

# Quelle taille minimale d'échantillon pour analyser des données microarray?

## Journées Statistiques du Sud

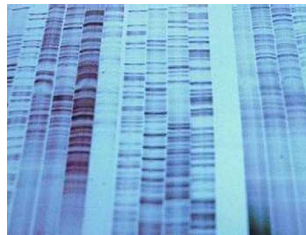
Nicolas Verzelen



23 Juin 2010

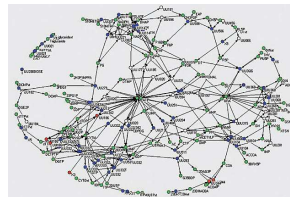
## De la génomique à la postgénomique

**Génomique** : Séquencage de différents organismes (génome humain  $\sim 04$ )

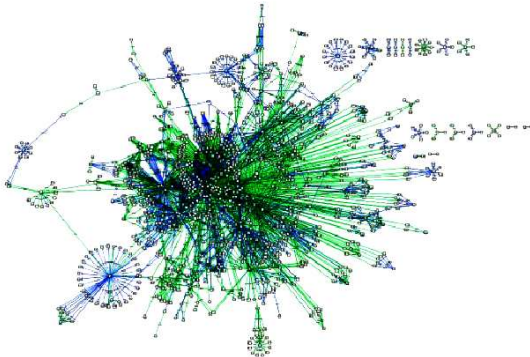


**Postgénomique** : Comprendre comment cela fonctionne

- Jeux de données massifs
- Inclut biologie, physique, informatique, mathématique



## Gènes de E. Coli



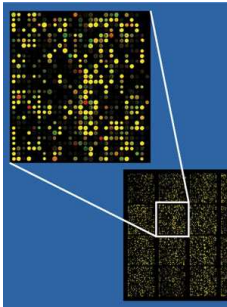
Intéactions complexes entre les gènes et leurs produits régissant l'activité de la cellule.

**Démarche** : Inférer une partie du réseau de gènes à partir de données transcriptomiques.

Proposer des **pistes** d'exploration concernant les interactions entre gènes.

# Analyse de données transcriptomiques

**Données transcriptomiques** = mesure du niveau d'expression des gènes (abondance en ARN messenger)



- **Analyse différentielle** des données obtenues dans différentes conditions : avec ou sans *deletion* d'un gène, avec ou sans stress (chaleur, chimique, biotique, etc...)
- **Analyse des structures de dépendance** dans les données récoltées (exploitation de toute la masse des données).

Réseaux biologiques

Modèles graphiques gaussiens

Risques minimax

Design fixe/design aléatoire

Conclusion

## Modèle graphique non orienté (Markov Random fields)

On considère  $X = (X_1, \dots, X_p) \sim \mathcal{N}_p(0, \Omega^{-1})$

$\Omega$  inversible

$\Gamma := \{1, \dots, p\}$

$g = (\Gamma, E)$  graphe non orienté

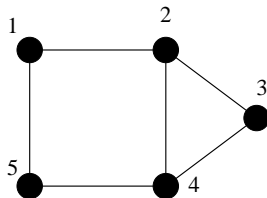
## Modèle graphique non orienté (Markov Random fields)

On considère  $X = (X_1, \dots, X_p) \sim \mathcal{N}_p(0, \Omega^{-1})$

$\Omega$  inversible

$\Gamma := \{1, \dots, p\}$

$g = (\Gamma, E)$  graphe non orienté



$X$  est un **modèle graphique gaussien** par rapport à  $g$  si pour tout sommet  $a$

$X_a$  **indépendant de**  $\{X_b : b \not\sim a\}$  conditionnellement à  $\{X_b : b \sim a\}$

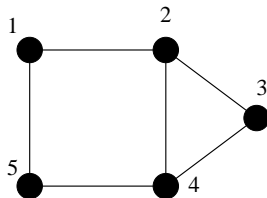
## Modèle graphique non orienté (Markov Random fields)

On considère  $X = (X_1, \dots, X_p) \sim \mathcal{N}_p(0, \Omega^{-1})$

$\Omega$  inversible

$\Gamma := \{1, \dots, p\}$

$g = (\Gamma, E)$  graphe non orienté



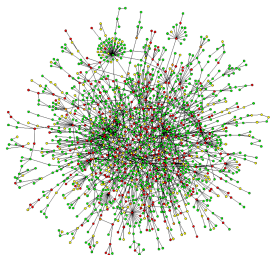
$X$  est un **modèle graphique gaussien** par rapport à  $g$  si pour tout sommet  $a$   
 $X_a$  **indépendant de**  $\{X_b : b \not\sim a\}$  conditionnellement à  $\{X_b : b \sim a\}$

**Unicité** du graphe minimal qui représente les dépendances conditionnelles.

**GGM** : modèle graphique gaussien.

Pas de notion d'**orientation** ou de **causalité** (problème difficile)

## Estimation de graphe



**Modélisation** : Les niveaux d'expression sont modélisés à l'aide d'un GGM de graphe  $\mathbf{g}$  inconnu. (le réseau de gène)

**Objectif** : estimer à partir des données transcriptomiques le graphe  $\mathbf{g}$  du GGM.

**Difficulté principale** :  $n \ll p$

- $p \approx 100$  à plusieurs 1000 gènes.
- $n \approx$  quelques 10.

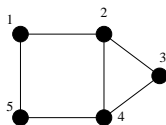
**Formulation statistique** :

$n$  observations de  $X$  de loi  $\mathcal{N}_p(0, \Omega^{-1})$  ( $\Omega$  inconnu).

Estimation de  $\mathbf{g}$ .

## Propriété de la précision

$$\Omega_{a,b} = 0 \iff (X_a \perp\!\!\!\perp X_b) | X_{-\{a,b\}}.$$



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

**Estimation** du graphe  $\iff$  Sélection des 0 de la précision.

$\Rightarrow$  Estimation de la précision par **maximum de vraisemblance pénalisé** :

**Ex1** : Pénalisation par complexité.

$$\hat{\Omega} = \arg \min_{\Omega'} -\mathcal{L}_n(\Omega') + \text{pen}[\|\Omega'\|_0] .$$

**Ex2** : Pénalisation  $l_1$  (Glasso)

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

## Régression conditionnelle

$$\mathbf{X}_a = \sum_{b \neq a} \theta_{a,b} \mathbf{X}_b + \epsilon_a ,$$

avec  $\epsilon_a \perp\!\!\!\perp (\mathbf{X}_b)_{b \neq a}$  et matrice  $\theta$  définie par

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

**Estimation** du graphe  $\iff$  Sélection des 0 de  $\theta$ .

$\Rightarrow$  Estimation dans modèle de régression linéaire à design gaussien :

**Ex1** : Pénalisation par complexité.

$$\hat{\theta}_{a,.} = \arg \min_{\theta'_{a,.}} \|\mathbf{X}_a - \sum_{b \neq a} \theta'_{a,b} \mathbf{X}_b\|^2 (1 + \text{pen}[\|\theta'_{a,.}\|_0]) .$$

**Ex2** : Pénalisation  $l_1$  (Lasso)

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

## Sujet de recherche actif

**Nouveaux algorithmes** : par seuillage ou par régularisation

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

## Sujet de recherche actif

**Nouveaux algorithmes** : par seuillage ou par régularisation

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

**Caractéristiques** :

- approches “souvent” algorithmiques.
- quelques résultats théoriques lorsque  $1 \ll n \ll p$   
+ **hypothèses** sur la matrice de covariance  $\Omega^{-1}$ .
- Performances pratiques parfois décevantes (ex : vraisemblance) et résultats non concordants.  $\rightsquigarrow$  [Villers *et al.* (08)]

## Limites de l'estimation de réseau par GGM

- 1 **Biais** de l'expérimentateur et normalisation des données.
- 2 Expériences pas toujours **indépendantes** (ex : séries temporelles)
- 3 Expériences **différentes** (ex : situations de stress, témoins)  
Les réseaux sont ils-différents ?  $\sim$  *étude de lois de mélange de GGM.*
- 4 **Limites structurelles liées à la grande dimension ?** ( $p \gg n$ )

## Limites de l'estimation de réseau par GGM

- ④ Limites structurelles liées à la grande dimension? ( $p \gg n$ )

Quelles performances peut-on espérer?

$p$  donné, quel  $n$  minimal pour estimer le graphe?

## Régression linéaire à design gaussien

$$Y = \sum_{a=1}^p \theta_a Z_a + \epsilon$$

avec

- $\theta \in \mathbb{R}^p$  inconnu.
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$  et  $\sigma^2 = \text{var}(Y|Z)$  inconnu.
- $(Z_i)_{1 \leq i \leq p} \sim \mathcal{N}(0, \Sigma)$  avec  $\Sigma$  inconnu mais non singulier.
- $\epsilon$  indépendant de  $Z$

DONNÉES :  $n$ -échantillon  $(\mathbf{Y}, \mathbf{Z}) \rightsquigarrow \mathbf{Y} = \mathbf{Z}\theta + \epsilon$

OBJECTIF :

- Estimation du support de  $\theta$  basé sur  $n$ -échantillon.
- Estimation de  $\theta$  (problème inverse)
- Estimation de  $Z\theta$  (**prédiction**)

## Régression linéaire à design gaussien

$$Y = \sum_{a=1}^p \theta_a Z_a + \epsilon$$

avec

- $\theta \in \mathbb{R}^p$  inconnu.
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$  et  $\sigma^2 = \text{var}(Y|Z)$  inconnu.
- $(Z_i)_{1 \leq i \leq p} \sim \mathcal{N}(0, \Sigma)$  avec  $\Sigma$  inconnu mais non singulier.
- $\epsilon$  indépendant de  $Z$

DONNÉES :  $n$ -échantillon  $(\mathbf{Y}, \mathbf{Z}) \rightsquigarrow \mathbf{Y} = \mathbf{Z}\theta + \epsilon$

OBJECTIF :

- Estimation du support de  $\theta$  basé sur  $n$ -échantillon.
- Estimation de  $\theta$  (problème inverse)
- Estimation de  $Z\theta$  (**prédiction**)

*Difficultés :*

- design aléatoire
- variance inconnue  $\sigma^2$ .

## Estimation par moindres carrés

OBJECTIF : Déterminer les **vitesse**s optimales en prédiction :

- 1 Etudier les performance d'une "*bonne*" procédure
- 2 Montrer qu'on ne peut faire mieux

## Estimation par moindres carrés

OBJECTIF : Déterminer les **vitesse optimale en prédiction** :

- 1 Etudier les performance d'une "*bonne*" procédure
- 2 Montrer qu'on ne peut faire mieux

Soit  $m \subset \{1, \dots, p\}$  et  $|m| < n$ ,

$$\hat{\theta}_m := \arg \min_{\theta_{-m}=0} \|\mathbf{Y} - \sum_{a=1}^p \theta'_a \mathbf{Z}_a\|_n^2 = (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1} \mathbf{Z}_m \mathbf{Y}$$

**Perte des moindres carrés intégrés** :  $\|\sqrt{\Sigma}(\theta' - \theta)\|_p^2 := \mathbb{E} [\sum_{a=1}^p (\theta'_a - \theta_a) Z_i]^2$ .

## Estimation par moindres carrés

OBJECTIF : Déterminer les **vitesse optimale en prédiction** :

- 1 Etudier la performance d'une "bonne" procédure
- 2 Montrer qu'on ne peut faire mieux

Soit  $m \subset \{1, \dots, p\}$  et  $|m| < n$ ,

$$\hat{\theta}_m := \arg \min_{\theta_{-m}=0} \|\mathbf{Y} - \sum_{a=1}^p \theta'_a \mathbf{Z}_a\|_n^2 = (\mathbf{Z}_m^* \mathbf{Z}_m)^{-1} \mathbf{Z}_m \mathbf{Y}$$

**Perte des moindres carrés intégrés** :  $\|\sqrt{\Sigma}(\theta' - \theta)\|_p^2 := \mathbb{E} [\sum_{a=1}^p (\theta'_a - \theta_a) Z_i]^2$ .

$$\begin{aligned} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta}_m - \theta)\|_p^2 \right] &= \|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2 + \left[ \|\sqrt{\Sigma}(\theta_m - \theta)\|_p^2 + \sigma^2 \right] \frac{|m|}{n-|m|-1} \\ &= \ll \text{biais} \gg + \ll \text{variance} \gg. \end{aligned}$$

Objectif en prédiction :

$\rightsquigarrow$  Choisir  $m \subset \{1, \dots, p\}$  qui **minimise le risque**  $\mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta}_m - \theta)\|_p^2 \right]$ .

## Description de la procédure

support  $m$  grand  $\implies$  variance importante  $\implies$  pénaliser les grands supports

**penalité** : fonction  $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}^+$

$\text{pen}(m)$  dépend de la "complexité" de  $m$  et de  $\mathcal{M}$ .

$$\begin{aligned} \text{Critère pénalisé} \quad \text{Crit}(m) &:= \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_m\|_n^2 [1 + \text{pen}(m)] \\ \hat{m} &:= \arg \min_{m \in \mathcal{M}} \text{Crit}(m) . \end{aligned}$$

Exemple :

- 1 BIC :  $\text{pen}(m) = e^{\frac{|m|}{n} \log(n)} - 1 \rightsquigarrow \hat{m} := \arg \min \log \left[ \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_m\|_n^2 \right] + \frac{|m|}{n} \log(n)$
- 2 AIC :  $\text{pen}(m) = e^{2\frac{|m|}{n}} - 1 \rightsquigarrow \hat{m} := \arg \min \log \left[ \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_m\|_n^2 \right] + 2\frac{|m|}{n}$

(BIC et AIC ne fonctionnent pas)

$\mathcal{M}_p^D$  : collection des sous-ensembles de  $\{1, \dots, p\}$  de taille plus petite que  $D$ .

## Description de la procédure

support  $m$  grand  $\implies$  variance importante  $\implies$  pénaliser les grands supports

**penalité** : fonction  $\text{pen} : \mathcal{M} \rightarrow \mathbb{R}^+$

$\text{pen}(m)$  dépend de la "complexité" de  $m$  et de  $\mathcal{M}$ .

$$\begin{aligned} \text{Critère pénalisé} \quad \text{Crit}(m) &:= \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_m\|_n^2 [1 + \text{pen}(m)] \\ \hat{m} &:= \arg \min_{m \in \mathcal{M}} \text{Crit}(m) . \end{aligned}$$

Exemple :

- 1 BIC :  $\text{pen}(m) = e^{\frac{|m|}{n} \log(n)} - 1 \rightsquigarrow \hat{m} := \arg \min \log \left[ \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_m\|_n^2 \right] + \frac{|m|}{n} \log(n)$
- 2 AIC :  $\text{pen}(m) = e^{2\frac{|m|}{n}} - 1 \rightsquigarrow \hat{m} := \arg \min \log \left[ \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_m\|_n^2 \right] + 2\frac{|m|}{n}$

(BIC et AIC ne fonctionnent pas)

$\mathcal{M}_p^D$  : collection des sous-ensembles de  $\{1, \dots, p\}$  de taille plus petite que  $D$ .

$\rightsquigarrow$  : le temps de calcul est exponentiel en  $p$ !!

# Inégalité oracle

Hypothèse :  $D \leq \frac{n}{4}$ ,

## Théorème

Prenons  $\text{pen}(m) = \square \frac{|m|}{n} \log \left( \frac{ep}{|m|} \right) \exp \left[ \square \frac{|m|}{n} \log \left( \frac{ep}{|m|} \right) \right]$ .

L'estimateur  $\hat{\theta}_{\hat{m}}$  satisfait alors

$$\begin{aligned} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta}_{\hat{m}} - \theta)\|_p^2 \right] &\leq \square \inf_{m \in \mathcal{M}} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta}_m - \theta)\|_p^2 \right] \log \left( \frac{ep}{|m|} \right) e^{\square \frac{|m|}{n} \log \left( \frac{ep}{|m|} \right)} \\ &\quad + \square \frac{\sigma^2}{n} \end{aligned}$$

# Inégalité oracle

Hypothèse :  $D \leq \frac{n}{4}$ ,

## Théorème

Prenons  $\text{pen}(m) = \square \frac{|m|}{n} \log \left( \frac{ep}{|m|} \right) \exp \left[ \square \frac{|m|}{n} \log \left( \frac{ep}{|m|} \right) \right]$ .

L'estimateur  $\hat{\theta}_{\hat{m}}$  satisfait alors

$$\begin{aligned} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta}_{\hat{m}} - \theta)\|_p^2 \right] &\leq \square \inf_{m \in \mathcal{M}} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta}_m - \theta)\|_p^2 \right] \log \left( \frac{ep}{|m|} \right) e^{\square \frac{|m|}{n} \log \left( \frac{ep}{|m|} \right)} \\ &\quad + \square \frac{\sigma^2}{n} \end{aligned}$$

**Commentaires :**

- Risque  $\frac{k}{n} [1 + \log(p/k)] \sigma^2$  si  $\theta$  est  $k$ -sparse et si  $k[1 + \log(p/k)] < n$ .
- Risque de l'ordre de  $(\frac{p}{k})^{\square k/n} \sigma^2$  si  $\theta$  est  $k$ -sparse et si  $k[1 + \log(p/k)] > n$ .
- Pas de connaissance de la parcimonie a priori.
- Aucune hypothèse portant sur  $\Sigma$ .

**Question :**

- **Optimalité** de cette borne de risque ?

## Minorations minimax

"Low dimension"

"High dimension"

"Ultra High dimension"

$$p \ll n$$

$$n = \mathcal{O}(p)$$

$$k \log(p) \geq n$$

$\Theta[k, p]$  : ensemble des vecteurs  $k$ -sparse.

### Proposition

Supposons  $\Sigma = I$  (covariables indépendentes) et prenons  $k \leq \sqrt{p}$ .

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta[k, p]} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_p^2 \right] \geq \square \frac{k}{n} \log \left( \frac{p}{k} \right) \exp \left[ \square \frac{k}{n} \log \left( \frac{p}{k} \right) \right] \sigma^2 .$$

## Minorations minimax

"Low dimension"

"High dimension"

"Ultra High dimension"

$$p \ll n$$

$$n = \mathcal{O}(p)$$

$$k \log(p) \geq n$$

$\Theta[k, p]$  : ensemble des vecteurs  $k$ -sparse.

### Proposition

Supposons  $\Sigma = I$  (covariables indépendentes) et prenons  $k \leq \sqrt{p}$ .

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta[k, p]} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_p^2 \right] \geq \square \frac{k}{n} \log \left( \frac{p}{k} \right) \exp \left[ \square \frac{k}{n} \log \left( \frac{p}{k} \right) \right] \sigma^2 .$$

**Commentaires pour  $k \log(p/k) \leq n$**

- Risque optimal (minimax)  $\frac{k}{n} \log \left( \frac{p}{k} \right) \sigma^2$ .
- Le prix à payer pour les  $p - k$  covariables non pertinentes est logarithmique.
- $\hat{\theta}_{\hat{m}}$  is minimax adaptatif à la taille du support.
- La "pire" covariance pour le risque minimax  $\Sigma = I$ .

# Très grande dimension

"Low dimension"

"High dimension"

"Ultra High dimension"

$$p \ll n$$

$$n = \mathcal{O}(p)$$

$$k \log(p) \geq n$$

## Proposition (Risque Minimax)

Si  $\Sigma = I$ ,

$$\inf_{\hat{\theta}} \sup_{\theta \in \theta[k,p]} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_p^2 \right] \approx \square \exp \left[ \square \frac{k}{n} \log \left( \frac{p}{k} \right) \right]$$

## Très grande dimension

"Low dimension"

"High dimension"

"Ultra High dimension"

$$p \ll n$$

$$n = \mathcal{O}(p)$$

$$k \log(p) \geq n$$

### Proposition (Risque Minimax)

Si  $\Sigma = I$ ,

$$\inf_{\hat{\theta}} \sup_{\theta \in \theta[k,p]} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_p^2 \right] \approx \square \exp \left[ \square \frac{k}{n} \log \left( \frac{p}{k} \right) \right]$$

**Commentaires :**

- Risque grandit **exponentiellement** /  $k$  et **polynomialement** /  $p$ .
- Interprétation : Le design empirique  $\mathbf{Z}^* \mathbf{Z}$  est éloigné de l'identité.

## Estimation du support de $\theta$

Cas particulier  $\Sigma = I$ .

$$\Theta_k[r] := \{\theta \in \Theta[k, p] : \#\{a : |\theta_a| \geq r\} = k\}$$

### Proposition (Wainwright(09))

Si  $r^2 \geq \square[1 + \log(p/k)] \frac{\sigma^2}{n}$  et si  $n \geq \square k[1 + \log(p/k)]$ , alors il existe un estimateur  $\hat{\theta}$  tel que

$$\mathbb{P} \left[ \text{supp}(\hat{\theta}) = \text{supp}(\theta) \right] \geq 1 - \square$$

Si  $r^2 \leq \square[1 + \log(p/k)] \frac{\sigma^2}{n}$ , alors

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k[r]} \mathbb{P} \left[ \text{supp}(\hat{\theta}) \neq \text{supp}(\theta) \right] \geq 1/8$$

## Estimation du support de $\theta$

Cas particulier  $\Sigma = I$ .

$$\Theta_k[r] := \{\theta \in \Theta[k, p] : \#\{a : |\theta_a| \geq r\} = k\}$$

### Proposition (Wainwright(09))

Si  $r^2 \geq \square[1 + \log(p/k)] \frac{\sigma^2}{n}$  et si  $n \geq \square k[1 + \log(p/k)]$ , alors il existe un estimateur  $\hat{\theta}$  tel que

$$\mathbb{P} \left[ \text{supp}(\hat{\theta}) = \text{supp}(\theta) \right] \geq 1 - \square$$

Si  $r^2 \leq \square[1 + \log(p/k)] \frac{\sigma^2}{n}$ , alors

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k[r]} \mathbb{P} \left[ \text{supp}(\hat{\theta}) \neq \text{supp}(\theta) \right] \geq 1/8$$

#### Commentaires :

- Estimation des voisins du gène  $a$  possible si  $\theta_{a,b}^2 \geq \square \log(p/k)/n\sigma^2$ .
- Minoration uniquement valable pour  $\Sigma = I$ . Pour des covariables liées, vitesses différentes.

## Estimation du support (suite)

"Low dimension"

"High dimension"

"Ultra High dimension"

$$p \ll n$$

$$n = \mathcal{O}(p)$$

$$k \log(p) \geq n$$

### Proposition (Très grande dimension)

Si  $r^2 \leq \exp\left[-\frac{k \log(p/k)}{n}\right] \sigma^2$ , alors

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k[r]} \mathbb{P} \left[ \text{supp}(\hat{\theta}) \neq \text{supp}(\theta) \right] \geq 1/8$$

## Estimation du support (suite)

"Low dimension"

"High dimension"

"Ultra High dimension"

$$p \ll n$$

$$n = \mathcal{O}(p)$$

$$k \log(p) \geq n$$

### Proposition (Très grande dimension)

Si  $r^2 \leq \exp\left[-\frac{k \log(p/k)}{n}\right] \sigma^2$ , alors

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta_k[r]} \mathbb{P} \left[ \text{supp}(\hat{\theta}) \neq \text{supp}(\theta) \right] \geq 1/8$$

#### Conséquence :

Si "l'intensité du lien" entre deux gènes est plus petite que  $\exp\left[-\frac{k \log(p/k)}{n}\right] \sigma^2$ , l'arête n'est pas détectable.

## Réduction de dimension

$\widehat{M}_\delta$  : sous-ensemble de  $\{1, \dots, p\}$  de taille  $p^\delta$  mesurable par rapport aux données.

### Proposition (Impossibilité de la réduction)

Si  $r^2 \leq C \frac{k}{n} \log\left(\frac{p}{k}\right) \exp\left[-\frac{k \log(p/k)}{n}\right] \sigma^2$ , alors il existe  $0 < \delta < 1$  tel que

$$\inf_{\widehat{M}_\delta} \sup_{\theta \in \Theta_k[r]} \mathbb{P} \left[ \text{supp}(\theta) \not\subseteq \widehat{M}_\delta \right] \geq 1/8$$

## Réduction de dimension

$\widehat{M}_\delta$  : sous-ensemble de  $\{1, \dots, p\}$  de taille  $p^\delta$  mesurable par rapport aux données.

### Proposition (Impossibilité de la réduction)

Si  $r^2 \leq C \frac{k}{n} \log \left( \frac{p}{k} \right) \exp \left[ -\frac{k \log(p/k)}{n} \right] \sigma^2$ , alors il existe  $0 < \delta < 1$  tel que

$$\inf_{\widehat{M}_\delta} \sup_{\theta \in \Theta_k[r]} \mathbb{P} \left[ \text{supp}(\theta) \not\subseteq \widehat{M}_\delta \right] \geq 1/8$$

#### Conséquence :

En très grande dimension, il est presque **impossible** de trouver un ensemble de gène de taille  $n$  qui contienne les vrais voisins du gène  $a$ .

Une préétude des données (ex : correlation screening, clustering,... ) a toute les chances de se tromper...

# Bilan

- En dimension raisonnable, le prix à payer pour la grande dimension est **logarithmique**.
- Vitesse atteinte par des **procédures** rapides (ex : lasso).... sous des hypothèses restrictives sur la covariance.
- Très grande dimension :  $\text{deg}(g) \geq n/[2 \log(p)]$ .  
↪ Il est presque **impossible** d'estimer le réseau ou même de faire de la réduction de dimension.  
ex :  $p = 3000$ ,  $n = 50$ , ↪  $\text{deg}(g) \leq 4$

# Bilan

- En dimension raisonnable, le prix à payer pour la grande dimension est **logarithmique**.
- Vitesse atteinte par des **procédures** rapides (ex : lasso).... sous des hypothèses restrictives sur la covariance.
- Très grande dimension :  $\deg(g) \geq n/[2 \log(p)]$ .  
↪ Il est presque **impossible** d'estimer le réseau ou même de faire de la réduction de dimension.  
ex :  $p = 3000$ ,  $n = 50$ , ↪  $\deg(g) \leq 4$

En pratique :

- Échantillon  $X$  non iid gaussien.
- Connaissances a priori.

## Extension à la régression à design fixe

$$\mathbf{Y} = \mathbf{Z}\theta + \boldsymbol{\epsilon},$$

où  $\theta \in \mathbb{R}^p$ ,  $\boldsymbol{\epsilon} \sim \mathcal{N}(0_n, \sigma^2 I_n)$ ,  $\mathbf{Z}$  de taille  $n \times p$ .

Le design  $\mathbf{Z}$  peut être considéré :

- 1 fixe
- 2 gaussien (GGM)

## Extension à la régression à design fixe

$$\mathbf{Y} = \mathbf{Z}\theta + \boldsymbol{\epsilon},$$

où  $\theta \in \mathbb{R}^p$ ,  $\boldsymbol{\epsilon} \sim \mathcal{N}(0_n, \sigma^2 I_n)$ ,  $\mathbf{Z}$  de taille  $n \times p$ .

Le design  $\mathbf{Z}$  peut être considéré :

- 1 fixe
- 2 gaussien (GGM)

La variance  $\sigma^2$  peut-être considéré :

- 1 connu.
- 2 inconnu (cas précédent)

## Extension à la régression à design fixe

$$\mathbf{Y} = \mathbf{Z}\theta + \epsilon ,$$

où  $\theta \in \mathbb{R}^p$ ,  $\epsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$ ,  $\mathbf{Z}$  de taille  $n \times p$ .

Le design  $\mathbf{Z}$  peut être considéré :

- 1 fixe
- 2 gaussien (GGM)

La variance  $\sigma^2$  peut-être considéré :

- 1 connu.
- 2 inconnu (cas précédent)

Risque minimax pour  $\theta$  parcimonieux :

- 1 **Prédiction**  $\rightsquigarrow$  Estimation de  $\mathbb{E}(\mathbf{Y})$  ou de  $\mathbb{E}(\mathbf{Y}|\mathbf{Z})$
- 2 **Problème inverse**  $\rightsquigarrow$  Estimation de  $\theta$
- 3 Estimation du support
- 4 Test d'hypothèses linéaire sur  $\theta$ .

## Prédiction

Risque minimax dépend du design  $\mathbf{Z}$  :  $\inf_{\hat{\theta}} \sup_{\theta \in [k,p]} \mathbb{E} \left[ \|\mathbf{Z}(\hat{\theta} - \theta)\|_n^2 \right] / (n\sigma^2)$

Quel risque peut- atteindre uniformément sur les designs ? Risque minimax sur le “pire design” :

$$R_{k,n,p}^{\text{Fix}} = \sup_{\mathbf{Z}} \inf_{\hat{\theta}} \sup_{\theta \in [k,p]} \mathbb{E} \left[ \|\mathbf{Z}(\hat{\theta} - \theta)\|_n^2 \right] / (n\sigma^2)$$

$$R_{k,n,p}^{\text{Rand}} = \sup_{\Sigma} \inf_{\hat{\theta}} \sup_{\theta \in [k,p]} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_p^2 \right] / \sigma^2$$

## Prédiction

Risque minimax dépend du design  $\mathbf{Z}$  :  $\inf_{\hat{\theta}} \sup_{\theta \in [k,p]} \mathbb{E} \left[ \|\mathbf{Z}(\hat{\theta} - \theta)\|_n^2 \right] / (n\sigma^2)$

Quel risque peut- atteindre uniformément sur les designs ? Risque minimax sur le “pire design” :

$$R_{k,n,p}^{\text{Fix}} = \sup_{\mathbf{Z}} \inf_{\hat{\theta}} \sup_{\theta \in [k,p]} \mathbb{E} \left[ \|\mathbf{Z}(\hat{\theta} - \theta)\|_n^2 \right] / (n\sigma^2)$$

$$R_{k,n,p}^{\text{Rand}} = \sup_{\Sigma} \inf_{\hat{\theta}} \sup_{\theta \in [k,p]} \mathbb{E} \left[ \|\sqrt{\Sigma}(\hat{\theta} - \theta)\|_p^2 \right] / \sigma^2$$

	Design Fixe	Design Gaussien
$\sigma^2$ connu	$\square \frac{k \log(p/k)}{n} \wedge 1$	$\square \frac{k \log(p/k)}{n} \exp \left[ \square \frac{k \log(p/k)}{n} \right]$
$\sigma^2$ inconnu	Adaptation est impossible	Pas d'effet

### Commentaires :

- design fixe, estimation d'un vecteur de dimension  $n$ .
- design fixe, la **parcimonie ne joue plus** en très grande dimension.
- En variance inconnue, impossible d'être adaptatif en très grande dimension (Baraud *et al.* 2009)
- Design fixes correspondent à des réalisations d'un design gaussien standard.

## Problèmes inverses

Pour certains designs  $\mathbf{Z}$ , le risque minimax est infini.

Le risque minimax proportionnel à un facteur d'échelle sur  $\mathbf{Z}$ .

$\rightsquigarrow \mathcal{D}_{n,p}$  : designs  $\mathbf{Z}$   $n \times p$  dont chaque colonne est **normée**.

$$\inf_{\mathbf{Z} \in \mathcal{D}_{n,p}} \inf_{\hat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E} \left[ \frac{\|\hat{\theta} - \theta\|_p^2}{\sigma^2} \right].$$

Que peut-on espérer avec le **"meilleur" design  $\mathbf{Z}$**  ?

## Problèmes inverses

Pour certains designs  $\mathbf{Z}$ , le risque minimax est infini.

Le risque minimax proportionnel à un facteur d'échelle sur  $\mathbf{Z}$ .

↔  $\mathcal{D}_{n,p}$  : designs  $\mathbf{Z}$   $n \times p$  dont chaque colonne est **normée**.

$$\inf_{\mathbf{Z} \in \mathcal{D}_{n,p}} \inf_{\hat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E} \left[ \frac{\|\hat{\theta} - \theta\|_p^2}{\sigma^2} \right].$$

Que peut-on espérer avec le "meilleur" design  $\mathbf{Z}$  ?

$(\mathbf{k}, \mathbf{n}, \mathbf{p})$	$k \log(p/k) \leq Cn$	$k \log(p/k) \gg n \log(n)$
Risque minimax	$Ck \log(p/k)$	$\exp[Ck \log(p/k)/n]$

### Commentaires :

- Similaire au design gaussien avec  $\Sigma = I$
- Dimension raisonnable : prix  $\log(p)$ .
- Très grande dimension : prix plus élevé.
- Argument **géométriques** :  
 Impossibilité de construire un design  $\mathbf{Z}$  séparant bien les covariables en TGD.  
 $\iff$  : impossibilité de placer  $p$  points "bien" séparés sur la sphère  $S_{n-1}$ .

# Résumé

## Estimation de réseaux :

- 1 En dimension raisonnable, un prix **logarithmique** ( $\log(p)$ ) à payer ....
- 2 une limite structurelle :  **$\deg(g) \geq n/[2 \log(p)]$** .

# Résumé

## Estimation de réseaux :

- 1 En dimension raisonnable, un prix **logarithmique** ( $\log(p)$ ) à payer ....
- 2 une limite structurelle :  **$\deg(g) \geq n/[2 \log(p)]$** .

## Plus généralement, **estimation en très grande dimension** :

- 1 La forme du **design** et connaissance de la **variance** parfois primordiales.
- 2 un outil d'analyse des **limites** de problèmes pratiques et de procédures d'inférence.
- 3 Des questions **géométriques**.