Apprentissage Statistique
M2 Probabilités et Statistiques, Université Paris-Sud

Cours 5 : Sélection d'estimateurs linéaires

SYLVAIN ARLOT ET FRANCIS BACH

Table des matières

1. Estimateurs linéaires en régression

Référence : Section 4.1 du deuxième cours de [1].

1.1. Cadre de la régression sur un plan d'expérience déterministe.

- Observation $Y = F + \varepsilon \in \mathbb{R}^n$, F déterministe, ε centré.
- Prédicteur $t \in \mathbb{R}^n$. Risque $n^{-1} \|t - F\|^2$. Risque empirique $n^{-1} \|t - Y\|^2$.

1.2. Estimateurs par projection.

- Modèle $S \subset \mathbb{R}^n$, lien avec le cas du design aléatoire.
- Minimiser le risque empirique revient à faire une projection orthogonale sur S.
- Exemple : fonctions constantes par morceaux.
- Exemple : sélection de variables.
  $X = (x_i^j)$ matrice $n \times p$. Prédicteur de la forme $\widehat{F} = Xw$.
  Estimateur par projection associé : $X(X^\top X)^{-1} X^\top Y$.

Date: 9 Mars 2015.

1.3. **Estimateurs linéaires.**
- Définition générale : $\widehat{F} = AY$, A déterministe.
- Exemple : estimateurs par projection.
- Exemple : estimateurs ridge à noyaux (rappel du cours 4)

$$\widehat{F} = K(K + n\lambda I)^{-1}Y \ .$$

  Pour la régression ridge, $K = XX^\top$. Pour la version à noyaux, $K = (k(x_i, x_j))_{1 \leqslant i,j \leqslant n}$.
- Exemple : estimateur des k-plus proches voisins (voir cours 1).
- Exemple : estimateur « par noyau » (Nadaraya-Watson ; voir cours 1).

Problème : choix de la matrice A ?

2. Sélection d'un estimateur linéaire

Référence : Section 4.2 du deuxième cours de [1].

2.1. **Problème.**
- Étant donnée une famille de matrices $(A_m)_{m \in \mathcal{M}}$, comment choisir parmi les estimateurs $\widehat{F}_m = A_m Y$? (Généralise le problème de choix de modèles)

2.2. **Calcul du risque.**

$$\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{2}{n}\left\langle A_m\varepsilon,\, (A_m - I)F\right\rangle + \frac{1}{n}\left\|A_m\varepsilon\right\|^2 \ ,$$

$$\text{d'où} \quad \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2 \operatorname{tr}(A_m^\top A_m)}{n}$$
$$= \text{Erreur d'approximation} + \text{Erreur d'estimation}$$

si les ε_i sont indépendants, centrés et de variance σ^2.

Exemples :
- projection sur S_m e.v. de dimension D_m : erreur d'estimation $\sigma^2 D_m/n$
- ridge : l'erreur d'estimation est une fonction décroissante de λ
- k-plus proches voisins : erreur d'estimation σ^2/k

2.3. **Pénalisation.**
- Calcul du risque empirique :

$$\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 = \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{2}{n}\left\langle (A_m - I)F,\, \varepsilon\right\rangle - \frac{2}{n}\left\langle A_m\varepsilon,\, \varepsilon\right\rangle + \frac{1}{n}\left\|\varepsilon\right\|^2 \ ,$$

$$\text{d'où} \quad \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] - \frac{2\sigma^2 \operatorname{tr}(A_m)}{n} + \sigma^2$$

  si les ε_i sont indépendants, centrés et de variance σ^2.
- Pénalité idéale et son espérance :

$$\operatorname{pen}_{\mathrm{id}}(m) = \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{1}{n}\left\|\varepsilon\right\|^2$$
$$= \frac{2}{n}\left\langle A_m\varepsilon,\, \varepsilon\right\rangle + \frac{2}{n}\left\langle (A_m - I)F,\, \varepsilon\right\rangle$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

est la pénalité C_L (Mallows [7]).

– Cas des estimateurs par projection : si A_m est une matrice de projection orthogonale sur un sev S_m de $\mathbb{R}^n$, alors,

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 \dim(S_m)}{n}$$

est la pénalité C_p de Mallows [7]. Par analogie, $\operatorname{tr}(A_m)$ est appelé nombre de degrés de liberté généralisé associé à A_m.

– Cas des k-plus proches voisins :

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(k)\right] = \frac{2\sigma^2}{k} \ .$$

– Inégalité-oracle si σ^2 est connu (au moins approximativement) [2].

Théorème 1. *On suppose :*
- $\forall m \in \mathcal{M}$, $|||A_m||| \leqslant 1$ *et* $\operatorname{tr}\left(A_m^\top A_m\right) \leqslant \operatorname{tr}(A_m)$
- *les* ε_i *sont i.i.d. Gaussiens centrés et de variance* $\sigma^2 > 0$,
- $C > 0$ *vérifie* $\left|C\sigma^{-2} - 1\right| \leqslant \kappa\sqrt{\ln(n)/n}$ *pour une constante* $\kappa > 0$.

Alors, des constantes $n_0, L(\gamma, \kappa) > 0$ *existent telles que pour tout* $\gamma \geqslant 1$, *avec probabilité au moins* $1 - 6\operatorname{Card}(\mathcal{M})n^{-\gamma}$, *si* $n \geqslant n_0$, *pour tout*

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{2C\operatorname{tr}(A_m)}{n} \right\} \ ,$$

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|_2^2 \leqslant \left(1 + \frac{1}{\ln(n)}\right) \inf_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - F\right\|_2^2 \right\} + \frac{L(\gamma,\kappa)\left(\ln(n)\right)^3 \sigma^2}{n} \ .$$

Éléments de preuve.

– calculs d'espérance ci-dessus,

– inégalités de concentration Gaussiennes : Propositions 2 et 4 dont les preuves (admises) sont données en Section 2.4 ci-dessous.

– début de preuve général d'une inégalité-oracle pour une procédure de pénalisation :

Lemme 1. *Si pour tout* $m \in \mathcal{M}$,

$$-B(m) \leqslant \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m) \leqslant A(m) \ ,$$

alors,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \mathrm{pen}(m) \right\}$$
$$-B(\widehat{m}) + \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leqslant \inf_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + A(m) \right\} \ .$$

– majoration des termes $A(m) = B(m)$ (issus naturellement des résultats de concentration ci-dessus) par une petite fraction du risque

□

2.4. **Résultats de concentration Gaussienne utilisés.**

Proposition 2. *Soit $n \geqslant 1$ un entier, X un vecteur gaussien standard dans $\mathbb{R}^n$ et $\alpha \in \mathbb{R}^n$. Alors, pour tout $x \geqslant 0$,*

$$\mathbb{P}\left(\langle X,\, \alpha\rangle > \|\alpha\|\, x\right) \leqslant \frac{1}{2}\exp\left(-\frac{x^2}{2}\right) \quad . \tag{1}$$

Preuve de la Proposition 2. Le résultat est évident si $\alpha = 0$. Sinon, puisque X est un vecteur Gaussien standard, la variable aléatoire réelle

$$Z := \frac{\langle X,\, \alpha\rangle}{\|\alpha\|}$$

suit une loi normale standard. On en déduit le résultat à l'aide du Lemme 3. □

Lemme 3. *Soit $Z \sim \mathcal{N}(0,1)$. Alors, pour tout $x \geqslant 0$,*

$$\mathbb{P}\left(Z \geqslant x\right) \leqslant \frac{1}{2}\exp\left(-\frac{x^2}{2}\right) \quad . \tag{2}$$

Preuve du Lemme 3. Comme Z admet pour densité $t \mapsto \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right)$ par rapport à la mesure de Lebesgue, on a

$$\begin{aligned}
\mathbb{P}\left(Z \geqslant x\right) &= \int_x^{+\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right) dt \\
&= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2+2tx+t^2}{2}\right) dt \\
&\leqslant \exp\left(-\frac{x^2}{2}\right)\int_0^{+\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right) dt \\
&= \exp\left(-\frac{x^2}{2}\right) \times \frac{1}{2} \quad .
\end{aligned}$$

□

Proposition 4. *Soit $n \geqslant 1$ un entier, X un vecteur gaussien standard dans $\mathbb{R}^n$ et $A \in \mathcal{M}_n(\mathbb{R})$. Alors, pour tout $x \geqslant 0$,*

$$\mathbb{P}\left(\langle X,\, AX\rangle - \operatorname{tr}(A) > 2\sqrt{x\operatorname{tr}(A^\top A)} + 2\,|\!|\!|A|\!|\!|\, x\right) \leqslant e^{-x} \tag{3}$$

$$\mathbb{P}\left(\langle X,\, AX\rangle - \operatorname{tr}(A) < -2\sqrt{x\operatorname{tr}(A^\top A)} - 2\,|\!|\!|A|\!|\!|\, x\right) \leqslant e^{-x} \tag{4}$$

où l'on a noté $|\!|\!|A|\!|\!| = \sup_{x\neq 0}\frac{\|Ax\|}{\|x\|}$ la norme d'opérateur de A.

Preuve de la Proposition 4. Posons

$$B = \frac{1}{2}\left(A^\top + A\right) \quad \text{et} \quad Z = \langle X,\, AX\rangle - \operatorname{tr}(A) = \langle X,\, BX\rangle - \operatorname{tr}(B) \quad .$$

La matrice B étant symétrique réelle, elle est diagonalisable en base orthonormée :

$$\exists P \in O(n) \text{ t.q. } B = P^\top D P \quad \text{avec} \quad D = \operatorname{diag}\left(d_1, \ldots, d_n\right) \quad .$$

Ainsi,

$$Z = X^\top BX - \mathrm{tr}(B) = (PX)^\top D(PX) - \mathrm{tr}(D) = \sum_{i=1}^n d_i \left(\xi_i^2 - 1\right)$$

où $\xi = PX$ est un vecteur gaussien standard de $\mathbb{R}^n$, car la loi normale standard est invariante par rotation. Par conséquent, à l'aide du Lemme 5, on obtient que pour tout $\lambda > 0$, tel que $\lambda d_i < 1/2$ pour $i = 1, \ldots, n$,

$$\begin{aligned}
\ln \mathbb{E}\left[\exp\left(\lambda Z\right)\right] = \sum_{i=1}^n \ln \mathbb{E}\left[e^{\lambda d_i(\xi_i^2 - 1)}\right] &\leqslant \sum_{i=1}^n \frac{d_i^2 \lambda^2}{1 - 2\lambda (d_i)_+} \\
&\leqslant \frac{\lambda^2}{1 - 2\lambda \max_i |d_i|} \sum_{i=1}^n d_i^2 \\
&= \frac{\lambda^2 \,\mathrm{tr}(B^\top B)}{1 - 2\lambda \left\|\!\left|B\right|\!\right\|} = \psi(2\lambda)
\end{aligned}$$

où l'on a posé

$$\forall u \in \left]0, 1/c\right[\quad , \qquad \psi(u) := \frac{u^2 v}{2(1 - uc)} \quad \text{avec} \quad v = \frac{\mathrm{tr}(B^\top B)}{2} \quad \text{et} \quad c = \left\|\!\left|B\right|\!\right\| \ .$$

Or, d'après le Lemme 7, on a l'inverse de la transformée de Fenchel-Legendre $\psi^\star$ de ψ :

$$\forall x > 0 \ , \qquad \psi^{\star\,-1}(x) = \sqrt{2vx} + cx \ .$$

Donc, pour tout $x > 0$, en posant $t = 2\psi^{\star\,-1}(x)$, pour tout $\lambda \in \left]0, 1/(2\left\|\!\left|B\right|\!\right\|)\right[$,

$$\mathbb{P}\left(Z \geqslant t\right) = \mathbb{P}\left(e^{\lambda Z} \geqslant e^{\lambda t}\right) \leqslant \exp\left(-\lambda t + \ln \mathbb{E}\left[e^{\lambda Z}\right]\right) \leqslant \exp\left(-\lambda t + \psi(2\lambda)\right) \ ,$$

soit

$$\begin{aligned}
\mathbb{P}\left(Z \geqslant t\right) &\leqslant \exp\left(- \sup_{u \in \left]0, 1/c\right[} \left\{ u\frac{t}{2} - \psi(u) \right\}\right) \\
&= \exp\left(-\psi^\star\left(\frac{t}{2}\right)\right) = e^{-x} \ .
\end{aligned}$$

On a donc prouvé que pour tout $x > 0$,

$$\mathbb{P}\left(\langle X,\, AX\rangle - \mathrm{tr}(A) \geqslant 2\sqrt{\mathrm{tr}(B^\top B)x} + 2\left\|\!\left|B\right|\!\right\| x\right) \leqslant e^{-x} \ .$$

Le résultat (3) s'en déduit car

$$\begin{aligned}
\left\|\!\left|B\right|\!\right\| &= \left\|\!\left|\frac{1}{2}\left(A^\top + A\right)\right|\!\right\| \leqslant \frac{1}{2}\left(\left\|\!\left|A^\top\right|\!\right\| + \left\|\!\left|A\right|\!\right\|\right) = \left\|\!\left|A\right|\!\right\| \\
\mathrm{tr}(B^\top B) &= \frac{1}{4}\,\mathrm{tr}\left[\left(A^\top + A\right)\left(A^\top + A\right)\right] \\
&= \frac{1}{4}\left[\mathrm{tr}\left(\left(A^\top\right)^2\right) + \mathrm{tr}\left(A^2\right) + 2\,\mathrm{tr}\left(A^\top A\right)\right] \\
&= \frac{1}{2}\left[\mathrm{tr}\left(A^2\right) + \mathrm{tr}\left(A^\top A\right)\right] \\
&\leqslant \mathrm{tr}\left(A^\top A\right) \ ,
\end{aligned}$$

où le dernier point provient du calcul suivant :

$$\mathrm{tr}(A^2) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}A_{j,i} \leqslant \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{A_{i,j}^2 + A_{j,i}^2}{2} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{i,j}^2 = \mathrm{tr}\left(A^\top A\right) \ .$$

On obtient (4) en remplaçant A par $-A$ dans (3). □

Lemme 5. *Soit $Z \sim \mathcal{N}(0,1)$. Alors, pour tout $\lambda < 1/2$,*

$$\mathbb{E}\left[\exp\left(\lambda Z^2\right)\right] = \frac{1}{\sqrt{1-2\lambda}} \tag{5}$$

si bien que

$$\ln \mathbb{E}\left[\exp\left(\lambda\left(Z^2-1\right)\right)\right] = \frac{-1}{2}\left[\ln\left(1-2\lambda\right) + 2\lambda\right] \leqslant \frac{\lambda^2}{1-(2\lambda)_+} \ . \tag{6}$$

Preuve du Lemme 5. Le membre de gauche de (5) existe (mais pourrait être égal à $+\infty$) car $\exp(\lambda Z^2) \geqslant 0$, et l'on a, pour tout $\lambda < 1/2$,

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda Z^2\right)\right] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(-t^2\left(\frac{1}{2}-\lambda\right)\right) dt \\ &= \sigma_\lambda \frac{1}{\sigma_\lambda\sqrt{2\pi}}\int_{\mathbb{R}} \exp\left(-\frac{t^2}{2\sigma_\lambda^2}\right) dt = \sigma_\lambda \ ,\end{aligned}$$

en posant $\sigma_\lambda = (1-2\lambda)^{-1/2}$, ce qui prouve (5). On en déduit (6) grâce au Lemme 6 (avec $\lambda = 2x$). □

Lemme 6. *Pour tout $x < 1$,*

$$\ln(1-x) + x \geqslant \frac{-x^2}{2(1-(x)_+)} \ .$$

Démonstration. Pour tout $x \in \ ]0,1[$, on pose

$$f(x) = \ln(1-x) + x + \frac{x^2}{2(1-x)} \ .$$

On a $f(0) = 0$ et pour tout $x \in [0,1[$,

$$f'(x) = \frac{-1}{1-x} + 1 + \frac{2x(1-x)+x^2}{2(1-x)^2} = \frac{x^2}{2(1-x)^2} \geqslant 0 \ ,$$

et donc $f(x) \geqslant 0$ pour $x \in [0,1[$.

Pour tout $x \leqslant 0$, on pose

$$g(x) = \ln(1-x) + x + \frac{x^2}{2} \ .$$

Alors, $g(0) = 0$ et pour tout $x < 0$,

$$g'(x) = \frac{-1}{1-x} + 1 + x = \frac{-x^2}{1-x} \leqslant 0$$

si bien que $g(x) \geqslant 0$ pour tout $x \leqslant 0$. □

Lemme 7 (voir la Section 2.4 de [5]). *Soit $v, c > 0$. Pour tout $\lambda \in]0, 1/c[$, on pose*

$$\psi(\lambda) := \frac{v\lambda^2}{2(1-c\lambda)} \ .$$

Alors, la transformée de Fenchel-Legendre de ψ s'écrit

$$\forall t > 0 \ , \qquad \psi^\star(t) := \sup_{\lambda \in]0,1/c[} \{t\lambda - \psi(\lambda)\} = \frac{v}{c^2} h_1\left(\frac{ct}{v}\right) \tag{7}$$

avec

$$\forall u \geqslant 0 \ , \qquad h_1(u) := 1 + u - \sqrt{1+2u} \ .$$

De plus, comme h_1 est une bijection strictement croissante de $[0, +\infty[$ dans lui-même et

$$\forall u \geqslant 0 \ , \qquad h_1^{-1}(u) = u + \sqrt{2u} \ ,$$

on a

$$\forall u > 0 \ , \qquad \psi^{\star\,-1}(u) = \sqrt{2vu} + cu \ . \tag{8}$$

3. Calibration de pénalités

Référence : deuxième cours de [1].

- Problème : σ^2 dans la pénalité doit être estimé à l'aide des données.
- Point-clé : l'espérance du risque empirique vaut

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \sigma^2\left(\frac{\operatorname{tr}(A_m^\top A_m) - 2\operatorname{tr}(A_m)}{n} + 1\right) \ .$$

3.1. Approches classiques.

- Estimation de σ^2 dans un « gros modèle ». Dans le cas des estimateurs par projection, on choisit m_0 tel que $\dim(S_{m_0})$ est grand, et on estime σ^2 par

$$\widehat{\sigma^2}(m_0) = \frac{\left\|\widehat{F}_{m_0} - Y\right\|^2}{n - \dim(S_{m_0})}$$

  en espérant que le biais $\frac{1}{n}\left\|(A_{m_0} - I)F\right\|^2$ est négligeable devant les autres termes.
- Validation croisée généralisée (GCV, [6]) : choisir m en minimisant

$$\left(\frac{\left\|\widehat{F}_m - Y\right\|}{1 - n^{-1}\operatorname{tr}(A_m)}\right)^2 \ .$$

  Dans le cas des estimateurs par projection, cela revient (à une approximation près) à estimer σ^2 dans chaque modèle, i.e., à utiliser la pénalité

$$\operatorname{pen}(m) = \frac{2\widehat{\sigma^2}(m)\dim(S_m)}{n} \ .$$

3.2. Heuristique de pente (cas des estimateurs par projection).

Références : [4, 3] ou Sections 1–3 du deuxième cours de [1].

– Décroissance linéaire asymptotique du risque empirique :

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 - \frac{\sigma^2}{n}\dim(S_m) + \sigma^2$$

décroît linéairement avec $\dim(S_m)$, dès lors que le biais $\frac{1}{n}\left\|(A_m - I)F\right\|^2$ ne varie plus.

– Notion de pénalité minimale. Saut de dimension associé : si

$$\widehat{m}(C) \in \arg\min_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\,\mathrm{pen}_{\min}(m)\right\} \ ,$$

alors $\dim(S_{\widehat{m}(C)})$ est très grande pour $C < 1$ (sur-apprentissage), et « raisonnablement grande » pour $C > 1$. Pour les estimateurs par projection,

$$\mathrm{pen}_{\min}(m) = \frac{\sigma^2}{n}\dim(S_m) \quad \text{d'où} \quad \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = 2\,\mathrm{pen}_{\min}(m) \ .$$

– Algorithme de calibration correspondant (voir [3] à propos d'un algorithme de calcul efficace de la trajectoire complète $(\widehat{m}(C))_{C>0}$) :

Entrée : $(A_m)_{m\in\mathcal{M}}$ famille finie de matrices de projections orthogonales, $Y \in \mathbb{R}^n$

– $\forall C > 0$, calculer

$$\widehat{m}(C) \in \operatorname*{argmin}_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{D_m}{n}\right\} \ .$$

– Trouver $\widehat{C}$ autour duquel $C \mapsto D_{\widehat{m}(C)}$ « saute ».

Sortie : $\widehat{m} = \widehat{m}(2\widehat{C})$.

Exercice 1. Montrer que $(\widehat{m}(C))_{C>0}$ est constante par morceaux et saute au plus $\mathrm{Card}(\mathcal{M}_n)$ fois.

– Théorème : existence d'une pénalité minimale et du saut de dimension.

Théorème 2. *On suppose :*

– *$\forall m \in \mathcal{M}$, A_m est une matrice de projection orthogonale sur un espace vectoriel de dimension D_m,*

– *les ε_i sont i.i.d. Gaussiens centrés et de variance $\sigma^2 > 0$,*

– *pour tout $C > 0$,*

$$\widehat{m}(C) \in \arg\min_{m\in\mathcal{M}}\left\{\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{CD_m}{n}\right\} \ ,$$

– *il existe $m_1, m_2 \in \mathcal{M}$ tels que*

$$A_{m_1} = I_n \quad D_{m_2} \leqslant \sqrt{n} \quad \textit{et} \quad \frac{1}{n}\left\|(A_{m_2} - I)F\right\|^2 \leqslant \frac{\sigma^2\sqrt{\ln(n)}}{\sqrt{n}} \ .$$

Alors, des constantes $n_1, L_2 > 0$ existent telles que si $n \geqslant n_1$, pour tout $\gamma \geqslant 1$, avec probabilité au moins $1 - 3\,\mathrm{Card}(\mathcal{M})n^{-\gamma}$,

$$\forall 0 \leqslant C < \left(1 - L_2\gamma\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2 \ , \qquad D_{\widehat{m}(C)} \geqslant \frac{9n}{10}$$

$$\text{et} \quad \forall C > \left(1 + L_2\gamma\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2 \ , \qquad D_{\widehat{m}(C)} \leqslant \frac{n}{10} \ .$$

Éléments de preuve. (Voir [1] pour une preuve complète.)

3.3. **Heuristique de pente (cas des estimateurs linéaires).** Référence : Sections 4–5 du deuxième cours de [1].

– Forme de la pénalité minimale :

$$\mathrm{pen}_{\min}(m) = \frac{\sigma^2(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))}{n}$$

n'est plus proportionnelle à $\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$.

– Modification correspondante de l'algorithme de calibration de pénalités :

Entrée : $(A_m)_{m\in\mathcal{M}}$ famille finie de matrices, $Y \in \mathbb{R}^n$
– $\forall C > 0$, calculer

$$\widehat{m}_0(C) \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + C\frac{2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)}{n} \right\} \ .$$

– Trouver $\widehat{C}$ autour duquel $C \mapsto \mathrm{tr}(A_{\widehat{m}_0(C)})$ « saute »
Sortie : $\widehat{m} \in \mathrm{argmin}_{m\in\mathcal{M}}\{n^{-1}\|\widehat{F}_m - Y\|^2 + 2\widehat{C}n^{-1}\,\mathrm{tr}(A_m)\}$.

– Résultat théorique :

Théorème 3. *On suppose :*
– *$\forall m \in \mathcal{M}$, $|||A_m||| \leqslant 1$ et $\mathrm{tr}\left(A_m^\top A_m\right) \leqslant \mathrm{tr}(A_m)$*
– *les ε_i sont i.i.d. Gaussiens centrés et de variance $\sigma^2 > 0$,*
– *pour tout $C > 0$,*

$$\widehat{m}(C) \in \arg\min_{m\in\mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{C(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m))}{n} \right\} \ ,$$

– *il existe $m_1, m_2 \in \mathcal{M}$ tels que*

$$A_{m_1} = I_n \quad \mathrm{tr}(A_{m_2}) \leqslant \sqrt{n} \quad \text{et} \quad \frac{1}{n}\|(A_{m_2} - I)F\|^2 \leqslant \frac{\sigma^2\sqrt{\ln(n)}}{\sqrt{n}} \ .$$

Alors, des constantes $n_1, L_2 > 0$ existent telles que si $n \geqslant n_1$, pour tout $\gamma \geqslant 1$, avec probabilité au moins $1 - 6\,\mathrm{Card}(\mathcal{M})n^{-\gamma}$,

$$\forall 0 \leqslant C < \left(1 - L_2\gamma\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2, \quad \mathrm{tr}(A_{\widehat{m}(C)}) \geqslant \frac{n}{3}$$

$$\text{et} \quad \forall C > \left(1 + L_2\gamma\sqrt{\frac{\ln(n)}{n}}\right)\sigma^2, \quad \mathrm{tr}(A_{\widehat{m}(C)}) \leqslant \frac{n}{10} \ .$$

Références

[1] Sylvain Arlot. Sélection de modèles et sélection d'estimateurs pour l'apprentissage statistique, January 2011. Cours Peccot. Collège de France. `http://www.di.ens.fr/~arlot/peccot.htm`.

[2] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[3] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10 :245–279 (electronic), 2009.

[4] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[5] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities : A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[6] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4) :377–403, 1978/79.

[7] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.