

# SÉLECTION DE MODÈLES EN RÉGRESSION

$$Y_i = \mu_i + \sigma \varepsilon_i, \quad i = 1, \dots, n$$

$$\varepsilon_1, \dots, \varepsilon_n \text{ i.i.d. } \mathcal{N}(0, 1)$$

Sélection de variables :

choix des variables explicatives pertinentes

$$(Y_i, X_i^1, \dots, X_i^p, p \leq n), \quad i = 1, \dots, n$$

modèle  $m = \{i_1, \dots, i_k\} \subset \{1, \dots, p\}$

$$m \leftrightarrow S_m = \text{e.v. } \{ \vec{X}^j, j \in m \}$$

$$\vec{\mu} \in S_m \leftrightarrow \vec{\mu} = \vec{\mu}_m = \sum_{j \in m} \beta_j \vec{X}^j$$

Cas ordonné :

Ex: régression polynomiale,  $\mu_i = \mu(x_i)$ ,  $X_i^j = x_i^{j-1}$ .

$k$ ,  $m_k = \{1, \dots, k\}$ ,  $k = 1, \dots$

Cas non ordonné :  $k$ ,  $C_p^k$   $k$ -uplets possibles

## Estimer un signal

en choisissant la meilleure partition

$$(Y_i, x_i), i = 1, \dots, n, \mu_i = \mu(x_i), 0 \leq x_1 < \dots < x_n \leq 1.$$

modèle  $m = (I_1, \dots, I_k)$ , partition de  $[0, 1]$ .

$$m \leftrightarrow S_m = \text{e.v. } \{ \vec{X}^j, j \in m \} \text{ où } X_i^j = I_{x_i \in I_j}$$

$$\vec{\mu} \in S_m \leftrightarrow \vec{\mu} = \vec{\mu}_m = \sum_{j \in m} \beta_j \vec{X}^j$$

Analyse différentielle : estimer le nombre de com-

posantes non nulles et leur emplacement

$$\text{modèle } m = \{i_1, \dots, i_k\} \in \{1, \dots, n\}$$

$$m \leftrightarrow S_m = \text{e.v. } \{ \vec{e}^j, j \in \{1, \dots, n\} \}$$

$$\vec{\mu} \in S_m \leftrightarrow \vec{\mu} = \vec{\mu}_m = \sum_{j \in m} \beta_j \vec{e}^j, \mu_i = 0 \text{ si } i \in \bar{m}$$

$$Y_i = \mu_i + \sigma \varepsilon_i, \quad i = 1, \dots, n$$

$\varepsilon_1, \dots, \varepsilon_n$  i.i.d.  $\mathcal{N}(0, 1)$ ,  $\sigma$  connu

famille de modèles  $\{S_m\}_{m \in \mathcal{M}}$

$\hat{\boldsymbol{\mu}}_m = \Pi_{S_m} Y$ , e.m.c.  $\boldsymbol{\mu}$  sur  $S_m$ . On note  $\tilde{\boldsymbol{\mu}}_m = \Pi_{S_m} \tilde{\boldsymbol{\mu}}$ .

Risque de  $\hat{\boldsymbol{\mu}}_m$  :

$$\mathbb{E} \|\hat{\boldsymbol{\mu}}_m - \tilde{\boldsymbol{\mu}}\|^2 = \|\tilde{\boldsymbol{\mu}}_m - \tilde{\boldsymbol{\mu}}\|^2 + \sigma^2 D_m, \quad \text{où } D_m = \dim(S_m)$$

Chercher  $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{\hat{m}}$ , qui minimise le risque  $\mathbb{E} \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2$ .

$$\mathbb{E} \|\hat{\boldsymbol{\mu}} - \tilde{\boldsymbol{\mu}}\|^2 \leq C \left( \inf_m \mathbb{E} \|\hat{\boldsymbol{\mu}}_m - \tilde{\boldsymbol{\mu}}\|^2 + q \right)$$

Heuristique :

$$\begin{aligned}
\min \|\vec{\mu}_m - \vec{\mu}\|^2 + \sigma^2 D_m &= \min \|\vec{\mu}_m - \vec{\mu}\|^2 + \sigma^2 D_m - \|\vec{\mu}\|^2 \\
&= \min \underbrace{-\|\vec{\mu}_m\|^2}_{-\|\hat{\mu}_m\|^2 + \sigma^2 D_m} + \sigma^2 D_m \\
&= \min \|Y - \hat{\mu}_m\|^2 + 2\sigma^2 D_m \\
&= \min \|Y - \hat{\mu}_m\|^2 + \sigma^2 \mathbf{pen}(m)
\end{aligned}$$

Théorème :  $\{L_m\}_{m \in \mathcal{M}}$ ,  $L_m > 0$ ,  $\Sigma = \sum_{m \in \mathcal{M}} e^{-D_m L_m} < \infty$

$$\mathbf{pen}(m) \geq K \sigma^2 D_m \left(1 + \sqrt{2L_m}\right)^2, \quad K > 1$$

$$\begin{aligned}
\mathbb{E}\|\hat{\mu} - \vec{\mu}\|^2 &\leq C(K) \left( \inf_m \left\{ \|\vec{\mu}_m - \vec{\mu}\|^2 + \sigma^2 \mathbf{pen}(m) \right\} \right. \\
&\quad \left. + (K + 1)\sigma^2 \Sigma \right)
\end{aligned}$$

ou bien

$$\mathbf{pen}(m) = K \sigma^2 D_m \left(1 + \sqrt{2L_m}\right)^2$$

$$\mathbb{E}\|\hat{\mu} - \vec{\mu}\|^2 \leq C(K) \left( \inf_m \left\{ \|\vec{\mu}_m - \vec{\mu}\|^2 + \sigma^2 D_m (1 + L_m) \right\} + \sigma^2 \Sigma \right)$$

Preuve : comment apparaissent les poids  $L_m$  ?

D'abord, on montre  $\forall m, \forall \xi > 0$

$$\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C(K) (\{\|\vec{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 + \mathbf{pen}(m)\} + \sigma^2(K+1)\xi)$$

sauf sur  $\Omega_\xi$ , t.q.  $\mathbb{P}(\overline{\Omega}_\xi) \leq \Sigma e^{-\xi}$

On en déduira

$$\mathbb{E}\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C(K) (\{\|\vec{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 + \mathbf{pen}(m)\} + \sigma^2(K+1)\Sigma)$$

On fixe  $m$ ,

$$\mathbf{crit}(m') \leq \mathbf{crit}(m)$$

$$\begin{aligned} \|Y - \widehat{\boldsymbol{\mu}}_{m'}\|^2 + \mathbf{pen}(m') &\leq \|Y - \widehat{\boldsymbol{\mu}}_m\|^2 + \mathbf{pen}(m) \\ &\leq \|Y - \boldsymbol{\mu}_m\|^2 + \mathbf{pen}(m) \end{aligned}$$

Truc :

$$\|Y - \boldsymbol{\mu}_m\|^2 = \|\boldsymbol{\mu} - \boldsymbol{\mu}_m\|^2 - 2 \langle \boldsymbol{\varepsilon}, \boldsymbol{\mu}_m \rangle + \sigma^2 \|\boldsymbol{\varepsilon}\|^2 + 2 \langle \boldsymbol{\varepsilon}, \boldsymbol{\mu} \rangle$$

d'où

$$\begin{aligned} \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{m'}\|^2 &\leq \|\boldsymbol{\mu} - \boldsymbol{\mu}_m\|^2 + \mathbf{pen}(m) \\ &\quad + 2\sigma \langle \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\mu}}_{m'} - \boldsymbol{\mu}_m \rangle - \mathbf{pen}(m') \end{aligned}$$

Etudier  $Z(t) = \frac{\langle \boldsymbol{\varepsilon}, \vec{t} - \boldsymbol{\mu}_m \rangle}{\|\vec{t} - \boldsymbol{\mu}_m\|}$ , Proc. gaussien, variance 1

$$\mathbb{P}\left(\sup_{\vec{t} \in S_{m'}} Z(t) \geq \underbrace{\mathbb{E} \sup_{\vec{t} \in S_{m'}} Z(t)}_{\sqrt{D_m + D_{m'}}} + x_{m'}\right) \leq \exp -\frac{x_{m'}^2}{2}$$

On fixe  $m$ ,  $\mathbf{crit}(m') \leq \mathbf{crit}(m)$

$$\begin{aligned} \|\vec{\mu} - \widehat{\vec{\mu}}_{m'}\|^2 &\leq \|\vec{\mu} - \vec{\mu}_m\|^2 + \mathbf{pen}(m) \\ &\quad + 2\sigma \langle \vec{\varepsilon}, \widehat{\vec{\mu}}_{m'} - \vec{\mu}_m \rangle - \mathbf{pen}(m') \end{aligned}$$

Pour chaque  $m'$ ,  $\forall t \in S_{m'}$

$$\langle \vec{\varepsilon}, \vec{t} - \vec{\mu}_m \rangle \leq \|\vec{t} - \vec{\mu}_m\| \left( \sqrt{D_m + D_{m'}} + x_{m'} \right)$$

avec proba  $\geq 1 - \exp(-x_{m'}^2/2)$

Bonferroni  $\Rightarrow \forall m', \forall t \in S_{m'}$

$$\langle \vec{\varepsilon}, \vec{t} - \vec{\mu}_m \rangle \leq \|\vec{t} - \vec{\mu}_m\| \left( \sqrt{D_m + D_{m'}} + x_{m'} \right)$$

avec proba  $\geq 1 - \sum_{m' \in \mathcal{M}} \exp(-x_{m'}^2/2)$

D'où

$$x_{m'} = \sqrt{L_{m'} D_{m'} + \xi \sqrt{2}} \Rightarrow \mathbf{proba} \geq 1 - \Sigma \exp(-\xi)$$

On en est là :

$$\begin{aligned} \|\hat{\vec{\mu}} - \hat{\vec{\mu}}_{m'}\|^2 &\leq \|\vec{\mu} - \vec{\mu}_m\|^2 + \mathbf{pen}(m) - \mathbf{pen}(m') \\ + 2\sigma \|\hat{\vec{\mu}}_{m'} - \vec{\mu}_m\| &\left( \sqrt{D_m + D_{m'}} + \sqrt{2} \sqrt{L_{m'} D_{m'}} + \xi \right) \end{aligned}$$

- $\|\hat{\vec{\mu}}_{m'} - \vec{\mu}_m\| \leq \|\vec{\mu} - \hat{\vec{\mu}}_{m'}\| + \|\vec{\mu} - \vec{\mu}_m\|$
- $2ab \leq ca^2 + b^2/c, \forall c > 0$
- $\mathbf{pen}(m') \geq K\sigma^2 D_{m'} (1 + \sqrt{2L_{m'}})$

$\rightsquigarrow$  résultat :  $\forall m, \forall \xi > 0$

$$\|\hat{\vec{\mu}} - \vec{\mu}\|^2 \leq C(K) \left( \|\vec{\mu}_m - \vec{\mu}\|^2 + \mathbf{pen}(m) \right) + \sigma^2 (K + 1) \xi$$

sauf sur un ensemble de proba  $\leq \Sigma e^{-\xi}$ .

Théorème :  $\{L_m\}_{m \in \mathcal{M}}$ ,  $L_m > 0$ ,  $\Sigma = \sum_{m \in \mathcal{M}} e^{-D_m L_m} < \infty$

si  $\mathbf{pen}(m) \geq K \sigma^2 D_m (1 + \sqrt{2L_m})^2$ ,  $K > 1$ , alors

$$\mathbb{E} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C_1 \left( \inf_m \{ \|\widehat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 + \sigma^2 \mathbf{pen}(m) \} + C_2 \sigma^2 \Sigma \right)$$

ou bien si  $\mathbf{pen}(m) = K \sigma^2 D_m (1 + \sqrt{2L_m})^2$ , alors

$$\mathbb{E} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 \leq C_1 \left( \inf_m \{ \|\widehat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 + \sigma^2 D_m (1 + L_m) \} + \sigma^2 \Sigma \right)$$

Sélection de variables ordonnées :  $\vec{X}^1, \dots, \vec{X}^p$ ,  $p = p(n) \uparrow n$

$$S_m = \text{e.v.} \left\{ \vec{X}^1, \dots, \vec{X}^m \right\}, m = 1, \dots, p, D_m = m.$$

$$\Sigma = \sum_{m=1}^p e^{-mL_m}, L_m = L, \Sigma \leq (e^L - 1)^{-1}$$

Si  $K = 2 / \left( 1 + \sqrt{2L} \right)$  ( $K > 1$  si  $L < 3/2 - \sqrt{2}$ ), et

$\mathbf{pen}(m) = 2\sigma^2 m$ , alors,

$$\begin{aligned} \mathbb{E} \|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 &\leq C(L) \left( \inf_m \{ \|\widehat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 + \sigma^2 m (1 + L) \} + \sigma^2 \Sigma \right) \\ &\leq C(L) \inf_m \{ \|\widehat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 + \sigma^2 m \} \\ &\leq C(L) \inf_m \mathbb{E} \|\widehat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 \end{aligned}$$

Sélection de variables non ordonnées :  $\vec{X}^1, \dots, \vec{X}^p$

$$m = \{i_1, \dots, i_{|m|}\} \subset \{1, \dots, p\}, S_m = \text{e.v. } \{\vec{X}^1, \dots, \vec{X}^{|m|}\}$$

$$\Sigma = \sum_{m \in \mathcal{M}} e^{-|m|L_m} = \sum_{|m|=1}^p C_p^{|m|} e^{-|m|L_m}$$

$$\text{si } L_m = L, \Sigma = (1 + e^{-L})^p - 1 \quad (p = p(n) \nearrow n)$$

$$\text{si } L = \log(p/c), \text{ alors } \Sigma < e^c - 1.$$

Autre choix : les  $L_m$  varient avec  $|m|$

$$\text{si } L_m = 1 + \log(c) + \log(p/|m|), \text{ alors } \Sigma < 1/(c - 1).$$

Théorème  $\Rightarrow$

$$\text{si } \text{pen}(m) = K|m|\sigma^2(1 + c_1\sqrt{\log(p/|m|)} + c_2\log(p/|m|))$$

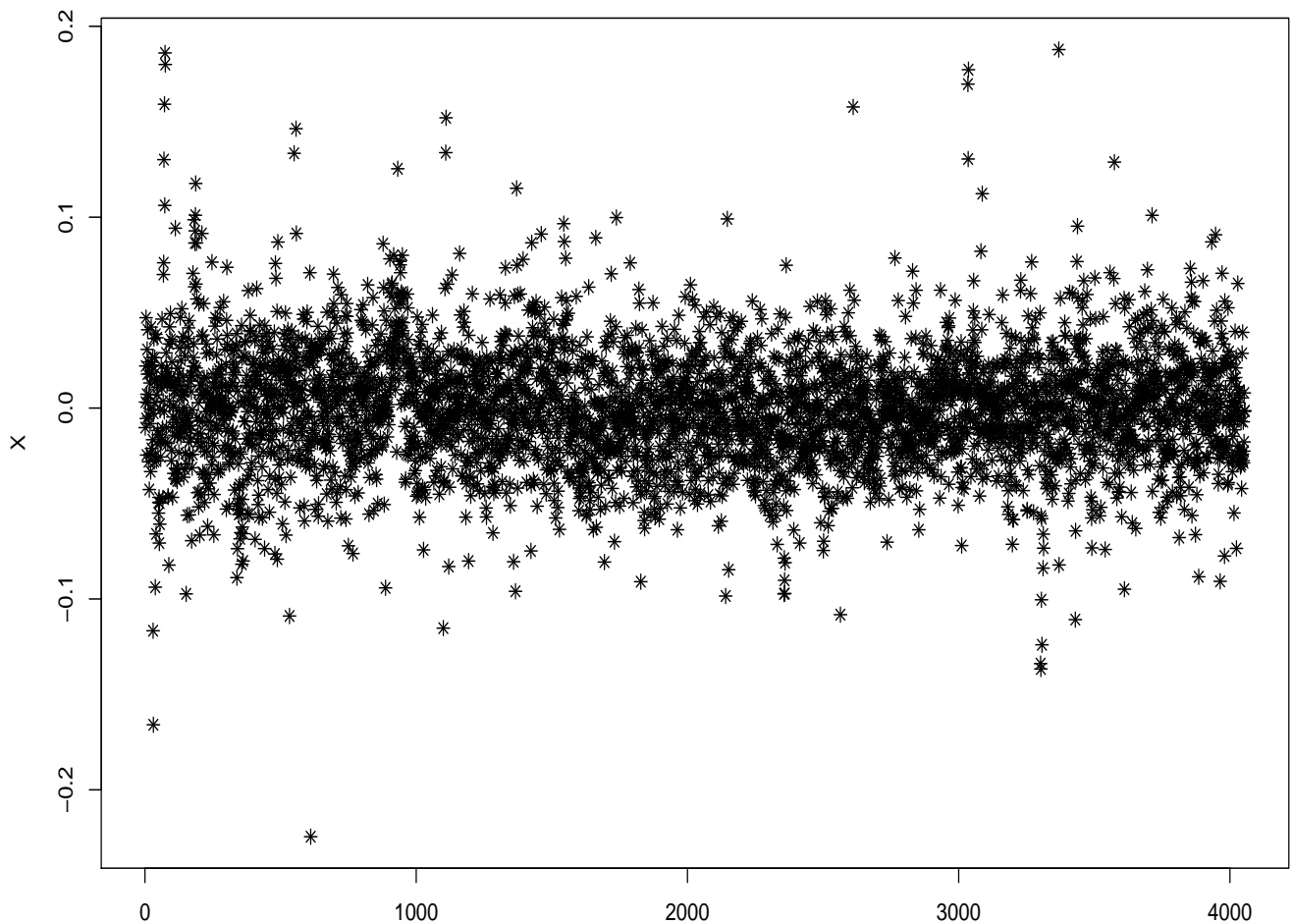
$$\mathbb{E}\|\hat{\vec{\mu}} - \vec{\mu}\|^2 \leq C \inf_m \left\{ \|\vec{\mu}_m - \vec{\mu}\|^2 + \sigma^2|m| \left( 1 + \log\left(\frac{p}{|m|}\right) \right) \right\}$$

Un autre Théorème :

$$\bar{m} \in \mathcal{M}, \text{ pen}(\bar{m}) \leq c\sigma^2|\bar{m}|\log(p), \quad 0 < c < 2, \text{ alors}$$

$$\mathbb{E}\|\hat{\vec{\mu}} - \vec{\mu}\|^2 \geq C\sigma^2\log(p)$$

Analyse différentielle : Etude des différences d'expression des gènes selon le milieu de culture de *B. subtilis*, **methionine** or ou **methythioribose** à partir de “puces à ADN”.



$\mu = \mu_{m_0}$ ,  $m_0 = \{1, \dots, k_0\}$ ,  $k_0$  inconnu,  $k_0 < n$

$\mu_i \neq 0$  si  $1 \leq i \leq k_0$   $\mu_i = 0$  si  $i > k_0$

$\hat{\boldsymbol{\mu}}_m$  :  $(\hat{\boldsymbol{\mu}}_m)_i = Y_i, i \in m$ ,  $(\hat{\boldsymbol{\mu}}_m)_i = 0$  sinon.

$$\|Y - \hat{\boldsymbol{\mu}}_m\|^2 = \sum_{i \in \bar{m}} Y_i^2$$

Remarque :

$$\inf_{m \in \mathcal{M}} \mathbb{E} \|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{k_0} \min \{\mu_i, \sigma^2\} < \mathbb{E} \|\hat{\boldsymbol{\mu}}_{m_0} - \boldsymbol{\mu}\|^2 = k_0 \sigma^2$$

$\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_{\hat{m}}$  :  $\hat{m}$  minimise  $\mathbf{crit}(m) = \sum_{i \in \bar{m}} Y_i^2 + \sigma^2 \mathbf{pen}(|m|)$

En pratique :

$$\min_{m \in \mathcal{M}} \mathbf{crit}(m) = \min_{1 \leq k \leq k_n} \min_{|m|=k} \sum_{i \in \bar{m}} Y_i^2 + \sigma^2 \mathbf{pen}(k)$$

Si  $Y_{\ell_1}^2 > \dots > Y_{\ell_n}^2$ ,

$$\min_{m \in \mathcal{M}} \mathbf{crit}(m) = \min_{1 \leq k \leq k_n} \sum_{i \geq k+1} Y_{\ell_i}^2 + \sigma^2 \mathbf{pen}(k)$$

et donc  $\hat{m} = m_{\hat{k}} = (\ell_1, \dots, \ell_{\hat{k}})$

Critère pénalisé = seuillage

Seuillage : Choisir  $t(1) > \dots > t(n)$ , et comparer les

$|Y_{\ell_i}|$  aux  $t(i)$  :

$$\hat{k} = 0 \text{ si } |Y_{\ell_k}| < t(k), \forall k \text{ sinon } \hat{k} = \max_k \{|Y_{\ell_k}| \geq t(k)\}$$

Lien avec critère pénalisé :

$$t^2(k) = \mathbf{pen}(k) - \mathbf{pen}(k-1), \text{ ou bien } \mathbf{pen}(k) = \sum_{l=1}^k t^2(l).$$

Donoho et Johnston (94, Biometrika),

$$t^2(k) = t^2 = 2\sigma^2 \log(n) \iff \mathbf{pen}(k) = 2\sigma^2 k \log(n)$$

## Comment calibrer la pénalité ?

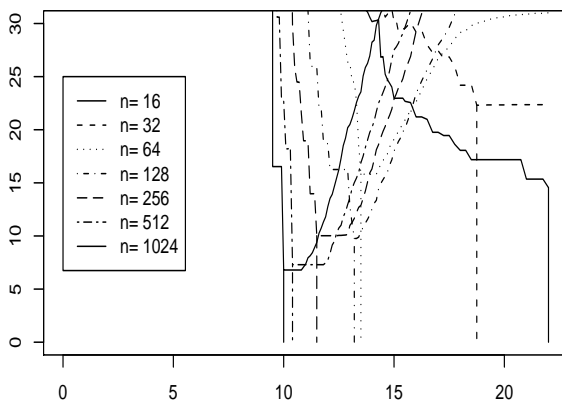
$$\text{pen}(m) = K|m|\sigma^2(1 + \alpha\sqrt{\log(n/|m|)} + \beta\log(n/|m|))$$

$$\text{pen}(k; c_1, c_2) = k\sigma^2(c_1 \log(n/k) + c_2)$$

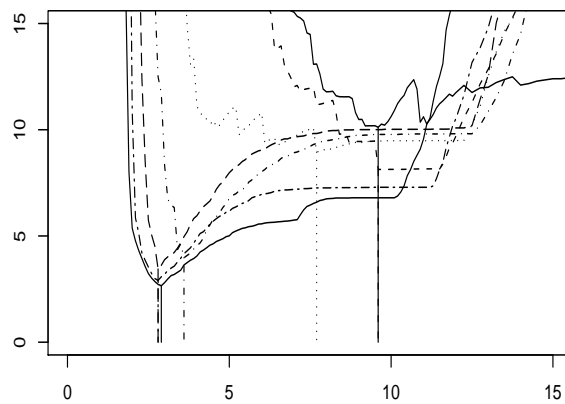
Pour chaque  $n$ , minimiser le rapport des risques :

$$r_n(c_1, c_2) = \sup_{\vec{\mu} \in R^n} \frac{\mathbb{E} \|\hat{\vec{\mu}}_{c_1, c_2} - \vec{\mu}\|^2}{\inf_{1 \leq k \leq k_n} \mathbb{E} \|\hat{\vec{\mu}}_m - \vec{\mu}\|^2}$$

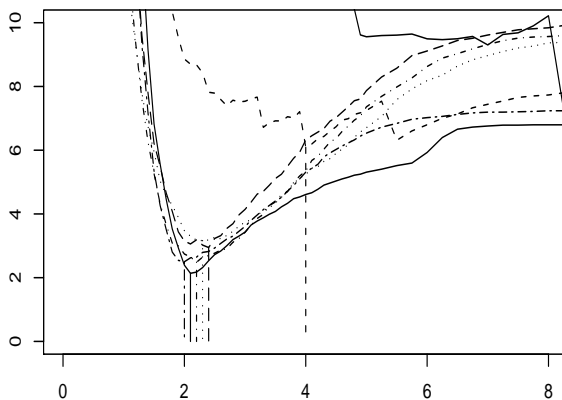
$c_2=0$



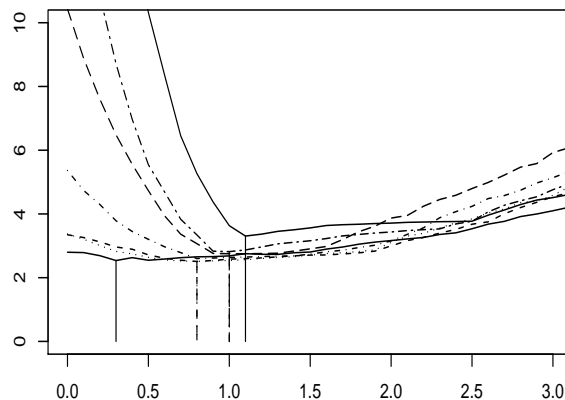
$c_2=2$



$c_2=4$



$c_2=8$



$$\text{pen}(k) = k\sigma^2(2\log(n/k) + 4)$$

