Apprentissage statistique et grande dimension

TD 2 : régression en grande dimension

Dans tout le TD, nous considérons $Y = (Y_1, ..., Y_n)^t$, $X = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq d}$ non aléatoire vérifiant

$$Y = X\beta^* + \varepsilon,$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ et $\beta^* = (\beta_1^*, ..., \beta_d^*)^t \in \mathbb{R}^d$.

Exercice 1 (Moindres carrés et ridge et Lasso en dimension 1)

Nous supposons ici que $d = 1$, le modèle considéré peut-être réecrit

$$Y_i = \beta^* x_i + \varepsilon_i, \quad i = 1, ..., n,$$

avec $\beta \in \mathbb{R}$. L'objectif de cet exercice est de comparer dans ce cadre-là l'estimateur des moindres carrés

$$\widehat{\beta}^{(MC)} \in \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta x_i)^2,$$

l'estimateur *ridge* $\widehat{\beta}_\lambda^{(R)}$ et l'estimateur Lasso $\widehat{\beta}_\lambda^{(L)}$.

1. Donner l'écriture de l'estimateur $\widehat{\beta}^{(MC)}$ en fonction de $\{(Y_i, x_i), i = 1, ..., n\}$. Calculer le biais et la variance de cet estimateur.
2. Écrire le problème de minimisation que doit vérifier l'estimateur *ridge* dans ce cadre-là et donner son écriture.
3. Calculer son biais, sa variance et son risque quadratique.
4. Nous considérons maintenant l'estimateur Lasso, c'est-à-dire la solution du critère de minimisation

$$\widehat{\beta}_\lambda^{(L)} \in \arg\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta x_i)^2 + \lambda |\beta|.$$

Calculer la solution du problème de minimisation.

Réponse de l'exercice 1.

1. L'estimateur des moindres carrés minimise

$$\Delta = (Y - \boldsymbol{x}\beta)^t (Y - \boldsymbol{x}\beta) = (Y^t Y - 2Y^t \boldsymbol{x}\beta + \beta^t \boldsymbol{x}^t \boldsymbol{x}\beta)$$

La condition du premier ordre est

$$\Delta' = 0 \Leftrightarrow -2Y^t \boldsymbol{x} + 2\boldsymbol{x}^t \boldsymbol{x}\beta = 0 \Leftrightarrow \boldsymbol{x}^t \boldsymbol{x}\beta = Y^t \boldsymbol{x}$$

La condition du second ordre est la positivité de la dérivée seconde

$$\Delta'' = 2\boldsymbol{x}^t \boldsymbol{x}$$

Par conséquent, si $\boldsymbol{x}^t \boldsymbol{x} = ||\boldsymbol{x}||^2$ est une norme strictement positive, alors la fonction est convexe et possède un unique minimum

$$\beta^{(MC)} = (\boldsymbol{x}^t \boldsymbol{x})^{-1} Y^t \boldsymbol{x} = \frac{< Y, \boldsymbol{x} >}{||\boldsymbol{x}||^2}.$$

Déterminons le biais et la variance sous l'hypothèse que ces moments existent pour Y (homoscédasticité).

$$E(\beta^{(MC)}) = \frac{E(< Y, \boldsymbol{x} >)}{||\boldsymbol{x}||^2} = \frac{< E(Y), \boldsymbol{x} >}{||\boldsymbol{x}||^2} = \frac{< \boldsymbol{x}\beta^*, \boldsymbol{x} >}{||\boldsymbol{x}||^2} = \beta^*.$$

$$Var(\beta^{(MC)}) = \frac{Var(< Y, \boldsymbol{x} >)}{||\boldsymbol{x}||^4} = \frac{\boldsymbol{x}' Var(Y) \boldsymbol{x}}{||\boldsymbol{x}||^4} = \frac{\boldsymbol{x}' \sigma^2 I_n \boldsymbol{x}}{||\boldsymbol{x}||^4} = \frac{\sigma^2}{||\boldsymbol{x}||^2}.$$

2. Le problème est

$$\widehat{\beta}_\lambda^{(R)} \in \arg\min_{\beta\in\mathbb{R}} \sum_{i=1}^{n} (Y_i - \beta x_i)^2 + \lambda|\beta|^2.$$

L'estimateur minimise

$$\Delta = (Y - \boldsymbol{x}\beta)^t(Y - \boldsymbol{x}\beta) + \lambda\beta^2 = (Y^tY - 2Y^t\boldsymbol{x}\beta + \beta^t\boldsymbol{x}^t\boldsymbol{x}\beta) + \lambda\beta^2$$

La condition du premier ordre est

$$\Delta' = 0 \Leftrightarrow -2Y^t\boldsymbol{x} + 2\boldsymbol{x}^t\boldsymbol{x}\beta + 2\lambda\beta = 0 \Leftrightarrow (\boldsymbol{x}^t\boldsymbol{x} + \lambda)\beta = Y^t\boldsymbol{x}$$

La condition du second ordre est la positivité de la dérivée seconde

$$\Delta'' = 2\boldsymbol{x}^t\boldsymbol{x} + 2\lambda$$

Par conséquent, si $||\boldsymbol{x}||^2 + n\lambda$ est une norme strictement positive, alors la fonction est convexe et possède un unique minimum

$$\beta^{(R)} = (\boldsymbol{x}^tX)^{-1}Y^t\boldsymbol{x} = \frac{< Y, \boldsymbol{x} >}{||\boldsymbol{x}||^2 + \lambda}.$$

3. Déterminons le biais et la variance sous l'hypothèse que ces moments existent pour Y (homoscédasticité).

$$E(\beta^{(R)}) = \frac{< E(Y), \boldsymbol{x} >}{||\boldsymbol{x}||^2 + \lambda} = \frac{\beta^* ||\boldsymbol{x}||^2}{||\boldsymbol{x}||^2 + \lambda},\ Var(\beta^{(R)}) = \frac{\boldsymbol{x}'Var(Y)\boldsymbol{x}}{(||\boldsymbol{x}||^2 + \lambda)^2} = \frac{\sigma^2 ||\boldsymbol{x}||^2}{(||\boldsymbol{x}||^2 + \lambda)^2}.$$

L'EQM ou MSE se déduit par

$$MSE(\beta^{(R)}) = (E(\beta^{(R)}) - \beta^\star)^2 + Var(\beta^{(R)}) = \left(\frac{\beta^* ||\boldsymbol{x}||^2}{||\boldsymbol{x}||^2 + \lambda} - \beta^*\right)^2 + \frac{\sigma^2 ||\boldsymbol{x}||^2}{(||\boldsymbol{x}||^2 + \lambda)^2}$$

$$= (\beta^*)^2 \left(\frac{n\lambda}{||\boldsymbol{x}||^2 + \lambda}\right)^2 + \frac{\sigma^2 ||\boldsymbol{x}||^2}{(||\boldsymbol{x}||^2 + \lambda)^2}$$

Il existe une valeur de λ minimisant l'EQM.

4. **Question intermédiaire :**
Cherchons le minimum de la fonction suivante $f(x) = a|x| + bx^2 + cx$ sur $\mathbb{R}$ avec $a, b > 0$ et $c \in \mathbb{R}$.
— Elle est strictement convexe en tant que somme de fonction strictement convexe.
— Elle possède donc une dérivée à gauche et à droite en tout point (égales pour tout $x \neq 0$).
— Donc sa sous-différentielle est

$$\forall x > 0, \partial f(x) = \{a + 2bx + c\},\ \forall x < 0, \partial f(x) = \{-a + 2bx + c\},\ \partial f(0) = [-a + c, a + c].$$

L'optimalité est atteinte en $x^\star$ lorsque $0 \in \partial f(x^\star)$. Résolvons les trois cas séparément.
— cas positif

$$\left\{ \begin{array}{l} x > 0 \\ a + 2bx + c = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = (-a - c)/2b \\ x > 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = \frac{-c}{2b}\left(1 - \frac{a}{-c}\right) \\ -a > c \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = \frac{-c}{2b}\left(1 - \frac{a}{|c|}\right) \\ 0 > -a > c \end{array} \right.$$

— cas négatif

$$\left\{ \begin{array}{l} x < 0 \\ -a + 2bx + c = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = (a - c)/2b \\ x < 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = \frac{-c}{2b}\left(1 - \frac{a}{c}\right) \\ a < c \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = \frac{-c}{2b}\left(1 - \frac{a}{|c|}\right) \\ 0 < a < c \end{array} \right.$$

— cas nul

$$\left\{ \begin{array}{l} x = 0 \\ -a + c \leq 0 \leq a + c \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = 0 \\ -a \leq c \leq a \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} x = 0 \\ |c| \leq a \end{array} \right.$$

Ainsi la solution se réécrit comme la partie positive suivante

$$x^\star = \frac{-c}{2b}\left(1 - \frac{a}{|c|}\right)_+.$$

Retour à la question :
L'estimateur minimise

$$\Delta = (Y - \boldsymbol{x}\beta)^t(Y - \boldsymbol{x}\beta) + \lambda|\beta| = (Y^tY - 2Y^t\boldsymbol{x}\beta + \boldsymbol{x}^t\boldsymbol{x}\beta^2) + \lambda|\beta|$$

Autrement dit c'est la fonction $f(\beta)$ avec $a = \lambda$, $b = \boldsymbol{x}^t\boldsymbol{x}$, $c = -2Y^t\boldsymbol{x}$ plus une constante Y^tY sans importance. Donc la solution est

$$\beta^\star = \frac{2Y^t\boldsymbol{x}}{2\boldsymbol{x}^t\boldsymbol{x}}\left(1 - \frac{\lambda}{|2Y^t\boldsymbol{x}|}\right)_+ = \frac{Y^t\boldsymbol{x}}{\boldsymbol{x}^t\boldsymbol{x}}\left(1 - \frac{\lambda}{2|Y^t\boldsymbol{x}|}\right)_+ = \beta^{(MC)}\left(1 - \frac{\lambda}{2|Y^t\boldsymbol{x}|}\right)_+.$$

Exercice 2 (Propriétés de l'estimateur Ridge)

Nous considérons, pour $\lambda > 0$, l'estimateur Ridge

$$\widehat{\beta}_\lambda^{(R)} = \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

L'objectif de cet exercice est de prouver les propriétés de l'estimateurs Ridge vues en cours.

1. Montrer que le problème de minimisation

$$\min_{\beta\in\mathbb{R}^d}\left\{\sum_{i=1}^n(Y_i - X_i\beta)^2 + \lambda\|\beta\|^2\right\} \tag{1}$$

admet $\widehat{\beta}_\lambda^{(R)}$ comme unique solution.

2. Montrer que toute solution du problème d'optimisation sous contrainte suivant

$$\min_{\beta\in\mathbb{R}^d,\|\beta\|\le M_\lambda}\left\{\sum_{i=1}^n(Y_i - X_i\beta)^2\right\}, \tag{2}$$

avec $M_\lambda = \left\|\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbf{Y}\right\|$ est aussi solution du problème (1). En déduire que $\widehat{\beta}_\lambda^{(R)}$ est aussi l'unique solution du problème (2).

3. Exprimer la norme au carré du biais de $\widehat{\beta}_\lambda^{(R)}$ en fonction des valeurs propres $\lambda_1, \ldots, \lambda_d$ (comptées avec multiplicité) de $\mathbf{X}^t\mathbf{X}$.

$$B_\lambda^{(R)} := \left\|\mathbb{E}\left[\widehat{\beta}_\lambda^{(R)}\right] - \beta^*\right\|^2$$

4. Exprimer le variance

$$V_\lambda^{(R)} = \mathbb{E}\left[\left\|\widehat{\beta}_\lambda^{(R)} - \mathbb{E}\left[\widehat{\beta}_\lambda^{(R)}\right]\right\|^2\right]$$

de $\widehat{\beta}_\lambda^{(R)}$ en fonction de la variance du bruit σ^2 et des valeurs propres $\lambda_1, \ldots, \lambda_d$.

Réponse de l'exercice 2.

1. La fonction $\beta \mapsto \lambda\|\beta\|^2$ est fortement convexe car, pour tout $\beta_1, \beta_2 \in \mathbb{R}^d$, $t \in]0,1[$,

$$\begin{aligned}\lambda\|t\beta_1 + (1-t)\beta_2\|^2 &= \lambda\left(t^2\|\beta_1\|^2 + 2t(1-t)\langle\beta_1,\beta_2\rangle + (1-t)^2\|\beta_2\|^2\right)\\ &= t^2\lambda\|\beta_1\|^2 + (1-t)^2\lambda\|\beta_2\|^2 + \lambda t(1-t)(-\|\beta_1-\beta_2\|^2 + \|\beta_1\|^2 + \|\beta_2\|^2)\\ &= t\lambda\|\beta_1\|^2 + (1-t)\lambda\|\beta_2\|^2 - \lambda t(1-t)\|\beta_1-\beta_2\|^2.\end{aligned}$$

Comme la fonction à minimiser

$$f_\lambda^{(R)}(\beta) = \sum_{i=1}^n(Y_i - X_i\beta)^2 + \lambda\|\beta\|^2,$$

est fortement convexe (car somme d'une fonction convexe $f_0(\beta) = \sum_{i=1}^n(Y_i - X_i\beta)^2$ et d'une fonction fortement convexe $\beta \mapsto \lambda\|\beta\|^2$) et de classe $\mathcal{C}^1$, elle admet un unique minimum $\overline{\beta}$ sur $\mathbb{R}^d$ qui vérifie l'équation d'Euler

$$\nabla f_\lambda^{(R)}(\overline{\beta}) = 0.$$

Nous avons

$$\nabla f_\lambda^{(R)}(\beta) = -2\sum_{i=1}^{n} X_i(Y_i - X_i\beta) + 2\lambda\beta = -2\mathbf{X}^t\mathbf{Y} + 2\mathbf{X}^t\mathbf{X}\beta + 2\lambda\beta,$$

nous obtenons bien que $\overline{\beta}$ est l'unique solution du système linéaire

$$(\mathbf{X}^t\mathbf{X} + \lambda I)\overline{\beta} = \mathbf{X}^t\mathbf{Y},$$

c'est-à-dire que $\overline{\beta} = \widehat{\beta}_\lambda^{(R)}$.

2. Remarquons déjà que la fonction f_0 étant continue sur le compact $\{\beta \in \mathbb{R}^d, \|\beta\| \leq M_\lambda$, le problème (2) admet au moins une solution. Soit $\widetilde{\beta}$ une telle solution. Nous savons par définition que :

$$\|\widetilde{\beta}\| \leq M_\lambda \text{ et } f_0(\widetilde{\beta}) \leq f_0(\beta),$$

pour tout $\beta \in \mathbb{R}^d$ tel que $\|\beta\| \leq M_\lambda$.
Comme

$$\|\widehat{\beta}_\lambda^{(R)}\| = \| \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1} \mathbf{X}^t\mathbf{Y}\| = M_\lambda,$$

$\widehat{\beta}_\lambda^{(R)}$ vérifie la contrainte ce qui implique que

$$f_0(\widetilde{\beta}) \leq f_0(\widehat{\beta}_\lambda^{(R)})$$

et donc que

$$f_\lambda^{(R)}(\widetilde{\beta}) = f_0(\widetilde{\beta}) + \lambda\|\widetilde{\beta}\|^2 \leq f_0(\widehat{\beta}_\lambda^{(R)}) + \lambda\|\widetilde{\beta}\|^2.$$

Comme $\|\widetilde{\beta}\| \leq M_\lambda$, nous avons donc

$$f_\lambda^{(R)}(\widetilde{\beta}) \leq f_0(\widehat{\beta}_\lambda^{(R)}) + \lambda M_\lambda^2 = f_0(\widehat{\beta}_\lambda^{(R)}) + \lambda\|\widehat{\beta}_\lambda^{(R)}\|^2 = f_\lambda^{(R)}(\widehat{\beta}_\lambda^{(R)}) \leq f_\lambda^{(R)}(\beta),$$

pour tout $\beta \in \mathbb{R}^d$ car $\widehat{\beta}_\lambda^{(R)}$ minimise $f_\lambda^{(R)}$ sur $\mathbb{R}^d$. Donc, $\widetilde{\beta}$ est aussi solution de (1) et, par unicité, $\widetilde{\beta} = \widehat{\beta}_\lambda^{(R)}$.

3.

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_\lambda^{(R)}] &= \mathbb{E}[\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbf{Y}] = \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbb{E}[\mathbf{Y}] = \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbf{X}\beta^*. \\
&= \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}(\mathbf{X}^t\mathbf{X} + \lambda I - \lambda I)\beta^* \\
&= \beta^* + \lambda\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\beta^*.
\end{aligned}$$

D'où

$$B_\lambda^{(R)} = \left\|\lambda\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\beta^*\right\|^2$$

Soit P une matrice orthogonale telle que $\mathbf{X}^t\mathbf{X} = PDP^t$ avec D la matrice diagonale de coefficients $(\lambda_1, \ldots, \lambda_d)$, nous avons

$$\begin{aligned}
\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\beta^* &= \left(PDP^t + \lambda I\right)^{-1}\beta^* = \left(PDP^t + \lambda PP^t\right)^{-1}\beta^* = \left(P(D + \lambda I)P^t\right)^{-1}\beta^* \\
&= P(D + \lambda I)^{-1}P^t\beta^*.
\end{aligned}$$

D'où, comme P est orthogonale,

$$B_\lambda^{(R)} = \lambda^2\left\|P(D + \lambda I)^{-1}P^t\beta^*\right\|^2 = \lambda^2\left\|(D + \lambda I)^{-1}P^t\beta^*\right\|^2 = \sum_{j=1}^{d}\frac{\lambda^2}{(\lambda_j + \lambda)^2}[P^t\beta^*]_j^2.$$

4. Nous remarquons que, étant donné que $\mathbf{Y}$ est un vecteur gaussien de moyenne $\mathbf{X}^t\beta^*$ et de matrice de covariance $\sigma^2 I$, $\widehat{\beta}_\lambda^{(R)} = \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbf{Y}$ est aussi un vecteur gaussien, de moyenne $\mathbb{E}[\widehat{\beta}_\lambda^{(R)}]$ (calculée dans la question précédente) et de matrice de covariance

$$\Sigma_\lambda^{(R)} = \left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\sigma^2 I\left(\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\right)^t = \sigma^2\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}\mathbf{X}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \lambda I\right)^{-1}.$$

Cela nous donne

$$V_\lambda^{(R)} = \sum_{j=1}^{d}\mathbb{E}\left[\left([\widehat{\beta}_\lambda^{(R)}]_j - \mathbb{E}\left[[\widehat{\beta}_\lambda^{(R)}]_j\right]\right)^2\right] = \sum_{j=1}^{d}\mathrm{Var}([\widehat{\beta}_\lambda^{(R)}]_j) = \sum_{j=1}^{d}\left[\Sigma_\lambda^{(R)}\right]_{j,j} = \mathrm{tr}(\Sigma_\lambda^{(R)}) = \sigma^2\sum_{j=1}^{d}\frac{\lambda_j}{(\lambda_j + \lambda)^2}.$$

Exercice 3 (Propriétés de l'estimateur Lasso)

L'objectif est de montrer que les problèmes d'optimisation

$$\min_{\beta\in\mathbb{R}^d}\left\{\sum_{i=1}^n(Y_i-X_i\beta)^2+\lambda\|\beta\|_1\right\} \tag{3}$$

et

$$\min_{\beta\in\mathbb{R}^d,\|\beta\|_1\leq M_\lambda}\left\{\sum_{i=1}^n(Y_i-X_i\beta)^2\right\}, \tag{4}$$

sont équivalents lorsque M_λ est bien choisi dans le sens où l'ensemble des solutions des deux problèmes est identique (on rappelle que pour le LASSO on n'a pas unicité de la solution).

1. Montrer que, pour toutes solutions $\widehat{\beta}_\lambda^{(L,1)}$ et $\widehat{\beta}_\lambda^{(L,2)}$ du problème pénalisé (3),
$$\left\|\widehat{\beta}_\lambda^{(L,1)}\right\|_1=\left\|\widehat{\beta}_\lambda^{(L,2)}\right\|_1.$$
Nous noterons par la suite cette valeur commune N_λ.
2. Montrer que toute solution de (3) est aussi solution de (4) lorsque $M_\lambda=N_\lambda$.
3. Montrer que toute solution de (4) est aussi solution de (3) lorsque $M_\lambda=N_\lambda$.

Réponse de l'exercice 3.

1. Notons
$$f_\lambda^{(L)}(\beta):=\sum_{i=1}^n(Y_i-X_i\beta)^2+\lambda\|\beta\|_1=f_0(\beta)+\|\beta\|_1,$$
la fonction à minimiser dans le problème (3). Comme $\widehat{\beta}_\lambda^{(L,1)}$ et $\widehat{\beta}_\lambda^{(L,2)}$ minimisent $f_\lambda^{(L)}$, nécessairement :
$$f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,1)})=f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,2)}).$$
— Montrons que, pour tout $i=1,\dots,n$, $X_i\widehat{\beta}_\lambda^{(L,1)}=X_i\widehat{\beta}_\lambda^{(L,2)}$, ce qui implique que $f_0(\widehat{\beta}_\lambda^{(L,1)})=f_0(\widehat{\beta}_\lambda^{(L,2)})$. Supposons par l'absurde qu'il existe j tel que $X_j\widehat{\beta}_\lambda^{(L,1)}\neq X_j\widehat{\beta}_\lambda^{(L,2)}$ et notons
$$\tilde{\beta}=\frac{1}{2}\widehat{\beta}_\lambda^{(L,1)}+\frac{1}{2}\widehat{\beta}_\lambda^{(L,2)}.$$
Dans ce cas, par stricte convexité de la fonction $x\mapsto x^2$,
$$(Y_j-X_j\tilde{\beta})^2=\left(\frac{1}{2}(Y_j-X_j\widehat{\beta}_\lambda^{(L,1)})+\frac{1}{2}(Y_j-X_j\widehat{\beta}_\lambda^{(L,2)})\right)^2<\frac{1}{2}(Y_j-X_j\widehat{\beta}_\lambda^{(L,1)})^2+\frac{1}{2}(Y_j-X_j\widehat{\beta}_\lambda^{(L,2)})^2,$$
et pour $i\neq j$,
$$(Y_i-X_i\tilde{\beta})^2\leq\frac{1}{2}(Y_i-X_i\widehat{\beta}_\lambda^{(L,1)})^2+\frac{1}{2}(Y_i-X_i\widehat{\beta}_\lambda^{(L,2)})^2,$$
ce qui implique que
$$f_0(\tilde{\beta})<\frac{1}{2}f_0(\widehat{\beta}_\lambda^{(L,1)})+\frac{1}{2}f_0(\widehat{\beta}_\lambda^{(L,2)}).$$
D'un autre côté, l'inégalité triangulaire nous donne
$$\|\tilde{\beta}\|_1\leq\frac{1}{2}\|\widehat{\beta}_\lambda^{(L,1)}\|_1+\frac{1}{2}\|\widehat{\beta}_\lambda^{(L,2)}\|_1.$$
D'où
$$f_\lambda^{(L)}(\tilde{\beta})<\frac{1}{2}f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,1)})+\frac{1}{2}f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,2)})\leq f_\lambda^{(L)}(\tilde{\beta}),$$
la dernière inégalité provenant du fait que, comme $\widehat{\beta}_\lambda^{(L,1)}$ et $\widehat{\beta}_\lambda^{(L,2)}$ minimisent $f_\lambda^{(L)}$, alors $f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,1)})=f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,2)})\leq f_\lambda^{(L)}(\tilde{\beta}_\lambda)$. On aboutit à une absurdité.
— Puisque $f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,1)})=f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L,2)})$ et $f_0(\widehat{\beta}_\lambda^{(L,1)})=f_0(\widehat{\beta}_\lambda^{(L,2)})$ alors nécessairement
$$\left\|\widehat{\beta}_\lambda^{(L,1)}\right\|_1=\left\|\widehat{\beta}_\lambda^{(L,2)}\right\|_1,$$
d'où le résultat.

2. Soit $\widehat{\beta}_\lambda^{(L)}$ une solution du problème (3). Si $M_\lambda = N_\lambda$, $\widehat{\beta}_\lambda^{(L)}$ vérifie la contrainte $\|\widehat{\beta}_\lambda^{(L)}\|_1 \leq M_\lambda$. D'autre part, soit $\beta \in \mathbb{R}^d$ quelconque vérifiant aussi la contrainte $\|\beta\|_1 \leq M_\lambda$, alors, comme $\widehat{\beta}_\lambda^{(L)}$ minimise $f_\lambda^{(L)}$,

$$f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L)}) \leq f_\lambda^{(L)}(\beta).$$

D'où

$$f_0(\widehat{\beta}_\lambda^{(L)}) + \lambda\|\widehat{\beta}_\lambda^{(L)}\|_1 \leq f_0(\beta) + \lambda\|\beta\|_1,$$

ce qui implique

$$f_0(\widehat{\beta}_\lambda^{(L)}) \leq f_0(\beta) + \lambda(\|\beta\|_1 - \|\widehat{\beta}_\lambda^{(L)}\|_1) = f_0(\beta) + \lambda(\|\beta\|_1 - N_\lambda) \leq f_0(\beta),$$

car $\|\beta\|_1 \leq M_\lambda = N_\lambda$. Cela implique bien que $\widehat{\beta}_\lambda^{(L)}$ est également solution du problème (4).

3. Soit $\widetilde{\beta}_{M_\lambda}$ une solution du problème (4) et $\widehat{\beta}_\lambda^{(L)}$ une solution du problème (3). Nous allons montrer que

$$f_\lambda^{(L)}(\widetilde{\beta}_{M_\lambda}) \leq f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L)}),$$

ce qui implique bien que $\widetilde{\beta}_\lambda$ est aussi solution du problème (3). Notons que, par définition, $M_\lambda = \|\widehat{\beta}_\lambda^{(L)}\|_1$, donc $\widehat{\beta}_\lambda^{(L)}$ vérifie bien la contrainte $\|\widehat{\beta}_\lambda^{(L)}\|_1 \leq M_\lambda$ ce qui implique que

$$f_0(\widetilde{\beta}_{M_\lambda}) \leq f_0(\widehat{\beta}_\lambda^{(L)}).$$

D'où

$$f_\lambda^{(L)}(\widetilde{\beta}_{M_\lambda}) = f_0(\widetilde{\beta}_{M_\lambda}) + \|\widetilde{\beta}_{M_\lambda})\|_1 \leq f_0(\widehat{\beta}_\lambda^{(L)}) + M_\lambda = f_\lambda^{(L)}(\widehat{\beta}_\lambda^{(L)}).$$

Exercice 4 (Elastic net)

Nous considérons le problème de minimisation suivant :

$$\widehat{\beta}_{\lambda,\mu}^{(EN)} \in \underset{\beta\in\mathbb{R}^d}{\arg\min}\ \mathcal{L}_{EN}(\beta),$$

avec

$$\mathcal{L}_{EN}(\beta) = \sum_{i=1}^n (Y_i - x_i^t\beta)^2 + \lambda\|\beta\|^2 + \mu\|\beta\|_1.$$

1. Que se passe-t'il dans le cas $\lambda = 0$? Dans le cas $\mu = 0$?
2. Supposons que la condition ORT est vérifiée. Calculer la valeur minimale de $\mathcal{L}_{EN}$.
3. Pour tout $j = 1, ..., d$, calculer la dérivée partielle de $\mathcal{L}_{EN}(\beta)$ par rapport à $\beta_j \neq 0$.
4. En déduire un algorithme de descente de gradient coordonnées par coordonnées pour approcher $\widehat{\beta}_{\lambda,\mu}^{(EN)}$ lorsque, pour tous $j = 1, ..., d$, $n^{-1}\sum_{i=1}^n x_{i,j}^2 = 1$.

Réponse de l'exercice 4.

1. $\lambda = 0$ on retrouve l'estimateur lasso ; $\mu = 0$ on retrouve l'estimateur de Ridge.
2. Notons $f(\beta) = \sum_{i=1}^n (Y_i - x_i^t\beta)^2 + \lambda\|\beta\|^2$ la partie différentiable de la fonction objective. D'après le cours, le point optimal β^* vérifie qu'il existe un point de la sous-différentiel $\delta^* \in \partial N_1(\beta^*)$ tel que $\nabla f(\beta^*) + \mu\delta^* = 0$. Déterminons le gradient

$$\frac{\partial f}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial}{\partial\beta_j}(Y_i - x_i^t\beta)^2 + \lambda\frac{\partial}{\partial\beta_j}\|\beta\|^2 = -\sum_{i=1}^n 2x_{i,j}(Y_i - x_i^t\beta) + \lambda 2\beta_j = -2(X^tY)_j + 2(X^tX\beta)_j + \lambda 2\beta_j$$

Donc $\nabla f = -2X^tY + 2X^tX\beta + \lambda 2\beta$. Ainsi la condition d'optimalité devient avec $N_1(\beta) = \|\beta\|_1$

$$\exists\delta^* \in \partial N_1(\beta^*),\ -2X^tY + 2X^tX\beta^* + 2\lambda\beta^* + \mu\delta^* = 0$$

En utilisant la condition ORT $X^tX = nI_d$, on trouve pour la première équation

$$n\beta^* + \lambda\beta^* = X^tY - \frac{\mu\delta^*}{2} \Leftrightarrow \beta^* = \frac{1}{n+\lambda}X^tY - \frac{\mu\delta^*}{2(n+\lambda)}.$$

Rappelons que la sous différentielle de la valeurs absolue au point x est

$$\partial|x| \stackrel{def}{=} \mathrm{co}\{f'_+(x), f'_-(x)\} = \left\{ \begin{array}{ll} \{1\} & \text{si } x > 0 \\ [-1, 1] & \text{si } x = 0 \\ \{-1\} & \text{si } x < 0 \end{array} \right. = \left\{ \begin{array}{ll} \{\mathrm{sign}(x)\} & \text{si } x \neq 0 \\ [-1, 1] & \text{si } x = 0 \end{array} \right.$$

Donc pour la fonction vectorielle N_1, la sous-différentielle est le produit cartésien $\partial N_1(x) = \partial|x_0| \times \cdots \times \partial|x_d|$.

Réécrivons le système

$$\left\{ \begin{array}{l} \delta^* \in \partial N_1(\beta^*) \\ \beta^* = \frac{1}{n+\lambda} X^t Y - \frac{\mu \delta^*}{2(n+\lambda)} \end{array} \right.$$

suivant le signe de β_j^*. Si $\beta_j^* \neq 0$ alors $\delta_j^* = \mathrm{sign}(\beta_j^*)$. Donc le système se réécrit

$$\beta_j^* = \frac{1}{n+\lambda}(X^t Y)_j - \frac{\mu \,\mathrm{sign}(\beta_j^*)}{2(n+\lambda)}$$

Sinon $\beta_j^* = 0$ et $\delta^* \in [-1, 1]$.
Réécrivons en 3 situations

— cas négatif

$$\left\{ \begin{array}{l} \beta_j^* < 0 \\ \beta_j^* = \frac{1}{(n+\lambda)}(X^t Y)_j + \frac{\mu}{2(n+\lambda)} \\ \delta_j^* = -1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \beta_j^* = \frac{1}{(n+\lambda)}(X^t Y)_j + \frac{\mu}{2(n+\lambda)} \\ \frac{1}{(n+\lambda)}(X^t Y)_j + \frac{\mu}{2(n+\lambda)} < 0 \\ \delta_j^* = -1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \beta_j^* = \frac{(X^t Y)_j}{(n+\lambda)} \left(1 - \frac{\mu}{-2(X^t Y)_j}\right) \\ (X^t Y)_j < -\frac{\mu}{2} \\ \delta_j^* = -1 \end{array} \right.$$

— cas positif

$$\left\{ \begin{array}{l} \beta_j^* > 0 \\ \beta_j^* = \frac{1}{(n+\lambda)}(X^t Y)_j - \frac{\mu}{2(n+\lambda)} \\ \delta_j^* = 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \beta_j^* = \frac{1}{(n+\lambda)}(X^t Y)_j - \frac{\mu}{2(n+\lambda)} \\ \frac{1}{(n+\lambda)}(X^t Y)_j - \frac{\mu}{2(n+\lambda)} > 0 \\ \delta_j^* = 1 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \beta_j^* = \frac{(X^t Y)_j}{(n+\lambda)} \left(1 - \frac{\mu}{2(X^t Y)_j}\right) \\ (X^t Y)_j > \frac{\mu}{2} \\ \delta_j^* = 1 \end{array} \right.$$

— cas nul

$$\left\{ \begin{array}{l} \beta_j^* = 0 \\ \delta_j^* \in [-1, 1] \\ \beta_j^* = \frac{1}{(n+\lambda)}(X^t Y)_j - \frac{\mu \delta_j^*}{2(n+\lambda)} \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \beta_j^* = 0 \\ (X^t Y)_j = \frac{\mu \delta_j^*}{2} \\ \delta_j^* \in [-1, 1] \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \beta_j^* = 0 \\ (X^t Y)_j = \frac{\mu \delta_j^*}{2} \\ -\frac{\mu}{2} < (X^t Y)_j < \frac{\mu}{2} \end{array} \right.$$

Autrement dit, on a

$$\beta_j^* = \frac{(X^t Y)_j}{(n+\lambda)} \left(1 - \frac{\mu}{2|(X^t Y)_j|}\right)_+, \delta_j^* = \mathrm{sign}\left((X^t Y)_j - \frac{\mu}{2}\right).$$

3.

$$\frac{\partial \mathcal{L}_{EN}(\beta)}{\partial \beta_j} = -2(X^t Y)_j + 2(X^t X \beta)_j + 2\lambda\beta_j + \mu \,\mathrm{sign}(\beta_j)$$

4. Le méta-algorithme de descente coordonnée par coordonnée est le suivant
   (a) initialisation : $\beta^{(0)} \in \mathbb{R}^{d+1}$
   (b) répéter l'itération $t+1$ jusqu'à convergence
   pour $j = 0, \ldots, d$,

$$\beta_j^{(t+1)} \in \arg\max_{x \in \mathbb{R}} \mathcal{L}_{EN}(\beta_0^{(t)}, \ldots, \beta_{j-1}^{(t)}, x, \beta_{j+1}^{(t)}, \ldots, \beta_d^{(t)})$$

   fin pour

Autrement dit on optimise la $j$ème coordonnée en fixant la valeurs des d autres composantes. Notons $\beta_{-j} = (\beta_0, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_d)$ le vecteur β sans sa $j$ème composante. Pour β_{-j} donné, définissons la fonction suivante

$$\begin{aligned} g(\beta_j) &= \mathcal{L}_{EN}(\beta_j, \beta_{-j}) = \sum_{i=1}^n (Y_i - \boldsymbol{x}_{i,-j}^t \beta_{-j} - x_{i,j}\beta_j)^2 + \lambda\|\beta_{-j}\|_2^2 + \lambda(\beta_j)^2 + \mu\|\beta_{-j}\|_1 + \lambda|\beta_j| \\ &= \beta_j^2 \sum_{i=1}^n x_{i,j}^2 - 2\beta_j \sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t \beta_{-j}) + \mu|\beta_j| + \lambda(\beta_j)^2 + \sum_{i=1}^n (Y_i - \boldsymbol{x}_{i,-j}^t \beta_{-j})^2 + \lambda\|\beta_{-j}\|_2^2 + \mu\|\beta_{-j}\|_1 \\ &= \beta_j^2 \sum_{i=1}^n x_{i,j}^2 - 2\beta_j \sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t \beta_{-j}) + \mu|\beta_j| + \lambda(\beta_j)^2 + \text{constante} \end{aligned}$$

En utilisant l'exercice précédent on sait que la fonction $x \mapsto a|x| + bx^2 + cx$ sur $\mathbb{R}$ avec $a, b > 0$ et $c \in \mathbb{R}$ atteint son minimum en $x^\star = \frac{-c}{2b}\left(1 - \frac{a}{|c|}\right)_+$. Ici

$$a = \mu, b = \sum_{i=1}^n x_{i,j}^2 + \lambda, c = -2\sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t \beta_{-j}).$$

Donc la solution est

$$\beta_j^\star = 2\sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t\beta_{-j})\frac{1}{2\sum_{i=1}^n x_{i,j}^2 + 2\lambda}\left(1 - \frac{\mu}{|2\sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t\beta_{-j})|}\right)_+.$$

Posons

$$R_j(\beta_{-j}) = \sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t\beta_{-j}) = \sum_{i=1}^n x_{i,j}(Y_i - \sum_{k\neq j} x_{i,k}\beta_k).$$

Ainsi

$$\beta_j^\star = \frac{R_j(\beta_{-j})}{\sum\limits_{i=1}^n x_{i,j}^2 + \lambda}\left(1 - \frac{\mu}{2|R_j(\beta_{-j})|}\right)_+.$$

Sous ORT $\frac{1}{n}\sum_{i=1}^n x_{i,j}^2 = 1$, l'expression se simplifie légèrement $\beta_j^\star = \frac{R_j(\beta_{-j})}{n+\lambda}\left(1 - \frac{\mu}{2|R_j(\beta_{-j})|}\right)_+$.

L'algorithme de descente coordonnée par coordonnée est le suivant

(a) initialisation : choisir $\beta^{(0)} \in \mathbb{R}^{d+1}$.

(b) répéter l'itération $t+1$ jusqu'à convergence

pour $j = 0, \ldots, d$,

i. calcul de $R_j(\beta_{-j}^{(t)})$

$$R_j^{(t)} = R_j(\beta_{-j}^{(t)}) = \sum_{i=1}^n x_{i,j}(Y_i - \boldsymbol{x}_{i,-j}^t\beta_{-j}^{(t)}).$$

ii. mise à jour de $\beta_j^{(t+1)}$

$$\beta_j^{(t+1)} = \frac{R_j^{(t)}}{\sum\limits_{i=1}^n x_{i,j}^2 + \lambda}\left(1 - \frac{\mu}{2|R_j^{(t)}|}\right)_+.$$

fin pour

Exercice 5 (Group-Lasso)

Nous supposons maintenant que les covariables se répartissent en deux groupes c'est-à-dire que nous séparons l'ensemble $\{1, ..., d\}$ en deux sous-ensembles $I_1 := \{1, ..., d_1\}$ et $I_2 := \{d_1 + 1, ..., d\}$ avec $2 \leq d_1 \leq d - 1$. Nous souhaitons soit garder tous les coefficients β_j pour $j \in I_1$ non nuls, soit les annuler tous en même temps, de même pour les coefficients β_j pour $j \in I_2$. Nous considérons que l'ordonnée à l'origine est nulle ($\beta_0 = 0$ et $X = (x_{i,j})_{1\leq i\leq n, 1\leq j\leq d}$). Nous considérons le problème de minimisation suivant

$$\widehat{\beta}_{\lambda_1,\lambda_2}^{(GL)} \in \arg\min_{\beta\in\mathbb{R}^d} \mathcal{L}_G(\beta), \tag{5}$$

avec

$$\mathcal{L}_G(\beta) = \sum_{i=1}^n (Y_i - x_i^t\beta)^2 + \lambda_1\|\beta\|_{I_1} + \lambda_2\|\beta\|_{I_2}$$

et

$$\|\beta\|_{I_1} = \sqrt{\sum_{j=1}^{d_1}\beta_j^2} \text{ et } \|\beta\|_{I_2} = \sqrt{\sum_{j=d_1+1}^{d}\beta_j^2}.$$

1. Montrer que $\mathcal{L}_G$ est de classe $\mathcal{C}^1$ sur l'ensemble

$$D_G = \{\beta \in \mathbb{R}^d, \|\beta\|_{I_1} \neq 0 \text{ et } \|\beta\|_{I_2} \neq 0\}.$$

et calculer son gradient de $\nabla\mathcal{L}_G(\beta)$ en tout point $\beta \in D_G$.

2. Montrer que le problème de minimisation (5) admet une solution et que toute solution $\widehat{\beta}$ vérifie

$$\begin{cases} \widehat{\beta}_{I_1} = \left(1 - \frac{\lambda_1}{2\|X_{I_1}^t R_2\|}\right)_+ \frac{1}{n} X_{I_1}^t R_2 \\ \widehat{\beta}_{I_2} = \left(1 - \frac{\lambda_2}{2\|X_{I_2}^t R_1\|}\right)_+ \frac{1}{n} X_{I_2}^t R_1, \end{cases}$$

avec, pour $k = 1, 2$, $\beta_{I_k} = (\beta_i, i \in I_k)$, $X_{I_k} = (X_{i,j})_{i=1,\dots,n, j \in I_k}$, et $R_j = Y - X_{I_j}\beta_{I_j}$.

3. Proposer un algorithme de descente de gradient coordonnées par coordonnées pour approcher $\widehat{\beta}^{(G)}_{\lambda_1,\lambda_2}$.

Réponse de l'exercice 5.

1. Les fonctions $\beta \mapsto \sum_{i=1}^n (Y_i - x_i^t\beta)^2$ et $\beta \mapsto \|\beta\|_{I_1}$, $\beta \mapsto \|\beta\|_{I_2}$ sont de classe $\mathcal{C}^\infty$ sur $\mathbb{R}^d$. D'autre part la fonction racine est C^1 sur $]0, +\infty[$ donc, comme somme et composée de fonctions de classe $\mathcal{C}^1$, $\mathcal{L}_G$ est de classe $\mathcal{C}^1$ sur D_G.
Soit $\beta \in D_G$, on a, pour $j \in \{1, \dots, d_1\}$,

$$\frac{\partial \mathcal{L}_G}{\partial \beta_j}(\beta) = -2\sum_{i=1}^n x_{i,j}(Y_i - x_i^t\beta) + \lambda_1/2 \frac{2\beta_j}{\sqrt{\sum_{j=1}^{d_1}\beta_j^2}} = -2(X^tY)_j + 2(X^tX\beta)_j + \lambda_1 \frac{\beta_j}{\|\beta\|_{I_1}}.$$

Dans ce cadre là, on a pour $j \in \{d_1 + 1, \dots, d\}$,

$$\frac{\partial \mathcal{L}_G}{\partial \beta_j}(\beta) = -2\sum_{i=1}^n x_{i,j}(Y_i - x_i^t\beta) + \lambda_2/2 \frac{2\beta_j}{\sqrt{\sum_{j=d_1+1}^{d}\beta_j^2}} = -2(X^tY)_j + 2(X^tX\beta)_j + \lambda_2 \frac{\beta_j}{\|\beta\|_{I_2}}.$$

Autrement dit

$$\nabla \mathcal{L}_G(\beta) = \begin{pmatrix} \nabla_1 \mathcal{L}_G(\beta) \\ \nabla_2 \mathcal{L}_G(\beta) \end{pmatrix} = -2X^tY + 2X^tX\beta + \frac{\lambda_1}{\|\beta\|_{I_1}} \begin{pmatrix} I_1 & 0_{d_1,d_2} \\ 0_{d_2,d_1} & 0_{d_2,d_2} \end{pmatrix} \beta + \frac{\lambda_2}{\|\beta\|_{I_2}} \begin{pmatrix} 0_{d_1,d_1} & 0_{d_1,d_2} \\ 0_{d_2,d_1} & I_2 \end{pmatrix} \beta$$

2. Rappelons les notations

$$X = (x_{i,j})_{1 \le i \le n, 1 \le j \le d} = (X_{I_1}, X_{I_2}) \text{ avec } X_{I_1} = (x_{i,j})_{1 \le i \le n, 1 \le j \le d_1}, X_{I_2} = (x_{i,j})_{1 \le i \le n, d_1+1 \le j \le d}$$

et

$$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} = \begin{pmatrix} \beta_{I_1} \\ \beta_{I_2} \end{pmatrix} \text{ avec } \beta_{I_1} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_{d_1} \end{pmatrix}, \beta_{I_2} = \begin{pmatrix} \beta_{d_1+1} \\ \vdots \\ \beta_d \end{pmatrix}.$$

Donc

$$X^tY = \begin{pmatrix} X_{I_1}^tY \\ X_{I_2}^tY \end{pmatrix}, X^tX = \begin{pmatrix} X_{I_1}^t \\ X_{I_2}^t \end{pmatrix} \times (X_{I_1}, X_{I_2}) = \begin{pmatrix} X_{I_1}^tX_{I_1} & X_{I_1}^tX_{I_2} \\ X_{I_2}^tX_{I_1} & X_{I_2}^tX_{I_2} \end{pmatrix}$$

Les conditions du premier ordre donnent

$$\begin{cases} \nabla_1 \mathcal{L}_G(\beta) = 0 \\ \nabla_2 \mathcal{L}_G(\beta) = 0 \end{cases} \Leftrightarrow \begin{cases} -2X_{I_1}^tY + 2\begin{pmatrix} X_{I_1}^tX_{I_1} & X_{I_1}^tX_{I_2} \end{pmatrix}\beta + \frac{\lambda_1}{\|\beta\|_{I_1}}\begin{pmatrix} I_1 & 0_{d_1,d_2} \end{pmatrix}\beta = 0 \\ -2X_{I_2}^tY + 2\begin{pmatrix} X_{I_2}^tX_{I_1} & X_{I_2}^tX_{I_2} \end{pmatrix}\beta + \frac{\lambda_2}{\|\beta\|_{I_2}}\begin{pmatrix} 0_{d_2,d_1} & I_2 \end{pmatrix}\beta = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} -2X_{I_1}^tY + 2X_{I_1}^tX_{I_1}\beta_{I_1} + 2X_{I_1}^tX_{I_2}\beta_{I_2} + \frac{\lambda_1}{\|\beta\|_{I_1}}\beta_{I_1} = 0 \\ -2X_{I_2}^tY + 2X_{I_2}^tX_{I_1}\beta_{I_1} + 2X_{I_2}^tX_{I_2}\beta_{I_2} + \frac{\lambda_2}{\|\beta\|_{I_2}}\beta_{I_2} = 0 \end{cases}$$

On suppose les conditions ORT partielles $\frac{1}{n}X_{I_1}^tX_{I_1} = I_1$ et $\frac{1}{n}X_{I_2}^tX_{I_2} = I_2$.

$$\Leftrightarrow \begin{cases} -2X_{I_1}^tY + 2n\beta_{I_1} + 2X_{I_1}^tX_{I_2}\beta_{I_2} + \frac{\lambda_1}{\|\beta\|_{I_1}}\beta_{I_1} = 0 \\ -2X_{I_2}^tY + 2X_{I_2}^tX_{I_1}\beta_{I_1} + 2n\beta_{I_2} + \frac{\lambda_2}{\|\beta\|_{I_2}}\beta_{I_2} = 0 \end{cases} \Leftrightarrow \begin{cases} \beta_{I_1}(2n + \frac{\lambda_1}{\|\beta\|_{I_1}}) = 2X_{I_1}^tY - 2X_{I_1}^tX_{I_2}\beta_{I_2} \\ \beta_{I_2}(2n + \frac{\lambda_2}{\|\beta\|_{I_2}}) = 2X_{I_2}^tY - 2X_{I_2}^tX_{I_1}\beta_{I_1} \end{cases}$$

$$\Leftrightarrow \begin{cases} \beta_{I_1}(1 + \frac{\lambda_1}{2n\|\beta\|_{I_1}}) = \frac{1}{n}X_{I_1}^t(Y - X_{I_2}\beta_{I_2}) \\ \beta_{I_2}(1 + \frac{\lambda_2}{2n\|\beta\|_{I_2}}) = \frac{1}{n}X_{I_2}^t(Y - X_{I_1}\beta_{I_1}) \end{cases}$$

Par conséquent l'estimation n'existe que sous certaines conditions et se fait de manière cyclique. Calculons les normes pour déterminer les contraintes

$$\begin{cases} \|\beta_{I_1}\|_{I_1} \times |1 + \frac{\lambda_1}{2n\|\beta\|_{I_1}}| = \frac{1}{n}\|X_{I_1}^t(Y - X_{I_2}\beta_{I_2})\| \\ \|\beta_{I_2}\|_{I_2} \times |1 + \frac{\lambda_2}{2n\|\beta\|_{I_2}}| = \frac{1}{n}\|X_{I_2}^t(Y - X_{I_1}\beta_{I_1})\| \end{cases} \Rightarrow \begin{cases} \|\beta_{I_1}\|_{I_1} + \frac{\lambda_1}{2n} = \frac{1}{n}\|X_{I_1}^tR_2\| \\ \|\beta_{I_2}\|_{I_2} + \frac{\lambda_2}{2n} = \frac{1}{n}\|X_{I_2}^tR_1\|. \end{cases}$$

Par conséquent la positivité de la norme oblige à choisir λ_j sous la contrainte

$$\lambda_1/2 \leq \|X_{I_1}^t R_2\|, \lambda_2/2 \leq \|X_{I_2}^t R_1\|.$$

Donc si $\lambda_1 \leq \|X_{I_1}^t R_2\|$ le système a une solution. Sinon lorsque $\lambda_1 \geq \|X_{I_1}^t R_2\|$, la fonction $f_1 : \beta_{I_1} \mapsto \mathcal{L}_G(\beta)$ pour β_{I_2} fixé admet un minimum en tant que fonction continue et concave sur le support D_G. Comme $f_1(\beta_{I_1}) \geq \lambda_1 \|\beta\|_{I_1} \to +\infty$, alors $\beta_{I_1} = 0$.
On peut donc réécrire le système de la façon suivante

$$\begin{cases} \beta_{I_1} = \left(1 - \frac{\lambda_1}{2\|X_{I_1}^t R_2\|}\right)_+ \frac{1}{n} X_{I_1}^t R_2 \\ \beta_{I_2} = \left(1 - \frac{\lambda_2}{2\|X_{I_2}^t R_1\|}\right)_+ \frac{1}{n} X_{I_2}^t R_1 \end{cases}$$

3. Par la dépendance β_{I_1} sur β_{I_2} l'algorithme va réaliser des cycles d'itération
   (a) Initialisation de β_{I_2} à $\beta_{I_2}^{(0)}$
   (b) Itération pour $k+1$ jusqu'à convergence
   - Calcul de $R_2^{(k)} = Y - X_{I_2}^t \beta_{I_2}^{(k)}$.
   - Calcul de $V_1^{(k)} = X_{I_1}^t R_2^{(k)}$.
   - Mise à jour de β_{I_1} par

$$\beta_{I_1}^{(k+1)} = \frac{1}{n}\left(1 - \frac{\lambda_1}{2\|V_1^{(k)}\|}\right)_+ V_1^{(k)}.$$

   - Calcul de $R_1^{(k+1)} = Y - X_{I_1}^t \beta_{I_1}^{(k+1)}$.
   - Calcul de $V_2^{(k+1)} = \frac{1}{n} X_{I_2}^t R_1^{(k+1)}$.
   - Mise à jour de β_{I_2} par

$$\beta_{I_2}^{(k+1)} = \frac{1}{n}\left(1 - \frac{\lambda_2}{2\|V_2^{(k+1)}\|}\right)_+ V_2^{(k+1)}.$$

Exercice 6 (Modèle logit)

Nous nous plaçons dans un premier temps dans un cadre de classification binaire où $Y_i \in \{0,1\}$ dépend de $X_i = (X_{i,1}, ..., X_{i,d})^t \in \mathbb{R}^d$ non aléatoire. Nous supposons que

$$g(\mathbb{E}[Y_i]) = X_i^t \beta^*,$$

avec $g(x) = \log(x/(1-x))$ la fonction de lien *logit*, $\beta^* \in \mathbb{R}^d$ inconnu. Nous souhaitons estimer β^* à partir de l'observation de $\{(X_i, Y_i), i = 1, ..., n\}$. Pour cela, nous considérons deux critères basés sur la log-vraisemblance des données, que nous noterons par la suite $\ell_n(\beta)$. Un critère de type *ridge*

$$\widehat{\beta}_\lambda^{(R)} = \arg\min_{\beta \in \mathbb{R}^d} \left\{-\ell_n(\beta) + \lambda \|\beta\|^2\right\},$$

et un critère de type Lasso,

$$\widehat{\beta}_\lambda^{(L)} = \arg\min_{\beta \in \mathbb{R}^d} \left\{-\ell_n(\beta) + \lambda \|\beta\|_1\right\},$$

dépendant tous deux d'un paramètre $\lambda \geq 0$.

1. Nous commençons par étudier la log-vraisemblance $\ell_n(\beta)$.
   (a) Soit Y une variable aléatoire de loi binomiale de paramètre $\mu \in]0,1[$, calculer la densité de la loi de Y par rapport à la mesure de comptage sur $\{0,1\}$.
   (b) En déduire la vraisemblance de $Y_1, ..., Y_n$ en fonction de $(\mu_1, ..., \mu_n)$ telle que, pour tout i, $\mu_i = \mathbb{E}[Y_i]$.
   (c) Montrer que

$$\ell_n(\beta) = \sum_{i=1}^n \left(Y_i X_i^t \beta - \log\left(1 + e^{X_i^t \beta}\right)\right).$$

   (d) La fonction ℓ_n est-elle convexe ? concave ? dérivable ?
2. Peux-t'on calculer facilement l'estimateur $\widehat{\beta}_\lambda^{(R)}$? Que vaut-il pour $\lambda = 0$?

3. Supposons que nos données sont des données génomiques, c'est-à-dire que pour n femmes ayant ou non un cancer du sein, nous avons prélevé et analysé l'ARN présent dans un échantillon de leurs cellules et nous disposons des données suivantes :
— $Y_i = 1$ si le i-ème individu est atteint d'un cancer du sein ($Y_i = 0$ sinon),
— $X_{i,j}$ le nombre de fois où de l'ARN correspondant au j-ème gène étudié pour l'individu i a été retrouvé.
Nous souhaitons déterminer les gènes ayant de l'influence sur l'apparition d'un cancer du sein. Devons-nous calculer l'estimateur *ridge* ou l'estimateur Lasso ?

Réponse de l'exercice 6.

1. (a) Soit $\phi : \{0,1\} \to \mathbb{R}$ une fonction test, nous avons

$$\begin{aligned}\mathbb{E}[\phi(Y)] &= \phi(0)\mathbb{P}(Y=0) + \phi(1)\mathbb{P}(Y=1) = \phi(0)(1-\mu) + \phi(1)\mu \\ &= \sum_{y\in\{0,1\}} \phi(y)(1-\mu)^{1-y}\mu^y.\end{aligned}$$

La densité de Y par rapport à la mesure de comptage sur $\{0,1\}$ est donc

$$f_Y(y) = (1-\mu)^{1-y}\mu^y.$$

(b)

$$V_{Y_1,\dots,Y_n}(\beta) = \prod_{i=1}^n f_Y(Y_i) = \prod_{i=1}^n (1-\mu_i)^{1-Y_i}\mu_i^{Y_i}$$

(c) Comme $Y_i \in \{0,1\}$, Y_i suit par définition une loi binomiale de paramètre $\mu_i = \mathbb{E}[Y_i] = \mathbb{P}(Y_i = 1)$ donc par (a),

$$\begin{aligned}\ell_n(\beta) &= \log V_{Y_1,\dots,Y_n}(\beta) = \log\left(\prod_{i=1}^n (1-\mu_i)^{1-Y_i}\mu_i^{Y_i}\right) = \sum_{i=1}^n \log\left((1-\mu_i)^{1-Y_i}\mu_i^{Y_i}\right) \\ &= \sum_{i=1}^n \left((1-Y_i)\log(1-\mu_i) + Y_i\log(\mu_i)\right) \\ &= \sum_{i=1}^n \left(\log(1-\mu_i) + Y_i\left(\log(\mu_i) - \log(1-\mu_i)\right)\right) \\ &= \sum_{i=1}^n \left(\log(1-\mu_i) + Y_i\log(\mu_i/(1-\mu_i))\right).\end{aligned}$$

Comme $g(\mu_i) = \log(\mu_i/(1-\mu_i)) = X_i^t\beta$ nous avons :

$$\mu_i = \frac{1}{1+e^{-X_i^t\beta}} \text{ et } 1-\mu_i = 1 - \frac{1}{1+e^{-X_i^t\beta}} = \frac{e^{-X_i^t\beta}}{1+e^{-X_i^t\beta}} = \frac{1}{1+e^{X_i^t\beta}}.$$

Cela nous donne

$$\ell_n(\beta) = \sum_{i=1}^n \left(\log\left(\frac{1}{1+e^{X_i^t\beta}}\right) + Y_i X_i^t\beta\right),$$

d'où le résultat.

(d) La fonction ℓ_n est de classe $\mathcal{C}^\infty$ comme sommes de fonctions de classe C^∞ (nous remarquons bien que $1+e^{X_i^t\beta} \in]0,+\infty[$ quelle que soient les valeurs de β et de X_i). Elle est de plus concave car : $\beta \mapsto Y_i X_i^t\beta$ est linéaire donc convexe et concave et $u : \beta \mapsto -\log(1+e^{X_i^t\beta})$ est concave. En effet, regardons les dérivées secondes de u :

$$\frac{\partial u}{\partial\beta_j} = \frac{\partial}{\partial\beta_j}\left(-\log(1+e^{X_i^t\beta})\right) = -\frac{X_{i,j}e^{X_i^t\beta}}{1+e^{X_i^t\beta}} = -\frac{X_{i,j}}{1+e^{-X_i^t\beta}}.$$

et donc

$$\frac{\partial^2 u}{\partial\beta_j\partial\beta_k} = -\frac{\partial^2}{\partial\beta_j\partial\beta_k}\left(\frac{X_{i,j}}{1+e^{-X_i^t\beta}}\right) = -\frac{X_{i,j}X_{i,k}}{(1+e^{-X_i^t\beta})^2}.$$

La matrice Hessienne de u est donc :

$$H(u) = -X_i^t X_i/(1+e^{-X_i^t\beta})^2,$$

qui est une matrice symétrique négative (car $X_i^t X_i$ est une matrice symétrique positive).

2. $\widehat{\beta}_\lambda^{(R)}$ est solution du problème de minimisation de

$$f_\lambda(\beta) = -\ell_n(\beta) + \lambda\|\beta\|^2.$$

La fonction f_λ est une fonction convexe et dérivable, nous calculons donc ses dérivées partielles :

$$\frac{\partial f_\lambda}{\partial \beta_j} = -\frac{\partial \ell_n}{\partial \beta_j}(\beta) + 2\lambda\beta_j.$$

Par 1.(c), nous avons

$$\begin{aligned}\frac{\partial f_\lambda}{\partial \beta_j} &= -\sum_{i=1}^{n}\left(Y_i X_{i,j}\beta_j - \frac{X_{i,j}}{1+e^{-X_i^t\beta}}\right) + 2\lambda\beta_j \\ &= \left(-\sum_{i=1}^{n} Y_i X_{i,j} + 2\lambda\right)\beta_j + \sum_{i=1}^{n}\frac{X_{i,j}}{1+e^{-X_i^t\beta}}\end{aligned}$$

et trouver β tel que $\frac{\partial f_\lambda}{\partial \beta_j} = 0$ pour tout j n'est pas évident ! Le cas $\lambda = 0$ correspond à l'estimateur du maximum de vraisemblance.

3. En observant les coefficients non nuls de l'estimateur Lasso, $\widehat{\beta}_\lambda^{(L)}$, nous pouvons aider à déterminer les gènes ayant une influence dans l'apparition d'un cancer du sein.