SORBONNE UNIVERSITÉ

LPSM LABORATOIRE DE PROBABILITÉS STATISTIQUE & MODÉLISATION

Sparse Gaussian Graphical Models for Structured Biological Networks

Roland Sogan

Sorbonne Université
Laboratoire de Probabilités, Statistique et Modélisation (LPSM)

Journées NetBio

November 24, 2025

Contents

Context

Complex networks in biology

Complex networks are omnipresent in biology.

- **Molecular Networks** Gene regulatory networks, protein-protein interaction networks, and metabolic pathways describe how genes, proteins, and metabolites coordinate cellular functions.

- **Cellular Networks** Signaling networks mediate communication within and between cells, orchestrating processes such as growth, differentiation, and immune response.

Data: Gene expression matrix

$$X \in \mathbb{R}^{n \times p}$$

n samples, p genes

X_{kj} = expression of gene j in sample k

Goal: Infer gene regulatory network

- Nodes = genes (p nodes)
- Edges = direct interactions

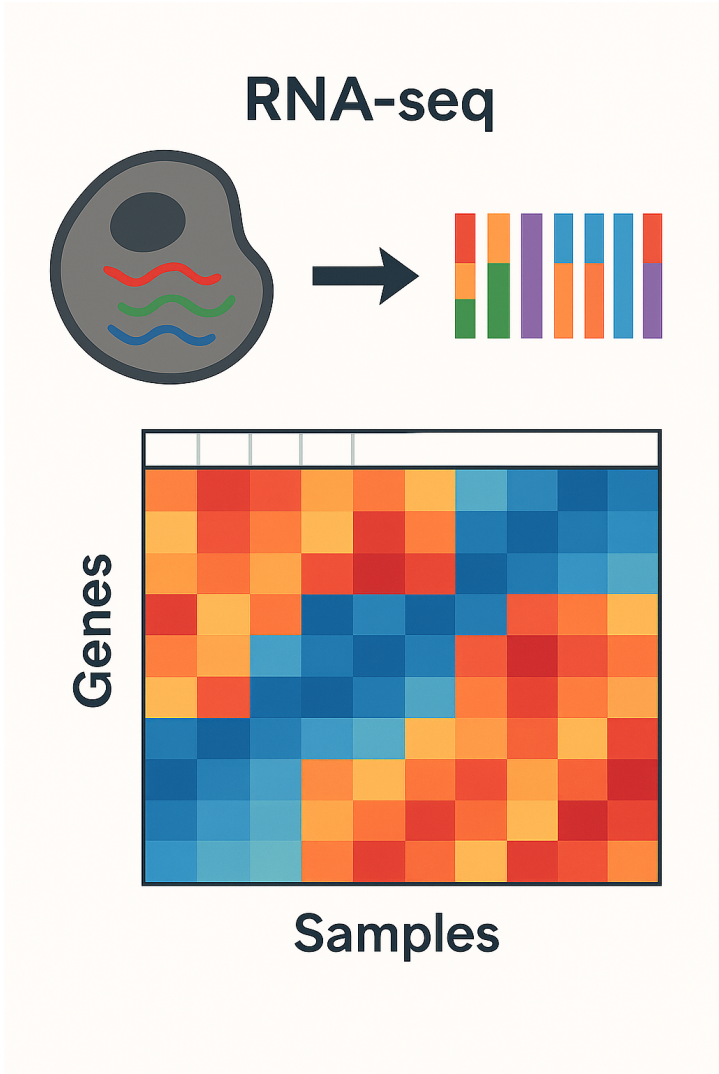


Figure: Gene expression matrix: rows = genes, columns = samples

Correlation Networks: A Naive Approach

Simple idea: Connect genes with high correlation

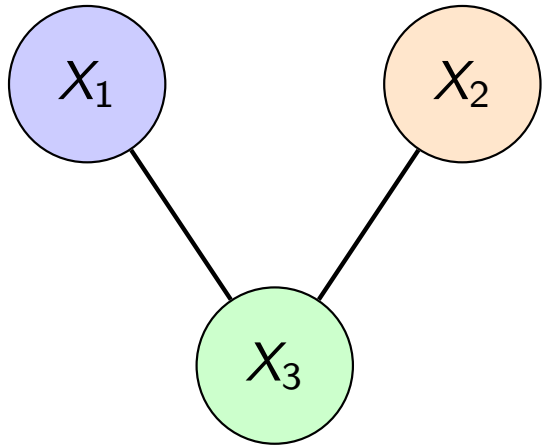
$$\text{Edge } (i,j) \Leftrightarrow |\text{corr}(X_i, X_j)| > \tau$$

Limitation: Correlation $\neq$ Direct interaction

- Captures indirect dependencies
- Cannot distinguish direct effects

Better approach: Conditional independence networks

Example:



True: $X_1 - X_3 - X_2$
But: $\text{corr}(X_1, X_2) \neq 0$

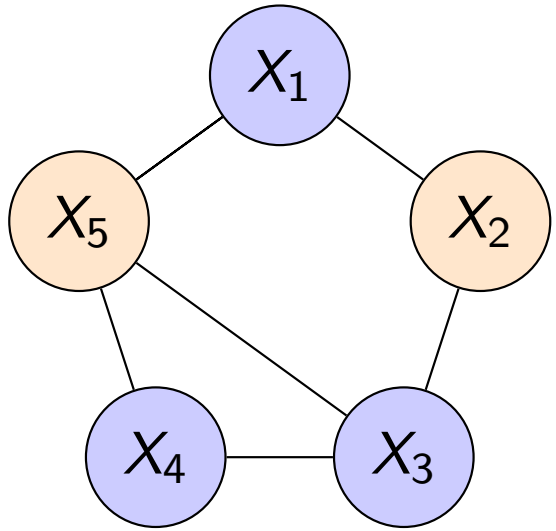
Undirected Graphical Models

conditional independence graph

An Undirected Graphical Model: represents $X = (X_1, ..., X_p)$ as graph $G = (V, E)$, where

- Nodes V = variables;
- Edges E = the pairwise relationships.

Example:



$X_2 \perp\!\!\!\perp X_5 \mid X_{\setminus\{2,5\}}$

Adjacency Matrix:

$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$$

This is called conditional independence graph, when

$$\text{No edge } (i,j) \Leftrightarrow X_i \perp\!\!\!\perp X_j \mid X_{\setminus\{i,j\}}$$

Gaussian Graphical Models

Gaussian Graphical Models (GGMs)

- **GGM Definition**: A conditional independence graph where

$$X \sim \mathcal{N}(\nu, \Sigma)$$

- **Precision matrix**:

$$K = \Sigma^{-1} \quad \text{(encodes conditional dependencies)}$$

Key Relationship:

$$X_i \perp\!\!\!\perp X_j \mid X_{\setminus\{i,j\}} \Leftrightarrow \operatorname{corr}(X_i, X_j \mid X_{\setminus\{i,j\}}) = 0$$
$$\Leftrightarrow K_{ij} = 0$$

(Missing edge $\Leftrightarrow$ Zero in K $\Leftrightarrow$ Conditional independence)

Application: Gene Regulatory Networks

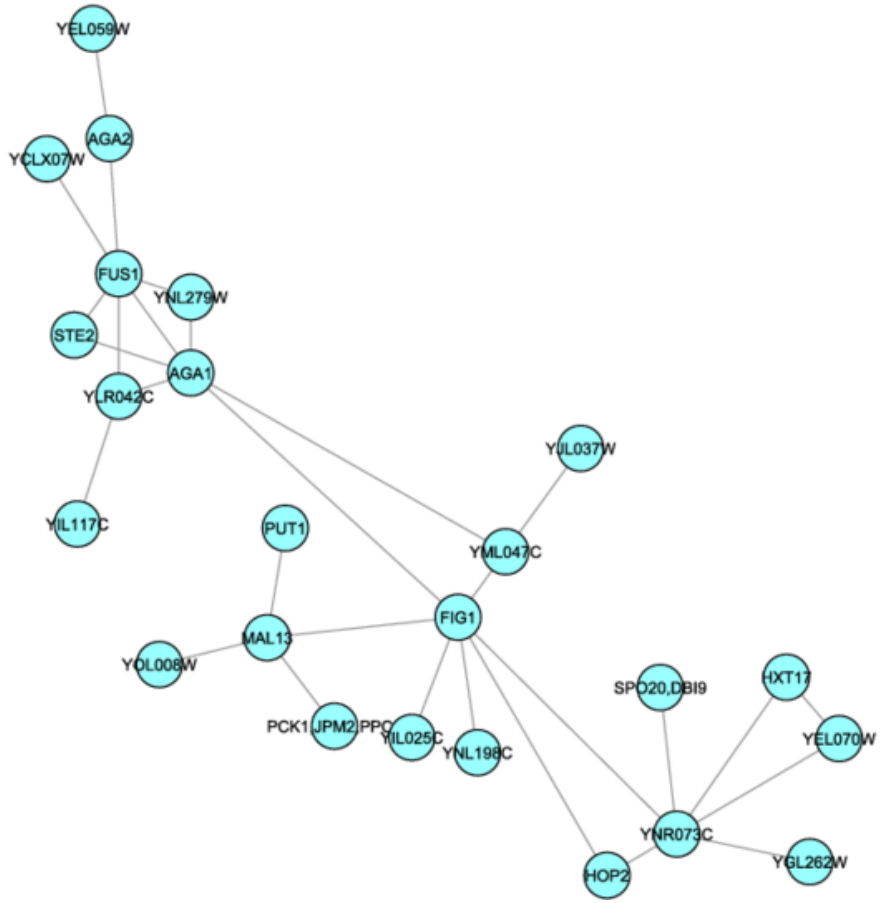


Figure: Iconix Microarray Data (Banerjee et al. (2008)).

Application: US Senate Voting Patterns

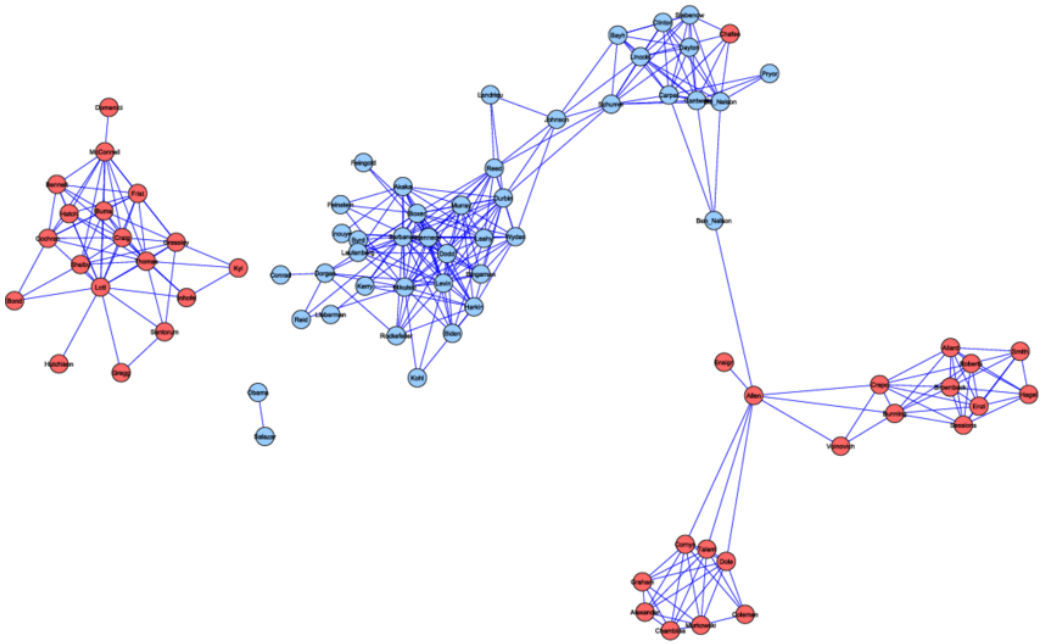


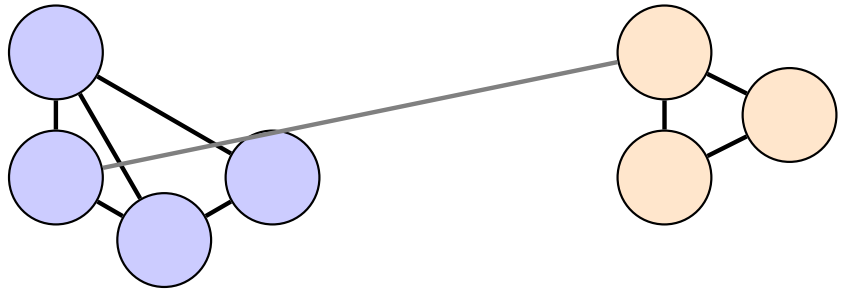
Figure: Graph structure of the US Senate (109th Congress, 2004–2006)(Banerjee et al. (2008)).

Beyond Sparsity: Modular Structure in Gene Networks

Sparse precision matrix identifies direct interactions

Biological networks exhibit modular organization:

- Genes cluster into functional modules
- Dense connections within modules
- Sparse connections between modules



$\Rightarrow$ **Block structure** in precision matrix

Setup

Data

- n samples $\times$ p biological variables matrix:

$$X \in \mathbb{R}^{n \times p}, \quad X_i \sim \mathcal{N}(0, K^{-1}) \text{ i.i.d.}$$

Assumptions

- **Sparsity**: few biological interactions
- **Clustering**: functional variable groups

Goal

Reconstruct biological network with cluster structure

Inference of GGM in the Literature

Maximum Likelihood Estimation (MLE)

Standard Approach

The log-likelihood for precision matrix $K \succ 0$:

$$\mathcal{L}(K) = \frac{n}{2} (\log \det K - \operatorname{tr}(SK))$$

where S is the sample covariance.

MLE solution: $\widehat{K}_{\text{MLE}} = S^{-1}$

Why MLE Fails in High Dimensions ($p \gg n$)

- × **Non-identifiability**: S singular when $p > n$
- × **No structure exploitation**: Ignores sparsity

$\Rightarrow$ **Need for regularization**

1. Graphical Lasso (Glasso)

Friedman et al. (2008)

Penalized Likelihood

$$\widehat{K} = \underset{K \succ 0}{\arg\max} \quad \mathcal{L}(K) - \rho \sum_{i \neq j} |K_{ij}|$$

Key features:

- ℓ_1 **penalty** encourages sparsity
- **Uniform penalization**: same ρ for all edges
- **Efficient optimization**: block coordinate descent

Limitation:

- $\times$ Assumes homogeneous sparsity
- $\times$ All edges penalized equally

2. SiMoNe (Ambroise et al., 2008)

Objective: Estimate the precision matrix K and the clustering $Z = (Z_1, \ldots, Z_p)$.

Penalized Likelihood

$$\underset{K \succ 0}{\text{maximize}} \quad \mathcal{L}(K) - \|\rho_{\mathbf{Z}}(K)\|_{\ell_1} + \text{const}$$

Adaptive penalty term:

$$\rho_{\mathbf{Z}_i, \mathbf{Z}_j}(K_{ij}) = \begin{cases} \sum_{q,\ell} Z_{iq} Z_{j\ell} \dfrac{|K_{ij}|}{\lambda_{q\ell}} & \text{if } i \neq j, \\ \dfrac{|K_{ii}|}{\lambda_0} & \text{if } i = j. \end{cases}$$

$\rightsquigarrow$ Optimization performed via a variational EM algorithm.

3. BAGUS (Gan et al., 2019)

Objective: Estimate the precision matrix K and A.

$$A_{ij} \overset{iid}{\sim} \mathcal{B}(\eta),$$
$$K_{ij} \mid A_{ij} \sim (1 - A_{ij})\, g_{\lambda_0} + A_{ij}\, g_{\lambda_1}$$

Penalized Log-Likelihood

$$\mathcal{L}(K) - \sum_{i<j} \text{pen}_{SS}(K_{ij}) - \sum_{i} \tau |K_{ii}|$$

Spike-and-slab penalty:

$$\text{pen}_{SS}(K_{ij}) = -\log\left[\frac{\eta}{2v_1} e^{-\frac{|K_{ij}|}{v_1}} + \frac{1-\eta}{2v_0} e^{-\frac{|K_{ij}|}{v_0}}\right]$$

- $\rightsquigarrow$ Penalization adapts to edge probability via A_{ij}.
- $\rightsquigarrow$ EM algorithm and block coordinate descent are used to infer K.

Our procedure

Limitations of existing approaches:

- Most methods do not provide **guarantees** on the inferred graph structure (e.g., no control over false positives).
- Only SiMoNE handles **clustering**, but with no guarantee.

Our Aim:

- **Graph inference:** Estimate the conditional independence graph A while controlling the False Discovery Rate (FDR), i.e., the proportion of errors among the discovered edges.

- **Clustering:** Recover the latent group structure underlying the variables.

Stochastic Block Model (SBM) for Binary Graphs

- Each node belongs to one of Q latent groups.

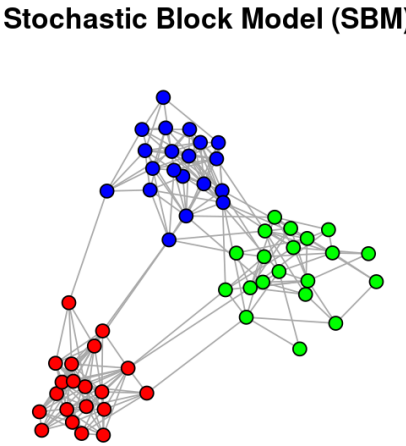
$$Z_1, \ldots, Z_p \overset{i.i.d.}{\sim} \mathcal{M}(1, \pi)$$

- Conditionally on Z, the variables A_{ij} are independent Bernoulli variables with parameters characterized by latent groups:

$$A_{ij} \mid (Z_i = q, Z_j = \ell) \sim \text{Bernoulli}(\omega_{q\ell})$$

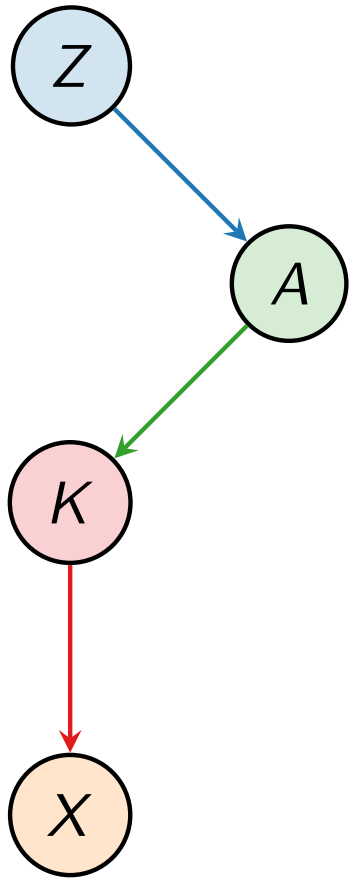
Example:

- $Q = 3$, $\pi = (0.4, 0.3, 0.3)$
- $\omega = \begin{pmatrix} 0.5 & 0.01 & 0.01 \\ 0.01 & 0.4 & 0.05 \\ 0.01 & 0.05 & 0.3 \end{pmatrix}$



Complete Hierarchical Model

- **Group Membership**
  $Z_i \sim \mathcal{M}(1, \pi)$
- **Connections**
  $A_{ij}|Z \sim \mathcal{B}(\omega_{Z_i Z_j})$
- **Precision Matrix**
  $$\begin{cases} K_{ii} \sim \mathcal{E}(\mu) \\ K_{ij}|A_{i,j} \sim (1 - A_{ij})\text{Laplace}(\lambda_0) + A_{ij}\mathcal{N}(0, \sigma^2) \end{cases}$$
- **Observed Data**
  $X|K \sim \mathcal{N}(0, K^{-1})$



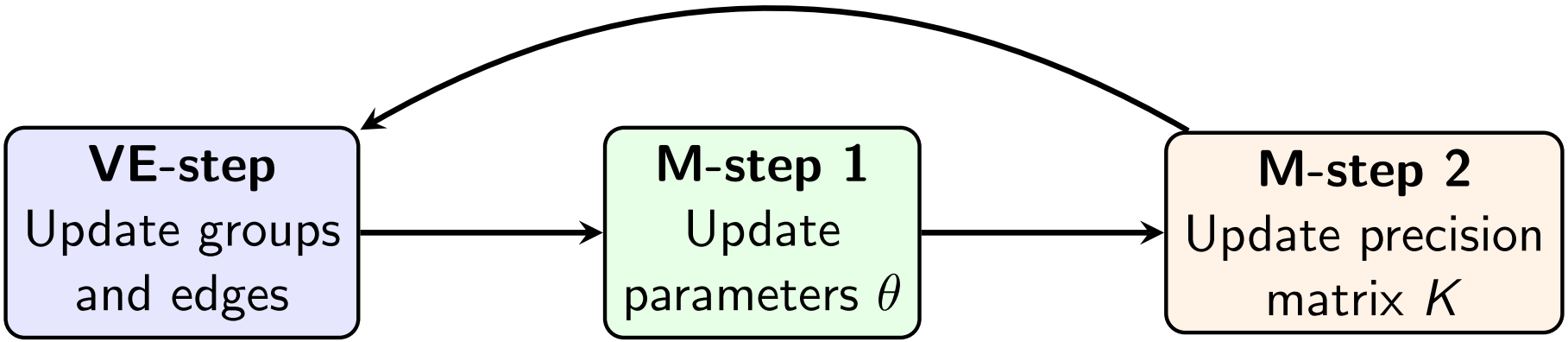
- Latent groups
- Network structure
- Spike-and-slab
- Observed data

Model Inference: Variational EM Approach

Challenge: Complex latent structure

- Precision matrix K
- Model parameters $\theta = (\pi, \omega, \sigma, \lambda_0)$
- Latent modules Z and edges A

Solution: Variational EM algorithm



VE-step: Updating Latent Structure

- **Edge probabilities** $\rho_{q\ell}^{i,j}$

$$\rho_{q\ell}^{i,j} = \mathbb{P}(A_{ij} = 1 \mid K_{i,j}, Z_i = q, Z_j = \ell)$$

Probability of connection given modules and precision

- **Group assignments** $\tau_{i,q}$
Fixed-point equation:

$$\widehat{\tau}_{i,q} = C_i \widehat{\pi}_q \exp\left(\sum_{j \neq i} \sum_{l=1}^{Q} \widehat{\tau}_{j,l} \, \widehat{d}_{q,l}^{i,j} \right)$$

Estimating the Precision Matrix K

Cluster-adaptive penalization

Key idea: Adaptive penalization based on module membership

For each gene i, solve:

$$\widehat{\beta}_i = \arg\min_{\beta} \frac{1}{2n}\|X_i - X_{-i}\beta\|^2 + \sum_{j\neq i}\left[\frac{p_{ij}}{2\sigma^2}\beta_j^2 + \frac{1-p_{ij}}{\lambda_0}|\beta_j|\right]$$

where:

$$p_{ij} = \sum_{q,\ell} \tau_{i,q}\tau_{j,\ell}\rho_{q\ell}^{i,j}$$

(edge probability weighted by module membership)

Interpretation: Higher connection probability $\Longrightarrow$ lower penalty

Graph inference with FDR Control

Edge probability:

$$\ell_{ij} = \mathbb{P}(A_{ij} = 0 \mid \widehat{Z}_i, \widehat{Z}_j, \widehat{K}_{ij}) = \frac{(1 - \widehat{\omega}_{\widehat{Z}_i \widehat{Z}_j})\, g_{\widehat{\lambda}_0}(\widehat{K}_{ij})}{(1 - \widehat{\omega}_{\widehat{Z}_i \widehat{Z}_j})\, g_{\widehat{\lambda}_0}(\widehat{K}_{ij}) + \widehat{\omega}_{\widehat{Z}_i \widehat{Z}_j}\, \phi(\widehat{K}_{ij}; 0, \widehat{\sigma}^2)}.$$

FDR control procedure:

1. Sort ℓ-values: $\ell_{(1)} \leq \ell_{(2)} \leq \cdots \leq \ell_{(m)}$
2. Find largest k such that: $\frac{1}{k} \sum_{r=1}^{k} \ell_{(r)} \leq \alpha$
3. Select edges with $\ell_{ij} \leq \ell_{(k)}$

Output : a final adjacency matrix $\widehat{A}$ with **FDR-controlled support recovery**.

Model Selection: Parameters and Number of Clusters

Parameter Selection (λ_0, σ):

- Test discrete grid of values
- Select combination minimizing BIC:

$$\text{BIC}_{\lambda_0,\sigma} = -2 \cdot \text{pseudoLik}(K) + \log(n) \cdot \|\widehat{K}\|_0$$

where $\|\widehat{K}\|_0$ = number of non-zero edges

Number of Clusters (Q):

- Choose Q minimizing cluster BIC:

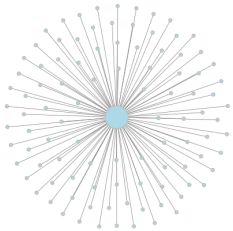
$$\text{BIC}_Q = -2 \cdot \text{pseudoLik}(K) + \underbrace{(Q-1)\log(p) + \frac{Q(Q+1)}{2}\log\left(\frac{p(p-1)}{2}\right)}_{\text{cluster complexity penalty}}$$

Simulations Experiments

Generated Network Structures

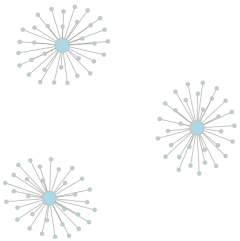
Three distinct topological configurations

- Single central hub



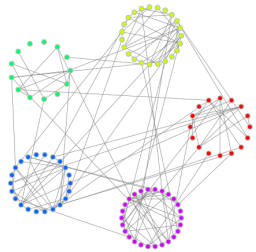
(a) 1-Hub Network

- 3 hubs



(b) 3-Hub Network

- 5 communities
- $w_{\text{in}} = 0.1,\ w_{\text{out}} = 0.01$



(c) SBM 5 Communities

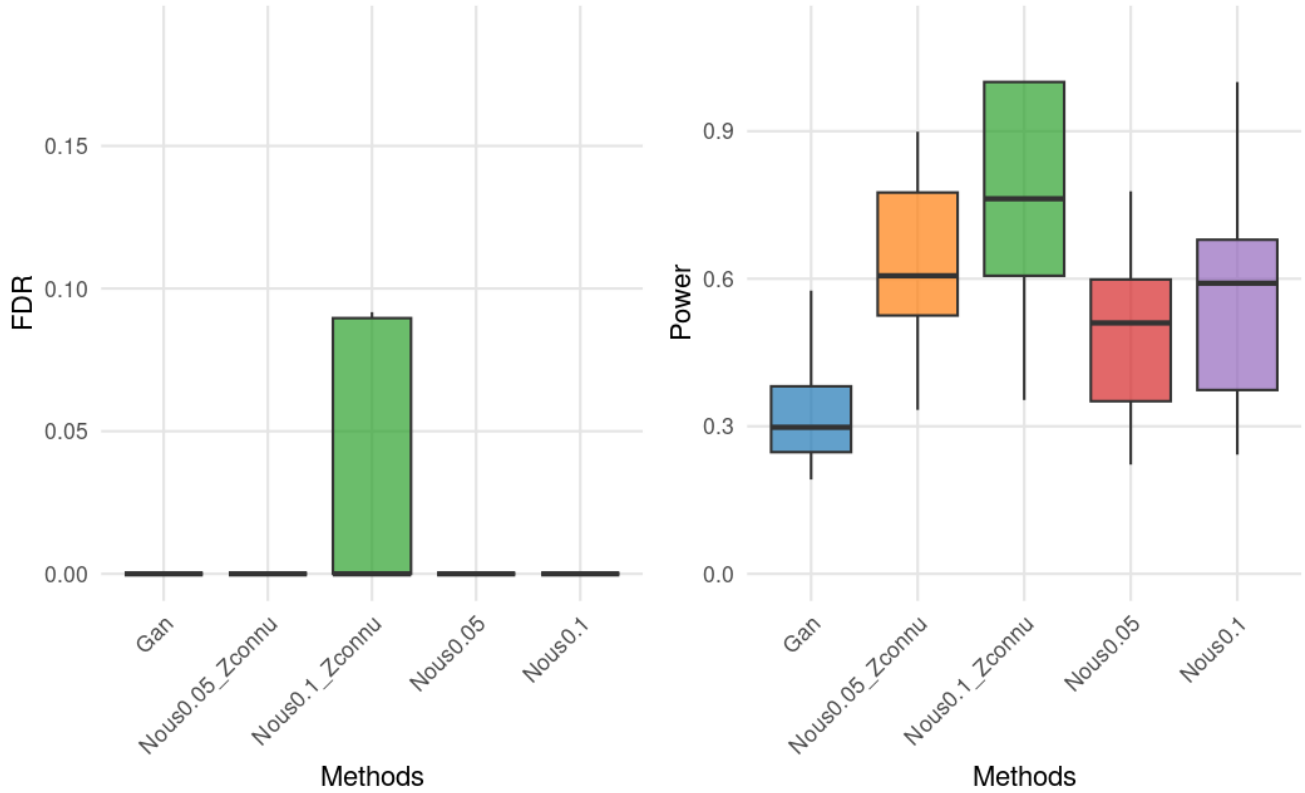
Simulation parameters: $n = 100,\ p = 100$

SBM (5 blocks)

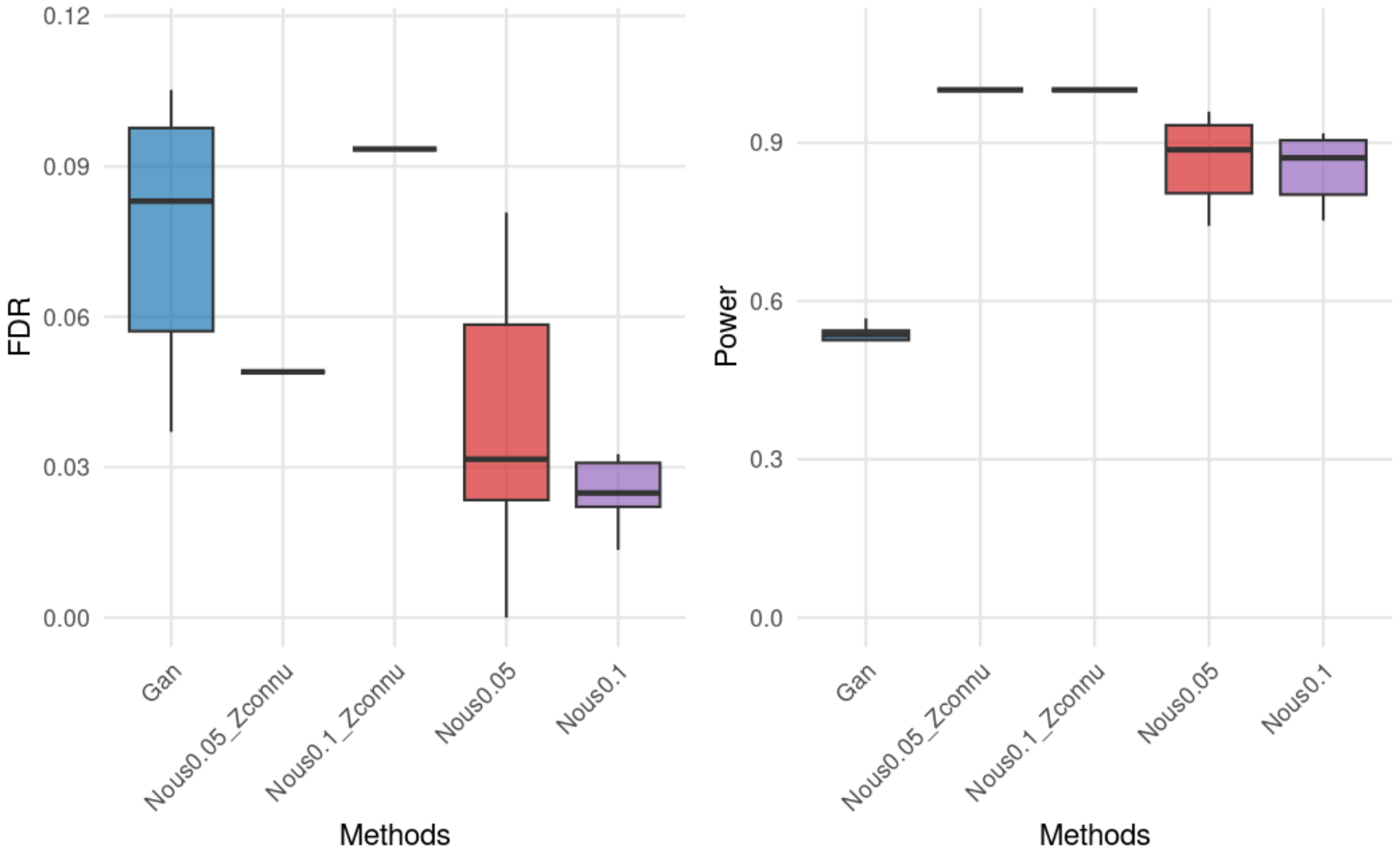
Evaluation metrics:

- **FDR (False Discovery Rate)**
- **Power (True Positive Rate)**

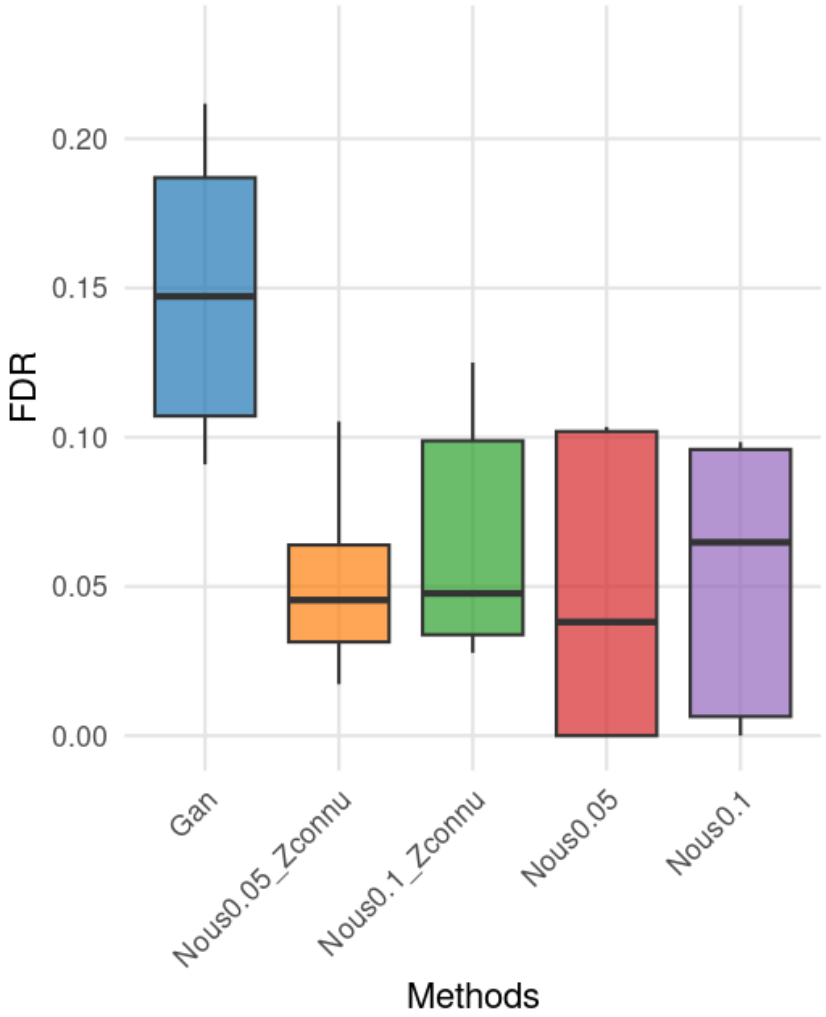
1-hub topology

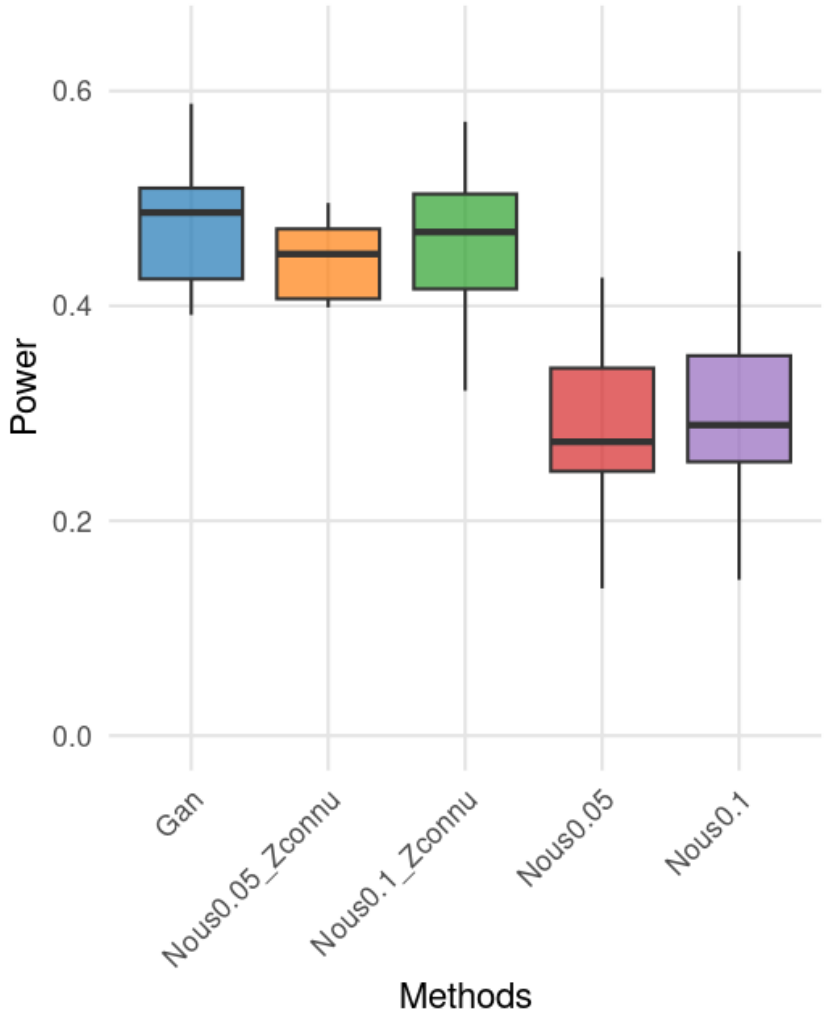


3-hub topology



SBM (5 blocks)


FDR
0.20
0.15
0.10
0.05
0.00
Gan
Nous0.05_Zconnu
Nous0.1_Zconnu
Nous0.05
Nous0.1
Methods


Power
0.6
0.4
0.2
0.0
Gan
Nous0.05_Zconnu
Nous0.1_Zconnu
Nous0.05
Nous0.1
Methods

Take-home messages

- We proposed a method for graph inference in Gaussian Graphical Models (GGMs), with control over the False Discovery Rate (FDR).
- This method also allows inference of a clustering structure among the variables.
- The method still needs to be applied to various datasets, including real-world data.

Références

Ambroise, C., Chiquet, J., and Matias, C. (2008). Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9(15):485–516.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9 3:432–41.

Gan, L., Narisetty, N. N., and Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, 114(527):1218–1231.