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

Journées Réseaux

Nicolas Verzelen

20 Janvier 2011

Modèle graphique non orienté (Markov Random fields)

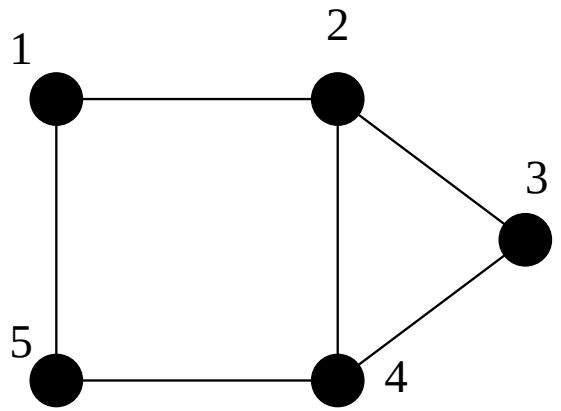
On considère $X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Omega^{-1})$

Ω inversible
$\Gamma := \{1, \ldots, p\}$
$g = (\Gamma, E)$ graphe non orienté

Modèle graphique non orienté (Markov Random fields)

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

Ω inversible
$\Gamma := \{1, \ldots, p\}$
$g = (\Gamma, E)$ graphe non orienté



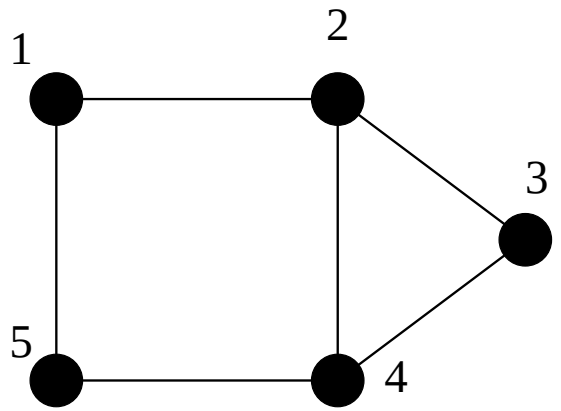
X est un modèle graphique gaussien par rapport à g si pour tout sommet a

X_a indépendant de $\{X_b \ : \ b \nsim a\}$ conditionellement à $\{X_b \ : \ b \sim a\}$

Modèle graphique non orienté (Markov Random fields)

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

Ω inversible
$\Gamma := \{1, \ldots, p\}$
$g = (\Gamma, E)$ graphe non orienté



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

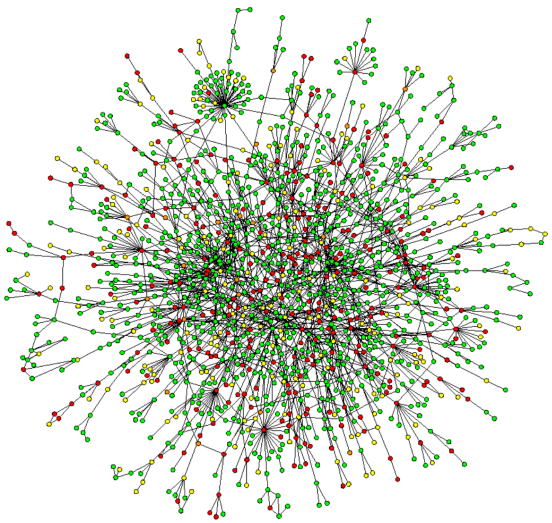
X_a indépendant de $\{X_b \ : \ b \nsim a\}$ conditionellement à $\{X_b \ : \ b \sim a\}$

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 modèlisés à l'aide d'un GGM de graphe **g** inconnu. (le réseau de gène)

Objectif : estimer à partir des données transcriptomiques le graphe **g** du GGM.

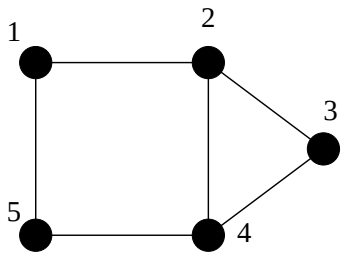
Difficulté principale : $n \ll p$

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

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

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

Ex2 : Pénalisation l_1 (Glasso)

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

Régression conditionelle

$$X_a = \sum_{b \neq a} \theta_{a,b} X_b + \epsilon_a \ ,$$

avec $\epsilon_a \perp\!\!\!\perp (X_b)_{b \neq a}$ et matrice θ définie par

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

Estimation du graphe $\Longleftrightarrow$ Sélection des 0 de θ.

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

Ex1 : Pénalisation par complexité.

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

Ex2 : Pénalisation l_1 (Lasso)

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

Sujet de recherche actif

Nouveaux algorithmes : par seuillage ou par régularisation

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

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/Huet/V. (09) ...	- Yuan/Lin (06) - Banerjee *et al.* (07) - Ambroise *et al.*(09) ...

Caractéristiques :

- approches "souvent" algorithmiques.
- quelques résultats théoriques lorsque $1 \ll n \ll p$
  + hypothèses sur la matrice de covariance Ω^{-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 ?

Modèle de régression linéaire

$$Y = X\theta + \epsilon \ ,$$

avec

- $\theta \in \mathbb{R}^p$ est inconnu.
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ et σ^2 inconnu
- $X \sim \mathcal{N}(0_p, \Sigma)$ et Σ inconnu.

Données : n observations indépendantes.

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

avec

- $\mathbf{Y}$ réponse de taille n.
- Design $\mathbf{X}$ de taille $n \times p$.

Liens avec les GGMs : le support de θ correspond aux voisins dans un modèle graphique gaussien.

Problèmes statistiques classiques– liens avec les GGMs

$$\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$$

($\mathbf{P_1}$) : **Prédiction**. Estimer un signal $\mathbf{X}\theta = \mathbb{E}[\mathbf{Y}|\mathbf{X}]$.
($\mathbf{P'_1}$) : **Prédiction à design aléatoire**. Estimation de $\mathbb{E}[Y_{new}|X_{new}]$.
$\rightsquigarrow$ but : comprendre la loi d'expression d'un gène conditionnellement aux autres.

($\mathbf{P_2}$) : **Test d'hypothèse linéaire**. Tester l'hypothèse nulle. $\mathbf{H_0}$: "$\theta = 0$".
$\rightsquigarrow$: but : tester une hypothèse sur le voisinage

($\mathbf{P_3}$) : **Problème inverse**. Estimer θ.
$\rightsquigarrow$: but : Estimer la contribution de chacun des gènes à l'expression d'un gène A

($\mathbf{P_4}$) : **Estimation du support**. Retrouver le support de θ. $\{i,\ \theta_i \neq 0\}$.
$\rightsquigarrow$ But : Estimer le voisinage d'un gène A dans le graphe.

($\mathbf{P'_4}$) : **réduction de dimension** . Estimer un ensemble de covariables $\widehat{M} \subset \{1, \dots p\}$ de taille raisonnables qui contienne le support de θ avec grande probabilité .
$\rightsquigarrow$But : sélectionner un sous-ensemble de gènes potentiellement voisins de A.

Parcimonie et grande dimension

Dans beaucoup d'applications (e.g., postgénomiques, fMRI), le nombre p de covariables est beaucoup plus grand que n.

Sparsité (parcimonie) : la plupart des composantes de θ sont nulles.
Notation : $\Theta[k, p]$ ensemble des vecteurs k-sparse.

Statistique en grande dimension : $k \leq n \leq p$.

- Difficultés Théoriques (analyse non-asymptotique).
- Difficultés Computationels : e.g. Lasso, Dantzig selector, ...

$$\widehat{\theta} := \arg\inf_{\theta'} \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 + \lambda\|\theta'\|_1$$

Parcimonie et grande dimension

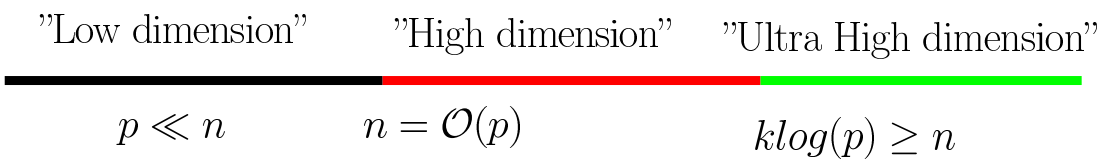
Dans beaucoup d'applications (e.g., postgénomiques, fMRI), le nombre p de covariables est beaucoup plus grand que n.

Sparsité (parcimonie) : la plupart des composantes de θ sont nulles.
Notation : $\Theta[k, p]$ ensemble des vecteurs k-sparse.

Statistique en grande dimension : $k \leq n \leq p$.

- Difficultés Théoriques (analyse non-asymptotique).
- Difficultés Computationels : e.g. Lasso, Dantzig selector, ...

$$\widehat{\theta} := \arg\inf_{\theta'} \|\mathbf{Y} - \mathbf{X}\theta'\|_n^2 + \lambda\|\theta'\|_1$$



Propriétés minimax et adaptations

Comprendre les limitations structurelles de ces problèmes :

1. Pour un problème donné, quel est le plus petit risque possible ?
2. Est-il possible d'obtenir un risque faible pour p arbitrairement grand ?
   $\rightsquigarrow$ Que peut-on faire avec $p = 5000$ genes et $n = 40$ expériences microarray ?

Propriétés minimax et adaptations

Comprendre les limitations structurelles de ces problèmes :

1. Pour un problème donné, quel est le plus petit risque possible ?
2. Est-il possible d'obtenir un risque faible pour p arbitrairement grand ?
   $\rightsquigarrow$ Que peut-on faire avec $p = 5000$ genes et $n = 40$ expériences microarray ?

Étant donnés une fonction de perte $l(.,.)$ et un estimateur $\widehat{\theta}$, le risque maximal de $\widehat{\theta}$ sur $\Theta[k,p]$ est défini par

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta},\theta)] \ .$$

Le risque minimax sur $\Theta[k,p]$ vaut

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta},\theta)] \ .$$

Objectif principal : Calculer le risque minimax sur $\Theta[k,p]$ for différentes fonctions de pertes associées aux problèmes ($\mathbf{P_1} - \mathbf{P_4}$).

Propriétés minimax et adaptations

Comprendre les limitations structurelles de ces problèmes :

1. Pour un problème donné, quel est le plus petit risque possible ?
2. Est-il possible d'obtenir un risque faible pour p arbitrairement grand ?
   $\rightsquigarrow$ Que peut-on faire avec $p = 5000$ genes et $n = 40$ expériences microarray ?

Étant donnés une fonction de perte $l(.,.)$ et un estimateur $\widehat{\theta}$, le risque maximal de $\widehat{\theta}$ sur $\Theta[k,p]$ est défini par

$$\sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta},\theta)] \ .$$

Le risque minimax sur $\Theta[k,p]$ vaut

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[l(\widehat{\theta},\theta)] \ .$$

Objectif principal : Calculer le risque minimax sur $\Theta[k,p]$ for différentes fonctions de pertes associées aux problèmes ($\mathbf{P_1} - \mathbf{P_4}$).

En pratique, le sparsité k est inconnue et la variance σ^2 est souvent inconnue.
Peut-on s'adapter à k ? Peut-on s'adapter à σ^2 ?

Risque minimax pour le pire des designs

Objectif : Estimer $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$. **Objectif** : Estimer $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$.

Fonction de perte : $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)$

$(P_1') \rightsquigarrow E_{X_{new}}\left[\left(X_{new}(\widehat{\theta} - \theta)\right)^2\right]/\sigma^2 = \|\sqrt{\Sigma}(\widehat{\theta} - \theta)\|_p^2/\sigma^2$

Risque minimax pour le pire des designs

Objectif : Estimer $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$. **Objectif** : Estimer $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$.

Fonction de perte : $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)$

$(P_1') \rightsquigarrow E_{X_{new}}\left[\left(X_{new}(\widehat{\theta} - \theta)\right)^2\right]/\sigma^2 = \|\sqrt{\Sigma}(\widehat{\theta} - \theta)\|_p^2/\sigma^2$

Si le support de θ est connu $\rightsquigarrow$ Paramétrique risque k/n.

Risque minimax pour le pire des designs

Objectif : Estimer $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$. **Objectif** : Estimer $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\theta$.

Fonction de perte : $\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)$

$(P_1') \rightsquigarrow E_{X_{new}}\left[\left(X_{new}(\widehat{\theta} - \theta)\right)^2\right]/\sigma^2 = \|\sqrt{\Sigma}(\widehat{\theta} - \theta)\|_p^2/\sigma^2$

Si le support de θ est connu $\rightsquigarrow$ Paramétrique risque k/n.

Dépendance complexe du risque minimax $\inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)]$ en le design $\mathbf{X}$.

Objectif : mettre en lumière le role (k, n, p).
$\rightsquigarrow$ Risques Minimax uniformément sur tous les designs $\mathbf{X}$ de taille $n \times p$.

$$\mathcal{R}[k] := \sup_{\mathbf{X}} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\mathbf{X}(\widehat{\theta} - \theta)\|_n^2/(n\sigma^2)]$$

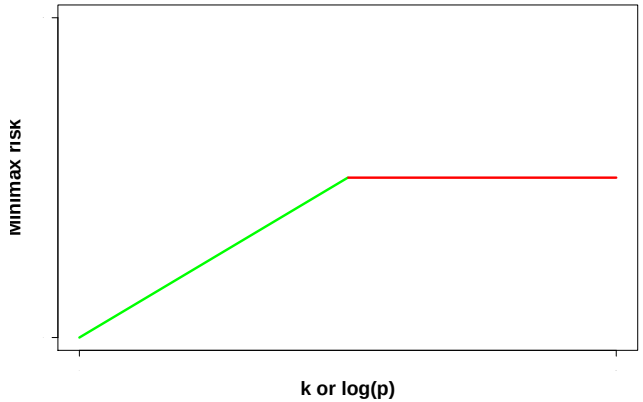
$$\mathcal{R}_R[k] := \sup_{\Sigma} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\sqrt{\Sigma}(\widehat{\theta} - \theta)\|_p^2/\sigma^2]$$

Proposition

Pour tout $k \leq n \wedge p$, on a

$$\square \frac{k}{n} \log(ep/k) \wedge 1 \leq \mathcal{R}[k] \leq \square' \frac{k}{n} \log(ep/k) \wedge 1$$

$$\mathcal{R}[k] \simeq \square \frac{k}{n} \log(ep/k) \wedge 1.$$



Commentaires :

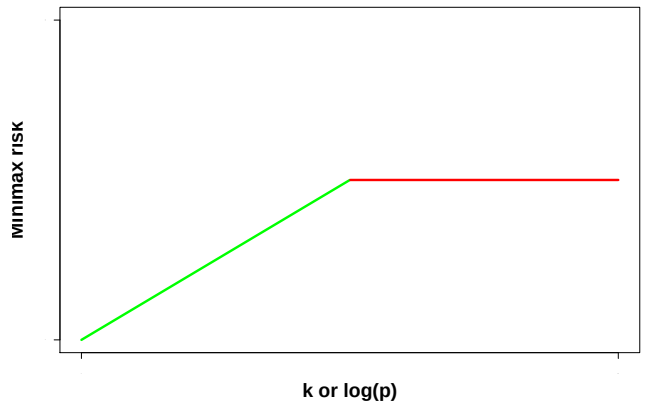
- En dimension raisonnable, "prix logarithmique" pour la non-connaissance du support.
  $\rightsquigarrow$ analogue au modèle de séquence Gaussienne (*Johnstone (94)*).
- En très grande dimension, le problème est aussi complexe qu'estimer un vecteur dans $\mathbb{R}^n$.

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \qquad \text{si } k \leq k^*$$

$$\widehat{\theta}_n := \arg \inf_{\theta} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \qquad \text{si } k \leq k^*$$

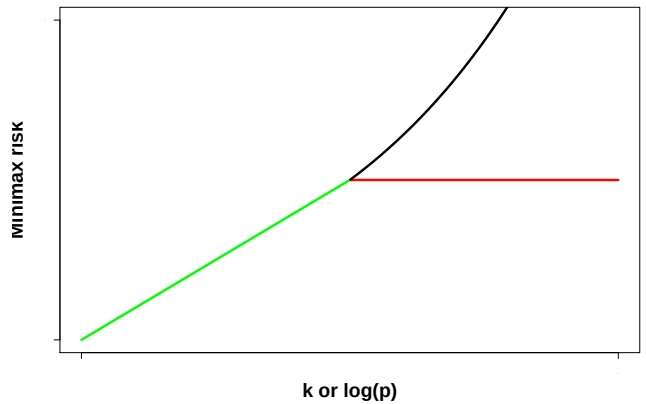
- Critères convexes (LASSO, Dantzig Selector) n'atteignent ces bornes que sous des hypothèses restrictives sur $\mathbf{X}$.

Adaptation à la variance et à la sparsité.



- Adaptation à la variance est possible (estimateur des moindre-carrés).
- Adaptation à la sparsité est possible (estimateurs des moindre-carrés pénalisés.

Adaptation à la variance et à la sparsité.



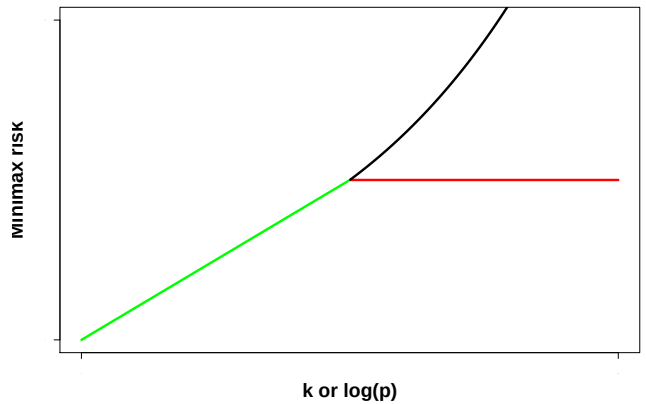
- Adaptation à la variance est possible (estimateur des moindre-carrés).
- Adaptation à la sparsité est possible (estimateurs des moindre-carrés pénalisés.
- Est-it possible d'être simultanémnent adaptatif à la sparsité and à la variance ? *Baraud/Giraud/Huet (09)*

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \qquad \text{si } k \leq k^*$$

$$\tilde{k}_{BGH} := \arg \inf_{k \leq n/2} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_k\|_n^2 \, [1 + \psi(k)] \ ,$$

ψ joue le role d'une pénalité.

Adaptation à la variance et à la sparsité.



- Adaptation à la variance est possible (estimateur des moindre-carrés).
- Adaptation à la sparsité est possible (estimateurs des moindre-carrés pénalisés.
- Est-it possible d'être simultanémnent adaptatif à la sparsité and à la variance ? *Baraud/Giraud/Huet (09)*

$$\widehat{\theta}_k := \arg \inf_{\theta \in \Theta[k,p]} \|\mathbf{Y} - \mathbf{X}\theta\|_n^2 \qquad \text{si } k \leq k^*$$

$$\tilde{k}_{BGH} := \arg \inf_{k \leq n/2} \|\mathbf{Y} - \mathbf{X}\widehat{\theta}_k\|_n^2 \, [1 + \psi(k)] \ ,$$

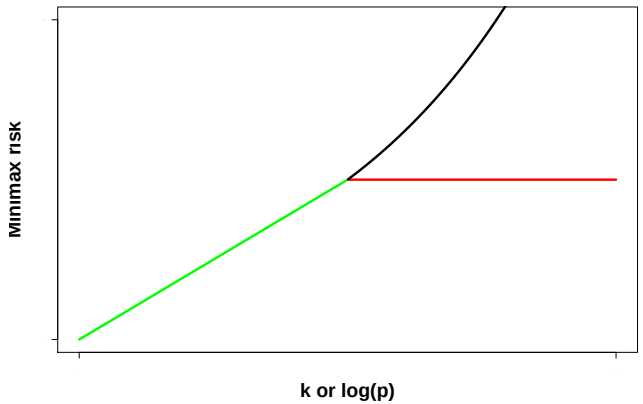
ψ joue le role d'une pénalité.
Non, c'est impossible, BGH est optimal.

Prédiction à design aléatoire.

Proposition

Pour tout $k \leq n \wedge p^{1/3}$, *on a*

$$\mathcal{R}_R[k] \simeq \square \frac{k}{n} \log(ep/k) \exp\left[\square \frac{k}{n} \log(ep/k)\right].$$



Commentaires :

- En dimension raisonnable, "prix logarithmique" pour la non-connaissance du support.
- En très grande dimension, explosion, estimer $\sqrt{\Sigma}\theta$ devient presque impossible.

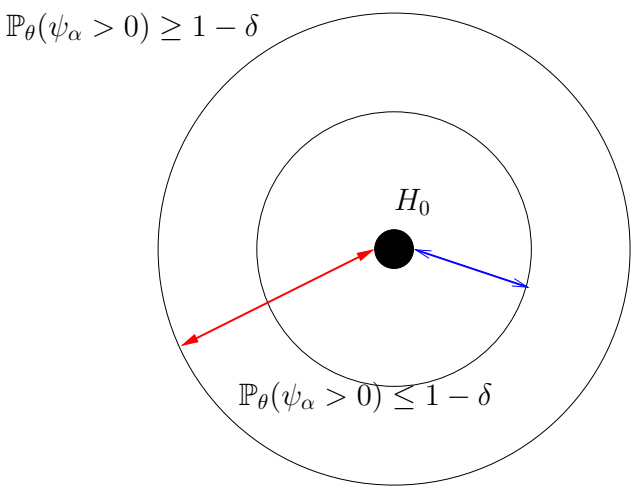
Distance de séparation Minimax

$\mathbf{H_0}$: $\theta = 0$ contre $\mathbf{H_1}$: $\theta \in \Theta[k, p] \setminus \{0\}$.
Test de l'hypothese : "le gène A n'a pas de voisins" contre "le géne A a au plus k voisins".

Fix $\delta > 0$. ψ_α test de Level α.

Distance de Séparation distance de ψ_α :

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



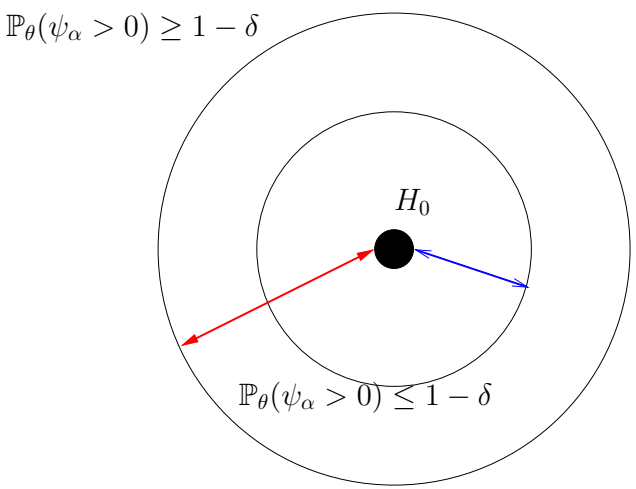
Distance de séparation Minimax

$\mathbf{H_0}$: $\theta = 0$ contre $\mathbf{H_1}$: $\theta \in \Theta[k, p] \setminus \{0\}$.
Test de l'hypothese : "le gène A n'a pas de voisins" contre "le géne A a au plus k voisins".

Fix $\delta > 0$. ψ_α test de Level α.

Distance de Séparation distance de ψ_α :

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



Distance Minimax de séparation

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

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

Variance connue σ^2

Si le support de θ est connu $\rightsquigarrow$ carré de la distance de séparation paramétrique $\sqrt{k}/n$.

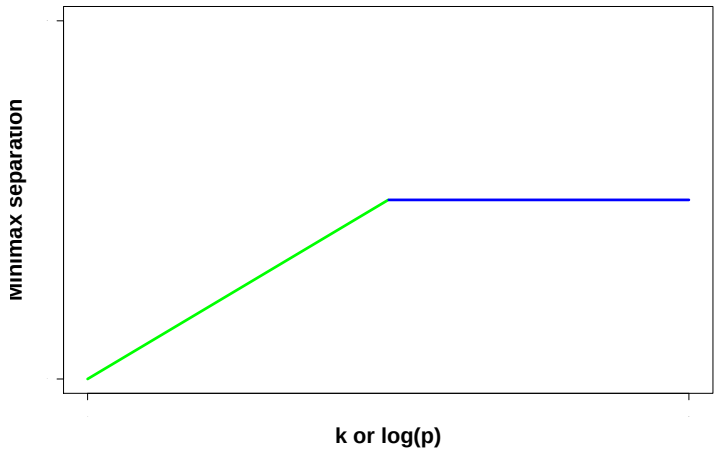
Variance connue σ^2

Si le support de θ est connu $\rightsquigarrow$ carré de la distance de séparation paramétrique $\sqrt{k}/n$.

Théorème

Pour $p \geq n \geq \square(\alpha, \delta)$ and $k \leq p^{1/3}$, nous avons

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



Comments :

- Si $k \log(ep/k)$ est petit par rapport à $\sqrt{n}$, analogue au risque minimax en prédiction minimax prediction risk.
  analogue au modèle de séquence gaussienne (*Baraud (02), Donoho/Jin (04)*).
- Des grands $(k, p) \Rightarrow$ distance séparation paramétrique dans $\mathbb{R}^n$.
- Adaptation à la sparsité est possible. (procédure de test multiple de Bonferroni).

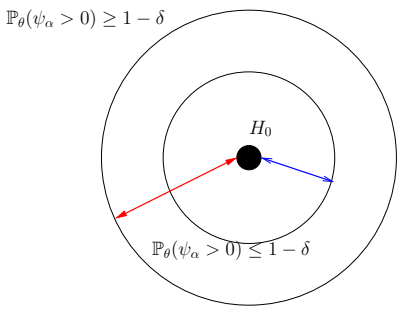
Variance inconnue σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Distance de séparation lorsque la variance inconnue.

Variance inconnue σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Distance de séparation lorsque la variance inconnue.

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



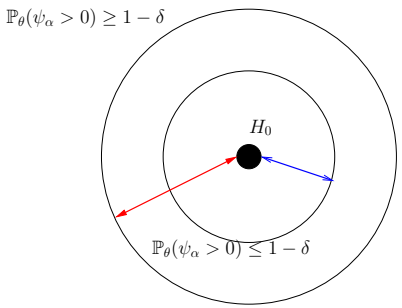
$$\rho^*_{\boldsymbol{U}}[k, \Sigma] := \inf_{\psi_\alpha} \rho_{\boldsymbol{U}}[\psi_\alpha, k, \Sigma] .$$

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

Variance inconnue σ^2

ψ_α : $\sup_{\sigma>0} \mathbb{P}_{0,\sigma}[\psi_\alpha = 1] \leq \alpha$. Distance de séparation lorsque la variance inconnue.

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



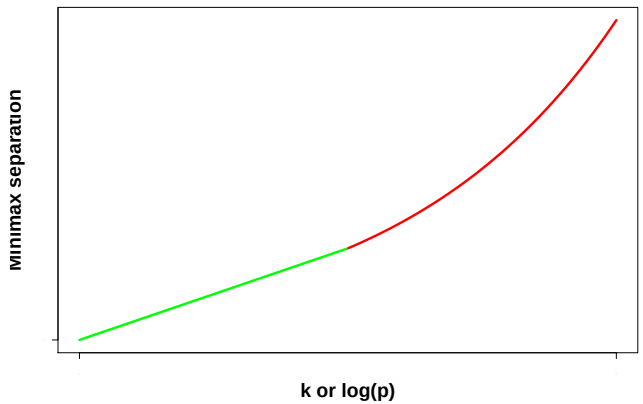
$$\rho^*_U[k, \Sigma] := \inf_{\psi_\alpha} \rho_U[\psi_\alpha, k, \Sigma] .$$

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

Théorème

Pour $p \geq n \geq \square(\alpha, \delta)$ and $k \leq p^{1/3}$, nous avons

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



Commentaire

- Si $k \log(ep/k)$ petit en comparaison de $\sqrt{n}$, même distance de séparation que pour la variance connue.
- Explosion en très grande dimension. Majoration $\rightsquigarrow$ (*Baraud/Huet/Laurent (03)*).

Problème inverse

Fonction de perte : $\|\theta - \widehat{\theta}\|_p^2/\sigma^2$. (estimation de l'influences des autres gènes sur l'expression du gène A)

$$\mathcal{RI}[k, \Sigma] := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\theta - \widehat{\theta}\|_p^2/\sigma^2] \ .$$

$\mathcal{RI}[k, \Sigma]$ est inversement proportionnel à Σ.

Problème inverse

Fonction de perte : $\|\theta - \widehat{\theta}\|_p^2/\sigma^2$. (estimation de l'influences des autres gènes sur l'expression du gène A)

$$\mathcal{RI}[k, \Sigma] := \inf_{\widehat{\theta}} \sup_{\theta \in \Theta[k,p]} \mathbb{E}_{\theta,\sigma}[\|\theta - \widehat{\theta}\|_p^2/\sigma^2] \ .$$

$\mathcal{RI}[k, \Sigma]$ est inversement proportionnel à Σ.

$\rightsquigarrow$ Collection $\mathcal{D}_{n,p}$ de lois Σ tel que la disagonale vaut 1..

$$\mathcal{RI}[k] := \inf_{\Sigma \in \mathcal{D}_{n,p}} \mathcal{RI}[k, \Sigma] \ .$$

$\rightsquigarrow$ Pour le "meilleur design possible", quel est le risque minimax ?

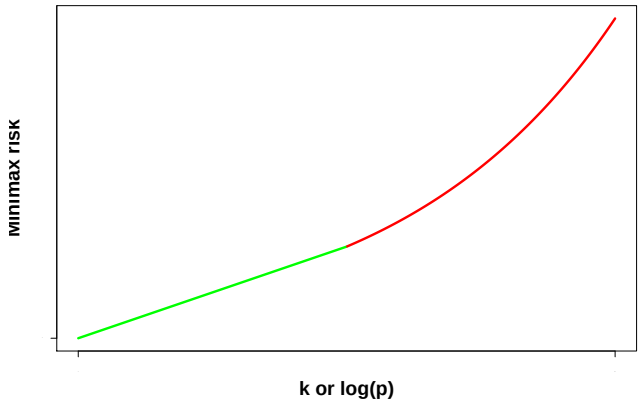
Théorème

Supposons que $k \log(ep/k) \leq \square n$. *Alors,*

$$\mathcal{RI}[k] \simeq \square k \log\left(\frac{ep}{k}\right) \ .$$

Supposons que $k \log(ep/k) \gg n \log(n)$, *alors*

$$\mathcal{RI}[k] \simeq \square \exp\left[\square \frac{k}{n} \log(p/k)\right]$$



Commentaires :

- En dimension "raisonnable", il existe des designs tels que le risque minimax est de l'ordre de $k \log\left(\frac{ep}{k}\right)$
  ex : Dantzig selector si Σ satisfait propriété d'isométrie restreinte.
- Explosion en très grande dimension.
  $\rightsquigarrow$ aucun design ne permet de retrouver θ.

Estimation du support et réduction de dimension

Definition

L'ensemble $\mathcal{C}_k^p(\rho)$ *corresponds au* $\theta \in \theta[k,p]$ *tels que* θ *contienne exactement* k *coefficient non nuls tous égaux à* $\rho/\sqrt{k}$.

Estimation du support et réduction de dimension

Definition

L'ensemble $\mathcal{C}_k^p(\rho)$ corresponds au $\theta \in \theta[k,p]$ tels que θ contienne exactement k coefficient non nuls tous égaux à $\rho/\sqrt{k}$.

Hypothèse : $k \leq p^{1/3}$
$\Sigma = 1$ suit une distribution gaussienne standard.
$\sigma^2 = 1$.

Proposition (réduction de dimension preque impossible)

$\rho^2 = \square \frac{k}{n} \log\left(\frac{p}{k}\right) \exp\left[\square' \frac{k}{n} \log\left(\frac{p}{k}\right)\right]$.

Il existe une constantes $0 < \delta < 1$ tel que pour tout ensemble $\widehat{M}$ de $\{1,\ldots,p\}$ de taille $p_0 \leq p^\delta$, on a

$$\sup_{\theta \in \mathcal{C}_k^p(\rho)} \mathbb{P}_{\theta,1}\left[\mathrm{supp}(\theta) \not\subseteq \widehat{M}\right] \geq 1/8 \ .$$

Commentaires :

- En très grande dimension, il est presque impossible d'estimer le support de θ.
- Il est même presque impossible de réduire efficacement la dimension du problème.

Simulations

$p = 5000$ and $p = 200$, $n = 50$.
$\sigma = 1$.
$\mathbf{X}$ suit une loi Gaussienne standard
$k = 1, \ldots, 15$.
$\theta_1 = \ldots = \theta_k = 4\sqrt{\log(p)/n} \approx 1.30$ (resp. 1.65) pour $p = 200$ (resp. $p = 5000$) et
$\theta_{k+1} = \ldots = \theta_p = 0$.
on a $\|\theta\|^2 = 16k \log(p)/n$.

Simulations

$p = 5000$ and $p = 200$, $n = 50$.
$\sigma = 1$.
X suit une loi Gaussienne standard
$k = 1, \ldots, 15$.
$\theta_1 = \ldots = \theta_k = 4\sqrt{\log(p)/n} \approx 1.30$ (resp. 1.65) pour $p = 200$ (resp. $p = 5000$) et $\theta_{k+1} = \ldots = \theta_p = 0$.
on a $\|\theta\|^2 = 16k \log(p)/n$.

Procédure de réduction de dimension. On applique les méthodes SIS (Lv et Fan) et Lasso pour réduire la dimension à un ensemble $\widehat{M}^S$ de taille $p_0 = 50$.
calcul de la puissance des procédures :

$$\text{Puissance} := \frac{\text{Card}[\widehat{M}^S \cap \{1, \ldots, k\}]}{k} .$$

Simulations

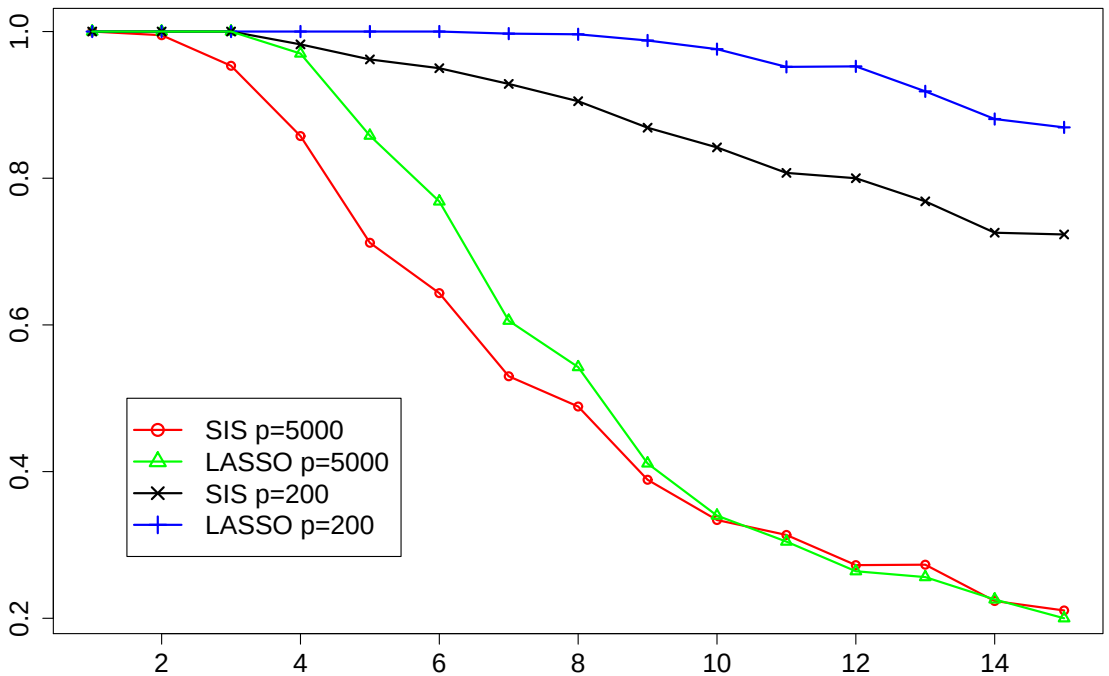


Figure: Puissance des procédures de réduction de dimension (SIS and LASSO)

θ tel que $\theta_1 = \ldots = \theta_k = u\sqrt{\log(p)/n}$ et $\theta_{k+1} = \ldots = \theta_p = 0$.
Calcul u_k^* le plus petit u tel que $\widehat{M}^L$ a une puissance plus grande que 0.9.

$\rightsquigarrow$ u_k^* correspond à l'intensité minimale du signal pour que la méthode de réduction de dimension n'oublie pas res covariables pertinentes.

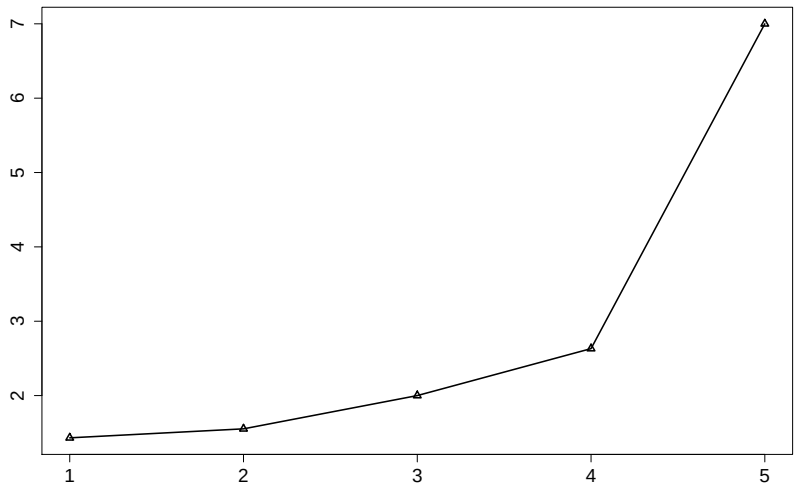


Figure: Signal minimal u_k^* en fonction de k.

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.
- un critère simple pour la très grande dimension :

$$\boxed{\frac{k \log(p/k)}{n} \geq 1/2.}$$

$\rightsquigarrow$ Il est presque impossible d'estimer θ ou même de faire de la réduction de dimension.
ex : $p = 5000$ and $n = 50$, $\rightsquigarrow k > 4$.
$p = 200$ and$n = 50$, $\rightsquigarrow k > 8$.

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.
- un critère simple pour la très grande dimension :

$$\boxed{\frac{k \log(p/k)}{n} \geq 1/2.}$$

$\rightsquigarrow$ Il est presque impossible d'estimer θ ou même de faire de la réduction de dimension.
ex : $p = 5000$ and $n = 50$, $\rightsquigarrow$ $k > 4$.
$p = 200$ and$n = 50$, $\rightsquigarrow$ $k > 8$.

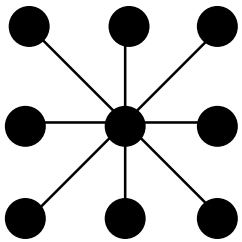
En pratique :

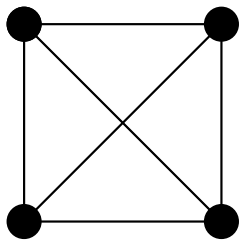
- Échantillon X non iid gaussien $\rightsquigarrow$ c'est encore pire
- Connaissances a priori.

Implications pour l'estimation des réseaux de gènes

En dimension raisonnable, un prix logarithmique $\log(p)$ à payer ....

GGM $\neq$ régression conditionelles... : estimation des clusters

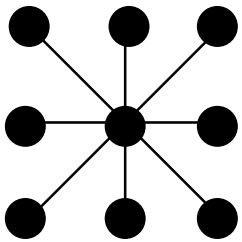


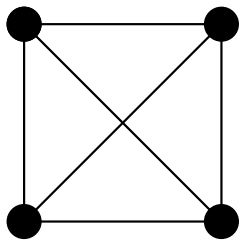


Implications pour l'estimation des réseaux de gènes

En dimension raisonnable, un prix logarithmique $\log(p)$ à payer ....

GGM $\neq$ régression conditionelles... : estimation des clusters





La limite structurelle

$$\boxed{\frac{k\log(p/k)}{n} \geq 1/2.}$$

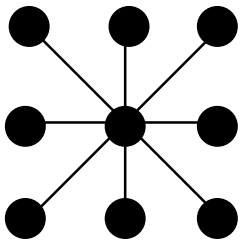
ne correspond pas tant au degré du graphe que la quantité :

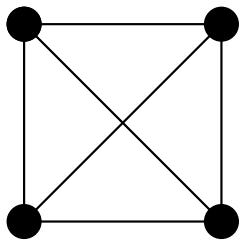
$$\sup_{a\in\Gamma}\left[\deg_g(a)\wedge\left(\sup_{b\sim_g a}\deg_g(b)\right)\right]$$

Implications pour l'estimation des réseaux de gènes

En dimension raisonnable, un prix logarithmique $\log(p)$ à payer ....

GGM $\neq$ régression conditionelles... : estimation des clusters





La limite structurelle

$$\boxed{\frac{k \log(p/k)}{n} \geq 1/2.}$$

ne correspond pas tant au degré du graphe que la quantité :

$$\sup_{a \in \Gamma} \left[\deg_g(a) \wedge \left(\sup_{b \sim_g a} \deg_g(b) \right) \right]$$

En pratique :

- Échantillon X non iid gaussien $\rightsquigarrow$ c'est encore pire
- Connaissances a priori/ intégration de d'autres types de données.