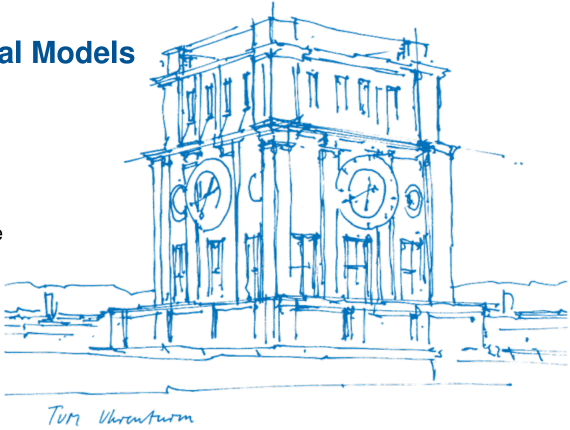
Learning Linear Non-Gaussian Polytree Models

ETH-UCPH-TUM Workshop on Graphical Models

Daniele Tramontano, Anthea Monod and Mathias Drton

Chair of Mathematical Statistics
Department of Mathematics and Munich Data Science Institute
Technical University of Munich

October 10-15, 2022



Problem and setup

- **Structure Learning Problem**
  Learn the DAG $\mathcal{G}$ of a structural causal model from observational data. Data are n i.i.d. copies of a p-dimensional random vector X satisfying

$$X_i = f_i(X_{\mathrm{pa}(i)}, \varepsilon_i),\ i \in [p],$$

  where the ε_i are independent noise terms.

- **Challenges**
  1. The graph $\mathcal{G}$ can be non identifiable,
  2. If p is large , then usual algorithms are too slow.

- **Assumptions**
  1. $\mathcal{G}$ is a *Polytree*,
  2. f_i are *Linear*,
  3. ε_i are *non-Gaussian*.

Polytrees

- The skeleton of the graph is a tree, i.e there are no undirected cycles,

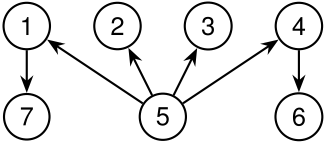


Figure 1 A directed tree[1]

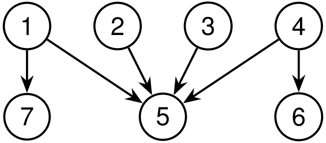


Figure 2 A polytree

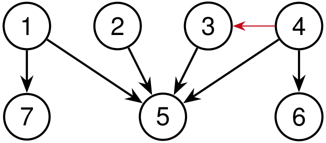


Figure 3 Not a polytree

- Why? The graph can be recovered using any (reasonable) bivariate dependence measures.[2]

[1] Jakobsen et al. [2022]

[2] Rebane and Pearl [1987]

Orientation

- **Proposed Two-Step Approach**
  1. Learn the skeleton with Chow-Liu algorithm,
  2. Three different orientation schemes.
- **Orientation Matrix**[3] For a potential edge $e : i \to j$ and $K \in \mathbb{N}$, define the matrix $A^{e,K}$ as:

$$\left(\frac{c_m^{e,k}}{c_{m-1}^{e,k}} \mid 2 \leq m \leq k \leq K \right),$$

  where $c_m^{e,k}$ is the kth cumulant $\mathrm{cum}(X_i, \ldots, X_{i_k})$ with $i_1 = \cdots = i_m = i$ and $i_{m+1} = \ldots i_k = j$.
- **Cumulants**
  1. $\mathrm{cum}(X_i) = \mathbb{E}[X_i] = 0$
  2. $\mathrm{cum}(X_{i_1}, X_{i_2}) = \mathbb{E}[X_{i_1} X_{i_2}]$
  3. $\mathrm{cum}(X_{i_1}, X_{i_2}, X_{i_3}) = \mathbb{E}[X_{i_1} X_{i_2} X_{i_3}]$
  4. $\mathrm{cum}(X_{i_1}, X_{i_2}, X_{i_3}, X_{i_4}) = \mathbb{E}[X_{i_1} X_{i_2} X_{i_3} X_{i_4}] - \mathbb{E}[X_{i_1} X_{i_2}]\mathbb{E}[X_{i_3} X_{i_4}] - \mathbb{E}[X_{i_1} X_{i_3}]\mathbb{E}[X_{i_2} X_{i_4}] - \mathbb{E}[X_{i_1} X_{i_4}]\mathbb{E}[X_{i_2} X_{i_3}]$.

[3] Améndola et al. [2021]

Orientation Matrix

- **Example**($\mathcal{G}: 1 \to 2$, $K = 3$, zero means) Let $X_1 = \varepsilon_1$ and $X_2 = \lambda_{1,2} X_1 + \varepsilon_2$. Thus:

$$A^{1\to 2,3} = \begin{bmatrix} \mathbb{E}(X_1^2) & \mathbb{E}(X_1^3) & \mathbb{E}(X_1^2 X_2) \\ \mathbb{E}(X_1 X_2) & \mathbb{E}(X_1^2 X_2) & \mathbb{E}(X_1 X_2^2) \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\varepsilon_1^2) & \mathbb{E}(\varepsilon_1^3) & \lambda_{1,2}\mathbb{E}(\varepsilon_1^3) \\ \lambda_{1,2}\mathbb{E}(\varepsilon_1^2) & \lambda_{1,2}\mathbb{E}(\varepsilon_1^3) & \lambda_{1,2}^2\mathbb{E}(\varepsilon_1^3) \end{bmatrix}$$

  whereas :

$$A^{2\to 1,3} = \begin{bmatrix} \mathbb{E}(X_2^2) & \mathbb{E}(X_2^3) & \mathbb{E}(X_2^2 X_1) \\ \mathbb{E}(X_2 X_1) & \mathbb{E}(X_2^2 X_1) & \mathbb{E}(X_2 X_1^2) \end{bmatrix} = \begin{bmatrix} \lambda_{1,2}^2\mathbb{E}(\varepsilon_1^2) + \mathbb{E}(\varepsilon_2^2) & \lambda_{1,2}^3\mathbb{E}(\varepsilon_1^3) + \mathbb{E}(\varepsilon_2^3) & \lambda_{1,2}^2\mathbb{E}(\varepsilon_1^3) \\ \lambda_{1,2}\mathbb{E}(\varepsilon_1^2) & \lambda_{1,2}^2\mathbb{E}(\varepsilon_1^3) & \lambda_{1,2}\mathbb{E}(\varepsilon_1^3) \end{bmatrix}.$$

  so $rk(A^{1\to 2,3}) = 1$ while $rk(A^{2\to 1,3}) = 2$ (in general).

- **Matrix Rank Reveals Orientation**
  1. Let $e : i \to j$ be an edge of G. Then
     (i) $\mathrm{rank}(A^{i\to j,K}) = 1$,
     (ii) $\mathrm{rank}(A^{j\to i,K}) = 2$, in general.
  2. Suppose the skeleton of $\mathcal{G}$ contains the subgraph $i - j - l$ with $\rho_{i,j}, \rho_{j,l} \neq 0$. Then the corresponding subgraph of G is $i \to j \leftarrow l$ iff $\rho_{i,l} = 0$.

Orientation Schemes and Consistency

- **Orientation Schemes**
  1. ***PO*** Orient all the edges independently using lemma 1,
  2. ***TPO*** Orient the first edge using lemma 1, and then use lemma 2 to orient the next edges,
  3. ***PTO*** Learn the the CPDAG first using lemma 2, and then use lemma 1 to orient the remainig edges.
- **Consistency** Suppose the data are an n-sample drawn from a distribution in the model given by a polytree G. Let $\hat{G}$ be the polytree obtained by applying the orientation scheme ***PO*** to the (undirected) edge set of the Chow–Liu tree $\mathcal{M}(R_n)$. There is a set of constants $\{\delta', M_K, L\}$ such that $\hat{G} = G$ with probability greater than

$$1 - 4B(K)(p-1)\exp\left\{-\frac{2}{LK^2\sqrt{M_K}}\left(\delta'\sqrt{n}\right)^{\frac{1}{K}}\right\} + -\frac{3p(p-1)}{2}\exp\left\{-\frac{1}{2L\sqrt{M_2}}\left(\frac{\lambda\gamma\sqrt{n}}{2+\lambda}\right)^{\frac{1}{2}}\right\},$$

for all n greater than:

$$\max\left\{\frac{e^2(2+\lambda)^2(4L^2\sqrt{M_2})^4}{\lambda^2\gamma^2}, \frac{e^2(LK^2\sqrt{M_K})^{2K}}{\delta'^2}\right\},$$

where p is the size of the tree and n is the sample size.

Simulations

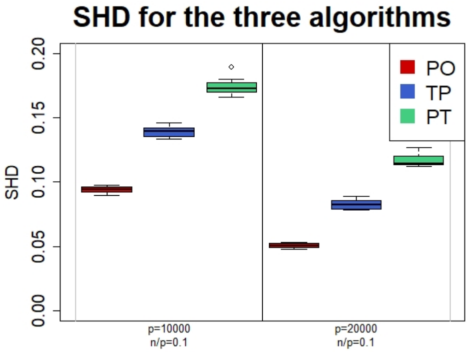


Figure 4 Structural Hamming distance on simulated polytrees, $p = 10000, 20000$, $n/p = 0.1$ and ε_i drawn from a gamma distribution

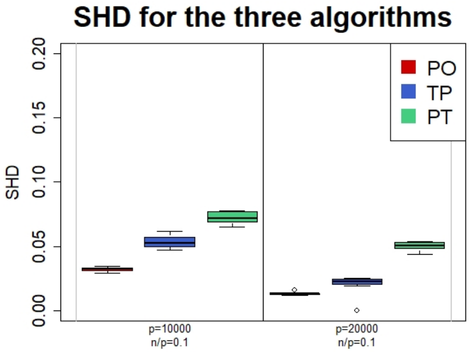


Figure 5 Structural Hamming distance on simulated polytrees, $p = 10000, 20000$, $n/p = 0.1$ and ε_i drawn from a uniform distribution

Future Work?

1. Other identifiable settings?
2. How to avoid Chow–Liu?
3. Which tree structures are the most difficult to learn?[4]
4. What happens when the graph is not a tree?[5]

[4] Tan et al. [2009]
[5] Acid and de Campos [1994], Dasgupta [1999], Grüttemeier et al. [2021]

Bibliography I

Silvia Acid and Luis M. de Campos. Approximations of causal networks by polytrees: an empirical study. In *Advances in Intelligent Computing - IPMU'94, 5th International Conference on Processing and Management of Uncertainty in Knowledge-Based Systems*, volume 945 of *Lecture Notes in Computer Science*, pages 149–158. Springer, 1994.

Carlos Améndola, Mathias Drton, Alexandros Grosdos, Roser Homs, and Elina Robeva. Third-order moment varieties of linear non-Gaussian graphical models, 2021.

Sanjoy Dasgupta. Learning polytrees. In Kathryn B. Laskey and Henri Prade, editors, *UAI '99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30 - August 1, 1999*, pages 134–141. Morgan Kaufmann, 1999.

Niels Grüttemeier, Christian Komusiewicz, and Nils Morawietz. On the parametrized complexity of polytree learning. *CoRR*, abs/2105.09675, 2021.

Martin E. Jakobsen, Rajen D. Shah, Peter Bühlmann, and Jonas Peters. Structure learning for directed trees. *Journal of Machine Learning Research*, 23(159):1–97, 2022.

Bibliography II

George Rebane and Judea Pearl. The recovery of causal poly-trees from statistical data. In *Int. J. Approx. Reason.*, 1987.

Vincent Y. F. Tan, Animashree Anandkumar, and Alan S. Willsky. How do the structure and the parameters of gaussian tree models affect structure learning? In *47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 684–691, 2009.