Improving Gaussian Graphical Model inference by modeling the graph structure

Valentin Kilian, Tabea Rebafka, Fanny Villers

Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université, Paris

Netbio, 15 novembre 2023

Outline

Data

Biological data :

- gene expression data
- or quantitative amounts of proteins

p = number of entities (genes, proteins)
n = number of repeating observations

Aim : infer the direct links between entities $\Leftrightarrow$ infer a graph:
- nodes = entities (genes, proteins)
- edge = direct relation between two entities

- regulations between genes
- protein-protein interactions

Gaussian Graphical model (GGM)

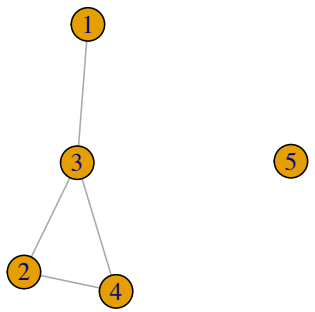
random variables $Y_1, \ldots, Y_p$: expression of the p genes or proteins

Assumption GGM : $(Y_1, \ldots Y_p) \sim \mathcal{N}(0, \Sigma)$

Direct links

Denote $\Omega = \Sigma^{-1} = (w_{ij})_{1 \leq i,j \leq p}$: precision matrix

$$i \sim j \text{ (edge between } i \text{ and } j) \Leftrightarrow \text{corr}(Y_i, Y_j | (Y_k)_{k \neq i,j}) \neq 0$$
$$\Leftrightarrow \omega_{ij} \neq 0$$



Graph inference in GGM

Inference of the graph edges based on a n-sample of $(Y_1, \dots Y_p)$
High-dimensional setting : $p \gg n$

Literature:

- infer the precision matrix Ω (glasso)
- infer the neighboors of each node (Meinshausen Bühlmann)
- multiple-testing approach $H_{0,ij} : w_{ij} = 0$ against $H_{1,ij} : w_{ij} \neq 0$

Graph inference in GGM

Inference of the graph edges based on a n-sample of $(Y_1, \dots Y_p)$
High-dimensional setting : $p \gg n$

Literature:

- infer the precision matrix Ω (glasso)
- infer the neighboors of each node (Meinshausen Bühlmann)
- multiple-testing approach $H_{0,ij} : w_{ij} = 0$ against $H_{1,ij} : w_{ij} \neq 0$

Inference is difficult:

- lack of power
- graph inferred can be different according to the method
- in general, no control on the inferred graph

Multiple-testing approach

$$H_{0,ij} : \underbrace{w_{ij} = 0}_{i \nsim j} \quad \text{against} \quad H_{1,ij} : \underbrace{w_{ij} \neq 0}_{i \sim j}$$

Multiple-testing approach

$$H_{0,ij} : \underbrace{w_{ij} = 0}_{i \nsim j} \quad \text{against} \quad H_{1,ij} : \underbrace{w_{ij} \neq 0}_{i \sim j}$$

Test statistics ?

- if $p \ll n$: natural test statistics based on the inverse of the sample covariance matrix $\widehat{\Sigma}$

Multiple-testing approach

$$H_{0,ij} : \underbrace{w_{ij} = 0}_{i \nsim j} \quad \text{against} \quad H_{1,ij} : \underbrace{w_{ij} \neq 0}_{i \sim j}$$

Test statistics ?

- if $p \ll n$: natural test statistics based on the inverse of the sample covariance matrix $\widehat{\Sigma}$
- in high-dimensional setting :
  Ref: Liu et al 2013, Ren et al 2015, Jankova et al 2018
  - estimators for the entries of the precision matrix w_{ij}
  - based on different modifications of initial Lasso-regularized estimators
  - proved to be asymptotically normal a sparcity condition
  - enables the construction of test statistics to test $H_{0,ij}$

Multiple-testing approach

$$H_{0,ij} : \underbrace{w_{ij} = 0}_{i \nsim j} \quad \text{against} \quad H_{1,ij} : \underbrace{w_{ij} \neq 0}_{i \sim j}$$

Test statistics ?

- if $p \ll n$: natural test statistics based on the inverse of the sample covariance matrix $\widehat{\Sigma}$
- in high-dimensional setting :
  Ref: Liu et al 2013, Ren et al 2015, Jankova et al 2018
  - estimators for the entries of the precision matrix w_{ij}
  - based on different modifications of initial Lasso-regularized estimators
  - proved to be asymptotically normal a sparcity condition
  - enables the construction of test statistics to test $H_{0,ij}$

Simultaneous tests: test $H_{0,ij}$ for all pairs of variables (i, j).
$\hookrightarrow$ multiple testing problem

Aim

Inference of the graph : detect significant edges

- with control on the inferred graph in term of False Discovery Rate (FDR: proportion of errors among the discovered edges)
  - (Bonferroni)
  - Benjamini and Hochberg
  - Liu et al 2013: asymtotic FDR control under sparcity assumption

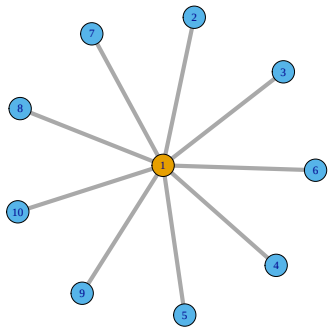
Aim

Inference of the graph : detect significant edges

- with control on the inferred graph in term of False Discovery Rate (FDR: proportion of errors among the discovered edges)
  - (Bonferroni)
  - Benjamini and Hochberg
  - Liu et al 2013: asymtotic FDR control under sparcity assumption
- with high ability to detect true edges
  - multiple testing literature : *Ref: Efron & al, 2001, Efron, 2004, Sun & Cai, 2007, Cai & Sun, 2009, Sun & Cai, 2009*
  - incorporating some latent dependence structure may allow more detections

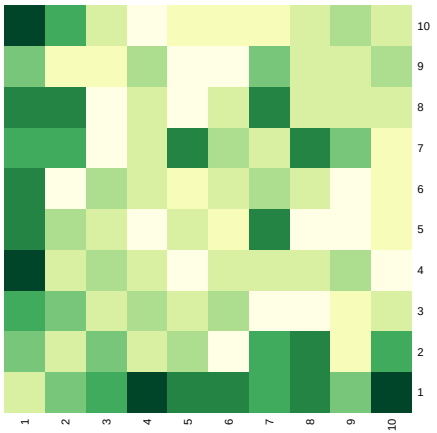
incorporating some latent structure ?

Graph to infer



from

Matrice with test statistics for each pairs of variables (i, j)



- learning the graph structure (nodes clustering)
- incorporating it in the multiple-testing procedure

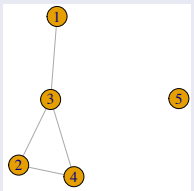
learning the graph structure ?

$\hookrightarrow$ modeling the graph structure through the adjacency matrix A

Adjacency matrix A of a graph

$A = (A_{ij})_{1 \leq i,j \leq p}$ with

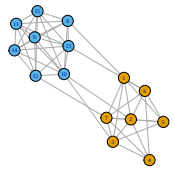
$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between i and j} : i \sim j \\ 0 & \text{otherwise} : i \nsim j \end{cases}$$



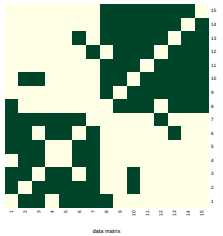
$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$\hookrightarrow$ random graph model on A

- random graph model on A : stochastic block model SBM

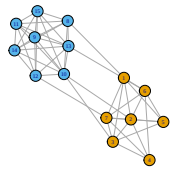


$\Longleftrightarrow$

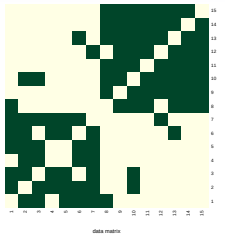


Graph to infer $A \in \{0,1\}^{p \times p}$

- random graph model on A : stochastic block model SBM



$\Longleftrightarrow$



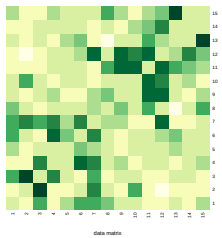
Graph to infer $A \in \{0,1\}^{p \times p}$

- Or A is unknown $\rightarrow$ NSBM model : Noisy SBM

$X =$ Noisy version of A

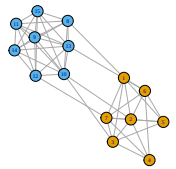
Observed: X : (p, p) matrix

with X_{ij}: test statistic

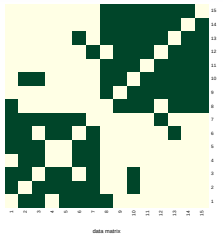


$X \in \mathbb{R}^{p \times p}$

- random graph model on A : stochastic block model SBM



$\Longleftrightarrow$



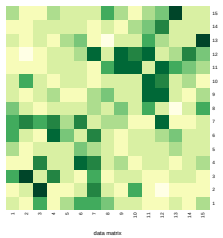
Graph to infer $A \in \{0,1\}^{p \times p}$

- Or A is unknown $\rightarrow$ NSBM model : Noisy SBM

$X =$ Noisy version of A

Observed: X : (p, p) matrix

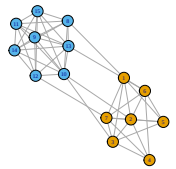
with X_{ij}: test statistic



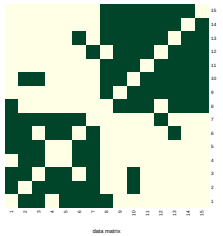
$X \in \mathbb{R}^{p \times p}$

- Estimation of the parameters of the model (nodes clustering)

- random graph model on A : stochastic block model SBM



$\Longleftrightarrow$



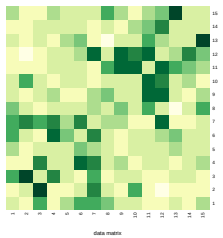
Graph to infer $A \in \{0,1\}^{p \times p}$

- Or A is unknown $\rightarrow$ NSBM model : Noisy SBM

$X =$ Noisy version of A

Observed: X : (p,p) matrix

with X_{ij}: test statistic



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

- Estimation of the parameters of the model (nodes clustering)
- Multiple-testing procedure incorporating the estimated parameters

$$H_{0,ij} : \underbrace{A_{ij} = 0}_{i \nsim j} \quad \text{against} \quad H_{1,ij} : \underbrace{A_{ij} = 1}_{i \sim j}$$

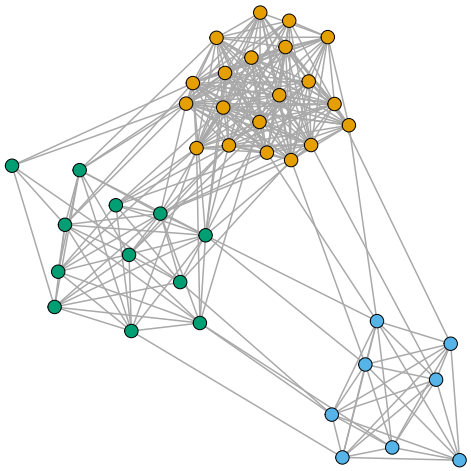
Outline

SBM

Stochastic Block Model - SBM

- Each node belongs to one of Q latent groups.
  Latent variables $Z_1, \ldots, Z_p$ i.i.d. with values $\{1, \ldots, Q\}$ and probability $\pi_q = \mathbb{P}(Z_1 = q)$
- Conditionally on Z, the variables A_{ij} are independent Bernoulli variables with parameters characterized by latent groups :

$$A_{ij}|(Z_i = q, Z_j = l) \sim Bernoulli(\gamma_{q,l})$$



Model

Noisy Stochastic Block Model - NSBM

NSBM

- The true underlying binary graph A is a SBM
  - with Q groups
  - connectivity parameters $\gamma = (\gamma_{q,l})_{1 \leq q,l \leq Q}$
  - group proportions $\pi = (\pi_q)_{1 \leq q \leq Q}$
  - latent variables $Z_i \in \{1, \ldots, Q\}$ for $i = 1, \ldots, p$
- Conditionally on A and Z, the observations X_{ij} are independent with

$$X_{ij}|Z, A \sim \begin{cases} \mathcal{N}(0, \sigma_0^2) & \text{if } A_{i,j} = 0 \quad (\text{if } i \nsim j) \\ \mathcal{N}(\mu_{ql}, \sigma_{ql}^2) & \text{if } A_{i,j} = 1 \quad (\text{if } i \sim j), Z_i = q, Z_j = l \end{cases}$$

NSBM model

Mixture model :

Observations : $X = (X_{ij})_{1 \leq i,j \leq p}$

Latent variables : Z, A

Unknown parameters : $\theta = (\pi, \gamma, \mu, \sigma)$

with $\pi = (\pi_q), \gamma = (\gamma_{ql}), \mu = (\mu_{ql}), \sigma = (\sigma_{ql})$ $q, l \in \{1, \dots Q\}$

we suppose that σ_0 is known ($\sigma_0 = 1$)

NSBM model

Mixture model :

Observations : $X = (X_{ij})_{1 \leq i,j \leq p}$

Latent variables : Z, A

Unknown parameters : $\theta = (\pi, \gamma, \mu, \sigma)$

with $\pi = (\pi_q), \gamma = (\gamma_{ql}), \mu = (\mu_{ql}), \sigma = (\sigma_{ql})$ $q, l \in \{1, \dots Q\}$

we suppose that σ_0 is known ($\sigma_0 = 1$)

- Estimate the parameters θ and make clustering (recover the latent groups = estimate Z)
- Estimate $A \in \{0,1\}^{pxp} \Leftrightarrow$ infer the graph G by using $\hat{\theta}$ and $\hat{Z}$

Multiple testing :

$$H_{0,ij} : \underbrace{A_{ij} = 0}_{i \nsim j} \quad \text{against} \quad H_{1,ij} : \underbrace{A_{ij} = 1}_{i \sim j}$$

Estimation and clustering

NSBM = mixture model with latent variables $\rightarrow$ MLE can not be computed

- Variational Expectation Maximization (VEM) algorithm to estimate $\hat{\theta}$
  + MAP rule to estimate Z
  + model selection to select the number of groups Q

- ICL_{ex} : Integrated complete-data log likelihood
  baysesian framework
  greedy algorithm for optimization in Z
  automatic estimation of the number of groups Q

Estimation and clustering

ref: Côme and Latouche, 2015 in SBM model

- Start from a initial partition of the nodes in Q_{up} groups (Q_{up} large)
- For each node : move the node from its group to another group ?
- Criteria : integrated complete-data log likelihood ICL_{ex}
- Some groups become empty
- At the end, we obtain a clustering of the nodes $\hat{Z}$ and an estimation of the number of groups $\hat{Q}$

Estimation and clustering

Integrated complete-data log likelihood ICL_{ex}:

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

- Bayesian framework
- all the parameters in $\theta = (\pi, \gamma, \mu, \sigma)$ are integrated out
- conjugate priors for π, γ, μ, σ

$\Rightarrow$ analytical expression of ICL_{ex}, which involves the number of nodes in group q, the number of edges between groups q and l...

Estimation and clustering

Greedy Algorithm:

- For each node i^*, we evaluate the variation $\Delta_{g \to h}$ of ICL_{ex} if i^* moves from its group g to a new group h.
- $\Delta_{g \to h}$ can be evaluated in a computationally efficient way

Estimation and clustering

Greedy Algorithm:

- For each node i^*, we evaluate the variation $\Delta_{g \to h}$ of ICL_{ex} if i^* moves from its group g to a new group h.
- $\Delta_{g \to h}$ can be evaluated in a computationally efficient way
- Difference with Côme and Latouche in the SBM : A is latent
  $\hookrightarrow$ *we estimate the posterior probability that there is an edge between i and j*
  $\hookrightarrow$ *depends on Z and θ that are estimated at each step of the algorithm*
  $\hookrightarrow$ *estimator of θ have the form of traditional ML estimators with weighted means*

Estimation and clustering

Greedy Algorithm:

- For each node i^*, we evaluate the variation $\Delta_{g \to h}$ of ICL_{ex} if i^* moves from its group g to a new group h.
- $\Delta_{g \to h}$ can be evaluated in a computationally efficient way
- Difference with Côme and Latouche in the SBM : A is latent
  $\hookrightarrow$ *we estimate the posterior probability that there is an edge between i and j*
  $\hookrightarrow$ *depends on Z and θ that are estimated at each step of the algorithm*
  $\hookrightarrow$ *estimator of θ have the form of traditional ML estimators with weighted means*
- At the end : merge groups ?

Estimation and clustering

Greedy Algorithm:

- For each node i^*, we evaluate the variation $\Delta_{g \to h}$ of ICL_{ex} if i^* moves from its group g to a new group h.
- $\Delta_{g \to h}$ can be evaluated in a computationally efficient way
- Difference with Côme and Latouche in the SBM : A is latent
  $\hookrightarrow$ *we estimate the posterior probability that there is an edge between i and j*
  $\hookrightarrow$ *depends on Z and θ that are estimated at each step of the algorithm*
  $\hookrightarrow$ *estimator of θ have the form of traditional ML estimators with weighted means*
- At the end : merge groups ?

Output : node clustering $\hat{Z}$, number of groups $\hat{Q}$, estimator $\hat{\theta}$

Graph inference

Aim : infer the adjacency matrix $A \in \{0,1\}^{pxp} \Leftrightarrow$ infer graph edges

Simultaneous test of : $H_{0,ij} : \underbrace{A_{ij} = 0}_{i \nsim j}$ against $H_{1,ij} : \underbrace{A_{ij} = 1}_{i \sim j}$

Graph inference

Aim : infer the adjacency matrix $A \in \{0,1\}^{pxp} \Leftrightarrow$ infer graph edges

Simultaneous test of : $H_{0,ij} : \underbrace{A_{ij} = 0}_{i \nsim j}$ against $H_{1,ij} : \underbrace{A_{ij} = 1}_{i \sim j}$

ℓ-values. (also called the local FDR. Efron, 2001)

$$\ell_{ij}(X, Z; \theta) = \mathbb{P}_\theta(A_{ij} = 0 \mid X, Z)$$

- $\ell_{ij}(X, Z; \theta)$ calculated in the NSBM with Bayes formula
- Reject $H_{0,ij}$ when $\ell_{ij}(X, Z; \theta) \leq t$
- Control of the **FDR** : proportion of errors among the discovered edges
  $\hookrightarrow$ threshold t such that the **FDR** is controlled at level α.

Graph inference

- Threshold t such that the **MFDR** is controlled at level α.

$$\text{MFDR}_\theta(t) = \frac{\mathbb{E}[\text{nb of falsely detected edges}]}{\mathbb{E}[\text{nb of detected edges}]}$$

$\text{MFDR}_\theta(t)$ explicitly calculated

- Choose largest threshold t such that $\text{MFDR}_\theta(t) \leq \alpha$
- $t = t_\theta(\alpha)$ generalized inverse of MFDR_θ en α.

Graph inference

- Threshold t such that the **MFDR** is controlled at level α.

$$\text{MFDR}_\theta(t) = \frac{\mathbb{E}[\text{nb of falsely detected edges}]}{\mathbb{E}[\text{nb of detected edges}]}$$

$\text{MFDR}_\theta(t)$ explicitly calculated

- Choose largest threshold t such that $\text{MFDR}_\theta(t) \leq \alpha$
- $t = t_\theta(\alpha)$ generalized inverse of MFDR_θ en α.

qvalues. (Storey, 2003)

$$q_{ij}(X, Z; \theta) = \text{MFDR}_\theta(\ell_{ij}(X, Z; \theta))$$

- Decision rule : Reject $H_{0,ij}$ provided that

$$q_{ij}(X, Z; \theta) \leq \alpha$$

Algorithm 1: Estimation and Graph inference in NSBM

Input: X, level α

Apply greedy algorithm to get $\widehat{\theta}$ and $\widehat{Z}$

Compute the q-values $q_{ij}(X, \widehat{Z}, \widehat{\theta})$

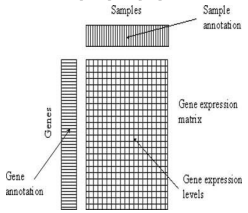
Output: Infer a graph

$$\widehat{A_{ij}} = \mathbb{1}\{q_{ij}(X, \widehat{Z}, \widehat{\theta}) \leq \alpha\}$$

Outline

Our Procedure

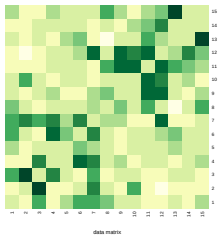
Gene expression data



p genes
n samples

SILGGM package

Test Statistics X_{ij}



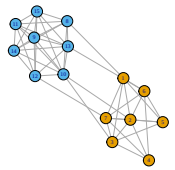
$X \in \mathbb{R}^{pxp}$

Algorithm 1

parameters: $\hat{\theta}$

clustering: $\hat{Z}$

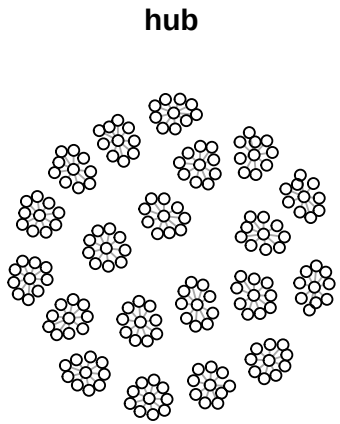
graph: $\hat{A} \Leftrightarrow \hat{G}$

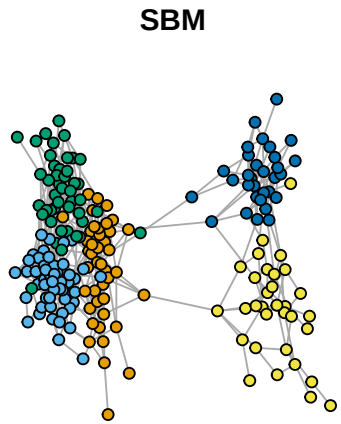


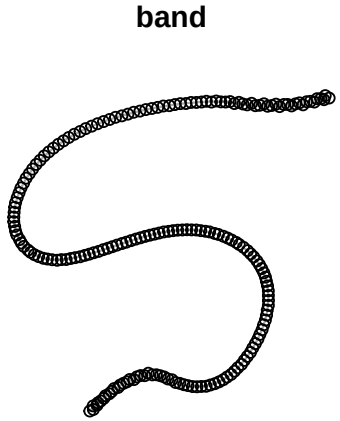
Outline

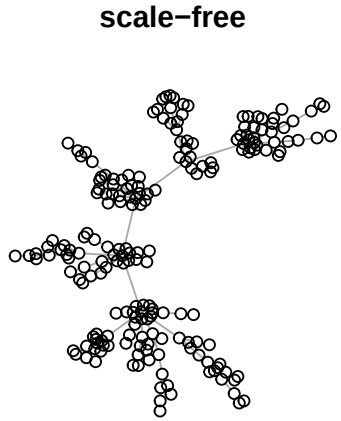
Simulations

- Different graph structures:









Simulations

- Different GGM inference methods:
  - test statistics provided by the package SILGGM : without and with our procedure
  - Glasso procedure
  - Meinshausen and Bühlmann procedure

Simulations

- Different GGM inference methods:
  - test statistics provided by the package SILGGM : without and with our procedure
  - Glasso procedure
  - Meinshausen and Bühlmann procedure

- Estimation of the FDP and the power with 200 Simulations

FDP = proportion of errors among the edges declared significant

TDP (power)= the proportion of edges declared significant among the true edges

Simulations

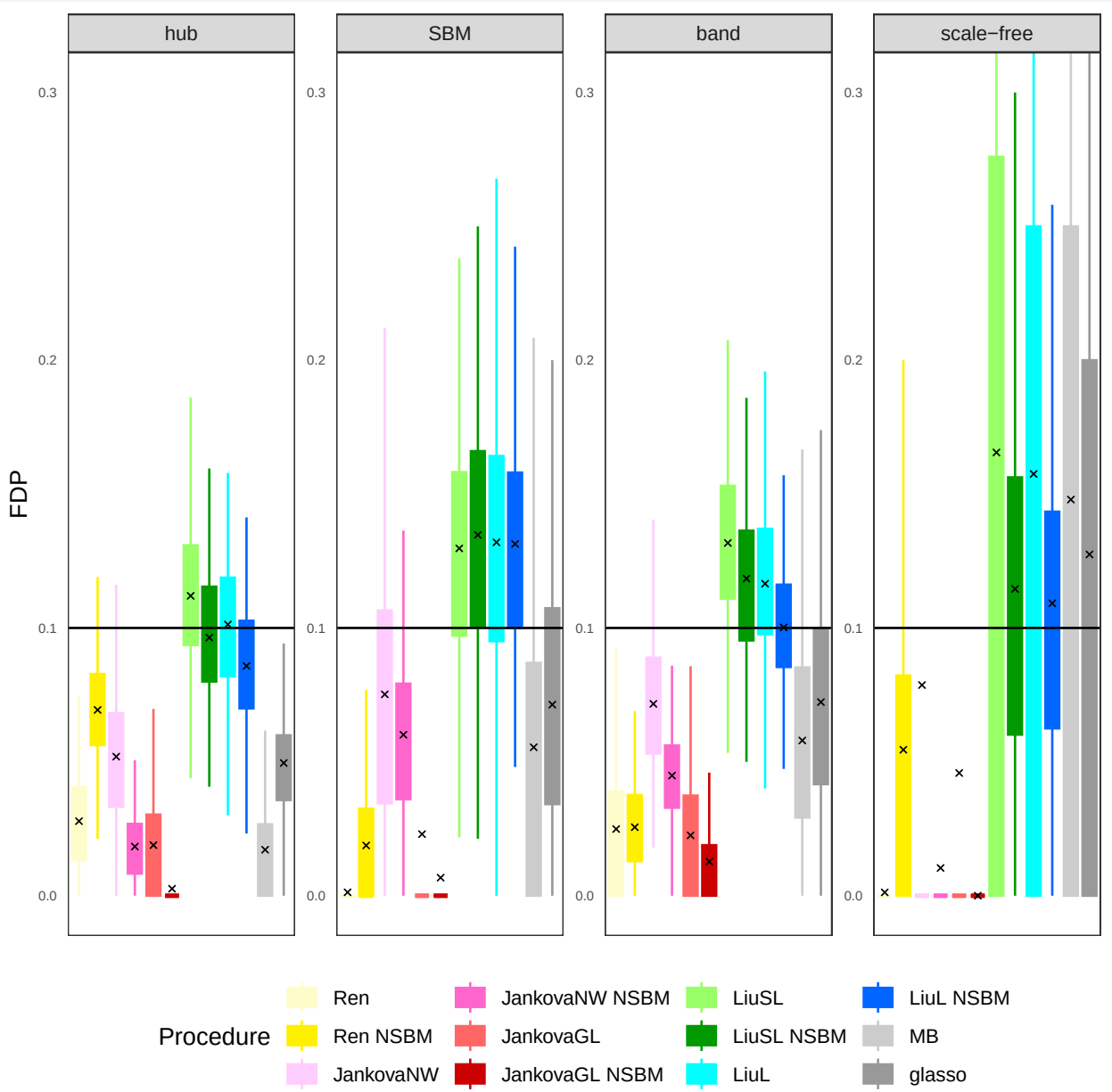
- Different GGM inference methods:
  - test statistics provided by the package SILGGM : without and with our procedure
  - Glasso procedure
  - Meinshausen and Bühlmann procedure

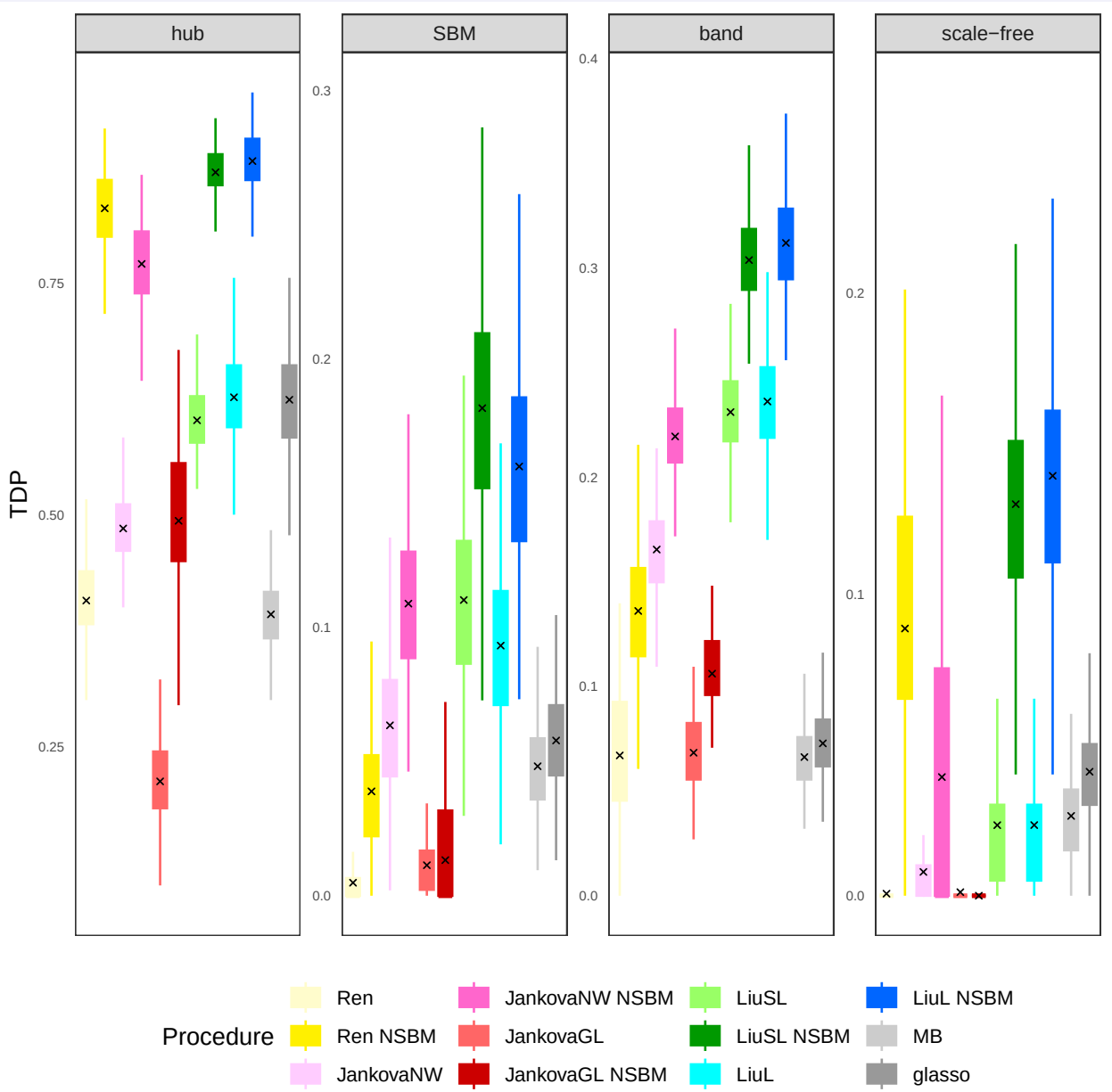
- Estimation of the FDP and the power with 200 Simulations

  FDP = proportion of errors among the edges declared significant

  TDP (power)= the proportion of edges declared significant among the true edges

- $n = 100, p = 200, \alpha = 0.1$

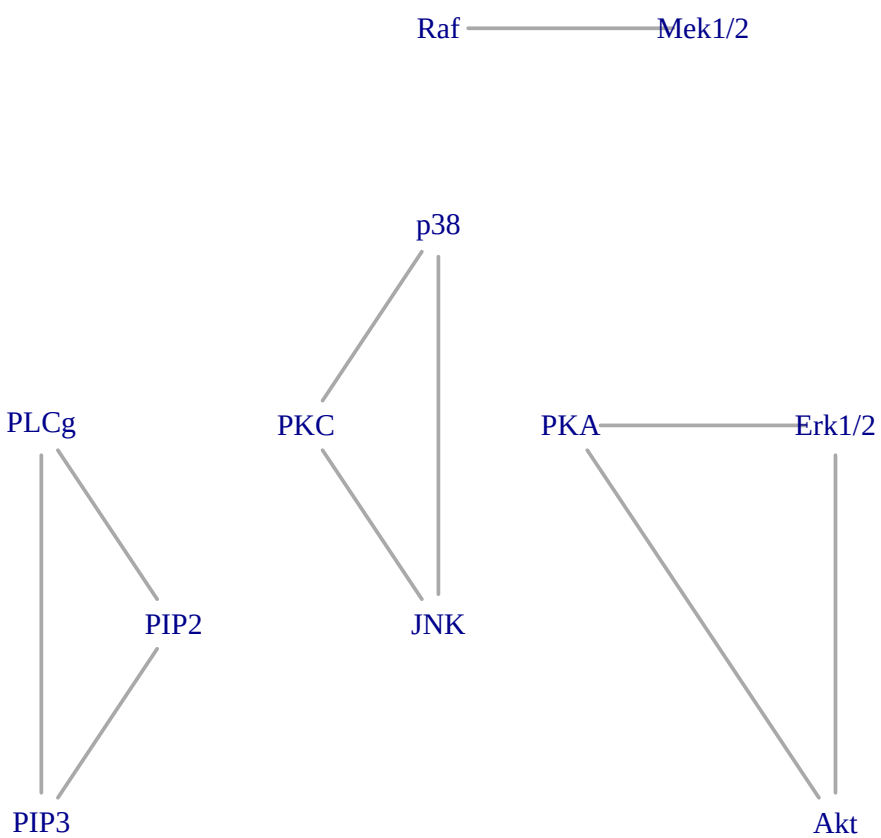

hub
SBM
band
scale−free
FDP
0.0
0.1
0.2
0.3
Procedure
Ren
Ren NSBM
JankovaNW
JankovaNW NSBM
JankovaGL
JankovaGL NSBM
LiuSL
LiuSL NSBM
LiuL
LiuL NSBM
MB
glasso


hub
SBM
band
scale−free
TDP
0.75
0.50
0.25
0.3
0.2
0.1
0.0
0.4
0.3
0.2
0.1
0.0
0.2
0.1
0.0
Procedure
Ren
Ren NSBM
JankovaNW
JankovaNW NSBM
JankovaGL
JankovaGL NSBM
LiuSL
LiuSL NSBM
LiuL
LiuL NSBM
MB
glasso

Real data

- flow cytometry data produced by Sachs et al.
- quantitative amounts of 11 proteins measured in 902 cells.

  - Inference with the full dataset (*LiuL*'s test statistics, $\alpha = 0.05$)



Real data

- Subsampling to test performance of our procedure

	n=10	
edge	LiuL	LiuL NSBM
Raf - Mek1/2	178	187
PLCg - PIP2	18	39
PLCg - PIP3	57	94
PIP2 - PIP3	114	147
Erk1/2 - Akt	178	185
Erk1/2 - PKA	14	43
Akt - PKA	44	79
PKC - p38	95	117
PKC - JNK	69	96
p38 - JNK	70	100

Number of times the 10 edges are detected over 200 simulations

Take-home message

- Inference in the NSBM :
  - faster alternative to the VEM algorithm
  - automatic selection of the number of groups
- Application to graph inference in GGM
  - use test statistics proposed in the literature on GGM as entries of our procedure
- Simulations
  - almost control in term of FDR on the inferred graph
  - increase in power
- Real dataset ?

References

- Côme, Latouche. Model selection and clustering in stochastic block models with the exact integrated complete data likelihood. Statistical Modelling 2015
- Rebafka, Roquain, Villers. Powerful multiple testing of paired null hypotheses using a latent graph model. EJS 2022
- Liu. Gaussian graphical model estimation with false discovery rate control. AoS 2013
- Ren, Sun, Zhang, Zhou. Asymptotic normality and optimalities in estimation of large gaussian graphical models. Aos 2015
- Jankova, Van de Geer. Inference in high-dimensional graphical models. Handbook of Graphical Models 2018.
- Sachs, Perez,Pe'er, Lauffenburger, Nolan. Causal protein signaling networks derived from multiparameter single-cell data. Science 2005