On Learning Necessary and Sufficient Causal Graphs

Hengrui Cai, Yixin Wang, Michael Jordan, Rui Song (2023+, arXiv)

Presented by Rohit Kanrar

June 14, 2023

Table of Contents

Table of Contents

Preliminaries: Causal Discovery

- Suppose we have p variables; $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$. Consider a Directed Acyclic Graph (DAG), $\mathcal{G} = (\boldsymbol{Z}, E)$, which has edge set E that contains only directed edges and no cyclic path.
- Causal Discovery (or Structure Learning) aims to identify $\widehat{\mathcal{G}}$ based on the data, $\{\boldsymbol{Z}_i\}_{i=1}^{n}$ that closely resembles the true causal relationships between the variables, denoted by $\mathcal{G}$.

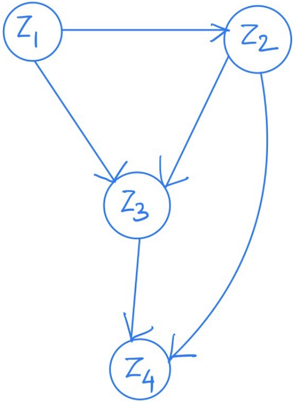

Figure: Example of a DAG

- Here we have 4 variables (or nodes).
- E contains 5 directed edges. There are no cyclic paths.
- If we include another edge, $Z_4 \to Z_1$, then it is no longer a DAG.
- Let $PA_{Z_i}(\mathcal{G})$ is the parent of Z_i in the DAG $\mathcal{G}$. For example, $PA_{Z_1}(\mathcal{G}) = \phi$ and $PA_{Z_3}(\mathcal{G}) = \{Z_1, Z_2\}$.

- **Bayesian Networks:** Consider a parametric density $f(.)$ and a DAG $\mathcal{G} = (\boldsymbol{Z}, E)$. The density factorizes according to a DAG $\mathcal{G}$ if there exists a set of parameter values $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p\}$ such that,

$$f(z_1, \ldots, z_p) = \prod_{i=1}^{p} f_i\left(z_i | PA_{Z_i}(\mathcal{G}) = PA_{z_i}(\mathcal{G}); \boldsymbol{\theta}_i\right)$$

  Such a pair $(\mathcal{G}, \boldsymbol{\Theta})$ is a Bayesian Network that defines the joint distribution.
- The distribution of $\boldsymbol{Z}$ is *DAG-perfect* if there exists a DAG $\mathcal{G}$ such that,
  - i (Markov) Every independence constraints encoded by $\mathcal{G}$ holds in $(\mathcal{G}, \boldsymbol{\Theta})$.
  - ii (Faithfulness) Every independence constraints encoded by $(\mathcal{G}, \boldsymbol{\Theta})$ holds in the DAG $\mathcal{G}$.

- **Linear Structural Equation Model (LSEM):** Suppose $\mathcal{G} = (\boldsymbol{Z}, E)$ is DAG-perfect. Under the LSEM assumption, the following holds,

$$\boldsymbol{Z} = B^T \boldsymbol{Z} + \epsilon$$

where $B = ((b_{i,j}))_{i,j=1}^{p}$ is called the *adjacency matrix*.

- For j-th variable, we have,

$$Z_j = b_{1,j} Z_1 + \ldots + b_{p,j} Z_p + \epsilon_j$$

- $Z_i \rightarrow Z_j$ is present in E iff $b_{i,j} \neq 0$
- Without further assumption on ϵ, LSEM can be identified upto a markov equivalence class of multiple DAGs.
- For the rest of this discussion, we will assume ϵ is *multivariate Gaussian* (LSEM with Gaussian Error) to enforce the identifiability of DAG from the observational data.

- Let us consider a specific case where, we have p variables, $\{Z_1, \ldots, Z_{p-1}, Y\}$, where Y is a response variable and is not expected to causally affect rest of the $p-1$ variables.
- Identifying *full* causal graph based without this added information may lead to spurious causal effects.
- For LSEM with Gaussian Error, we may consider the following parameterization,

$$\begin{bmatrix} \boldsymbol{Z} \\ Y \end{bmatrix} = \begin{bmatrix} B_{\boldsymbol{Z}}^T & 0 \\ \boldsymbol{\theta}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{Z} \\ Y \end{bmatrix} + \begin{bmatrix} \epsilon_{\boldsymbol{Z}} \\ \epsilon_Y \end{bmatrix}$$

where $\boldsymbol{Z} = (Z_1, \ldots, Z_{p-1})^T$.
- However, it only leads to only *sufficient* causal graphs and include many spurious causal variables that does not affect the response.

Table of Contents

Necessary and Sufficient Causal Graph (NSCG)

What is Sufficient Causal Graph? To present the idea, we first define some notations.

- $O = (\boldsymbol{Z}, Y)$ a collection of nodes containing a massive amount of features $\boldsymbol{Z} = [Z_1, \ldots, Z_p]^T \in \mathcal{Z} \subset \mathbb{R}^p$ and the outcome of interest as $Y \in \mathcal{Y} \subset \mathbb{R}$.
- $Y(\boldsymbol{Z} = \boldsymbol{z})$ be the potential value of Y that would be observed after setting variable $\boldsymbol{Z}$ as $\boldsymbol{z}$.
- Similarly, define the potential outcome $Y(Z_i = z_i)$ by setting individual variable Z_i as z_i, while keeping the rest unchanged.

- Let $\boldsymbol{X} \subset \mathcal{X} \in \mathbb{R}^d$ ($d \ll p$) be either subset of $\boldsymbol{Z}$ or $\boldsymbol{X} = f(\boldsymbol{X})$ that indeed captures the causal relationship between $\boldsymbol{Z}$ and Y.
- Here $f : \mathbb{R}^p \to \mathbb{R}^d$ is an unknown function that selects a low-dimensional latent variable $\boldsymbol{X}$. Often, $\boldsymbol{X}$ is treated as a low-dimensional representation of high-dimensional $\boldsymbol{Z}$.
- Denote the full graph, $\mathcal{G}_O = (O, e_O)$ and the sub-graph $\mathcal{G}_V = (V, e_V)$ with $O = (\boldsymbol{Z}, Y)$ and $V = (\boldsymbol{X}, Y)$ are the respective causal nodes and e_V, e_O are independent noise.
- Let $\mathbb{P}_{\mathcal{G}_V}$ and $\mathbb{P}_{\mathcal{G}_O}$ are the density functions under the respective causal graph. e.g., for LSEM with Gaussian error, $\mathbb{P}_{\mathcal{G}_V}$ and $\mathbb{P}_{\mathcal{G}_O}$ will be the Gaussian density function.

Sufficient Graph: $\mathcal{G}_V$ is a sufficient causal graph to capture relationship among $\boldsymbol{Z}$ and Y with $\boldsymbol{X} \subset \boldsymbol{Z}$ or $\boldsymbol{X} = f(\boldsymbol{Z})$ if

$$\mathbb{P}_{\mathcal{G}_V}\{Y|PA_Y(\mathcal{G}_V)\} \prod_{X_i \in PA_Y(\mathcal{G}_V)} \mathbb{P}_{\mathcal{G}_V}\{X_i|PA_{X_i}(\mathcal{G}_V)\}$$
$$= \mathbb{P}_{\mathcal{G}_O}\{Y|PA_Y(\mathcal{G}_O)\} \prod_{Z_i \in PA_Y(\mathcal{G}_O)} \mathbb{P}_{\mathcal{G}_O}\{Z_i|PA_{Z_i}(\mathcal{G}_O)\}$$

- Sufficient graphs may contain spurious relations that are not causally relevant to the outcome Y.
- Necessary graphs avoids such spurious relations.

Necessary Graph: $\mathcal{G}_V$ is a necessary causal graph to capture the causal relationship among $\boldsymbol{Z}$ and Y with $\boldsymbol{X} \subset \boldsymbol{Z}$ or $\boldsymbol{X} = f(\boldsymbol{Z})$ if for any subset $\boldsymbol{W} \subset \boldsymbol{X}$ or $\boldsymbol{W} = g(\boldsymbol{X})$, we have

$$\mathbb{P}_{\mathcal{G}_V}\{Y|PA_Y(\mathcal{G}_V)\} \prod_{X_i \in PA_Y(\mathcal{G}_V)} \mathbb{P}_{\mathcal{G}_V}\{X_i|PA_{X_i}(\mathcal{G}_V)\}$$
$$\neq \mathbb{P}_{\mathcal{G}_U}\{Y|PA_Y(\mathcal{G}_U)\} \prod_{W_i \in PA_Y(\mathcal{G}_U)} \mathbb{P}_{\mathcal{G}_U}\{W_i|PA_{W_i}(\mathcal{G}_U)\}$$

where $\mathcal{G}_U$ is the causal graph for causal nodes $U = (\boldsymbol{W}, Y)$.

- Now, we will discuss some motivating examples on why it is important to incorporate necessity in estimating causal graphs.

Y: Growth yield of yeast; $\boldsymbol{Z}$: Gene expressions

YMR105C $\rightarrow$ YMR090W is a spurious relation that holds no caual impact on Y.

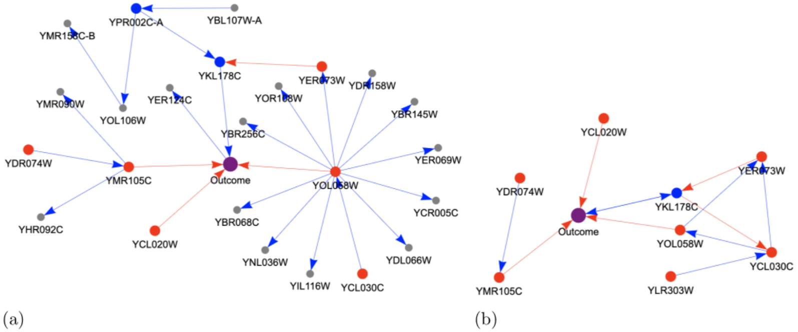

Figure: Causal graphs for candidate genes that affect the growth yield of yeast: (a) a sufficient graph; (b) a necessary and sufficient graph.

Y: Admission outcome of a student in a Graduate program.
$\boldsymbol{Z}$: Other information of the student.

$G \to A$ is a spurious relation that holds no causal impact on Y.

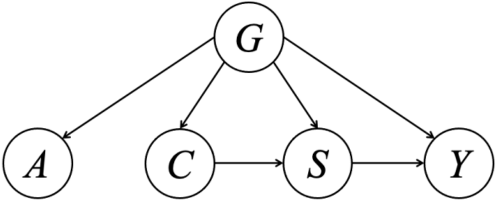

Figure: Causal relationships to understand the effect of gender in Gradate admissions. Node G defines the applicants' gender in the profile; node C is their pre-enrollment career objectives; node S is their choice of the department for study; node A is their appearance in the profile; node Y is the admission outcome.

Table of Contents

How to find a Necessary and Sufficient Causal Graph (NSCG)?

- It is difficult to find an NSCG directly using the definitions, primarily due to the latent representation $\boldsymbol{X}$ is unknown.
- A naive approach is to force search all different combinations of $\boldsymbol{Z}$ to find a representation $\boldsymbol{X}$ that satisfies the definitions; it is a nondeterministic polynomial (NP) hard problem.
- Instead, the authors suggest to evaluate the necessity and sufficiency of features on the prediction of outcome, using a property known as the probability of causation (POC).

Assumptions

Before presenting the definition of POC, we now present the assumptions needed for identification of causal quantities.

A1 **Consistency:** (or SUTVA)

$$\boldsymbol{Z} = \boldsymbol{z} \leftrightarrow Y(\boldsymbol{Z} = \boldsymbol{z}) = Y,\ \forall\ \boldsymbol{z} \in \mathcal{Z} \tag{1}$$

A2 **Ignorability:** (or NUC)

$$\left\{Y(\boldsymbol{Z} = \boldsymbol{z}), Y(\boldsymbol{Z} = \boldsymbol{z}^{'})\right\} \perp \boldsymbol{Z}$$
$$\left\{Y(Z_i = z_i), Y(Z_i = z_i^{'})\right\} \perp Z_i | PA_{Z_i \cup Y}(\mathcal{G}_O)$$

A3 **Monotonicity:**

$$\{Y(\boldsymbol{Z} \neq \boldsymbol{z}) = y\} \wedge \{Y(\boldsymbol{Z} = \boldsymbol{z}) \neq y\} = False$$
$$\{Y(Z_i \neq z_i) = y\} \wedge \{Y(Z_i = z_i) \neq y\} = False$$

Probability of Causation (POC) in Tian and Pearl (2000)

POC for the feature $\boldsymbol{Z}$ can be quantified using the probability of necessity and sufficiency.

Probability of Necessity and Sufficiency (PNS):

$$\begin{aligned} PNS &\stackrel{\text{def}}{=} \mathbb{P}\{Y(\boldsymbol{Z} \neq z) \neq y, Y(\boldsymbol{Z} = \boldsymbol{z}) = y\} \\ &\stackrel{\text{by(A1)}}{=} \mathbb{P}\{\boldsymbol{Z} = \boldsymbol{z}, Y = y\} \; PN + \mathbb{P}\{\boldsymbol{Z} \neq \boldsymbol{z}, Y \neq y\} \; PS \end{aligned}$$

where,

- the probability of necessity (PN) is defined as

$$PN \stackrel{\text{def}}{=} \mathbb{P}\{Y(\boldsymbol{Z} \neq \boldsymbol{z}) \neq y | \boldsymbol{Z} = \boldsymbol{z}, Y = y\}$$

- the probability of sufficiency (PS) is defined as

$$PS \stackrel{\text{def}}{=} \mathbb{P}\{Y(\boldsymbol{Z} = \boldsymbol{z}) = y | \boldsymbol{Z} \neq \boldsymbol{z}, Y \neq y\}$$

Intuition for PN and PS

Consider the feature $\mathbb{I}\{\boldsymbol{Z} = \boldsymbol{z}\}$ and the following notions,

$$Y \begin{cases} = y & \text{is a "good" outcome} \\ \neq y & \text{is a "bad" outcome} \end{cases}$$

and

$$\boldsymbol{Z} \begin{cases} = \boldsymbol{z} & \text{is the feature being present} \\ \neq \boldsymbol{z} & \text{is the feature being absent} \end{cases}$$

- **Intuition of PN:** Given the good outcome observed (with the feature being present), PN measures the probability of a bad outcome after excluding the feature.

$$PN \stackrel{\text{def}}{=} \mathbb{P}\{Y(\boldsymbol{Z} \neq \boldsymbol{z}) \neq y | \boldsymbol{Z} = \boldsymbol{z}, Y = y\}$$

- **Intuition of PS:** Given the bad outcome observed (with the feature being absent) PS measures the probability of a good outcome after including the feature.

$$PS \stackrel{\text{def}}{=} \mathbb{P}\{Y(\boldsymbol{Z} = \boldsymbol{z}) = y | \boldsymbol{Z} \neq \boldsymbol{z}, Y \neq y\}$$

POC for Individual Feature Z_i

PNS can be further generalized to quantify the POC of an individual feature Z_i.

Conditional POC (C-POC):

$$C - POC_i \stackrel{\text{def}}{=} \mathbb{P}\left\{Y\left(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}\right) \neq \boldsymbol{y}, Y\left(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}\right) = \boldsymbol{y}\right\}$$

Marginal POC (M-POC):

$$M - POC_i \stackrel{\text{def}}{=} \mathbb{P}\left\{Y\left(Z_i \neq z_i\right) \neq \boldsymbol{y}, Y\left(Z_i \neq z_i\right) = \boldsymbol{y}\right\}$$

where, $\boldsymbol{Z}_{-i} \stackrel{\text{def}}{=} \boldsymbol{Z} \backslash Z_i$ be the complementary variable set of Z_i.

Identifiability of C-POC & M-POC

Theorem 4.4: Suppose (A1)-(A3) hold. Then the POC of an individual feature Z_i are all indentifiable as,

$$C - POC_i = \mathbb{P}(Y = y|Z_i = z_i, \boldsymbol{Z}_{-i} \neq \boldsymbol{z}_{-i}) - \mathbb{P}(Y = y|Z_i \neq z_i, \boldsymbol{Z}_{-i} \neq \boldsymbol{z}_{-i})$$

$$M - POC_i = \mathbb{P}(Y = y|Z_i = z_i) - \mathbb{P}(Y = y|Z_i \neq z_i)$$

- Theorem 4.4 enables us to estimate POC from the observed data.
- But, estimating the conditional probabilities of Y based on high-dimensional features is challenging.
- This motivates the authors to consider expected mean outcome given different combinations of the confounders.

POC and Causal Effects

Corollary 4.5: Suppose (A1)-(A3) hold. Then we have

$$\int_{y\in\mathcal{Y}} y\ M-POC_i\ dy = \mathbb{E}\{Y(Z_i = z_i)\} - \mathbb{E}\{Y(Z_i \neq z_i)\} \overset{\text{def}}{=} \delta_M(Z_i)$$

$$\begin{aligned}\int_{y\in\mathcal{Y}} y\ C-POC_i\ dy &= \mathbb{E}\{Y(Z_i = z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-1})\} - \\ &\qquad \mathbb{E}\{Y(Z_i \neq z_i, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-1})\} \\ &\overset{\text{def}}{=} \delta_C(Z_i)\end{aligned}$$

where $\delta_M(Z_i)$ and $\delta_C(Z_i)$ are defined as the marginal and conditional causal effects using the differences of expectations based on the corresponding POC.

Causal Effects for Non-binary Confounders

- The causal effects $\delta_M(.)$ and $\delta_C(.)$ are related to the POC only for binary confounders and positive outcome.
- To generalize the notion, authors extended the idea of total and direct effect for a variable of interest Z_i originally defined by Pearl (2009).

Natural Causal Effects:

$$TE_i = \partial \mathbb{E}\left\{Y(Z_i = z_i^{'})\right\} / \partial z_i^{'}$$
$$DE_i = \partial \mathbb{E}\left\{Y(Z_i = z_i^{'}, \boldsymbol{Z}_{-i} = \boldsymbol{z}_{-i}^{(z_i)})\right\} / \partial z_i^{'}$$

where $\boldsymbol{z}_{-i}^{(z_i)}$ is the value of $\boldsymbol{Z}_{-i}$ if setting $do(Z_i = z_i)$.
Note: For binary confounders, the causal effects δ_M and δ_C match TE and DE, respectively.

Table of Contents

Necessary and Sufficient Causal Structure Learning (NSCSL)

- A simple solution is to conduct a pre-screening process to find $\widehat{\boldsymbol{X}} \subset \boldsymbol{Z}$, which achieve high C-POC/M-POC.
- Followed by employing a causal discovery method to obtain $\widehat{\mathcal{G}} = \left(\left\{\widehat{\boldsymbol{X}}, Y\right\}, e\right)$ to approximate the NSCG, $\mathcal{G}_V$.
- Instead of such a two-step approach, the authors suggested a single-step procedure, NSCSL, which we will present now.

- Let $g : \mathbb{R}^p \to \mathbb{R}^d$ ($d \ll p$) be a selector function.
- We assume the nodes $\{g(\boldsymbol{Z}), Y\}$ have the following causal structure.

$$\begin{bmatrix} g(\boldsymbol{Z}) \\ Y \end{bmatrix} = \begin{bmatrix} B_{\boldsymbol{Z}}^T & 0 \\ \boldsymbol{\theta}^T & 0 \end{bmatrix} \begin{bmatrix} g(\boldsymbol{Z}) \\ Y \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{\boldsymbol{Z}} \\ \epsilon_Y \end{bmatrix}$$

  where $\boldsymbol{\theta} \in \mathbb{R}^d$, $B_{\boldsymbol{Z}} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{\epsilon}$ is a $d+1$ dimensional random vector of joint independent error.

Note: This is the LSEM with Gaussian error assumption. Additionally, we impose the constraint of no descendent of Y.

How to estimate the POC for i-th selected feature $g_i(\boldsymbol{Z})$?

- Estimation of $\mathbb{P}(Y = y|g_i(\boldsymbol{Z}) = z_i)$ can be achieved by either parametric models (e.g., logistic for binary outcome) or non-parametric models (e.g., random forest/neural network)
- Data: $\{o^{(j)} \;=\; z^{(j)}, y^{(j)}\}_{1\leq j\leq n}$.
- The marginal POC can be estimated as,

$$\widehat{M-POC}\left(g_i|\left\{o^{(j)}\right\}\right) = \prod_{j=1}^{n}\Big|\widehat{\mathbb{P}}\left\{Y = y^{(j)}|g_i(\boldsymbol{Z}) = g_i(\boldsymbol{z}^{(j)})\right\} - \widehat{\mathbb{P}}\left\{Y = y^{(j)}|g_i(\boldsymbol{Z}) \neq g_i(\boldsymbol{z}^{(j)})\right\}\Big|$$

- Similarly, the conditional POC can be estimated as,

$$\widehat{C-POC}\left(g_i \middle| \left\{o^{(j)}\right\}\right)$$
$$= \prod_{j=1}^{n} \left| \widehat{\mathbb{P}}\left\{Y = y^{(j)} | g_i(\boldsymbol{Z}) = g_i(\boldsymbol{z}^{(j)}), g_{-i}(\boldsymbol{Z}) = g_{-i}(\boldsymbol{z}_{-i}^{(j)})\right\} - \right.$$
$$\left. \widehat{\mathbb{P}}\left\{Y = y^{(j)} | g_i(\boldsymbol{Z}) \neq g_i(\boldsymbol{z}^{(j)}), g_{-i}(\boldsymbol{Z}) = g_{-i}(\boldsymbol{z}_{-i}^{(j)})\right\} \right|$$

where $g_{-i}(.) \stackrel{\text{def}}{=} g(.) \backslash g_i(.)$ is the complementary of $g_i(.)$.
- Even for $d \ll p$, the estimation of M-POC/C-POC are difficult (Wang and Jordan (2022)).
- Authors proposed to use TE/DE as it has close form expression under LSEM.

Close Form Expressions of TE and DE under LSEM: Recall,

$$\begin{bmatrix} g(\boldsymbol{Z}) \\ Y \end{bmatrix} = \begin{bmatrix} B_{\boldsymbol{Z}}^T & 0 \\ \boldsymbol{\theta}^T & 0 \end{bmatrix} \begin{bmatrix} g(\boldsymbol{Z}) \\ Y \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{\boldsymbol{Z}} \\ \epsilon_Y \end{bmatrix}$$

- Then we have,

$$DE_i = \theta_i$$

  as θ_i presents the weight of the direct edge $g_i(\boldsymbol{Z}) \to Y$.
- Total causal effect TE_i of the i-th selected feature $g_i(\boldsymbol{Z})$ can be expressed simplified using the path method, which we describe next.

- Let $\pi_i = \{g_i(\boldsymbol{Z}) \to \ldots \to Y\}$ be the set of directed paths that starts with $g_i(\boldsymbol{Z})$ and ends with Y. Suppose there are m_i directed paths in π_i.
- The causal effect of $g_i(\boldsymbol{Z})$ on Y through the directed path $\pi_i^{(k)} = \{i, l_1, \ldots, l_{e_k}, d+1\} \in \pi_i$ with length $e_k + 1$ is

$$PE\left\{\pi_i^{(k)}\right\} = \prod b_{i,l_1} \ldots b_{l_{e_k-1},l_{e_k}} \theta_{l_{e_k}}$$

where $B_{\boldsymbol{Z}} = ((b_{i,j}))_{p\times p}$.
- Then, the total effect of $g_i(\boldsymbol{Z})$ on Y is defined as,

$$TE_i = \sum_{k=1}^{m_i} PE\left\{\pi_i^{(k)}\right\}$$

NSCSL: Learning Algorithm

- **Step 1: Causal Discovery**
  - Main idea is to estimate B with the acyclicity constraint.
  - Authors proposed to use the following constraint proposed by Yu et al. (2019).

$$h_1(B) \stackrel{\text{def}}{=} tr\left[(I_{d+1} + tB \circ B)^{d+1}\right] - (d+1)$$

    where $B \circ B$ is the element-wise square operator.
  - Theorem 1 (Yu et al. (2019)) $\implies$ $h_1(B) = 0$ iff B is acyclic.
  - First part of the proposed loss,

$$L_1(B, g, \theta, \lambda_1 | \left\{o^{(j)}\right\}) = f(B, g, \theta | \left\{o^{(j)}\right\}) + \lambda_1 h_1(B)$$

    where $f(.)$ is any suitable loss function (such as least square for LSEM with Gaussian error).

- **Step 2: Constraint based on POC/Natural Effects**
  - Based on POC:

$$L_2^P(B, g, \gamma | \left\{ o^{(j)} \right\}) = -\sum_{i=1}^{d} \widehat{P}(g_i | \left\{ o^{(j)} \right\}) + \gamma R(g)$$

    where $\widehat{P}$ is either $C \widehat{- PO}C$ or $M \widehat{- PO}C$.
  - Based on Natural Effects:

$$L_2^{CE}(B, g, \gamma | \left\{ o^{(j)} \right\}) = -\sum_{i=1}^{d} \widehat{CE}_i(B) + \gamma R(g)$$

    where $\widehat{CE}_i$ is either DE_i or TE_i.
- Here γ is a penalty term and $R(.)$ is some norm to control the complexity of $g(.)$.

- Finally, the proposed optimization problem is,

$$\min_{B,g} f\left(B, g, \theta | \left\{o^{(j)}\right\}\right) + \lambda_1 h_1(B) + \gamma R(g)$$

- The estimated NSCG, $\widehat{g}_V$ can be obtained by using the estimated adjacency matrix, $\widehat{B}$.
- Authors proposed to use a black-box stochastic optimizer, Adam for estimation purposes.

References

Pearl, J. (2009). "Causal inference in statistics: An overview."

Tian, J. and Pearl, J. (2000). "Probabilities of causation: Bounds and identification." *Annals of Mathematics and Artificial Intelligence*, **28(1-4)**, 287–313.

Wang, Y. and Jordan, M.I. (2022). "Desiderata for representation learning: A causal perspective."

Yu, Y., Chen, J., Gao, T., and Yu, M. (2019). "Dag-gnn: Dag structure learning with graph neural networks." In "International Conference on Machine Learning," pages 7154–7163. PMLR.

Thank You