Optimal Estimation of Gaussian (Poly)trees

Yuhao Wang, Ming Gao, Wai Ming Tai, Bryon Aragam, Arnab Bhattacharyya

National University of Singapore, University of Chicago, Nanyang Technological University

Introduction

Goal: Minimum number of samples required to learn the graph from data.

Examples

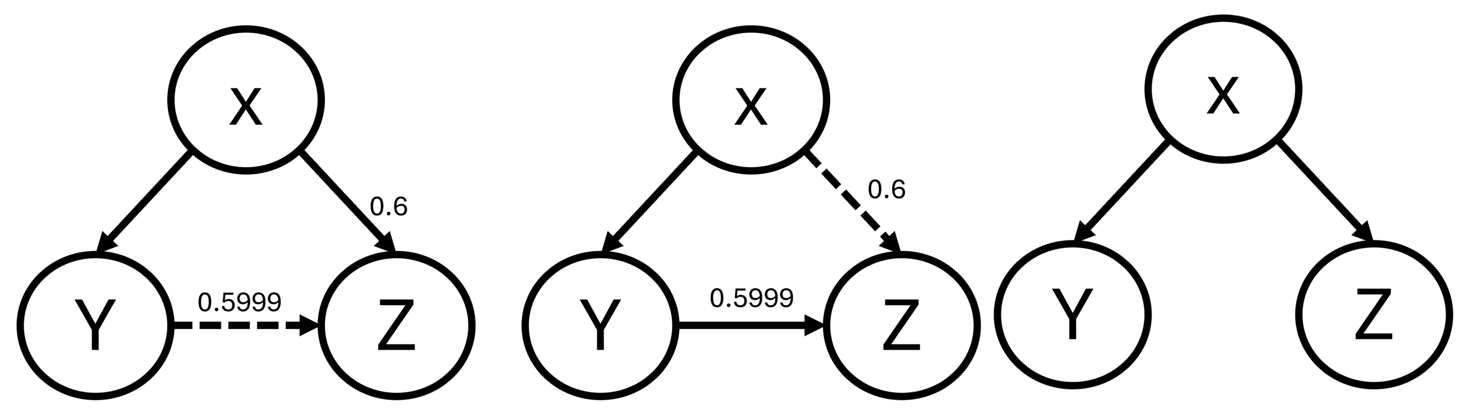

Figure 1:(a) and (b) Distribution learning and (c) Structure learning

Questions

1. (Non-realizable setting) P might not representable by any tree), how many samples are required to learn a tree-structured distribution Q?
2. (Realizable setting)P itself is tree-structured, how many samples are required to learn a tree-structured distribution Q?
3. (Faithful setting) P is faithful to some tree T, how many samples are required to learn T up to Markov equivalence?

Bayesian networks and Tree-Faithfulness

Distribution learning: For a distribution P and a directed tree T, let

$$P_T := \underset{T\text{-structured distribution}Q}{\arg\min} D_{\mathrm{KL}}(P||Q),$$

where $D_{\mathrm{KL}}\{\cdot||\cdot\}$ denotes the KL-divergence.

Definition 1 : [Tree-faithfulness] We say distribution P is tree-faithful to a polytree T if

1. For any two nodes connected $X_j - X_k$, we have $X_k \not\perp\!\!\!\perp X_j \,|\, X_\ell$ for all $\ell \in V \cup \{\emptyset\} \setminus \{k, j\}$;
2. For any v-structure $X_k \to X_\ell \leftarrow X_j$, we have $X_k \not\perp\!\!\!\perp X_j \,|\, X_\ell$.

Definition 2 : [c-strong tree-faithfulness] We say that P is c-strong tree-faithful to a polytree T if

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

Our Contribution

Non-realizable Setting

Without making additional assumptions on P, we show that

$$n = \widetilde{\Theta}\left(\frac{d^2}{\varepsilon^2}\right) \tag{1}$$

samples are necessary and sufficient to learn (with probability at least $2/3$) a tree-structured distribution that is ε-close to the closest tree-structured distribution for P.

Realizable Setting When P itself is Markov to a tree T (i.e. it is *tree-structured*), then

$$n = \widetilde{\Theta}\left(\frac{d}{\varepsilon}\right) \tag{2}$$

samples are necessary and sufficient to learn (with probability at least $2/3$) a tree-structured distribution that is ε-close to P itself.

Faithful Polytrees Assuming that P is faithful to some *polytree* T, we show that the optimal sample complexity of learning $\overline{T}$, the CPDAG of T, is

$$n = \Theta\left(\frac{\log d}{c^2}\right), \tag{3}$$

where c is the strong faithfulness parameter.

Learning Tree-Structured Gaussians

Theorem 1 : [Non-realizable setting] Let P be a Gaussian distribution. Given n i.i.d. samples from P, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d^2}{\varepsilon^2}\log\frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 1 satisfies

$$D_{\mathrm{KL}}(P||P_{\widehat{T}}) \leq \min_{T\in\mathcal{T}} D_{\mathrm{KL}}(P||P_T) + \varepsilon,$$

with probability at least $1-\delta$. Besides, if $n = o(d^2/\varepsilon^2)$, no algorithm returns a directed tree $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P||P_{\widehat{T}}) \leq \min_{T\in\mathcal{T}} D_{\mathrm{KL}}(P||P_T) + \varepsilon$$

with probability at least $2/3$.

Theorem 2 : [Realizable setting] Let T^* be a directed tree and P_{T^*} be a T^*-structured Gaussian. Given n i.i.d. samples from P_{T^*}, for any $\varepsilon, \delta > 0$, if $n \gtrsim \frac{d}{\varepsilon}\log\frac{d}{\delta}$, then $\widehat{T}$ returned by Algorithm 1 satisfies

$$D_{\mathrm{KL}}(P_{T^*}||P_{\widehat{T}}) \leq \varepsilon,$$

with probability at least $1-\delta$. Besides, suppose P is an unknown Gaussian distribution such that $P = P_{T^*}$. Given n i.i.d. samples drawn from P. For any small $\varepsilon > 0$, if $n = o(d/\varepsilon)$, no algorithm returns a directed tree $\widehat{T}$ such that

$$D_{\mathrm{KL}}(P||P_{\widehat{T}}) \leq \varepsilon$$

with probability at least $2/3$.

Algorithm 1 Modified `Chow-Liu` algorithm

- **Input**: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$
- For each $j = 1, \ldots, d$:
  - $\widehat{\sigma}_j^2 \leftarrow \frac{1}{n}\sum_{i=1}^n (X_j^{(i)})^2$
- For each pair $(j,k), 1 \leq j < k \leq d$:
  - $\widehat{\rho}_{jk} \leftarrow \frac{1}{n}\sum_{i=1}^n X_j^{(i)} X_k^{(i)}$
- For each pair $(j,k), 1 \leq j < k \leq d$:
  - $\widehat{I}(X_j; X_k) \leftarrow -\frac{1}{2}\log\left(1 - \frac{\widehat{\rho}_{jk}^2}{\widehat{\sigma}_j^2\widehat{\sigma}_k^2}\right)$ which is same as $\frac{1}{2}\log(1 + \frac{\widehat{\beta}_{jk}^2\widehat{\sigma}_j^2}{\widehat{\sigma}_{k|j}})$
- $G \leftarrow$ the weighted complete undirected graph on $[d]$ whose edge weight for (j,k) is $\widehat{I}(X_j; X_k)$
- $\widehat{S} \leftarrow$ the maximum weighted spanning tree of G
- $\widehat{T} \leftarrow$ any directed tree with skeleton to be $\widehat{S}$
- **return** A directed tree $\widehat{T}$

Optimal Faithful Tree Learning

Algorithm 2 `PC-Tree` algorithm

- **Input**: n i.i.d. samples $(X_1^{(i)}, \ldots, X_d^{(i)})$
- Let $\widehat{E} = \emptyset$.
- For each pair (j,k), $0 \leq j < k \leq d$:
  - For all $\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\}$:
    - Test $H_0 : X_j \perp\!\!\!\perp X_k \,|\, X_\ell$ vs. $H_1 : X_j \not\perp\!\!\!\perp X_k \,|\, X_\ell$, store the results.
  - If all tests reject, then $\widehat{E} \leftarrow \widehat{E} \cup \{j - k\}$.
  - Else (if some test accepts), let $S(j,k) = \{\ell \in [d] \cup \{\emptyset\} \setminus \{j,k\} : X_j \perp\!\!\!\perp X_k \,|\, X_\ell\}$.
  - **Return**: $\widehat{T} = ([d], \widehat{E})$, separation set S

Theorem 3 : [Structure learning] For any $T \in \widetilde{\mathcal{T}}$, assuming P is c-strong tree-faithful to T, applying Algorithm 2 with sample correlation for CI testing, if the sample size

$$n \gtrsim \frac{1}{c^2}\left(\log d + \log(1/\delta)\right),$$

then $\Pr(\widehat{T} = \mathrm{sk}(T)) \geq 1-\delta$, and $\Pr(\mathrm{Orient}(\widehat{T}, S) = \overline{T}) \geq 1-\delta$.

Besides, assuming $c^2 \leq 1/5$, $d \geq 4$, if the sample size is bounded as

$$n \leq \frac{1-2\delta}{8} \times \frac{\log d}{c^2},$$

then for any estimator $\widehat{T}$ for $\overline{T}$,

$$\inf_{\widehat{T}} \sup_{\substack{T\in\widetilde{\mathcal{T}} \\ P \text{ is } c\text{-strong} \\ \text{tree-faithful to } T}} \Pr(\widehat{T} \neq \overline{T}) \geq \delta - \frac{\log 2}{\log d}.$$

Experiment results

- PC-Tree algorithm does perform the best, especially on PRR over the baselines.
- We have not analyzed the performance of Chow-Liu under the goal of structure learning, and we conjecture a similar sample complexity is shared with PC-Tree.

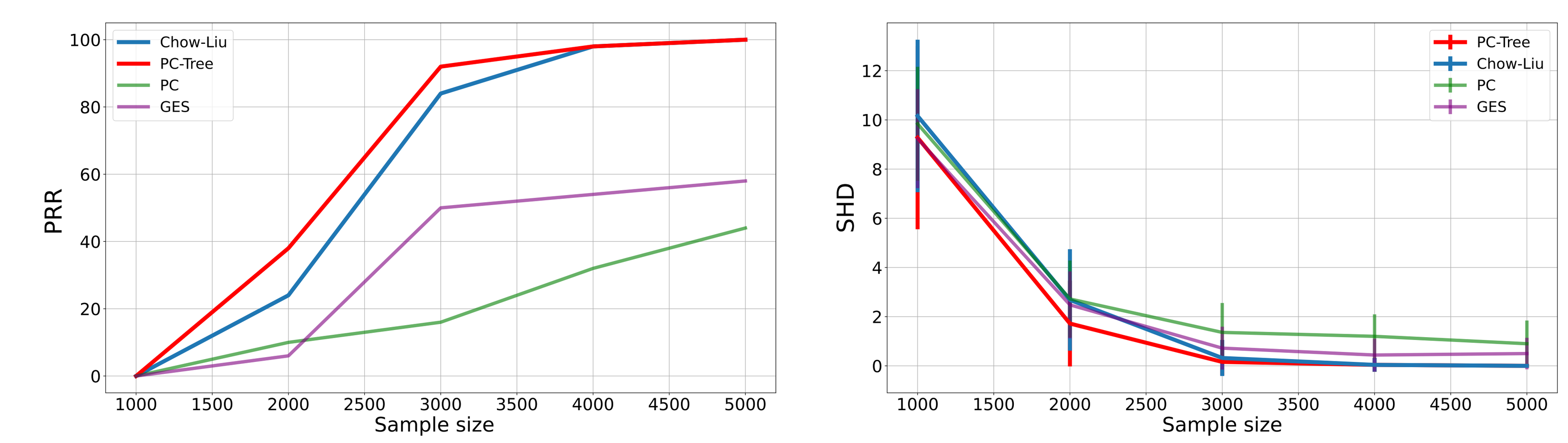

Figure 2:Performance comparison for PC-Tree, Chow-Liu, PC and GES algorithm evaluated on SHD and PRR. The red, blue, green, purple lines are for PC-Tree, Chow-Liu, PC and GES respectively.

Conclusion and Future Work

We treat both problems in a unified setting, allowing for an explicit comparison of these problem:

1. In regime $\varepsilon \ll dc^2$, distribution learning is harder (in terms of sample size needed);
2. In regime $c^2 \ll \varepsilon \ll dc^2$, distribution learning does not automatically imply structure learning;
3. Extending these results beyond the Gaussians we consider here (as well as finite alphabets as in previous work) is a promising direction for future research.