Devoir maison II

Exercice 1.

On suppose $y = f + \xi$ avec ξ vecteur σ-SG . Soient $\Theta \subseteq \mathbf{B}_1(L) := \{\theta : |\theta|_1 \leq L\}$, $\tilde{f} := X\hat{\theta}^{\mathrm{MC}}(\Theta)$ et $\delta \in \,]0\,;1[$ fixé.

Montrer qu'avec une probabilité d'au moins $1-\delta$, on a l'inégalité suivante:

$$\forall f \in \mathbb{R}^n : \|\tilde{f} - f\|^2 \leq \inf_{\theta\in\Theta} \|X\theta - f\|^2 + U_\delta$$

où U_δ désigne un terme indépendant de f et que l'on précisera.

Solution 1.

Par le lemme 1.1, on a

$$\forall f \in \mathbb{R}^p, \forall \theta \in \Theta \quad \|\tilde{f} - f\|^2 \leq \|f - X\theta\|^2 + 2\left(\xi \,|\, \tilde{f} - X\theta\right)$$

Concentrons nous sur le terme de gauche. En utilisant la relation $\tilde{f} := X\hat{\theta}^{\mathrm{MC}}$ et en notant $\hat{\theta} := \hat{\theta}^{\mathrm{MC}}(\Theta)$, on a successivement

$$\left(\xi \,|\, \tilde{f} - X\theta\right) = \left(X^T\xi \,|\, \hat{\theta} - \theta\right) \leq \frac{1}{n}\sum_{j=1}^{p} [X^T\xi]_j(\hat{\theta}_i - \theta_i) \leq \frac{1}{n}\max_{1\leq j\leq p}\left|x_j^T\xi\right| \cdot \left|\hat{\theta} - \theta\right|_1$$

avec x_j désignant la $j$ème colonne de la matrice X. Comme θ et $\hat{\theta}$ appartiennent à $\Theta \subseteq \mathbf{B}_1(L)$, on a par inégalité triangulaire que $\left|\hat{\theta} - \theta\right|_1 \leq 2L$.

En introduisant $v_j := \frac{x_j}{|x_j|_2}$ et $x^\star := \max_{1\leq j\leq p}\left\|x_j\right\| = \frac{1}{\sqrt{n}}\max_{1\leq j\leq p}\left|x_j\right|_2$, on a par ailleurs

$$\max_{1\leq j\leq p}\left|x_j^T\xi\right| \leq \sqrt{n}\,x^\star \cdot \max_{1\leq j\leq p}\left|v_j^T\xi\right|$$

Par définition de ξ vecteur σ-SG , on a $v_j^T\xi$ variable σ-SG . On en déduit grâce au lemme 1.4 l'inégalité

$$\max_{1\leq j\leq p}\left|v_j^T\xi\right| \leq \sigma\sqrt{2\log\left(\frac{2p}{\delta}\right)}$$

vérifiée avec une probabilité d'au moins $1-\delta$.

En posant $U_\delta = 4\sigma L x^\star\sqrt{\dfrac{2\log\left(\frac{2p}{\delta}\right)}{n}}$ on a donc avec une probabilité d'au moins $1-\delta$ l'inégalité suivante:

$$\forall f \in \mathbb{R}^p, \forall \theta \in \Theta \quad \|\tilde{f} - f\|^2 \leq \|f - X\theta\|^2 + U_\delta$$

Il suffit alors de prendre l'infimum sur θ dans Θ pour conclure.

Exercice 2.

Montrer que

$$\hat{\theta}^H = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}}\left(\sum_{j=1}^{p}\left(Y_j-\theta_j\right)^2+\tau^2|\theta|_0\right) \tag{1}$$

$$\hat{\theta}^S = \underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}}\left(\sum_{j=1}^{p}\left(Y_j-\theta_j\right)^2+2\tau|\theta|_1\right) \tag{2}$$

Solution 2.

(1) Notons $\hat{\theta}$ un argmin de la fontion à minimiser. En réécrivant cette dernière, on a successivement:

$$\sum_{j=1}^{p}\left(Y_j-\theta_j\right)^2+\tau^2|\theta|_0=\sum_{j=1}^{p}\left\{\left(Y_j-\theta_j\right)^2+\tau^2\mathbb{1}_{\theta_j\neq 0}\right\}=\sum_{j=1}^{p}\psi\left(\theta_j,Y_j\right)$$

où $\psi:\begin{cases}\mathbb{R}^2 & \mapsto & \mathbb{R}\\ (\theta,Y) & \mapsto & (Y-\theta)^2+\tau^2\mathbb{1}_{\theta\neq 0}\end{cases}.$

Minimiser cette somme sur θ à Y donné reviens simplement à minimiser la fonction ψ à Y fixé et on aura $\hat{\theta}_j=\underset{\theta\in\mathbb{R}}{\operatorname{argmin}}\psi\left(\theta,Y_j\right)$.

On cherche à connaître la position relative de $\psi(\theta,Y)$ par rapport à $\psi(0,Y)$:

$$\psi(\theta,Y)-\psi(0,Y)=(Y-\theta)^2+\tau^2-Y^2=\theta^2-2Y\theta+\tau^2.$$

En posant $\Delta'=Y^2-\tau^2$, les trois configurations suivantes sont alors possibles:

- $\Delta<0$, *i.e.* $Y^2<\tau^2$. On a alors

$$\forall\theta\in\mathbb{R}-\{0\},\quad \psi(\theta,Y)>\psi(0,Y).$$

  L'argmin est donc unique est vaut 0.

- $\Delta=0$, *i.e.* $Y^2:\tau^2$. On a alors

$$\forall\theta\in\mathbb{R}-\{0\},\quad \psi(\theta,Y)-\psi(0,Y)=(\theta-Y)^2.$$

  On a donc $\theta=0$ et $\theta=Y$ qui minimisent la fonction objectif.

- $\Delta>0$, *i.e.* $Y^2>\tau^2$. On a alors

$$\forall\theta\in\left]Y-\sqrt{\Delta'};Y+\sqrt{\Delta'}\right[,\quad \psi(\theta,Y)<\psi(0,Y).$$

  Seul $\theta=Y$ minimise $(\theta,Y)=(Y-\theta)^2+\tau^2$ sur cet intervalle.

  Finalement les coordonnées de $\hat{\theta}_j^H:=Y_jI\left(\left|Y_j\right|>\tau\right)$ vérifient bien les conditions obtenues ci-dessus [1].

(2) Posons $G(\theta)=\|Y-\theta\|^2+2\tau|\theta|_1$. Par somme, on a bien G fonction convexe et l'équivalence suivante est vérifiée:

$$\hat{\theta}\in\underset{\theta\in\mathbb{R}^p}{\operatorname{argmin}}G(\theta)\Longleftrightarrow 0\in\partial G\left(\hat{\theta}\right). \tag{3}$$

Calculons la sous-différentielle de G:

$$\partial G(\theta)=2(\theta-Y)+2\tau\cdot\partial|\theta|_1$$

1. On aurait aussi pu choisir $\hat{\theta}_j^H:=Y_jI\left(\left|Y_j\right|\geq\tau\right)$ puisque l'argmin contient aussi Y lorsque $Y^2=\tau^2$.

avec

$$\forall j \in \{1, \dots p\} \quad (\partial |\theta|_1)_j = \begin{cases} \left\{\text{signe}\left(\theta_j\right)\right\}, & \text{si } \theta_j \neq 0 \\ [-1\,;1], & \text{sinon} \end{cases}.$$

En utilisant 3, on en déduit donc:

$$\forall j \in \{1, \dots p\} \qquad \begin{cases} 0 = 2\left(\hat{\theta}_j - Y_j\right) + 2\tau \,\text{signe}\left(\hat{\theta}_j\right), & \text{si } \theta_j \neq 0 \\ 0 = 2\left(\hat{\theta}_j - Y_j\right) + 2\tau\alpha_j, & \text{sinon} \end{cases}.$$

avec $\left|\alpha_j\right| \leq 1$ ce qui est équivalent à

$$\hat{\theta}_j = \begin{cases} Y_j - \tau, & \text{si } Y_j > \tau \\ 0, & \text{si } \left|Y_j\right| \leq \tau \\ Y_j + \tau, & \text{sinon} \end{cases}$$

pour tout $j \in \{1, \dots p\}$. On retrouve bien l'expression attendue de l'estimateur de *soft-thresholding* $\hat{\theta}^S_j := \left(1 - \frac{\tau}{|Y_j|}\right)_+ Y_j$.

Exercice 3.
Montrer que sous l'hypothèse ORT $\left(\dfrac{1}{n} X^T \cdot X = I_p\right)$, on a les égalités suivantes:

$$\hat{\theta}^D = \hat{\theta}^L = \hat{\theta}^S.$$

Solution 3. Rappelons la définition de chacun d'entre eux:

$$\hat{\theta}^S = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left(\sum_{j=1}^{p} \left(Y_j - \theta_j\right)^2 + 2\tau|\theta|_1 \right) \tag{4}$$

$$\hat{\theta}^L = \underset{\theta \in \mathbb{R}^p}{\text{argmin}} \left(\left\| y - X\theta \right\| + 2\tau|\theta|_1 \right) \tag{5}$$

$$\hat{\theta}^D = \underset{\theta \in \mathbb{R}^p \text{ tq } \left|\frac{1}{n} X^T (y - X\theta)\right|_\infty \leq \tau}{\text{argmin}} |\theta|_1 \tag{6}$$

- Montrons que $\hat{\theta}^L = \hat{\theta}^S$. En utilisant les relations $\frac{1}{n} X^T \cdot X = I_p$ et $Y = \frac{1}{n} X^T y$, on a successivement:

$$\begin{aligned} \left\| y - X\theta \right\|^2 &= \frac{1}{n} \left| y - X\theta \right|_2^2 \\ &= \frac{1}{n} \left(y^T y - y^T X\theta - (X\theta)^T y + (X\theta)^T X\theta \right) \\ &= \frac{1}{n} y^T y - Y^T\theta - \theta^T Y + \theta^T\theta \\ &= |Y - \theta|_2^2 + C \end{aligned}$$

  où $C = \frac{1}{n} y^T y - Y^T Y$ indépendant de θ variable d'optimisation.
  Les deux fonctions objectifs définissant les estimateurs $\hat{\theta}^L$ et $\hat{\theta}^S$ ne différant que d'une constante, on a bien égalité de ces derniers [2].
- Montrons que $\hat{\theta}^D = \hat{\theta}^S$. Commençons par réécrire la contrainte de Dantzig en terme de Y. Les conditions d'orthonormalité sur X et la relation $Y = \frac{1}{n} X^T y$ nous permettent d'écrire:

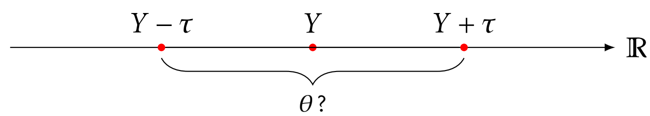
$$\frac{1}{n} X^T (y - X\theta) = Y - \theta.$$

2. Les deux fonctions objectifs étant strictement convexes, on a bien unicité de ce dernier, son existence étant garanti par la coercivité des fonctions objectifs.

On doit donc montrer que

$$\hat{\theta}^S = \underset{\theta \in \mathbb{R}^p \text{ tq } |Y-\theta|_\infty \leq \tau}{\operatorname{argmin}} |\theta|_1 .$$

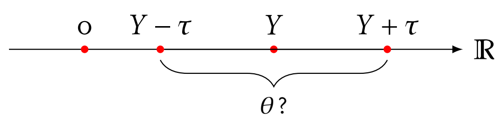
Notons $\hat{\theta}$ l'armin de la fonction ci-dessus. Minimiser $|\theta|_1 = \sum_{j=1}^{p} |\theta_i|$ étant donné les contraintes indépendantes sur les θ_i reviens à résoudre le problème de minimisation ci-dessous pour $p = 1$, *i.e.*du type et illustrer figure 3: $\underset{\theta \in \mathbb{R} \text{ tq } |Y-\theta| \leq \tau}{\operatorname{argmin}} |\theta|$.



On cherche à minimiser la valeur absolue de θ sous la contrainte de ne pas s'éloigner de Y de plus d'un τ.

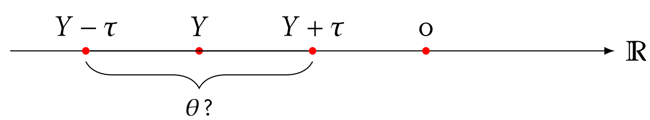
Trois configurations sont alors possibles:

- $Y - \tau \geq 0$:



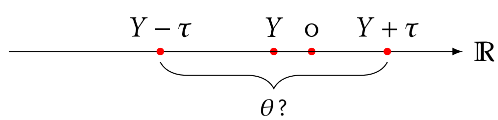
  Dans ce cas $\hat{\theta} = Y - \tau$ minimise bien la valeur absolue.

- $Y + \tau \leq 0$:



  Dans ce cas $\hat{\theta} = Y + \tau$ minimise bien la fonction objectif.

- $Y - \tau \geq 0$:



  Dans ce cas $\hat{\theta} = 0$ est bien la valeur optimale.

Finalement, on a bien $\hat{\theta}_j^D = \hat{\theta}_j^S$ pour tout j dans $\{1, \dots, p\}$ ce qui conclut la preuve.

Exercice 4.

Soit $y = X\theta^\star + \xi$ avec $\theta^\star$ sparse. On suppose que les hypothèses d'orthogonalité de X $\left(\frac{1}{n}X^T \cdot X = I_p\right)$ et de gaussianité du bruit : $\xi \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$. On peut donc écrire: $Y_j = \theta_i^\star + \epsilon + \eta_j$ avec $\epsilon = \frac{\sigma}{\sqrt{n}}$ et $\eta_j \sim \mathcal{N}(0,1)$.

Soit $\tau = A\sigma\sqrt{\frac{2\log\left(\frac{p}{\delta}\right)}{n}}$ pour $\delta \in]0;1[$ et $A \geq 1$. Montrer que l'on a avec une probabilité supérieure à $1-\delta$ l'**inégalité d'oracle en probabilité** suivante :

$$\left|\hat{\theta}^S - \theta^\star\right|_2^2 \leq C\frac{\sigma^2}{n}\left|\theta^\star\right|_0 \log\left(\frac{p}{\delta}\right).$$

Solution 4.

Comme pour l'estimateur de *hard-thresholding*, considérons l'évènement suivant :

$$\mathcal{A} := \left\{\max\left|Y_j - \theta_j^\star\right| \leq r \mid \forall j = 1\ldots p\right\}.$$

où $r := \sigma\sqrt{\frac{2\log\left(\frac{p}{\delta}\right)}{n}} = \epsilon\sqrt{2\log\left(\frac{p}{\delta}\right)}$ [3]. Procédons par étape:

- Montrons que $\mathbb{P}\left(\mathcal{A}^C\right) \leq \delta$. En utlisant la relation $Y_j = \theta_j^\star + \epsilon\eta_j$, on a successivement:

$$\begin{aligned}\mathbb{P}\left(\mathcal{A}^C\right) &= \mathbb{P}\left(\max\left|Y_j - \theta_j^\star\right| > r \mid \forall j = 1\ldots p\right)\\ &= \mathbb{P}\left(\max\left|Y_j - \theta_j^\star\right| > \epsilon\sqrt{2\log\left(\frac{p}{\delta}\right)}\right)\\ &= \mathbb{P}\left(\max\left|\eta_j\right| > \sqrt{2\log\left(\frac{p}{\delta}\right)}\right) \leq \delta\end{aligned}$$

  en vertu du lemme 2.3 a) [4].
- Soient $\hat{J} = \left\{j \text{ tq } \hat{\theta}_j^S \neq 0\right\} = \left\{j \text{ tq } \left|Y_j\right| > \tau\right\}$ [5]. Montrons que $\hat{J} \subseteq J^\star$ sur $\mathcal{A}$.
  En effet, pour $j \in \hat{J}$, on a par inégalité triangulaire:

$$\left|\theta_j^\star\right| \geq \left|Y_j\right| - \left|\theta_j^\star - Y_j\right| = \left|Y_j\right| - \left|\epsilon\eta_j\right| \geq \left|Y_j\right| - r > \tau - Ar = 0.$$

- Reste maintenant à majorer $\left|\hat{\theta}^S - \theta^\star\right|_2$. On a:

$$\left|\hat{\theta}^S - \theta^\star\right|_2^2 = \sum_{j\in J^\star}\left|\hat{\theta}_j^S - \theta_j^\star\right|^2 + \sum_{j\notin J^\star}\left|\hat{\theta}_j^S - \theta_j^\star\right|^2.$$

  ☞ $\sum_{j\notin J^\star}\left|\hat{\theta}_j^S - \theta_j^\star\right|^2$. Comme $\hat{J} \subseteq J^\star$, on a $\left(J^\star\right)^C \subseteq \left(\hat{J}\right)^C$. D'où la nullité de cette somme.

  ☞ $\sum_{j\in J^\star}\left|\hat{\theta}_j^S - \theta_j^\star\right|^2$. Faisons une disjonction de cas:

  - Si $\left|Y_j\right| > \tau$, alors $\hat{\theta}_j^S = Y_j \pm \tau = \theta_j^\star + \epsilon\eta_j \pm \tau$. On a donc la majoration:

$$\left|\hat{\theta}_j^S - \theta_j^\star\right|^2 \leq \left(\epsilon\left|\eta_j\right| + \tau\right)^2 \leq 2\left(\epsilon^2\eta_j^2 + \tau^2\right).$$

3. On a alors la relation $\tau = Ar$.
4. Ceci est inchangé par rapport à la preuve avec l'estimateur de *hard-thresholding* puisqu'aucun estimateur n'intervient dans la définition de $\mathcal{A}$.
5. On rappelle que $\hat{\theta}_j^S = \left(1 - \frac{\tau}{|Y_j|}\right)_+ Y_j$

- Si $\left|Y_j\right| \le \tau$. Dans ce cas $\hat{\theta}_j^S = 0$ et on a:

$$\left|\hat{\theta}_j^S - \theta_j^\star\right|^2 \;=\; \left|\theta_j^\star\right|^2 \;\le\; \left(\left|Y_j\right| + \epsilon\left|\eta_j\right|\right)^2 \;\le\; \left(\tau + \epsilon\left|\eta_j\right|\right)^2 \;\le\; 2\left(\epsilon^2\eta_j^2 + \tau^2\right).$$

Finalement, en utilisant $\left|Y_j - \theta_j^\star\right| = \left|\epsilon\eta_j\right| \le r$ sur $\mathcal{A}$ et la relation $\tau = Ar$:

$$\begin{aligned}
\left|\hat{\theta}^S - \theta^\star\right|_2^2 &\le 2\left|\theta^\star\right|_0\left(\epsilon^2\eta_j^2 + \tau^2\right)\\
&\le 2\left|\theta^\star\right|_0(r^2 + \tau^2)\\
&\le 2\left|\theta^\star\right|_0(1 + A^2)\,r^2\\
&\le C\frac{\sigma^2}{n}\left|\theta^\star\right|_0\log\left(\frac{p}{\delta}\right)
\end{aligned}$$

où $C = 4(1 + A^2)$ est une constante absolue, d'où le résultat.