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

Alexander Reisach
Maastricht University

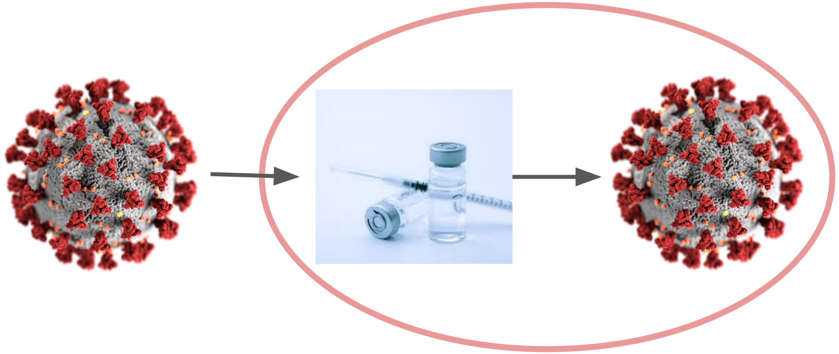
Christof Seiler
Maastricht University

Sebastian Weichwald
Copenhagen University

$$X_1 \to X_2 \to X_3$$

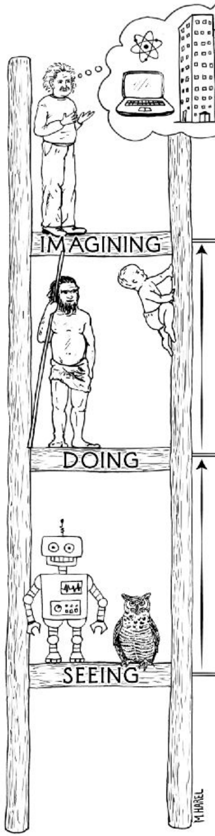
tl;dr: In additive noise models, variance tends to increase along the causal order
⇝ sorting by variance is SOTA 😱

Causality

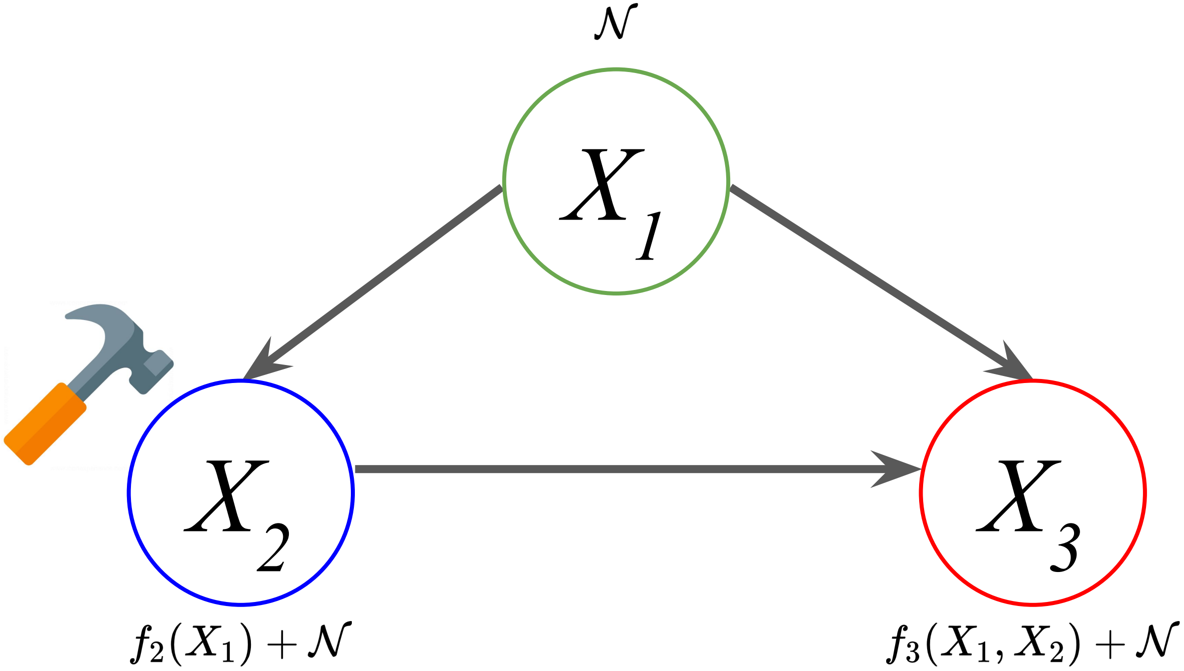


- Mechanism of data generating process
- Distribution under interventions

  ⇝ “Causal Model = structured set of distributions”
- Effect of interventions
- Counterfactuals



Causality and Structural Causal Models



Adjacency Matrix of Directed Acyclic Graph

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

Causal Structure Learning

Structural Equations of an
Additive Noise Model (ANM)

$$X_j = f_j(\text{Parents}(X_j)) + N_j$$

$$P(X_1, \dots, X_n) = \prod_{j=1}^{d} P(X_j | \text{Parents}(X_j))$$

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

Causal Structure

Data generating process

$\longrightarrow$

$\longleftarrow$

Causal Structure Learning

$$\text{minimize } -\mathcal{L}$$
$$\text{s.t. } B \text{ acyclic}$$

$$\begin{pmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} \\ \vdots & \vdots & \vdots \\ x_1^{(n)} & x_2^{(n)} & x_3^{(n)} \end{pmatrix}$$

Observations $\mathbf{X}$

Causal Structure Learning

- Dataset
  - Real-world
    - Structural Biology
  - Synthetic
    - Graph structure
    - Additive noise type/parameters
    - Functional relationships
- Performance Measures
  - Structural Hamming Distance
  - Structural Intervention Distance

- Algorithms
  - Constraint-based (conditional independence testing)
  - Score-based (goodness-of-fit score)
    - Combinatorial optimization
    - Continuous optimization (“NoTears”)*

Given observations $\mathbf{X}$ and graph adjacency matrix W:

$$\underset{W \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \quad \underbrace{\|\mathbf{X} - \mathbf{X}W\|_F^2}_{\text{Mean squared error}}$$

$$\text{s.t.} \quad \underbrace{\operatorname{tr}(\exp(W \times W)) - d = 0}_{\text{Acyclicity measure}}$$

*Zheng, Xun, et al. "Dags with no tears: Continuous optimization for structure learning." *Advances in Neural Information Processing Systems* 31 (2018).

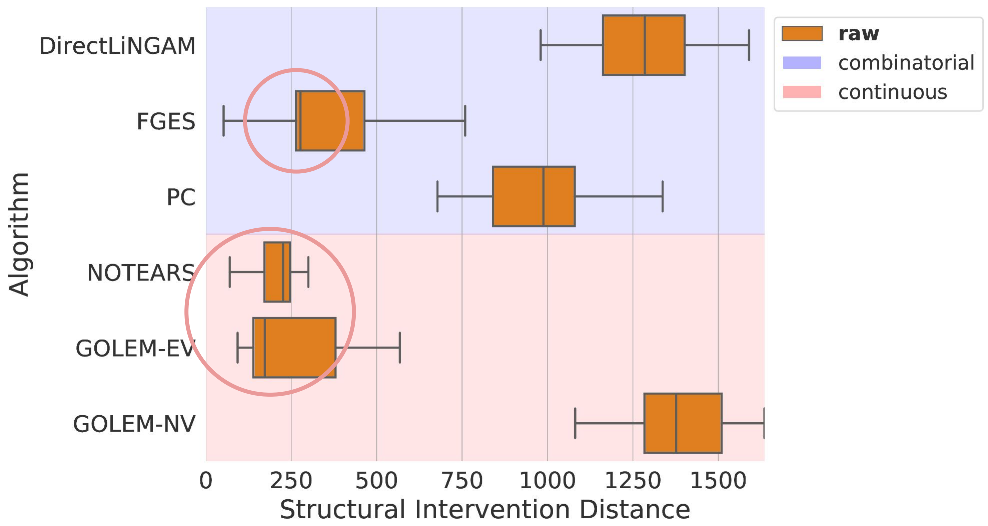
Continuous Causal Structure Learning is Popular!

Method	Year	Data	Acycl.	Interv.	Output
CMS [152]	2014	low	-	no	Bi
NO TEARS [267]	2018	low	yes	no	DAG
CGNN [75]	2018	low	yes	no	DAG
Graphite [83]	2019	low/medium	no	no	UG
SAM [122]	2019	low/medium	yes	no	DAG
DAG-GNN [262]	2019	low	yes	no	DAG
GAE [177]	2019	low	yes	no	DAG
NO BEARS [142]	2019	low/medium/high	yes	no	DAG
Meta-Transfer [19]	2019	Bi	yes	yes	Bi
DEAR [214]	2020	high	yes	no	-
CAN [167]	2020	low/medium/high	yes	no	DAG
NO FEARS [251]	2020	low	yes	no	DAG
GOLEM [176]	2020	low	yes	no	DAG
ABIC [20]	2020	low	yes	no	ADMG/PAG
DYNOTEARS [178]	2020	low	yes	no	SVAR
SDI [124]	2020	low	yes	yes	DAG
AEQ [64]	2020	Bi	-	no	direction
RL-BIC [272]	2020	low	yes	no	DAG
CRN [125]	2020	low	yes	yes	DAG
ACD [151]	2020	low	Granger	no	time-series DAG
V-CDN [145]	2020	high	Granger	no	time-series DAG
CASTLE (reg.) [138]	2020	low/medium	yes	no	DAG
GranDAG [139]	2020	low	yes	no	DAG
MaskedNN [175]	2020	low	yes	no	DAG
CausalVAE [257]	2020	high	yes	yes	DAG
CAREFL [126]	2020	low	yes	no	DAG / Bi
Varando [244]	2020	low	yes	no	DAG
NO TEARS+ [268]	2020	low	yes	no	DAG
ICL [250]	2020	low	yes	no	DAG
LEAST [271]	2020	low/medium/high	yes	no	DAG

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

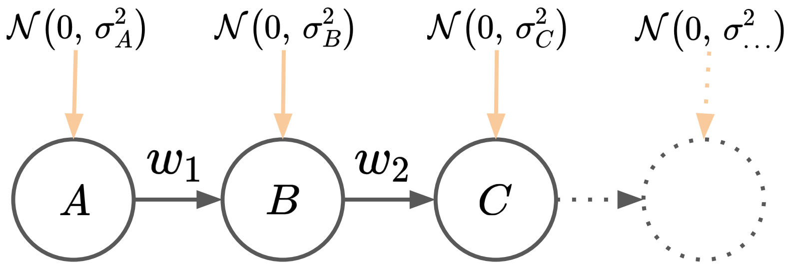
Current State of Causal Structure Learning

Graph	ER-2
Nodes	50
Samples	1000
Edge weights	iid uniform, std. (.5, 2), (-.5, -2)
Noise	iid Gaussian, std. (.4, .8)

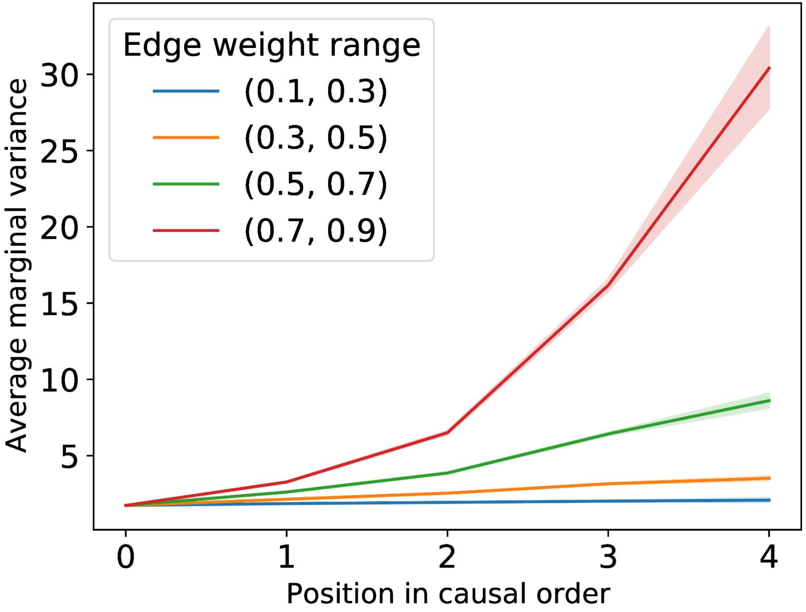


Marginal Variances in ANMs

Generation of a causal chain



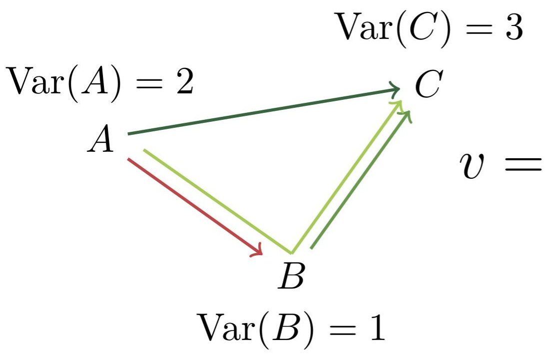
Weights and noise parameters are drawn i.i.d!



Var-sortability

1. Marginal variances carry information about the causal order
2. Sorting by variance gives a complete ordering
3. Causal ordering is a partial ordering

To what extent is the ordering by variance a valid causal ordering?



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

Var-sortability:
Fraction of all cause-effect paths where the effect has a higher variance than the cause.

Empirical Var-sortability

Linear

graph	noise	varsortability min	mean	max
ER-1	Gauss-EV	0.94	0.97	0.99
	exponential	0.94	0.97	0.99
	gumbel	0.94	0.97	1.00
ER-2	Gauss-EV	0.97	0.99	1.00
	exponential	0.97	0.99	1.00
	gumbel	0.98	0.99	0.99
ER-4	Gauss-EV	0.98	0.99	0.99
	exponential	0.98	0.99	0.99
	gumbel	0.98	0.99	0.99

Nonlinear

graph	ANM-type	varsortability min	mean	max
ER-1	Additive GP	0.81	0.91	1.00
	GP	0.72	0.86	0.96
	MLP	0.55	0.79	0.96
	Multi Index Model	0.62	0.82	1.00
ER-2	Additive GP	0.79	0.91	0.98
	GP	0.82	0.89	0.97
	MLP	0.46	0.71	0.87
	Multi Index Model	0.65	0.79	0.89
ER-4	Additive GP	0.90	0.95	0.98
	GP	0.74	0.88	0.93
	MLP	0.59	0.72	0.85
	Multi Index Model	0.57	0.73	0.85

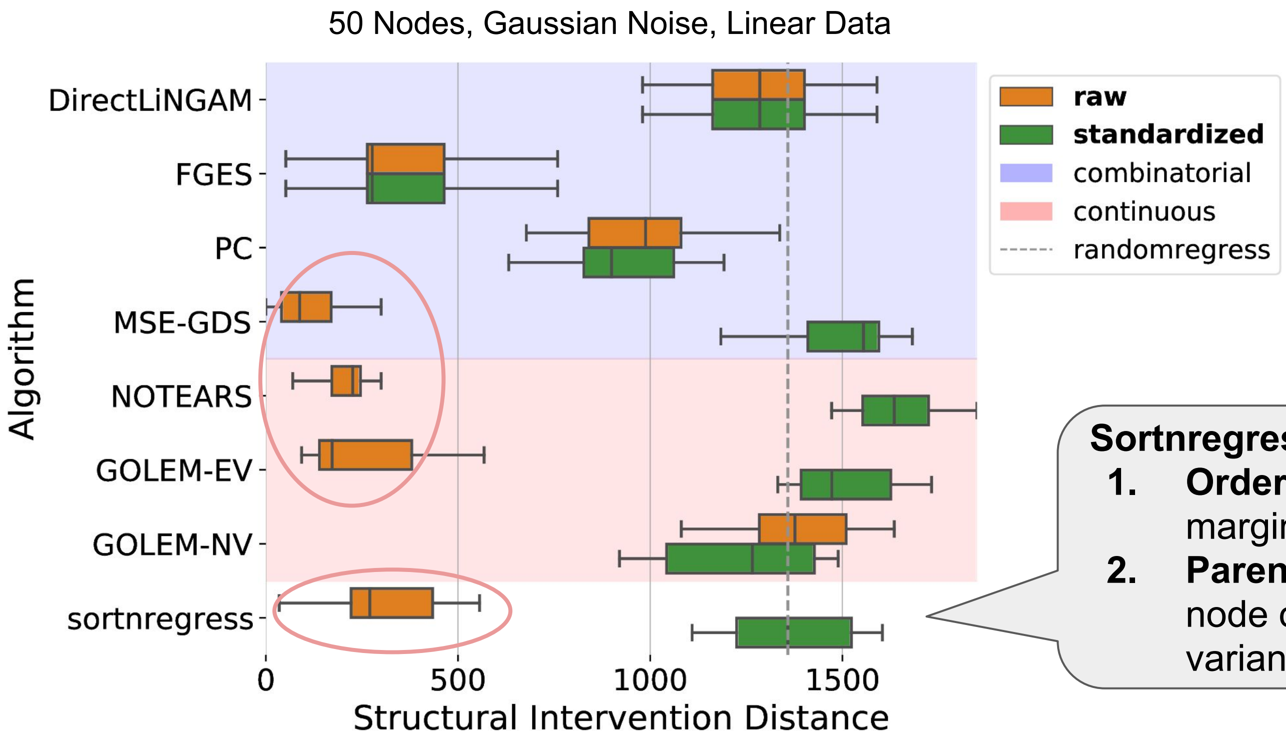
Exploiting Var-sortability

1. Order Search
   a. Sort by increasing (decreasing) marginal variance
2. Parent Search
   a. Regress each node on its predecessors in the variance order
      i. BIC/Lasso regularization
      ii. Edge thresholds

"Sort-and-regress" – a diagnostic tool for the effect of varsortability

(Not the only way of exploiting var-sortability!)

Empirical Results

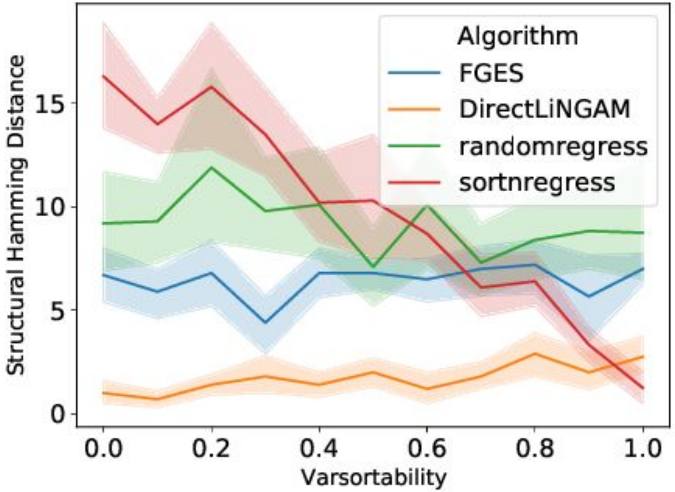


Sortnregress – A diagnostic tool:

1. **Order Search** Sort by increasing marginal variance
2. **Parent Search** Regress each node on its predecessors in the variance order

Explanation and Consequences

- Intuition
  - MSE prioritizes high-variance nodes
  - High-variance nodes tend to be effects
  - Acyclicity constraint keeps structure
- Identifiability
  - Var-sortability reveals causal order
  - Iid sampling & parameters drive varsortability
- Performance
  - MSE-based methods perform well regardless of optimization
  - Drop upon standardization

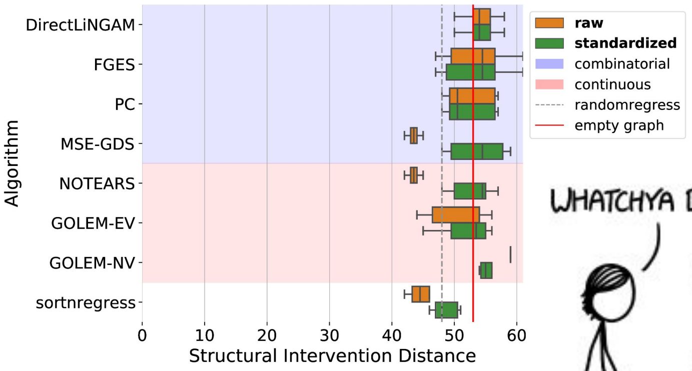


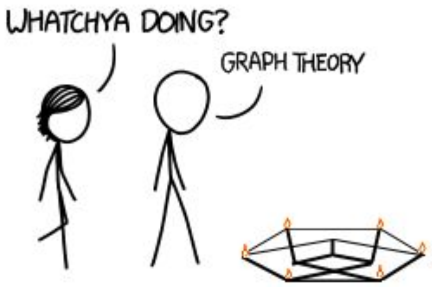
Takeaway: Beware of the Simulated DAG!

- Benchmarking
  - Data scale matters
  - Var-sortability arises easily in synthetic data
  - Does benchmark performance translate?
  - Is var-sortability real?
- Best practices
  - Report var-sortability/sortnregress
  - Simulate real-world processes
  - Use real-world data

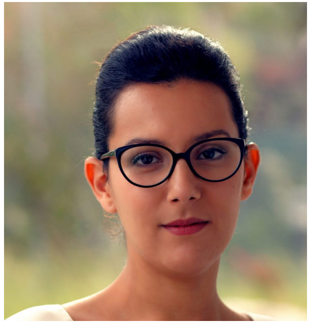
empty graph

Real-world data





Follow-up Work at MICS



Myriam Tami
CentraleSupélec

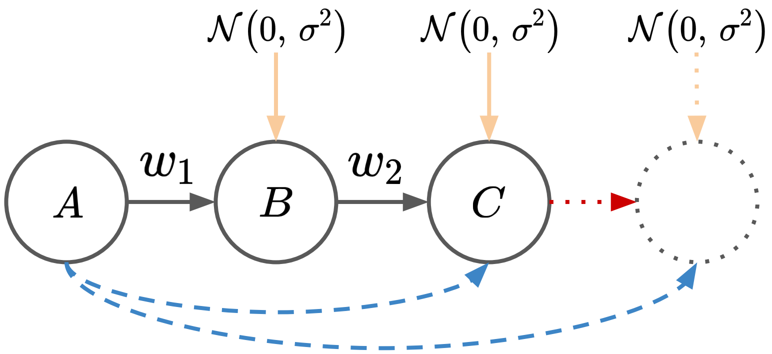
Céline Hudelot
CentraleSupélec

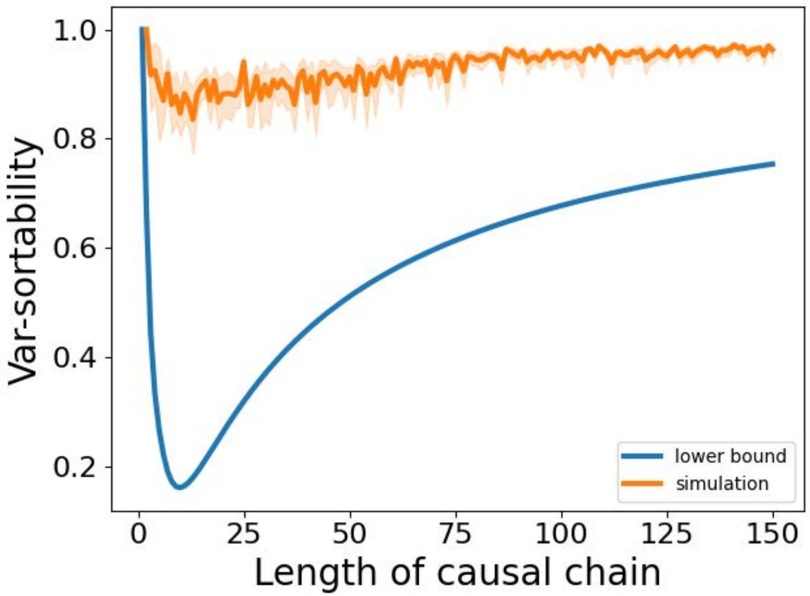
Expected Var-sortability in Causal Chains

Does var-sortability arises for a given distribution of model parameters? *Assumption: Diverging expectation of weight product distribution*

Direct paths: Var-sortability is lower bounded by weight distribution

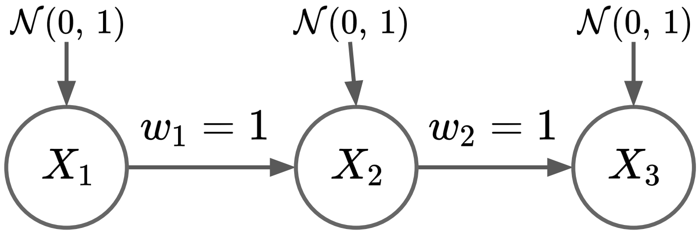
In-direct paths: Increasingly var-sortable (from a point), a lower bound can be found





Weights in [.5, 2], Gaussian noise variance in [.4, .8]

Marginal Variances and Signal-to-Noise ratio



Covariance Matrix

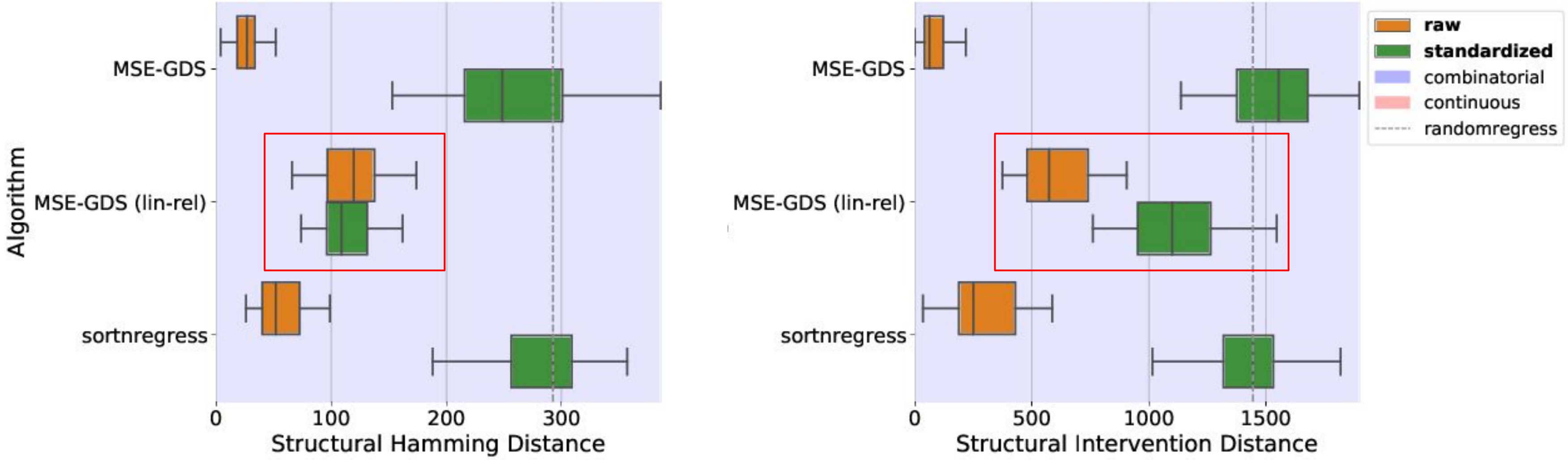
$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} \xrightarrow{\text{standardization}} \begin{pmatrix} 1 & .7 & .6 \\ .7 & 1 & .8 \\ .6 & .8 & 1 \end{pmatrix}$$

Greedy relative MSE DAG search

1. Greedy forward search over new edge insertions
2. Explainable fraction of each node's variance as score Criterion

Preliminary Empirical Results

50 node ER-2 Gaussian linear ANM



Open questions

1. Sufficient and necessary criteria for high (low) var-sortability
    a. In relation to the length of causal chains
    b. In general graphs
2. Signal-to-noise ratio & associated algorithms
    a. Impact of different scaling schemes
    b. Efficient exploitation for structure learning
    c. Empirical comparison on common benchmarks
3. Identifiability after standardization
    a. Given var-sortability of 1 (on raw data)
    b. Partial identifiability for smaller var-sortability (on raw data)

Thank you for your attention!