

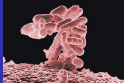
Statistical Inference in Gaussian Graphical Models

Y. Baraud⁽¹⁾, **C. Giraud**^(1,2), S. Huet⁽²⁾, N. Verzelen⁽³⁾

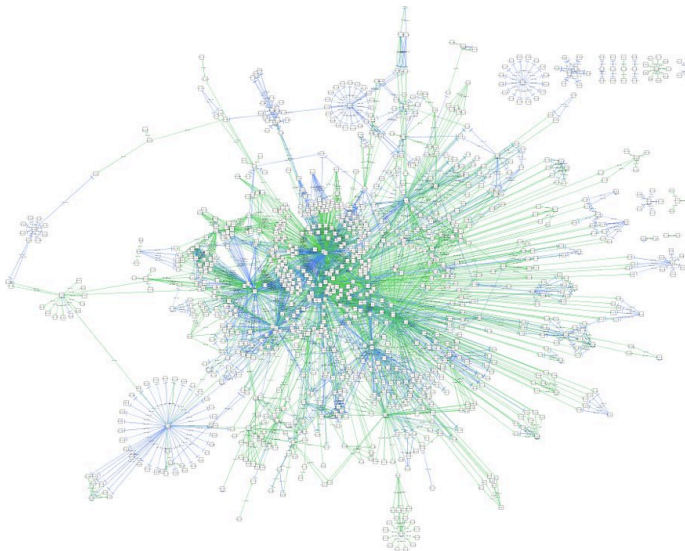
(1) Université de Nice, (2) INRA Jouy-en-Josas, (3) Université Paris Sud

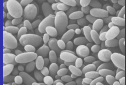
Vienna 2008.





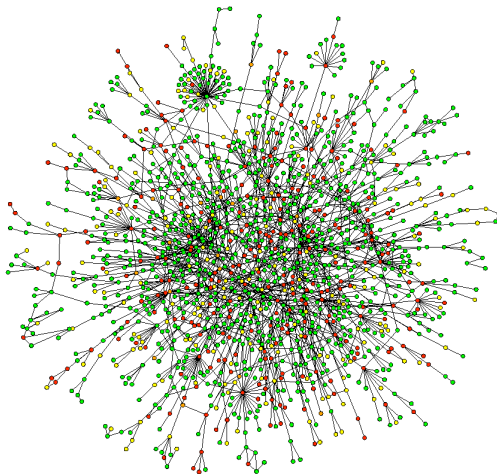
Gene - gene regulation network of *E. coli*



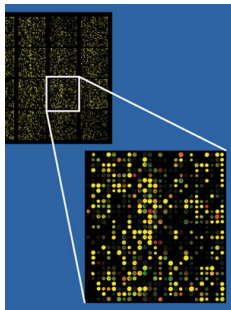


Protein - protein network of *S. cerevisiae*

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



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).

Descriptive tools:

- Kernel methods (supervised learning)

Model based tools:

- Bayesian Networks
- 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

an edge $i \stackrel{\mathbf{g}}{\sim} j$ between the genes i and j
iff
 $X^{(i)}$ and $X^{(j)}$ are not independent given $\{X^{(k)}, k \neq i, j\}$

regulation network \longleftrightarrow graph \mathbf{g}

The task of the statistician

Goal: estimate \mathbf{g} from a sample X_1, \dots, X_n .

Main difficulty: $n \ll p$

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

New algorithms: based on thresholding or regularization

- many of them have quite disappointing numerical performances (Villers *et al.* 2008)
- no theoretical results or in an asymptotic framework (with strong hypotheses on the covariance)

Estimation by model selection

Partial correlations

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

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

Skeleton of θ : we have $\theta_i^{(j)} = \frac{\text{Cov}(X^{(i)}, X^{(j)} | X^{(k)}, k \neq i, j)}{\text{Var}(X^{(i)} | X^{(k)}, k \neq j)}$ so

$$\theta_i^{(j)} \neq 0 \iff i \stackrel{g}{\sim} j$$

Goal: Estimate θ from a sample X_1, \dots, X_n with quality criterion

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

Partial correlations

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

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

Skeleton of θ : we have $\theta_i^{(j)} = \frac{\text{Cov}(X^{(i)}, X^{(j)} | X^{(k)}, k \neq i, j)}{\text{Var}(X^{(i)} | X^{(k)}, k \neq j)}$ so

$$\theta_i^{(j)} \neq 0 \iff i \underset{g}{\sim} j$$

Goal: Estimate θ from a sample X_1, \dots, X_n with quality criterion

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

Partial correlations

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

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

Skeleton of θ : we have $\theta_i^{(j)} = \frac{\text{Cov}(X^{(i)}, X^{(j)} | X^{(k)}, k \neq i, j)}{\text{Var}(X^{(i)} | X^{(k)}, k \neq j)}$ so

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

Goal: Estimate θ from a sample X_1, \dots, X_n with quality criterion

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

Partial correlations

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

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

Skeleton of θ : we have $\theta_i^{(j)} = \frac{\text{Cov}(X^{(i)}, X^{(j)} | X^{(k)}, k \neq i, j)}{\text{Var}(X^{(i)} | X^{(k)}, k \neq j)}$ so

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

Goal: Estimate θ from a sample X_1, \dots, X_n with quality criterion

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

Estimation procedure

- 1 Choose a collection \mathcal{G} of candidate graphs

e.g. all the graphs with p vertices and degree $\leq D$,

- 2 Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$

$$\hat{\theta}_g = \operatorname{argmin}_{A \sim g} \|X(I - A)\|_{n \times p}^2 \quad (\text{empirical MSEP})$$

- 3 Select one $\hat{\theta}_{\hat{g}}$ by minimizing a penalized empirical risk

with a criterion inspired by that in Baraud *et al.*

Estimation procedure

- 1 Choose a collection \mathcal{G} of candidate graphs

e.g. all the graphs with p vertices and degree $\leq D$,

- 2 Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$

$$\hat{\theta}_g = \operatorname{argmin}_{A \sim g} \|X(I - A)\|_{n \times p}^2 \quad (\text{empirical MSE})$$

- 3 Select one $\hat{\theta}_{\hat{g}}$ by minimizing a penalized empirical risk

with a criterion inspired by that in Baraud *et al.*

Estimation procedure

- 1 Choose a collection \mathcal{G} of candidate graphs

e.g. all the graphs with p vertices and degree $\leq D$,

- 2 Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$

$$\hat{\theta}_g = \operatorname{argmin}_{A \sim g} \|X(I - A)\|_{n \times p}^2 \quad (\text{empirical MSEP})$$

- 3 Select one $\hat{\theta}_{\hat{g}}$ by minimizing a penalized empirical risk

with a criterion inspired by that in Baraud *et al.*

Estimation procedure

- 1 Choose a collection \mathcal{G} of candidate graphs

e.g. all the graphs with p vertices and degree $\leq D$,

- 2 Associate to each graph $g \in \mathcal{G}$ an estimator $\hat{\theta}_g$

$$\hat{\theta}_g = \operatorname{argmin}_{A \sim g} \|X(I - A)\|_{n \times p}^2 \quad (\text{empirical MSEP})$$

- 3 Select one $\hat{\theta}_{\hat{g}}$ by minimizing a penalized empirical risk

with a criterion inspired by that in Baraud *et al.*

Theorem: risk bound.

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

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

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

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

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

Theorem: risk bound.

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

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

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

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

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

Theorem: risk bound.

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

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

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

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

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

Theorem: risk bound.

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

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

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

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

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

Theorem: risk bound.

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

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

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

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

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

Theory

Condition on the degree

How far can we trust the empirical MSEP?

Prediction error:

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

Proposition: From empirical to population MSEP

Under the previous condition on the degree, we 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 $\hat{\theta} \in \bigcup_{g \in \mathcal{G}} \Theta_g$.

How far can we trust the empirical MSEP?

Prediction error:

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

Proposition: From empirical to population MSEP

Under the previous condition on the degree, we 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 $\hat{\theta} \in \bigcup_{g \in \mathcal{G}} \Theta_g$.

Lemma: Restricted Inf / Sup of Random Matrices

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

Then for any $x > 0$

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

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

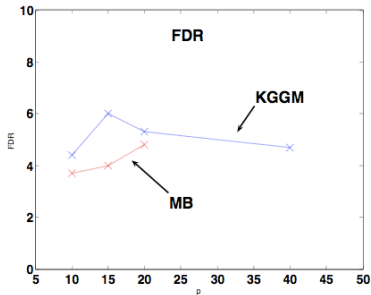
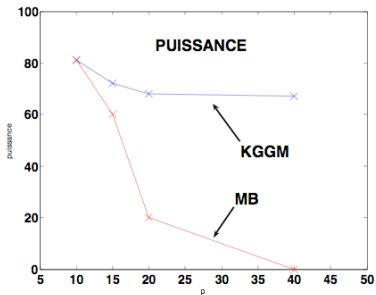
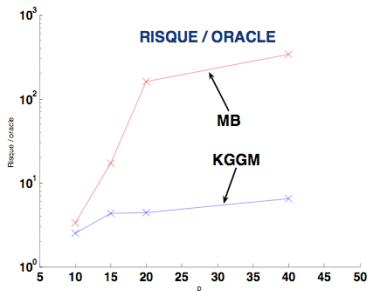
$$(1 - \delta) \|C^{1/2}(I - \hat{\theta})\| \leq \frac{1}{\sqrt{n}} \|X(I - \hat{\theta})\| \leq (1 + \delta) \|C^{1/2}(I - \hat{\theta})\|$$

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

In practice

Numerical performance

Random graphs, $n = 15$ and p increases



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 $n \times p^{\deg(\mathcal{G})+1}$

\implies cannot be used in practice when $p > 50 \dots$

Ongoing work: with S. Huet and N. Verzelen

Reduction of the size of the collection of graph, using data-driven collections.

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 $n \times p^{\deg(\mathcal{G})+1}$

\implies cannot be used in practice when $p > 50 \dots$

Ongoing work: with S. Huet and N. Verzelen

Reduction of the size of the collection of graph, using data-driven collections.

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 $n \times p^{\deg(\mathcal{G})+1}$

\implies cannot be used in practice when $p > 50 \dots$

Ongoing work: with S. Huet and N. Verzelen

Reduction of the size of the collection of graph, using data-driven collections.

Main reference of the talk

C. Giraud. *Estimation of Gaussian graphs by model selection*.
Electronic Journal of Statistics. Vol. 2 (2008) pp. 542–563

Related references

- 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*.
Personal communication