TD 11 : Ridge et Lasso

Les questions marquées d'un astérisque (*) *sont facultatives.*

Soient $Y \in \mathbb{R}^n$ et $X \in \mathbb{R}^{n\times d}$ des variables aléatoires dont on suppose qu'elles obéissent au modèle linéaire

$$Y = X\beta + \varepsilon.$$

L'objectif est d'estimer $\beta \in \mathbb{R}^d$. Dans ce TD, on considèrera les trois estimateurs suivants :

— Ridge : $\hat{\beta}^{\mathrm{Ridge}} \in \arg\min_\beta(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2)$,

— Lasso : $\hat{\beta}^{\mathrm{Lasso}} \in \arg\min_\beta(\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1)$,

— Pénalité ℓ_0 : $\hat{\beta}^0 \in \arg\min_\beta(\|Y - X\beta\|_2^2 + \lambda|\beta|_0)$ où $|\beta|_0 = \mathrm{Card}\{i \in \{1,\dots,d\} \text{ t.q. } \beta_i \neq 0\}$.

Deux remarques :

— il n'existe pas en général de formule explicite pour ces deux derniers estimateurs,

— l'application $\beta \longmapsto |\beta|_0$ n'est pas une norme.

Exercice 1. (Formulations contrainte et pénalisée) L'objectif de cet exercice est de montrer l'équivalence entre les deux formulations des problèmes Ridge, Lasso et ℓ_0 :

$$\hat{\beta}_1(\lambda) \in \arg\min_\beta(\|Y - X\beta\|_2^2 + \lambda h(\beta)) \tag{1}$$

et

$$\hat{\beta}_2(t) \in \underset{\beta \text{ t.q. } h(\beta)\leqslant t}{\arg\min} \|Y - X\beta\|_2^2 \tag{2}$$

où $h(\beta) = \|\beta\|_2^2$ pour l'estimateur Ridge, $h(\beta) = \|\beta\|_1$ pour l'estimateur Lasso et $h(\beta) = |\beta|_0$ pour l'estimateur par pénalité ℓ_0.

Soit $\lambda > 0$, soit $\hat{\beta}_1(\lambda)$ défini dans (1) et posons $t = h(\hat{\beta}_1(\lambda))$.

1. Montrer que la quantité $\hat{\beta}_2(t)$ définie dans (2) vérifie $\|Y - X\hat{\beta}_2(t)\|_2^2 \leqslant \|Y - X\hat{\beta}_1(\lambda)\|_2^2$ et $h(\hat{\beta}_2(t)) \leqslant h(\hat{\beta}_1(\lambda))$.
2. En déduire que $\hat{\beta}_2(t)$ est une solution de (1).
3. Déduire de (1) et des question précédentes que $\|Y - X\hat{\beta}_2(t)\|_2^2 = \|Y - X\hat{\beta}_1(\lambda)\|_2^2$ et $h(\hat{\beta}_1(\lambda)) = h(\hat{\beta}_2(t))$, puis que $\hat{\beta}_1(\lambda)$ est une solution de (2).

Nous venons de montrer que pour tout $\lambda \geqslant 0$, il existe $t(\lambda)$ tel que les solutions du problème (1) *sont des solutions du* (2) *et réciproquement.*

Exercice 2. (Convexité et unicité des estimateurs) Dans la suite, on note $X^\top$ la matrice transposée de X.

1. Notons $\hat{\beta}^{LS}$ un estimateur des moindres carrés (LS = Least Squares), c'est-à-dire un minimiseur de $\beta \mapsto \|Y - X\beta\|_2^2$.
   (a) Justifier que $(Y - X\hat{\beta}^{LS})^\top X = 0$.
   Indication : passez par le gradient.
   (b) En déduire que pour tout $\beta \in \mathbb{R}^d$, $\|Y - X\beta\|_2^2 = \|Y - X\hat{\beta}^{LS}\|_2^2 + \|X(\hat{\beta}^{LS} - \beta)\|_2^2$.
   (c) Montrer que l'application
   $$\beta \longmapsto \|Y - X\beta\|_2^2 \tag{3}$$
   est convexe.
   Indication : on pourra montrer que si f est une fonction convexe à valeurs positives, alors f^2 est convexe.

Dans la suite, on pourra admettre les quelques propriétés suivantes. **(*)** Les montrer.

— Si f et g sont des fonctions convexes, alors $f+g$ est toujours convexe. Si l'une d'entre elles est strictement convexe, alors $f+g$ est aussi strictement convexe. Une fonction f est dite strictement convexe si pour tout $x \neq y$ et tout $t \in (0,1)$, $f(tx+(1-t)y) < tf(x)+(1-t)f(y)$.

— Une fonction continue strictement convexe qui tend vers l'infini en l'infini admet un unique minimiseur.

— L'application (3) est strictement convexe lorsque $X^\top X$ est définie positive.

Le grand avantage des fonctions convexe est qu'elles sont faciles à minimiser : il suffit de suivre la pente pour tomber au fond de la cuvette ! C'est le principe de la descente de gradient.

2. **(Ridge)** Montrer que l'estimateur Ridge est unique dès lors que $\lambda > 0$.
3. **(Lasso)** Montrer que l'estimateur Lasso est unique dès lors que $X^\top X$ est définie positive.
   L'estimateur Lasso n'est pas toujours unique ! Par contre, on peut toujours le calculer par descente de gradient.
4. **(Pénalité ℓ_0)** La fonction $\beta \longmapsto |\beta|_0$ est-elle convexe ?

Exercice 3. (Parcimonie) On se place ici dans le cas où d est grand et où β n'a que $r < d$ composantes non nulles, autrement dit $|\beta|_0 = r < d$. *Lorsque r est beaucoup plus petit de d, on dit que le vecteur est* parcimonieux. *Tirer profit de l'information qu'un vecteur est parcimonieux est crucial quand d est très grand.*

On suppose les $(\varepsilon_i)_{1\leqslant i\leqslant n}$ i.i.d. gaussiens centrés de variance σ^2. On se place dans le cas orthogonal où $X^\top X = n \cdot I_d$. Attention, il y a un facteur n en plus comparé au cours.

1. Montrer que l'estimateur des moindres carrés est $\hat\beta^{LS} = \frac{1}{n}X^\top Y$.
2. Quelle est la loi de $\hat\beta^{\mathrm{LS}}$? En déduire que presque sûrement, $|\hat\beta^{\mathrm{LS}}|_0 = d$.
3. Montrer que $\mathbb{E}[\|\hat\beta^{LS} - \beta\|^2] = \frac{\sigma^2 d}{n}$. En déduire $\hat\beta^{LS} \underset{n\to\infty}{\longrightarrow} \beta$ en probabilité (en supposant r et d constants).
4. **(Ridge)** Montrer que $\hat\beta^{\mathrm{Ridge}} = \frac{1}{1+\lambda/n}\hat\beta^{\mathrm{LS}}$. En déduire que quel que soit λ, presque sûrement, $|\hat\beta^{\mathrm{Ridge}}|_0 = d$.
5. **(Lasso)**
   (a) Montrer que $\hat\beta^{\mathrm{Lasso}}$ vérifie $-2\hat\beta_j^{LS} + 2\hat\beta_j^{\mathrm{Lasso}} + (\lambda/n)\cdot \mathrm{signe}(\hat\beta_j^{\mathrm{Lasso}}) = 0$ pour tout j tel que $\hat\beta_j^{\mathrm{Lasso}} \neq 0$.
   (b) En déduire que pour tout j,
$$\hat\beta_j^{\mathrm{Lasso}} = \begin{cases} \hat\beta_j^{LS} - \lambda/(2n) & \text{si } \hat\beta_j^{LS} \geqslant \lambda/(2n), \\ \hat\beta_j^{LS} + \lambda/(2n) & \text{si } \hat\beta_j^{LS} \leqslant -\lambda/(2n), \\ 0 & \text{si } \hat\beta_j^{LS} \in [-\lambda/(2n), \lambda/(2n)]. \end{cases} \tag{4}$$
   (c) Montrer que à n fixé, $|\hat\beta^{\mathrm{Lasso}}|_0 \underset{\lambda\to\infty}{\longrightarrow} 0$ et $|\hat\beta^{\mathrm{Lasso}}|_0 \underset{\lambda\to 0}{\longrightarrow} d$.
   (d) Montrer que si la valeur de λ/n est fixée (et strictement positive) et à r et d fixés, $\lim_{n\to\infty} |\hat\beta^{\mathrm{Lasso}}|_0 \leqslant r$ (avec égalité dès que λ/n est « assez petit » ; préciser le seuil).
   En pratique, le bon choix de λ est d'ordre $\sqrt{n}$ et non n comme précisé ici, mais la constante de proportionnalité n'admet pas de formule explicite et dépend de la situation.
6. **(Pénalité ℓ_0)** Montrer que $|\hat\beta^0|_0 \underset{\lambda\to\infty}{\longrightarrow} 0$ et $|\hat\beta^0|_0 \underset{\lambda\to 0}{\longrightarrow} d$.