Generalized Independent Noise Condition for Estimating Latent Variable Causal Graphs

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, Kun Zhang

Mohammadsajad Abavisani - March 5 2021

Introduction

- Most causal discovery approaches focus on the situation without latent variables.
- In practice, observed variable are not all the underlying causes.
- The causal model is linear acyclic graph where:

$\mathbf{V} = \mathbf{X} \cup \mathbf{L}$ | X= observed variables, L= latent variables

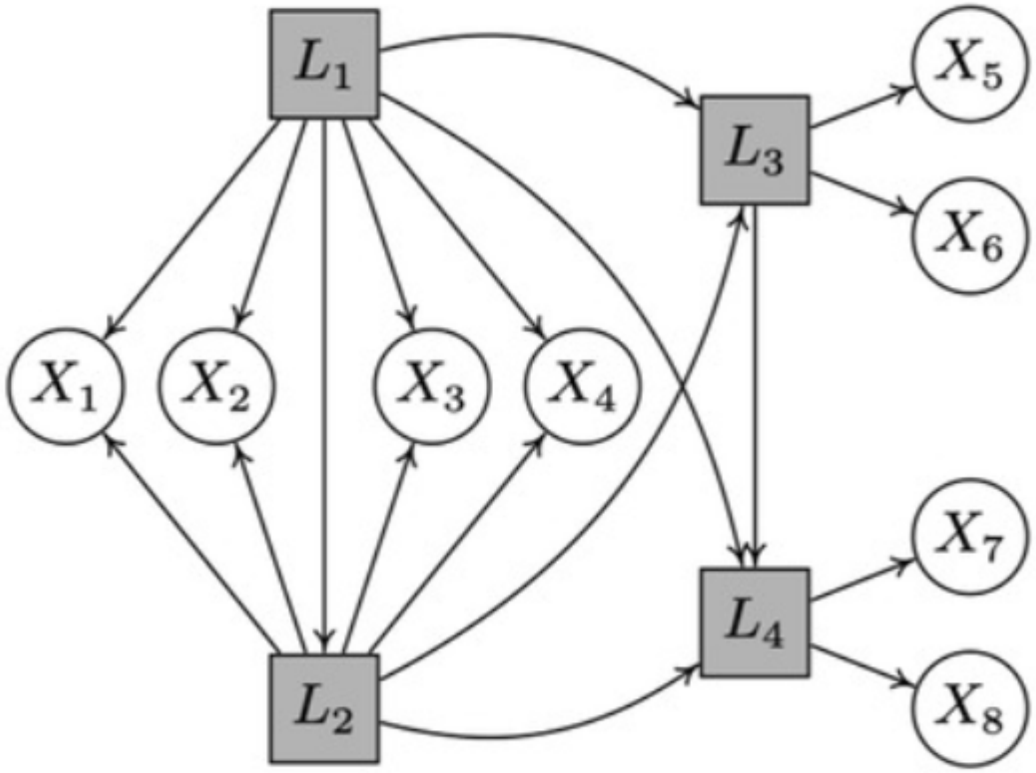
- Variables are generate using this process :

$$V_i = \sum_{k(j)<k(i)} b_{ij} V_j + \varepsilon_{V_i}, i = 1, 2, ..., m + n,$$

- k(i) is the causal order of variables and $b_{i,j}$ is the strengths of link and ϵv_i is and independent noise term.

Problem Definition

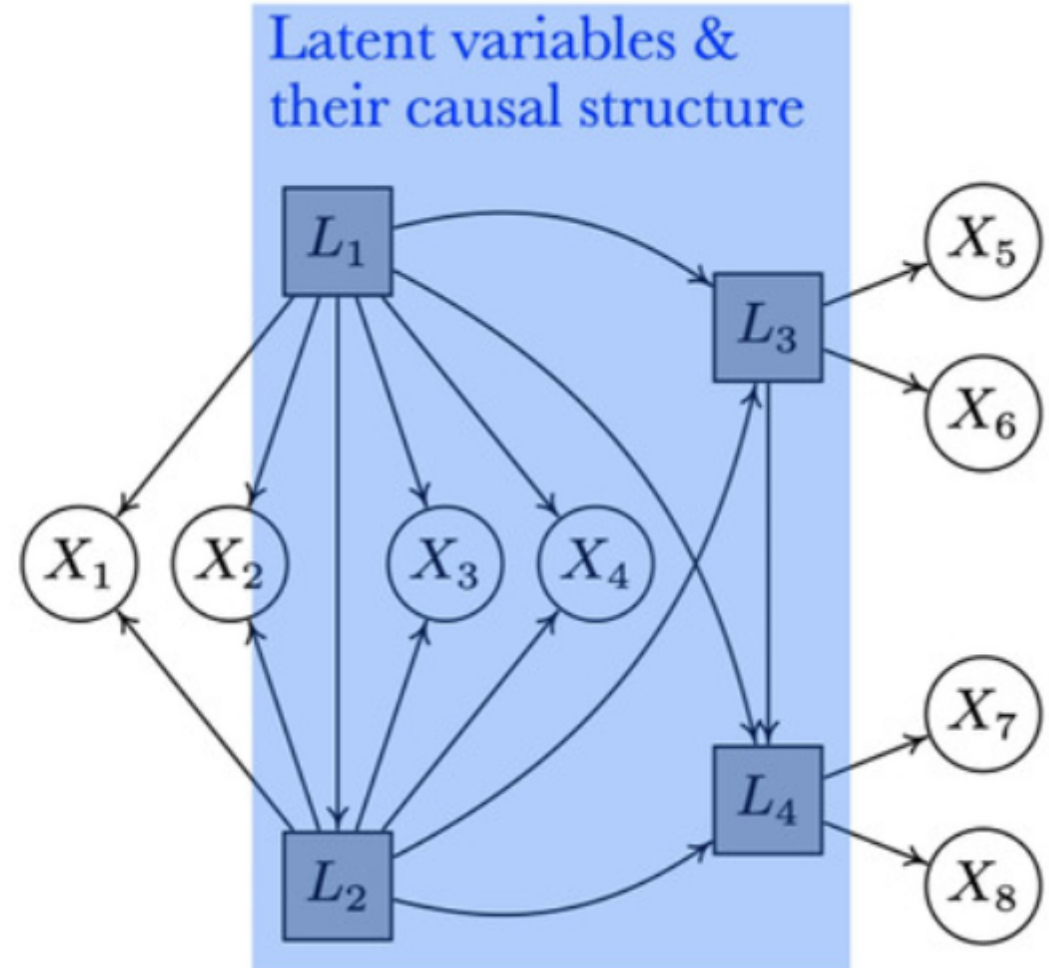
X1	X2	X3	X4	X5	X6	X7	X8
4.2	3.6	6.5	6.8	9.6	7.6	2.7	4.8
3.8	1.9	6.5	7.3	8.9	6.9	1.1	4.6
4.2	3.4	6.5	6.9	9.5	7.4	2.5	4.6
4.2	2.2	6.2	6.9	9.6	7.2	1.9	4.8
3.9	1.9	6.5	6.8	9.0	6.8	1.7	4.4
4.0	2.0	6.4	7.2	9.1	7.0	1.0	4.6
3.8	1.7	6.4	7.3	9.0	6.7	0.8	4.3
4.1	2.8	6.5	6.9	9.3	6.7	2.7	4.6
...	...	...	...	...	...	...	...



- Find latent variables L_i and their causal relations from measured variables X_i?

Problem Definition

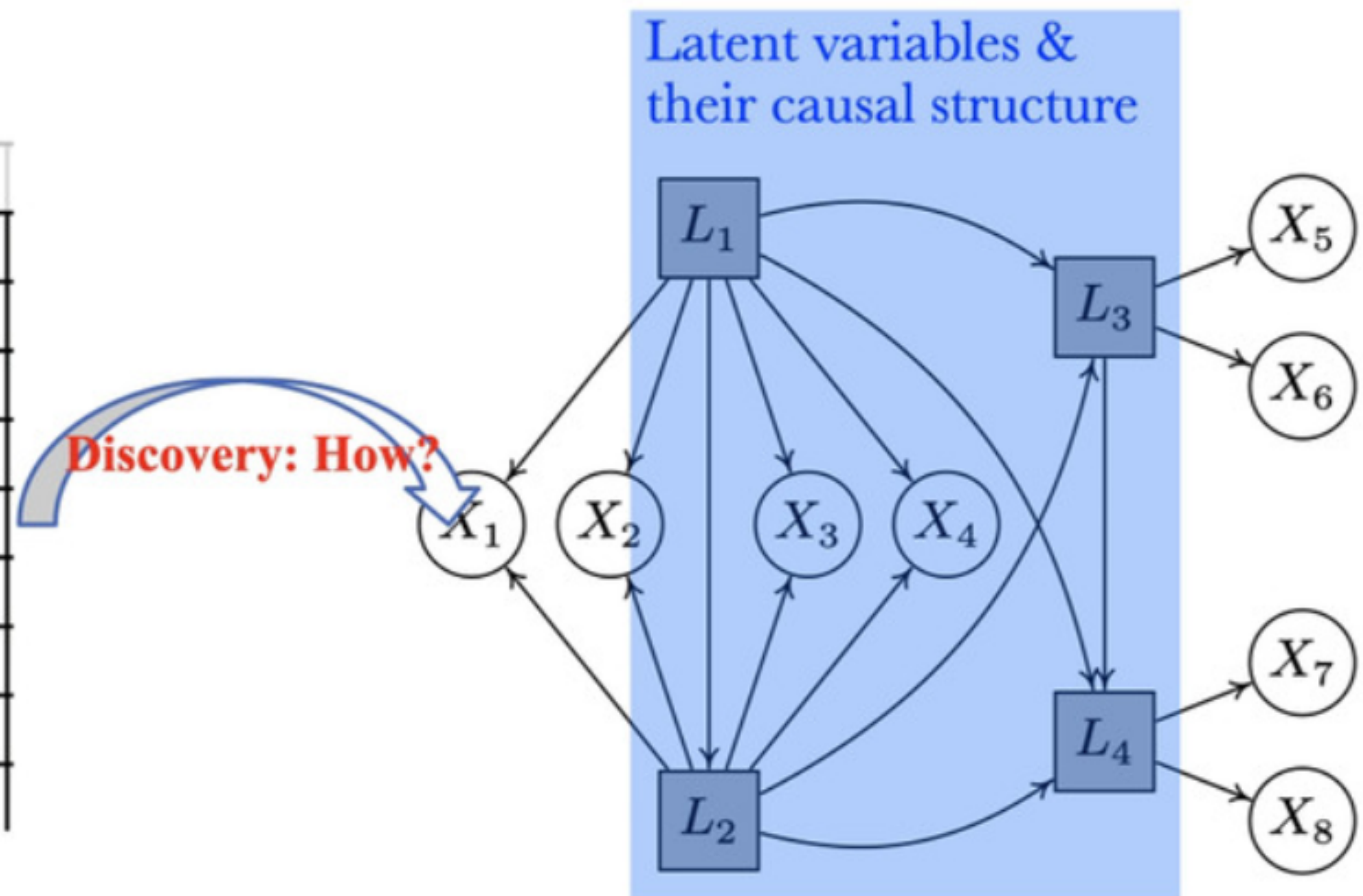
X1	X2	X3	X4	X5	X6	X7	X8
4.2	3.6	6.5	6.8	9.6	7.6	2.7	4.8
3.8	1.9	6.5	7.3	8.9	6.9	1.1	4.6
4.2	3.4	6.5	6.9	9.5	7.4	2.5	4.6
4.2	2.2	6.2	6.9	9.6	7.2	1.9	4.8
3.9	1.9	6.5	6.8	9.0	6.8	1.7	4.4
4.0	2.0	6.4	7.2	9.1	7.0	1.0	4.6
3.8	1.7	6.4	7.3	9.0	6.7	0.8	4.3
4.1	2.8	6.5	6.9	9.3	6.7	2.7	4.6
...	...	...	...	...	...	...	...



- Find latent variables L_i and their causal relations from measured variables X_i?

Problem Definition

X1	X2	X3	X4	X5	X6	X7	X8
4.2	3.6	6.5	6.8	9.6	7.6	2.7	4.8
3.8	1.9	6.5	7.3	8.9	6.9	1.1	4.6
4.2	3.4	6.5	6.9	9.5	7.4	2.5	4.6
4.2	2.2	6.2	6.9	9.6	7.2	1.9	4.8
3.9	1.9	6.5	6.8	9.0	6.8	1.7	4.4
4.0	2.0	6.4	7.2	9.1	7.0	1.0	4.6
3.8	1.7	6.4	7.3	9.0	6.7	0.8	4.3
4.1	2.8	6.5	6.9	9.3	6.7	2.7	4.6
...	...	...	...	...	...	...	...

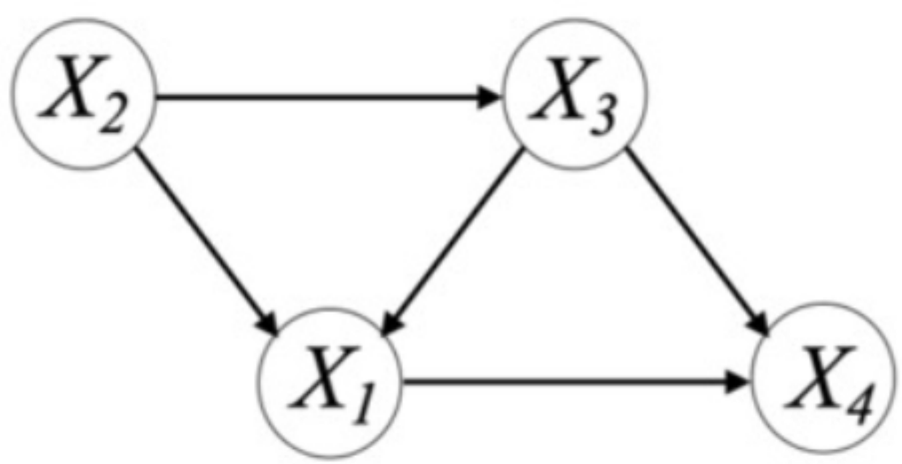


- Find latent variables L_i and their causal relations from measured variables X_i?

Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

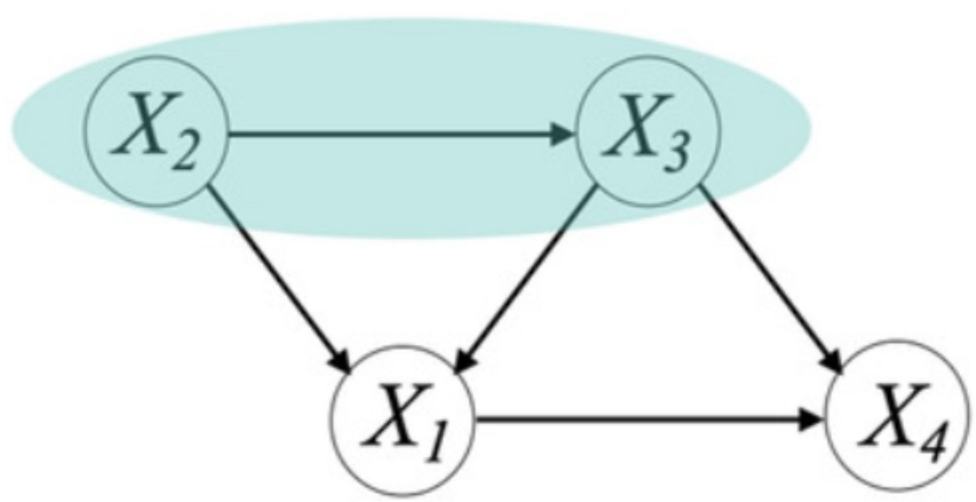
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

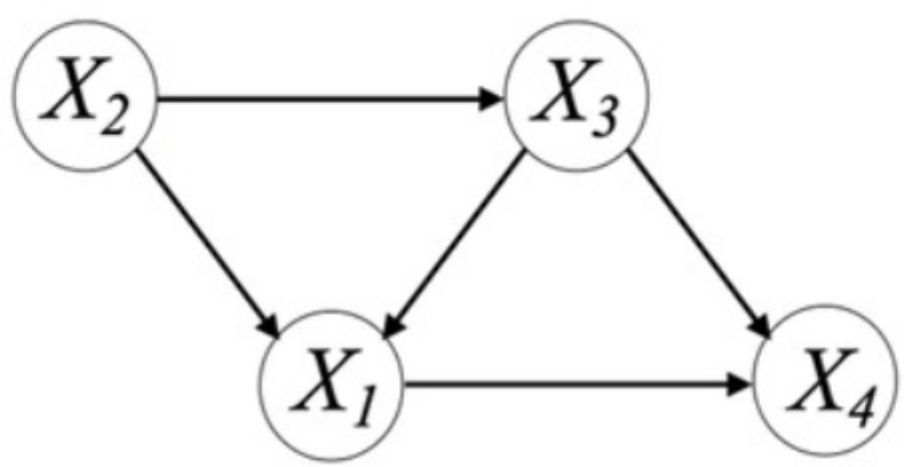
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^{\intercal}\mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

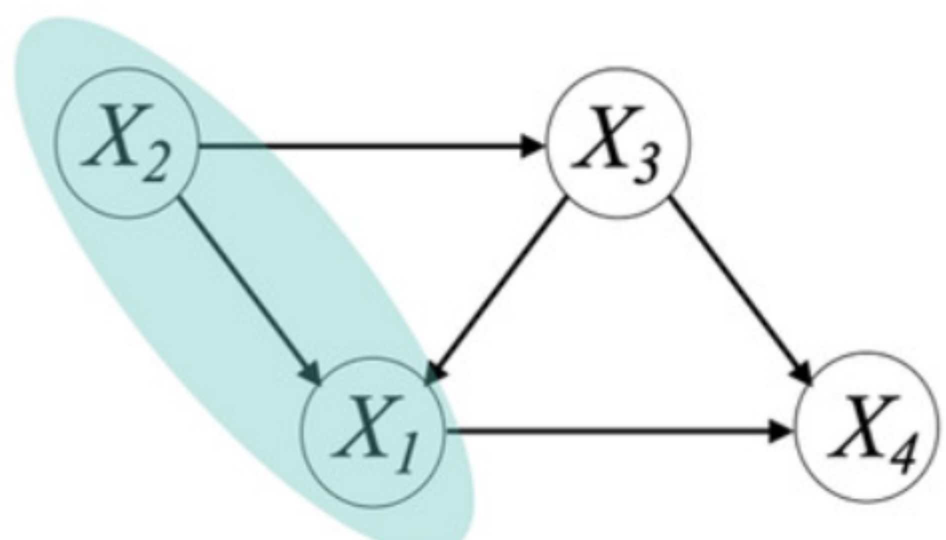
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

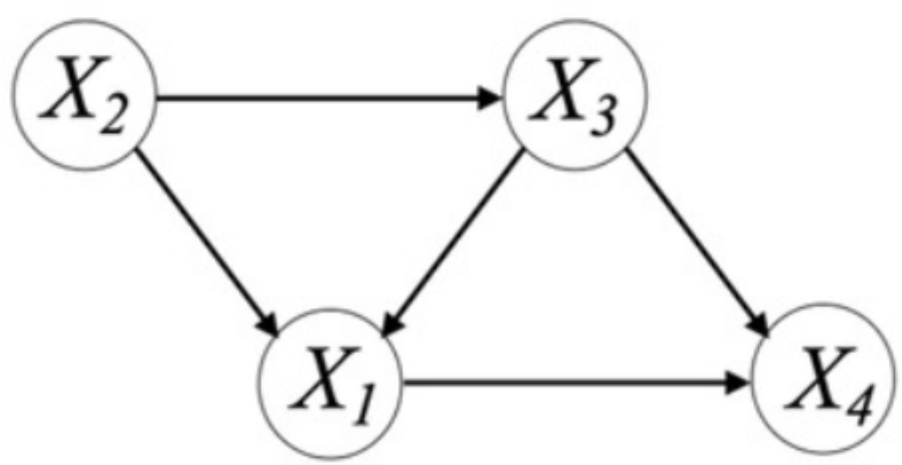
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

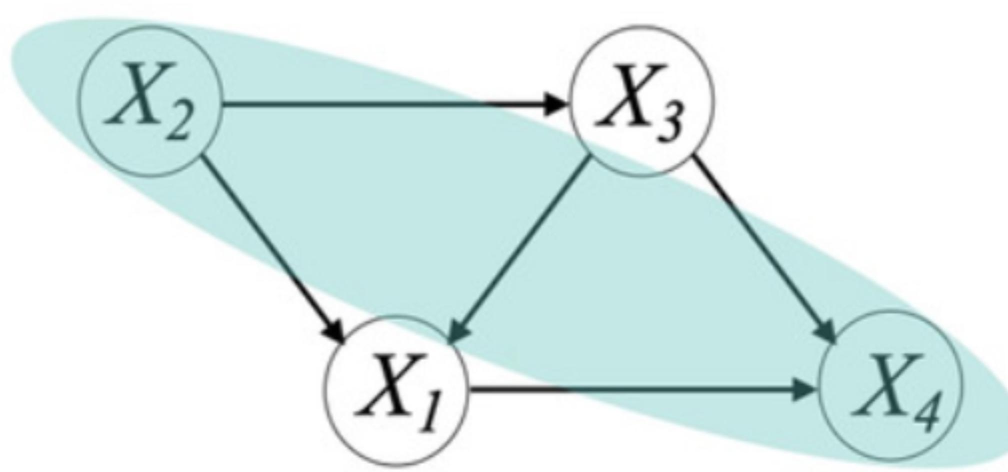
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

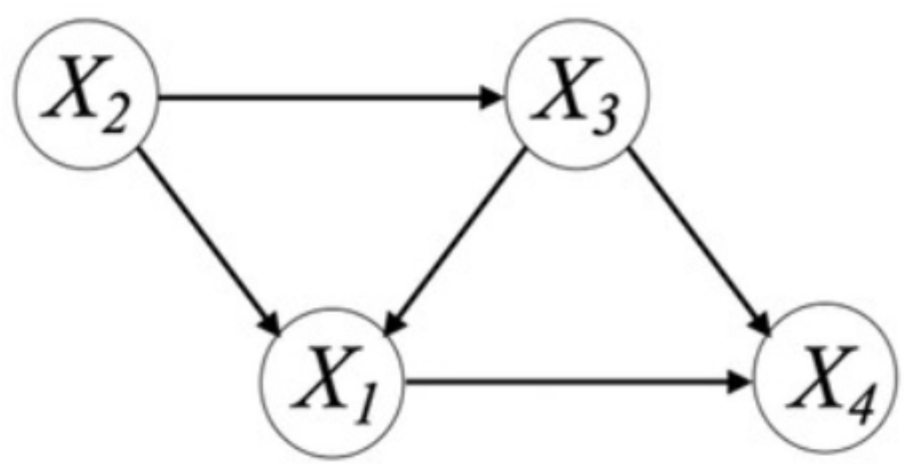
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

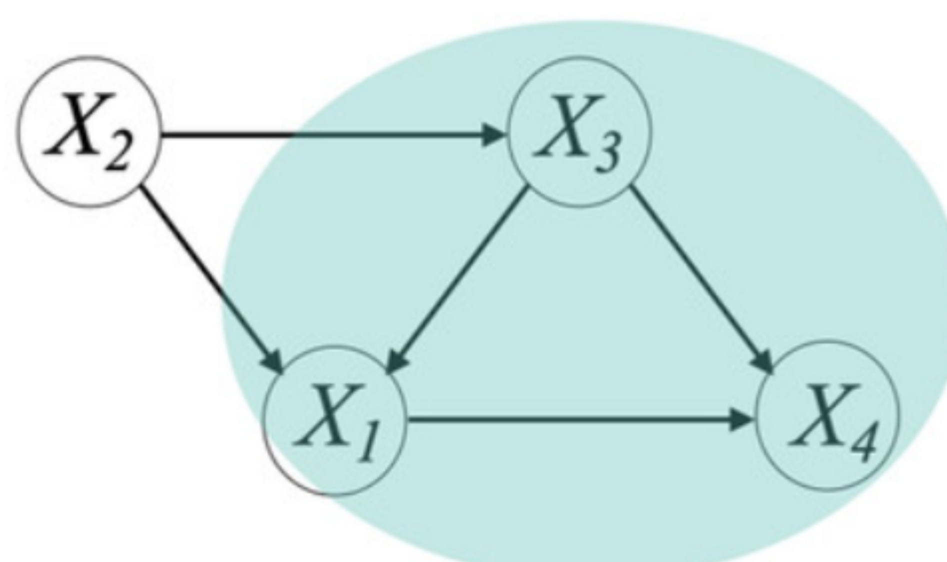
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

$$\mathbf{Z} \longrightarrow Y$$

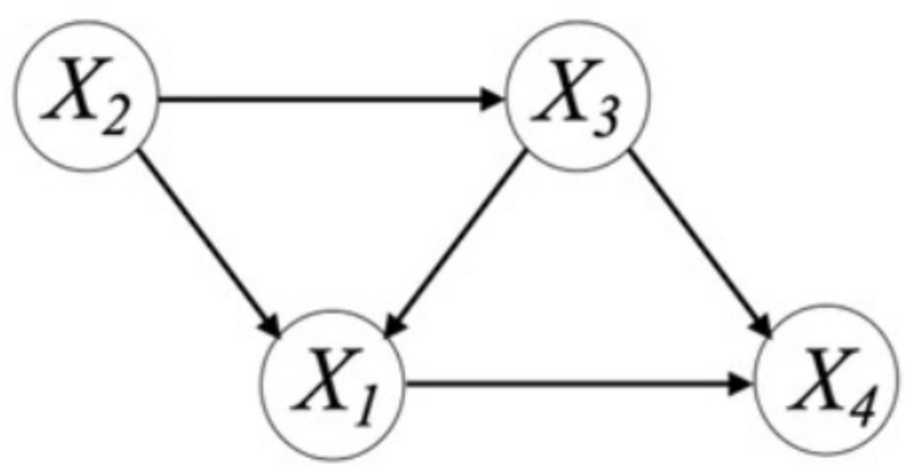
- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Independent Noise (IN) Condition

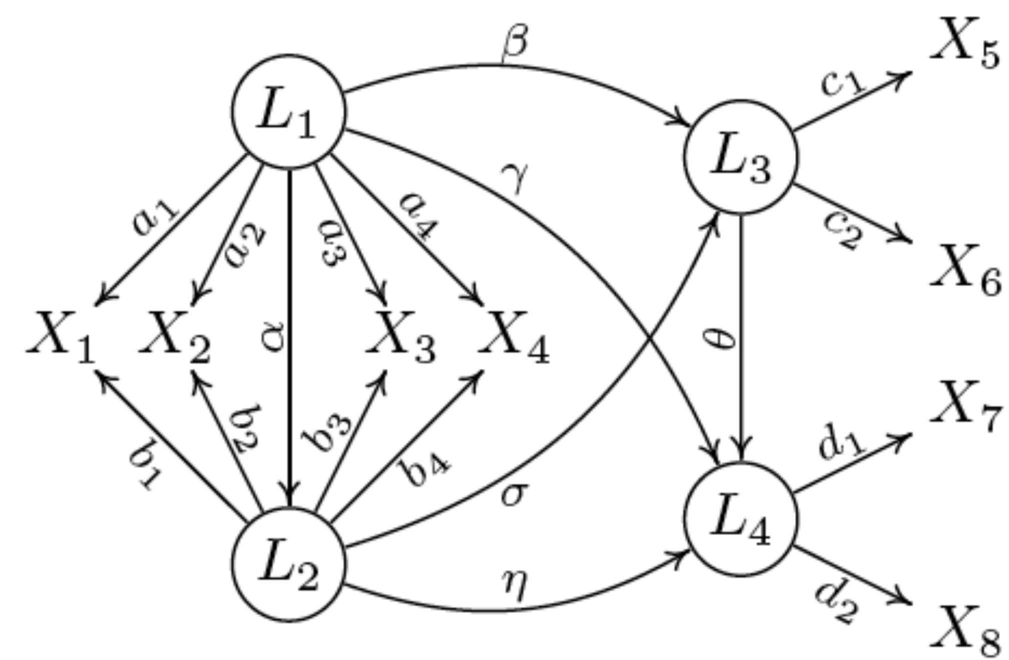
$$\mathbf{Z} \longrightarrow Y$$

- $(\mathbf{Z}, Y)$ follows the <u>IN condition</u> iff regression residual $Y - \tilde{w}^\intercal \mathbf{Z}$ is independent from $\mathbf{Z}$
- Can be used to estimate the Linear, Non-Gaussian Acyclic Causal model (LiNGAM), as in the DirectLiNGAM algorithm



Linear, non-Gaussian Latent Variable (LiNGLaM) Causal Model

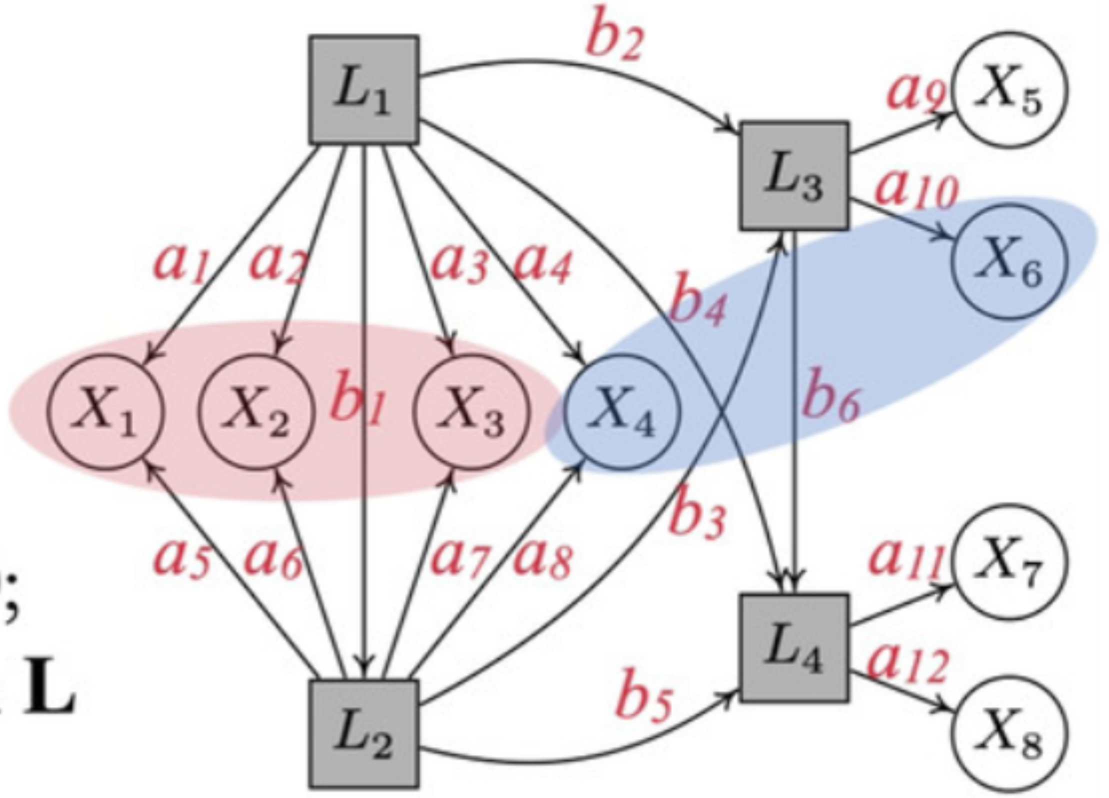
- [Measurement Assumption] There is no observed variable in X being an ancestor of any latent variables in L.
- The noise terms are non-Gaussian.
- [Double-Pure Child Variable Assumption] Each latent variable set L′, in which every latent variable directly causes the same set of observed variables, has at least 2Dim(L′) pure measurement variables as children.
- [Purity Assumption] There is no direct edge between observed variables.



Using Measured Variables as Surrogate

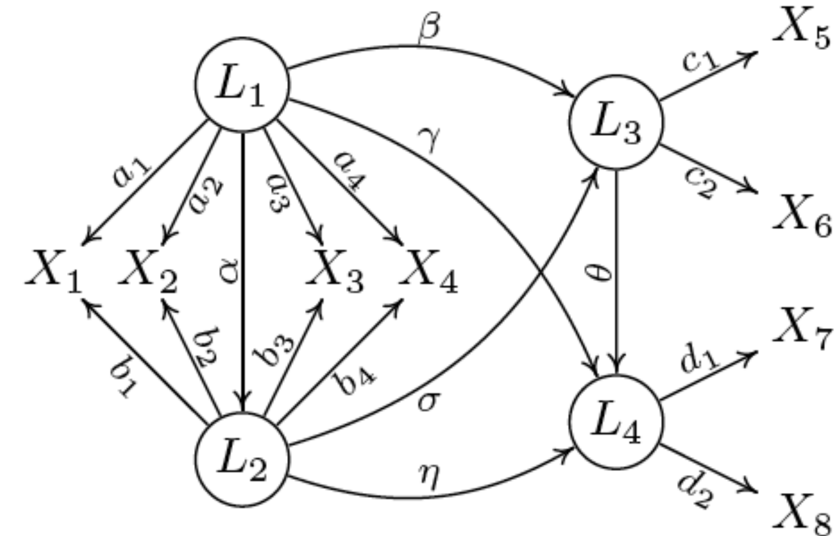
- IN does not hold on measured variables X_i
- If we have access to latent variables:

$$\underbrace{\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}}_{\mathbf{Y}} = \underbrace{\begin{bmatrix} a_1 & a_5 \\ a_2 & a_6 \\ a_3 & a_7 \end{bmatrix}}_{\mathbf{A}} \cdot \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} + \underbrace{\begin{bmatrix} E_{X_1} \\ E_{X_2} \\ E_{X_3} \end{bmatrix}}_{\mathbf{E}}$$



- ∃ nonzero vector $\mathbf{w}$ s.t. $\mathbf{w} \cdot \mathbf{Cov}(\mathbf{Y}, L_{1,2}) = \mathbf{0}$; then $\mathbf{wA} = \mathbf{0}$, so $\mathbf{w}^\intercal \mathbf{Y} = \mathbf{w}^\intercal \mathbf{E}$ is ind. from $\mathbf{L}$
- But we don't have access to L_1 and L_2
- Use $\mathbf{Z} = (X_4, X_6)^\intercal$ instead: ∃ nonzero vector $\mathbf{w}$ s.t. $\mathbf{w} \cdot \mathbf{Cov}(\mathbf{Y}, \mathbf{Z}) = \mathbf{0} \Rightarrow \mathbf{w}^\intercal \mathbf{Y}$ is ind. from $\mathbf{Z}$

Using Measured Variables as Surrogate



$Y = \{X_1, X_2, X_3\}, Z = \{X_4, X_5\}$

$$L_1 = \varepsilon_{L_1}, L_2 = \alpha L_1 + \varepsilon_{L_2} = \alpha\varepsilon_{L_1} + \varepsilon_{L_2},$$

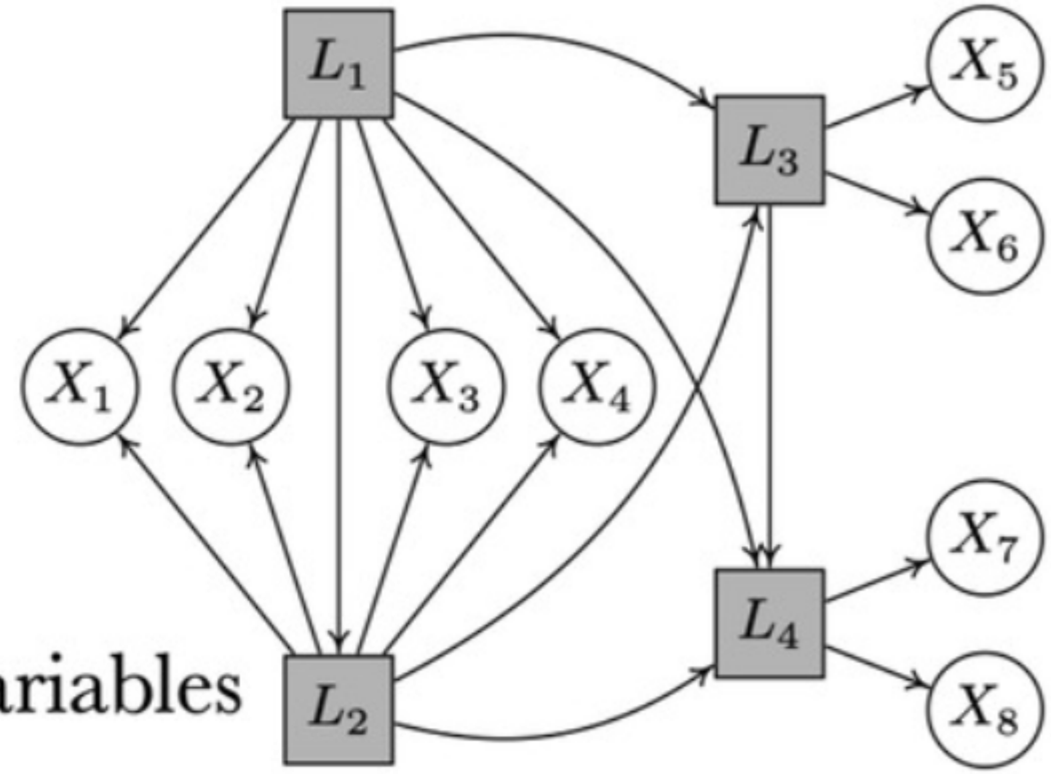
$$L_3 = \beta L_1 + \sigma L_2 + \varepsilon_{L_3} = (\beta + \alpha\sigma)\varepsilon_{L_1} + \sigma\varepsilon_{L_2} + \varepsilon_{L_3},$$

$$\underbrace{\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}}_{\mathbf{Y}} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} + \underbrace{\begin{bmatrix} \varepsilon_{X_1} \\ \varepsilon_{X_2} \\ \varepsilon_{X_3} \end{bmatrix}}_{\mathbf{E_Y}}, \qquad \underbrace{\begin{bmatrix} X_4 \\ X_5 \end{bmatrix}}_{\mathbf{Z}} = \begin{bmatrix} a_4 & b_4 \\ \beta c_1 & \sigma c_1 \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \end{bmatrix} + \underbrace{\begin{bmatrix} \varepsilon_{X_4} \\ \varepsilon_{X_5'} \end{bmatrix}}_{\mathbf{E_Z}},$$

With $\omega = [a_2 b_3 - b_2 a_3, b_1 a_3 - a_1 b_3, a_1 b_2 - b_1 a_2]^\intercal$

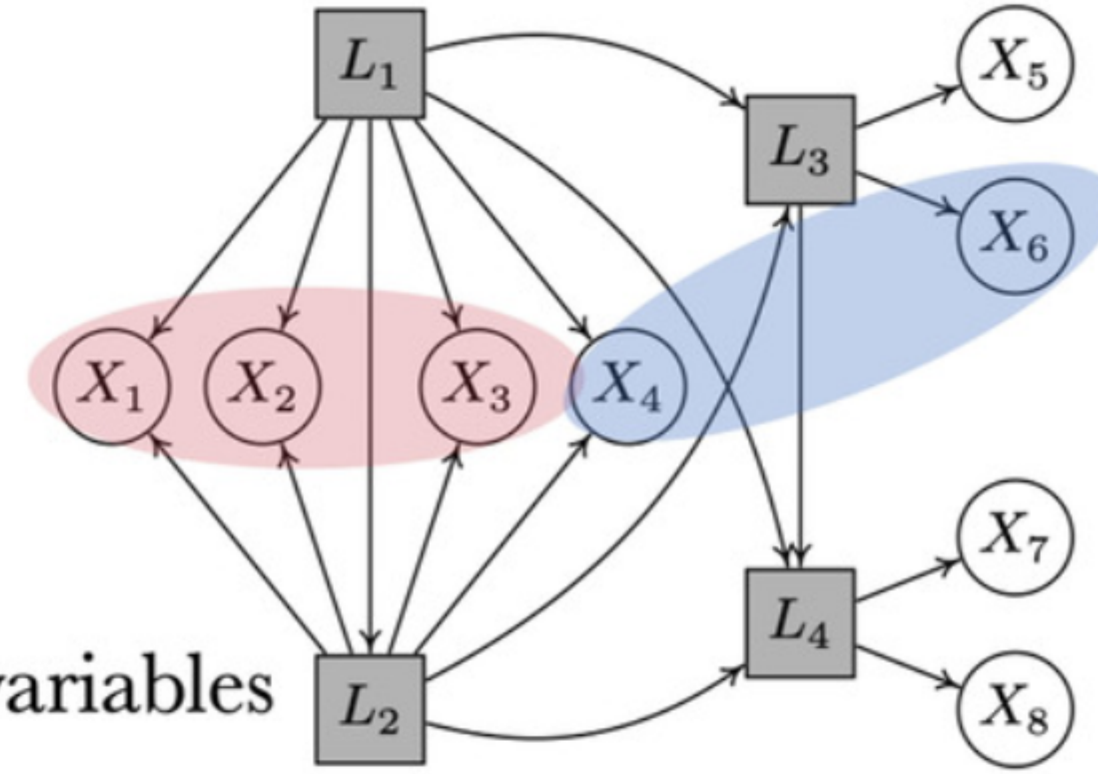
them $\omega^T Y = \omega^T$EY and it will be independent from Z.

GIN Condition



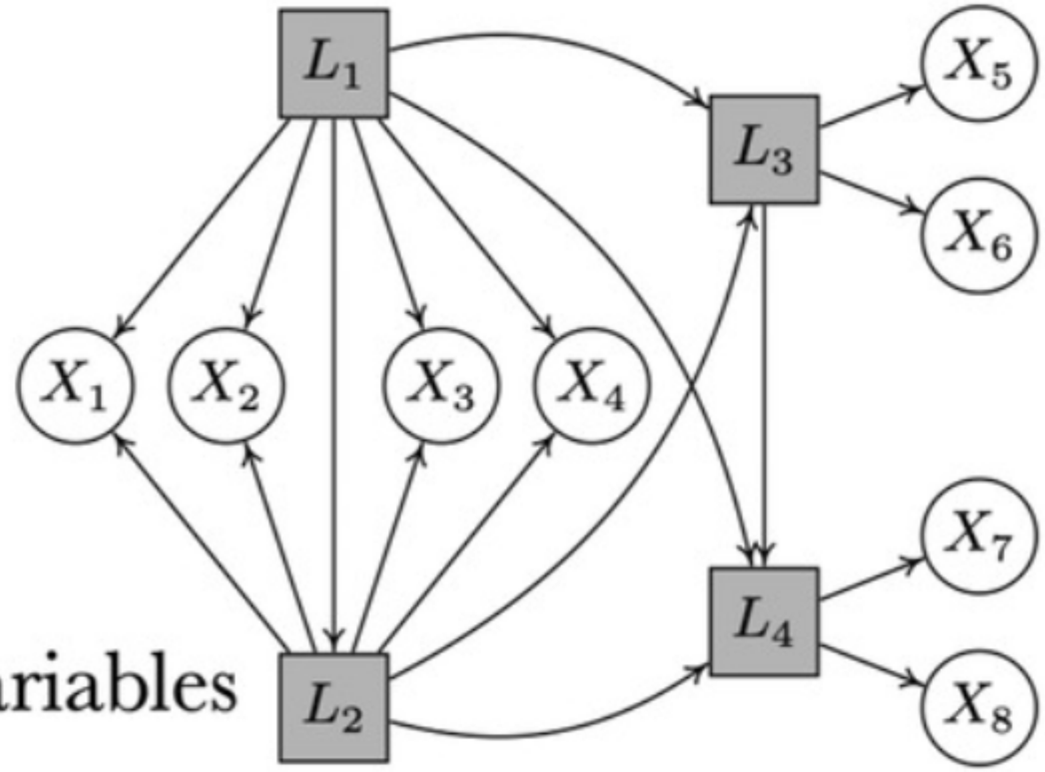
- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal\mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
  - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
  - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

GIN Condition



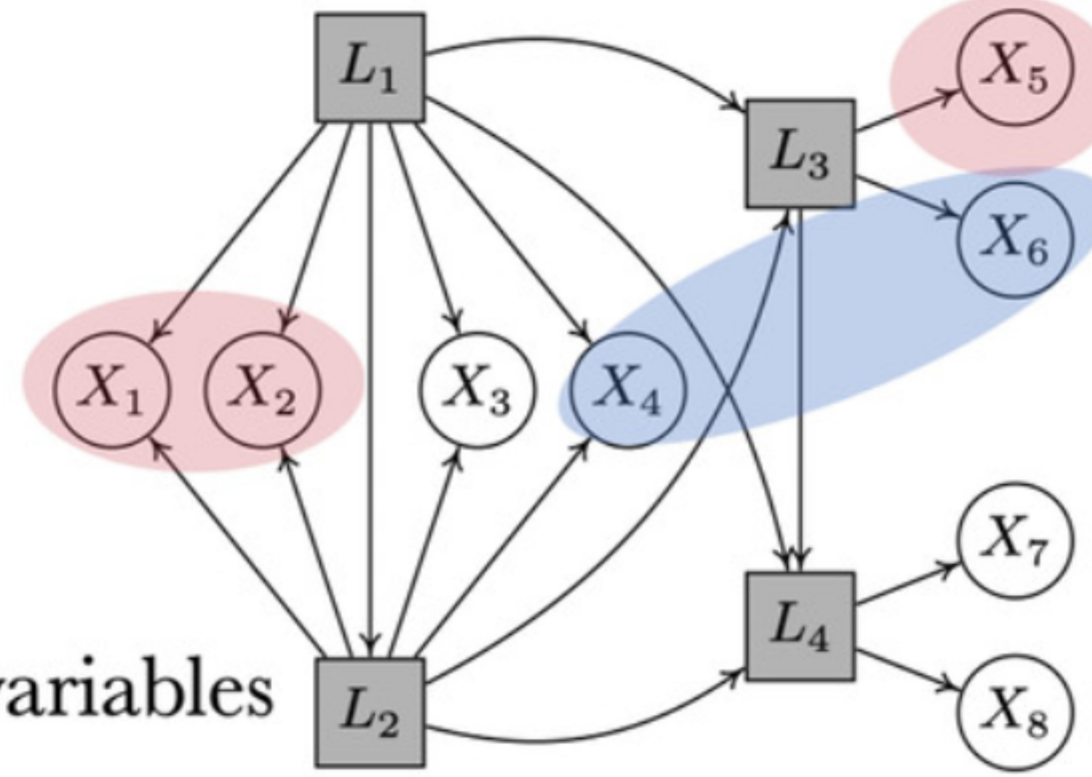
- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal\mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
  - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
  - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

GIN Condition



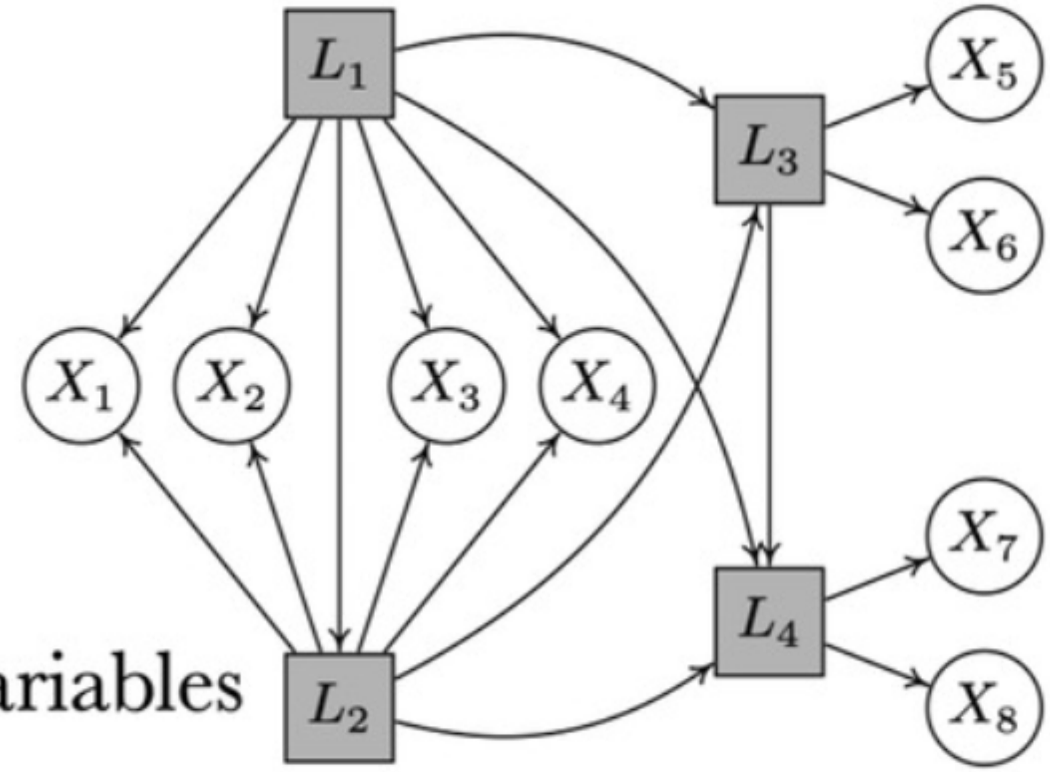
- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal\mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
  - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
  - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

GIN Condition



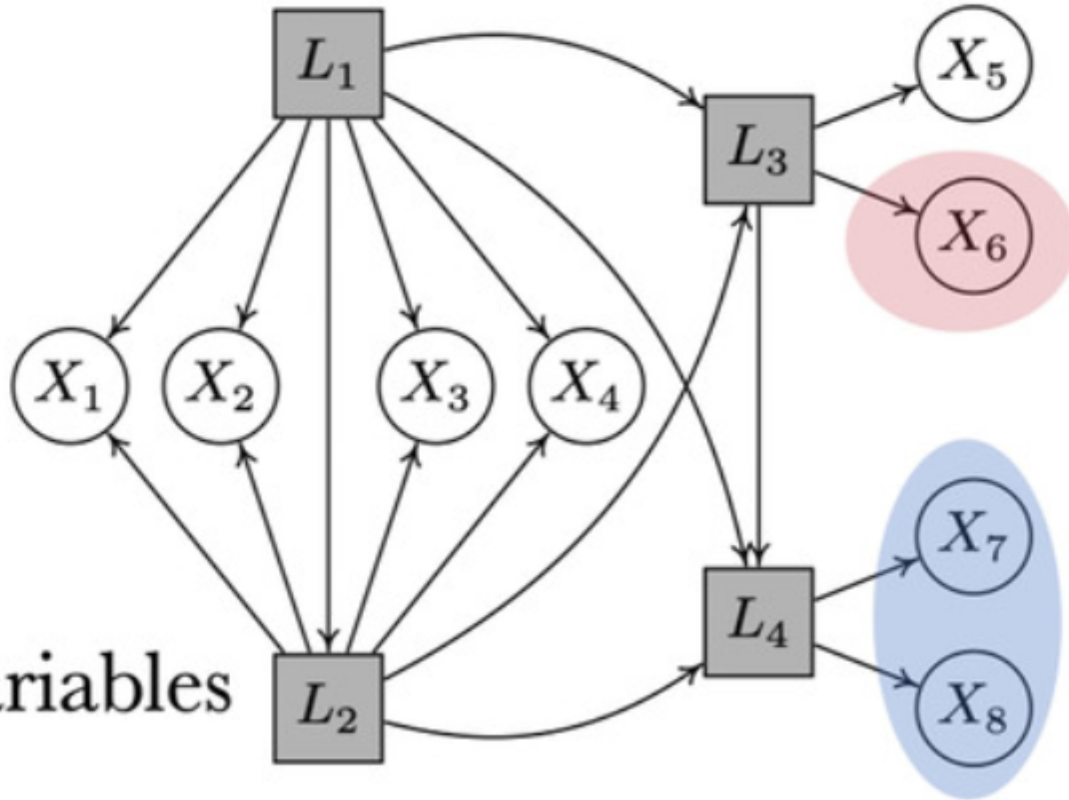
- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal\mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
    - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
    - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

GIN Condition



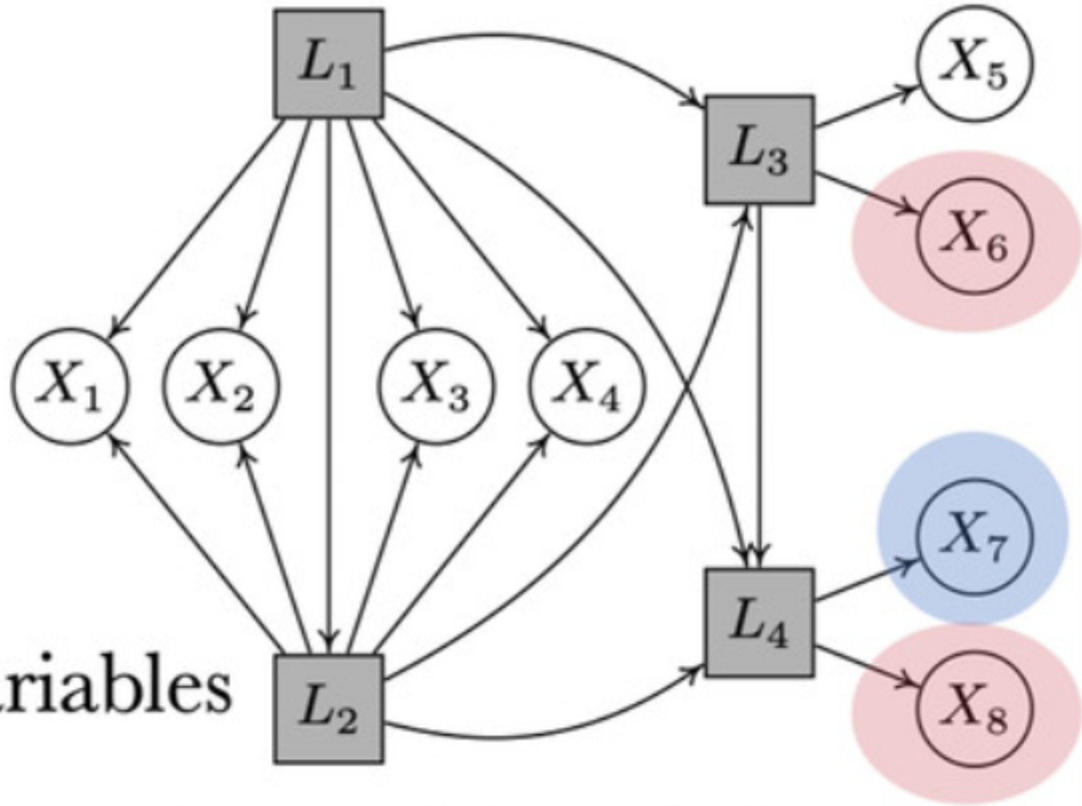
- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal\mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
    - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
    - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

GIN Condition



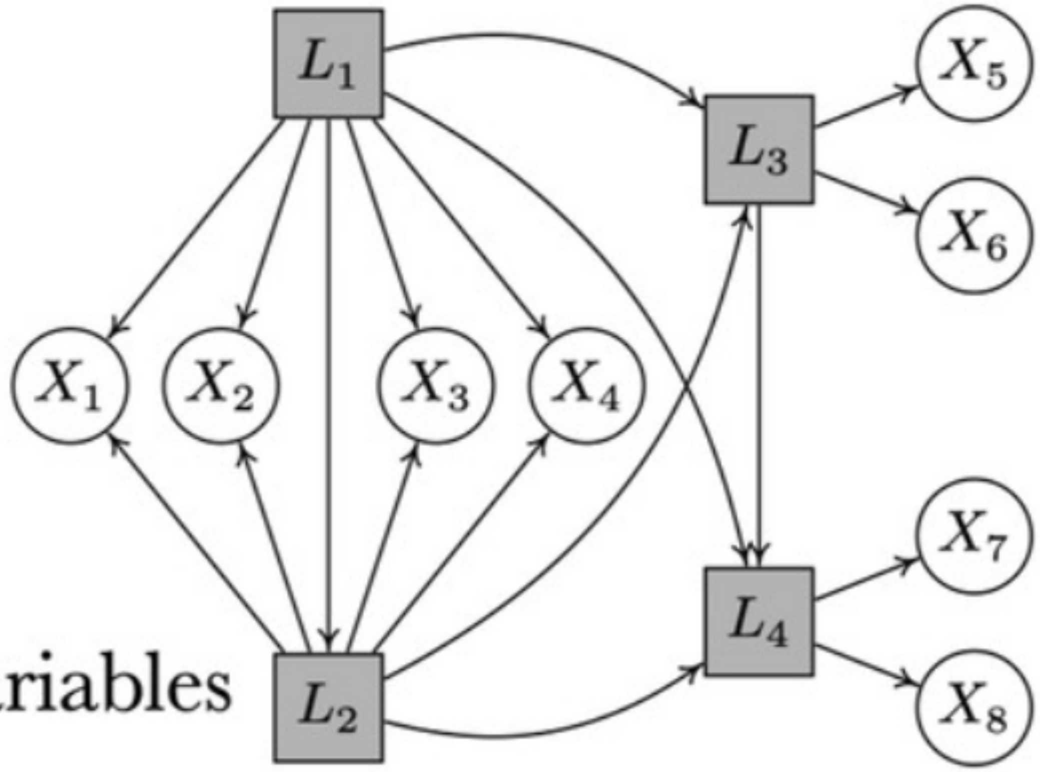
- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal\mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
  - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
  - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

GIN Condition



- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal \mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
  - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
  - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

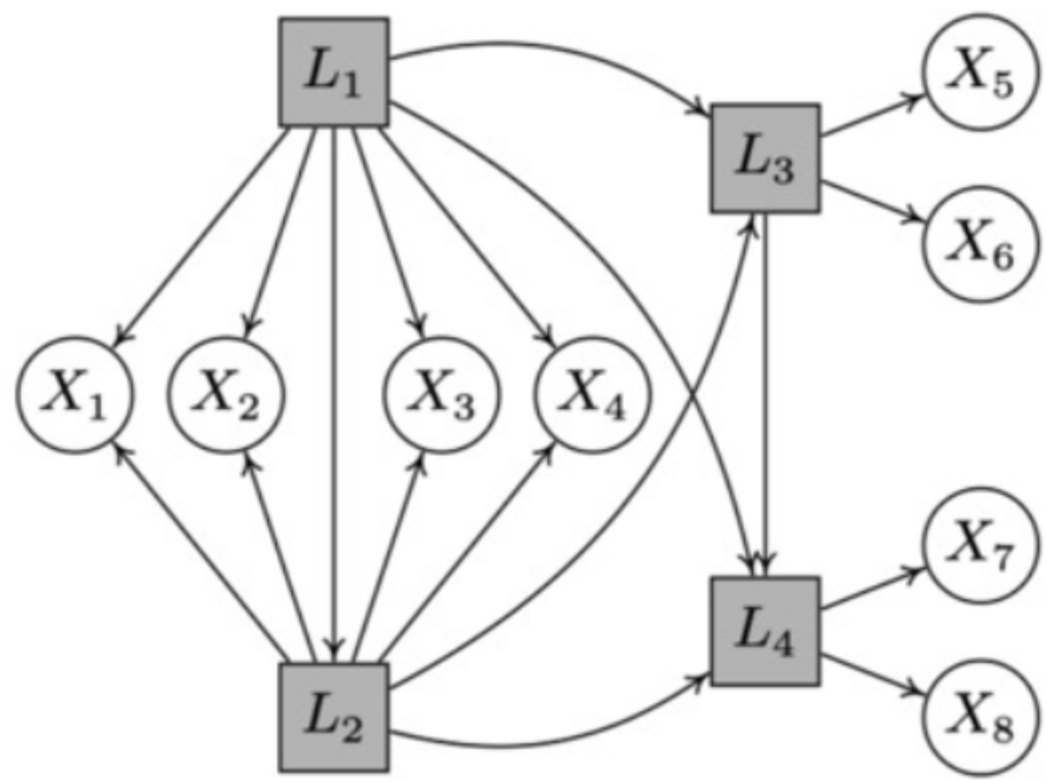
GIN Condition



- **Z** and **Y**: disjoint subsets of observed variables
- (**Z**, **Y**) satisfies the GIN condition iff there exists $\mathbf{w} \neq \mathbf{0}$ such that $\mathbf{w}^\intercal \mathbf{Y}$ is independent from **Z**
- In LiNGLaM, (**Z**, **Y**) satisfies GIN iff **Y** and **Z** are d-separated by the latent common causes of **Y** *(roughly speaking; see paper for precise conditions)*
  - $(\{X_4, X_6\}, \{X_1, X_2, X_3\})$; $(\{X_6\}, \{X_7, X_8\})$… satisfy GIN 😀
  - $(\{X_4, X_6\}, \{X_1, X_2, X_5\})$; $(\{X_7\}, \{X_6, X_8\})$… violate GIN 😞
- Determine where the latent variables are and their causal order by testing GIN conditions

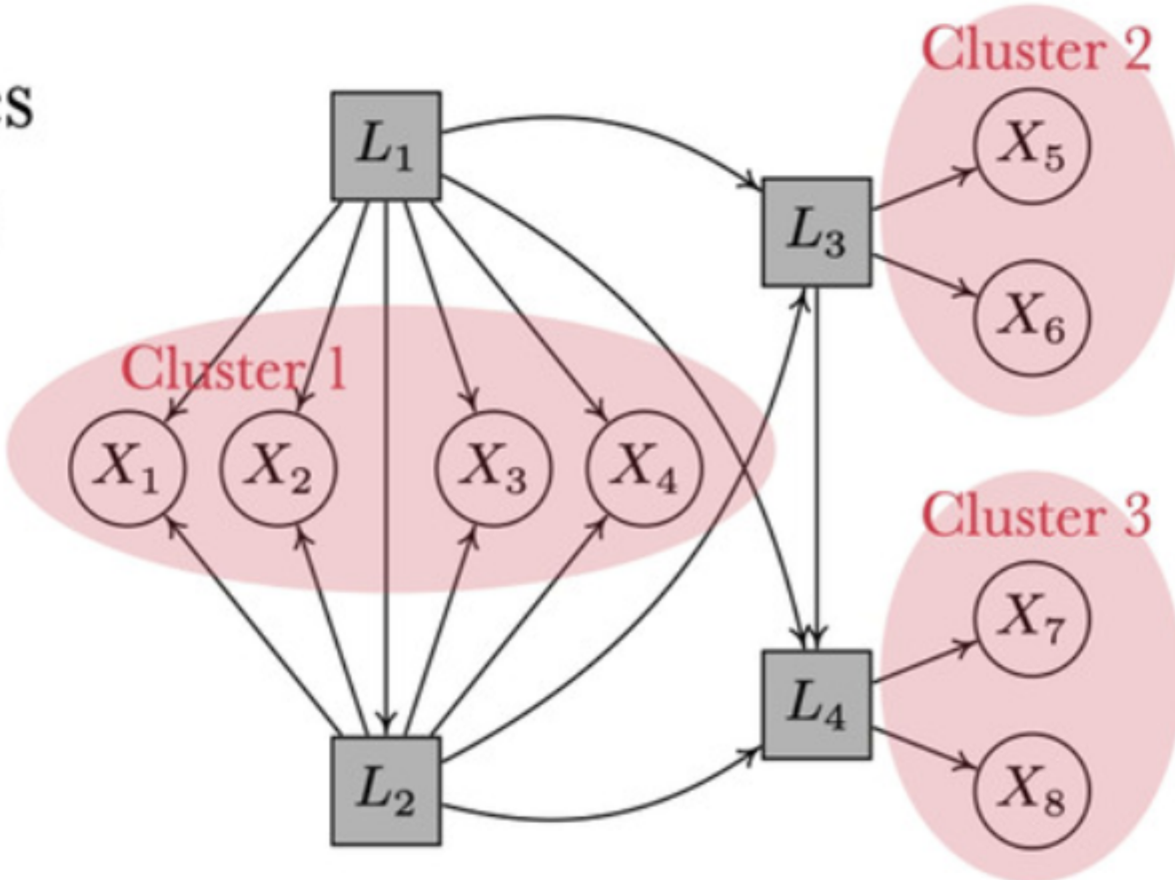
Estimating LiNGLaM Based on GIN

- Step 1: find causal clusters (variables sharing the same latent variables as parents)
- Step 2: determine causal order of the latent variables
- Estimate the coefficients if needed



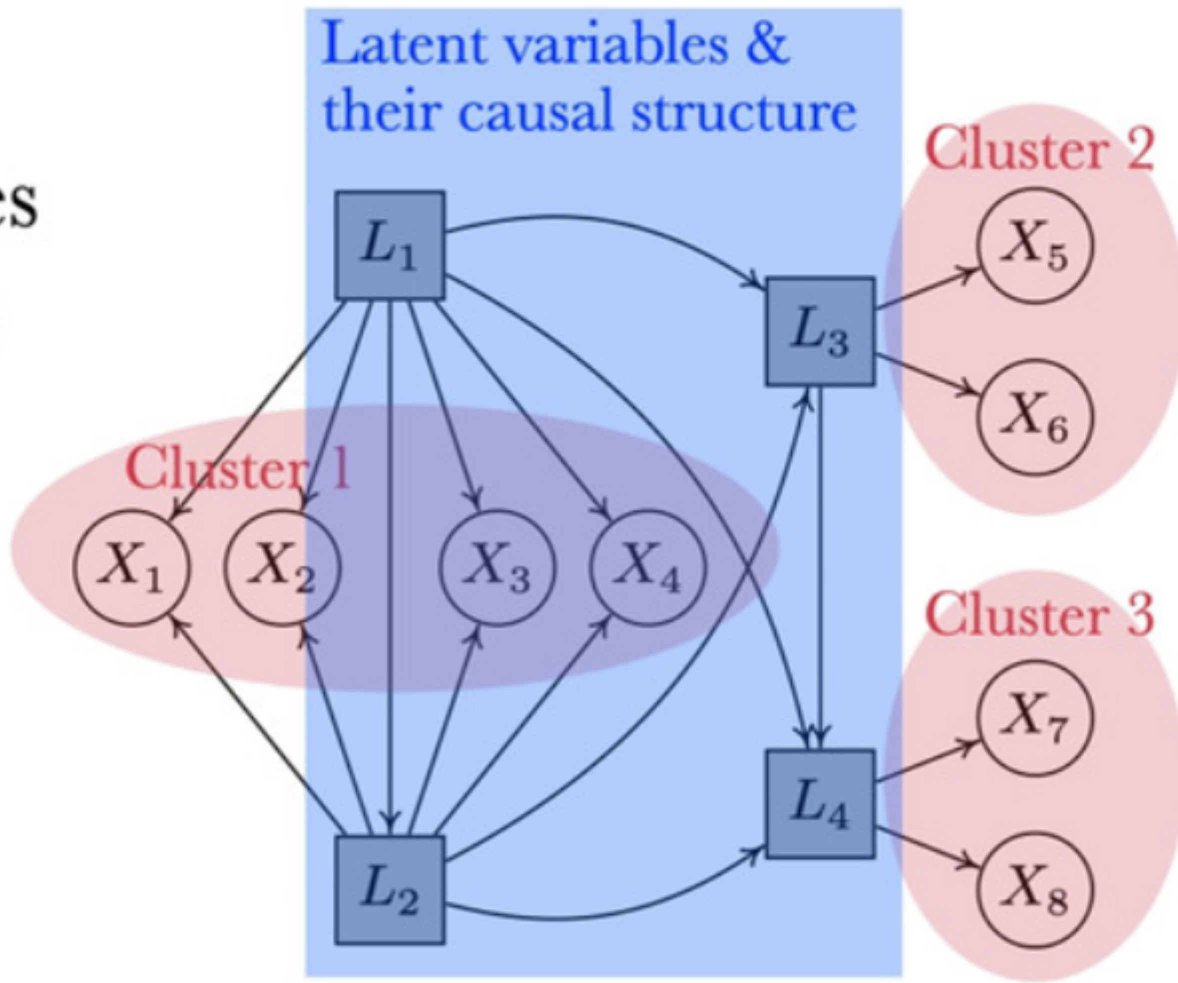
Estimating LiNGLaM Based on GIN

- Step 1: find causal clusters (variables sharing the same latent variables as parents)
- Step 2: determine causal order of the latent variables
- Estimate the coefficients if needed



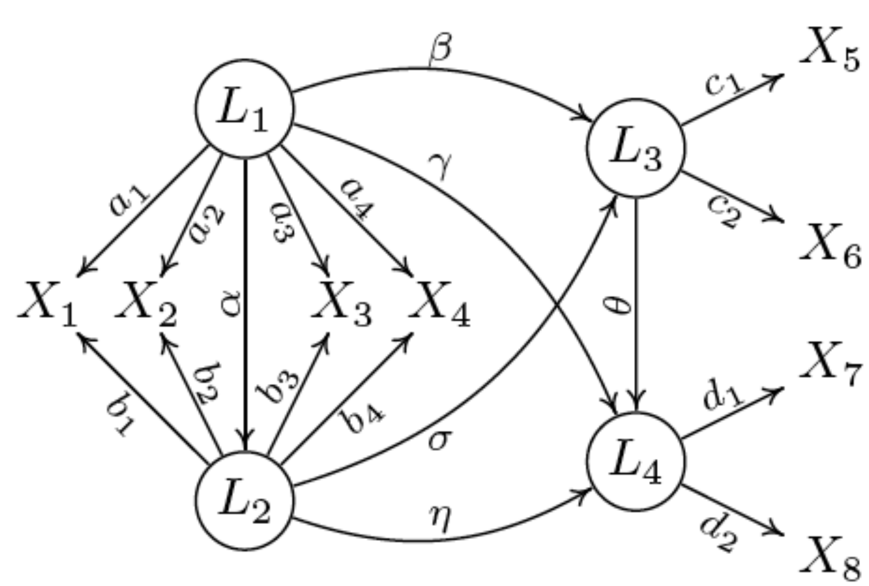
Estimating LiNGLaM Based on GIN

- Step 1: find causal clusters (variables sharing the same latent variables as parents)
- Step 2: determine causal order of the latent variables
- Estimate the coefficients if needed



Finding Causal Clusters

- Theorem #3 : Let X be the set of all observed variables in a LiNGLaM and Y be a proper subset of X. If (X \ Y,Y) follows the GIN condition and there is no subset Y' ⊆ Y such that (X \ Y' , Y') follows the GIN condition, then Y is a causal cluster and Dim(L(Y)) = Dim(Y) − 1.
- Consider the example in Figure 1, for $\{X_5, X_6\}$, one can find ($\{X_5, X_6\}$, $\{X_1, \ldots X_4, X_7, X_8\}$) follows the GIN condition, so $\{X_5, X_6\}$ is a causal cluster and Dim(L($\{X_5, X_6\}$)) = Dim($\{X_5, X_6\}$) − 1 = 1(i.e., L_3). But, for $\{X_1, X_2, X_5\}$, ($\{X_3, X_4, X_6, X_7, X_8\}$, $\{X_1, X_2, X_5\}$) violates the GIN condition, thus $\{X_1, X_2, X_5\}$ is not a causal cluster.



Finding Causal Clusters

- Start with finding clusters with a single latent variable and merge the overlapping clusters, and then increase the number of allowed latent variables until all variables are put in the clusters.

Algorithm 1 Identifying Causal Clusters

Input: Data set $\mathbf{X} = \{X_1, ..., X_m\}$
Output: Causal cluster set $\mathcal{S}$

1: Initialize $\mathcal{S} = \varnothing$, $Len = 1$, and $\mathbf{P} = \mathbf{X}$;
2: **repeat**
3:     **repeat**
4:         Select a variable subset $\mathcal{P}$ from $\mathbf{P}$ such that $\mathrm{Dim}(\mathcal{P}) = Len$;
5:         **if** $E_{\mathcal{P}||(\mathbf{P}\setminus\mathcal{P})} \perp\!\!\!\perp (\mathbf{P} \setminus \mathcal{P})$ holds **then**
6:             $\mathcal{S} = \mathcal{S} \cup \mathcal{P}$;
7:         **end if**
8:     **until** all subsets with length Len in $\mathbf{P}$ have been selected;
9:     Merge all the overlapping sets in $\mathcal{S}$;
10:     $\mathbf{P} \leftarrow \mathbf{P} \setminus \mathcal{S}$, and $Len \leftarrow Len + 1$;
11: **until** $\mathbf{P}$ is empty or $\mathrm{Dim}(\mathbf{P}) \leq Len$;
12: **Return:** $\mathcal{S}$

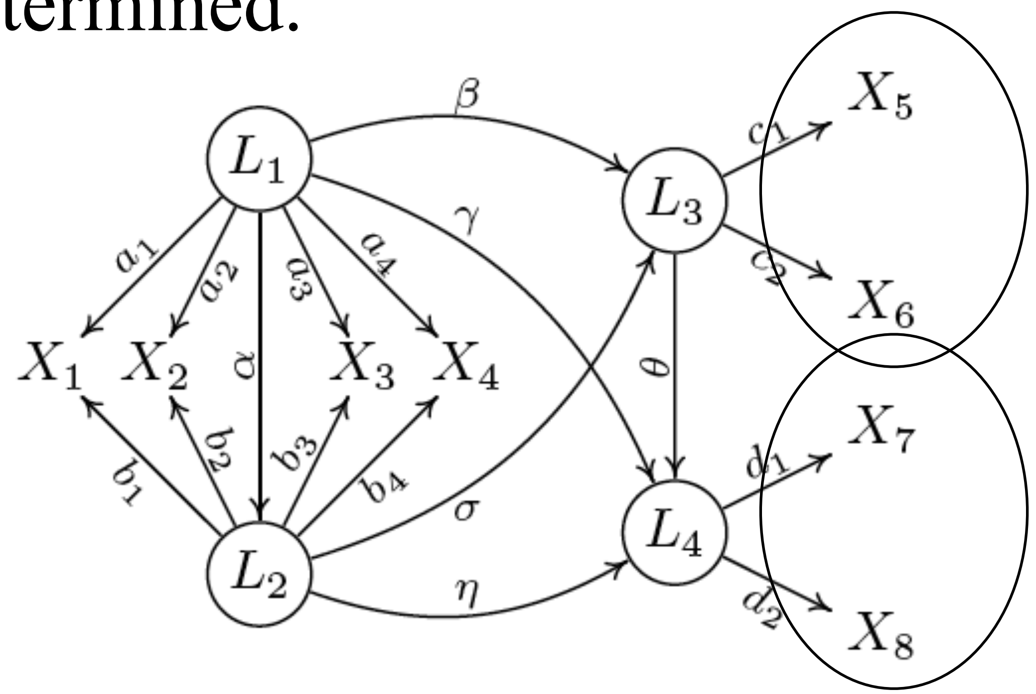
- Example : Consider the example in last Figure. First, we set Len = 1 to find the clusters with a single latent variable, i.e., we find $\{X_5, X_6\}$ and $\{X_7, X_8\}$ based on Theorem #3 (Line 4-9). Then we set Len = 2 and find the clusters $\{X_1, X_2, X_3, X_4\}$ with two latent variables.

Learning the Causal Order of Latent Variables

- Lemma #1: Let Sr be a cluster and S_k, k ≠ r, be any other cluster of a LiNGLaM. Suppose that S_r contains 2Dim(L(S_r)) number of variables with $S_r = \{R_1, R_2, \ldots, R_{2\times Dim(L(S_r))}\}$ and that S_k contains 2Dim(L(S_k)) number of variables with $S_k = \{K_1, K_2, \ldots, K_{2\times Dim(L(S_k))}\}$. if $(\{R_{Dim(L(S_r))+1}, \ldots R_{2\times Dim(L(S_r))}\}$, $\{R_1, \ldots, R_{Dim(L(S_r))}, K_1, \ldots K_{2\times Dim(L(S_k))}\}$) follows the GIN condition, then $L(S_r)$is a root latent variable set.

- Use this lemma to recursively discover the "root variable" until the causal order of latent variables is fully determined.

- Example: For L_3, we find that $(\{X_6, X_3, X_4\}, \{X_5, X_7, X_1, X_2\})$ satisfies the GIN condition, while for L_4, $(\{X_8, X_3, X_4\}, \{X_1, X_2, X_5, X_7\})$ violates the GIN condition, which means that L_3 is the "root variable".



Learning the Causal Order of Latent Variables

Algorithm 2 Learning the Causal Order of Latent Variables

Input: Set of causal clusters $\mathcal{S}$
Output: Causal order $\mathcal{K}$

1: Initialize $\mathcal{L}$ with the root variable sets of each cluster, $\mathbf{T} = \varnothing$, and $\mathcal{K} = \varnothing$;
2: **while** $\mathcal{L} \neq \varnothing$ **do**
3: Find the root node $L(\mathcal{S}_r)$ according to Proposition 4;
4: $\mathcal{L} = \mathcal{L} \setminus L(\mathcal{S}_r)$;
5: Include $L(\mathcal{S}_r)$ into the $\mathcal{K}$;
6: $\mathbf{T} = \mathbf{T} \cup \mathcal{S}_r$;
7: **end while**
8: **Return:** Causal order $\mathcal{K}$

Proposition 4. *Suppose that $\{\mathcal{S}_1, ...\mathcal{S}_i, ..., \mathcal{S}_n\}$ contains all clusters of the LiNGLaM. Denote $\mathbf{T} = \{L(\mathcal{S}_1), ...L(\mathcal{S}_i)\}$ and $\mathbf{R} = \{L(\mathcal{S}_{i+1}), ...L(\mathcal{S}_n)\}$, where all elements in $\mathbf{T}$ are causally earlier than those in $\mathbf{R}$. Let $\hat{\mathbf{Z}}$ contain the elements from the half set of the children of each latent variable set in $\mathbf{T}$, and $\hat{\mathbf{Y}}$ contain the elements from the other half set of the children of each latent variable set in $\mathbf{R}$. Furthermore, Let $L(\mathcal{S}_r)$ be a latent variable set of $\mathbf{R}$ and $\mathcal{S}_r = \{R_1, R_2, ..., R_{2Dim(L(\mathcal{S}_r))}\}$. If for any one of the remaining elements $L(\mathcal{S}_k) \in \mathbf{R}$, with $k \neq r$ and $\mathcal{S}_k = \{K_1, K_2, ..., K_{2Dim(L(\mathcal{S}_k))}\}$ such that $(\{R_{Dim(L(\mathcal{S}_r))+1}, ...R_{2Dim(L(\mathcal{S}_r))}, \hat{\mathbf{Z}}\}, \{R_1, ..., R_{Dim(L(\mathcal{S}_r))}, K_1, ...K_{Dim(L(\mathcal{S}_k))}, \hat{\mathbf{Y}}\})$ follows the GIN condition, then $L(\mathcal{S}_r)$ is a root latent variable set in $\mathbf{R}$.*

Simulation

- 4 cases, with different DAG structures for latent variables
- Can we find clusters (determine the location of latent variables)?
  - Latent omission: measure omitted latent variables
  - Latent commission: measure falsely detected latent variables
  - Mismeasurements: measure the misclassification of observed variables

Table 1. Performance of cluster recovery by different methods (the lower, the better)

		Latent omission				Latent commission				Mismeasurements			
Algorithm		GIN	LSTC	FOFC	BPC	GIN	LSTC	FOFC	BPC	GIN	LSTC	FOFC	BPC
Case 1	500	0.00(0)	0.00(0)	1.00(10)	0.50(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)
	1000	0.00(0)	0.00(0)	1.00(10)	0.50(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)
	2000	0.00(0)	0.00(0)	1.00(10)	0.50(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)
Case 2	500	0.10(2)	0.20(4)	0.9(10)	0.50(10)	0.00(0)	0.05(1)	0.00(0)	0.00(0)	0.12(2)	0.12(4)	0.00(0)	0.20(10)
	1000	0.05(1)	0.15(3)	1.00(10)	0.50(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.04(1)	0.12(3)	0.00(0)	0.20(10)
	2000	0.00(0)	0.00(0)	1.00(10)	0.50(10)	0.00(0)	0.02(2)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.20(10)
Case 3	500	0.20(3)	0.20(3)	0.13(9)	0.10(1)	0.00(0)	0.03(3)	0.00(0)	0.00(0)	0.19(3)	0.17(3)	0.00(0)	0.00(0)
	1000	0.06(2)	0.13(2)	0.16(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.06(2)	0.00(0)	0.00(0)	0.00(0)
	2000	0.00(0)	0.00(0)	0.50(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.00(0)
Case 4	500	0.13(4)	0.40(6)	0.90(10)	0.63(10)	0.00(0)	0.23(5)	0.00(0)	0.00(0)	0.04(2)	0.15(6)	0.02(2)	0.06(4)
	1000	0.10(3)	0.26(6)	0.93(10)	0.66(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.05(3)	0.11(2)	0.01(1)	0.02(2)
	2000	0.03(1)	0.32(6)	1.00(10)	0.70(10)	0.00(0)	0.00(0)	0.00(0)	0.00(0)	0.04(1)	0.11(3)	0.00(10)	0.00(0)

Note: The number in parentheses indicates the number of occurrences that the current algorithm *cannot* correctly solve the problem.

Application to Teacher's Burnout Data

- Contains 28 measured variables
- Discovered clusters and causal order of the latent variables:

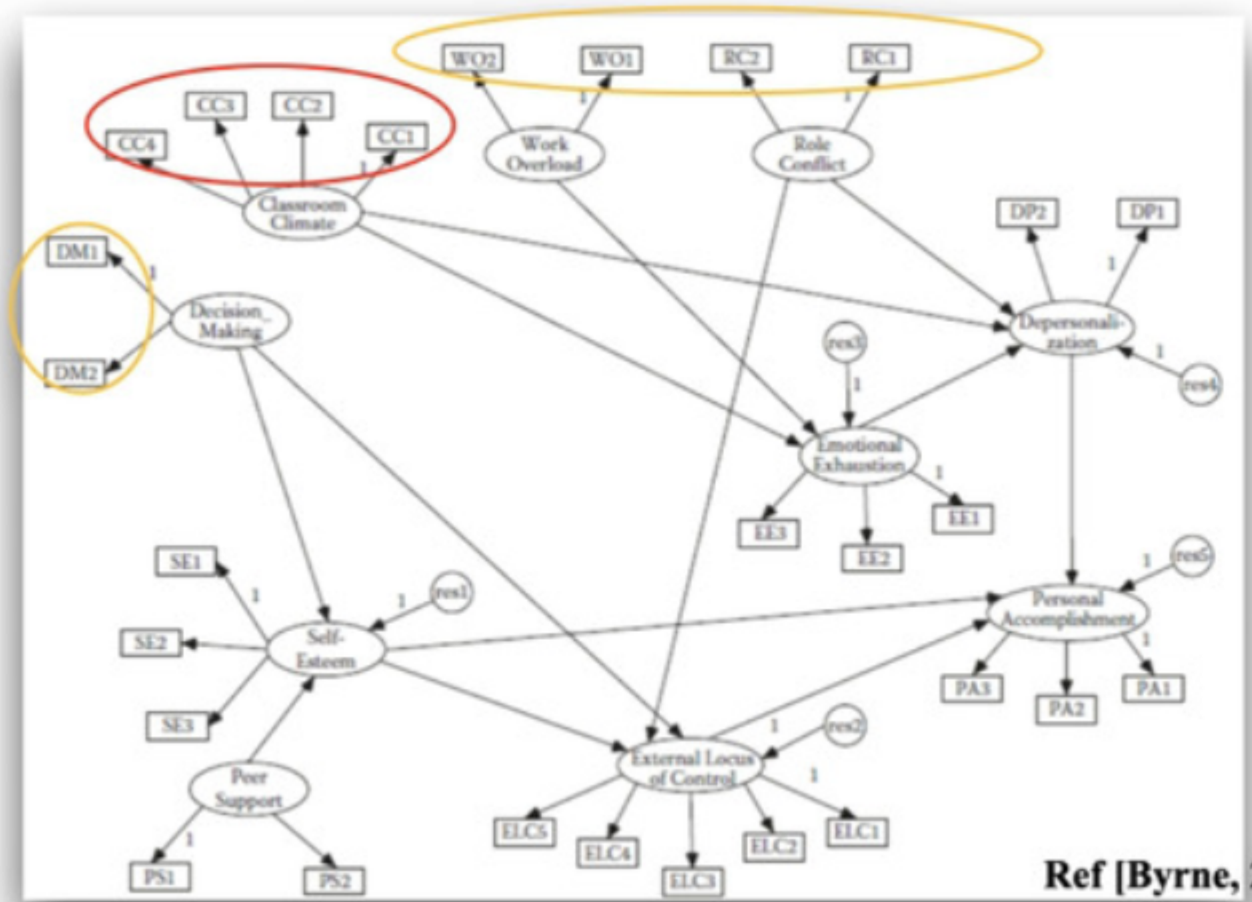
Causal Clusters	Observed variables
$\mathcal{S}_1$ (1)	$RC_1, RC_2, WO_1, WO_2, DM_1, DM_2$
$\mathcal{S}_2$ (1)	CC_1, CC_2, CC_3, CC_4
$\mathcal{S}_3$ (1)	PS_1, PS_2
$\mathcal{S}_4$ (1)	$ELC_1, ELC_2, ELC_3, ELC_4, ELC_5$
$\mathcal{S}_5$ (2)	$SE_1, SE_2, SE_3, EE_1, EE_2, EE_3, DP_1, PA_3$
$\mathcal{S}_6$ (3)	DP_2, PA_1, PA_2

$$L(\mathcal{S}_1) \succ L(\mathcal{S}_2) \succ L(\mathcal{S}_3) \succ L(\mathcal{S}_5) \succ L(\mathcal{S}_4) \succ L(\mathcal{S}_6).$$

(from root to leaf)

- Consistent with the hypothesized model

Hypothesized model by experts



Ref [Byrne, 2010]

Conclusion

- Essential to learn hidden causal representation
- The whole latent variable causal model is identifiable under suitable assumptions
- GIN condition is a powerful extension of IN condition
- Future work: more general mechanisms?