

# Learning gene regulation networks with Gaussian Graphical Model

**Christophe Giraud**<sup>(1)</sup>, Sylvie Huet<sup>(2)</sup>, Nicolas Verzelen<sup>(3)</sup>

(1) Ecole Polytechnique

(2) INRA Jouy-en-Josas

(3) INRA Montpellier

Luminy, 14 janvier 2010



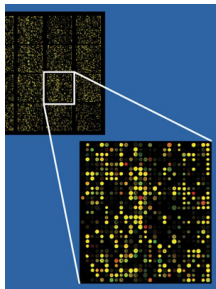
- 1 Introduction
- 2 Gaussian Graphical Models
- 3 GGMselect: estimation procedure for GGM
- 4 Final comments

**Goal:** Learn (part of) the gene regulation network of a given organism from transcriptomic data.

**Output:** a graph whose nodes represent the genes and whose edges represent some strong "statistical connection" between them.

**Exploratory point of view:**

- to suggest possible interactions between certain genes
- to suggest possible functions of orphan genes



- **Differential analysis** of data sets collected in different conditions (deletion, stress, etc).
  - Finding genes differentially expressed ( $\approx 5$  to 10% of the genes)
  - Usual analysis.
- **Analysis of the statistical dependences** between the transcriptomic levels (second order analysis). The whole data set is exploited.

## Supervised learning

- kernel methods (e.g. SIRENE), etc

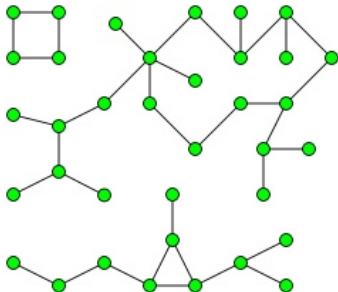
## Unsupervised learning: with Gaussian Graphical Models

- with no *a priori* (glasso, SPLICE, GGMselect, UPC, etc)
- with structural *a priori*: SIMoNe

# Graphical models

## Undirected models

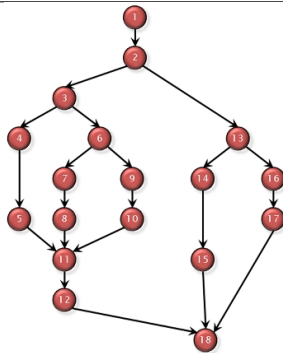
Markov random Fields / Gibbs Fields



$X_a$  independent of  $\{X_b : b \approx a\}$   
given  $\{X_b : b \sim a\}$

## Directed models

DAG models / Bayesian networks



$X_a$  indpt of  $\{X_b : a \nrightarrow \dots \nrightarrow b\}$   
given  $\{X_b : b \rightarrow a\}$

Directed models are natural models for time series (ordered variables).



## Be aware for unordered variables!

Non uniqueness of the minimal graphs  $\implies$  be cautious with the interpretation

**Ex:**  $X_{a+1} = \alpha X_a + \varepsilon_a$  with  $\varepsilon_a$  independent of  $X_1, \dots, X_{a-1}$ .  
Minimal directed graphs for AR(1):

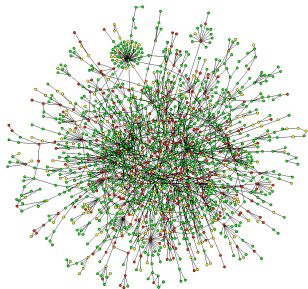
$$1 \rightarrow 2 \rightarrow \dots \rightarrow p \quad \text{and also} \quad 1 \leftarrow 2 \leftarrow \dots \leftarrow p$$

**Undirected models:** uniqueness of the minimal graph representing the conditional dependencies (under a weak hypothesis).

⇒ suited for unordered variables

**GGM:** Gaussian Graphical Models. The  $(X_1, \dots, X_p)$  are distributed according to a multivariate Gaussian law.

# GGM et gene regulation networks



**Model:** the gene expression levels are modeled with a GGM with unknown graph  $\mathbf{g}$ .

**Statistical goal:** estimate from transcriptomic data the graph  $\mathbf{g}$  of the GGM

**Main statistical challenge:** sample size  $n \ll p$  (number of genes)

- $p \approx$  a few 100 to a few 1000 genes
- $n \approx$  a few 10 experiments

**New algorithms:** based on multiple testing ideas or convex optimization.

**Main issues:**

- low algorithmic complexity
- statistical accuracy for small  $n$  + strong correlations
- choice of the "tuning parameters"

# GGMselect

**GGMselect:** R package

(available on <http://cran.r-project.org/>)

- Input:  $n \times p$  data matrix

$$X = \left[ X_i^{(a)} \right]_{\substack{i=1,\dots,n \\ a=1,\dots,p}} = \left[ X^{(1)}, \dots, X^{(p)} \right] = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$

Hypothesis:  $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \Sigma)$

- Output: estimated graph  $\hat{g}$

**Our goal:** to propose a procedure

- powerful in practice (power versus FDR)
- theoretically grounded in a non asymptotic setting

**Estimation scheme:**

- 1 build a family of candidate graphs from data
- 2 select one of the candidate graphs by minimizing some penalized empirical risk

# Family of candidate graphs

**Ideal:** family of candidate graphs = all possible graphs

**In practice:** cannot handle the  $2^{p(p-1)/2}$  possible graphs.

**Strategy:** use heuristics and existing estimation procedures to build a family  $\hat{\mathcal{G}}$  of candidate graphs from data

**Example of family  $\hat{\mathcal{G}}$  of candidate graphs:**

- **QE family:** Quasi-Exhaustive family
- **LA family:** based on the LASSO
- **C01 family:** based on 0-1 partial Correlation
- **EW family:** based on Exponential Weighting
- **mixed family:** any mix of the above families

**Quality criterion:** for a graph  $g$  compared to the "true" graph  $\mathbf{g}$ .

$$\begin{aligned}\text{MSEP}(g) &= \text{Mean Square Error of Prediction related to } g \\ &= \text{bias}(g) + \text{variance}(g)\end{aligned}$$

where

- $\text{bias}(g)$  quantifies how important are the missing edges
- $\text{variance}(g)$  is roughly proportional to the number of edges in  $g$  divided by  $n$ .

## Why MSEP?

It is a way to *quantify* the importance of each edge.

## Ideal:

select  $g^*$  such that  $\text{MSEP}(g^*)$  is minimal in  $\{\text{MSEP}(g) : g \in \hat{\mathcal{G}}\}$   
→  $g^*$  unknown!

## Selection criterion:

"select  $\hat{g}$  which minimizes some *penalized empirical* MSEP"

where the penalty term:

- roughly penalizes each node of  $\hat{g}$  according to its degree (number of edges),
- depends on a "scale-free" parameter  $\lambda > 1$ , typically  $\lambda = 2$
- is based on quantiles of Fisher random variables.

## Théorème (C.G., S.Huet, N.Verzelen 09)

If  $\max_{g \in \hat{\mathcal{G}}} \{\deg(g)\} \leq \rho \frac{n}{2(1.1 + \sqrt{\log \rho})^2}$ , for some  $\rho < 1$ ,

then the estimated graph  $\hat{g}$  fulfills

$$\text{MSEP}(\hat{g}) \leq c_{\rho, \lambda} \log(\rho) \mathbb{E} \left[ \inf_{g \in \hat{\mathcal{G}}} \text{MSEP}(g) \right] + R_n,$$

where  $R_n = O(\text{Tr}(\Sigma)e^{-c'_\rho n} + \text{CVar}(\Sigma) \log(\rho)/n)$   
with  $\text{CVar}(\Sigma) = \sum_a (\Sigma_{aa}^{-1})^{-1}$ .

## Théorème (C.G., S.Huet, N.Verzelen 09)

If  $\max_{g \in \hat{\mathcal{G}}} \{\deg(g)\} \leq \rho \frac{n}{2(1.1 + \sqrt{\log p})^2}$ , for some  $\rho < 1$ ,

then the estimated graph  $\hat{g}$  fulfills

$$\text{MSEP}(\hat{g}) \leq c_{\rho, \lambda} \log(p) \mathbb{E} \left[ \inf_{g \in \hat{\mathcal{G}}} \text{MSEP}(g) \right] + R_n,$$

where  $R_n = O(\text{Tr}(\Sigma)e^{-c'_{\rho}n} + \text{CVar}(\Sigma) \log(p)/n)$   
with  $\text{CVar}(\Sigma) = \sum_a (\Sigma_{aa}^{-1})^{-1}$ .

- **Optimal selection criterion?**
  - "minimal" size of the penalty to avoid overfitting
  - minimax estimation rates when  $\hat{\mathcal{G}}$  contains good graphs
- **What about the condition on the degree?**  $(n/2 \log p)$   
unavoidable, otherwise estimation rate gets worse.
- **What about the tuning parameter?**  $\lambda$   
scale-free: does not depend on unknown values.

## Family $\hat{\mathcal{G}}$ :

- **QE:** Quasi Exhaustive family
- **LA:** based on the LASSO
- **C01:** based on 0-1 partial Correlation
- **EW:** based on Exponential Weighting

## Competing procedures:

- **WB:** Wille & Bühlmann (06) with 0-1 partial correlation.
- **MB.or / MB.and:** Meinshausen & Bühlmann (06)  $\rightarrow$  LASSO
- **Aglasso:** Fan *et al.* (08), with adaptive glasso.

## Simulation scheme:

- heterogeneous random graph with 3 clusters.
- random covariance matrix  $\Sigma$ .
- sparsity index  $I_s =$  mean number of nodes per node

## Power versus FDR

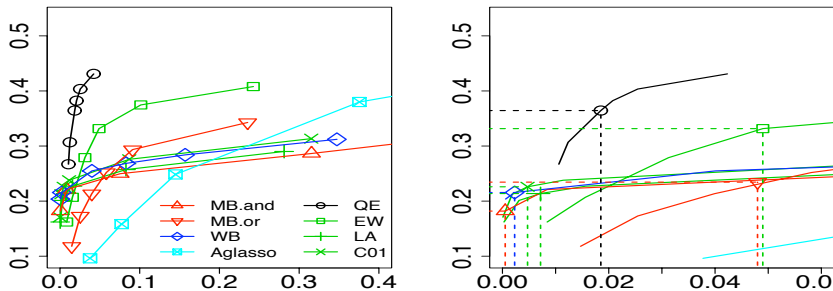


Figure: Power versus FDR for  $p = 100$ ,  $n = 50$  and  $I_s = 3$ .

# Effect of the sample size $n$

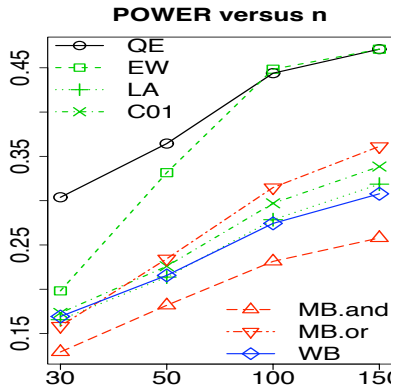
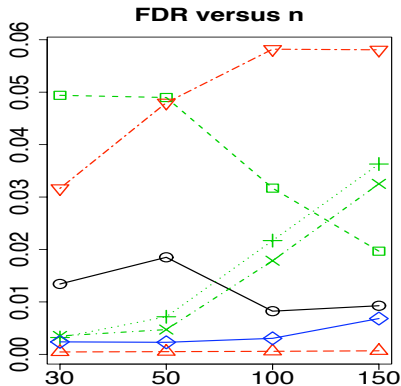


Figure: FDR versus  $n$  and power versus  $n$  for  $p = 100$  and  $I_s = 3$ .

# Conclusion

## Positive points:

- good theoretical and practical results
- no problem with the choice of the tuning parameter
- enable to calibrate any procedure and compare different ones

## Negative points: (on GGM)

- need enough repetitions
- hard to estimate hubs
- need to select a subfamily of genes
- no use of biological informations
- adequacy of Gaussian modeling?

## Positive points:

- good theoretical and practical results
- no problem with the choice of the tuning parameter
- enable to calibrate any procedure and compare different ones

## Negative points: (on GGM)

- need enough repetitions
- hard to estimate hubs
- need to select a subfamily of genes
- no use of biological informations
- adequacy of Gaussian modeling?

## In progress:

- cluster experiments which behave as repetitions (to enlarge the sample size). With E. Le Pennec and N. Verzelen.
- tests to compare different samples. With N. Verzelen.

**Future:** validate the different methods on a biological example which is fully known.