Filter, Rank, and Prune: Learning Linear Cyclic Gaussian Graphical Models

Soheun Yi and Sanghack Lee

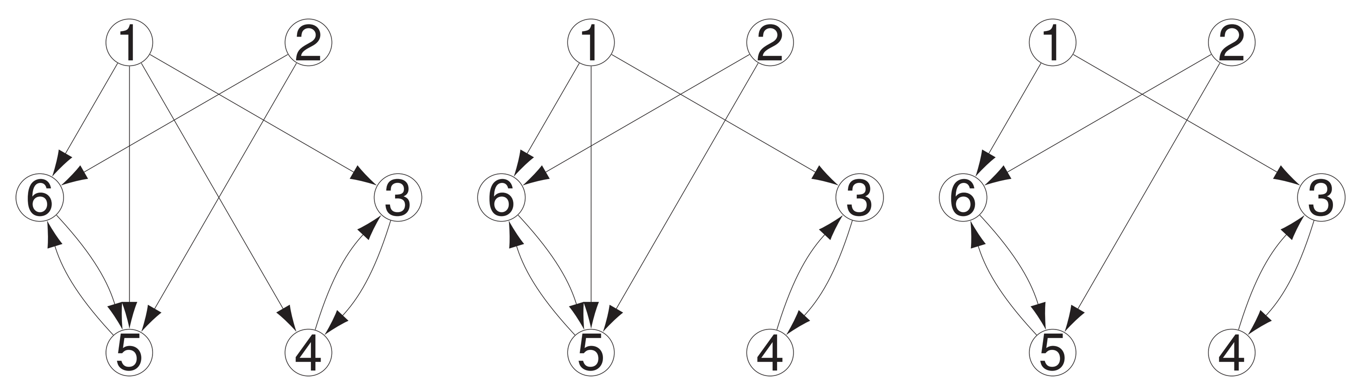
Department of Statistics and Data Science, Carnegie Mellon University
Graduate School of Data Science, Seoul National University

Motivation & Background

Linear Gaussian SEM

- Linear Gaussian SEM: $X = XW + \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathrm{diag}(\Omega))$ where Ω is a positive diagonal matrix.
- The corresponding graph has $\{(i,j) \mid W_{ij} \neq 0\}$ as edges.
- The precision matrix (inverse covariance matrix) is $\Theta = (I - W)\Omega^{-1}(I - W)^\top$.

Equivalence Classes of Linear Gaussian SEMs



- There are equivalent graphs in the sense that they can represent the same set of Gaussian distributions (Ghassami et al., 2020).
- We choose a graph with the smallest number of edges as the "best" graph in the equivalence class.

Performance Measure

- $\mathcal{L}(E, \Theta)$ measures "how close can E represent Θ", given by the minimum of KL divergence between $\mathcal{N}(0, \Theta^{-1})$ and $\mathcal{N}(0, (I_p - \widehat{W})^{-1}\widehat{\Omega}(I_p - \widehat{W})^{-\top})$ over all $\widehat{W}$ supported on E and positive diagonal matrices $\widehat{\Omega}$.
- $\mu|E|$ measures "how simple is E", given by the number of edges in E times $\mu = \frac{\log n}{2n}$, which matches with the BIC.
- $\mathcal{L}_\mu(E, \Theta) = \mathcal{L}(E, \Theta) + \mu|E|$.

Relaxing Faithfulness Assumption

- Faithfulness: $X_i \perp\!\!\!\perp X_j \mid X_{\mathbf{S}} \Leftrightarrow$ i and j are independent given $\mathbf{S}$ in a SEM.
- Strong faithfulness: $|\mathrm{corr}(X_i, X_j \mid X_{\mathbf{S}})| \leq \lambda \Leftrightarrow$ i and j are independent given $\mathbf{S}$ in a SEM.
- The strong faithfulness condition is too restrictive (Uhler et al., (2013)).
- We relax this by suggesting λ-edge faithfulness: $(i,j) \in E \Rightarrow |\mathrm{corr}(X_i, X_j \mid X_{V\setminus\{i,j\}})| > \lambda$.
- The first phase of our method is **consistent** given the λ-edge faithfulness assumption.

Filter Stage

- We first "filter" edges by removing (i,j) pairs with small
$$\psi_{ij} = |\mathrm{corr}(X_i, X_j \mid X_{V\setminus\{i,j\}})|,$$
reflecting our λ-edge faithfulness assumption.
- These values for all (i,j) can be efficiently estimated by
$$\widehat{\psi}_{ij} = \left| \frac{\widehat{\Theta}_{ij}}{\sqrt{\widehat{\Theta}_{ii}\widehat{\Theta}_{jj}}} \right|,$$
where $\widehat{\Theta}$ is the MLE of the precision matrix.
- The threshold is set proportional to $n^{-1/4}$, where n is the sample size.
- This procedure is consistent due to concentration of the partial correlations.

Rank and Prune Stage

The filter stage gives us a set of candidates of edges $\widehat{E}_0$: we take iterative steps of "ranking" and "pruning" to obtain the final estimate $\widehat{E}$.

Ranking Stage

To "rank" how important each edge candidate is, we get a solution $\widehat{Q}$ for

$$\begin{aligned} \underset{Q \in \mathbb{R}^{p\times p}}{\text{minimize}} \quad & \mathrm{KL}_{\mathcal{N}}(\widehat{\Theta} \,\|\, QQ^\top) + \mathrm{reg}(Q), \\ \text{subject to} \quad & Q_{ij} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin \widehat{E} \end{aligned}$$

and rank $(i,j) \in \widehat{E}$ with respect to $|\widehat{Q}_{ij}|$ in ascending order.

Pruning Stage

- An edge candidate (i,j) with $\widehat{Q}_{ij}$ can be removed without sacrificing the accuracy measure $\mathcal{L}(E, \widehat{\Theta})$.
- We remove edges starting from the first rank until the accuracy worsening exceeds a threshold (ϵ_{tol}).

Complete Algorithm

Run the following N_{init} times in a parallel manner:

1. Run Filter stage to obtain initial edge candidates.
2. Run Rank and Prune stages until we cannot remove more edges.

Report output edges with smallest (sample) loss.

Experiments & Results

Methods

- **FRP**: Our proposed method.
- **NOTEARS**: A continuous optimization-based algorithm to estimate an acyclic structure from observational data.
- **DGLEARN**: A combinatorial approach to estimate a Gaussian graphical model from observational data (Ghassami et al.).
- **GOLEM**: A continuous optimization-based algorithm to estimate a possibly cyclic structure based on observational data.
- **NODAGS-Flow**: A continuous optimization-based algorithm to estimate a possibly cyclic structure based on both observational and interventional data.

Datasets

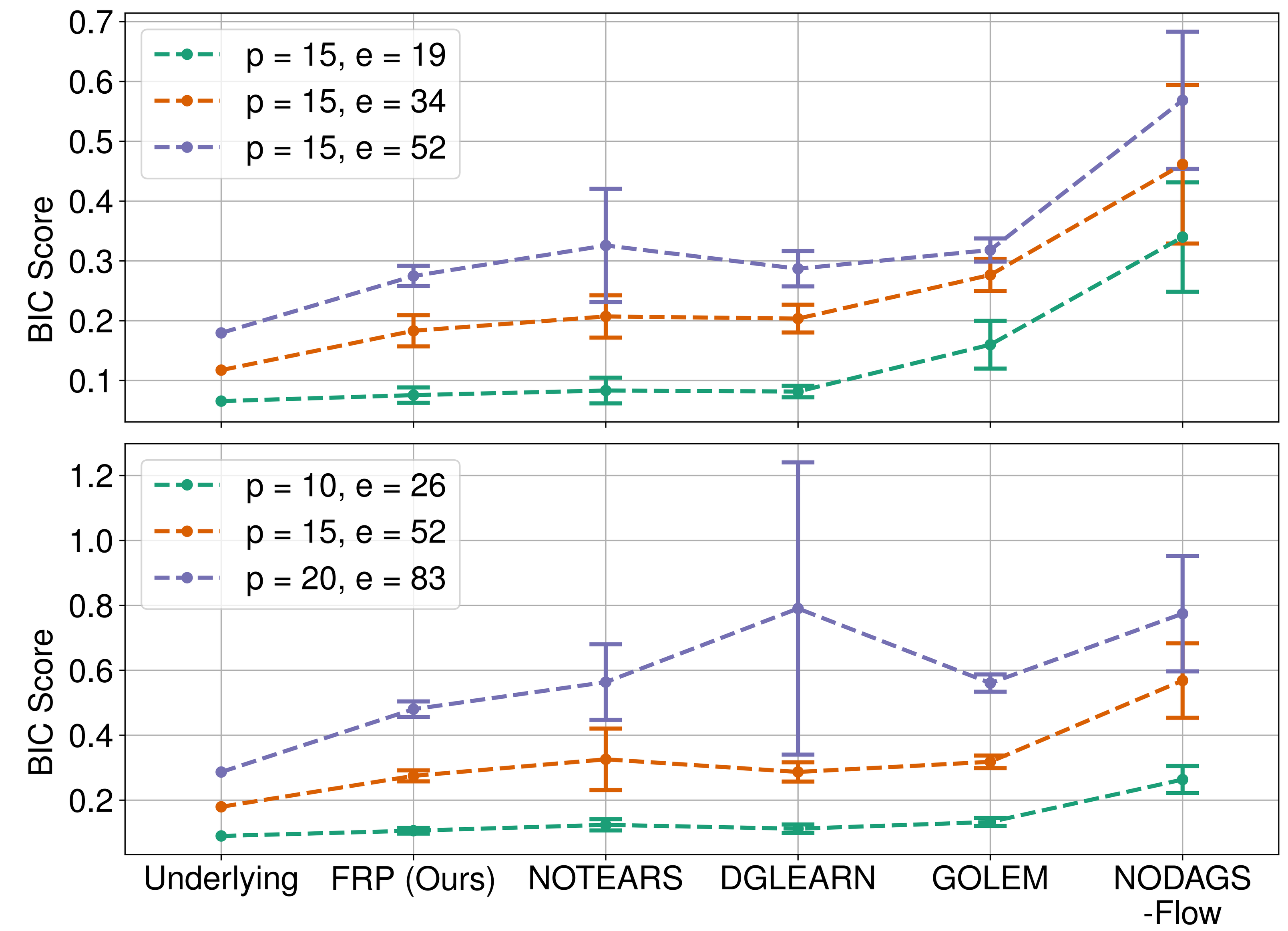
- Synthetic dataset with $n = 1000$ and dimension $p = 10, 12, \ldots, 20$.
- 4 different number of edges.
- Random weights for edges following (Ghassami et al.).

Metric

- $\mathcal{L}_\mu(E, \Theta)$: A BIC-like performance measure we previously defined.

Result: Better Performance than All Baselines

(Smaller the better performance)



https://github.com/soheunyi/frp