# Causal Structure Discovery between Clusters of Nodes Induced by Latent Factors

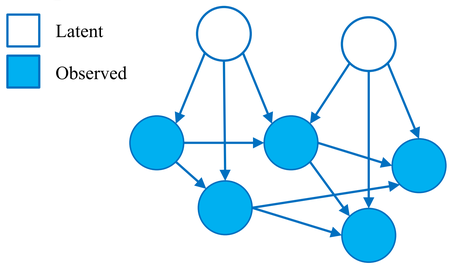
Chandler Squires[123], Annie Yun[12], Eshaan Nichani[12], Raj Agrawal[123], and Caroline Uhler[12]

[1]LIDS, MIT [2]IDSS, MIT [3]CSAIL, MIT

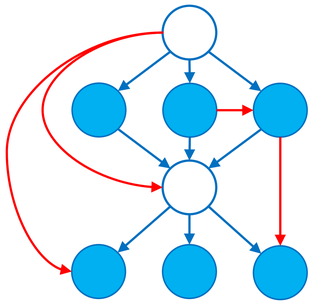
## GOAL:

### RECOVER *NON-EXOGENOUS* LATENT VARIABLES WHEN LEARNING A CAUSAL STRUCTURE FROM DATA.

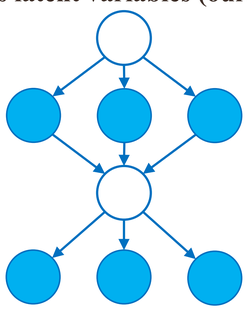
**Exogenous latent variables (most other methods)**




**Non-exogenous latent variables (general)**



**Non-exogenous latent variables (ours)**



## MOTIVATION:

### GENE REGULATORY NETWORKS

**Observed variables = gene expression levels**

**Latent variables = activity of *transcription factories* [1]**

**Clusters = spatially co-located genes**

## LATENT FACTOR CAUSAL MODEL (LFCM) ASSUMPTIONS:

### (a) UNIQUE CLUSTER

Each observed node has one latent parent.

### (b) BIPARTITE

No (observed → observed) or (latent → latent) edges.

### (c) TRIPLE-CHILD

At least 3 children per latent node.

### (d) DOUBLE-PARENT

At least two parents per non-source latent.

## PROPOSED METHOD: EstimateLFCM

$H_{vt}(A, B)$: the hypothesis that the submatrix $\Sigma_{A,B}$ of the covariance matrix is low-rank + diagonal.

- Corresponds to the **vanishing** of all **tetrads** ($2 \times 2$ non-diagonal subdeterminants).
- Use *Wishart* test: divide sample tetrad by standard error, using formula from [2].

### Phase 1: FindOrderedClusters

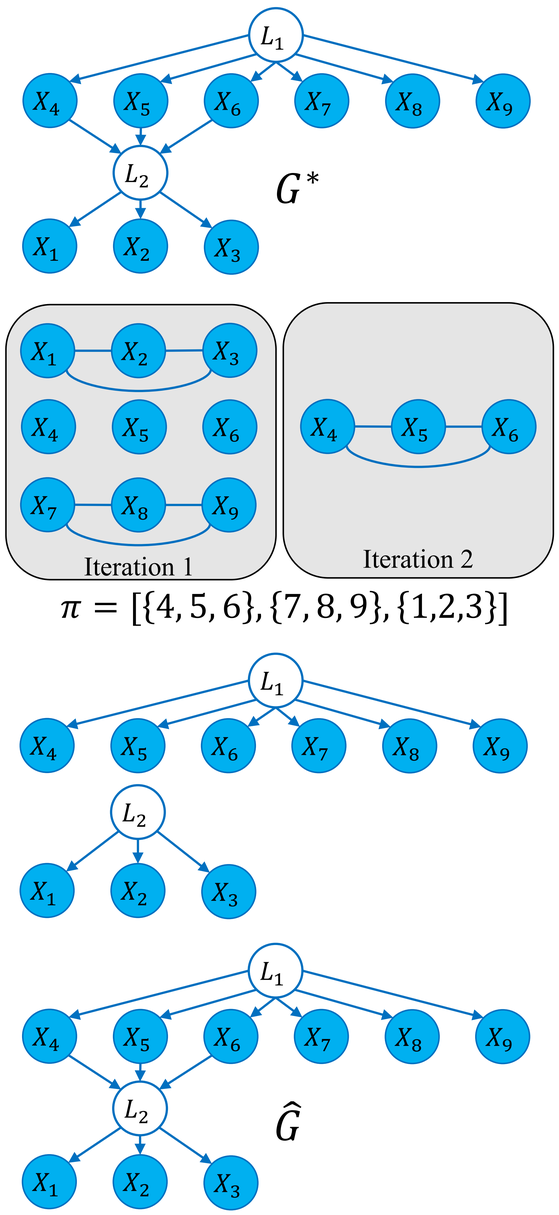
1. Add an undirected edge between observed nodes $i$ and $j$ if we accept $H_{vt}(\{i,j\}, [p] \setminus \{i,j\})$
2. Add cliques of size $\geq 3$ to the beginning of the ordering and remove them from the graph
3. Repeat until graph is empty

### Phase 2: MergeClusters

1. Merge cluster $c_1$ and $c_2$ if we accept $H_{vt}(c_1 \cup c_2, c_1 \cup c_2)$
2. Add a latent variable for each cluster, pointing to all nodes in the cluster.

### Phase 3: LearnDAG

For each node in a cluster, add an edge to the later clusters if it is not conditionally independent of the nodes in that cluster given the other nodes in its cluster and previous clusters.




## THEORETICAL RESULTS

### IDENTIFIABILITY + CONSISTENCY (THEOREM 2):

Under a **linear** LFCM model and three appropriate **faithfulness** assumptions, our algorithm consistently recovers $G^*$.

### COMPUTATIONAL COMPLEXITY (THEOREM 3):

Given $p$ variables, our algorithm takes time at most $\mathcal{O}(p^6)$. A more fine-grained analysis in terms of the maximum cluster size and number of clusters is given in the paper.

## EMPIRICAL RESULTS

### Comparisons for clustering:

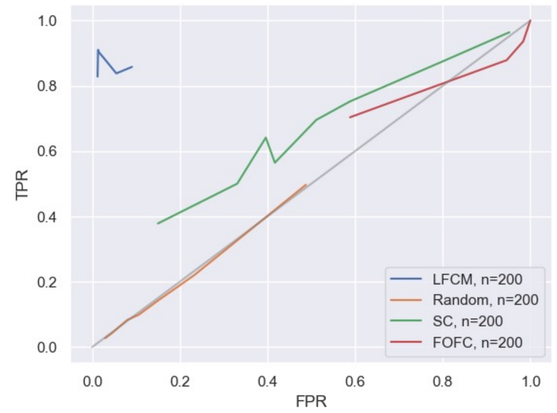
- Spectral Clustering (SC)
- FindOneFactorClusters (FOFC) [3]

### Setting:

- 50 random graphs
- 10 latent nodes
- Erdös-Rényi w/ density 0.5
- Other details in paper

### Other results in paper

- Edge recovery performance
- Semi-synthetic experiment using Sachs protein signaling data




**Our method recovers the ground truth clustering better than baselines.**

## REFERENCES

[1] Belyaeva, Anastasiya, et al. "Network analysis identifies chromosome intermingling regions as regulatory hotspots for transcription." *Proceedings of the National Academy of Sciences* 114.52 (2017): 13714-13719.

[2] Drton, Mathias, Hélène Massam, and Ingram Olkin. "Moments of minors of Wishart matrices." *The Annals of Statistics* 36.5 (2008): 2261-2283.

[3] Kummerfeld, Erich, and Joseph Ramsey. "Causal clustering for 1-factor measurement models." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.