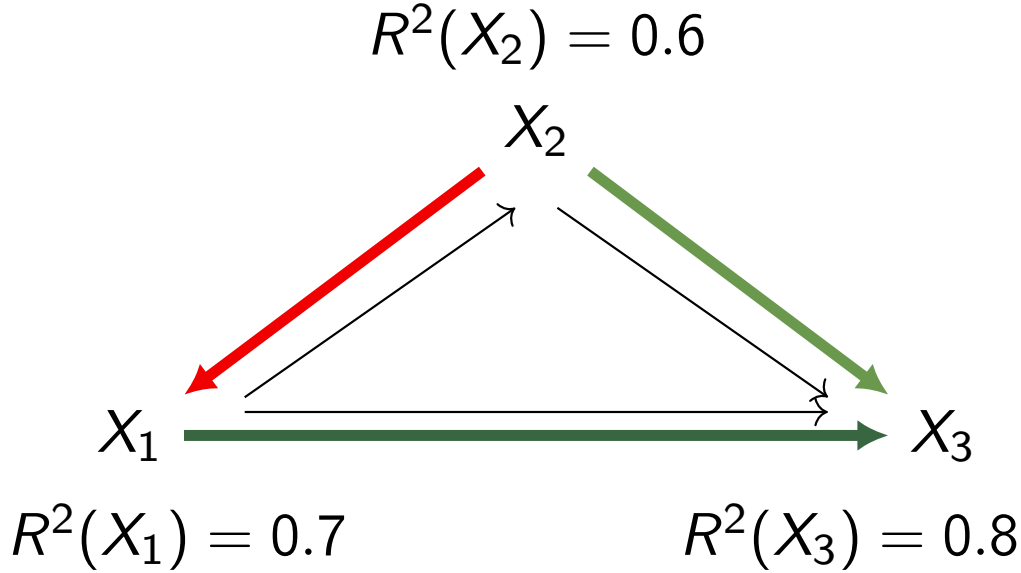


$$v = \frac{1}{1+1}$$

Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

Example R^2-sortability: $\tau(X, i) = R^2(X_i)$

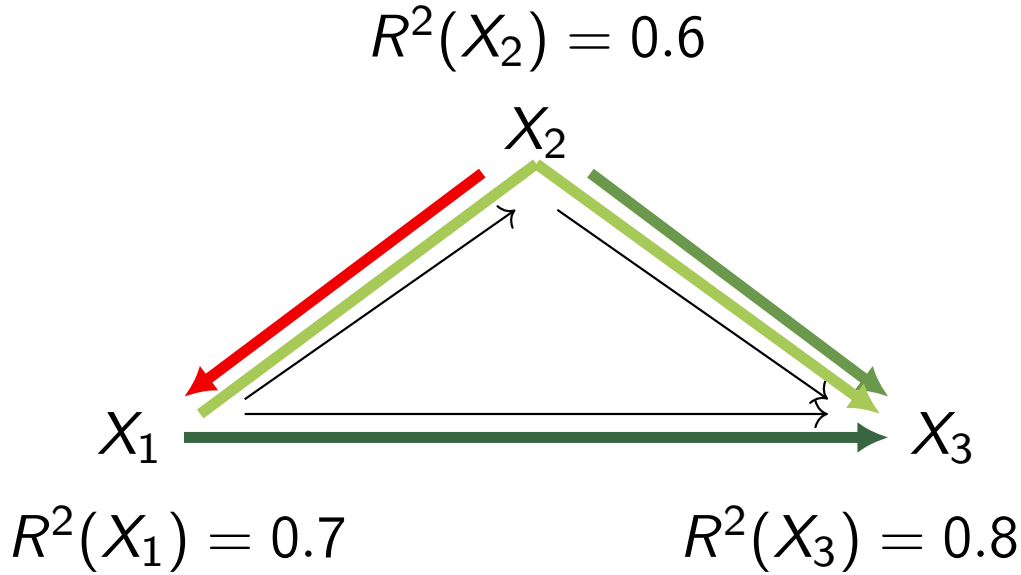


$$v = \frac{1+1}{1+1+1}$$

Sortability

τ-sortability: The fraction of all cause-effect pairs for which the τ-criterion is higher for the effect than for the cause.

Example R^2-sortability: $\tau(X, i) = R^2(X_i)$



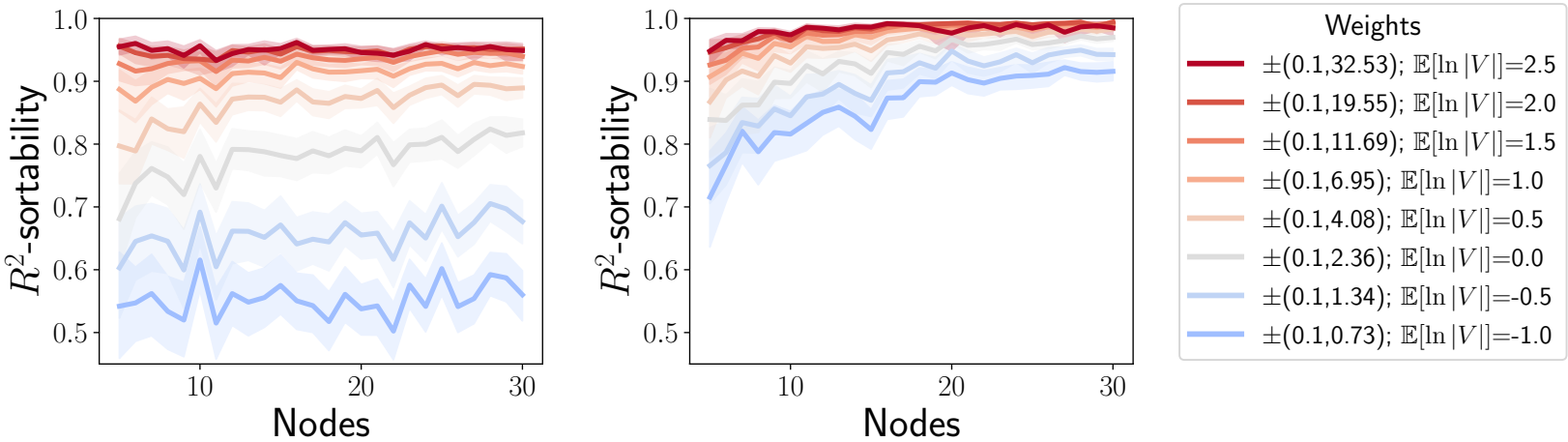
$$v = \frac{1+1+1}{1+1+1+1} = \frac{3}{4}$$

R^2-Sortability in Random DAGs

R^2-sortability measures the agreement between the R^2 ordering and a causal order. A value of 0.5 amounts to a random ordering; a value of 1 amounts to a perfect causal ordering.

R^2-Sortability in Random DAGs

R^2-sortability measures the agreement between the R^2 ordering and a causal order. A value of 0.5 amounts to a random ordering; a value of 1 amounts to a perfect causal ordering.



(a) Erdős–Rényi graphs (b) Scale-free graphs

Figure: Relationship of size and R^2-sortability in random graphs with an average in-degree of 2.

Exploiting R^2-Sortability

R^2SortnRegress – a simple causal discovery algorithm

1. For each variable, compute the R^2 given all others.
2. Sort variables by increasing R^2.
3. Perform sparse regression of each node onto on all its predecessors.

R^2-SortnRegress is simple, fast, and scale-invariant.

Causal Discovery Performance of R^2-SortnRegress

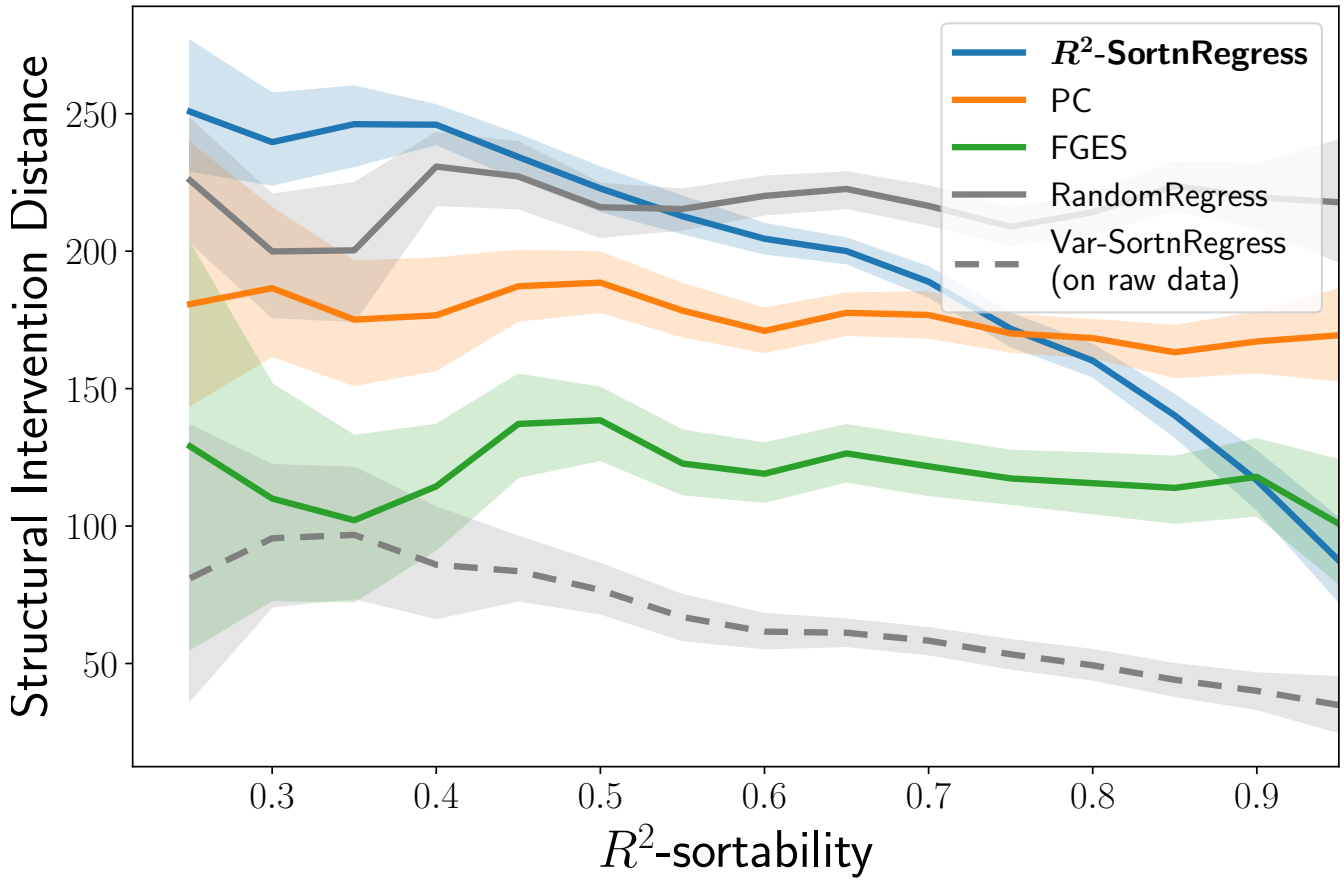


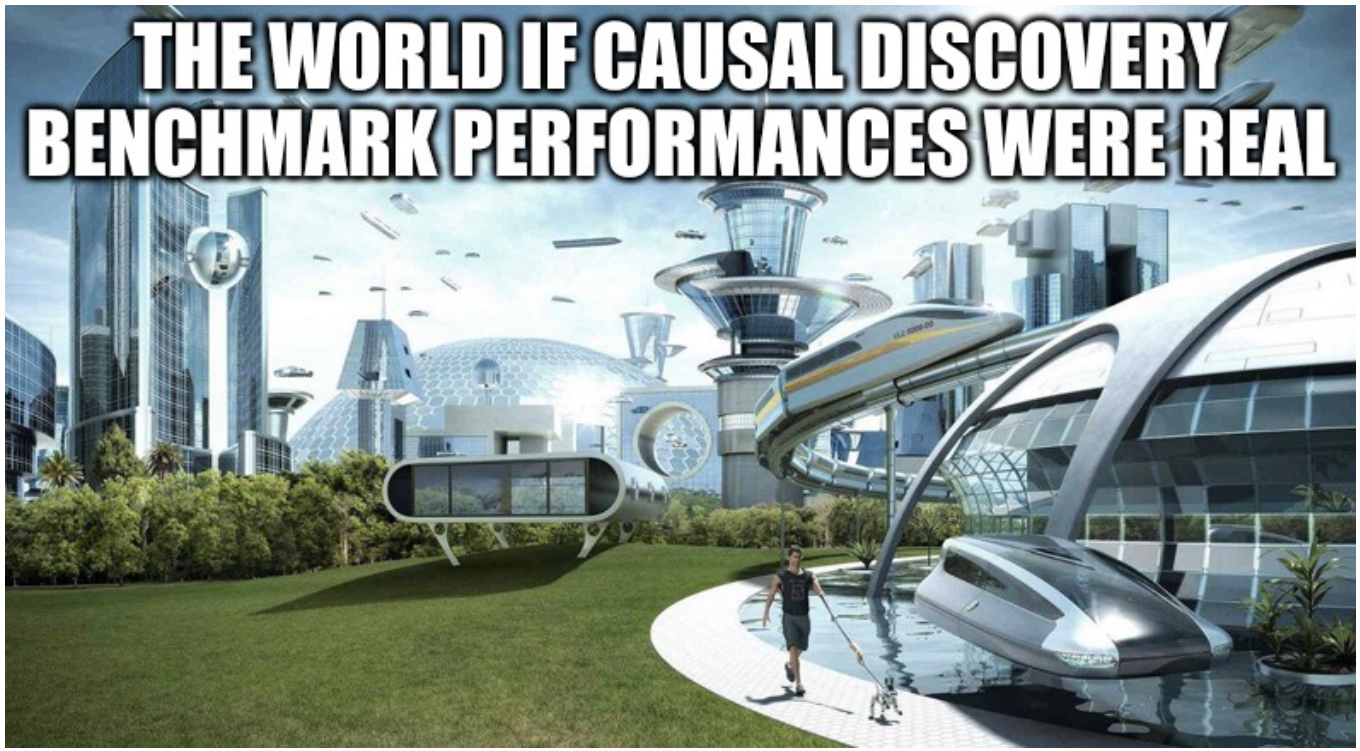
Figure: Causal Discovery results on 500 Erdős–Rényi DAGs with 20 nodes and an avg. in-degree of 2, Gaussian noise with standard deviations drawn iid from Unif$(0.5, 2)$, weights drawn iid from Unif$(\pm(0.5, 1))$.

Take-Away

- Parameter choices can leave distinct patterns in causal models (and they do on many simulated benchmarks).
- **Sortability** is a measure to evaluate the presence of such patterns for a given criterion.
- What (if any) SCM parameterizations are realistic?

Take-Away

- Parameter choices can leave distinct patterns in causal models (and they do on many simulated benchmarks).
- **Sortability** is a measure to evaluate the presence of such patterns for a given criterion.
- What (if any) SCM parameterizations are realistic?



Thank you for your attention!


Lauritzen, Steffen L (1996). *Graphical models*. Vol. 17. Clarendon Press.


Messerli, Franz H (2012). *Chocolate consumption, cognitive function, and Nobel laureates*.


Peters, Jonas et al. (2011). "Identifiability of Causal Graphs using Functional Models". In: *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*. AUAI Press, pp. 589–598.


Reisach, Alexander G., Christof Seiler, and Sebastian Weichwald (2021). "Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game". In: *Advances in Neural Information Processing Systems 34 (NeurIPS)*.


Reisach, Alexander G., Myriam Tami, et al. (2023). "A Scale-Invariant Sorting Criterion to Find a Causal Order in Additive Noise Models". In: *Advances in Neural Information Processing Systems 36 (NeurIPS)*.