# Statistical Inference of biological networks

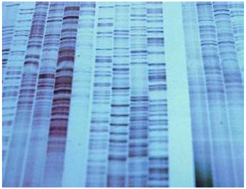
Y. Baraud[1], **C. Giraud**[1,2], S. Huet[2], N. Verzelen[3]

(1) Université de Nice, (2) INRA Jouy-en-Josas, (3) Université Paris Sud

Toulouse, juin 2008.

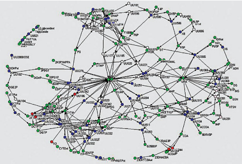
# From genomics to postgenomics

**Genomics:** genome sequencing of various organisms (human genome $\sim$ 04)



**Postgenomics:** try to understand how it works!

- Massive "omics" data sets
- Involves biology, physics, computer sciences, maths...



# Size of genomes

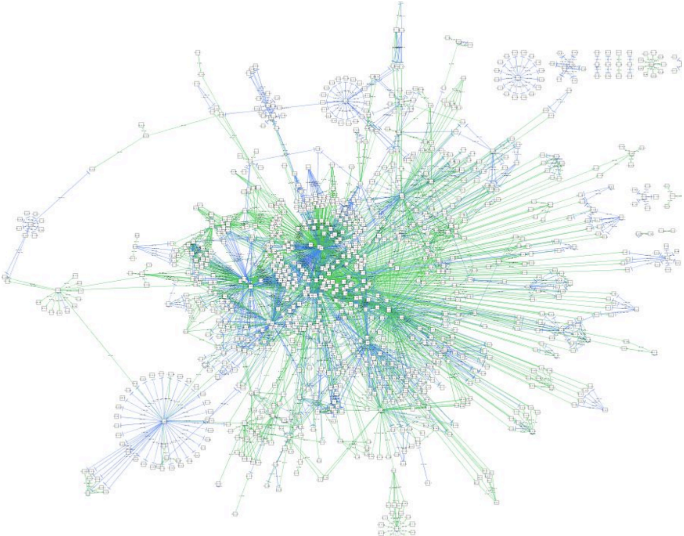
**Genome of a bacteria:**

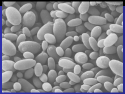
- 1 to 5 millions of base pairs
- 1 to 5 thousands genes
- a few millions of possible gene-gene interactions

**Human genome:**

- around 3.4 billions of base pairs
- around 25 000 genes
- more than 100 millions of possible gene-gene interactions

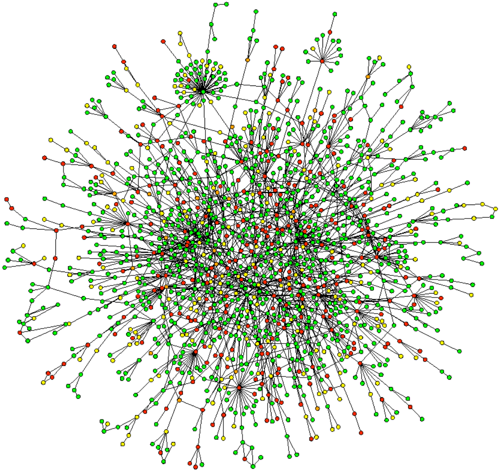
# Gene - gene regulation network of *E. coli*

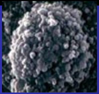




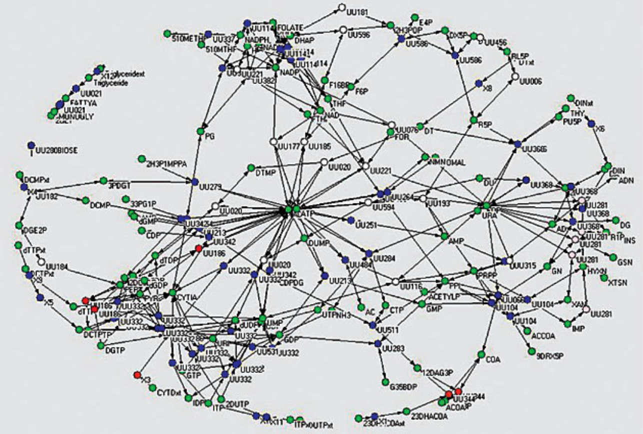
# Protein - protein network of *S. cerevisiae*

1458 proteins (vertices) and their 1948 known interactions (edges)





# Metabolic pathway of *U. urealyticum*



# Why focusing on regulation networks?

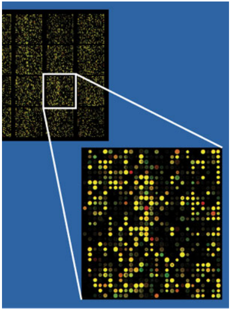
- **Traditional biology** studies the specific functions of individual genes, proteins or cells.

- **System biology** tries to understand how the whole system works by investigating the interaction network between genes, proteins, metabolites, etc.

$\longrightarrow$ emergent properties

# Inferring gene regulation networks

**Data:** massive transcriptomic data sets produced by microarrays.



- **Differential analysis** of data obtained in different conditions: with or without deletion of a gene, with or without stress, etc.
- **Analysis of the conditional dependences** in the data (exploits the whole data set).

# A few statistical tools

**Descriptive tools:**

- Kernel methods (supervised learning)

**Model based tools:**

- Bayesian Networks
- Gaussian Graphical Models

# Gaussian Graphical Models

# Gaussian Graphical Models

**Statistical model:** The transcription levels $(X^{(1)}, \dots, X^{(p)})$ of the $p$ genes are modeled by a Gaussian law in $\mathbb{R}^p$.

**Graph of the conditional dependences:** graph $\mathbf{g}$ with

$$\boxed{\begin{array}{c}\text{an edge } i \overset{\mathbf{g}}{\sim} j \text{ between the genes } i \text{ and } j \\ \underline{\text{iff}} \\ X^{(i)} \text{ and } X^{(j)} \text{ are not independent given } \left\{X^{(k)},\ k \neq i,j\right\}\end{array}}$$

regulation network $\longleftrightarrow$ graph $\mathbf{g}$

# The task of the statistician

**Goal:** estimate $\mathbf{g}$ from a sample $X_1, \ldots, X_n$.

**Main difficulty:** $n \ll p$

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

# Some new algorithms

**New algorithms:** based on thresholding or regularization

| **Multiple testing** | **Convex minimization** |
|---|---|
| - Drton & Perlman (2004) | - Meinshausen & Bühlmann (2006) |
| - Schäfer & Strimmer (2005) | - Huang *et al.* (2006) |
| - Wille & Bühlmann (2006) | - Yuan & Lin (2007) |
| - Verzelen & Villers (2007) | - Banerjee *et al.* (2007) |
| - Bühlmann & Kalisch (2008) | - Friedman *et al.* (2007) |
| ... | ... |

$\longrightarrow$ quite disappointing numerical performances (Villers *et al.* 2008)

$\longrightarrow$ no theoretical results or in an asymptotic framework (with strong hypotheses on the covariance)

# Penalized empirical risk estimation

# Partial correlations

**Hypothesis:** $(X^{(1)},\dots,X^{(p)}) \sim \mathcal{N}(0,C)$ in $\mathbb{R}^p$, with $C$ positive definite.

**Notation:** We write $\theta = \left(\theta_k^{(j)}\right)$ for the $p \times p$ matrix such that $\theta_j^{(j)} = 0$ and $\mathbb{E}\left(X^{(j)} \mid X^{(k)}, k \neq j\right) = \sum_{k \neq j} \theta_k^{(j)} X^{(k)}$.

**Skeleton of $\theta$:** we have $\theta_i^{(j)} = \frac{\mathbf{Cov}\left(X^{(i)}, X^{(j)} | X^{(k)},\ k \neq i,j\right)}{\mathbf{Var}\left(X^{(j)} | X^{(k)},\ k \neq j\right)}$ so

$$\boxed{\theta_i^{(j)} \neq 0 \iff i \overset{\mathbf{g}}{\sim} j}$$

# Partial correlations

**Hypothesis:** $(X^{(1)}, \dots, X^{(p)}) \sim \mathcal{N}(0, C)$ in $\mathbb{R}^p$, with $C$ positive definite.

**Notation:** We write $\theta = \left(\theta_k^{(j)}\right)$ for the $p \times p$ matrix such that $\theta_j^{(j)} = 0$ and $\quad \mathbb{E}\left(X^{(j)} \mid X^{(k)}, k \neq j\right) = \sum_{k \neq j} \theta_k^{(j)} X^{(k)}$.

**Skeleton of $\theta$:** we have $\quad \theta_i^{(j)} = \frac{\mathbf{Cov}\left(X^{(i)}, X^{(j)} | X^{(k)},\ k \neq i,j\right)}{\mathbf{Var}\left(X^{(j)} | X^{(k)},\ k \neq j\right)}$ so

$$\boxed{\theta_i^{(j)} \neq 0 \iff i \overset{\mathrm{g}}{\sim} j}$$

# Partial correlations

**Hypothesis:** $(X^{(1)}, \dots, X^{(p)}) \sim \mathcal{N}(0, C)$ in $\mathbb{R}^p$, with $C$ positive definite.

**Notation:** We write $\theta = \left(\theta_k^{(j)}\right)$ for the $p \times p$ matrix such that $\theta_j^{(j)} = 0 \text{ and } \quad \mathbb{E}\left(X^{(j)} \mid X^{(k)}, k \neq j\right) = \sum_{k \neq j} \theta_k^{(j)} X^{(k)}$.

**Skeleton of $\theta$:** we have $\theta_i^{(j)} = \dfrac{\mathbf{Cov}\left(X^{(i)}, X^{(j)} | X^{(k)},\ k \neq i,j\right)}{\mathbf{Var}\left(X^{(j)} | X^{(k)},\ k \neq j\right)}$ so

$$\boxed{\theta_i^{(j)} \neq 0 \iff i \overset{\mathbf{g}}{\sim} j}$$

# Estimation strategy

**Goal:** Estimate $\theta$ from a sample $X_1, \ldots, X_n$ with quality criterion

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}\left[\|C^{1/2}(\hat{\theta}-\theta)\|^2_{p\times p}\right] = \mathbb{E}\left[\|X_{new}^T(\hat{\theta}-\theta)\|^2_{1\times p}\right]$$

**Estimation strategy:**

1. Choose a collection $\mathcal{G}$ of candidate graphs.
2. Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$ of $\theta$.
3. Select one of them by minimizing a penalized empirical risk.

# Estimation strategy

**Goal:** Estimate $\theta$ from a sample $X_1, \ldots, X_n$ with quality criterion

$$\mathrm{MSEP}(\hat{\theta}) \;=\; \mathbb{E}\left[\|C^{1/2}(\hat{\theta}-\theta)\|^2_{p\times p}\right] \;=\; \mathbb{E}\left[\|X^T_{new}(\hat{\theta}-\theta)\|^2_{1\times p}\right]$$

## Estimation strategy:

1. Choose a collection $\mathcal{G}$ of candidate graphs.
2. Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$ of $\theta$.
3. Select one of them by minimizing a penalized empirical risk.

# Estimation strategy

**Goal:** Estimate $\theta$ from a sample $X_1, \ldots, X_n$ with quality criterion

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}\left[\|C^{1/2}(\hat{\theta} - \theta)\|^2_{p\times p}\right] = \mathbb{E}\left[\|X_{new}^T(\hat{\theta} - \theta)\|^2_{1\times p}\right]$$

## Estimation strategy:

1. Choose a collection $\mathcal{G}$ of candidate graphs.
2. Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$ of $\theta$.
3. Select one of them by minimizing a penalized empirical risk.

# Estimation strategy

**Goal:** Estimate $\theta$ from a sample $X_1, \ldots, X_n$ with quality criterion

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}\left[\|C^{1/2}(\hat{\theta} - \theta)\|^2_{p\times p}\right] = \mathbb{E}\left[\|X_{new}^T(\hat{\theta} - \theta)\|^2_{1\times p}\right]$$

## Estimation strategy:

1. Choose a collection $\mathcal{G}$ of candidate graphs.
2. Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$ of $\theta$.
3. Select one of them by minimizing a penalized empirical risk.

# Estimation strategy

**Goal:** Estimate $\theta$ from a sample $X_1, \ldots, X_n$ with quality criterion

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}\left[\|C^{1/2}(\hat{\theta} - \theta)\|^2_{p \times p}\right] = \mathbb{E}\left[\|X^T_{new}(\hat{\theta} - \theta)\|^2_{1 \times p}\right]$$

## Estimation strategy:

1. Choose a collection $\mathcal{G}$ of candidate graphs.
2. Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$ of $\theta$.
3. Select one of them by minimizing a penalized empirical risk.

# 1- Collection of candidate graphs

**Choice of a collection $\mathcal{G}$ of candidate graphs**

**Examples**

- Set of all the graphs with $p$ vertices and degree $\leq D$,
- Set of all the graph with $p$ vertices and degree $\leq D$ containing a known subgraph $g_o$.

**Model associated to $g \in \mathcal{G}$ to estimate $\theta$:**

$$g \curvearrowright \Theta_g = \left\{ \theta \in \mathbb{R}^{p \times p} : \; i \overset{g}{\nsim} j \;\Rightarrow\; \theta_i^{(j)} = 0 \right\}$$

# 1- Collection of candidate graphs

**Choice of a collection $\mathcal{G}$ of candidate graphs**

**Examples**

- Set of all the graphs with $p$ vertices and degree $\leq D$,
- Set of all the graph with $p$ vertices and degree $\leq D$ containing a known subgraph $g_o$.

**Model associated to $g \in \mathcal{G}$ to estimate $\theta$:**

$$g \curvearrowright \Theta_g = \left\{ \theta \in \mathbb{R}^{p \times p} : \ i \overset{g}{\nsim} j \ \Rightarrow \ \theta_i^{(j)} = 0 \right\}$$

# 2- Estimator $\hat{\theta}_g$ associated to $g$

**Characterization:** $\theta = \mathrm{argmin}_{A\in\Theta} \|C^{1/2}(I-A)\|^2_{p\times p}$

where $\Theta$ = set of matrices with null diagonal.

**Empirical version:** $C^{1/2} \leftrightarrow X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = [X^{(1)}, \ldots, X^{(p)}]$

Estimator associated to $g$:

$$\hat{\theta}_g = \underset{A\in\Theta_g}{\mathrm{argmin}} \|X(I-A)\|^2_{n\times p}$$

# 2- Estimator $\hat{\theta}_g$ associated to $g$

**Characterization:** $\theta = \text{argmin}_{A\in\Theta} \|C^{1/2}(I-A)\|^2_{p\times p}$
where $\Theta =$ set of matrices with null diagonal.

**Empirical version:** $C^{1/2} \leftrightarrow X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \left[X^{(1)}, \ldots, X^{(p)}\right]$

Estimator associated to $g$:

$$\hat{\theta}_g = \underset{A\in\Theta_g}{\text{argmin}} \|X(I-A)\|^2_{n\times p}$$

# 2- Estimator $\hat{\theta}_g$ associated to $g$

**Characterization:** $\theta = \operatorname{argmin}_{A\in\Theta} \|C^{1/2}(I-A)\|^2_{p\times p}$
where $\Theta =$ set of matrices with null diagonal.

**Empirical version:** $C^{1/2} \leftrightarrow X = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix} = \left[X^{(1)},\ldots,X^{(p)}\right]$

Estimator associated to $g$:

$$\hat{\theta}_g = \operatorname*{argmin}_{A\in\Theta_g} \|X(I-A)\|^2_{n\times p}$$

# 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{nC_{jj}^{-1}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \underbrace{\|X(I - \hat{\theta}_g)\|^2}_{\text{empirical MSEP}} + \mathrm{pen}(g)$$

## 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{n C^{-1}_{jj}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \underbrace{\|X(I - \hat{\theta}_g)\|^2}_{\text{empirical MSEP}} (1 + \mathrm{pen}(g))$$

# 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{n C_{jj}^{-1}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \underbrace{\|X(I - \hat{\theta}_g)\|^2}_{\text{empirical MSEP}} \times (1 + \mathrm{pen}(g))$$

# 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{nC_{jj}^{-1}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \underbrace{\|X(I - \hat{\theta}_g)\|^2}_{\text{empirical MSEP}} \times (1 + \mathrm{pen}(g))$$

# 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{nC_{jj}^{-1}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \sum_{j=1}^{p} \left[\|X^{(j)} - X\hat{\theta}_g^{(j)}\|^2\right] \times \left(1 + \frac{\mathrm{pen}(\deg(j))}{n - \deg(j)}\right)$$

# 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{nC^{-1}_{jj}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \sum_{j=1}^{p} \left[\|X^{(j)} - X\hat{\theta}_g^{(j)}\|^2\right] \times \left(1 + \frac{\mathrm{pen}(\deg(j))}{n - \deg(j)}\right)$$

# 3- Which estimator should be chosen among $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

**Ideal:** choose $\hat{\theta}_{g_*}$ minimizing

$$\mathrm{MSEP}(\hat{\theta}_g) = \mathbb{E}\left(\|C^{1/2}(\theta - \hat{\theta}_g)\|^2\right) \approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{nC^{-1}_{jj}}$$

**Selection criterion:** set $\hat{\theta} = \hat{\theta}_{\hat{g}}$ where $\hat{g}$ minimizes over $\mathcal{G}$

$$\mathrm{crit}(g) = \sum_{j=1}^{p} \left[\|X^{(j)} - X\hat{\theta}_g^{(j)}\|^2\right] \times \left(1 + \frac{\mathrm{pen}(\deg(j))}{n - \deg(j)}\right)$$

which penalty $\mathrm{pen}$?

# The penalty

**Notation:** let EDkhi$(d, N, x)$ be the inverse of

$$x \mapsto \mathbb{P}\left(F_{d+2,N} \geq \frac{x}{d+2}\right) - \frac{x}{d}\,\mathbb{P}\left(F_{d,N+2} \geq \frac{N+2}{Nd}\,x\right)$$

where $F_{d,N}$ is a Fisher$(d,N)$.

**Penality:** For $K > 1$ we set

$$\begin{aligned} \mathrm{pen}(d) &= K\,\frac{n-d}{n-d-1}\,\mathrm{EDkhi}\left[d+1, n-d-1, \left(C_{p-1}^{d}(d+1)^2\right)^{-1}\right] \\ &\lesssim K\left(1+e^{\rho}\sqrt{2\log p}\right)^2(d+1) \quad \text{when} \quad d \leq \rho\,\frac{n}{2\log p}. \end{aligned}$$

# The penalty

**Notation:** let EDkhi$(d, N, x)$ be the inverse of

$$x \mapsto \mathbb{P}\left(F_{d+2,N} \geq \frac{x}{d+2}\right) - \frac{x}{d}\,\mathbb{P}\left(F_{d,N+2} \geq \frac{N+2}{Nd}\,x\right)$$

where $F_{d,N}$ is a Fisher$(d,N)$.

**Penality:** For $K > 1$ we set

$$\begin{aligned} \mathrm{pen}(d) &= K\,\frac{n-d}{n-d-1}\,\mathrm{EDkhi}\left[d+1, n-d-1, \left(C^{d}_{p-1}(d+1)^2\right)^{-1}\right] \\ &\lesssim K\left(1+e^{\rho}\sqrt{2\log p}\right)^2(d+1) \quad \text{when} \quad d \leq \rho\,\frac{n}{2\log p}. \end{aligned}$$

## Theorem: risk bound.

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho\, \frac{n}{2\left(1.1+\sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then the MSEP of $\hat{\theta}$ is bounded by

$$\mathrm{MSEP}(\hat{\theta}) \ \leq\ c_{K,\rho} \log(p) \inf_{g \in \mathcal{G}} \left\{ \mathrm{MSEP}(\hat{\theta}_g) \vee \frac{\|C^{1/2}(I-\theta)\|^2}{n} \right\} + R_n$$

where $R_n = O\big(\mathrm{Tr}(C) e^{-\kappa_\rho n}\big)$.

## Theorem: risk bound.

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho\, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then the MSEP of $\hat{\theta}$ is bounded by

$$\mathrm{MSEP}(\hat{\theta}) \;\leq\; c_{K,\rho} \log(p) \inf_{g \in \mathcal{G}} \left\{ \mathrm{MSEP}(\hat{\theta}_g) \vee \frac{\|C^{1/2}(I-\theta)\|^2}{n} \right\} + R_n$$

where $R_n = O\big(\mathrm{Tr}(C) e^{-\kappa_\rho n}\big)$.

## Theorem: risk bound.

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then the MSEP of $\hat\theta$ is bounded by

$$\mathrm{MSEP}(\hat\theta) \ \leq \ c_{K,\rho} \log(p) \inf_{g \in \mathcal{G}} \left\{ \mathrm{MSEP}(\hat\theta_g) \vee \frac{\|C^{1/2}(I-\theta)\|^2}{n} \right\} + R_n$$

where $R_n = O\big(\mathrm{Tr}(C) e^{-\kappa_\rho n}\big)$.

## Theorem: risk bound.

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho\, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then the MSEP of $\hat{\theta}$ is bounded by

$$\mathrm{MSEP}(\hat{\theta}) \;\leq\; c_{K,\rho} \log(p) \inf_{g \in \mathcal{G}} \left\{ \mathrm{MSEP}(\hat{\theta}_g) \vee \frac{\|C^{1/2}(I-\theta)\|^2}{n} \right\} + R_n$$

where $R_n = O\left(\mathrm{Tr}(C) e^{-\kappa_\rho n}\right)$.

## Theorem: risk bound.

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then the MSEP of $\hat{\theta}$ is bounded by

$$\mathrm{MSEP}(\hat{\theta}) \leq c_{K,\rho} \log(p) \inf_{g \in \mathcal{G}} \left\{ \mathrm{MSEP}(\hat{\theta}_g) \vee \frac{\|C^{1/2}(I - \theta)\|^2}{n} \right\} + R_n$$

where $R_n = O\left(\mathrm{Tr}(C) e^{-\kappa_\rho n}\right)$.

# Optimality?

1. **Where does the condition on the degree come from?**

2. **Can we choose a smaller penalty?**

## Theory

# Condition on the degree

# Which size of graph can we hope to estimate?

**Prediction error:**

$$\mathrm{MSEP}(\hat{\theta}) = \mathbb{E}(\|C^{1/2}(\theta-\hat{\theta})\|^2) = \mathbb{E}(\|C^{1/2}(I-\hat{\theta})\|^2)-\|C^{1/2}(I-\theta)\|^2$$

To control the MSEP, we would like to have with large probability

$$(1-\delta)\|C^{1/2}(I-\hat{\theta})\|_{p\times p} \leq \frac{1}{\sqrt{n}}\|X(I-\hat{\theta})\|_{n\times p} \leq (1+\delta)\|C^{1/2}(I-\hat{\theta})\|_{p\times p}$$

for all matrices $\hat{\theta} \in \bigcup_{g\in\mathcal{G}} \Theta_g$.

# Which size of graph can we hope to estimate?

**Prediction error:**

$$\mathrm{MSEP}(\hat\theta) = \mathbb{E}(\|C^{1/2}(\theta-\hat\theta)\|^2) = \mathbb{E}(\|C^{1/2}(I-\hat\theta)\|^2) - \|C^{1/2}(I-\theta)\|^2$$

To control the MSEP, we would like to have with large probability

$$(1-\delta)\|C^{1/2}(I-\hat\theta)\|_{p\times p} \leq \frac{1}{\sqrt{n}}\|X(I-\hat\theta)\|_{n\times p} \leq (1+\delta)\|C^{1/2}(I-\hat\theta)\|_{p\times p}$$

for all matrices $\hat\theta \in \bigcup_{g\in\mathcal{G}} \Theta_g$.

**Proposition:** From empirical to population MSEP

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho \,\frac{n}{2\left(1.1+\sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then for $\delta > \sqrt{\rho}$,

we have with probability $\geq 1 - 2\exp\left(-n(\delta-\sqrt{\rho})^2/2\right)$

$$(1-\delta)\|C^{1/2}(I-A)\| \leq \frac{1}{\sqrt{n}}\|X(I-A)\| \leq (1+\delta)\|C^{1/2}(I-A)\|$$

for all matrices $A \in \bigcup_{g\in\mathcal{G}} \Theta_g$.

**Proposition:** From empirical to population MSEP

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho\, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then for $\delta > \sqrt{\rho}$,

we have with probability $\geq 1 - 2\exp\left(-n(\delta - \sqrt{\rho})^2/2\right)$

$$(1-\delta)\|C^{1/2}(I-A)\| \leq \frac{1}{\sqrt{n}}\|X(I-A)\| \leq (1+\delta)\|C^{1/2}(I-A)\|$$

for all matrices $A \in \bigcup_{g\in\mathcal{G}} \Theta_g$.

## Proposition: From empirical to population MSEP

When $\deg(\mathcal{G}) = \max\{\deg(g),\ g \in \mathcal{G}\}$ fulfills

$$\boxed{\deg(\mathcal{G}) \leq \rho\, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}, \quad \text{for some} \quad \rho < 1,}$$

then for $\delta > \sqrt{\rho}$,

we have with probability $\geq 1 - 2\exp\left(-n(\delta - \sqrt{\rho})^2/2\right)$

$$(1-\delta)\|C^{1/2}(I-A)\| \leq \frac{1}{\sqrt{n}}\|X(I-A)\| \leq (1+\delta)\|C^{1/2}(I-A)\|$$

for all matrices $A \in \bigcup_{g \in \mathcal{G}} \Theta_g$.

## Lemma: Restricted Infimum of Random Matrices

Consider a $n \times p$ matrix $Z$ with $n \leq p$ and i.i.d. $Z_{i,j} \sim \mathcal{N}(0,1)$. Consider also a collection $V_1, \ldots, V_N$ of subspaces of $\mathbb{R}^p$ with dimension $d < n$.

Then for any $x > 0$

$$\mathbb{P}\left[\inf_{v \in V_1 \cup \ldots \cup V_N} \frac{\frac{1}{\sqrt{n}}\|Zv\|}{\|v\|} \leq 1 - \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N + x}{\sqrt{n}}\right] \leq e^{-x^2/2}$$

where $\delta_N = \frac{1}{N\sqrt{8\log N}}$.

## Lemma: Restricted Supremum of Random Matrices

Consider a $n \times p$ matrix $Z$ with $n \leq p$ and i.i.d. $Z_{i,j} \sim \mathcal{N}(0,1)$. Consider also a collection $V_1, \ldots, V_N$ of subspaces of $\mathbb{R}^p$ with dimension $d < n$.

Then for any $x > 0$

$$\mathbb{P}\left[\sup_{v \in V_1 \cup \ldots \cup V_N} \frac{\frac{1}{\sqrt{n}}\|Zv\|}{\|v\|} \leq 1 + \frac{\sqrt{d} + \sqrt{2\log N} + \delta_N + x}{\sqrt{n}}\right] \leq e^{-x^2/2}$$

where $\delta_N = \frac{1}{N\sqrt{8\log N}}$.

# A geometrical constraint

When $C = I$, there exists some constant $c(\delta) > 0$ such that for any $n, p, \mathcal{G}$ fulfilling

$$\deg(\mathcal{G}) \geq c(\delta) \frac{n}{1 + \log(p/n)},$$

there exists no $n \times p$ matrix $X$ fulfilling

$$(1-\delta)\|C^{1/2}(I-A)\| \leq \frac{1}{\sqrt{n}} \|X(I-A)\| \leq (1+\delta)\|C^{1/2}(I-A)\|$$

for all matrices $A \in \bigcup_{g \in \mathcal{G}} \Theta_g$.

## Theory

# Minimal penalty

# Minimal size of $\mathrm{pen}(d)$?

In the simple case where $\mathbf{g} = \emptyset$ we would like to select a "small" $\hat{g}$

## Minimal penalty

$$\Longrightarrow \text{"}\mathrm{pen}(d) \geq 2d\log(p)\text{"}$$

## Proposition: Over-fitting

For some $D < n < p$ and $p \geq e^{2/(1-\gamma)} + 1$, assume that:

- $\mathrm{pen}(d) = 2(1-\gamma) d \log(p-1) \quad \text{with } \gamma \in (0,1)$,
- $\{\text{graphs with at most } D \text{ edges}\} \subset \mathcal{G}$,
- $\mathbf{g} = \emptyset$.

Then,

$$\mathbb{P}\left(\deg(\hat{g}) \geq \frac{c(\gamma)\min(n, p^{\gamma/4})}{(\log p)^{3/2}} \wedge \lfloor \gamma D/8 \rfloor\right) \geq 1 - \frac{3}{p-1} - 2e^{-\gamma^2 n/8^3}$$

where $c(\gamma) > 0$ is an (explicit) constant depending on $\gamma$ only.

## Proposition: Over-fitting

For some $D < n < p$ and $p \geq e^{2/(1-\gamma)} + 1$, assume that:

- $\mathrm{pen}(d) = 2(1-\gamma)d\log(p-1) \quad \text{with } \gamma \in (0,1)$,
- $\{\text{graphs with at most } D \text{ edges}\} \subset \mathcal{G}$,
- $\mathbf{g} = \emptyset$.

Then,

$$\mathbb{P}\left(\deg(\hat{g}) \geq \frac{c(\gamma)\min(n, p^{\gamma/4})}{(\log p)^{3/2}} \wedge \lfloor \gamma D/8 \rfloor\right) \geq 1 - \frac{3}{p-1} - 2e^{-\gamma^2 n/8^3}$$

where $c(\gamma) > 0$ is an (explicit) constant depending on $\gamma$ only.

## In practice

# Numerical performance

# Numerical performance

**Comparison to Meinshausen & Buhlmann's procedure.**

**Setting:**

- random "Erdos-Reny" graphs & and random covariance matrices
- $n = 15$ observations

**Focus on two settings:**

- when the density of the graph increases
- when the number $p$ of covariables increases

# $n = 15$, $p = 10$, edges$=$ 10%, 30% & 33%

## edges $=$ 10%

| | KGGM | MB |
|---|---|---|
| risk/oracle | 2.5 | 3.3 |
| Puissance | 81% | 81% |
| FDR | 4.4% | 3.7% |

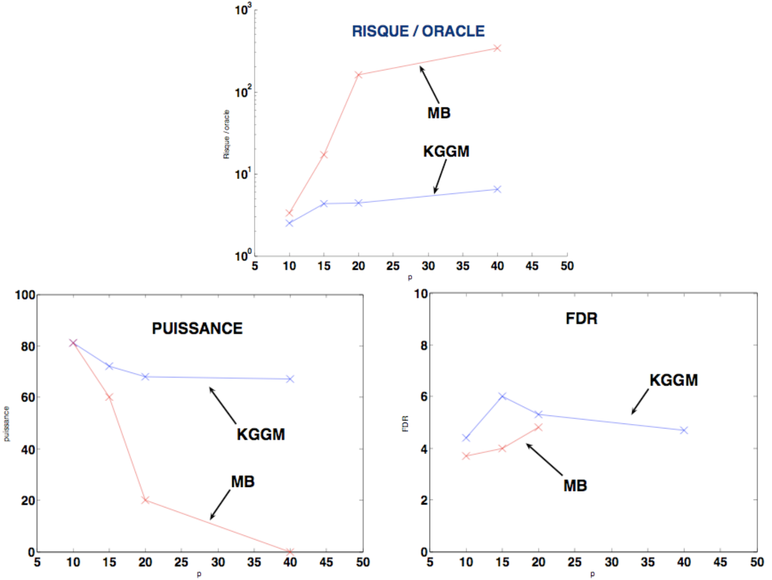
## edges $=$ 30%

| | KGGM | MB |
|---|---|---|
| risk/oracle | 4.9 | 6.9 |
| Puissance | 20% | 14% |
| FDR | 5.4% | 2.9% |

## edges $=$ 33%

| | KGGM | MB |
|---|---|---|
| risk/oracle | 4.9 | 6.4 |
| Puissance | 10% | 3.5% |
| FDR | 4.1% | 1.1% |

# When $p$ increases ($n = 15$, fixed sparsity)



# Conclusion

## Some nice features:

- **good theoretical properties:** non-asymptotic control of the MSEP with no condition on the covariance matrix $C$
- **good numerical performances:** even when $n \ll p$

## BUT

- **very high numerical complexity:** typically $n \times p^{\deg(\mathcal{G})+1}$

$\Longrightarrow$ cannot be used in practice when $p > 50 \ldots$

# Conclusion

## Some nice features:

- **good theoretical properties:** non-asymptotic control of the MSEP with no condition on the covariance matrix $C$
- **good numerical performances:** even when $n \ll p$

## BUT

- **very high numerical complexity:**
  typically $n \times p^{\deg(\mathcal{G})+1}$

$$\boxed{\Longrightarrow \text{ cannot be used in practice when } p > 50 \ldots}$$

# Practical implementation

**Ongoing work:** with S. Huet and N. Verzelen

**Reduction of the collection of graphs,** with the aim to find a balance between

- computational efficiency
- estimation accuracy

# References


- C. Giraud. *Estimation of Gaussian graphs by model selection.* Preprint (2007). http://hal.archives-ouvertes.fr/hal-00178275/fr/

- Y. Baraud, C. Giraud, S. Huet. *Gaussian model selection with unknown variance.* To appear in the Annals of Statistics (2008?).

- N. Verzelen. *High-dimensional Gaussian model selection on a Gaussian design.* Preprint (2008).