Improving Gaussian Graphical Model inference by learning the graph structure

Valentin Kilian [1] Tabea Rebafka [2] Fanny Villers [2]

[1]Department of Statistics, University of Oxford [2]LPSM, Sorbonne University

Motivation

Our goal is to infer the **interaction network** between a set of genes using only the easily available gene expression data.

The uncertainty in measuring means that we can only access a noisy version of the interaction. Our contribution in [1] is the use of latent structure to improve graph inference.

Gaussian Graphical Model (GGM)

Consider $p \in \mathbb{N}$, $p \geq 2$, and a random vector:

$$Y = (Y_1, \dots, Y_p)' \sim \mathcal{N}_p(0, \Sigma)$$

The **GGM** associated with Y is a graphical representation of the **conditional dependence relationships** between the variables.

An edge indicates a non-null **partial correlation** :

$$i \sim j \quad \Leftrightarrow \quad \text{Corr}(Y_i, Y_j | Y_{-(i,j)}) \neq 0 \quad \Leftrightarrow \quad \omega_{ij} \neq 0$$

where $\Omega = \Sigma^{-1} = (\omega_{ij})_{i,j}$.

The R package `SILGGM` provides some **test statistics** for

$$H_{0,i,j} : \omega_{i,j} = 0 \quad \textbf{vs} \quad H_{1,i,j} : \omega_{i,j} \neq 0,$$

We will focus on one of them introduced in [2].

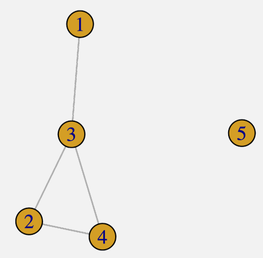


Figure: Example of a GGM

Objective: Detect graph edges based on an n-sample $\boldsymbol{Y}$ of $(Y_1, \dots, Y_p)'$ while controlling the proportion of false discoveries.

Noisy Stochastic Block Model (NSBM) [3]

- The number of nodes, $p \geq 2$. The number of latent groups, $Q \in \{1, \dots, p\}$.
- The block memberships of nodes $Z = (Z_1, \dots, Z_p)$, with $Z_i \overset{iid}{\sim} \pi$.
- Latent graph structure : for some parameter $w = (w_{kl})_{k,l} \in \mathcal{S}_Q([0,1])$,

$$A_{i,j} \mid Z \overset{cond.\ iid}{\sim} \mathcal{B}ern(w_{Z_i,Z_j}).$$

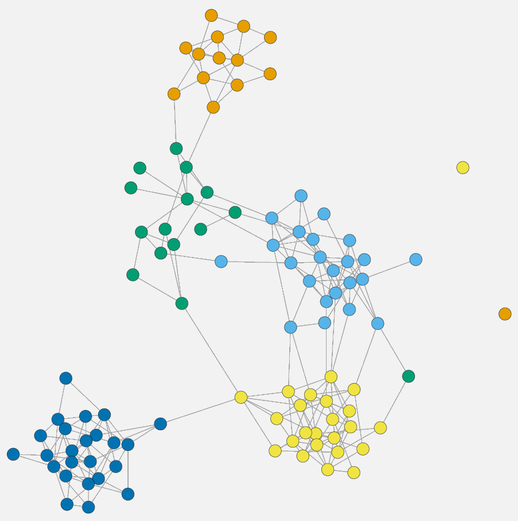


Figure: Example of a SBM with 5 groups and 50 nodes

- Observed variables : for some parameters $\mu,\ \sigma \in \mathcal{S}_Q(\mathbb{R})$ and $\sigma_0 \in \mathbb{R}$,

$$X_{i,j} \mid Z, A \sim (1 - A_{i,j})\mathcal{N}(0, \sigma_0^2) + A_{i,j}\mathcal{N}(\mu_{Z_i,Z_j}, \sigma^2_{Z_i,Z_j}).$$

The unknown global model parameter is then

$$\theta = (\pi, w, \mu, \sigma).$$

The observation is X, while both Z and A are unobserved and latent variables of the model.

Estimation in the NSBM: Greedy Algorithm (1/2)

The algorithm to estimate θ and Z, inspired by [4], operates as follows:

1. Start with an initial partition of nodes into Q_{up} groups Z.

2. Evaluate, for each node, whether it's beneficial to reassign it to a different group. To determine this, we efficiently compute the change in the **integrated complete-data log likelihood** ICL_{ex} for each potential group swap:

Estimation in the NSBM: Greedy Algorithm (2/2)

$$ICL_{ex}(Z, A) = \log\left(\int_{\pi,w,\mu,\sigma} p(X, A, Z, \pi, w, \mu, \sigma) d(\pi, w, \mu, \sigma)\right)$$

Use of conjugate priors for π, w, μ, σ.

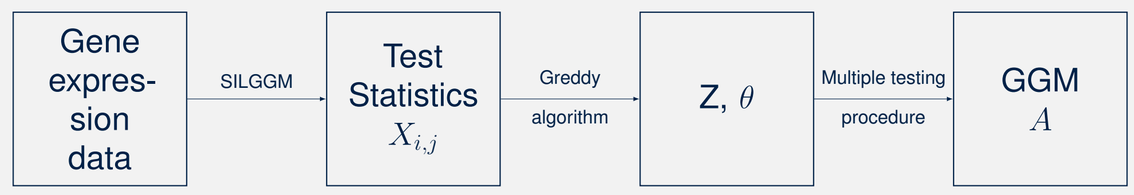
3. Whenever a node changes its group, both Z and the estimate of θ are updated. During this process, some groups may become empty.

4. After steps 2 and 3 have been completed for all nodes, we check if it's advantageous to merge certain groups to increase the ICL.

This results in a node clustering Z, an estimate of the number of latent groups Q, and an estimation of the model parameter θ.

Then, we apply a **multiple testing procedure** based on l-values to infer the graph while controlling the False Discovery Rate (FDR).

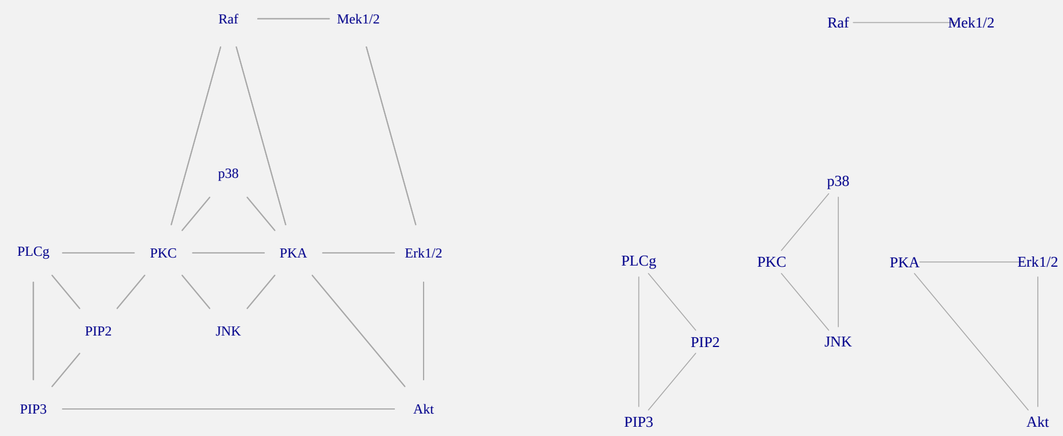
Strategy



The simulations using synthetic data demonstrate that our method outperforms the global procedures proposed in `SILGGM` and several other classic methods.

Human T cell

We applied our procedure to Sachs et al.'s data [5], that have been **extensively studied in the literature**. The dataset includes $p = 11$ protein measurements from 902 cells.



(a) Established graph

(b) Graph inferred with the full data set, level $\alpha = 0.05$.

Figure: Our inferred graph contains ten edges, nine of which are well-established in the literature. The last edge, $p38 - JNK$, was also detected by Sachs with low confidence and by other statisticians. This graph serves as a benchmark for full dataset inference.

To assess our method's ability to recover these edges with a smaller dataset, we randomly sample subsets.

	n=10			n=20		
Edge	LiuL	LiuL NSBM	LiuL NIG	LiuL	LiuL NSBM	LiuL NIG
Raf - Mek1/2	183	191	192	200	200	200
PLCg - PIP2	15	32	30	30	44	43
PLCg - PIP3	70	95	93	107	134	133
PIP2 - PIP3	119	140	147	168	176	176
Erk1/2 - Akt	178	180	187	197	198	198
Akt - PKA	59	85	88	118	136	139
...	...	...	...	...	...	...

Table: Over 200 simulations, we counted how often the 10 edges were detected when the procedures were applied to randomly chosen subsets of either $n = 10$ or $n = 20$ observations.

Our procedure detects all ten edges more frequently. **This confirms the efficacy of our procedure in improving GGM inference.**

References

[1] V. Kilian, T. Rebafka, and F. Villers, "**Improving Gaussian Graphical Model inference by learning the graph structure**," *ArXiv available soon*, 2023.

[2] W. Liu, "Gaussian graphical model estimation with false discovery rate control," *The Annals of Statistics*, vol. 41, no. 6, pp. 2948 – 2978, 2013.

[3] T. Rebafka, É. Roquain, and F. Villers, "Powerful multiple testing of paired null hypotheses using a latent graph model," *Electronic Journal of Statistics*, vol. 16, no. 1, pp. 2796 – 2858, 2022.

[4] E. Côme and P. Latouche, "Model selection and clustering in stochastic block models with the exact integrated complete data likelihood," *Statistical Modelling*, vol. 15, 03 2013.

[5] K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science (New York, N.Y.)*, vol. 308, pp. 523–9, 05 2005.