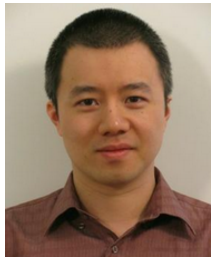
Multi-task Learning of Order-Consistent Causal Graphs

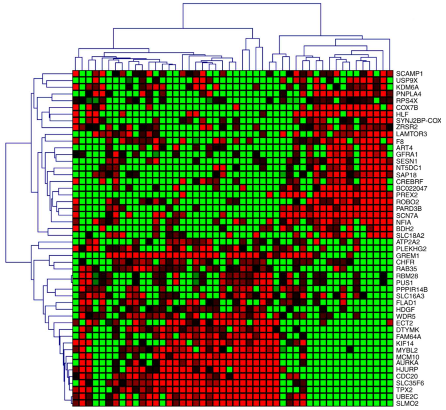
Xinshi Chen[1] Haoran Sun[1] Caleb Ellington[2] Eric Xing[2,3] Le Song[3,4]

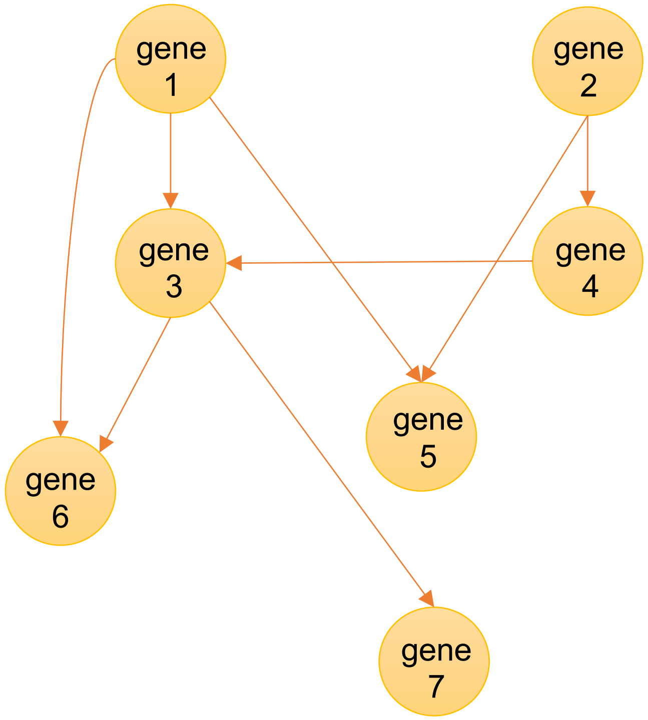
1 Georgia Institute of Technology
2 Carnegie Mellon University
3 Mohamed bin Zayed University of Artificial Intelligence
4 BioMap

Causal Discovery From Observational Data

Observational Data
(e.g., gene expression data)

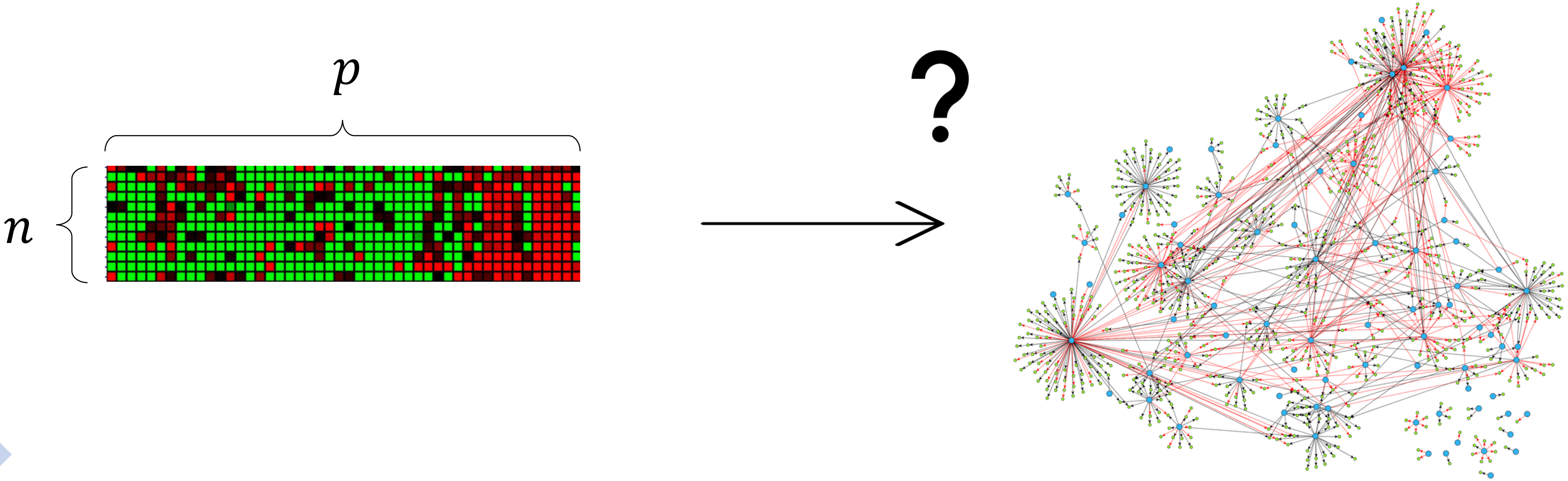


Directed Acyclic Graph (DAG)
(e.g., gene regulatory network)



Challenge 1 – Small Data

- Very few samples are observed for reconstructing the graph
- $n \ll p$



Challenge 2 – Non-identifiability

- Even with **infinite samples**, a DAG can be **non-identifiable**.
- Example:

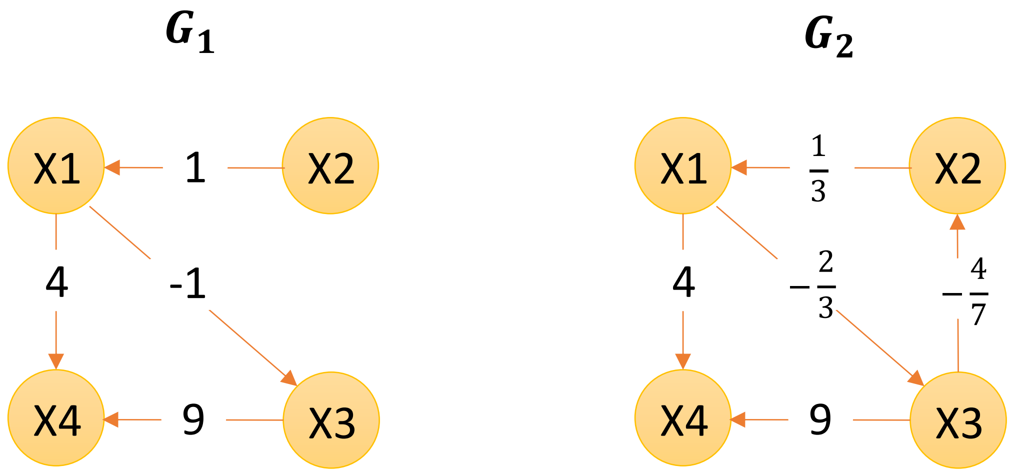


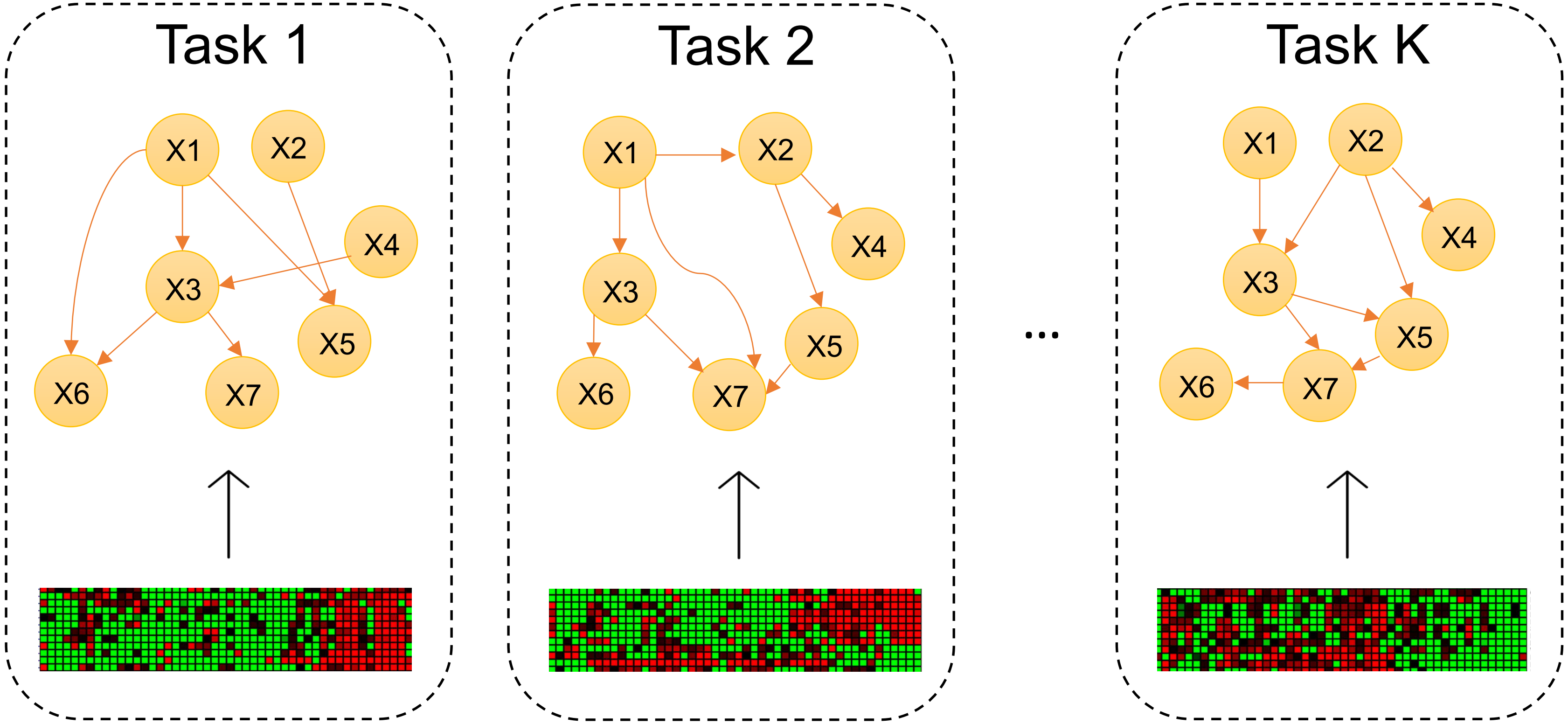
Figure 1 . *(From [Aragam, 2015])*.

Linear SEM $\begin{cases} X = G_1^\top X + W_1 \\ X = G_2^\top X + W_2 \end{cases} \Rightarrow \begin{array}{l} X \sim \mathcal{N}(0, \Sigma_1) \\ X \sim \mathcal{N}(0, \Sigma_2) \end{array}$ with $\Sigma_1 = \Sigma_2$ 

Aragam, Bryon, Arash A. Amini, and Qing Zhou. "Learning directed acyclic graphs with penalized neighbourhood regression." arXiv preprint (2015).

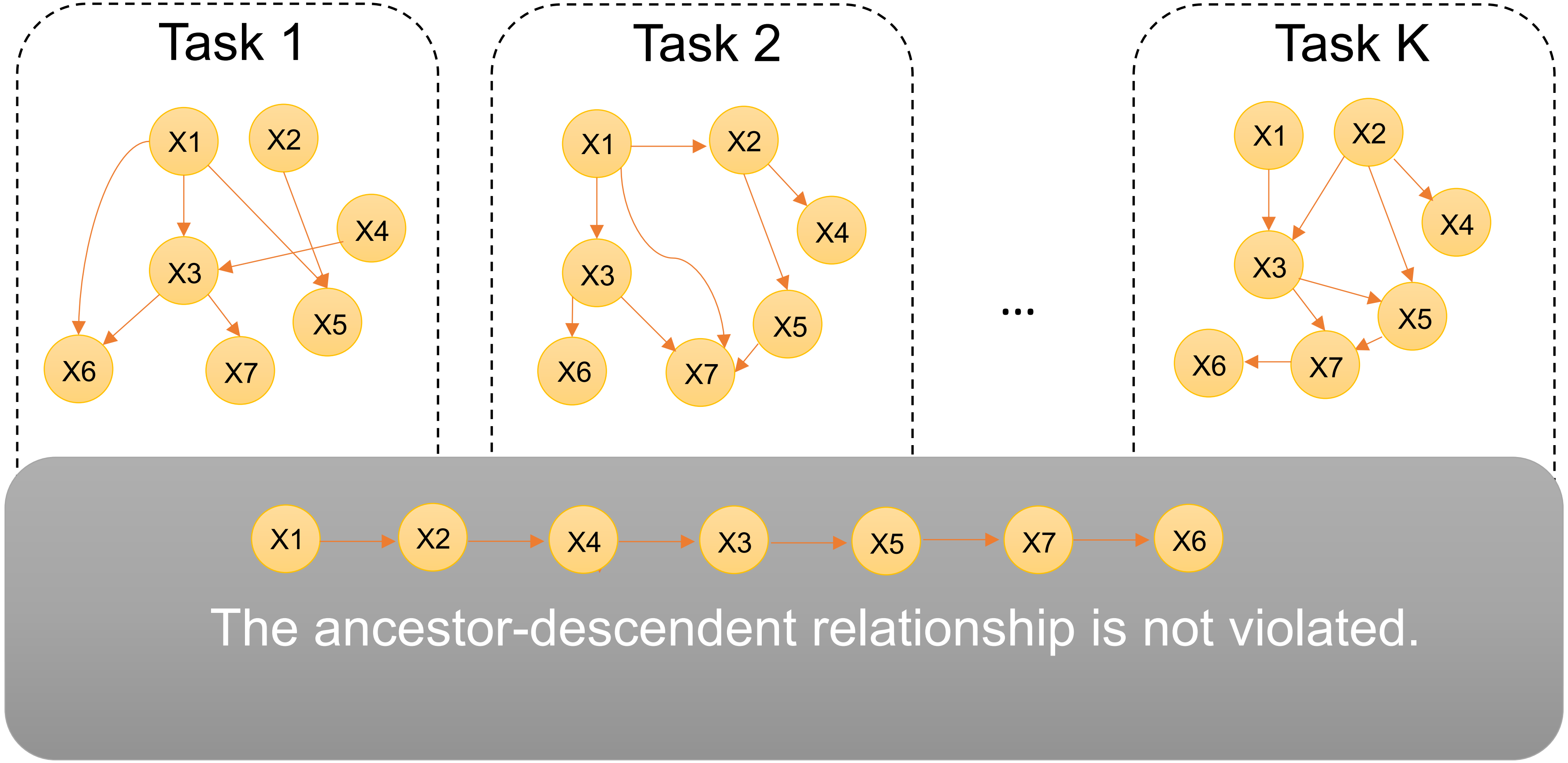
What can we utilize?

Similarity among multiple tasks!



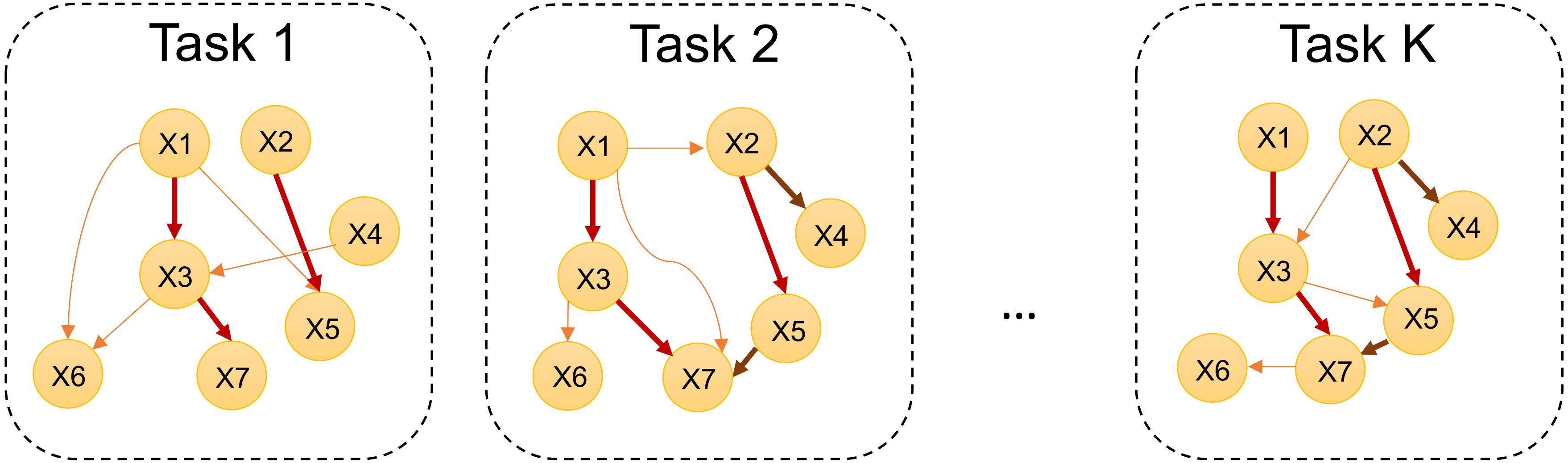
What can we utilize?

- Assumption 1 – Consistent Causal Ordering (Topological Ordering)



What can we utilize?

- Assumption 2 – Sparsity Pattern



Size of the support union of edges $|S| = s$

Multi-task Learning Setting

- K linear SEM (structural equation model)

$$\text{for } k = 1, \cdots, K, \qquad X^{(k)} = \underbrace{G_0^{(k)\top}}_{\text{DAG}} X^{(k)} + \underbrace{W^{(k)}}_{\sim \mathcal{N}(0, \Omega^{(k)})}$$

- Assume each task has n samples.

$$\underbrace{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots, \mathbf{X}^{(K)}}_{} \quad \text{where} \quad \mathbf{X}^{(k)} \in \mathbb{R}^{n \times p}$$

$$\Big\downarrow \text{ Jointly recover?}$$

$$G_0^{(1)}, G_0^{(2)}, \cdots, G_0^{(K)}$$

Joint Estimator

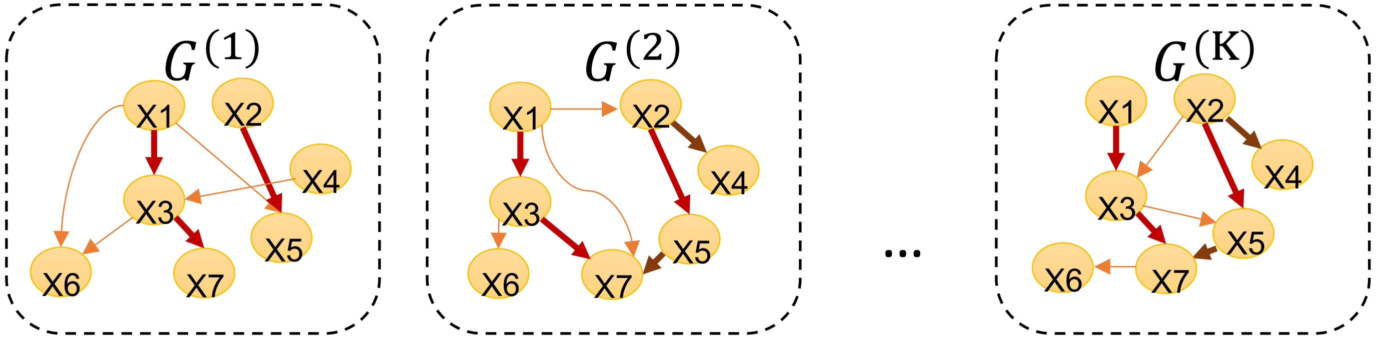
$$\min_{\pi,\ \{G^{(k)}\}_k^K} \sum_{k=1}^{K} \frac{1}{2n} ||\mathbf{X}^{(k)} - \mathbf{X}^{(k)} G^{(k)}||_F^2 + \lambda ||G^{(1:K)}||_{l_1/l_2}$$

$$\text{s.t.} \begin{cases} 1.\ \ \pi \in \mathbb{S}_p \\ 2.\ \ G^{(k)} \in \mathbb{DAG}(\pi) \end{cases}$$

1. $\pi \in \mathbb{S}_p$ represents the causal order (i.e., topological order)

$X1 \to X2 \to X4 \to X3 \to X5 \to X7 \to X6$

2. $G^{(k)} \in \mathbb{DAG}(\pi)$ is a DAG whose topological order is **consistent with π.**



Joint Estimator

$$\min_{\pi,\ \{G^{(k)}\}_k^K} \quad \sum_{k=1}^{K} \frac{1}{2n} ||\mathbf{X}^{(k)} - \mathbf{X}^{(k)} G^{(k)}||_F^2 + \lambda ||G^{(1:K)}||_{l_1/l_2}$$

$$\text{s.t.} \quad \begin{cases} \mathit{1.} \quad \pi \in \mathbb{S}_p \\ \mathit{2.} \quad G^{(k)} \in \mathbb{DAG}(\pi) \end{cases}$$

Different from separate estimation:

- It optimizes a single π *shared* across DAGs.
- The *group norm* $||G^{(1:K)}||_{l_1/l_2}$ penalizes the size of union support softly.

Questions

Theoretical questions:

- Is this joint estimator leading to an *improved sample complexity*?
- Can this joint estimator help to recover *non-identifiable DAGs*?

Practical question:

- How to compute the minimizer $\pi, \{G^{(k)}\}_k^K$ *efficiently?*

Main Result – Identifiable Case

- Assume for each k, $G_0^{(k)}$ *is a unique minimum-trace DAG*

- *(Theorem 3.1) Recovering the true causal order* π_0

  A sample complexity measure: $\theta(n, K, p, s) = \frac{p}{s}\sqrt{\frac{nK}{p \log p}}$

  the rate at which the sample size must grow

- *(Theorem 3.1) Recovering the DAGs*

  Averaged error: $\frac{1}{K}\sum_{k=1}^{K} ||G^{(k)} - G_0^{(k)}||_F^2 = \mathcal{O}(s\sqrt{\frac{p \log p}{nK}})$

Main Result – Non-identifiable Case

- Assume K' DAGs $\{G_0^{(1)}, \cdots, G_0^{(K')}\}$ are *identifiable*.
- The other $K - K'$ DAGs are *non-identifiable*.

- *Recovering the true causal order π_0*

A sample complexity measure: $\cancel{\theta(n,K,p,s) = \frac{p}{s}\sqrt{\frac{nK}{p\log p}}}$

$$\theta(n,K,p,s) = \frac{p}{s}\sqrt{\frac{1}{p\log p}\frac{nK'^2}{K}}$$ ✓

Practical Algorithm

$$\min_{\pi,\ \{G^{(k)}\}_k^K} \sum_{k=1}^{K} \frac{1}{2n} ||\mathbf{X}^{(k)} - \mathbf{X}^{(k)} G^{(k)}||_F^2 + \lambda ||G^{(1:K)}||_{l_1/l_2}$$

$$\text{s.t.} \begin{cases} \textit{1.} \quad \pi \in \mathbb{S}_p \\ \textit{2.} \quad G^{(k)} \in \mathbb{DAG}(\pi) \end{cases}$$

- How to compute the optimal solution $\pi, \{G^{(k)}\}_k^K$ *efficiently?*

Practical Algorithm

Key: an equivalent continuous formulation.

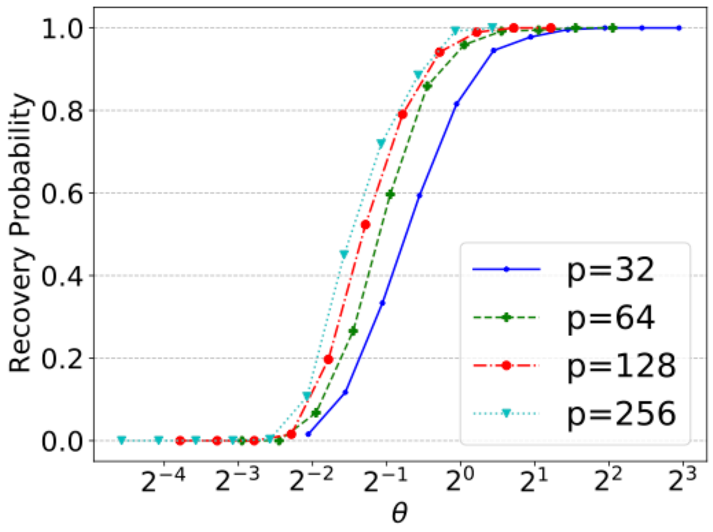
$$\min_{\substack{T \in \mathbb{R}^{p \times p} \\ G^{(1)}, \cdots, G^{(K)} \in \mathbb{R}^{p \times p}}} \sum_{k=1}^{K} \frac{1}{2n} \left\| \boldsymbol{X}^{(k)} - \boldsymbol{X}^{(k)} \overline{G}^{(k)} \right\|_F^2 + \lambda \| \overline{G}^{(1:K)} \|_{l_1/l_2} + \rho \| \mathbf{1}_{p \times p} - T \|_F^2$$

$$\text{subject to} \quad h(T) := \operatorname{trace}(e^{T \circ T}) - p = 0,$$

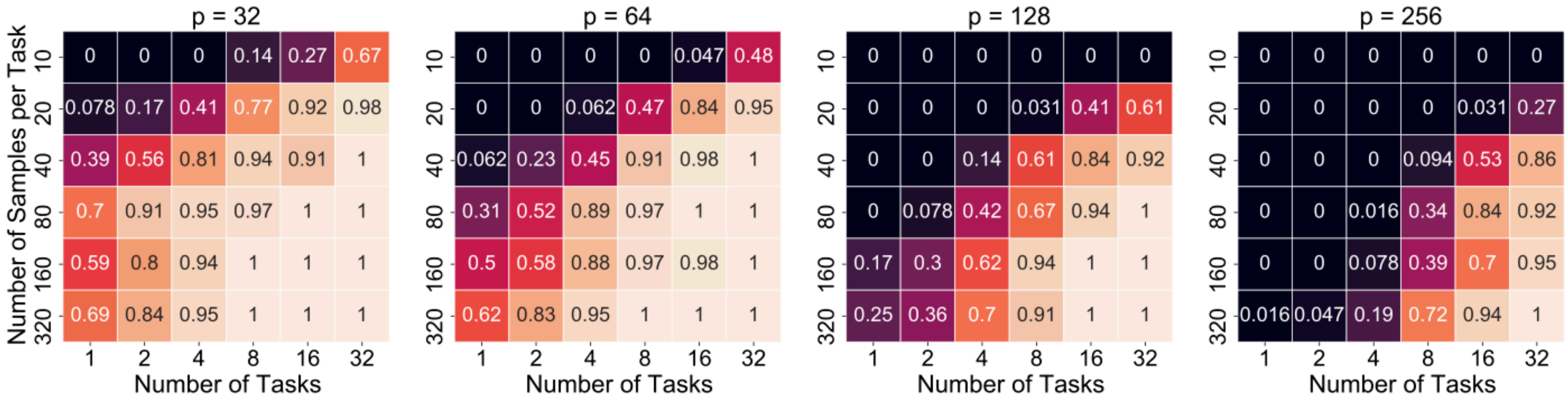
where $\overline{G}^{(k)} := G^{(k)} \circ T$ is element-wise multiplication between $G^{(k)}$ and T

Synthetic Experiment: Linear SEM

- Order recovery probability versus theoretical sample complexity: $\theta = p/s\sqrt{nK/(p\log p)}$

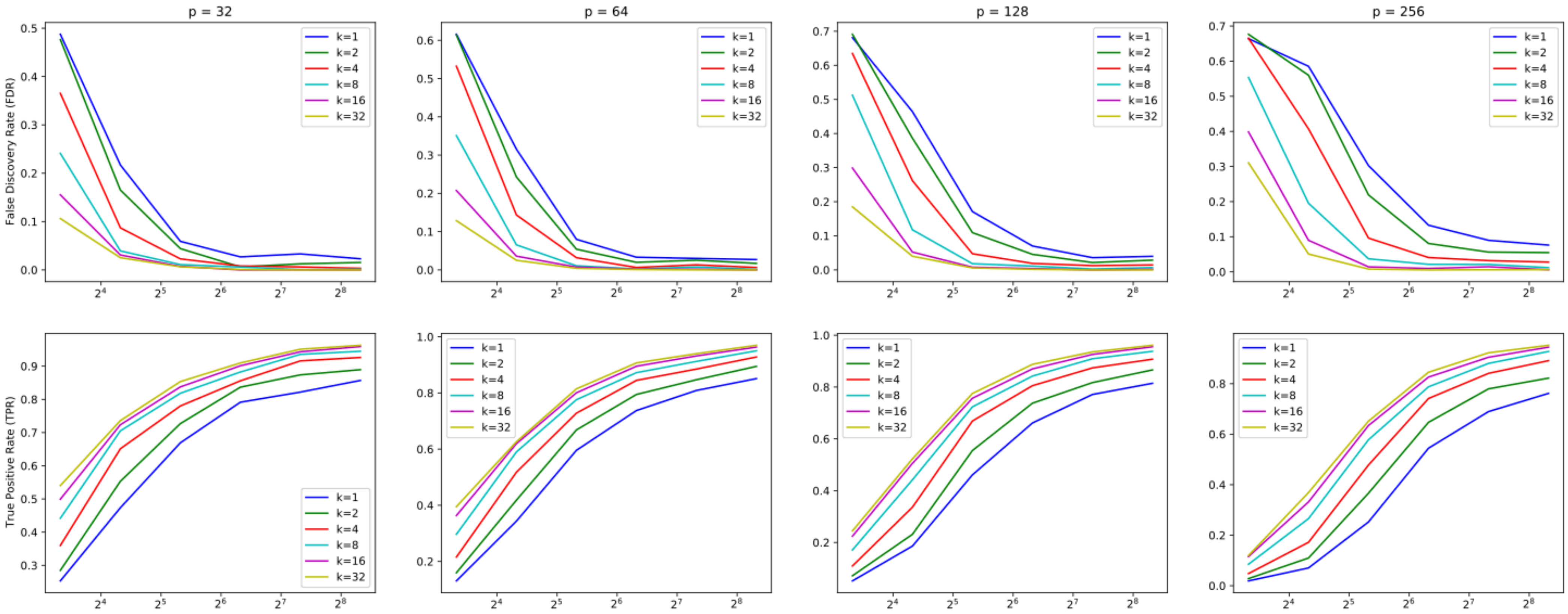


- Order recovery probability under different problem sizes, number of tasks, and number of samples per task.



Synthetic Experiment: Linear SEM

- Structure recovery quality with different numbers of tasks in False Discovery Rate (FDR), and True Predictive Rate (TPR)



Gene Expression Experiment using SERGIO

Rate of gene i expression

$$\frac{dx_i}{dt} = P_i x_i - \beta(x_i)$$

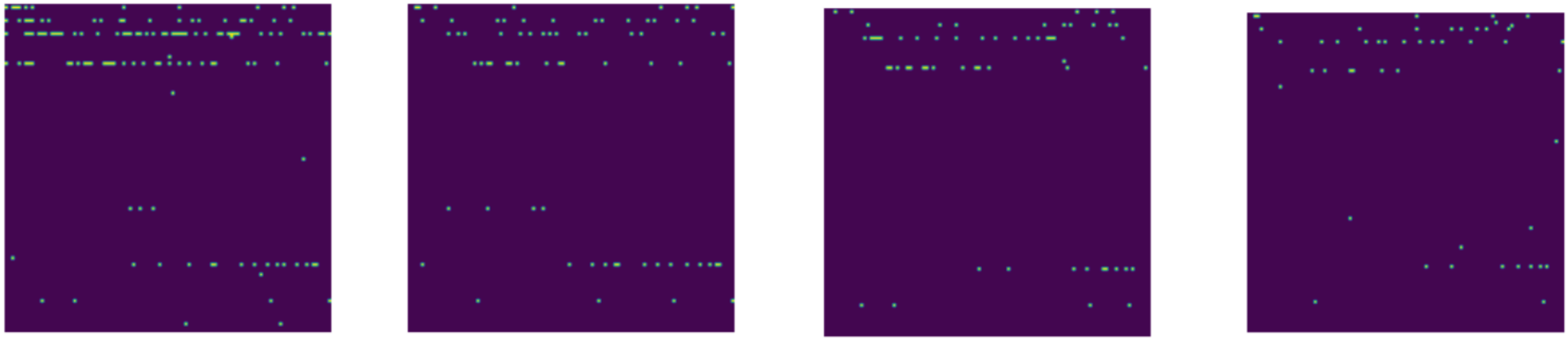
Total production rate of gene i

$$P_i = \sum_{j \in R_i} p_{ij} + b_i$$

Effect of regulator j on gene i

$$p_{ij} = K_{ij} \frac{x_j^{n_{ij}}}{h_{ij}^{n_{ij}} + x_j^{n_{ij}}}$$

(a) True: Ecoli 100   (b) $K = 1, n = 1000$ (c) $K = 10, n = 100$ (d) $K = 1, n = 100$



FDR: 0.11, TPR: 0.47   FDR:0.06, TPR:0.41   FDR: 0.18, TPR: 0.29