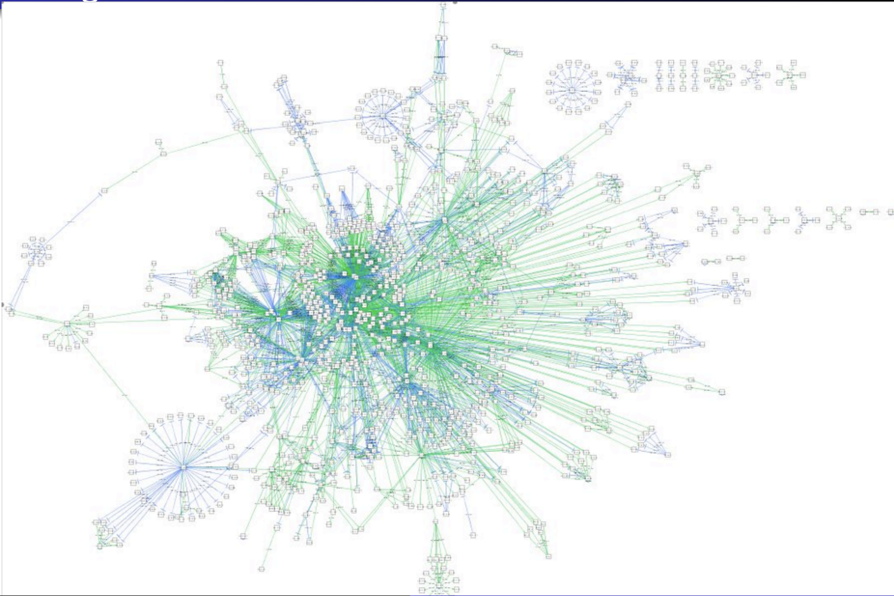
Inferring Biological Regulation Networks

Christophe Giraud

Université de Nice
et
INRA Jouy-en-Josas

(On the estimation of Gaussian graphs)

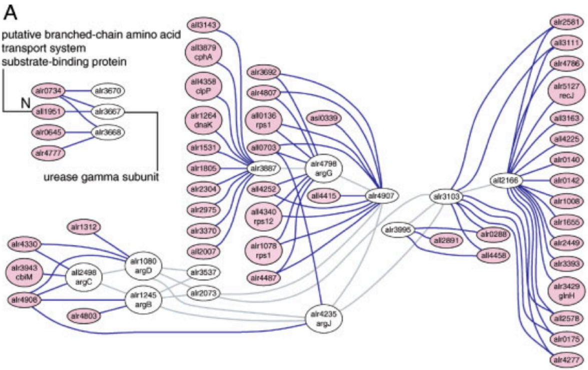
Gene regulation network of *E. coli*



Proteomic regulation network of a yeast

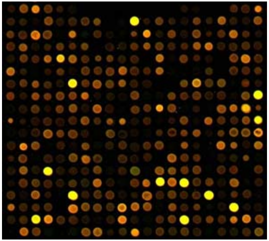
1458 proteins (nodes) and their 1948 interactions (edges)

Gene regulation network for *Anabaena* (nitrogen deprivation)



Goal

Detect such a regulation network from microarray experiments



Available data:

- $p \approx$ a few 100 or 1000 proteins or genes
- $n \approx$ a few 10 microarray

Some tools

- **Kernels methods:** supervised learning (need some learning) - can handle data of different nature
- **Bayesian network:** integration of *a priori* knowledge
- **Gaussian graphical models:** no need of *a priori* knowledge - purely explanatory

Estimation of Gaussian Graphs.

Theory.

http://hal.archives-ouvertes.fr/hal-00178275/fr/

Model: Gaussian Graphs

Statistical model: The gene expression levels $(X^{(1)}, \ldots, X^{(p)})$ are distributed according to $\mathcal{N}(0, C)$ 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

$$\text{and} \quad \mathbb{E}\left(X^{(j)} \mid X^{(k)}, k \neq j\right) = \sum_{k \neq j} \theta_k^{(j)} X^{(k)}.$$

Remark: $\theta_k^{(j)} = -\left(C^{-1}\right)_{k,j} / \left(C^{-1}\right)_{j,j}$ so

$$\theta_k^{(j)} \neq 0 \iff \theta_j^{(k)} \neq 0.$$

A few References

Multiple testing	L^1 **procedures**
- 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)
...	- Friedman *et al.* (2007)
	...

Goal

Goal: Estimate θ from a sample $X_1, \ldots, X_n$ with $n < p$, with quality criterion

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

Notations: For a matrix $A \in \mathbb{R}^{k\times q}$

- $A = [A^{(1)}, \ldots, A^{(q)}]$
- $\|A\|^2_{k\times q} = \sum_{i=1}^{k}\sum_{j=1}^{q}\left(A_i^{(j)}\right)^2$

Theory

Estimation strategy

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 θ.
3. Select one of these estimators by minimizing some penalized empirical risk.

Collection of candidate graphs

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

Examples

- Set of the graphs with p nodes of degree $\leq D$,
- Set of the graphs with p nodes including a known graph g_o.

Model for θ associated to $g \in \mathcal{G}$:

$$g \curvearrowright \Theta_g = \left\{\theta \in \mathbb{R}^{p\times p} : \ i \not\leftrightarrow j \ \Rightarrow \ \theta_i^{(j)} = 0\right\}$$

Estimator $\hat{\theta}_g$ associated to g

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

where Θ =span of matrices with 0 on the 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:

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

Which estimator shall we choose among the $\{\hat{\theta}_g,\ g \in \mathcal{G}\}$?

Risk: if $d_j = \deg(j) = \#\{i :\ i \leftrightarrow j\}$

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

$$= \sum_{j=1}^{p} \frac{n-1}{n-d_j-1}\left[\|C^{1/2}(\theta^{(j)} - \theta_g^{(j)})\|^2 + \frac{d_j}{C_{jj}^{-1}(n-1)}\right] \quad (1)$$

$$\approx \|C^{1/2}(\theta - \theta_g)\|^2 + \sum_{j=1}^{p} \frac{\deg(j)}{nC_{jj}^{-1}}$$

Oracle: the ideal would be to choose $\hat{\theta}_{g_*}$ minimizing (1).

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

où $d_j = \deg(j)$

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

où $d_j = \deg(j)$

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

où $d_j = \deg(j)$

Questions:

1. Which penalty $\mathrm{pen}(\cdot)$ shall we use?
2. Which "size" of graph can we hope to estimate?

Selection criterion: penalized empirical risk

We choose $\hat{\theta} = \hat{\theta}_{\hat{g}}$, where $\hat{g}$ minimizes over $\mathcal{G}$

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

où $d_j = \deg(j)$

Questions:

1. Which penalty $\mathrm{pen}(\cdot)$ shall we use?
2. Which "size" of graph can we hope to estimate?

Theory

Maximum "size" of the graph?

Which "size" of graph can we hope to estimate?

Risque:

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

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

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

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

Complexity of a graph

Different notions of complexity for a graph: number of edges, degree, exponent of the number of cycles, fractal dimension, etc

Natural notion here: the degree of g

$$\deg(g) = \max \{\deg(j),\ j = 1, \ldots, p\}.$$

Proposition: control of the empirical risk

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

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

and if $\delta > \sqrt{\eta}$,

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

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

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

Note: $\log(p)$ can be replaced by $\log(p/\deg(\mathcal{G}))$ in the box

Lemma: Restricted quasi-isometry

Let Z be a $n \times p$ (with $n \leq p$) matrix with i.i.d. $\mathcal{N}(0,1)$ entries. Consider any collection $V_1, \ldots, V_N$ of subspaces of $\mathbb{R}^p$ with dimension $d < n$. Then for any $x > 0$

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

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

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

Conversely...

For $C = I$, there exists some constant $c(\delta) > 0$ such that if

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

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.

Which penalty?

The minimal size of $\mathrm{pen}(d)$?

Selection criterion:

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

In the simple case where $\theta = 0$ (viz $C = I$) we would like that the selected graph $\hat{g}$ stay of "small" size.

Minimal penalty

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

The choosen penalty

Notation: we write EDkhi(d, N, x) for 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 random variable with d and N degrees of freedom.

Penalty: For $K > 1$ we set

$$\mathrm{pen}(d) = K\,\frac{n-d}{n-d-1}\,\mathrm{EDkhi}\left[d+1, n-d-1, \left(C_{p-1}^{d}(d+1)^2\right)^{-1}\right].$$

Size of the penalty

When

$$\deg(\mathcal{G}) \leq \eta \, \frac{n}{2\left(1.1 + \sqrt{\log p}\right)^2}.$$

we have

$$\mathrm{pen}(d) \lesssim K \left(1 + e^{\eta}\sqrt{2\log p}\right)^2 (d+1).$$

Theory

Performance

Modified estimator $\tilde{\theta}$

To handle any family $\mathcal{G}$ of graph we need to slightly modify $\hat{\theta}$.

Modification of $\hat{\theta}$: we set $\tilde{\theta}$ for

$$\tilde{\theta}^{(j)} = \hat{\theta}^{(j)}\, \mathbf{1}_{\{\|\hat{\theta}^{(j)}\| \leq \sqrt{p}\, T_n\}}, \text{ for } j \in \{1, \ldots, p\}, \quad \text{with} \quad T_n = n^{2 \log n}.$$

Theorem: non-asymptotic control of the risk

When $\mathcal{G}$ fulfills the condition

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

the risk of the estimator $\tilde{\theta}$ is bounded by

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

where $R_n(\eta, C) = O(p^2 n^{-4\log n})$.

Theorem: non-asymptotic control of the risk

When $\mathcal{G}$ fulfills the condition

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

the risk of the estimator $\tilde{\theta}$ is bounded by

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

where $R_n(\eta, C) = O(p^2 n^{-4\log n})$.

Theorem: non-asymptotic control of the risk

When $\mathcal{G}$ fulfills the condition

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

the risk of the estimator $\tilde{\theta}$ is bounded by

$$\mathbb{E}\left(\|C^{1/2}(\theta-\tilde{\theta})\|^2\right)$$

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

where $R_n(\eta, C) = O(p^2 n^{-4\log n})$.

Theorem: non-asymptotic control of the risk

When $\mathcal{G}$ fulfills the condition

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

the risk of the estimator $\tilde{\theta}$ is bounded by

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

where $R_n(\eta, C) = O(p^2 n^{-4 \log n})$.

Theorem: non-asymptotic control of the risk

When $\mathcal{G}$ fulfills the condition

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

the risk of the estimator $\tilde{\theta}$ is bounded by

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

where $R_n(\eta, C) = O(p^2 n^{-4\log n})$.

Numerical simulations

- Simulation of "Erdos-Reny" graphs, with random covariance matrices,
- $n = 15$ observations,
- p from 10 to 40,
- from 2.5% to 33% of edges,
- SNR more or less high,
- comparaison to SINful and L^1 (Meinschausen & Bühlmann).

$n = 15$, $p = 10$, edges $= 10\% - 20\%$, good SNR

Sparse graphs (10%)

	$K = 2$	L^1 "or"	L^1 "and"	SIN(0.05)	SIN(0.25)
risk/oracle	2.4	2.2	3000	15.10^3	12.10^3
Power	84%	86%	67%	41%	46%
FDR	4.6%	3%	1%	1.2%	6.4%

More connected graphs (20%)

	$K = 2$	L^1 "or"	L^1 "and"	SIN(0.05)	SIN(0.25)
risk/oracle	4.1	5.4	2000	8900	7900
Power	49%	44%	28%	7%	11%
FDR	7.7%	4.4%	1.8%	1.5%	7.2%

$n = 15$, $p = 10$, edges$= 20\%$, poor SNR

poor SNR

	$K = 2$	L^1 "or"	L^1 "and"	SIN(0.05)	SIN(0.25)
risk/oracle	5.8	12	11	13	12
Power	30%	7.9%	7%	3.8%	8.4%
FDR	5.5%	0.6%	0.5%	3.5%	7.9%

$n = 15$, $p = 10$, edges$= 30 - 33\%$

edges= 30%

	$K = 2$	L^1 "or"
risk/oracle	4.3	6.8
Power	24%	15%
FDR	7%	3.2%

edges= 33%

	$K = 2$	L^1 "or"
risk/oracle	4.5	6.5
Power	14%	4.5%
FDR	5.8%	1.3%

With a larger p and "less" edges

$n = 15$, $p = 15$ and 10% of edges

	$K = 2$	L^1 "ou"
risk/oracle	5.9	36
Power	57%	14%
FDR	5.7%	1%

$n = 15$, $p = 40$ and 2.5% of edges

	$K = 2$	L^1 "ou"
risk/oracle	4.1	330
Puissance	77%	0.0%
FDR	4.1%	0.0%

Estimation of gaussian graphs.

In pratice.

Reduce p

To avoid a poor estimation of a graph with degree d we need

$$\frac{2d(1+\log(p/d))}{n} < 1$$

viz $\boxed{p < de^{n/(2d)-1}.}$

To reduce p:

- Restrict to the genes that are differentially expressed and have a "high" variance
- Group genes with the same expression profile and search the interaction between the groups

Reduce the numerical computational complexity

- Reduce p...
- Compute approximate solutions.
- First select "candidate" edges with (e.g.) (adaptive-)LARS and apply the selection procedure to these edges

In progress...