Estimation and tests for Gaussian graphical models

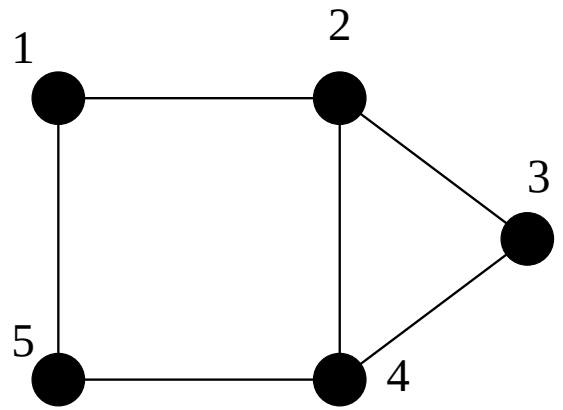
Nicolas Verzelen

Workshop on random graphs

Undirected graphical model

We consider $Z = (Z_1, \ldots, Z_{p+1}) \sim \mathcal{N}_{p+1}(0, \Omega^{-1})$

Ω non singular.
$\Gamma := \{1, \ldots, p+1\}$
$G = (\Gamma, E)$ finite undirected graph.
$\mathrm{ne}_G(a)$: neighbors of a in G.



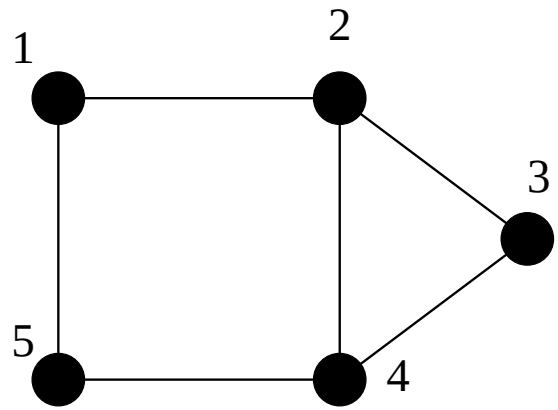
Undirected graphical model

We consider $Z = (Z_1, \ldots, Z_{p+1}) \sim \mathcal{N}_{p+1}(0, \Omega^{-1})$

Ω non singular.
$\Gamma := \{1, \ldots, p+1\}$
$G = (\Gamma, E)$ finite undirected graph.
$\mathrm{ne}_G(a)$: neighbors of a in G.

Z satisfies the Markov local property at a with respect to G if

$$(Z_a \perp\!\!\!\perp Z_{-\{a, \mathrm{ne}_G(a)\}}) | Z_{\mathrm{ne}_G(a)}$$



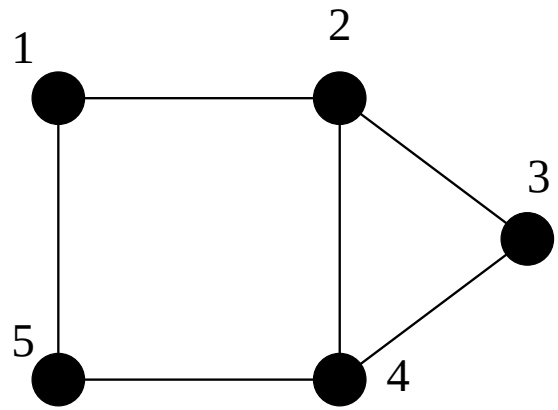
Undirected graphical model

We consider $Z = (Z_1, \ldots, Z_{p+1}) \sim \mathcal{N}_{p+1}(0, \Omega^{-1})$

Ω non singular.
$\Gamma := \{1, \ldots, p+1\}$
$G = (\Gamma, E)$ finite undirected graph.
$\mathrm{ne}_G(a)$: neighbors of a in G.

Z satisfies the Markov local property at a with respect to G if

$$(Z_a \perp\!\!\!\perp Z_{-\{a, \mathrm{ne}_G(a)\}}) | Z_{\mathrm{ne}_G(a)}$$

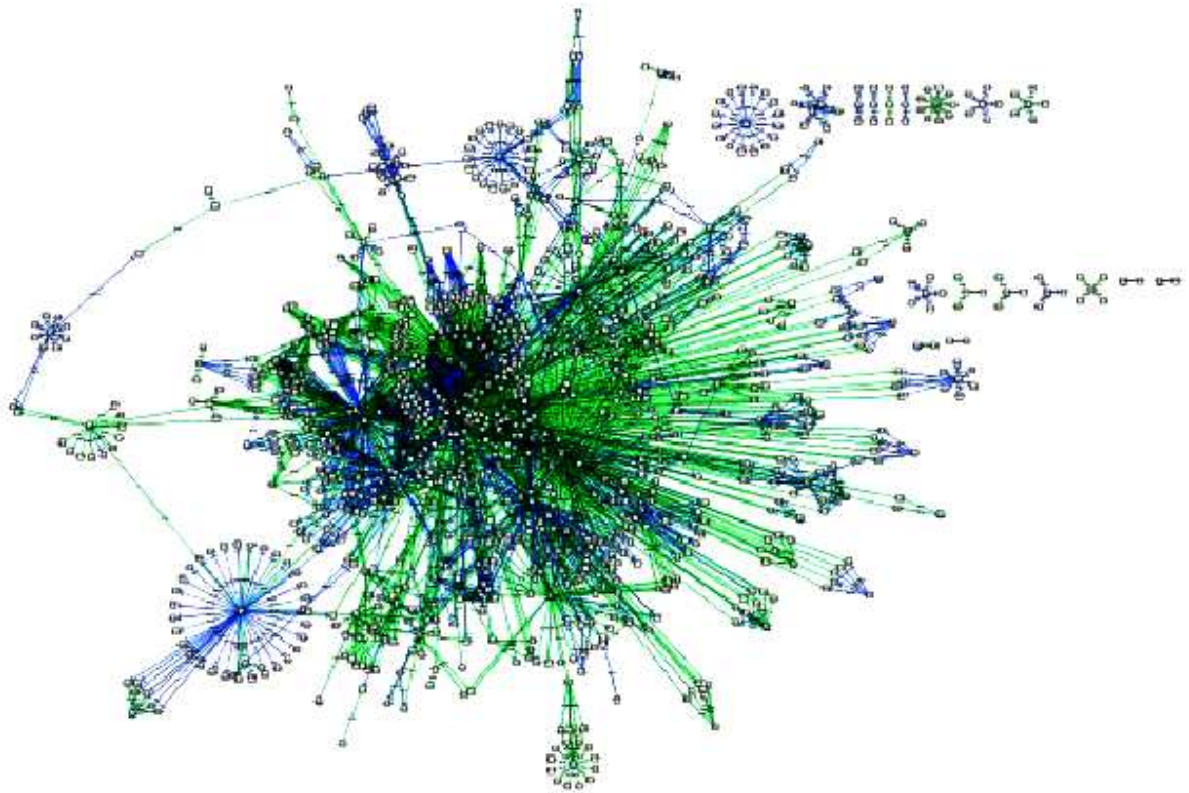


Z is a Gaussian graphical model with respect to G

$\Longleftrightarrow$

Z satisfies the local Markov property for any $a \in \Gamma$.

Ex. 1 : Genetic network of E. Coli

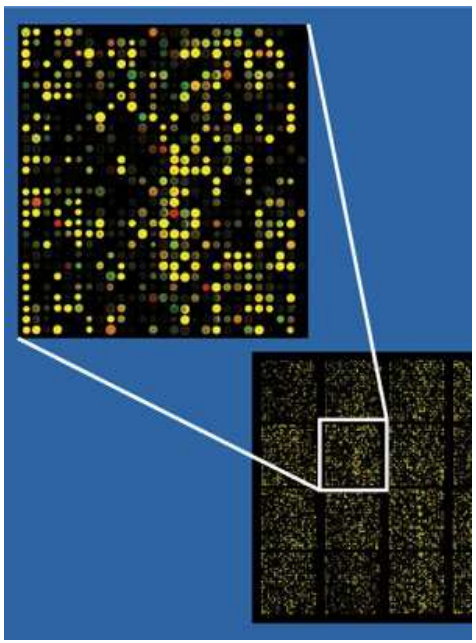


Nodes : genes

Vertices : interactions between two genes and their products that

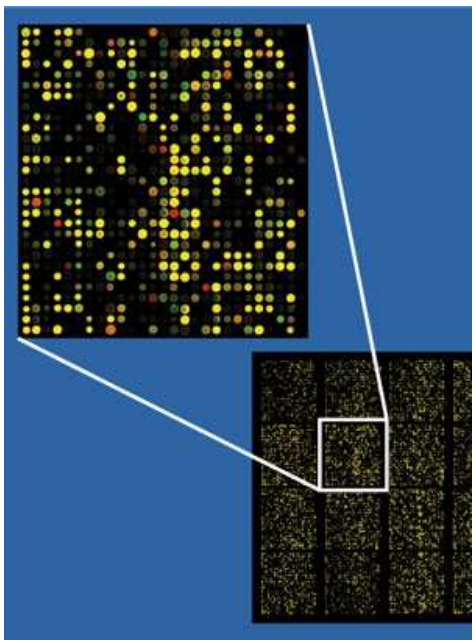
Transcriptomic data analysis

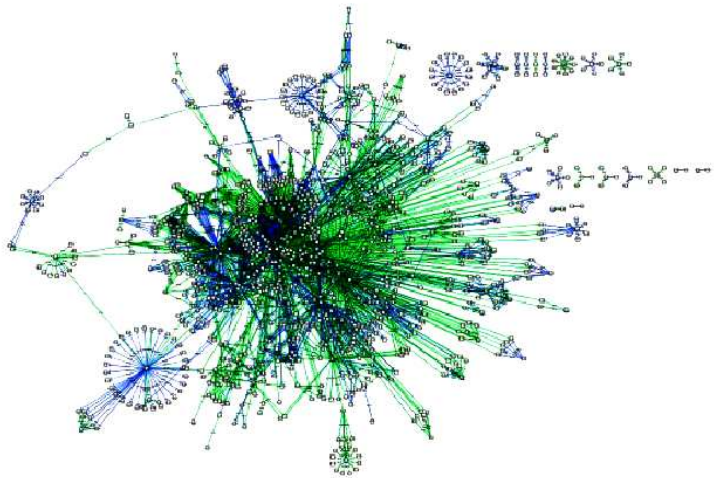
Transcriptomic data = measures of the gene expression levels (RNAm)



Transcriptomic data analysis

Transcriptomic data = measures of the gene expression levels (RNAm)



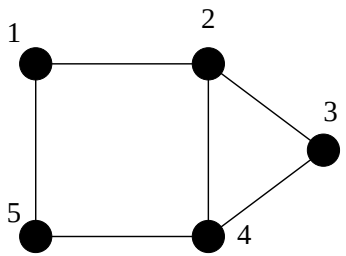


Analysis of the dependence structure in the data.

Goal : Inferring a part of this gene network using transcriptomic data.

Property

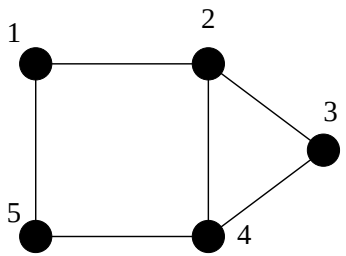
$$\Omega_{a,b} = 0 \iff (Z_a \perp\!\!\!\perp Z_b) | Z_{-\{a,b\}}.$$



	1	2	3	4	5
1	*	*	0	0	*
2	*	*	*	*	0
3	0	*	*	*	0
4	0	*	*	*	*
5	*	0	0	*	*

Property

$$\Omega_{a,b} = 0 \iff (Z_a \perp\!\!\!\perp Z_b) | Z_{-\{a,b\}}.$$

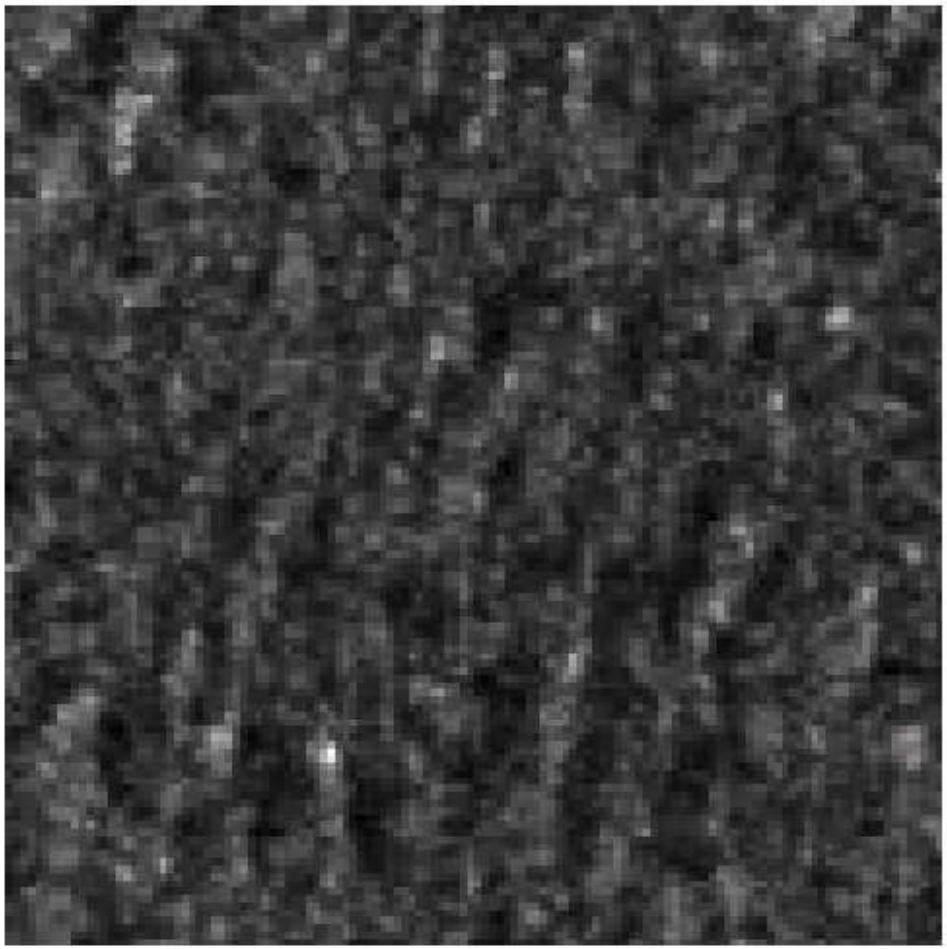


	1	2	3	4	5
1	*	*	0	0	*
2	*	*	*	*	0
3	0	*	*	*	0
4	0	*	*	*	*
5	*	0	0	*	*

Two benefits :

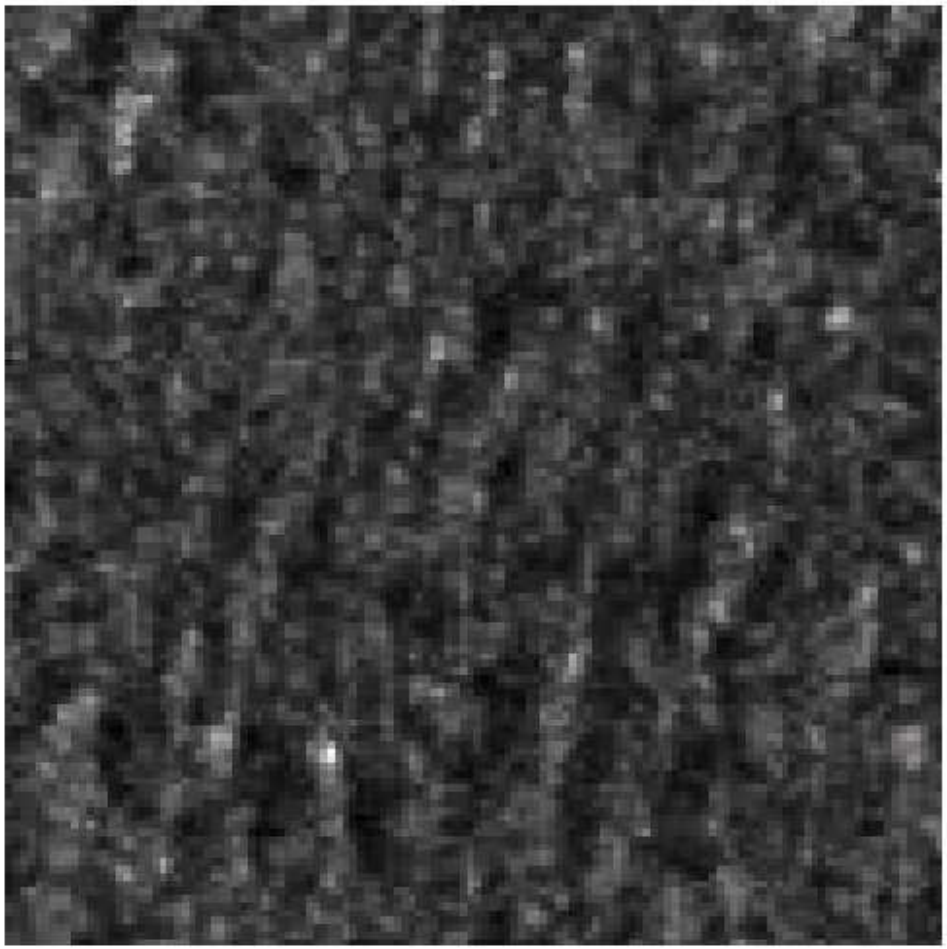
- scientific interpretation of the graph.
- dimension reduction in a statistical analysis.

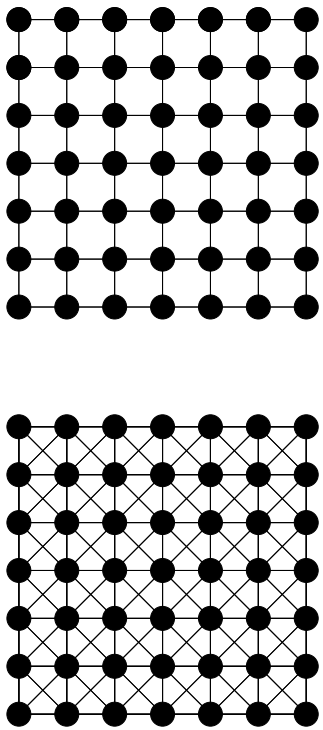
Ex. 2 : Image analysis



Rue and Tjelmeland (2002)

Ex. 2 : Image analysis





Rue and Tjelmeland (2002)

Statistics

Data : $Z \sim \mathcal{N}_{p+1}(0, \Omega^{-1})$.
n observations of $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{p+1})$

Statistics

Data : $Z \sim \mathcal{N}_{p+1}(0, \Omega^{-1})$.
n observations of $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{p+1})$

QUESTIONS :

1. Estimation or test of the structure.
2. Estimation of the distribution

Statistics

Data : $Z \sim \mathcal{N}_{p+1}(0, \Omega^{-1})$.
n observations of $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{p+1})$

QUESTIONS :

1. Estimation or test of the structure.
2. Estimation of the distribution

SETTING :

- High dimension $p \gg n$ (e.g., microarray)
- Optimality and adaptation
- Flexibility
- computationally fast

Outline of the talk

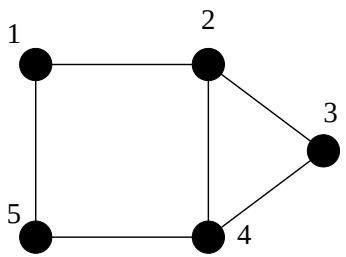
Graph estimation

Adaptive tests

Minimax bounds

Graph estimation : Precision matrix estimation

$$\Omega_{a,b} = 0 \iff (Z_a \perp\!\!\!\perp Z_b) | Z_{-\{a,b\}}.$$



	1	2	3	4	5
1	*	*	0	0	*
2	*	*	*	*	0
3	0	*	*	*	0
4	0	*	*	*	*
5	*	0	0	*	*

Graph Estimation $\iff$ Selecting the 0 of the precision matrix. la précision.

$\Rightarrow$ Estimation of the precision matrix by penalized maximum likelihood :

Penalized maximum likelihood

Ex1 : Complexity penalization.

Idea : For any graph G', $\mathcal{A}_{G'} = \{\Omega' : \Omega'_{ij} = 0 \text{ for } i \overset{G'}{\nsim} j\}$.

$$\widehat{\Omega}_{G'} = \arg \max_{\Omega' \in \mathcal{A}_{G'}} \mathcal{L}_n(\Omega')$$

Remarks : When the number of edges in G' increases, then

- the likelihood $\mathcal{L}_n(\widehat{\Omega}_{G'})$ always increases.
- the bias of $\widehat{\Omega}_{G'}$ decreases.
- the variance of $\widehat{\Omega}_{G'}$ increases.

Penalized maximum likelihood

Ex1 : Complexity penalization.

Idea : For any graph G', $\mathcal{A}_{G'} = \{\Omega' : \Omega'_{ij} = 0 \text{ for } i \overset{G'}{\nsim} j\}$.

$$\widehat{\Omega}_{G'} = \arg \max_{\Omega' \in \mathcal{A}_{G'}} \mathcal{L}_n(\Omega')$$

Remarks : When the number of edges in G' increases, then

- the likelihood $\mathcal{L}_n(\widehat{\Omega}_{G'})$ always increases.
- the bias of $\widehat{\Omega}_{G'}$ decreases.
- the variance of $\widehat{\Omega}_{G'}$ increases.

There exists (at least) one graph G^* that minimizes the Kullback risk

$$\mathbb{E}\left[\mathcal{K}(\widehat{\Omega}_{G'}; \Omega)\right] \approx \text{ bias } + \text{ variance } .$$

Penalized maximum likelihood

Ex1 : Complexity penalization.

Idea : For any graph G', $\mathcal{A}_{G'} = \{\Omega' : \Omega'_{ij} = 0 \text{ for } i \overset{G'}{\nsim} j\}$.

$$\widehat{\Omega}_{G'} = \arg \max_{\Omega' \in \mathcal{A}_{G'}} \mathcal{L}_n(\Omega')$$

Remarks : When the number of edges in G' increases, then

- the likelihood $\mathcal{L}_n(\widehat{\Omega}_{G'})$ always increases.
- the bias of $\widehat{\Omega}_{G'}$ decreases.
- the variance of $\widehat{\Omega}_{G'}$ increases.

There exists (at least) one graph G^* that minimizes the Kullback risk

$$\mathbb{E}\left[\mathcal{K}(\widehat{\Omega}_{G'}; \Omega)\right] \approx \text{ bias } + \text{ variance } .$$

We do not know G^* !

If n is large enough, then $G^* = G$.
If n is too small, G^* has less edge than G. Hopeless to estimate G.
A reasonnable goal is too aim for G^*.

Goal : Picking a graph $\widehat{G}$ that performs almost as well as G^*.

Heuristic : data-driven approximation of the Kullback risk

$$\mathbb{E}\left[\mathcal{K}(\widehat{\Omega}_{G'};\Omega)\right] \approx -\mathcal{L}_n(\widehat{\Omega}_G) + \frac{\square}{n}\left[\sum_{a=1}^{p} deg_{G'}(a)\right] .$$

Data-driven criterion :

$$\widehat{G} \quad = \quad \arg\min_{G'} -\mathcal{L}_n(\widehat{\Omega}_G) + K\frac{\log(p)}{n}\left[\sum_{a=1}^{p} deg_G(a)\right] ,$$

for some $K > 0$.

Analysis : Simultaneous control of the deviations of the estimators $\widehat{\Omega}_{G'}$
$\rightsquigarrow$ computing the metric entropy associated to the distributions in $\mathcal{A}_{G'}$.
$\rightsquigarrow$ Under some conditions....

$$\mathbb{E}\left[\mathcal{K}(\widehat{\Omega}_{\widehat{G}};\Omega)\right] \leq \square \log(p)\mathbb{E}\left[\mathcal{K}(\widehat{\Omega}_{G^*};\Omega)\right]$$

Computational complexity : $2^{p(p-1)/2}$ estimators to compute.

Convexifying the criterion

Ex2 : l_1 penalization (Glasso) [Banerjee et al. (07), Rothman et al. (08), Ravikumar et al. (09)]

$$\widehat{\Omega} = \arg\min_{\Omega'} -\mathcal{L}_n(\Omega') + \lambda \|\Omega'\|_1 \ .$$

The estimator $\widehat{\Omega}$ is usually sparse. l_1 penalization shrinks some coefficients to zero.

Convexifying the criterion

Ex2 : l_1 penalization (Glasso) [Banerjee et al. (07), Rothman et al. (08), Ravikumar et al. (09)]

$$\widehat{\Omega} = \arg\min_{\Omega'} -\mathcal{L}_n(\Omega') + \lambda\|\Omega'\|_1 \ .$$

The estimator $\widehat{\Omega}$ is usually sparse. l_1 penalization shrinks some coefficients to zero.

Analysis : the Kullback loss $\mathcal{K}(\widehat{\Omega};\Omega)$ is small ... under some conditions :

$$\mathbb{E}\left[\mathcal{K}(\widehat{\Omega};\Omega)\right] \leq \square \frac{\log(p)}{n}\left[\sum_{a=1}^{p} deg_G(a)\right]$$

$\rightsquigarrow$ tools : Behaviour of the spectrum of a Wishart matrix [Davidson and Szarek (01)] + deviation of the entries of a standard Wishart matrix.

Convexifying the criterion

Ex2 : l_1 penalization (Glasso) [Banerjee et al. (07), Rothman et al. (08), Ravikumar et al. (09)]

$$\widehat{\Omega} = \arg\min_{\Omega'} -\mathcal{L}_n(\Omega') + \lambda \|\Omega'\|_1 \ .$$

The estimator $\widehat{\Omega}$ is usually sparse. l_1 penalization shrinks some coefficients to zero.

Analysis : the Kullback loss $\mathcal{K}(\widehat{\Omega}; \Omega)$ is small ... under some conditions :

$$\mathbb{E}\left[\mathcal{K}(\widehat{\Omega}; \Omega)\right] \leq \square \frac{\log(p)}{n} \left[\sum_{a=1}^{p} deg_G(a)\right]$$

$\rightsquigarrow$ tools : Behaviour of the spectrum of a Wishart matrix [Davidson and Szarek (01)] + deviation of the entries of a standard Wishart matrix.

When n is small, the graph $\tilde{G}$ associated to $\widehat{\Omega}$ (empirically) performs very bad.

Graph estimation : conditional regression

$$Z_a = \sum_{b \neq a} \theta_{a,b} Z_b + \epsilon_a \ ,$$

with $\epsilon_a \perp\!\!\!\perp (X_b)_{b \neq a}$ and the matrix θ is defined by

$$\theta_{a,b} = -\Omega_{a,b}/\Omega_{a,a} \ .$$

Graph Estimation $\Longleftrightarrow$ Selection of the 0 of θ.

$\Rightarrow$ Estimation in a linear regression model with Gaussian design :

Graph estimation : conditional regression

$$Z_a = \sum_{b \neq a} \theta_{a,b} Z_b + \epsilon_a \ ,$$

with $\epsilon_a \perp\!\!\!\perp (X_b)_{b \neq a}$ and the matrix θ is defined by

$$\theta_{a,b} = -\Omega_{a,b}/\Omega_{a,a} \ .$$

Graph Estimation $\Longleftrightarrow$ Selection of the 0 of θ.

$\Rightarrow$ Estimation in a linear regression model with Gaussian design :

Ex1 : Complexity penalization

$$\widehat{\theta}_{a,.} = \arg\min_{\theta'_{a,.}} \|\mathbf{Z}_a - \sum_{b \neq a} \theta'_{a,b} \mathbf{Z}_b\|^2 \left(1 + K \frac{\log(p)}{n} \|\theta'_{a,.}\|_0\right) \ .$$

Graph estimation : conditional regression

$$Z_a = \sum_{b \neq a} \theta_{a,b} Z_b + \epsilon_a \ ,$$

with $\epsilon_a \perp\!\!\!\perp (X_b)_{b \neq a}$ and the matrix θ is defined by

$$\theta_{a,b} = -\Omega_{a,b}/\Omega_{a,a} \ .$$

Graph Estimation $\Longleftrightarrow$ Selection of the 0 of θ.

$\Rightarrow$ Estimation in a linear regression model with Gaussian design :

Ex1 : Complexity penalization

$$\widehat{\theta}_{a,.} = \arg\min_{\theta'_{a,.}} \|\mathbf{Z}_a - \sum_{b \neq a} \theta'_{a,b} \mathbf{Z}_b\|^2 \left(1 + K \frac{\log(p)}{n} \|\theta'_{a,.}\|_0\right) \ .$$

Ex2 : l_1 Penalization (Lasso)

$$\widehat{\theta}_{a,.} = \arg\min_{\theta'_{a,.}} \|\mathbf{Z}_a - \sum_{b \neq a} \theta'_{a,b} \mathbf{Z}_b\|^2 + \lambda \|\theta'_{a,.}\|_1 \ .$$

Analysis : (Gaussian) concentration + (a bit of) concentration for Wishart matrices.
[Giraud (08), Rothman et al. (08)]

Graph estimation : Bayesian approach

1. Define a prior distribution on the set of Graphs (e.g. Erdös-Rényi model)
2. For any graph G, Define a prior distribution on the set of precision matrices with prescribed zeros (typically : HyperWishart distribution).

$\rightsquigarrow$ Compute the posterior distribution of $\mathcal{L}(G, \Omega | Z_1, \ldots, Z_n)$.

Graph estimation : Bayesian approach

1. Define a prior distribution on the set of Graphs (e.g. Erdös-Rényi model)
2. For any graph G, Define a prior distribution on the set of precision matrices with prescribed zeros (typically : HyperWishart distribution).

$\rightsquigarrow$ Compute the posterior distribution of $\mathcal{L}(G, \Omega | Z_1, \ldots, Z_n)$.

Computational difficulties :

- If the graph G is chordal, then the HyperWishart distribution is conjugate easy computation of $\mathcal{L}(\Omega | G, Z_1, \ldots, Z_n)$
- If G is not chordal, one has to use MCMC-like algorithms :$\rightsquigarrow$ computationally intensive. [Dellaportas et al. (03), Scott and Carvahlho (09)]

Active research topic

tests multiples	**Pseudo-vraisemblance**	**Vraisemblance**
- Schäfer/Strimmer (04) - Wille/Bühlmann (06) - Bühlmann/Kalisch (08) ...	- Meinshausen/Bühlmann (06) - Giraud/Huet/V. (09) ...	- Yuan/Lin (06) - Banerjee *et al.* (07) - Friedman *et al.* (07) ...

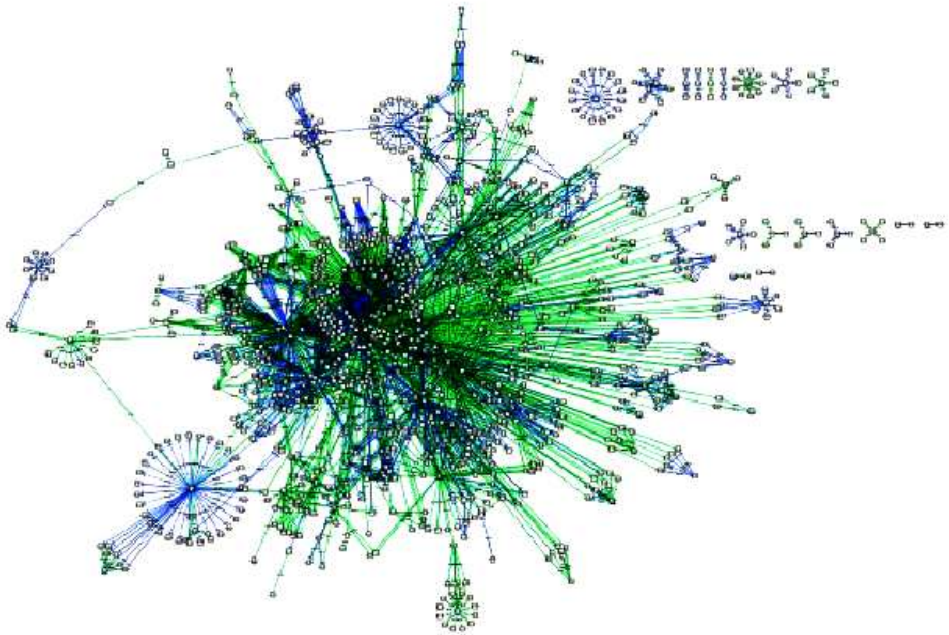
Active research topic

tests multiples	**Pseudo-vraisemblance**	**Vraisemblance**
- Schäfer/Strimmer (04) - Wille/Bühlmann (06) - Bühlmann/Kalisch (08) ...	- Meinshausen/Bühlmann (06) - Giraud/Huet/V. (09) ...	- Yuan/Lin (06) - Banerjee *et al.* (07) - Friedman *et al.* (07) ...

Common features :

- "often" algorithmic approaches.
- A few theoretical results $1 \ll n \ll p$
  + hypotheses on the covariance Ω^{-1}.
- Practical performance are (sometimes) disapointing and the results are not corroborating. $\rightsquigarrow$ [Villers *et al.* (08)]

Validation of a graph



Expression data : $p+1$ genes
n observations of $(\mathbf{Z}_1, \ldots, \mathbf{Z}_{p+1})$

GOAL : testing that no regulation between genes has been forgotten.

Validation of a graph

Graph : $G = (\Gamma, E)$ with $\Gamma = \{1, \dots, p+1\}$

DATA : n observations of $\mathbf{Z} = (\mathbf{Z_1}, \dots, \mathbf{Z_{p+1}}) \sim \mathcal{N}(0, \Omega^{-1})$

GOAL : testing

$$H_0 : \text{« } Z \text{ is Gaussian graphical model with respect to } G \text{. »}$$

Means : Testing for any vertex a :

$$H_{0,a} : \text{« } Z \text{ satisfies the local Markov property at } a \text{ with respect to } G \text{ »}$$
$$\Updownarrow$$
$$Z_a \perp\!\!\!\perp Z_{-\{a, \mathrm{ne}_G(a)\}} | Z_{\mathrm{ne}_G(a)}$$

Approach

Regression of Z_a with respecto Z_{-a} : $Z_a = \sum_{b \in \Gamma \setminus \{a\}} \theta_b^a Z_b + \epsilon_a$, $\epsilon_a \perp\!\!\!\perp Z_{-a}$

Neighborhood test for the vectex a : testing

$$H_{0,a} : \text{« } Z_a \perp\!\!\!\perp Z_{-\{a,\mathrm{ne}_G(a)\}} | Z_{\mathrm{ne}_G(a)} \text{ »}$$

$$\Longleftrightarrow \qquad H_{0,a} : \text{« } \theta^a_{-ne_G(a)} = 0 \text{ »}$$

Approach

Regression of Z_a with respecto Z_{-a} : $Z_a = \sum_{b \in \Gamma \setminus \{a\}} \theta_b^a Z_b + \epsilon_a$, $\epsilon_a \perp\!\!\!\perp Z_{-a}$

Neighborhood test for the vectex a : testing

$$H_{0,a} : \text{« } Z_a \perp\!\!\!\perp Z_{-\{a, \mathrm{ne}_G(a)\}} | Z_{\mathrm{ne}_G(a)} \text{ »}$$

$$\Longleftrightarrow \qquad H_{0,a} : \text{« } \theta^a_{-ne_G(a)} = 0 \text{ »}$$

$$Y = \sum_{i=1}^{p} \theta_i X_i + \epsilon$$

with :

- $\theta \in \mathbb{R}^p$ unknown
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = \mathrm{var}(Y|X)$ unknown
- $(X_i)_{1 \leq i \leq p} \sim \mathcal{N}(0, \Sigma)$ with Σ unknown but non singular.
- ϵ independent of X

Approach

Regression of Z_a with respecto Z_{-a} : $Z_a = \sum_{b \in \Gamma \setminus \{a\}} \theta_b^a Z_b + \epsilon_a$, $\epsilon_a \perp\!\!\!\perp Z_{-a}$

Neighborhood test for the vectex a : testing

$$H_{0,a} : \text{« } Z_a \perp\!\!\!\perp Z_{-\{a, \mathrm{ne}_G(a)\}} | Z_{\mathrm{ne}_G(a)} \text{ »}$$

$$\iff \qquad H_{0,a} : \text{« } \theta^a_{-ne_G(a)} = 0 \text{ »}$$

$$Y = \sum_{i=1}^{p} \theta_i X_i + \epsilon$$

with :

- $\theta \in \mathbb{R}^p$ unknown
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = \mathrm{var}(Y|X)$ unknown
- $(X_i)_{1 \leq i \leq p} \sim \mathcal{N}(0, \Sigma)$ with Σ unknown but non singular.
- ϵ independent of X

$V \subset \{1, \ldots, p\}$

Testing the support of θ :

data : n observations of $(\mathbf{Y}, \mathbf{X})$.

$$H_0 : \text{« } \theta_{-V} = 0 \text{ »}$$

Description of the procedure

$\rightsquigarrow$ For the sake of simplicity : $V = \emptyset$.

Consider $m \subset \{1, \ldots, p\}$

testing H_0 : « $\theta = 0$ ». against the alternative $H_{1,m}$: « $\theta_{-\{m\}} = 0$ »

Description of the procedure

$\rightsquigarrow$ For the sake of simplicity : $V = \emptyset$.

Consider $m \subset \{1, \dots, p\}$

testing H_0 : « $\theta = 0$ ». against the alternative $H_{1,m}$: « $\theta_{-\{m\}} = 0$ »

We note

- Π_m the orthogonal projection onto the space generated by $\mathbf{X}_i$, $i \in m$
- $\|.\|_n$: euclidean norm on $\mathbb{R}^n$.

Fisher test :

$$\phi_m(\mathbf{Y}, \mathbf{X}) = \frac{(n - |m|)\|\Pi_m \mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2}{|m|\|\mathbf{Y} - \Pi_m \mathbf{Y}\|_n^2}$$

We reject H_0 if $\phi_m(\mathbf{Y}, \mathbf{X}) >$ threshold

Under H_0, $\phi_m(\mathbf{Y}, \mathbf{X}) \sim Fisher(|m|, n - |m|)$

Description (fd)

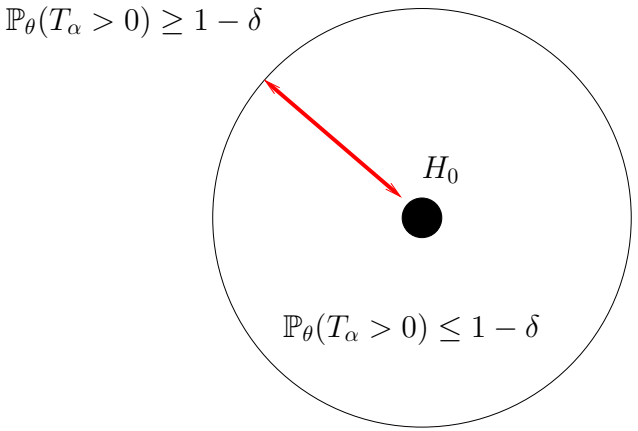
$\mathcal{M}(k,p)$: Collection of subset of $\{1,\ldots,p\}$ of size k. $(k \leq n-1)$

Definition

Let us reject the test T_α if

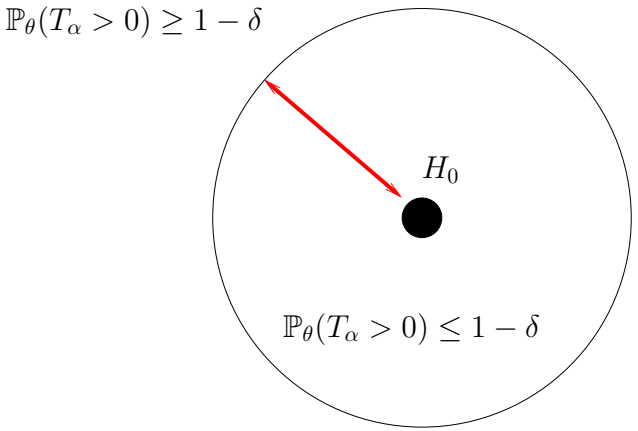
$$\exists m \in \mathcal{M}(k,p) : \phi_{\boldsymbol{m}}(\mathbf{Y},\mathbf{X}) > \bar{F}^{-1}_{|\boldsymbol{m}|,N_{\boldsymbol{m}}}(\alpha/|\mathcal{M}|)$$

$\bar{F}_{|\boldsymbol{m}|,N_{\boldsymbol{m}}}(u)$ is the probability that a Fisher distribution with $|m|$ et $N_{\boldsymbol{m}}$ degrees of freedom is larger than u.



$\Theta[k, p]$: Vectors $\theta \in \mathbb{R}^p$ with at most k non zero coordinates.
$H_0 : \theta = 0$ against $H_1 : \theta \in \Theta[k, p] \backslash \{0\}$

$\|\sqrt{\Sigma}\theta\|_p^2 := \text{var}[\sum_{i=1}^p \theta_i X_i]$



$\Theta[k, p]$: Vectors $\theta \in \mathbb{R}^p$ with at most k non zero coordinates.
$H_0 : \theta = 0$ against $H_1 : \theta \in \Theta[k, p] \backslash \{0\}$

$\|\sqrt{\Sigma}\theta\|_p^2 := \text{var}[\sum_{i=1}^p \theta_i X_i]$

Theorem

Assume that

$$n \geq \square[\alpha, \delta] k \log\left(\frac{p}{k}\right) \ ,$$

then $\mathbb{P}_\theta\left(T_\alpha > 0\right) \geq 1 - \delta$ *then for all* Σ *and for all* $\theta \in \Theta[k, p]$ *such that*

$$\frac{\|\sqrt{\Sigma}\theta\|_p^2}{\sigma^2} \geq \frac{\square[\alpha, \delta]}{n}\left[k \log\left(\frac{p}{k}\right)\right]$$

.

Summary

- The test T_α is powerful against a k-sparse alternative when

$$\frac{\|\sqrt{\Sigma}\theta\|_p^2}{\sigma^2} \geq \square \frac{k}{n} \log\left(\frac{p}{k}\right) .$$

Is it optimal ?

Summary

- The test T_α is powerful against a k-sparse alternative when

$$\frac{\|\sqrt{\Sigma}\theta\|_p^2}{\sigma^2} \geq \square \frac{k}{n} \log\left(\frac{p}{k}\right) .$$

  Is it optimal ?

- ... under the hypothesis that $k \log(p/k)$ is small compared to n.
  Why is this condition occuring ? Is it minimal ?

Summary

- The test T_α is powerful against a k-sparse alternative when

$$\frac{\|\sqrt{\Sigma}\theta\|_p^2}{\sigma^2} \geq \square \frac{k}{n} \log\left(\frac{p}{k}\right) .$$

  Is it optimal ?

- ... under the hypothesis that $k \log(p/k)$ is small compared to n.
  Why is this condition occuring ? Is it minimal ?
- It is possible to achieve the rate of testing $k \log\left(\frac{p}{k}\right)$ against $\Theta[k, p]$ simultaneously over all $k \in \{1, \ldots, n/2 \log(p)\}$.

Summary

- The test T_α is powerful against a k-sparse alternative when

$$\frac{\|\sqrt{\Sigma}\theta\|_p^2}{\sigma^2} \geq \square \frac{k}{n} \log\left(\frac{p}{k}\right) .$$

  Is it optimal ?

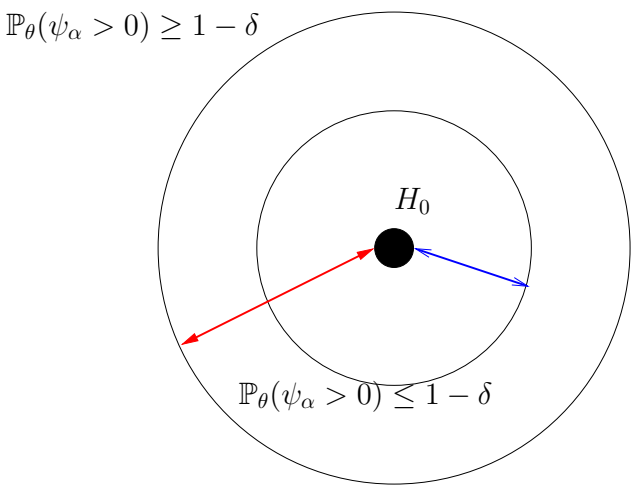
- ... under the hypothesis that $k \log(p/k)$ is small compared to n.
  Why is this condition occuring ? Is it minimal ?
- It is possible to achieve the rate of testing $k \log\left(\frac{p}{k}\right)$ against $\Theta[k, p]$ simultaneously over all $k \in \{1, \ldots, n/2 \log(p)\}$.
- The computational complexity of T_α is terrible. Proportionnal to $\binom{k}{p}$.
  How can we cope with this in practice ?.
  data-driven collection of models given by a fast procedure (e.g. : Lasso).

Minimax separation distance

$\mathbf{H_0}$: $\theta = 0$ against $\mathbf{H_1}$: $\theta \in \Theta[k, p] \setminus \{0\}$.
Fix $\delta > 0$. ψ_α test of Level α.

Separation distance of ψ_α :

$$\rho[\psi_\alpha, k, \Sigma] := \inf \left\{ \rho > 0, \inf_{\theta \in \Theta[k,p],\ \|\sqrt{\Sigma}\theta\|_p \geq \rho\sigma} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1 - \delta \right\} .$$

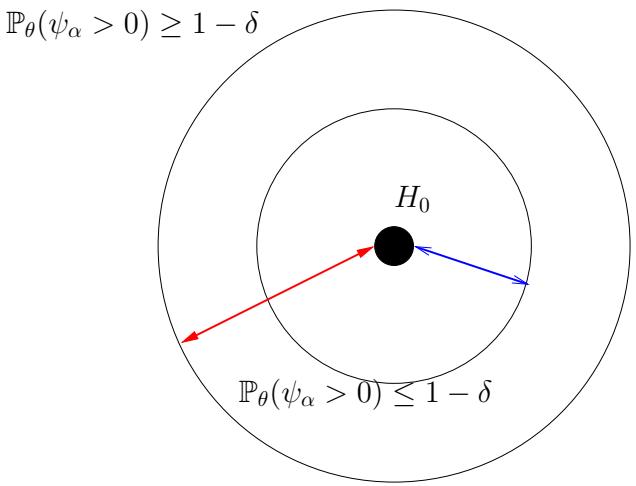


Minimax separation distance

$\mathbf{H_0}$: $\theta = 0$ against $\mathbf{H_1}$: $\theta \in \Theta[k, p] \setminus \{0\}$.
Fix $\delta > 0$. ψ_α test of Level α.

Separation distance of ψ_α :

$$\rho[\psi_\alpha, k, \Sigma] := \inf \left\{ \rho > 0, \inf_{\theta \in \Theta[k,p],\ \|\sqrt{\Sigma}\theta\|_p \geq \rho\sigma} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1 - \delta \right\} .$$



Minimax distance of separation

$$\rho^*[k, \Sigma] := \inf_{\psi_\alpha} \rho[\psi_\alpha, k, \Sigma] .$$

$$\rho^*[k] := \sup_{\mathbf{X}} \rho^*[k, \Sigma]$$

Known variance σ^2

If the support of θ is known $\rightsquigarrow$ square of the parametric separation distance $\sqrt{k}/n$.

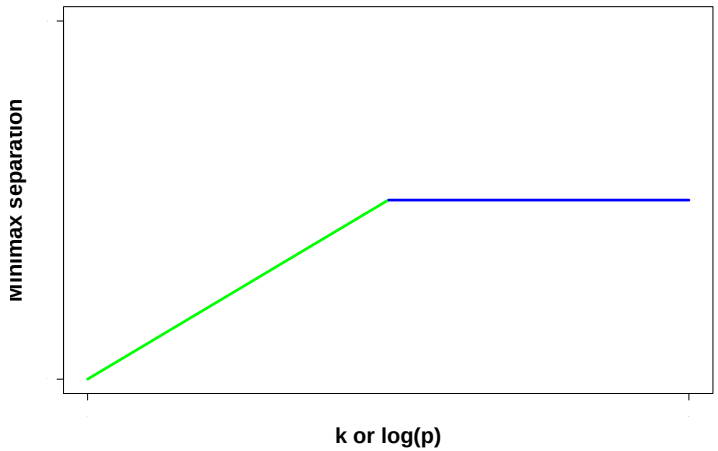
Known variance σ^2

If the support of θ is known $\rightsquigarrow$ square of the parametric separation distance $\sqrt{k}/n$.

Theorem

As long as $p \geq n \geq \square(\alpha, \delta)$ and $k \leq p^{1/3}$, we have

$$(\rho^*[k])^2 \simeq \square[\alpha, \delta] \left[\frac{k}{n} \log\left(\frac{p}{k}\right) \wedge \frac{1}{\sqrt{n}} \right] .$$



Comments :

- If $k \log(ep/k)$ is small before $\sqrt{n}$, analogous to minimax risk for prediction. Analogous to Gaussian sequence model (*Ingster (02), Baraud (02)*).
- Large $(k, p) \Rightarrow$ parametric separation distance over $\mathbb{R}^n$.
- Adaptation to the sparsity is possible. (Bonferroni multiple testing procedure).

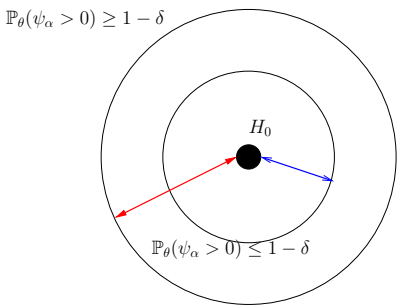
Unknown variance σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Separation distance when the variance is unknown.

Unknown variance σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Separation distance when the variance is unknown.

$$\rho_U[\psi_\alpha, k, \Sigma] := \inf \left\{ \rho > 0, \quad \inf_{\substack{\sigma>0,\ \theta\in\Theta[k,p], \\ \|\sqrt{\Sigma}\theta\|_n \geq \rho\sigma}} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1-\delta \right\} .$$



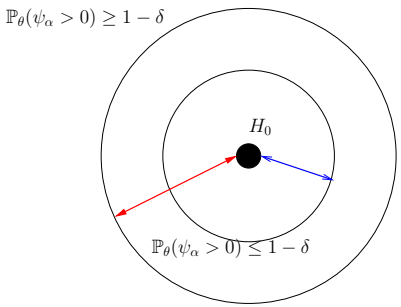
$$\rho^*_U[k, \Sigma] := \inf_{\psi_\alpha} \rho_U[\psi_\alpha, k, \Sigma] .$$

$$\rho^*_U[k] := \sup_{\mathbf{X}} \rho^*_U[k, \Sigma]$$

Unknown variance σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Separation distance when the variance is unknown.

$$\rho_U[\psi_\alpha, k, \Sigma] := \inf \left\{ \rho > 0, \quad \inf_{\substack{\sigma>0,\ \theta\in\Theta[k,p], \\ \|\sqrt{\Sigma}\theta\|_n \geq \rho\sigma}} \mathbb{P}_{\theta,\sigma}[\psi_\alpha = 1] \geq 1-\delta \right\} .$$



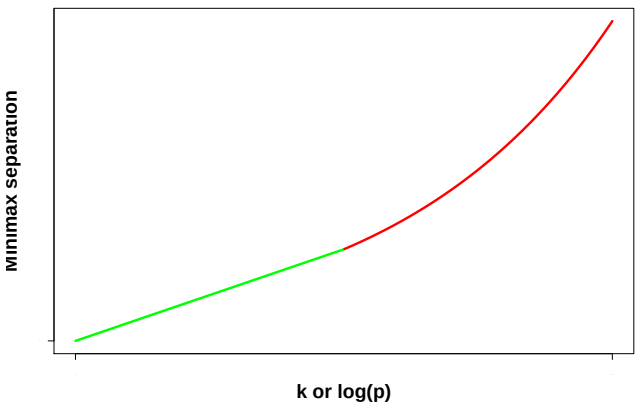
$$\rho^*_U[k, \Sigma] := \inf_{\psi_\alpha} \rho_U[\psi_\alpha, k, \Sigma] .$$

$$\rho^*_U[k] := \sup_{\mathbf{x}} \rho^*_U[k, \Sigma]$$

Theorem

If $p \geq n \geq \square(\alpha, \delta)$ and $k \leq p^{1/3}$, we have

$$(\rho^*_U[k])^2 \simeq \square[\alpha, \delta] \frac{k}{n} \log\left(\frac{p}{k}\right) \exp\left[\square[\alpha, \delta] \frac{k \log(ep/k)}{n}\right] .$$



Comments :

- If $k \log(ep/k) \leq \sqrt{n}$, same minimax separation distance as for known variance.
- Blow up in ultra-high dimension.

Statistical arguments $\Rightarrow$ geometrical results

X design matrix of size $n \times p$. Sparse eigenvalues

$$\Phi_{k,+}(\mathbf{X}) = \sup_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} \qquad \Phi_{k,-}(\mathbf{X}) = \inf_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2}$$

Statistical model : $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(0_n, \sigma^2 I_n)$. θ is sparse in some way.

Statistical arguments $\Rightarrow$ geometrical results

X design matrix of size $n \times p$. Sparse eigenvalues

$$\Phi_{k,+}(\mathbf{X}) = \sup_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} \qquad \Phi_{k,-}(\mathbf{X}) = \inf_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2}$$

Statistical model : $\mathbf{Y} = \mathbf{X}\theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. θ is sparse in some way.

Proposition

For any design $\mathbf{X}$,

$$n\sigma^2 \geq \inf_{\widehat{\theta}} \sup_{\theta : \|\theta\|_0 = k} \mathbb{E}\left[\|\mathbf{X}(\widehat{\theta} - \theta)\|^2\right] \geq \square \frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})} k \log\left(\frac{p}{k}\right) \sigma^2$$

Proof : Fano's lemma

Statistical arguments $\Rightarrow$ geometrical results

X design matrix of size $n \times p$. Sparse eigenvalues

$$\Phi_{k,+}(\mathbf{X}) = \sup_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2} \qquad \Phi_{k,-}(\mathbf{X}) = \inf_{\theta, \|\theta\|_0 \leq k} \frac{\|\mathbf{X}\theta\|_n^2}{\|\theta\|_p^2}$$

Statistical model : $\mathbf{Y} = \mathbf{X}\theta + \epsilon$ with $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$. θ is sparse in some way.

Proposition

For any design $\mathbf{X}$,

$$n\sigma^2 \geq \inf_{\widehat{\theta}} \sup_{\theta : \|\theta\|_0 = k} \mathbb{E}\left[\|\mathbf{X}(\widehat{\theta} - \theta)\|^2\right] \geq \square \frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})} k \log\left(\frac{p}{k}\right) \sigma^2$$

Proof : Fano's lemma

Corollary

No design $\mathbf{X}$ *satisfies* $\frac{\Phi_{2k,-}(\mathbf{X})}{\Phi_{2k,+}(\mathbf{X})}$ *is close to* 1 *if* $k \log(p/k) \gtrsim n$.

(Baraniuk et al. 2008).

Geometrical arguments $\Rightarrow$ statistical results

$D_{n,p}$: Design matrices whose columns have been normalized to 1

Proposition

For any design $\mathbf{X} \in \mathcal{D}_{n,p}$

$$\Phi_{2k,-}(\mathbf{X}) \leq Ck^2 \exp\left[-\frac{2k}{n}\log\left(\frac{p}{k}\right)\right] \vee 1 \ .$$

Geometrical arguments $\Rightarrow$ statistical results

$D_{n,p}$: Design matrices whose columns have been normalized to 1

Proposition

For any design $\mathbf{X} \in \mathcal{D}_{n,p}$

$$\Phi_{2k,-}(\mathbf{X}) \leq Ck^2 \exp\left[-\frac{2k}{n}\log\left(\frac{p}{k}\right)\right] \vee 1 \ .$$

Corollary (Minimax lower bound for estimation)

If $k\log(p/k) \gg n\log(n)$. *For any design* $\mathbf{X} \in \mathcal{D}_{n,p}$,

$$\inf_{\widehat{\theta}} \sup_{\theta:\|\theta\|_0=k} \mathbb{E}\left[\|\widehat{\theta}-\theta\|^2\right] \geq \square \exp\left[\square' \frac{k}{n}\log(\frac{p}{k})\right]\sigma^2$$

References

Graphical models

- Lauritzen. Graphical models. *Oxford university Press*.

Graph estimation

- Meinshausen and Bühlmann (2006). High-dimensional graphs and Variable Selection with the Lasso. *Annals of Statistics.*
- Rothman, Levina et al. (2008). Sparse permutant invariant covariance estimation. *Electronical Journal of Statistics.*
- Ambroise, Chiquet, and Matias (2009). Inferring Sparse Gaussian graphical models with latent structure. *Electronical Journal of Statistics.*
- Giraud, Huet, V.(2009). Graph selection with GGMSelect.
- Wong, Carter, and Kohn (2003) Efficient estimation of covariance selection models. *Biometrika*

Tests and Minimax rates

- V. and Villers (2010). Goodness-of-fit Tests for high-dimensional Gaussian linear models. *Annals of Statistics*.
- V. (2010). Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons. `http://arxiv.org/abs/1008.0526v2`