MAP 553
Apprentissage statistique

Christophe Giraud

CMAP, Ecole Polytechnique

PC6

1. Seuillage dur et C_p de Mallows
2. Estimateur Lasso et seuillage doux
3. Analyse Linéaire Discriminante

Seuillage dur

Régression linéaire

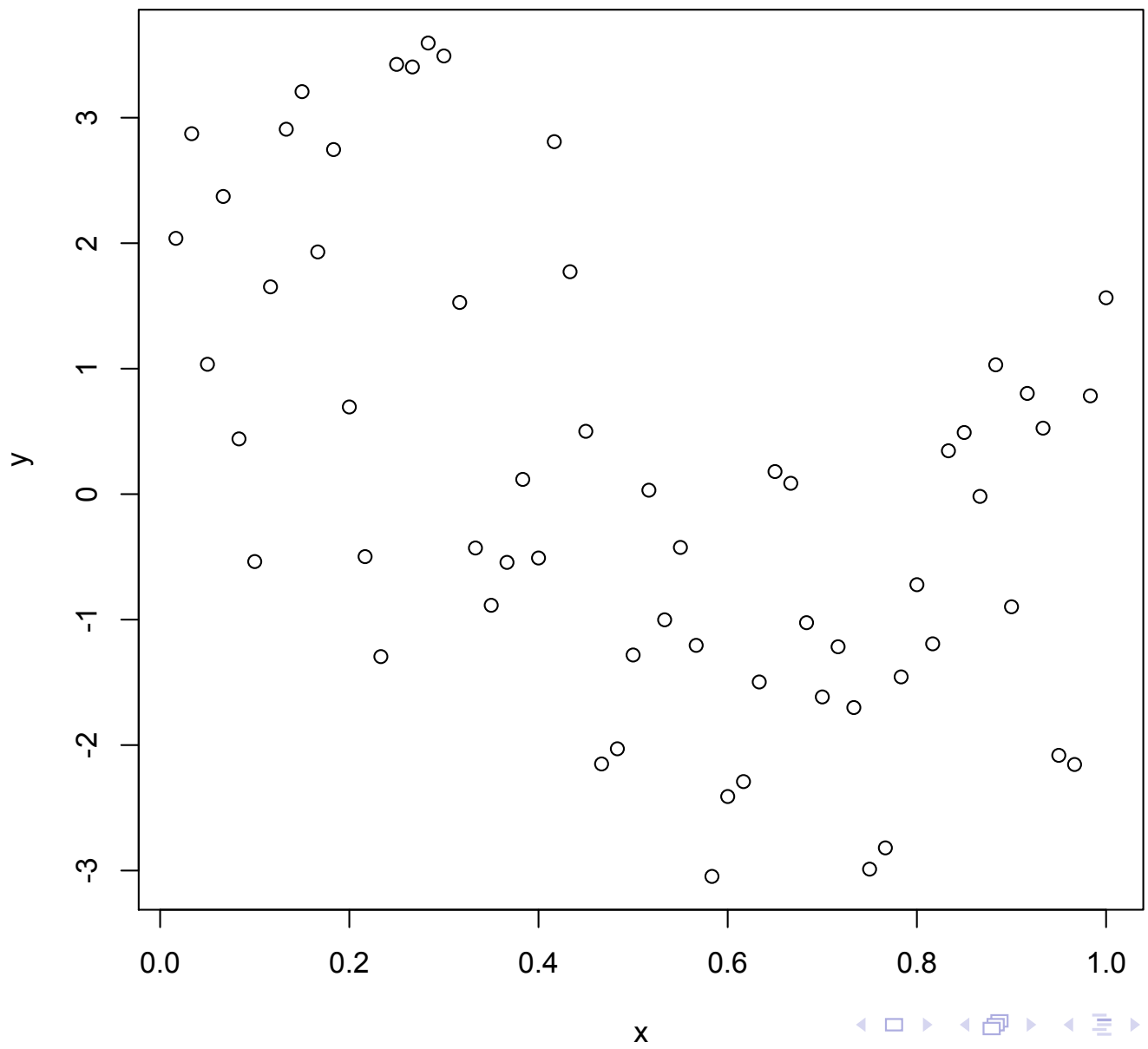
Modèle: $y = X\theta + \xi$

avec

- **observations:** $y \in \mathbb{R}^n$
- **design:** X matrice $n \times p$ connue (et fixe)
- **paramètre:** $\theta \in \mathbb{R}^p$ à estimer
- **bruit:** $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$

Exemple:

$X_{i,j} = \varphi_j(x_i)$ avec $x_i = i/n$ et φ_j la base trigonométrique.


observations
y
x
3
2
1
0
-1
-2
-3
0.0
0.2
0.4
0.6
0.8
1.0

Hypothèse (ORT)

Design orthogonal: $\frac{1}{n} X^T X = I_p$

Transformation des données:

$$z = \frac{1}{n} X^T y = \theta + \underbrace{\frac{1}{n} X^T \xi}_{= \zeta}$$

avec $\zeta \sim \mathcal{N}(0, \frac{\sigma^2}{n} I_p)$.

Notation

Pour $m \subset \{1, \dots, p\}$, on note $\hat{\theta}_m$ l'estimateur défini par

$$(\hat{\theta}_m)_j = z_j \mathbf{1}_{j \in m}, \quad \text{pour } j = 1, \dots, p.$$

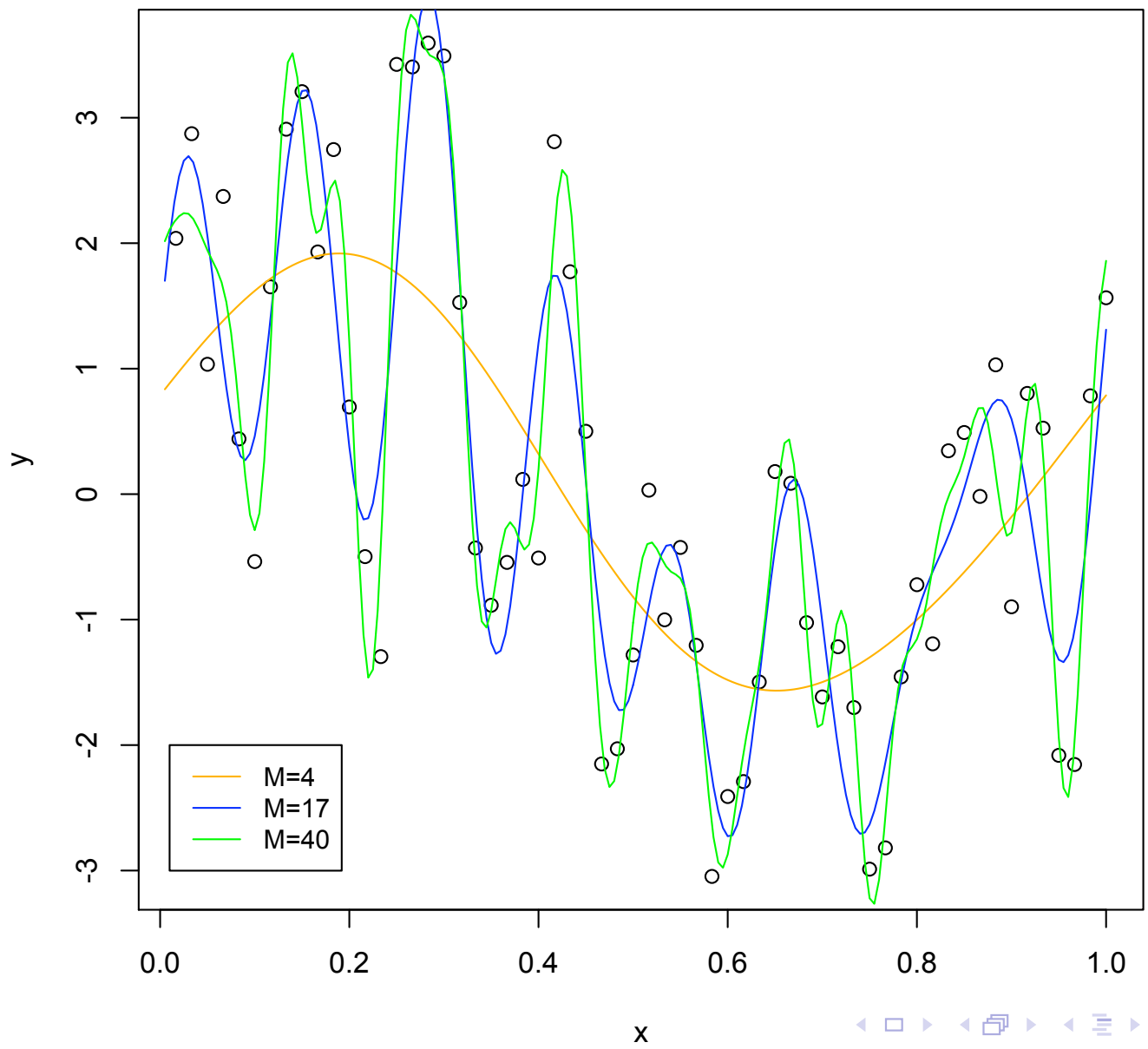
Autrement dit: $\hat{\theta}_m = \text{Proj}_{\text{vect}\{e_j,\ j \in m\}}(z)$ avec $(e_1, \dots, e_p)$ base canonique de $\mathbb{R}^p$.

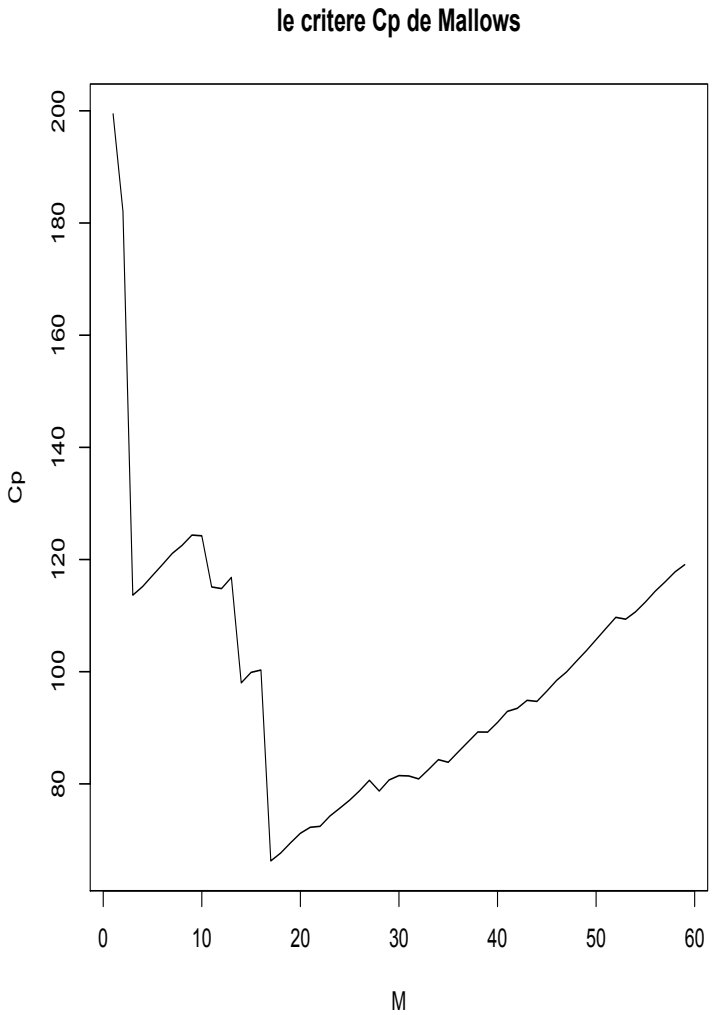
Estimateur "progressif"

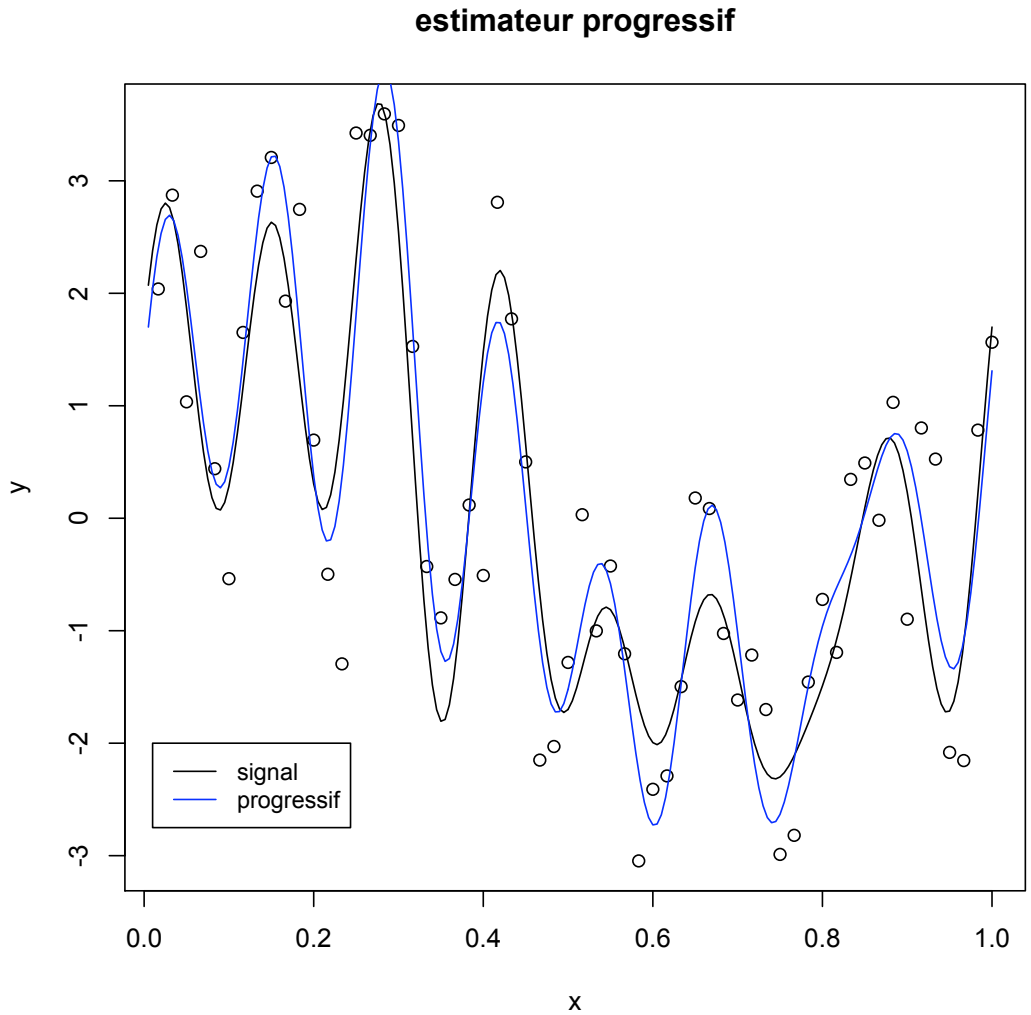
Estimateurs comme à la PC5: $\hat{\theta}_{\{1,\dots,M\}}, \quad M = 1, \dots, p.$

Estimateur progressif: $\hat{\theta}_{prog} = \hat{\theta}_{\{1,\dots,\hat{M}\}}$ avec $\hat{M}$ minimisant

$$\mathbf{C_p}\left(\hat{\theta}_{\{1,\dots,M\}}\right) = |z - \hat{\theta}_{\{1,\dots,M\}}|_2^2 + \frac{2\sigma^2 M}{n}.$$


pour differentes valeurs de M
y
x
M=4
M=17
M=40
3
2
1
0
-1
-2
-3
0.0
0.2
0.4
0.6
0.8
1.0


le critere Cp de Mallows
Cp
M
80
100
120
140
160
180
200
0
10
20
30
40
50
60

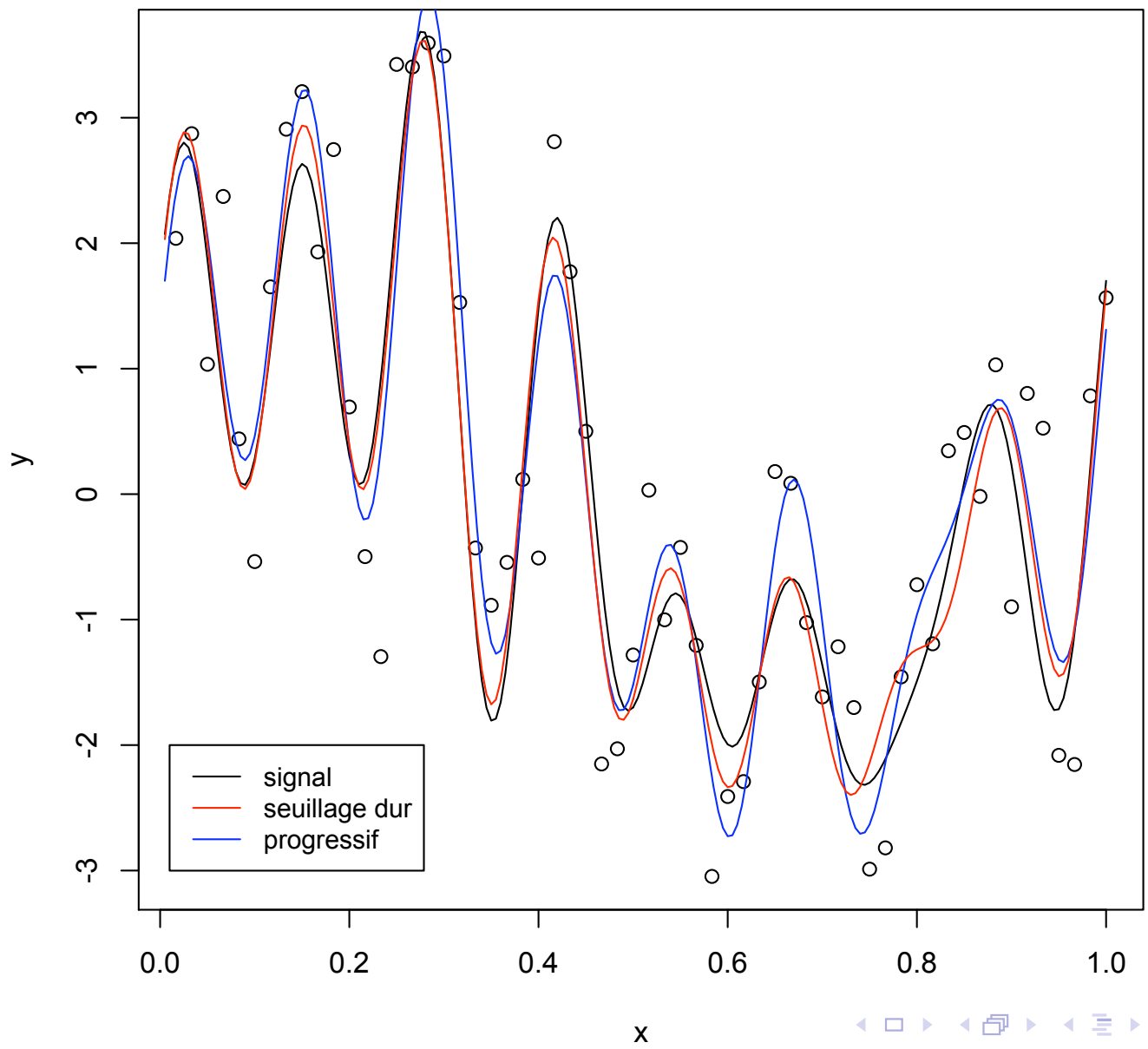

estimateur progressif
y
x
signal
progressif
-3
-2
-1
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0

Seuillage dur

Estimateur par seuillage dur: $\hat{\theta}^H$ défini pour $\tau > 0$ par

$$\hat{\theta}_j^H = z_j \mathbf{1}_{|z_j|>\tau}, \quad \text{pour } j = 1, \ldots, p.$$

Quel avantage de $\hat{\theta}^H$ sur $\hat{\theta}_{prog}$?


progressif versus seuillage dur
y
x
signal
seuillage dur
progressif
3
2
1
0
-1
-2
-3
0.0
0.2
0.4
0.6
0.8
1.0

Quel niveau de seuillage?

- ⚠ Le choix $\tau^2 = 2\sigma^2/n$ ne convient pas!
- Il faut prendre $\tau^2 = 2\sigma^2 \log(p)/n$ ou plus grand.

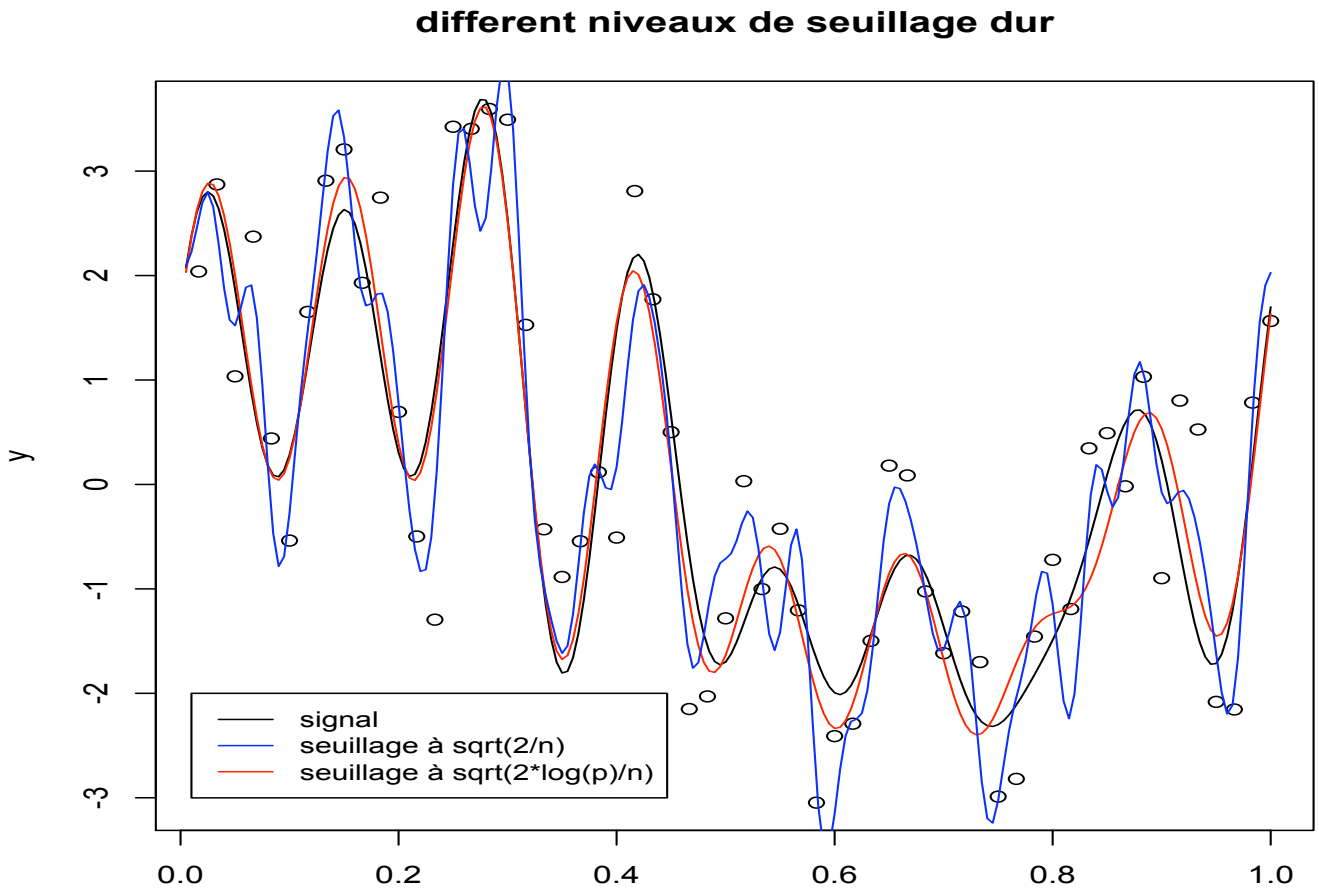


Figure: En bleu: $\tau^2 = 2\sigma^2/n$. En rouge: $\tau^2 = 2\sigma^2 \log(p)/n$.

Explication:

il faut seuiller à $\tau = \sqrt{2\sigma^2 \log(p)/n}$ pour ne pas prendre trop de "bruit".

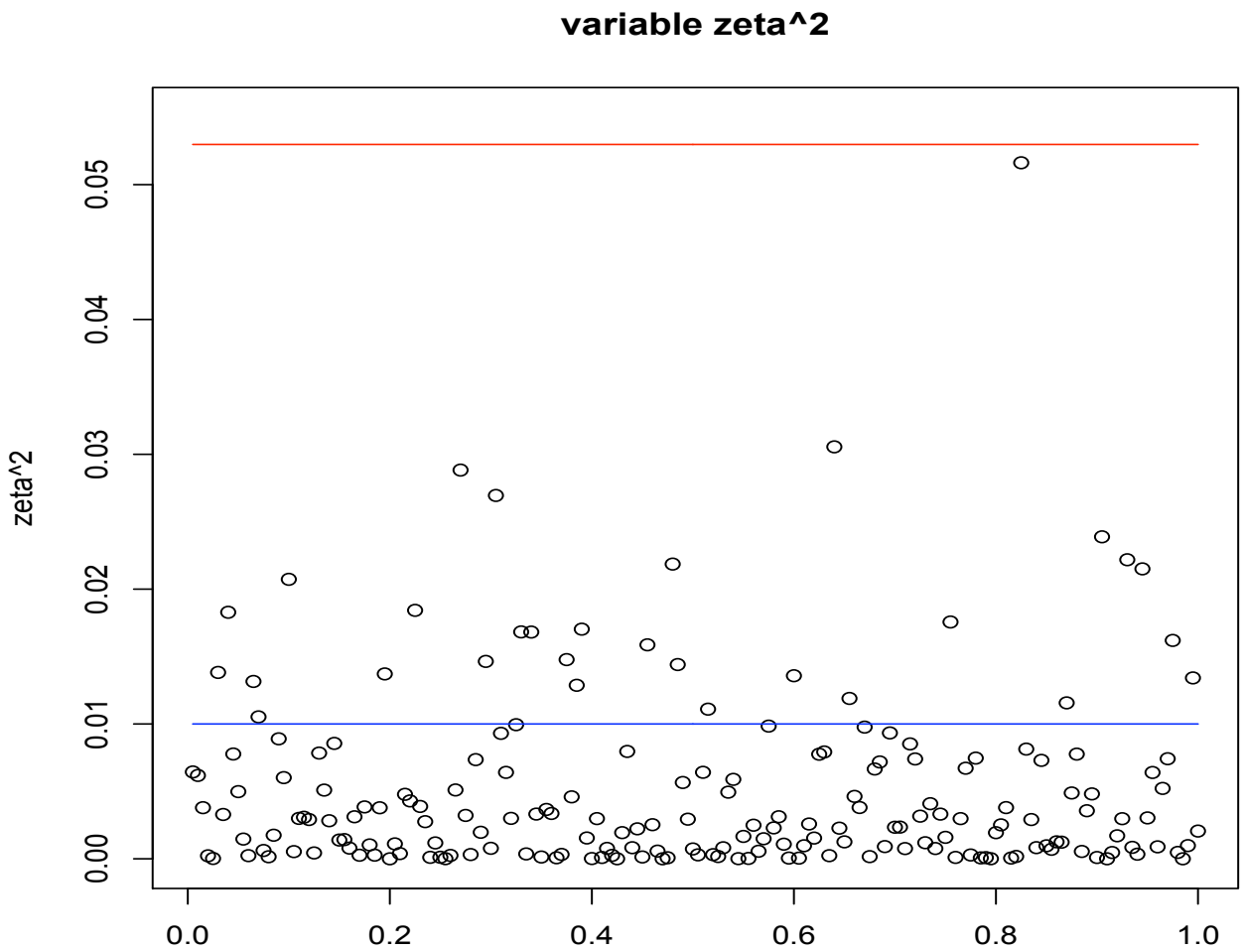


Figure: Valeurs de ζ_i^2. En bleu: seuil à $\tau^2 = 2\sigma^2/n$.
En rouge: seuil à $\tau^2 = 2\sigma^2 \log(p)/n$

Seuillage doux et Lasso

Estimateur Lasso

Défini pour $\tau > 0$ par:

$$\hat{\theta}^L = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} |y - X\theta|_2^2 + 2\tau \sum_{j=1}^{p} |\theta_j| \right\}. \qquad (1)$$

Design orthogonal

Sous l'hypothèse (ORT), le problème (1) est équivalent à

$$\hat{\theta}^L = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \sum_{j=1}^{p} (z_j - \theta_j)^2 + 2\tau \sum_{j=1}^{p} |\theta_j| \right\}.$$

avec $z = \frac{1}{n} X^T y$.

Estimateur Lasso avec Design Orthogonal

Sous l'hypothèse (ORT), l'estimateur Lasso

$$\hat{\theta}^L = \arg\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} |y - X\theta|_2^2 + 2\tau \sum_{j=1}^{p} |\theta_j| \right\}.$$

est donné par

$$\hat{\theta}_j^L = z_j \left(1 - \frac{\tau}{|z_j|} \right)_+, \quad j = 1, \ldots, p$$

où $(x)_+ = \max(x, 0)$ et $z = \frac{1}{n} X^T y$.

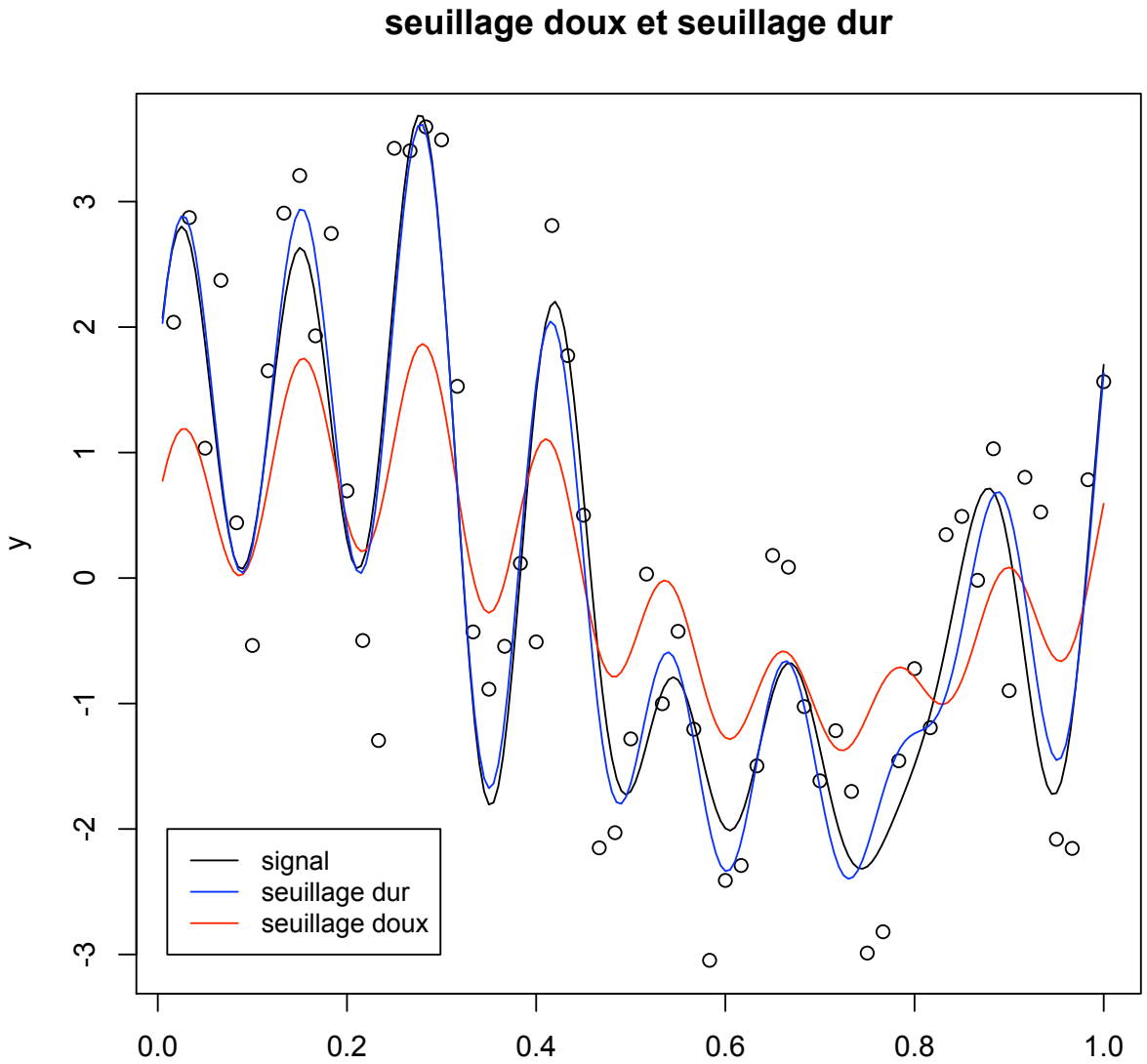


Figure: Seuillage dur (en bleu) et seuillage doux (en rouge) avec $\tau^2 = 2\sigma^2 \log(p)/n$

Design non-orthogonal

Cadre non-orthogonal: lorsque (ORT) n'est pas vérifiée, le problème d'optimisation

$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} |y - X\theta|_2^2 + \tau^2 \sum_{j=1}^{p} \mathbf{1}_{\{\theta_j \neq 0\}} \right\}$$

est **NP-hard** en général alors que le problème

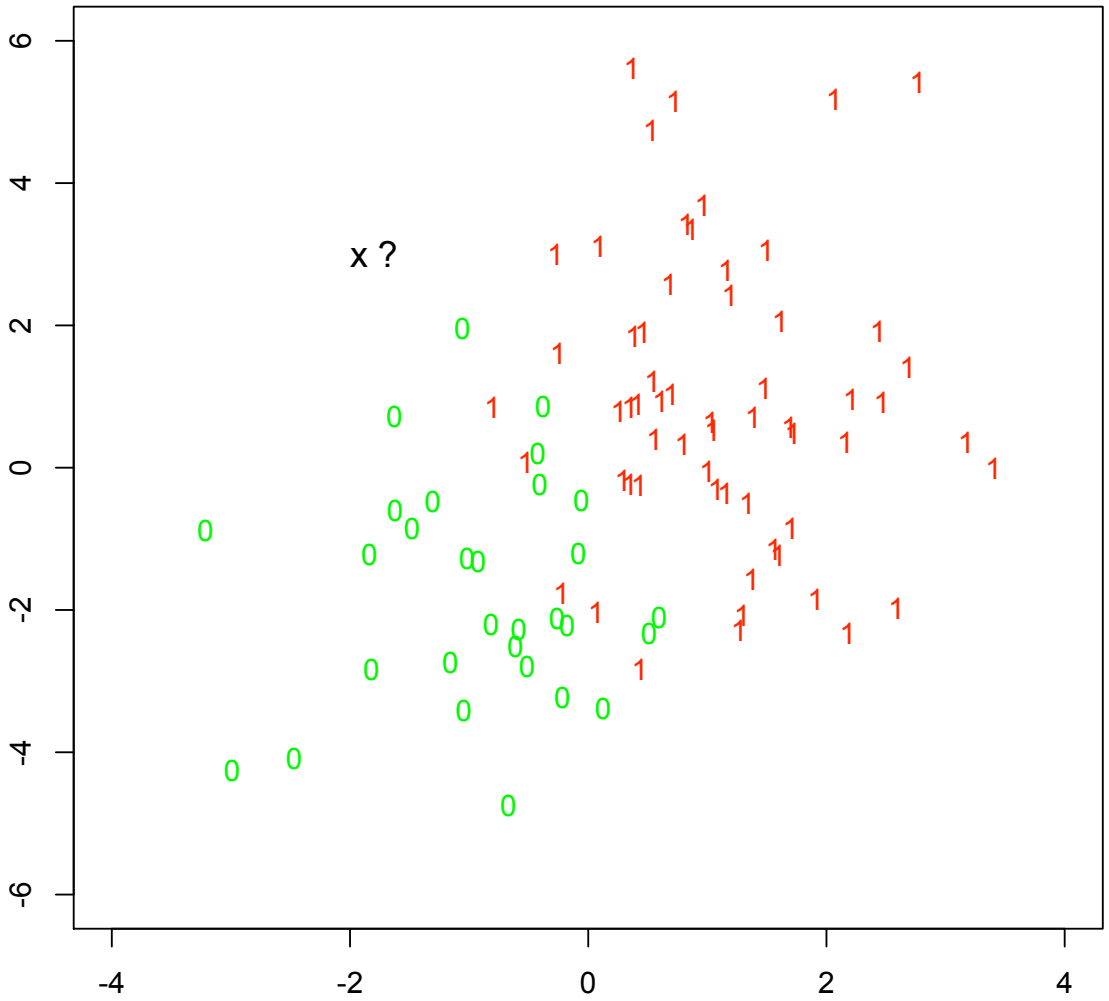
$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} |y - X\theta|_2^2 + 2\tau \sum_{j=1}^{p} |\theta_j| \right\}$$

est un problème d'**optimisation convexe** pour lequel il existe des algorithmes efficaces.

Classification

Données:

points $X_i \in \mathbb{R}^p$ avec label $Y_i \in \{0, 1\}$ pour $i = 1, \ldots, n$.



Objectif: prédire la classe d'un nouveau point x.

Cadre statistique

On supposera les (X_i, Y_i) i.i.d. avec

- $\mathbb{P}(Y_i = k) = \pi_k$, pour $k = 0, 1$
- $\text{Loi}(X_i | Y_i = k) = g_k(x)\,dx$, pour $k = 0, 1$.

Classifieur de Bayes

Le classifieur $h_* : \mathbb{R}^p \to \{0, 1\}$ défini par

$$h_*(x) = \mathbf{1}_{\{\pi_1 g_1(x) > \pi_0 g_0(x)\}}$$

vérifie

$$\mathbb{P}(h_*(X) \neq Y) = \min_h \mathbb{P}(h(X) \neq Y).$$

Preuve:

L'égalité $\min(\pi_0 g_0, \pi_1 g_1) = (\pi_0 g_0 \mathbf{1}_{h_*=1} + \pi_1 g_1 \mathbf{1}_{h_*=0})$ donne

$$\begin{aligned}
\mathbb{P}(h(X) \neq Y) &= \pi_0 \mathbb{P}(h(X) = 1 | Y = 0) + \pi_1 \mathbb{P}(h(X) = 0 | Y = 1) \\
&= \int \pi_0 g_0 \mathbf{1}_{h=1} + \int \pi_1 g_1 \mathbf{1}_{h=0} \\
&\geq \int (\pi_0 g_0 \mathbf{1}_{h_*=1} + \pi_1 g_1 \mathbf{1}_{h_*=0})(\mathbf{1}_{h=1} + \mathbf{1}_{h=0}) \\
&\geq \underbrace{\int \pi_0 g_0 \mathbf{1}_{h_*=1} + \int \pi_1 g_1 \mathbf{1}_{h_*=0}}_{= \mathbb{P}(h_*(X) \neq Y)} .
\end{aligned}$$

□

Cadre gaussien

$$\text{Loi}(X_i | Y_i = k) \;=\; \mathcal{N}(\mu_k, \Sigma_k), \quad \text{pour } k = 0, 1$$

c'est à dire

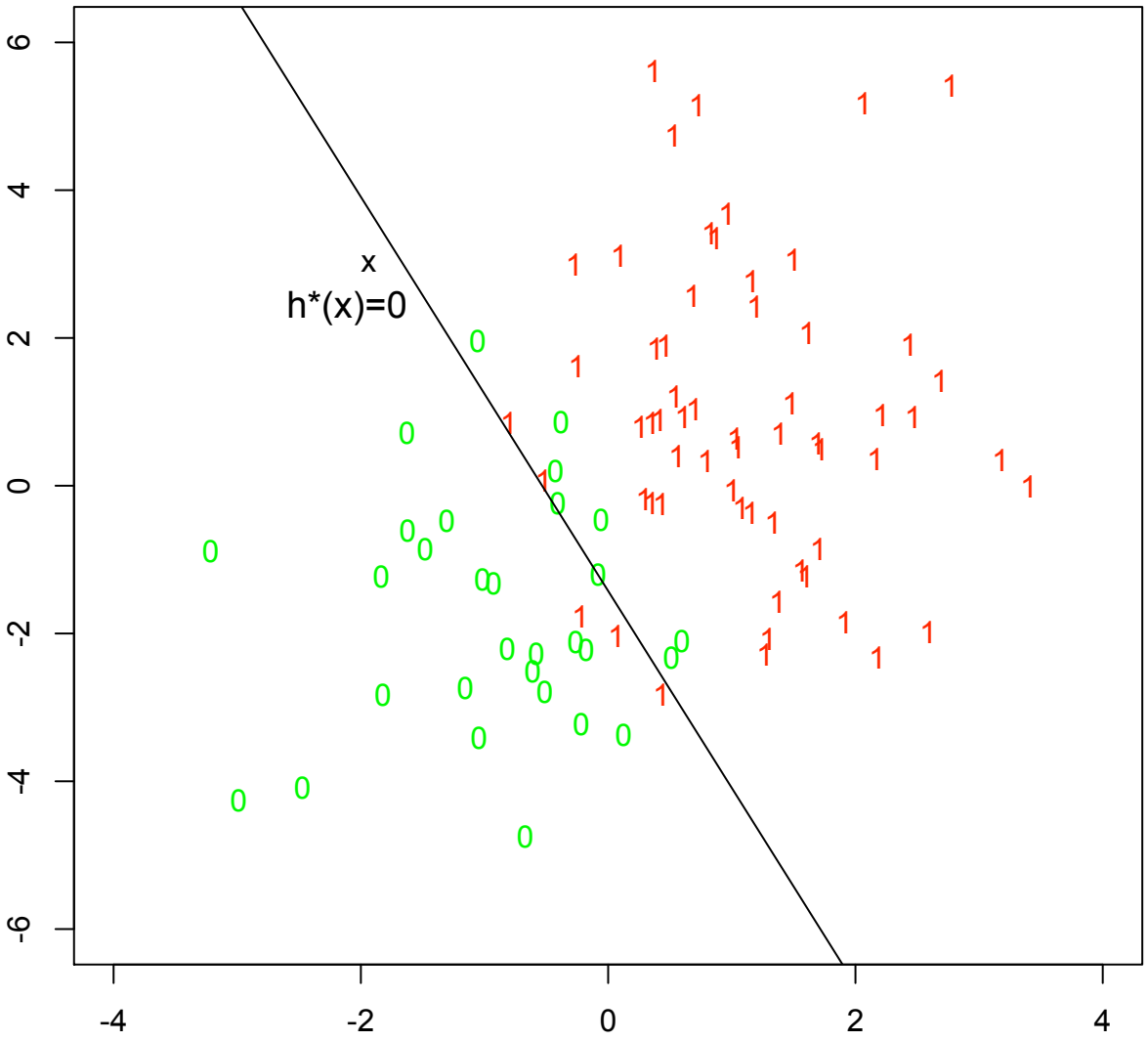
$$g_k(x) = (2\pi)^{-p/2} \sqrt{\det(\Sigma_k^{-1})} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right).$$

Cas où $\Sigma_0 = \Sigma_1 = \Sigma$

Lorsque $\Sigma_0 = \Sigma_1 = \Sigma$ on a

$$h_*(x) = 1 \iff (\mu_1 - \mu_0)^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_0}{2} \right) > \log(\pi_0/\pi_1).$$

Frontière du classifieur de Bayes



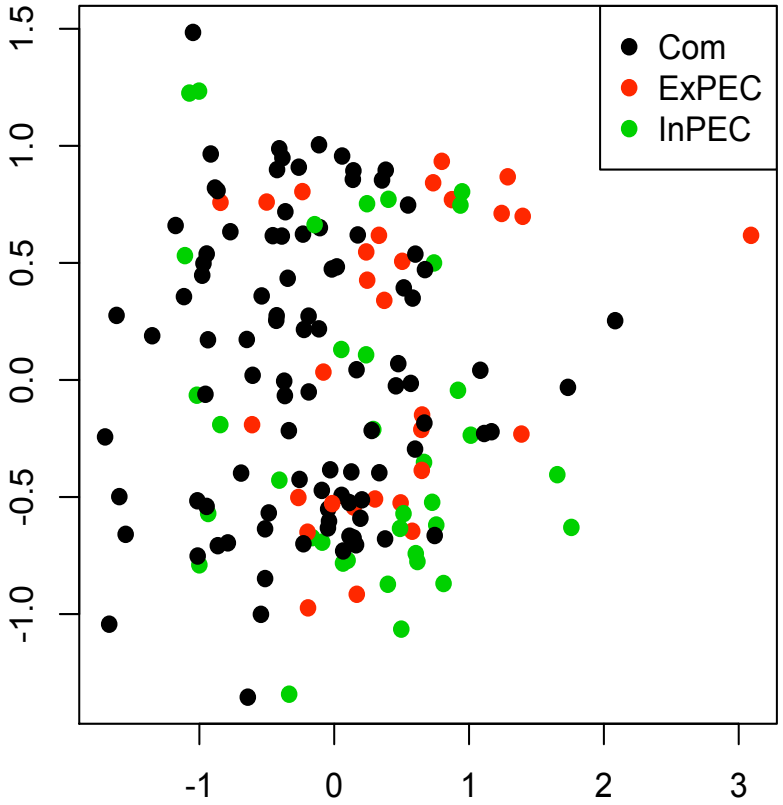
Cas où $\Sigma_1 = \Sigma_0 = \Sigma$

ACP versus ALD pour réduire la dimension

Exemple:

- X = mesure de la composition chimique (55 composés) de $n = 162$ souches de *E. coli*.
- Y = souche commensale ou pathogène

ACP



ALD