Structure Learning

Tri Nguyen

Internal Presentation
Oregon State University

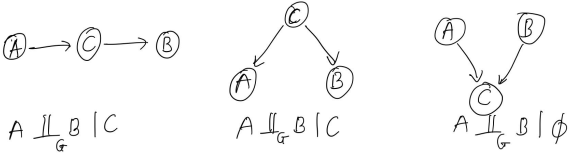
May 13, 2022

Main Reference

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

Some Definitions

- **Directed Acyclic Graph (DAG)**. A graph G is a DAG if it is directed and there is no cycle.
  - **d-separation**. 3 vertices is called $A \perp\!\!\!\perp_G B|C$ if they form either a chain, fork, or collider in G (in a particular order).



- **Markov assumption**. A joint probability P is Markov compatible to a DAG G iff

$$P(X_1, \ldots, X_p) = \prod_i P(X_i | \text{pa}_i)$$

  - P is Markov compatible to G iff

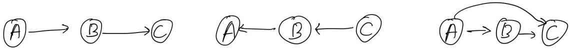
$$\boldsymbol{A} \perp\!\!\!\perp_G \boldsymbol{B} \mid \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp_P \boldsymbol{B} \mid \boldsymbol{C}$$

Some Definitions

- **Minimality** (informal). G is the "smallest graph" that is compatible with P.
- **Faithfulness assumption**. P is faithful to a DAG G iff

$$\boldsymbol{A} \perp\!\!\!\perp_P \boldsymbol{B} \mid \boldsymbol{C} \Rightarrow \boldsymbol{A} \perp\!\!\!\perp_G \boldsymbol{B} \mid \boldsymbol{C}$$

  - Faithfulness and Markov assumption leads to minimality.
- **Markov equivalence**. Set of all minimal DAG G that are Markov compatible to P.

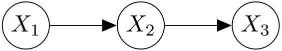


$$P(X_1, X_2, X_3) = P(X_1|X_2)P(X_3|X_2)P(X_2).$$

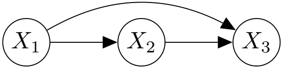
Problem

Structure Identification

Given n i.i.d data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ that are generated from some $P(X_1, \ldots, X_p)$, can we identify a minimal DAG G up to Markov equivalence?



G_1



G_2

If the ground truth $P(X_1, X_2, X_3) = P(X_1|X_2)P(X_3|X_2)P(X_2)$, can we recover G_1 (or its equivalence) from observational data?

Constraint-based Approach: The PC-Algorithm

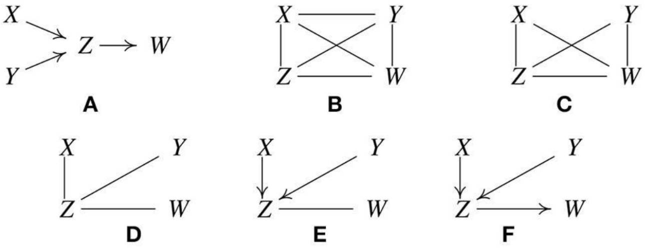


FIGURE 1 | Illustration of how the PC algorithm works. **(A)** Original true causal graph. **(B)** PC starts with a fully-connected undirected graph. **(C)** The $X - Y$ edge is removed because $X \perp\!\!\!\perp Y$. **(D)** The $X - W$ and $Y - W$ edges are removed because $X \perp\!\!\!\perp W \mid Z$ and $Y \perp\!\!\!\perp W \mid Z$. **(E)** After finding v-structures. **(F)** After orientation propagation.

[Glymour et al. 2019]

- Step 1: Identify the skeleton (A-D)
- Step 2: Identify v-structures and orient them (E)
- Step 3: Identify qualifying edges that are incident on collider (F)

Structural Equation Model

- Another representation named Structural Equation Model (SEM) is used to model relationship among variables.

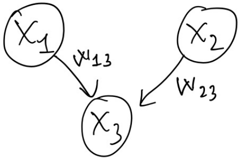
$$X_i = f(\mathsf{Pa}_i, z_i),$$

  where z_i is independent to all variables in Pa_i, and all z_is are mutually independent.
- One popular consideration is linear function, and some/all of z_i follow Gaussian distribution [Loh et al. 2014; Van de Geer et al. 2013].
- In [Zheng et al. 2018], f is assumed as

$$X_i = w_i^\top \mathsf{Pa}_i + z_i$$

  Then a DAG G can be represented by an adjacency matrix $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ such that
  - $w_{ij} \neq 0 \Leftrightarrow (i \to j)$ is an edge in G. Denote such constructed graph $G(\boldsymbol{W})$.
  - $X_i = \boldsymbol{W}(:, i)^\top X + z_i$.

Score-based Approach



A general formulation,

$$\begin{aligned} \underset{G}{\text{maximize}} \quad & s(\boldsymbol{W}) \\ \text{subject to} \quad & G(\boldsymbol{W}) \text{ is a DAG} \end{aligned}$$

- Many score function $s(\cdot)$ have been developed that guarantee identifiability of G, such as Bayesian information criterion (BIC). For example, $\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\|_{\mathrm{F}}^2 + \lambda r(\boldsymbol{W})$ is used in case of Gaussian linear structural model [Van de Geer et al. 2013].
- However, dealing with the constraint is difficult. The problem is NP-hard [M. Chickering et al. 2004].
- A pioneer work is greedy equivalence search (GES) [D. M. Chickering 2002].

Score-based Approach

$$\begin{aligned} \min_{\boldsymbol{W} \in \mathbb{R}^{d \times d}} \quad & s(\boldsymbol{W}) \\ \text{subject to} \quad & G(\boldsymbol{W}) \in \mathtt{DAG} \end{aligned} \quad \Leftrightarrow \quad \begin{aligned} \min_{\boldsymbol{W} \in \mathbb{R}^{d \times d}} \quad & s(\boldsymbol{W}) \\ \text{subject to} \quad & h(\boldsymbol{W}) = 0 \end{aligned}$$

where we wish h to be

- $h(\boldsymbol{W}) = 0$ if and only if $G(\boldsymbol{W})$ is acyclic.
- $h(\boldsymbol{W}) = 0$ measures the "DAG-ness" of the graph.
- $h(\boldsymbol{W})$ is smooth.
- $h(\boldsymbol{W})$ and its derivatives are easy to compute.

Binary Case

Proposition (Infinite series)

Suppose $\boldsymbol{B} \in \{0,1\}^{p \times p}$ *and* $|\lambda_{\max}(\boldsymbol{B})| < 1$. *Then* $G(\boldsymbol{B})$ *is a DAG if and only if*

$$\mathit{tr}(\mathbf{I} - \boldsymbol{B})^{-1} = p.$$

Proof.

- Number of length-2 paths from i to j is $\sum_{t=1}^{p} B(i,t)B(t,j) = \boldsymbol{B}^2(i,j)$.
- Number of length-k paths from i to j is $\boldsymbol{B}^k(i,j)$.
- Number of closed length-k paths from i to i is $\boldsymbol{B}^k(i,i)$.
- Number of closed length-k paths is $\mathrm{tr}(\boldsymbol{B}^k)$.
- A graph is acyclic if and only if $\sum_{k=1}^{\infty} \mathrm{tr}(\boldsymbol{B}^k) = 0$

□

For any square matrix $\boldsymbol{B}$,

$$\begin{aligned}(\boldsymbol{I}-\boldsymbol{B})^{-1} &= \boldsymbol{I}+(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{B} \\ &= \boldsymbol{I}+(\boldsymbol{I}+(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{B})\boldsymbol{B} \\ &= \dots \\ &= \boldsymbol{I}+\boldsymbol{B}+\boldsymbol{B}^2+\dots\end{aligned}$$

$$\mathsf{tr}\left((\boldsymbol{I}-\boldsymbol{B})^{-1}\right) = \mathsf{tr}(\boldsymbol{I}) + \sum_{k=1}^{\infty} \mathsf{tr}(\boldsymbol{B}^k) = p$$

A Better Formula

Proposition

A binary matrix $\boldsymbol{B} \in \{0,1\}^{d\times d}$ *is a DAG if and only if*

$$\mathit{tr}(e^{\boldsymbol{B}}) = d.$$

where

$$e^{\boldsymbol{B}} := \sum_{k=0}^{\infty} \frac{1}{k!}\boldsymbol{B}^k$$

Remark

- $e^{\boldsymbol{B}}$ is always well-defined for all square matrix $\boldsymbol{B}$.
- The equivalence of having no cyclic path and $\mathrm{tr}(\boldsymbol{B}^k) = 0$ for all k only hold if $\boldsymbol{B} > 0$.

Arbitrary Weight Matrix $\boldsymbol{B}$

Theorem

For $\boldsymbol{W} \in \mathbb{R}^{p \times p}$, $G(\boldsymbol{W})$ is a DAG iff

$$h(\boldsymbol{W}) := tr\left(e^{\boldsymbol{W} * \boldsymbol{W}}\right) - d = 0$$

Remark

- Gradient of h is $\nabla h(\boldsymbol{W}) = (e^{\boldsymbol{W} * \boldsymbol{W}})^\top * 2\boldsymbol{W}$.
- Evaluating $e^{\boldsymbol{W}}$ costs $O(p^3)$ [Al-Mohy et al. 2010].

To this end,

$$\begin{aligned} \underset{W}{\text{minimize}} \quad & \frac{1}{2n} \left\| \boldsymbol{X} - \boldsymbol{W}\boldsymbol{X} \right\|_{\mathrm{F}}^2 + \lambda \left\| \boldsymbol{W} \right\|_1 \\ \text{subject to} \quad & \mathrm{tr}(e^{\boldsymbol{W} * \boldsymbol{W}}) = d \end{aligned}$$

and [Zheng et al. 2018] solved it using augmented Lagrange method.

Experiment Result

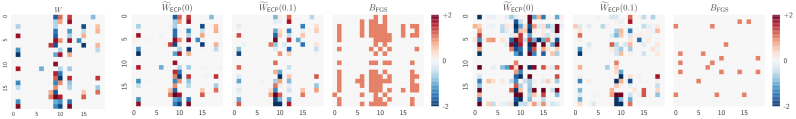
Baseline

- PC-algorithm is excluded since GES and NOTEARS outperforms it significantly.
- A fast version of GES named FGS is used [Ramsey et al. 2017]

Data

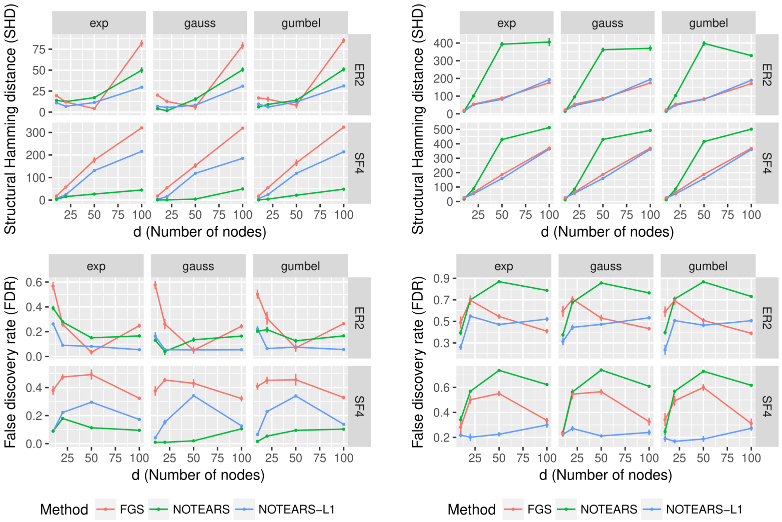
- Generate a random graph G by Erdös-Rényi (ER) or scale-free (SF) model.
- Generate uniform $\boldsymbol{W}$ respect to graph G.
- Sample noise according to Gaussian, Exponential, and Gumble distribution.
- Finally, generate data $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ for $p \in \{10, 20, 50, 100\}$, and $n \in \{20, 10000\}$.

Experiment Result



(a) true graph (b) estimate with $n = 1000$ (c) estimate with $n = 20$

Experiment Result



(a) SHD with $n = 1000$

(b) SHD with $n = 20$

Figure 3: Structure recovery in terms of SHD and FDR to the true graph (lower is better). Rows: random graph types, {ER,SF}-k = {Erdös-Rényi, scale-free} graphs with kd expected edges. Columns: noise types of SEM. Error bars represent standard errors over 10 simulations.

Reference I

[1] David Maxwell Chickering. "Optimal structure identification with greedy search". In: *Journal of machine learning research* 3.Nov (2002), pp. 507–554.

[2] Max Chickering et al. "Large-sample learning of Bayesian networks is NP-hard". In: *Journal of Machine Learning Research* 5 (2004), pp. 1287–1330.

[3] Clark Glymour et al. "Review of causal discovery methods based on graphical models". In: *Frontiers in genetics* 10 (2019), p. 524.

[4] Po-Ling Loh et al. "High-dimensional learning of linear causal networks via inverse covariance estimation". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3065–3105.

[5] Awad H Al-Mohy et al. "A new scaling and squaring algorithm for the matrix exponential". In: *SIAM Journal on Matrix Analysis and Applications* 31.3 (2010), pp. 970–989.

Reference II

[6] Joseph Ramsey et al. “A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images”. In: *International journal of data science and analytics* 3.2 (2017), pp. 121–129.

[7] Sara Van de Geer et al. “ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs”. In: *The Annals of Statistics* 41.2 (2013), pp. 536–567.

[8] Xun Zheng et al. “Dags with no tears: Continuous optimization for structure learning”. In: *Advances in Neural Information Processing Systems* 31 (2018).

Bayesian Network

A Bayesian network is a tuple of 2 components: $U, G =< V, E >$.

- $U = X_1, \ldots, X_p$: set of random variables.
- G is a directed acyclic graph, where vertex V_i represents X_i.

Altogether, a BN defines a joint distribution $P(X_1, \ldots, X_p)$ as

$$P(X_1, \ldots, X_p) = \prod_i^p P(X_i | \mathsf{pa}_i)$$

Assume X is satisfied

$$X_i = w_i^\top \mathsf{pa}_i + z_i$$

where z_i is some noise. All z_i are mutually independent.
Now, given dataset, how do we identify graph G (or find W)?