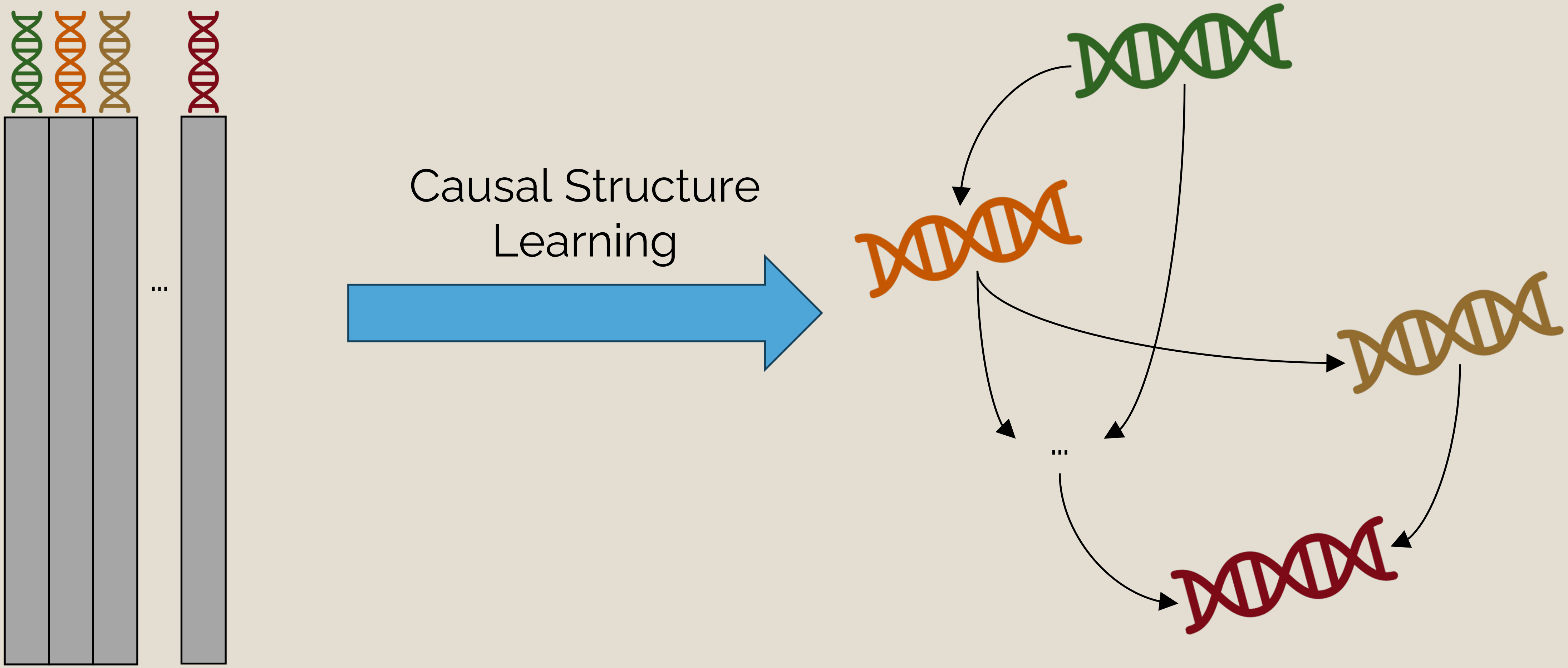
Causal Structure Discovery between Clusters of Nodes Induced by Latent Factors

Chandler Squires[*1] Annie Yun[*1] Eshaan Nichani[1] Raj Agrawal[1] Caroline Uhler[1]

[*]Equal Contribution

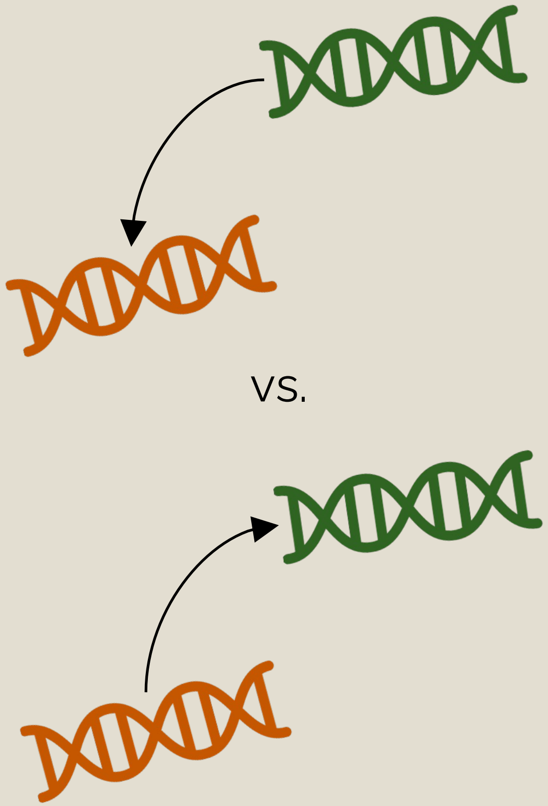
[1] MIT

Causal Structure Learning in Genomics

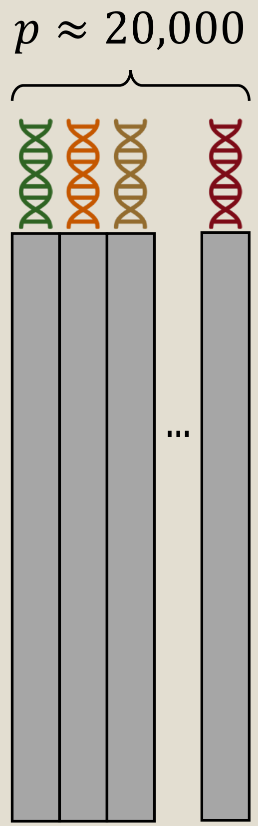


Challenges

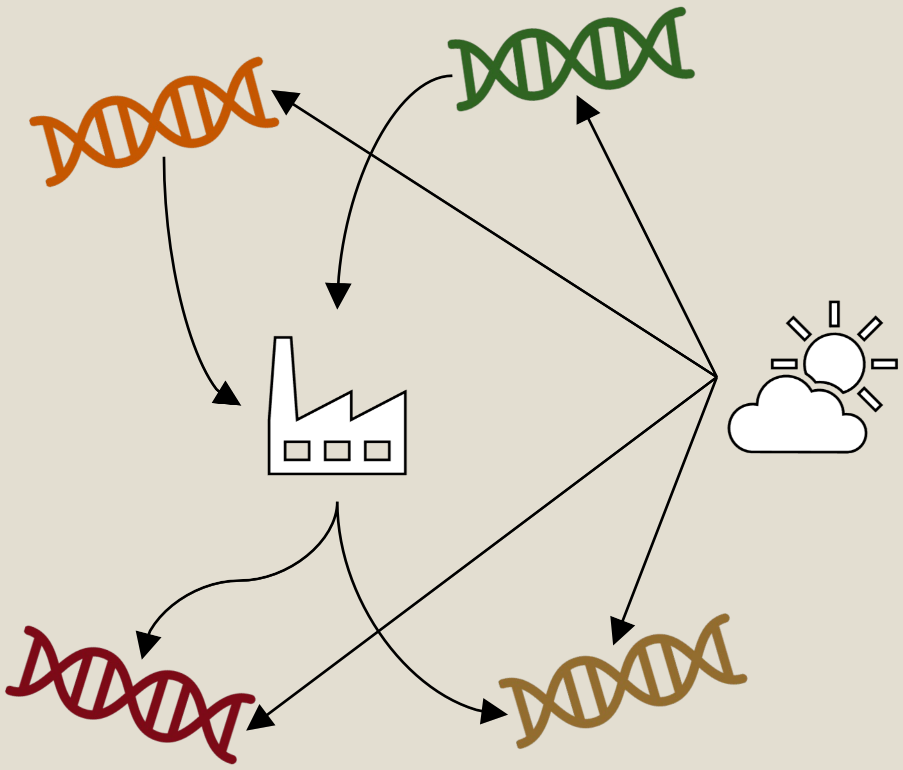
Identifiability



High dimensionality

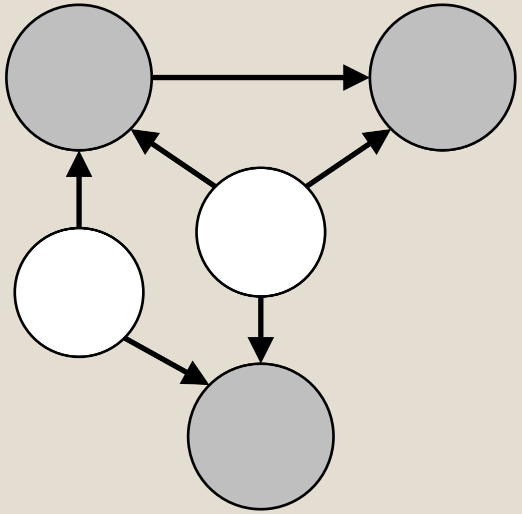


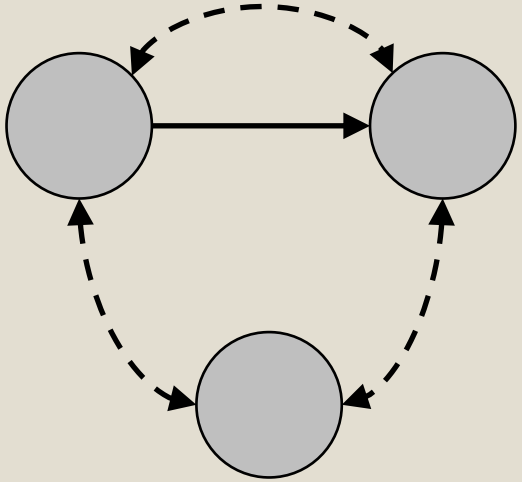
Latent variables



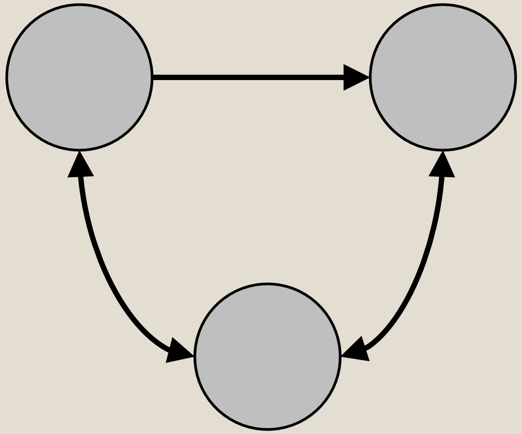
Latent Variables in Causal Discovery

Marginal Graphical Models

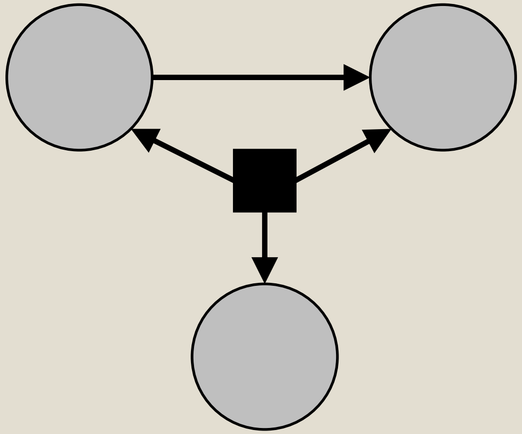




ADMGs



MAGs



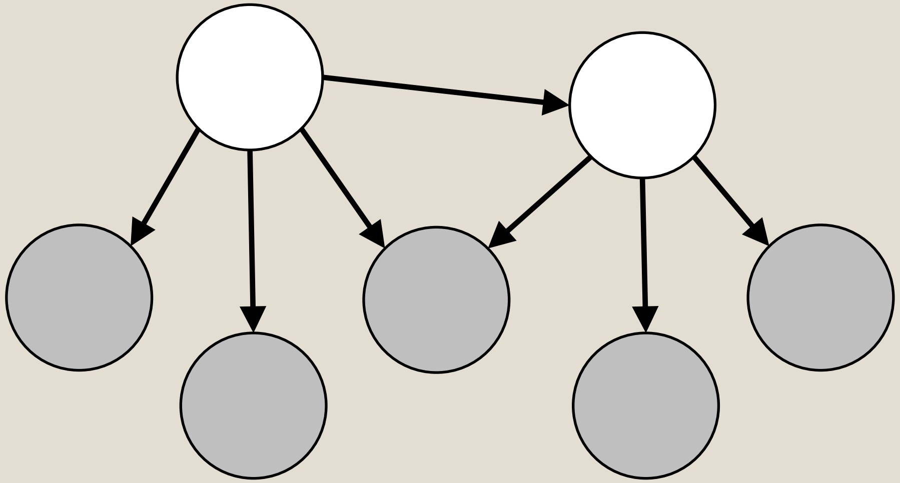
mDAGs

...

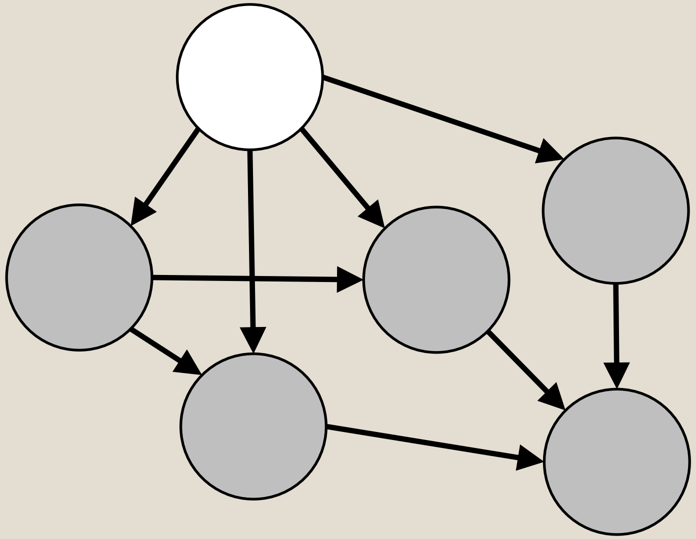
Not always **parsimonious,** and may be less scientifically useful.

Latent Recovery

Measurement models



Pervasive confounding

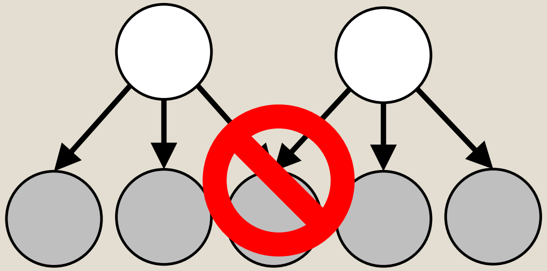


Both require latent variables to be **exogenous** (no observed variables as parents).

Goal: Recover non-exogenous latent variables and their causal relations to observed variables.

Latent Factor Causal Model (LFCM)

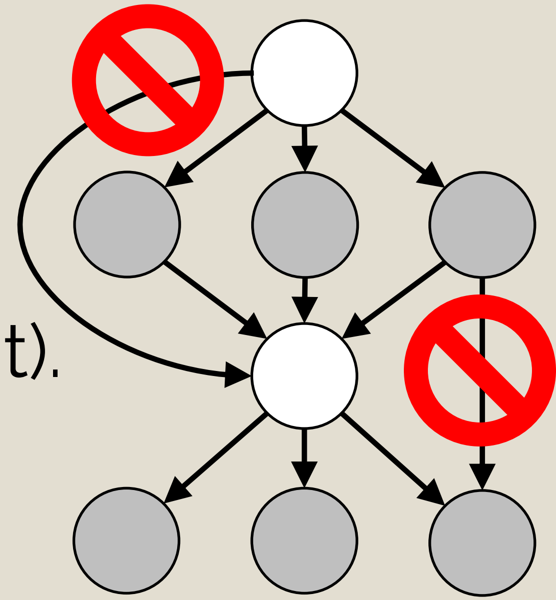
(1) Unique Cluster Assumption:
Each observed variable has 1 latent parent.



Latent Factor Causal Model (LFCM)

(1) Unique Cluster Assumption:
Each observed variable has 1 latent parent.

(2) **Bipartite Assumption:**
No (observed → observed) or (latent → latent).

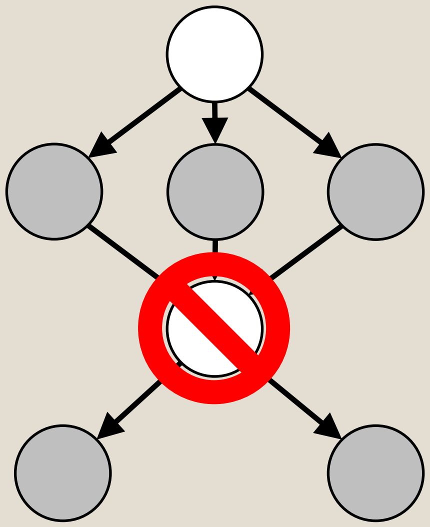


Latent Factor Causal Model (LFCM)

(1) Unique Cluster Assumption:
Each observed variable has 1 latent parent.

(2) **Bipartite Assumption:**
No (observed → observed) or (latent → latent).

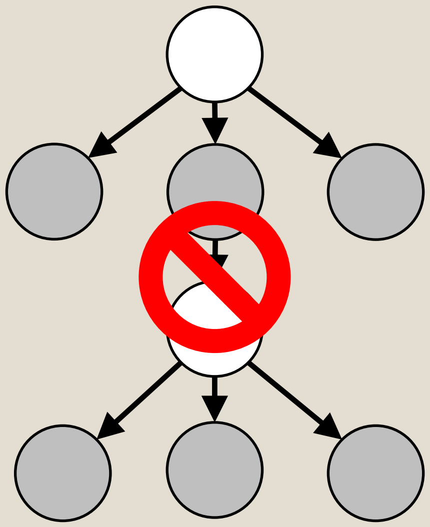
(3) **Triple child assumption:**
Each latent variable has ≥ 3 children.



Latent Factor Causal Model (LFCM)

(1) Unique Cluster Assumption:
Each observed variable has 1 latent parent.

(2) **Bipartite Assumption:**
No (observed → observed) or (latent → latent).



(3) **Triple child assumption:**
Each latent variable has ≥ 3 children.

(4) **Double parent assumption:**
Latent variable are “linked” by at least two observed variables.

Linear DAG model

We assume each variable is a linear function of its parents plus independent noise*.

$$X_j = \sum_{i \in pa(j)} \beta_{ij} X_i + \varepsilon_i$$

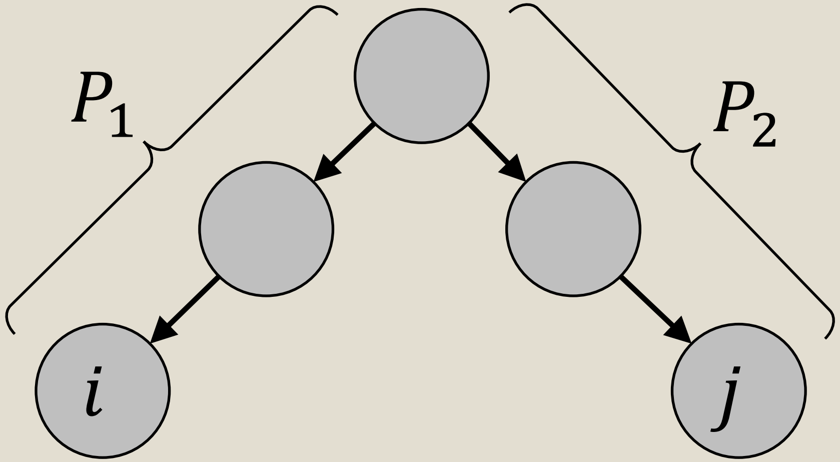
*No requirements beyond finite second moments, but hypothesis testing assumes Gaussianity.

Background: Trek separation in DAGs

Treks

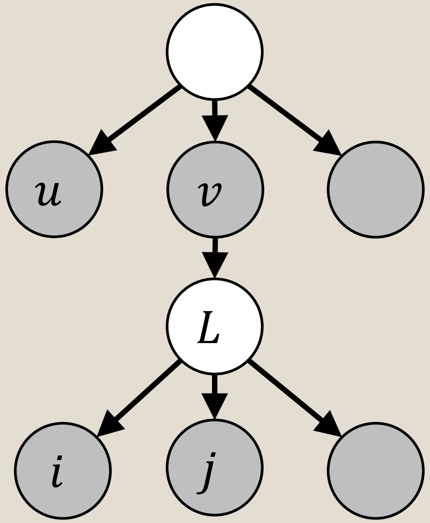
A **trek** from a node i to a node j is a tuple of directed paths (P_1, P_2) such that:

- i is the sink of P_1,
- j is the sink of P_2, and
- P_1 and P_2 have the same source



Trek separation

Two sets A and B are **t-separated** by the tuple of sets (C_A, C_B) if for every trek (P_1, P_2) between A and B, either P_1 contains a node in C_A, or P_2 contains a node in C_B.



$A = \{i, j\}$ and $B = \{u, v\}$ are t-separated by $(\{L\}, \emptyset)$.

Trek separation lemma

$$\mathrm{rank}(\Sigma_{AB}) \leq \min\{|C_A| + |C_B| : (C_A, C_B) \text{ t}-\text{separates } A \text{ and } B \text{ in } \mathcal{G}\}$$

Equality holds generically
for Σ consistent with $\mathcal{G}$.

Tetrads

For i, j, u, v distinct, define
$t_{ij,uv} = \det(\Sigma_{[ij],[uv]})$.

We call $t_{ij,uv}$ a **tetrad**.

If $A = \{i, j\}$ and $B = \{u, v\}$ are t-separated by a single node, we have $t_{ij,uv} = 0$.

Vanishing Tetrad Tests

Let $H_{vt}(A, B)$ denote the null hypothesis that all tetrads of the matrix $\Sigma_{A,B}$ vanish, i.e., $\Sigma_{\mathrm{A,B}}$ is rank one plus diagonal.

We use the **Wishart test** to compute p-values for each tetrad, dividing the sample tetrads $\hat{t}_{ij,uv}$ by their standard errors (using a formula for the sampling variance from [1]).

We aggregate the p-values using Sidak adjustment.

[1] Drton, Mathias, Hélène Massam, and Ingram Olkin. "Moments of minors of Wishart matrices." *The Annals of Statistics* (2008).

Method for Estimating Latent Factor Causal Models

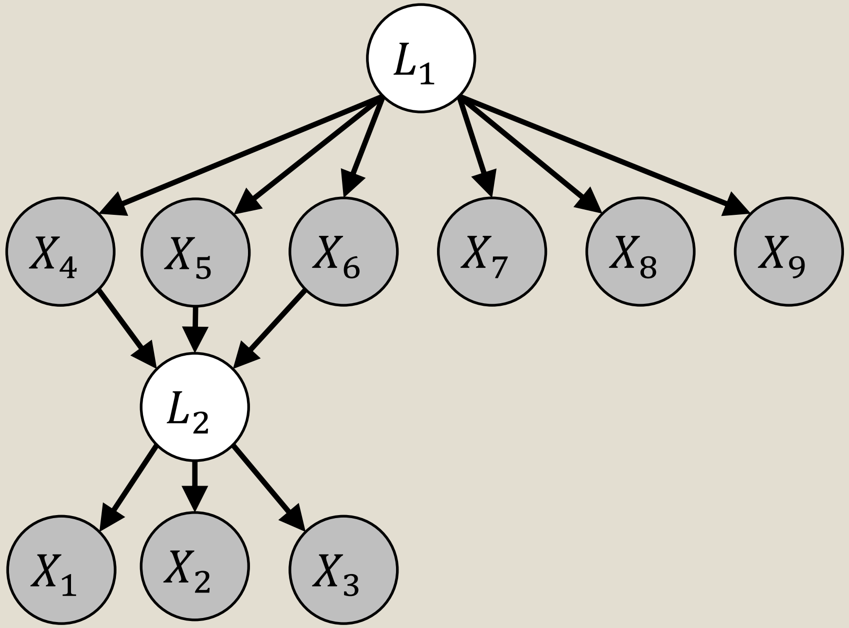
Method: EstimateLFCM

Phase 1:
Order Clusters

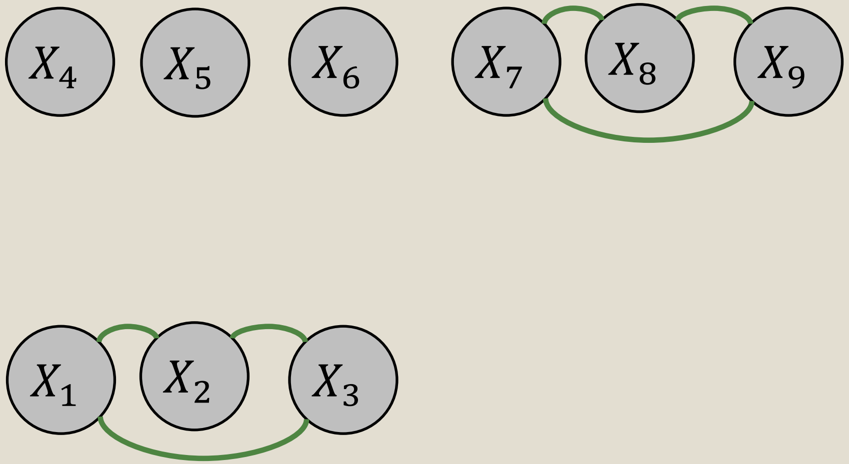
Phase 2:
Merge Clusters

Phase 3:
Learn DAG

Phase 1:
Order Clusters

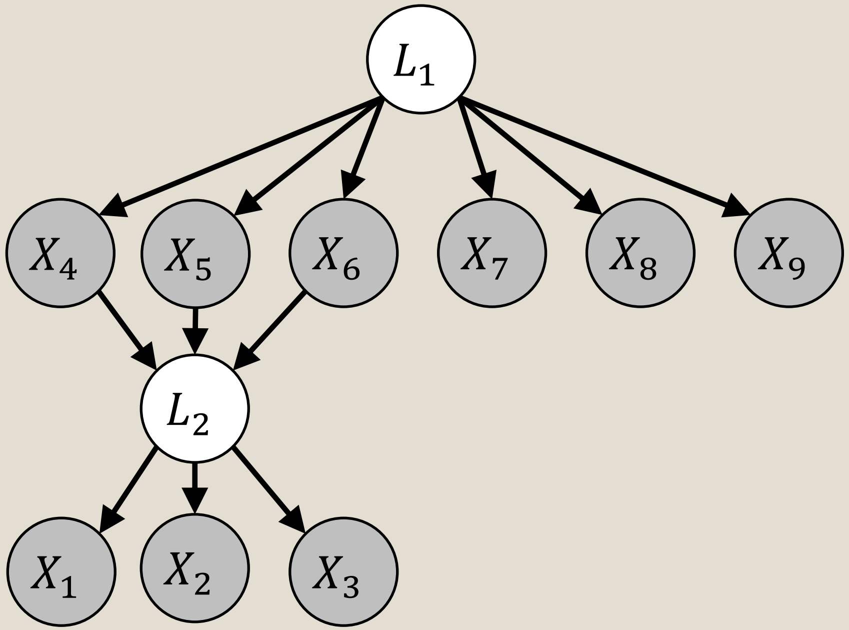


$\mathcal{G}^*$



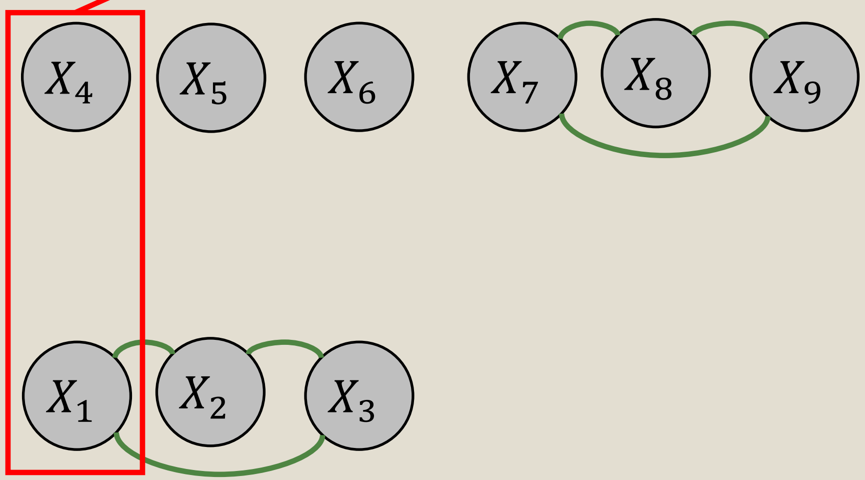
Add a temporary edge between pairs i, j such that $H_{vt}(\{i, j\}, \mathcal{V} \setminus \{i, j\})$ passes.

Phase 1: Order Clusters



$\mathcal{G}^*$

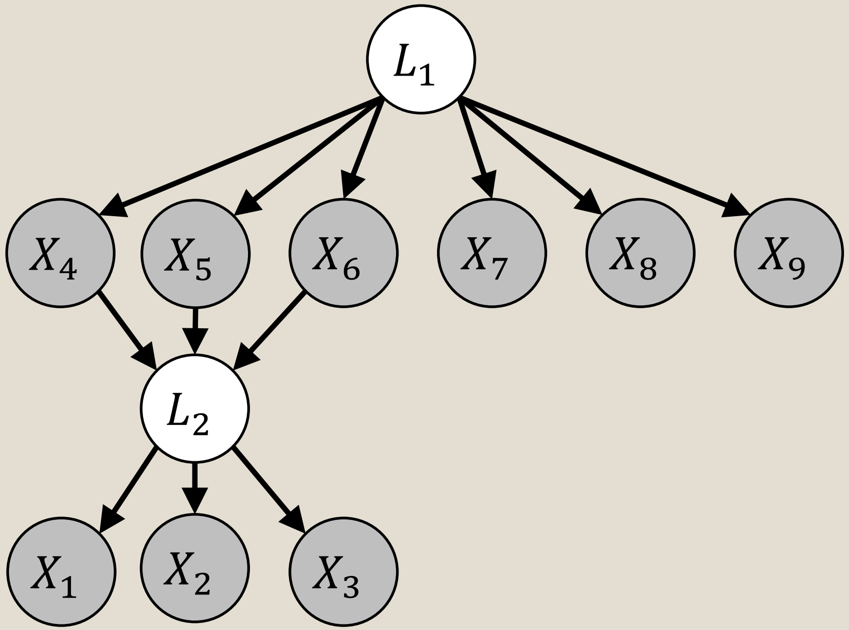
$\{X_1, X_4\}$ isn't t-separated from $\{X_2, X_5\}$ by a single node. The triple child assumption guarantees that only nodes in the same cluster will be connected.



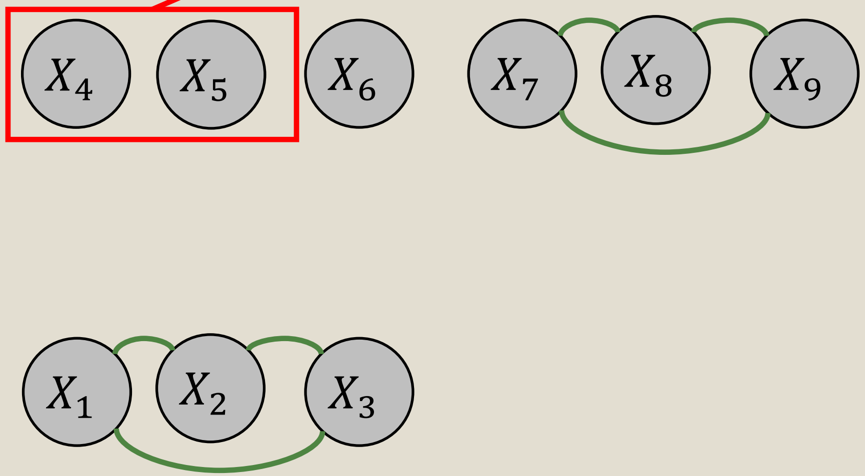
Add a temporary edge between pairs i, j such that $H_{vt}(\{i, j\}, \mathcal{V} \setminus \{i, j\})$ passes.

Phase 1: Order Clusters

$\{X_4, X_5\}$ isn't t-separated from $\{X_6, X_1\}$ by a single node. The triple child assumption guarantees that only sink nodes can be clustered.

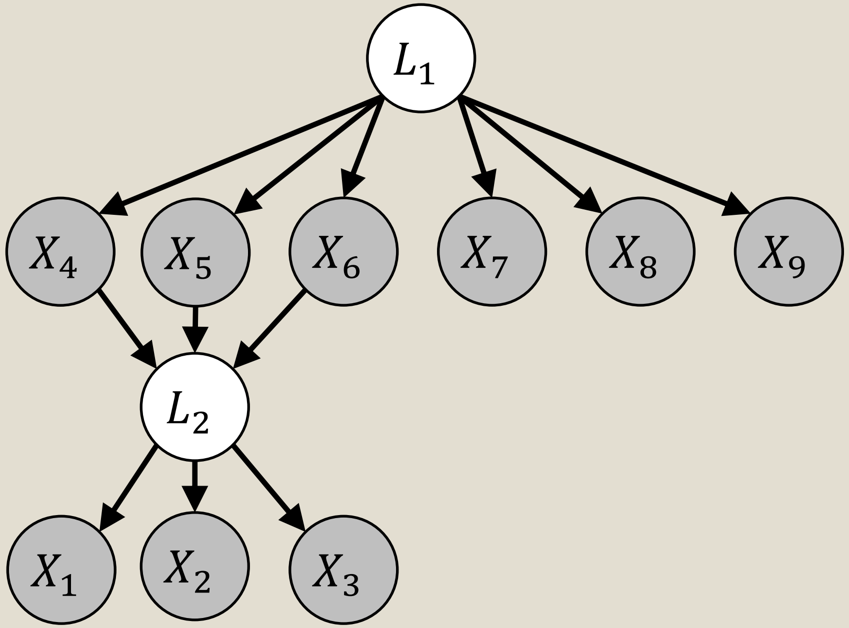


$\mathcal{G}^*$

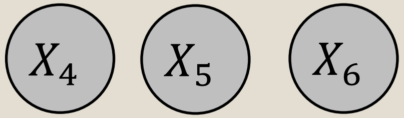


Add a temporary edge between pairs i, j such that $H_{vt}(\{i, j\}, \mathcal{V} \setminus \{i, j\})$ passes.

Phase 1:
Order Clusters



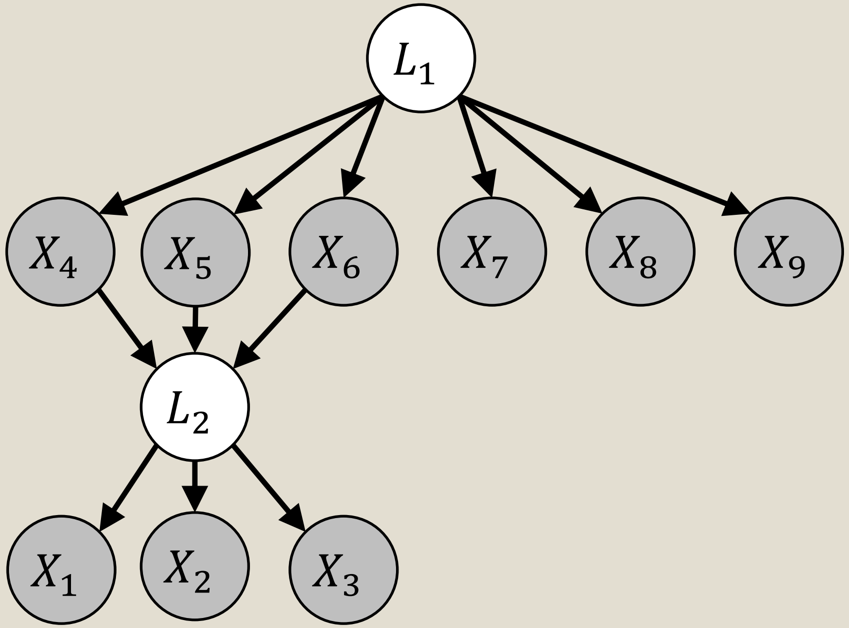
$\mathcal{G}^*$



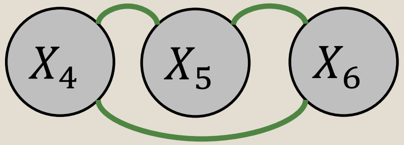
$\pi = [\{1,2,3\},\{7,8,9\}]$

Prepend cliques to an ordered list π
and remove them.

Phase 1:
Order Clusters



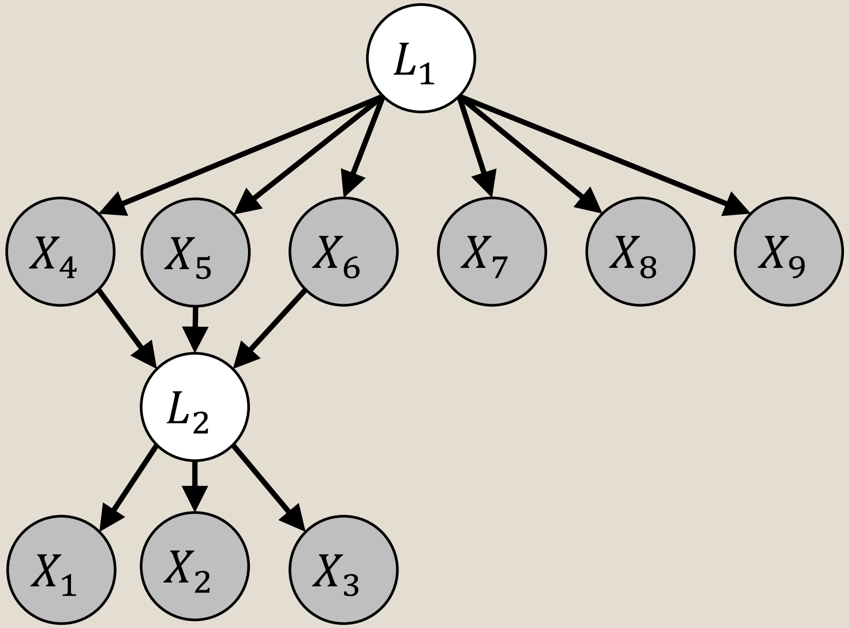
$\mathcal{G}^*$



$\pi = [\{1, 2, 3\}, \{7, 8, 9\}]$

Repeat!

Phase 1:
Order Clusters

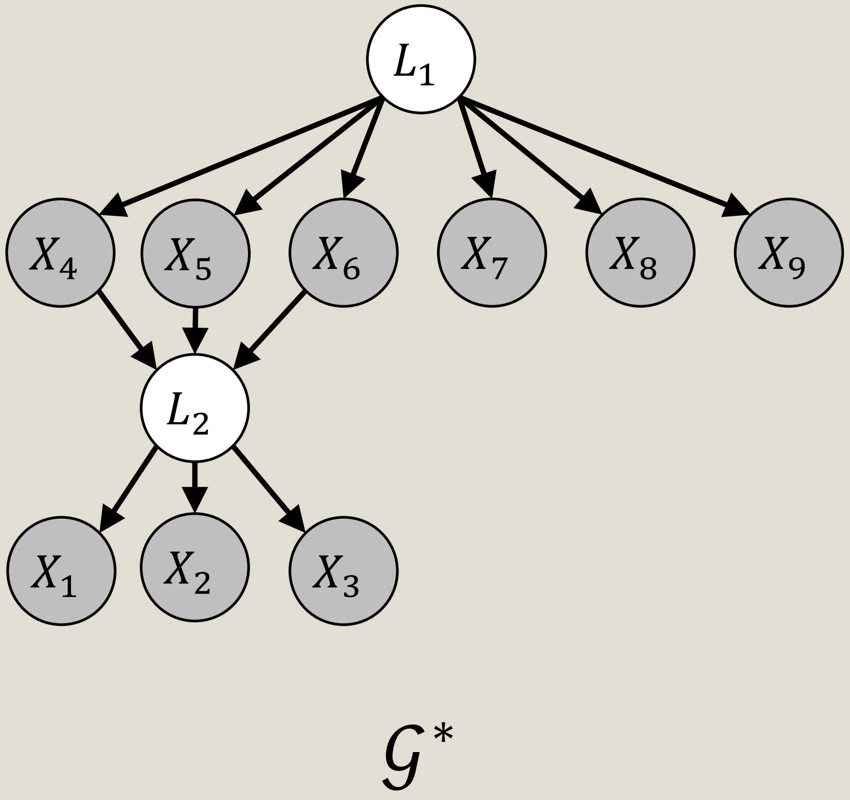


$\mathcal{G}^*$

$\pi = [\{4,5,6\}, \{1, 2, 3\}, \{7, 8, 9\}]$

Repeat!

Phase 2: Merge Clusters



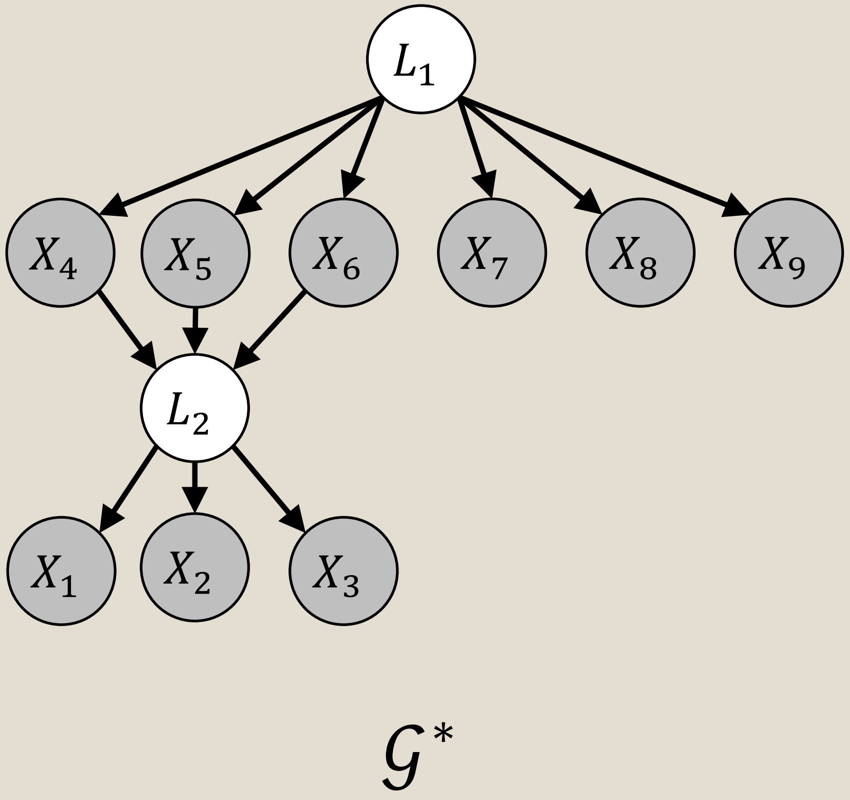
$$\pi = [\{4,5,6\}, \{1,2,3\}, \{7,8,9\}]$$

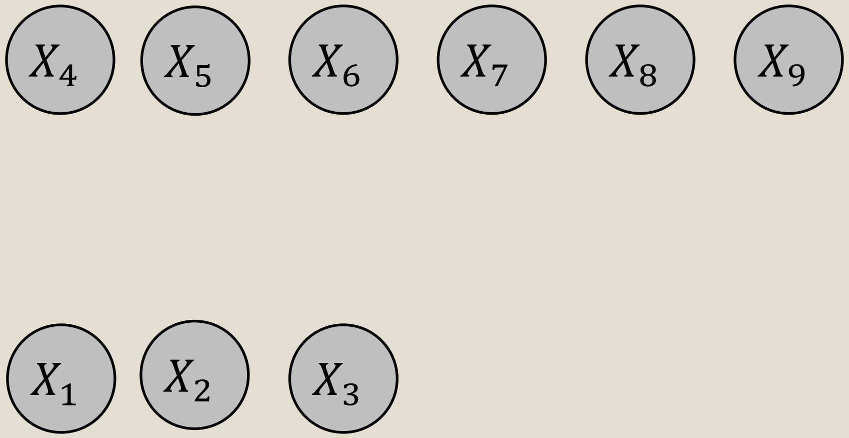
$$\downarrow$$

$$\pi = [\{4,5,6,7,8,9\}, \{1,2,3\}]$$

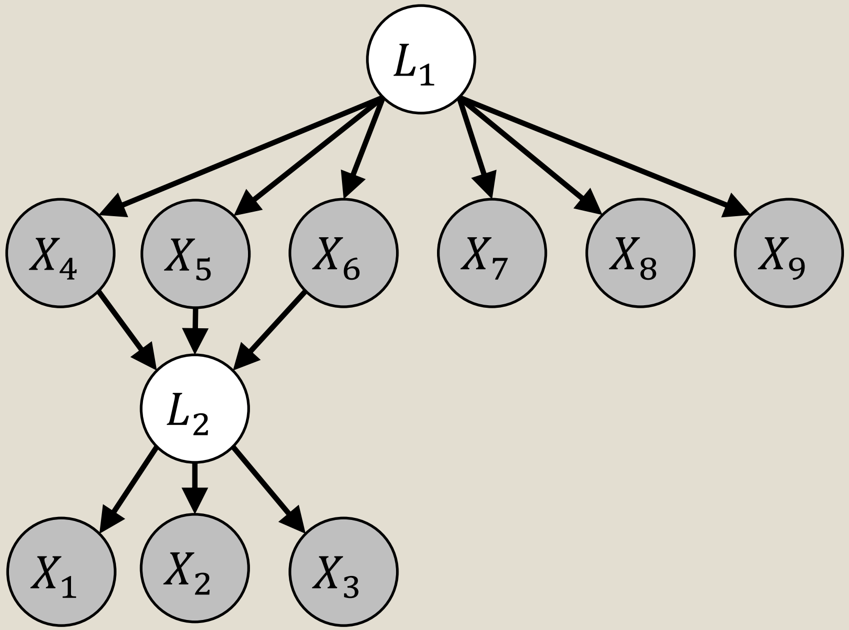
If two clusters C_1 and C_2 satisfy $H_{vt}(C_1 \cup C_2, C_1 \cup C_2)$, move nodes from C_2 to C_1.

Phase 3:
Learn DAG

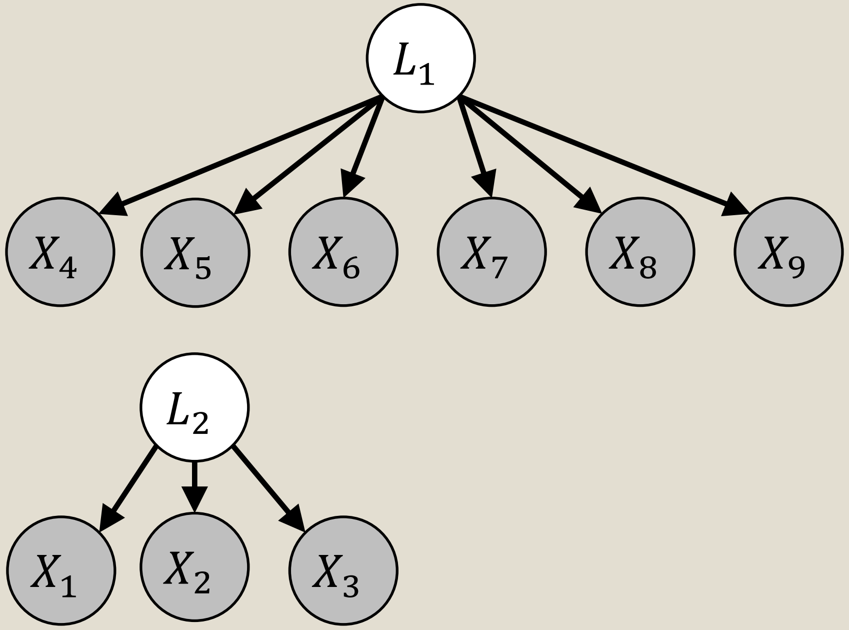




Phase 3:
Learn DAG

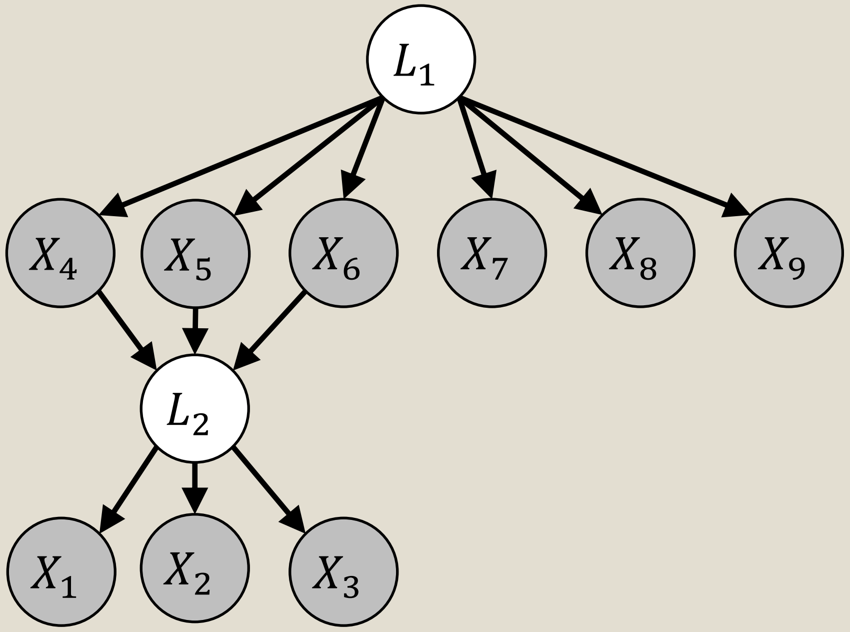


$\mathcal{G}^*$

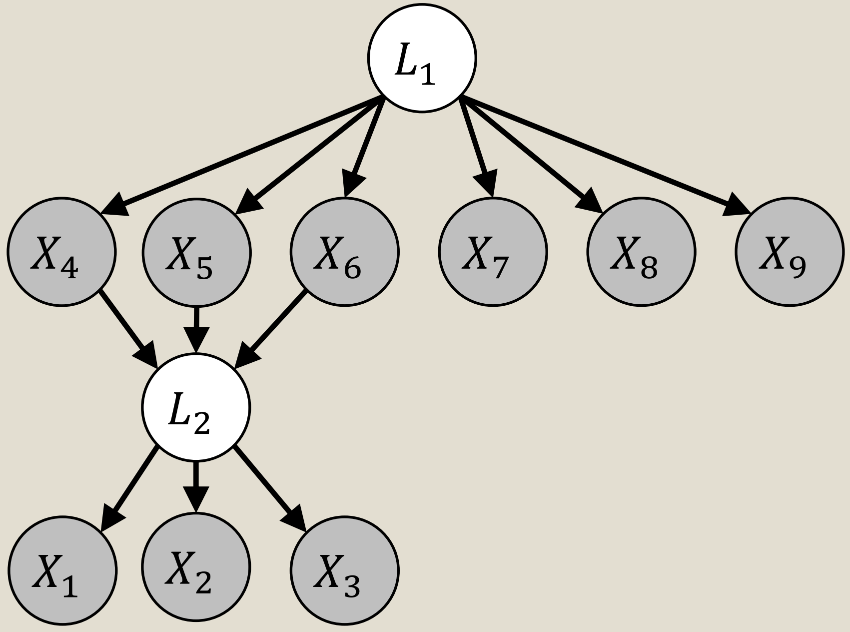


Add a latent parent for each cluster.

Phase 3:
Learn DAG



$\mathcal{G}^*$



Use conditional independence testing to recover edges from earlier cluster to later clusters.

Challenges

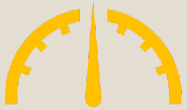
Identifiability

Under relevant faithfulness assumptions, our algorithm consistently recovers the true graph if it is a linear LFCM.

High dimensionality

$\mathcal{O}(p^5 + p^2 \cdot M^4)$ for M the largest cluster. Compare to $\mathcal{O}\left(p^{d+2}\right)$ for algorithms in the fully observed setting, with maximum in-degree d.

Latent variables



Allowed for non-exogenous latent variables.

Future Directions

- Nonlinear/non-parametric models
- Relaxing assumptions:
  - **Unique cluster assumption:** use higher rank testing.
  - **Bipartite assumption:** allow for sparse connections.
  - **Triple-child / double-parent:** modifications to the method?

Thanks!

Squires, Chandler, et al. "Causal Structure Discovery between Clusters of Nodes Induced by Latent Factors." *First Conference on Causal Learning and Reasoning*. 2021.