Optimal structure learning and conditional independence testing

Ming Gao, Yuhao Wang, Bryon Aragam

The University of Chicago Booth School of Business

Main Takeaway

Can we understand the statistical optimality of structure learning via CI testing ?

We establish a fundamental connection between the statistical complexity of CI testing and structure learning (SL):

- optimal testing radius (of CI) $\implies$ optimal sample complexity (of SL)
- achieved by an efficient variant of PC algo w/ the optimal test plugged in
- applied to several examples (Bernoulli, Gaussian, nonparam. cts.)

Motivation

Structure learning / causal discovery / DAG learning enjoys extensive algorithmic developments. However, the statistical optimality is less understood.

One special case: Equal-variance model [3, 5, 2]

What about faithfulness? nonparametric models? A general recipe?

Let's leverage CI testing!

- Closely related: e.g. workhorse to (constraint-based) structure learning.
- Extensively studied: e.g. discrete [1]; continuous [4].

Problem Setup

Given a distribution family $\mathcal{P}$ and a dependence measure m:

CI Testing aims to distinguish two hypotheses of distributions:

$$\begin{aligned} \mathcal{H}_0: &\quad p(X,Y,Z) \quad \text{s.t. } m(X;Y \mid Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y \mid Z \\ \mathcal{H}_1: &\quad p(X,Y,Z) \quad \text{s.t. } m(X;Y \mid Z) \geq r \end{aligned}$$

where $p \in \mathcal{P}$, and r is the signal strength. A CI test ψ is a function of data that outputs a binary decision.

The *optimal testing radius* is the infimum of $r = r_n$ in terms of n s.t. there exists a ψ whose Type-I and Type-II errors are controlled.

Structure Learning (for poly-forest) is to learn the Markov equivalence class of G (a poly-forest) given i.i.d. samples from p where

- p is Markov to G
- p is c-strong tree-faithful to G wrt. m
- $p_{X_j,X_k,X_\ell} \in \mathcal{P}, \forall j,k,\ell \in [d]$

The *optimal sample complexity* is the infimum of sample size n in terms of the number of nodes d and the strong tree-faithfulness parameter c s.t. G can be confidently learned.

A poly-forest is a DAG whose skeleton has any two nodes connected by at most one path.

Markov property: $p(X) = p(X_1,\ldots,X_d) = \prod_{k=1}^{d} p(X_k \mid \mathrm{pa}_G(k))$.

c-strong Tree-faithfulness:

- For any two nodes connected $j-k$, we have $m(X_k;X_j \mid X_\ell) \geq c$ for $\ell \in V \cup \{\emptyset\} \setminus \{k,j\}$;
- For any v-structure $k \to \ell \leftarrow j$, we have $m(X_k;X_j \mid X_\ell) \geq c$.

Main Result

CI testing ⟹ Structure learning for poly-forest

optimal testing radius $\implies$ *optimal sample complexity*

$r_n \asymp n^{-1/\alpha}$ $\qquad$ $n^* \asymp \frac{\log d}{c^\alpha}$

- $\alpha > 0$ depends on the modeling setup
- c is the strong faithfulness parameter
- Applied to Bernoulli and Gaussian ($\alpha = 2$), and nonparametric continuous distributions ($\alpha = \frac{5s+2}{2s} > 2$)
- Optimality achieved by PC-tree with the optimal test plugged in

Applications

Distribution Family $\mathcal{P}$	Parameter α	Testing Radius	Sample Complexity
Bernoulli	$\alpha = 2$	$r_n \asymp n^{-1/2}$	$n^* \asymp \log d / c^2$
Gaussian	$\alpha = 2$	$r_n \asymp n^{-1/2}$	$n^* \asymp \log d / c^2$
Nonparam. Cts. [4]	$\alpha = \frac{5s+2}{2s}$	$r_n \asymp n^{-2s/(5s+2)}$	$n^* \asymp \log d / c^{\frac{5s+2}{2s}}$

- s is a smoothness parameter in the nonparametric function class
- $\alpha = \frac{5s+2}{2s} > 2$ in the nonparametric continuous case is strictly larger than the two parametric cases $\implies$ intrinsic difficulty
- Optimality results for Bernoulli an Gaussian can be extended for poly-tree learning

PC-tree Algorithm

Algorithm 1 PC-Tree

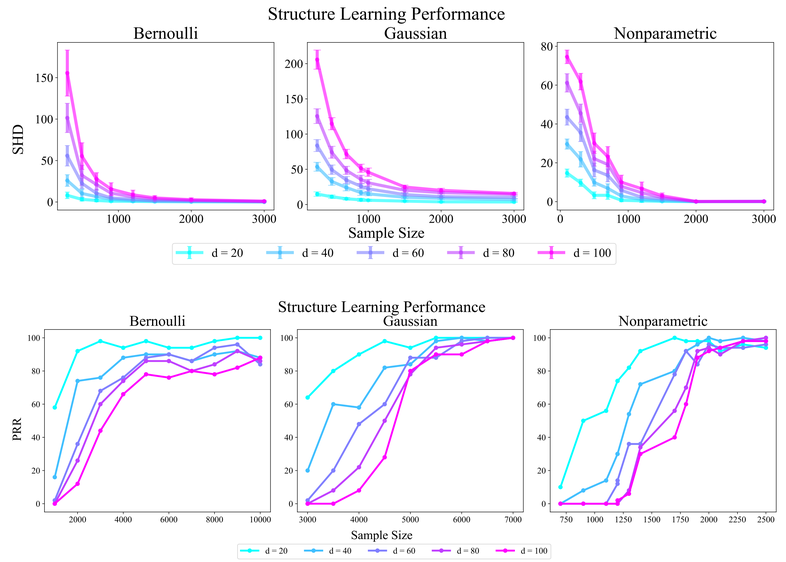
Input: n i.i.d. samples $\{X_1^{(i)},\ldots,X_d^{(i)}\}_{i=1}^n$, CI test ψ as function of data;

1. Let $\widehat{E} = \emptyset$.
2. For each pair (j,k), $0 \leq j < k \leq d$:
   1. For all $\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\}$:
      1. Test $H_0 : X_j \perp\!\!\!\perp X_k \mid X_\ell$ vs. $H_1 : X_j \not\perp\!\!\!\perp X_k \mid X_\ell$ using ψ, store the results.
   2. If all tests reject, then $\widehat{E} \leftarrow \widehat{E} \cup \{j-k\}$.
   3. Else (if some test accepts), let $S(j,k) = \{\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\} : X_j \perp\!\!\!\perp X_k \mid X_\ell\}$.

Output: $\widehat{G} = ([d], \widehat{E})$, separation set S.

The output is refined using orientation rules.

Experiments



More in paper!

References

[1] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart.
Testing conditional independence of discrete distributions.
In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 735--748, 2018.

[2] Constantinos Daskalakis, Vardis Kandiros, and Rui Yao.
Learning gaussian dag models without condition number bounds.
In *Forty-second International Conference on Machine Learning*, 2025.

[3] Ming Gao, Wai Ming Tai, and Bryon Aragam.
Optimal estimation of gaussian dag models.
In *International Conference on Artificial Intelligence and Statistics*, pages 8738--8757. PMLR, 2022.

[4] Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman.
Minimax optimal conditional independence testing.
The Annals of Statistics, 49(4):2151--2177, 2021.

[5] Mishfad Shaikh Veedu, Deepjyoti Deka, and Murti Salapaka.
Information theoretically optimal sample complexity of learning dynamical directed acyclic graphs.
In *International Conference on Artificial Intelligence and Statistics*, pages 4636--4644. PMLR, 2024.