A Polynomial-time Algorithm for Learning Nonparametric Causal Graphs

Ming Gao, Yi Ding, Bryon Aragam

University of Chicago

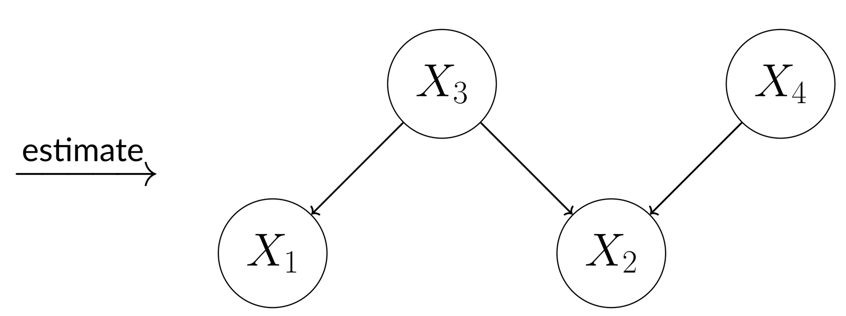
Summary

- A polynomial-time algorithm for learning nonparametric DAGs
- Simultaneous statistical and computational guarantees
- Code: `https://github.com/MingGao97/NPVAR`

Structure Learning

- Estimate a directed acyclic graph (DAG) that fits the data best:

$$\begin{array}{cccc} X_1 & X_2 & X_3 & X_4 \end{array}$$
$$\begin{bmatrix} 0.3 & -0.2 & 0.4 & 0.9 \\ 0.9 & 1.1 & 0.3 & -1.6 \\ 1.1 & 0.2 & -0.4 & 0.6 \\ -0.7 & 1.7 & -0.5 & 1.1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$



- Notation: $(X_1, \ldots, X_d)$ = data, G = graph, pa(j) = parent set of node j

Motivation

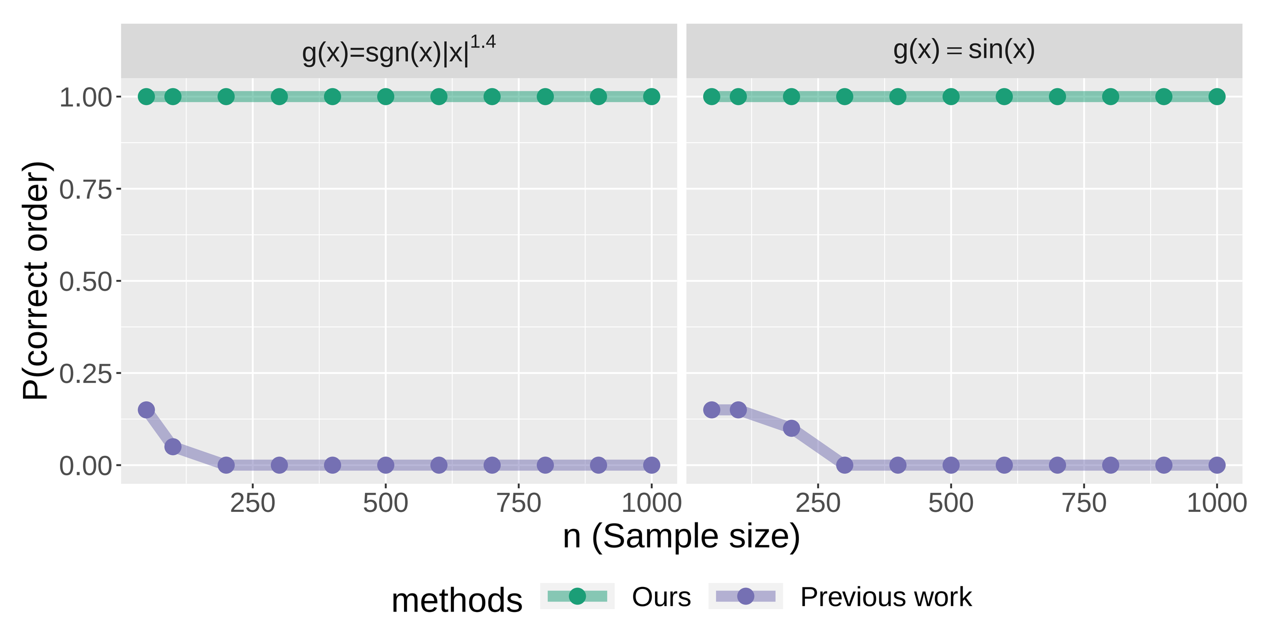


Figure 1: Existing methods fail to find a correct topological ordering in simple settings.

Well-studied problem in parametric settings: What about nonparametric?

Previous work:

- Score-based methods: Computationally expensive
  - Exact solution requires exponential time and space [4, 5]
  - GOBNILP: Integer programming with certificate of optimality [2]
- Constraint-based methods (PC, etc.) [6]: Strong assumptions
  - Complexity: $O(d^k)$, $k = \max|\text{adj}(G, j)|$ = largest neighbourhood
  - Faithfulness + multiple testing
- Order-based methods: Parametric
  - Polynomial-time and sample complexity for linear models [3, 1]
  - This work: Order-based method for nonparametric models

Our Contributions

Structure learning for general nonparametric $P(X)$.

- Provably polynomial-time algorithm
- Provably consistent with explicit sample complexity
- Weak assumptions: No faithfulness, linearity, additivity, independence

NPVAR: A Polynomial-time Algorithm

Basic idea: Order-based method for nonparametric models

- Under equal variance condition (Condition 2, below), source in G $\Longleftrightarrow$ minimize the residual variances.
- Iteratively identify and remove sources by minimizing residual variances
- Nodes can be organized into "layers" $L = (L_1, \ldots, L_{\widehat{r}})$, from which G can be learned via standard nonlinear variable selection methods

Input: $X^{(1)}, \ldots, X^{(n)}$, $\eta > 0$.

1. Set $\widehat{L}_0 = \emptyset$, $\widehat{\sigma}^2_{\ell 0} = \widehat{\text{var}}(X_\ell)$, $k_0 = \arg\min_\ell \widehat{\sigma}^2_{\ell 0}$, $\widehat{\sigma}^2_0 = \sigma^2_{k_0 0}$.
2. Set $\widehat{L}_1 := \{\ell : |\widehat{\sigma}^2_{\ell 0} - \widehat{\sigma}^2_0| < \eta\}$.
3. For $j = 2, 3, \ldots$:
   - Randomly split the n samples in half and let $\widehat{A}_j := \cup_{m=1}^{j} \widehat{L}_m$.
   - For each $\ell \notin \widehat{A}_j$, use the first half of the sample to estimate $f_{\ell j}(X_{\widehat{A}_j}) = \mathbb{E}[X_\ell \,|\, \widehat{A}_j]$ via a nonparametric estimator $\widehat{f}_{\ell j}$.
   - For each $\ell \notin \widehat{A}_j$, use the second half of the sample to estimate the residual variances via the plug-in estimator
   $$\widehat{\sigma}^2_{\ell j} = \frac{1}{n/2}\sum_{i=1}^{n/2}(X_\ell^{(i)})^2 - \frac{1}{n/2}\sum_{i=1}^{n/2}\widehat{f}_{\ell j}(X_{\widehat{A}_j}^{(i)})^2.$$
   - Set $k_j = \arg\min_{\ell \notin \widehat{A}_j} \widehat{\sigma}^2_{\ell j}$ and $\widehat{L}_{j+1} = \{\ell : |\widehat{\sigma}^2_{\ell j} - \widehat{\sigma}^2_{k_j j}| < \eta, \ \ell \notin \widehat{A}_j\}$.
4. Return $\widehat{L} = (\widehat{L}_1, \ldots, \widehat{L}_{\widehat{r}})$.

Identifiability + Sample Complexity

Theorem 1: Equal Variance Identifiability

If $\mathbb{E}\,\text{var}(X_j \,|\, \text{pa}(j)) \equiv \sigma^2$ does not depend on j, G is identifiable from $\mathbb{P}(X)$.

Condition 1: Regularity

For all j and all $\ell \notin A_j$, (a) $X_j \in [0, 1]$, (b) $f_{\ell j} : [0, 1]^{d_j} \to [0, 1]$, (c) $f_{\ell j} \in L^\infty([0, 1]^{d_j})$, and (d) $\text{var}(X_\ell \,|\, A_j) \le \zeta_0 < \infty$.

Condition 2: Identifiability

$\mathbb{E}\,\text{var}(X_j \,|\, \text{pa}(j)) \equiv \sigma^2$ does not depend on j.

Condition 3: Estimator

The nonparametric estimator $\widehat{f}$ satisfies (a) $\mathbb{E}[Y \,|\, Z] \in L^\infty \implies \widehat{f} \in L^\infty$ and (b) $\mathbb{E}_{\widehat{f}}\|\widehat{f}(Z) - \mathbb{E}[Y \,|\, Z]\|_2^2 \to 0$.

Theorem 2: Main Theorem

Assume Conditions 1-3. Let $\Delta_j > 0$ be such that $\mathbb{E}\,\text{var}(X_\ell \,|\, A_j) > \sigma^2 + \Delta_j$ for all $\ell \notin A_j$ and define $\Delta := \inf_j \Delta_j$. Let $\delta^2 := \sup_{\ell, j} \mathbb{E}_{\widehat{f}_{\ell j}}\|f_{\ell j}(X_{A_j}) - \widehat{f}_{\ell j}(X_{A_j})\|_2^2$. Then for any $\delta\sqrt{d} < \eta < \Delta/2$,

$$\mathbb{P}(\widehat{L} = L(\mathsf{G})) \gtrsim 1 - \frac{\delta^2}{\eta^2} r d$$

Remarks:

- Agnostic to the choice(s) of estimator(s) $\widehat{f}_{\ell j}$
- Sample complexity depends on δ = sample complexity of learning $\mathbb{E}[X_\ell \,|\, A_j]$
- No assumptions $\implies$ exponential, sparsity or smoothness $\implies$ polynomial

Experiments: Recovering Graph Structure

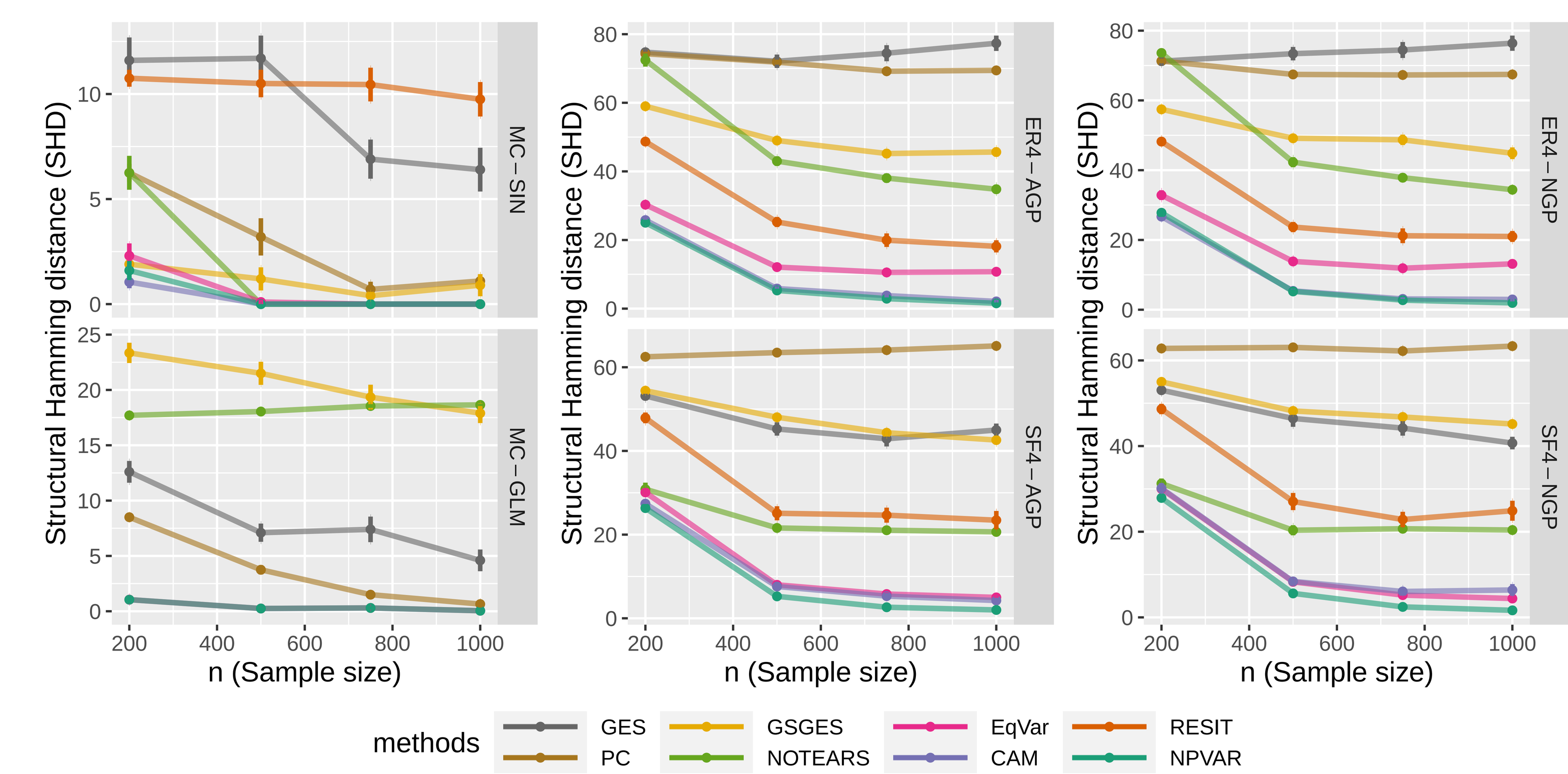


Figure 2: SHD vs. n ($d = 20$). Error bars denote ± 1 standard error.

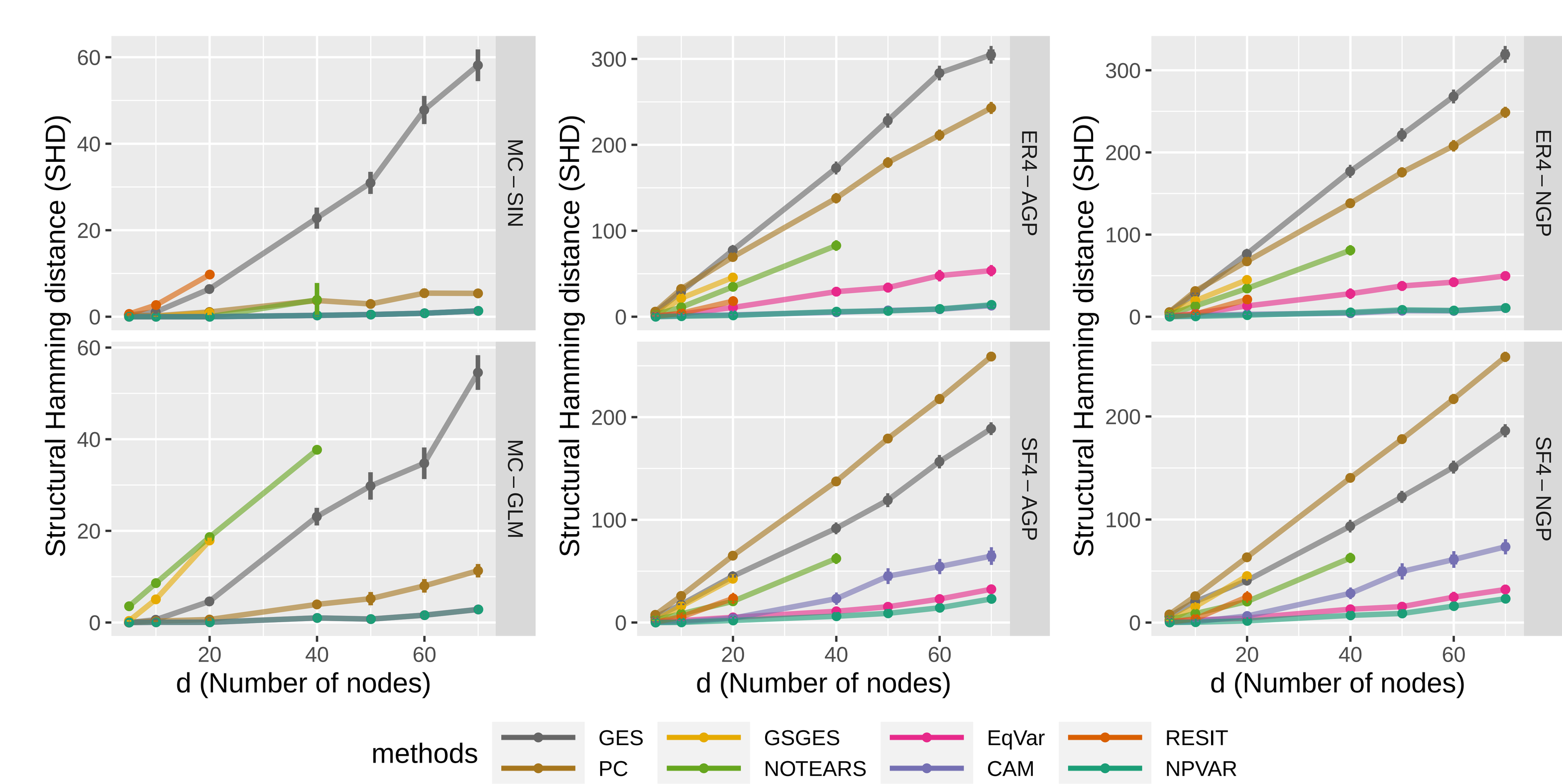


Figure 3: SHD vs. d ($n = 1000$). Error bars denote ± 1 standard error.

More in the paper!

References

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

[2] James Cussens. Bayesian network learning with cutting planes. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 153--160, 2011.

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

[4] Sascha Ott, Seiya Imoto, and Satoru Miyano. Finding optimal models for small gene networks. In *Pacific symposium on biocomputing*, volume 9, pages 557--567. Citeseer, 2004.

[5] Tomi Silander and Petri Myllymaki. A simple approach for finding the globally optimal bayesian network structure. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 2006.

[6] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62--72, 1991.