Statistique en Grande Dimension et Apprentissage - TD 3

Ce TD a pour but d'illustrer le chapitre 3 principalement sur la régression linéaire pénalisée.

Exercice 1 (Quelques propriétés standards sur le modèle linéaire classique)**.** On considère le modèle $\mathbf{Y} = \mathbf{X}\theta^\star + \varepsilon$ avec $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X}$ matrice $n \times p$, $\theta \in \mathbb{R}^p$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$. et $F(\theta) = \frac{1}{2n}\|\mathbf{Y} - \mathbf{X}\theta\|^2$. On suppose que $X^T X$ est inversible.

1. Retrouvez l'expression de $\hat{\theta} = \text{Argmin}_\theta F(\theta)$.
2. Montrez que $\hat{\theta} = \theta^\star + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon$. En déduire la loi de $\hat{\theta} - \theta^\star$.
3. Quel résultat retrouve-t-on lorsque $\mathbf{X} = (1, \ldots, 1)^T$?
4. On considère le modèle $Y = \theta_0 + \sum_{i=1}^p \theta_i X_i + \varepsilon$. A l'aide du théorème de Cochran, montrez que
$$\frac{SCR}{\sigma^2} \sim \chi^2(n-p-1)$$
Sous l'hypothèse $(H_0) := \theta_1 = \ldots = \theta_p = 0$, montrez que
$$\frac{SCM}{\sigma^2} \sim \chi^2(p).$$
En déduire que dans ce cas,
$$\frac{SCT}{\sigma^2} \sim \chi^2(n-1).$$

Exercice 2 (soft-shrinkage)**.** Le but de cet exercice est de comprendre l'expression de $\hat{\theta}^{LASSO}$ à partir de l'étude du cas unidimensionnel. On pose pour tout $\theta \in \mathbb{R}$,

$$\phi(\theta) = \frac{1}{2N}\sum_{i=1}^n (y_i - z_i\theta)^2 + \lambda|\theta|.$$

où y_i, z_i sont des réels et $\lambda > 0$. On fait également l'hypothèse que $\frac{1}{N}\sum_{i=1}^n z_i^2 = 1$.

1. Calculez la dérivée de $\theta \mapsto \phi(\theta)$ (sur $\mathbb{R}^*$).
2. Montrez que ϕ est strictement convexe (Indication : une fonction de classe $\mathcal{C}^2$ sur $\mathbb{R}^*$, continue sur $\mathbb{R}$ de dérivée seconde strictement positive sur $\mathbb{R}^*$ est strictement convexe).
3. Montrer sans calcul que ϕ a au moins un minimum. Pourquoi ce minimum est-il unique (illustrer par un dessin) ? On note $\hat{\theta}$ ce minimum dans la suite.
4. Etablir le tableau de variations de $\theta \mapsto \phi(\theta)$ lorsque $\frac{\langle \mathbf{z}, \mathbf{y} \rangle}{N} > \lambda$. En déduire $\hat{\theta}$ dans ce cas.

5. Montrez que
$$\hat{\theta} = \begin{cases} \frac{\langle \mathbf{z},\mathbf{y}\rangle}{N} - \lambda & \text{si } \frac{\langle \mathbf{z},\mathbf{y}\rangle}{N} > \lambda \\ 0 & \text{si } \frac{|\langle \mathbf{z},\mathbf{y}\rangle|}{N} < \lambda, \\ \frac{\langle \mathbf{z},\mathbf{y}\rangle}{N} + \lambda & \text{si } \frac{\langle \mathbf{z},\mathbf{y}\rangle}{N} < -\lambda. \end{cases}$$
6. En déduire que
$$\hat{\theta} = \mathcal{S}_\lambda\left(\frac{\langle \mathbf{z},\mathbf{y}\rangle}{N}\right)$$
où
$$\mathcal{S}_\lambda(x) = \mathrm{sgn}(x)(|x| - \lambda)_+.$$
$\mathcal{S}_\lambda$ est appelé l'opérateur de seuillage doux (soft-thesholding operator).

Exercice 3 (Coordinate Descent)**.** Pour rappel, la descente coordonnées par coordonnées est un algorithme de recherche de $\hat{\theta}_\lambda$ basé sur le principe de minimiser successivement la fonction L de l'exercice précédent sur une seule des variables. Plus précisément, considérons la fonction L définie par :
$$L(\theta) = \frac{1}{2N}\|\mathbf{y} - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1.$$
Remarque : Selon les exercices, le coefficient devant $\|\mathbf{y}-\mathbf{X}\theta\|_2^2$ peut changer $(1, 1/(2N), 1/N)$. Notez que quitte à changer la valeur de λ, les problèmes de minimisation restent équivalents.

1. On note $\theta^{-j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p)$. Pour un tel vecteur, on note $L_j^{\theta^{-j}}$ l'application de $\mathbb{R}$ dans $\mathbb{R}$ définie par : $L_j^{\theta^{-j}}(\beta) = L(\theta_1, \ldots, \theta_{j-1}, \beta, \theta_{j+1}, \ldots, \theta_p)$. Montrer que
$$L_j^{\theta^{-j}}(\beta) = \frac{1}{2N}\sum_{i=1}^{N}(r_i^{(j)} - z_i^{(j)}\beta)^2 + \lambda\sum_{k\neq j}|\theta_k| + \lambda|\beta|$$
où $r_i^{(j)}$ et $z_i^{(j)}$, $i = 1, \ldots, n$ désignent des réels que l'on explicitera.
2. Quitte à normaliser la matrice $\mathbf{X} = (x_{ij})_{i,j}$, on fait l'hypothèse que $\frac{1}{N}\sum_{j=1}^{N} x_{i,j}^2 = 1$ pour tout $i \in \{1, \ldots, n\}$. A l'aide de l'exercice précédent, montrez que la fonction $L_j^{\theta^{-j}}$ atteint son minimum en
$$\hat{\beta} = \mathcal{S}_\lambda\left(\frac{1}{N}\langle \mathbf{x}_j, \mathbf{r}^{(j)}\rangle\right)$$
où $\mathbf{r}^{(j)} = (r_1^{(j)}, \ldots, r_n^{(j)})^T$ et $\mathbf{x}_j$ désigne la j-ième colonne de la matrice $\mathbf{X}$.

L'algorithme de descente coordonnées par coordonnées fonctionne alors de la manière suivante. On initialise θ. On pose $\theta^{(0)} = 0$ par exemple. On considère alors la fonction
$$\beta \mapsto L_1^{(\theta^{(0)})^{-1}}(\beta)$$
dont on calcule le minimum $\hat{\beta}$ (Il se trouve qu'il est explicite !!). On définit alors $\theta^{(1)}$ par $\theta_1^{(1)} = \hat{\beta}$ et pour tout $j \neq 1$, $\theta_j^{(1)} = \theta_j^{(0)}$. A l'étape $j \leq p$, on dispose d'un $\theta^{(j)}$ et on reproduit le même schéma en figeant les variables différentes de j et en cherchant le minimum de $j \mapsto L_j^{(\theta^{(j)})^{-j}}$. Une fois les p variables passées, on recommence avec la première....

3. Montrez que la suite $(L(\theta^{(n)}))_n$ est décroissante.
4. Montrez que $(\theta^{(n)})$ admet une sous-suite convergente vers $\theta^{(\infty)} \in \mathbb{R}^p$.
5. Pourriez-vous expliquer sans le démontrer que $\theta^{(\infty)}$ est nécessairement un minimum ?
6. En déduire que $L(\theta^{(n)})$ converge vers $\min_\theta L(\theta)$.

Exercice 4.

1. Retrouvez la formule de $\hat{\theta}^{Ridge}$.
2. On admet les conditions de "stationnarité" : $\hat{\theta}_\lambda$ est minimum de $L(\theta) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|_1$ si et seulement si il est solution du système :

$$\begin{cases} \mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\theta) = \lambda \text{sgn}(\theta_j) & \text{si } j \in J(\theta) \\ |\mathbf{x}_j^T(\mathbf{y} - \mathbf{X}\theta)| \leq \lambda & \text{sinon.} \end{cases} \tag{1}$$

où $J(\theta) = \{j \in \{1, \dots, p\}, \theta_j \neq 0\}$. Montrez que dans le cas $\mathbf{X}^T\mathbf{X} = I_p$, alors

$$\hat{\theta}_\lambda(j) = \text{sgn}(\hat{\theta}_j)(|\theta_j| - \lambda)_+$$

où $\hat{\theta} = \hat{\theta}_0$, *i.e.* désigne la solution du système non pénalisé (bien définie dans ce cas).

3. On suppose que $\frac{\mathbf{X}^T\mathbf{X}}{n} = I_p$ [1] et que le modèle est de la forme $\mathbf{Y} = \mathbf{X}\theta^\star + \varepsilon$. Vérifiez que l'on peut écrire

$$(\hat{\theta}_\lambda)(j) = \mathcal{S}_\lambda\left(\frac{\langle \mathbf{x}_j, \mathbf{y}\rangle}{n}\right) = \mathcal{S}_\lambda(\theta_j^\star + \frac{(\mathbf{X}^T\varepsilon)_j}{n}).$$

Exercice 5 (λ_{max})**.** Dans cet exercice, on se pose la question suivante. Peut-on calculer la valeur de λ à partir de laquelle, le LASSO ne détecte aucune variable ? On pose

$$L(\theta) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|_1.$$

1. Montrez que

$$L(\theta) - L(0) = \frac{1}{2}\|\mathbf{X}\theta\|_2^2 - \langle \mathbf{y}, \mathbf{X}\theta\rangle + \lambda\|\theta\|_1.$$

2. Montrez que

$$\langle \mathbf{y}, \mathbf{X}\theta\rangle = \sum_{j=1}^{p} \theta_j \langle \mathbf{x}_j, \mathbf{y}\rangle.$$

3. On pose

$$\lambda_{max} = \max\{|\langle \mathbf{x}_j, \mathbf{y}\rangle|, j = 1 \dots, p\}.$$

Déduire de la question précédente que pour tout $\lambda > \lambda_{max}$, $L(\theta) - L(0) > 0$ pour tout $\theta \in \mathbb{R}^p$. Que vaut $\hat{\theta}_\lambda$ dans ce cas ?

4. Supposons maintenant $\lambda < \lambda_{max}$. Montrez que dans ce cas, il existe $j \in \{1, \dots, p\}$ tel que $\theta_j \mapsto L(0, \dots, 0, \theta_j, 0, \dots, 0)$ n'atteint pas son minimum en 0. Déterminez-le. Conclure que

$$\lambda_{max} = \sup\{\lambda, \hat{\theta}_\lambda \neq 0\}.$$

5. Interprétation : que pensez-vous de la valeur de λ_{max} ? Cela vous semble-t-il naturel que le LASSO détecte d'abord la variable qui maximise $|\langle \mathbf{x}_j, \mathbf{y}\rangle|$?

1. On omet souvent la division par n pour simplifier les notations mais elle est naturelle : pensez-par exemple au cas où $\mathbf{Y} = \theta^\star + \varepsilon$ (Que vaut $\mathbf{X}^T\mathbf{X}$ dans ce cas ?). Elle permet de faire apparaître la convergence vers $\theta^\star$.