Sparse network inference in presence of missing variables

Geneviève Robin Stéphane Robin Christophe Ambroise

UMR 518 AgroParisTech/INRA MIA
Équipe Statistique et génome

October 11, 2016

1. Gaussian Graphical Models (GGM)
2. GGM with missing variables
3. Inference
4. EM with aggregation of spanning trees
5. Experiments - simulations and flow cytometry data

Outline

Motivations

Biological problem : Gene/species networks describing direct actions between genes/species

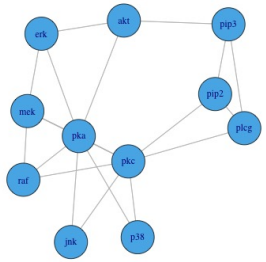
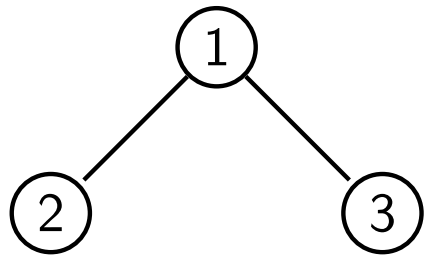


Figure: Raf network

Statistical problem : Independent observations $X^{(i)}$, $1 \leq i \leq n$. Inference of the conditional independence network.

$$\begin{pmatrix} X_1^{(1)} & X_1^{(2)} & X_1^{(3)} & X_1^{(4)} \\ X_2^{(1)} & X_2^{(2)} & X_2^{(3)} & X_2^{(4)} \\ X_3^{(1)} & X_3^{(2)} & X_3^{(3)} & X_3^{(4)} \end{pmatrix}$$

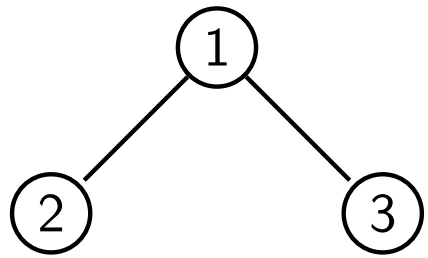


Gaussian graphical models

- $X^{(i)}$ Gaussian random vector of dimension p with parameters $\mu = 0$ and Σ.

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 1.5 & -0.5 \\ -1 & -0.5 & 1.5 \end{pmatrix}$$

$$K = \Sigma^{-1} = \begin{pmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix}, \qquad \mathcal{G} =$$



- Underlying graph $\mathcal{G} = (V, E)$, $V = \{1, \ldots, p\}$
- The edge $\{i, j\}$ is in E if $K_{ij} \neq 0$

Inferring $\mathcal{G} \Leftrightarrow$ inferring the support of K.

Inference of K

- Estimate K from data
- Maximum likelihood estimator:

$$\begin{aligned} \hat{K}^{MLE} &= \arg\max_{K} \log\det(K) - \mathrm{tr}(K\Sigma_n) \\ &= \Sigma_n^{-1} \end{aligned} \tag{1}$$

- Impossible to span all possible graphs: $2^{\frac{p(p-1)}{2}}$ of size p.
- Hypothesis on the structure of the support of K.

Inference methods

H1: Sparsity of the support of K

- Graphical lasso

$$\hat{K}^{GL} = \arg\max_{K} \log\det(K) - \mathrm{tr}(K\Sigma_n) + \lambda \|\mathbf{K}\|_1 \quad (2)$$

  - [Meinshausen and Bühlmann, 2006], [Friedman et al., 2008]

H2: The graph is a tree

- Chow-Liu algorithm

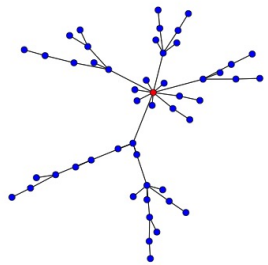


Figure: Tree

Outline

Effect of the missing variables

- Non measured variables
- Experimental conditions

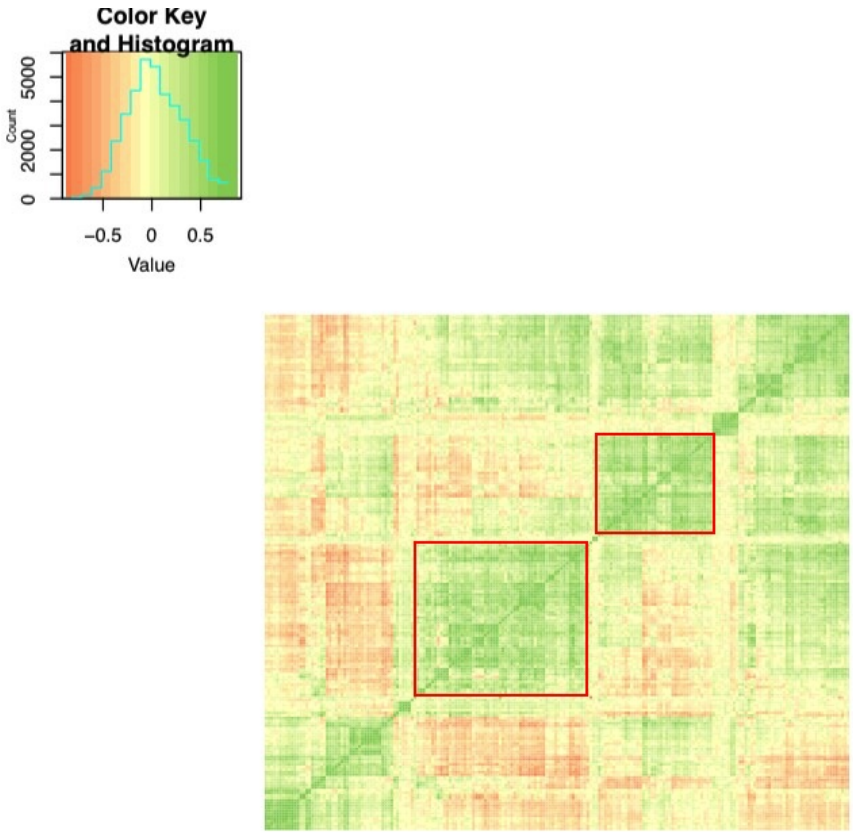
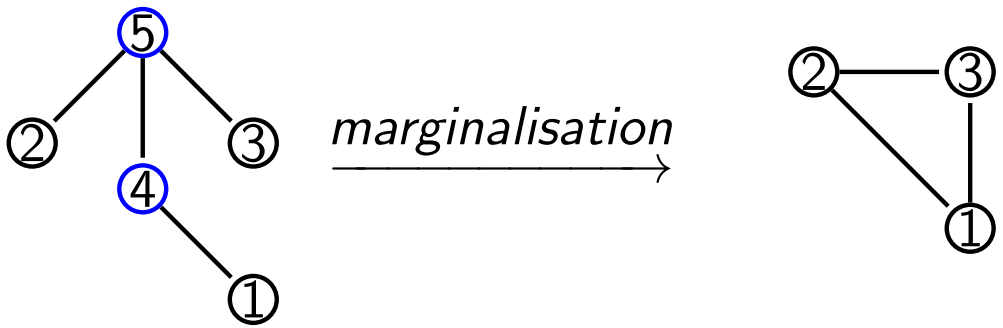


Figure: Covariance matrix. WGCNA data - 200 genes

Effect of the missing variables

- Missing variables involved in the process of interest but not measured
- $\mathcal{G} = (\{1, \ldots, p, p+1, \ldots, p+r\}, E)$, $\mathcal{G}_m = (\{1, \ldots, p\}, E_m)$
- Problem: inference of $\mathcal{G}_m$, $\quad \mathcal{G}$

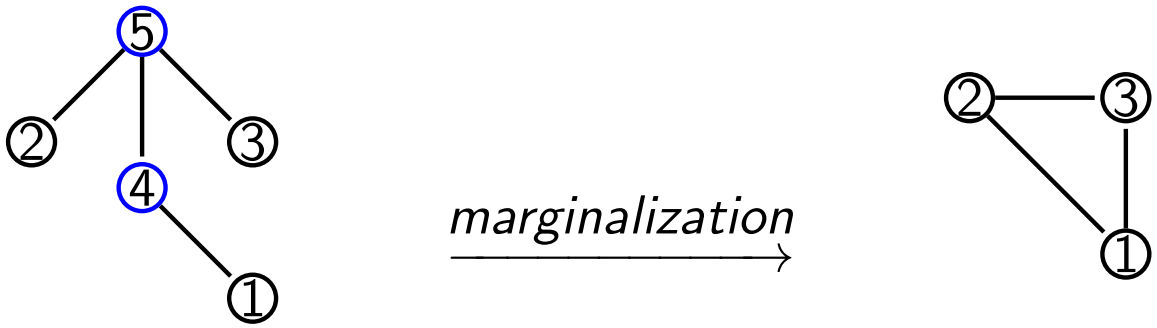
$$\begin{pmatrix} X_O \\ X_H \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix} :$$



- Apparition of cliques

O = Observed, H = Hidden

Effect of the missing variables



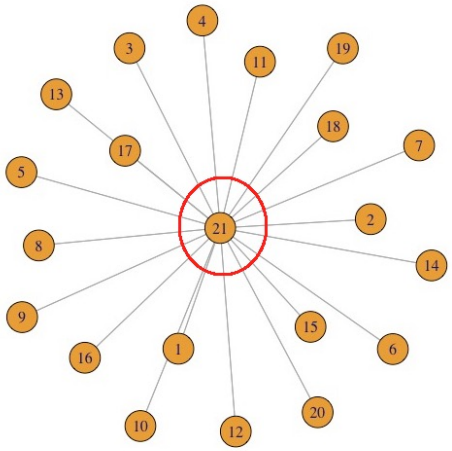
$$\mathcal{G} : K = \underbrace{\begin{pmatrix} K_{OO} & K_{OH} \\ K_{HO} & K_{HH} \end{pmatrix}}_{\text{arêtes de } E} \quad \Sigma = \begin{pmatrix} \Sigma_{OO} & \Sigma_{OH} \\ \Sigma_{HO} & \Sigma_{HH} \end{pmatrix}$$

$$\mathcal{G}_{\mathbf{m}} : K_m = \underbrace{K_{OO} - K_{OH} K_{HH}^{-1} K_{HO}}_{\text{arêtes de } E_m} \quad \Sigma_m = \Sigma_{OO}$$

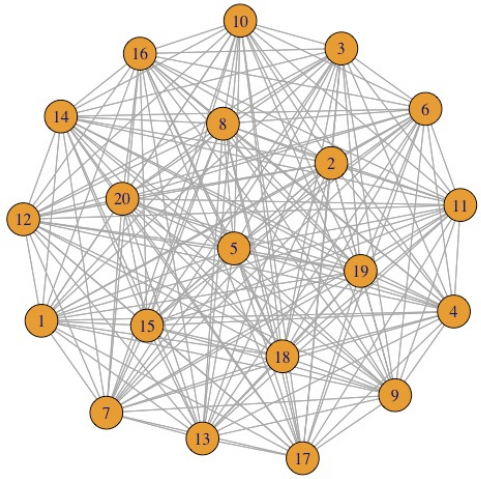
O = Observed, H = Hidden

Consequences

- [Chandrasekaran et al., 2012] $\mathcal{G}_m$ is not sparse



(a) Full graph



(b) Marginal graph

Consequences on interpretation + on quality of inference

Outline

EM algorithm

Missing data $\rightarrow$ EM algorithm

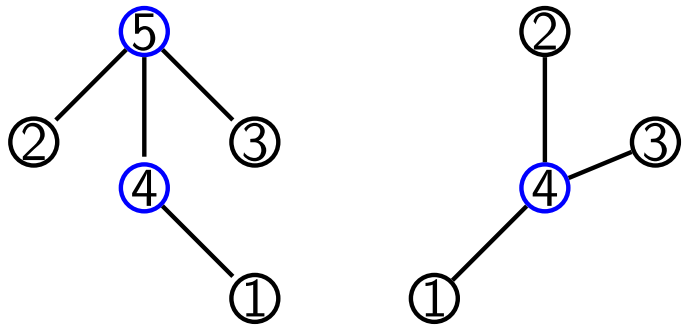
Parameters: T, $K = \begin{pmatrix} K_{OO} & K_{OH} \\ K_{HO} & K_{HH} \end{pmatrix}$, $\Sigma = \begin{pmatrix} \Sigma_{OO} & \Sigma_{OH} \\ \Sigma_{HO} & \Sigma_{HH} \end{pmatrix}$

E-step: $\mathbb{E}_{X_H|X_O;K^t}[l_c(X_H, X_O)] = \mathbb{E}_{X_H|X_O;K^t}[\log \det(K^t) - \text{tr}(K^t\Sigma)]$

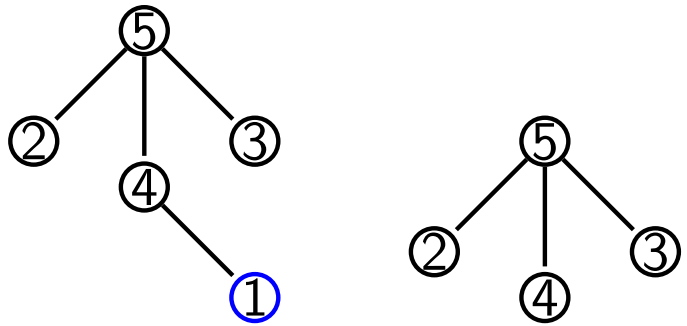
M-step: $K^{t+1} = \underbrace{\arg\max_K \log \det(K^t) - \text{tr}(K^t \mathbb{E}_{X_H|X_O;K^t}[\Sigma]) + \lambda \left\| K^t \right\|}_{\text{graphical lasso}}$

Identifiability issue

- $K_m \rightarrow K$?
- **chains**:



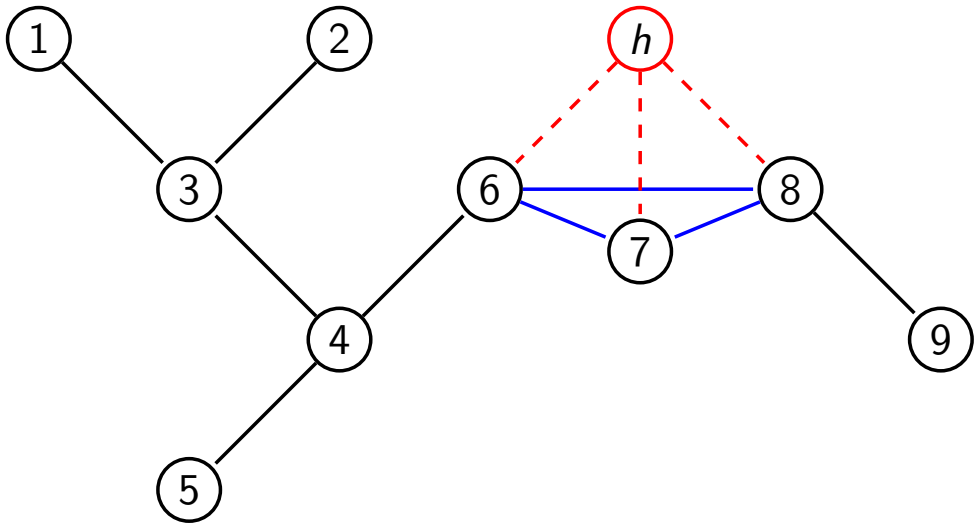
- **leaves**:



- trees [Choi et al., 2011]
- general case [Chandrasekaran et al., 2012]

Tree case

$\mathcal{G}$ **(rouge/noir) arbre**



Tree case

Sparsity

- p nodes $\rightarrow$ $p-1$ edges

Identifiability conditions

- No chain
- Every hidden variable has more than 3 neighbors

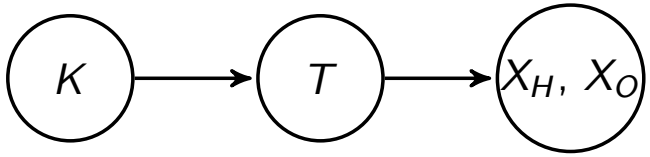
Inference methods

- Chow-Liu algorithm ($O(p^2 \log(p))$
- Tree of maximum likelihood

Outline

Latent Tree Model

Bayesian Inference of Graphical Model Structures Using Trees [Schwaller and Robin, 2015]



- Full matrix K fixed.
- Tree T random variable, a priori distribution $P(T) = \prod_{\{i,j\} \in E_T} \pi_{ij}$
- Probability of edge appearance $P(\{i,j\} \in E | X_O) = \underbrace{\sum_{\substack{T \in \mathcal{T} \\ \{i,j\} \in E_T}} P(T|X_O)}_{\text{Matrix-Tree theorem}}$
- EM: T, X_H latent variables

EM algorithm

E-step:

$$\mathbb{E}_{X_H,T|X_O;K^t}[L_c(X_O,X_H,T)|X_O] \quad = \mathbb{E}_{T|X_O;K^t}[\mathbb{E}_{X_H|X_O,T;K^t}[L_c(X_O,X_H,T)|X_O,T]|X_O].$$

- Computation using the Matrix-Tree theorem

M-step:

$$\begin{aligned} K^{t+1} &= \max_K \quad \mathbb{E}_{X_H,T|X_O;K^t}[L_c(X_O,X_H,T)|X_O] \\ K_{ij}^{t+1} &= \max_{K_{ij}} \quad \mathbb{E}_{X_H,T|X_O;K^t}[L_c(X_O,X_H,T)|X_O] \quad \text{(approx.)} \end{aligned}$$

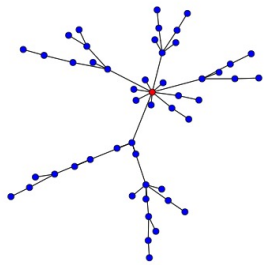
Output :

- Matrix A. A_{ij} interpreted as $P(\{i,j\} \in E \mid X_O)$, $i,j \in V$

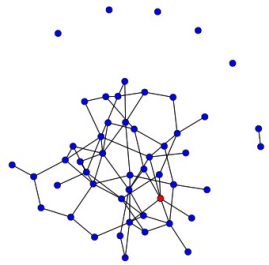
Outline

Simulated data

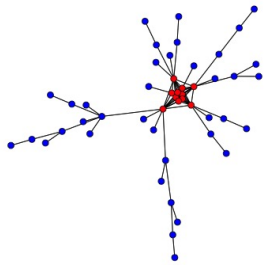
- Graphs of size $p = 50$: tree, erdös ($\pi = 0.1$)
- Samples of size $n = 200$



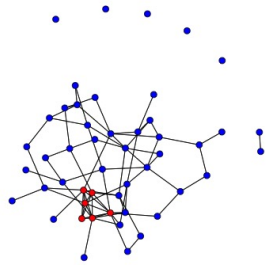
(a) Tree



(b) Erdös



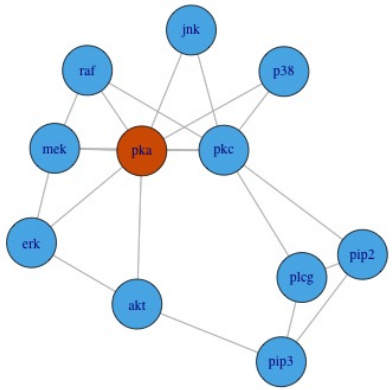
(c) Tree (marg.)



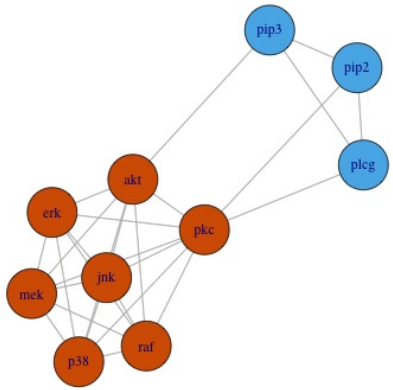
(d) Erdös (marg.)

Real data

- Raf network (regulation of cellular proliferation)
- Flow cytometry
- p=11, n=100



(a) Full graph



(b) Marginal graph

Experiments

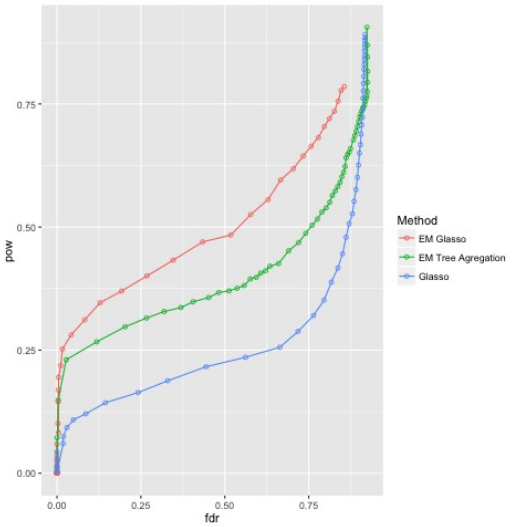
Compared methods

- Glasso (Meinshausen & Bühlmann approximation)
- EM-Glasso
- EM-aggregation

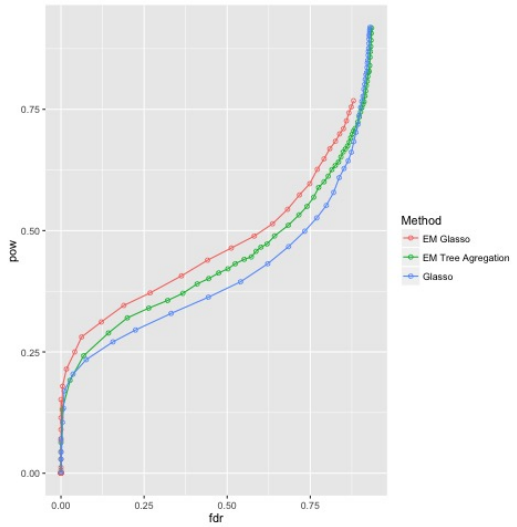
Evaluation criterion

$$\text{power} = \frac{TP}{FN + TP}, \text{ FDR} = \frac{FP}{FP + TP}$$

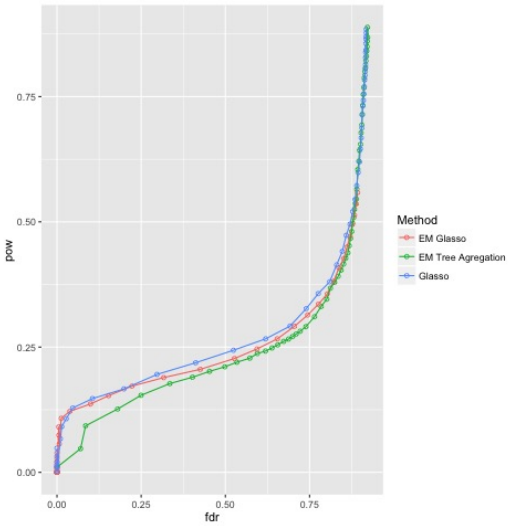
Results



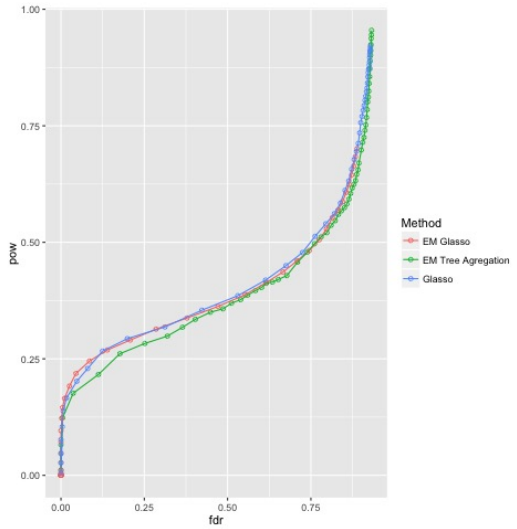
(a) tree



(b) erdös



(c) tree



(d) erdös

Flow cytometry

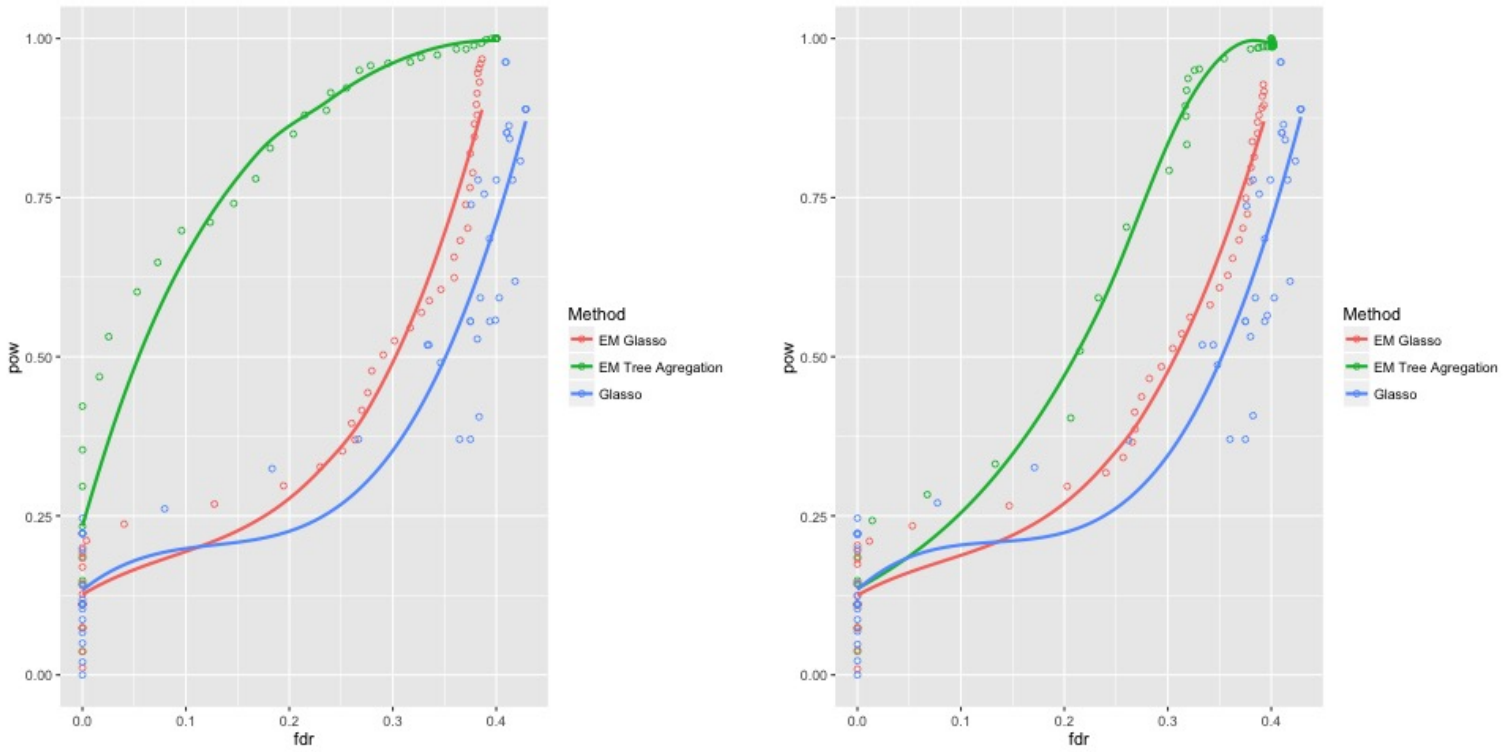


Figure: cytometry

Perspectives

Method for chosing the number of hidden variables

- Hierarchical classification

Evaluation of the problem complexity

- Entropy based criterion
- $\frac{H[P(T|X_O)] - \mathbb{E}_{X_H|X_O}[H[P(T|X_O,X_H)]]}{H[P(T|X_O)]}$.

Extension to non gaussian cases

- Count data
- Temporal data

References

V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4): 1935–1967, 2012.

M. J. Choi, V. Tan, A. Anandkumar, and A. S. Willsky. Learning latent tree graphical models. *The Journal of Machine Learning Research*, 12:1771–1812, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

L. Schwaller and S. Robin. Apprentissage de réseaux par agrégation bayésienne d'arbres couvrants. *Revue d'intelligence artificielle*, 2:153–172, 2015.

Initialisation

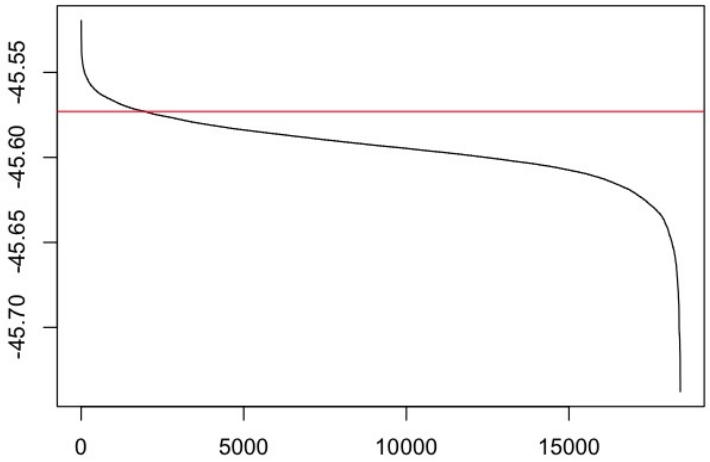


Figure: Vraisemblance des observations pour chaque triplet possible

Chow-Liu

Entrée Σ_n, sortie $\hat{T}^{CL}$, $\hat{K}^{CL}$

$$\hat{T}^{CL} = \arg\max_{T \text{ arbre}} \underbrace{\log(P(X;T))}_{\sum_{\{i,j\}\in E_T} I(X_i,X_j)+C} \tag{3}$$

1. Estimation des paramètres $\hat{I}(X_i, X_j)$
2. Arbre couvrant maximal pour poids $\hat{I}(X_i, X_j)$

Initialisation

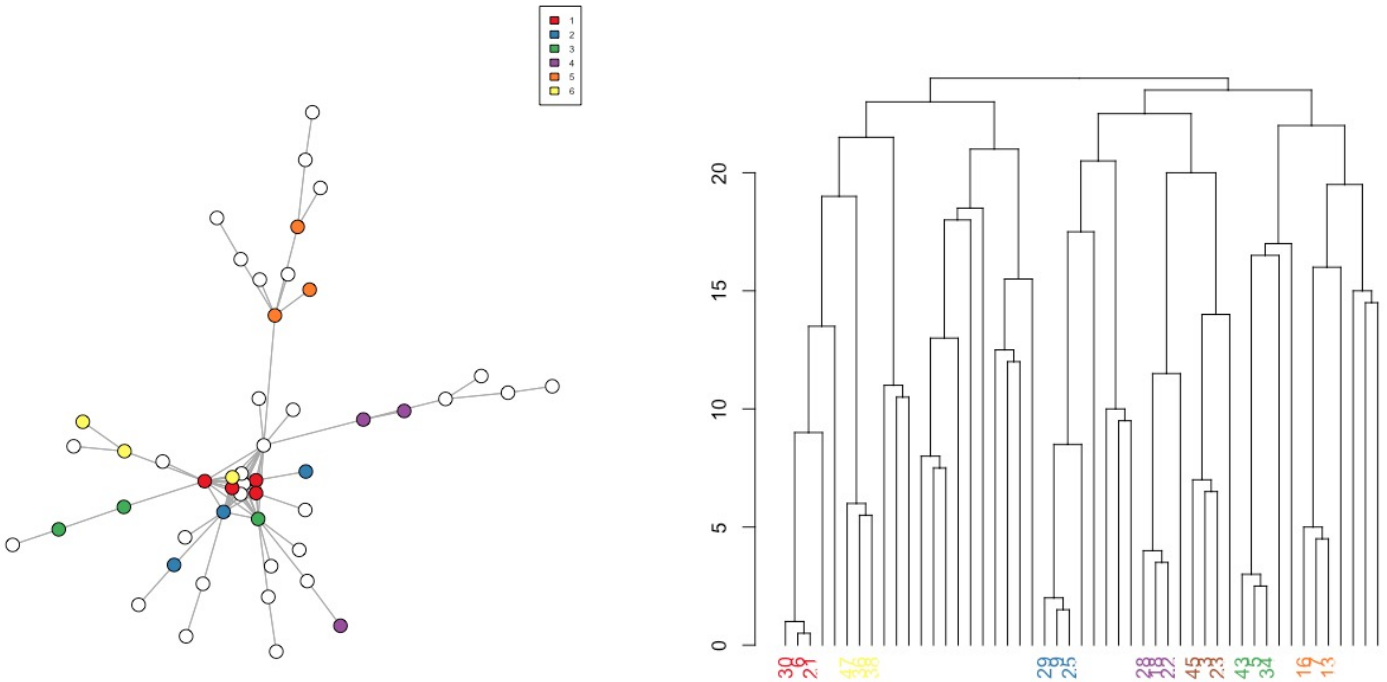


Figure: Classification hiérarchique au max du BIC