# Causal Discovery with Latent Confounders Based on Higher-Order Cumulants

Ruichu Cai, Zhiyi Huang, Wei Chen, Zhifeng Hao, Kun Zhang

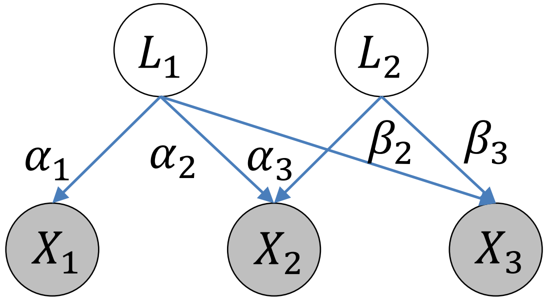
# Motivation

- Causal discovery with latent confounders is an important but challenging task
  - usually unable to collect or measure all the underlying causal variables.
  - Such latent confounders generally pose serious identifiability problems in causal discovery.
- Latent-Variable Linear, non-Gaussian Acyclic Model (lvLiNGAM): based on Overcomplete Independent Component Analysis (OICA)
  - OICA is hard to compute and may get stuck in local optima.

Observational Data

| | $X_1$ | $X_2$ | ,,, | $X_n$ |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | | | |
| … | | | | |
| … | | | | |
| m | | | | |

Causal Discovery




How to learn causal structure in the presence of latent confounders?

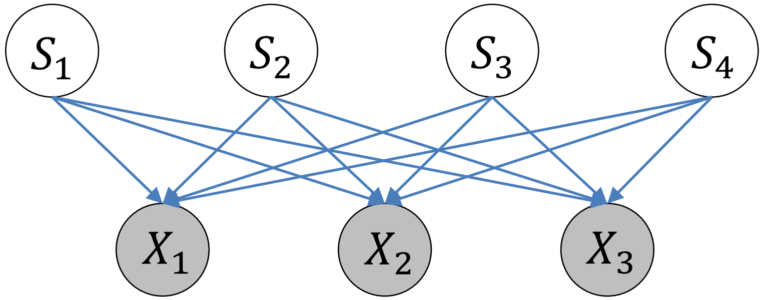
# Canonical lvLiNGAM

- We focus on causal discovery with latent confounders in the linear, non-Gaussian case, where the data generating process can be modeled as:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\Lambda}\boldsymbol{L} + \boldsymbol{S}_{\boldsymbol{X}}.$$

- The above model can be solve by Overcomplete Independent Component Analysis(OICA), which can be rewritten as:

$$\boldsymbol{X} = [(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Lambda} \quad (\boldsymbol{I} - \boldsymbol{B})^{-1}] \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{S}_{\boldsymbol{X}} \end{bmatrix} = \boldsymbol{A}\boldsymbol{S}.$$




$\boldsymbol{S}$: Independent components

$\boldsymbol{X}$ :Observed variables

# Higher-Order Cumulants

- **Cumulant**: The k-th cumulant tensor of a random vector $Z = (Z_1, \dots, Z_p)$ is the $p \times \cdots \times p$ ($k$ times) table with entry at position $(i_1, \dots, i_k)$ given by

$$cum(Z_{i1}, \dots, Z_{i_k}) = \sum_{(D_1,\dots,D_h)} (-1)^{h-1}(h-1)!\, \mathbb{E}\left[\prod_{j \in D_1} Z_j\right] \mathbb{E}\left[\prod_{j \in D_2} Z_j\right] \dots \mathbb{E}\left[\prod_{j \in D_h} Z_j\right],$$

where the sum is taken over all partitions $(D_1, \dots, D_h)$ of the set $\{i_1, \dots, i_k\}$.

- Two important properties of cumulants
  - **Linearity**: for any two random variables $X_i, X_j$ and constants $a, b$,

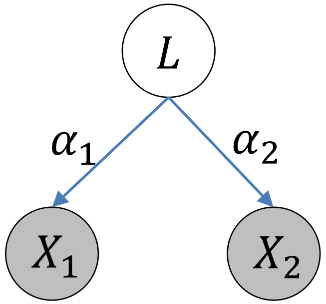
$$cum(X_i, X_i, X_i, aX_i + bX_j) = a\, cum(X_i, X_i, X_i, X_i) + b\, cum(X_i, X_i, X_i, X_j)$$

  - **Independence:** If any of the random variables $X_i, X_j, X_k, X_l$ is statistically independent of the others then $cum(X_i, X_j, X_k, X_l) = 0$

# Closed-Form Solution to OICA in Specific Cases

➢ One-Latent-Component structure

## Generation Process




$$X_1 = \alpha_1 L + S_1$$
$$X_2 = \alpha_2 L + S_2$$

## Expanding Cumulant

$$cum(X_1, X_1, X_1, X_2) = \alpha_1^3 \alpha_2 cum(L, L, L, L)$$
$$cum(X_1, X_1, X_2, X_2) = \alpha_1^2 \alpha_2^2 cum(L, L, L, L)$$
$$cum(X_1, X_2) = \alpha_1 \alpha_2 cum(L, L)$$

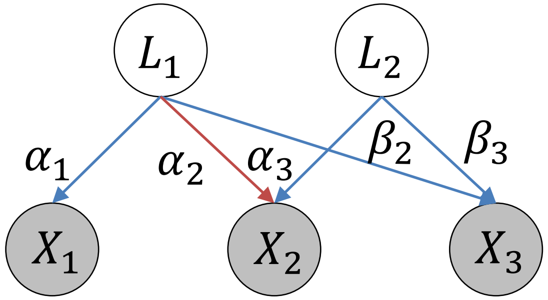
## Closed-Form Solution

$$\hat{\alpha}_1 = \sqrt{\frac{cum(X_1, X_1, X_1, X_2)}{cum(X_1, X_1, X_2, X_2)} \cdot cum(X_1, X_2)} = |\alpha_1| \sqrt{cum(L, L)}, \qquad \hat{\alpha}_2 = \frac{cum(X_1, X_2)}{\hat{\alpha}_1}$$

# Closed-Form Solution to OICA in Specific Cases

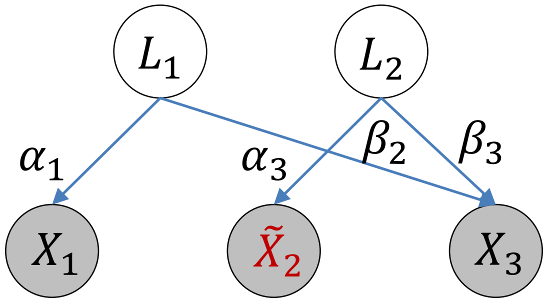
- More complex cases can be solved by eliminating independent components.

$X_2$ and $X_3$ are affected by independent components $L_1$ and $L_2$




$$\tilde{X}_2 = X_2 - \frac{\alpha_2}{\alpha_1} X_1$$

eliminating independent components




$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & 1 & 0 & 0 \\ \alpha_2 & \beta_2 & 0 & 1 & 0 \\ \alpha_3 & \beta_3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

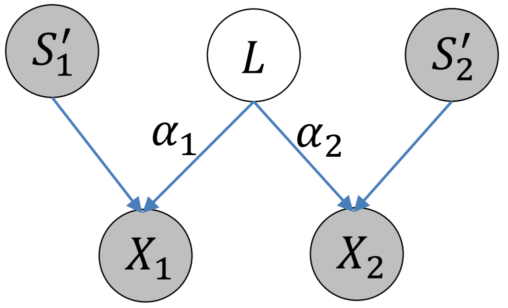
$$\begin{bmatrix} X_1 \\ \tilde{X}_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 & 1 & 0 & 0 \\ 0 & \beta_2 & -\frac{\alpha_2}{\alpha_1} & 1 & 0 \\ \alpha_3 & \beta_3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ E_1 \\ E_2 \\ E_3 \end{bmatrix}$$

# One-Latent-Component condition

- A testable framework: One-Latent-Component Condition

## Hypothesis

One-Latent-Component Structure




$$X_1 = \alpha_1 L + S_1'$$
$$X_2 = \alpha_2 L + S_2'$$

## Statistics

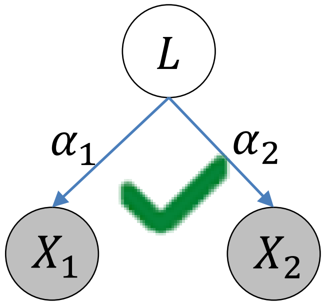
**One-Latent-Component Condition**

$$\boldsymbol{S_1' \perp S_2'}$$

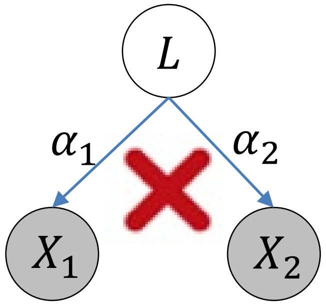
Accept

Reject

Accept
Hypothesized Structure



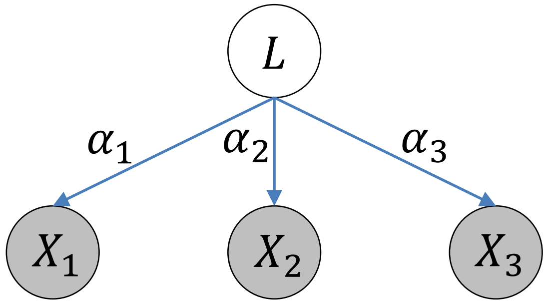

Reject
Hypothesized structure



# Identify the Latent Confounders

- Identify the latent confounders by using the one-latent-component condition.

$(\{X_3\}, \{X_1, X_2\})$ satifies One-Latent-Component condition



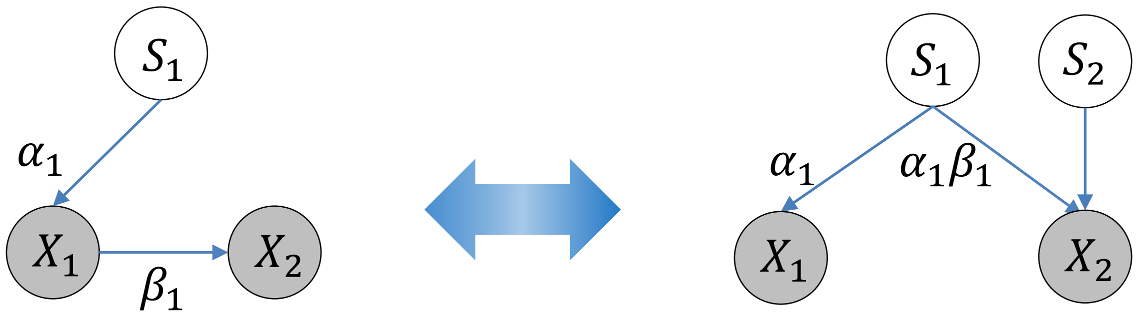

There is shared latent variable for $X_1, X_2, X_3$

**Identify the latent confounders based on one-latent-component structure**

# Determine the Causal Orders

- Determine the causal orders by using the one-latent-component condition.

$(\{X_1\},\{X_1,X_2\})$ satifies One-Latent-Component condition




There is a directed edge between $X_1$ and $X_2$

$X_1$ and $X_2$ share one latent component

**Determine the causal orders based on the special one-latent-component structure**

# Conclusion and Future Research

**Conclusion**

- We proposed a closed-form solution to OICA in specific cases, which explicitly estimates the mixing matrix using the higher-order cumulants.
- We further extended the results to estimate a canonical lvLiNGAM, by iteratively employing the proposed One-Latent-Component condition to test for the existence of latent confounders and determine the causal directions between observed variables.

**Future Research**

- Extend our methods to handle nonlinear causal relationships.
- Relax the independence assumption among the latent variables.