# La régression Lasso


Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau provisoire : 07 au BSHM
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

# Introduction

- Nous avons introduit la régression pénalisée et discuté de la régression ridge, dans laquelle la pénalité prend la forme d'une somme de carrés des coefficients de régression.
- Dans ce sujet, nous allons plutôt pénaliser les valeurs absolues des coefficients de régression, un changement apparemment simple mais aux conséquences étendues.

# Régression Lasso

- Plus précisément, considérons la fonction objective :

$$Q(\beta|X, y) = \frac{1}{2n}||y - X\beta||_2^2 + \lambda\, ||\beta||_1$$

où $||\beta||_1 = \sum_j |\beta_j|$ désigne la norme $L^1$ des coefficients de régression

- Comme précédemment, les estimations de $\beta$ sont obtenues en minimisant la fonction ci-dessus pour une valeur donnée de $\lambda$, ce qui donne $\hat{\beta}(\lambda)$
- Cette approche est une méthode de rétrécissement/contraction des coefficients de la régression développée par Robert Tibshirani (1996) [1], qui l'a appelé : *Least Absolute Shrinkage and Selection Operator (LASSO)*

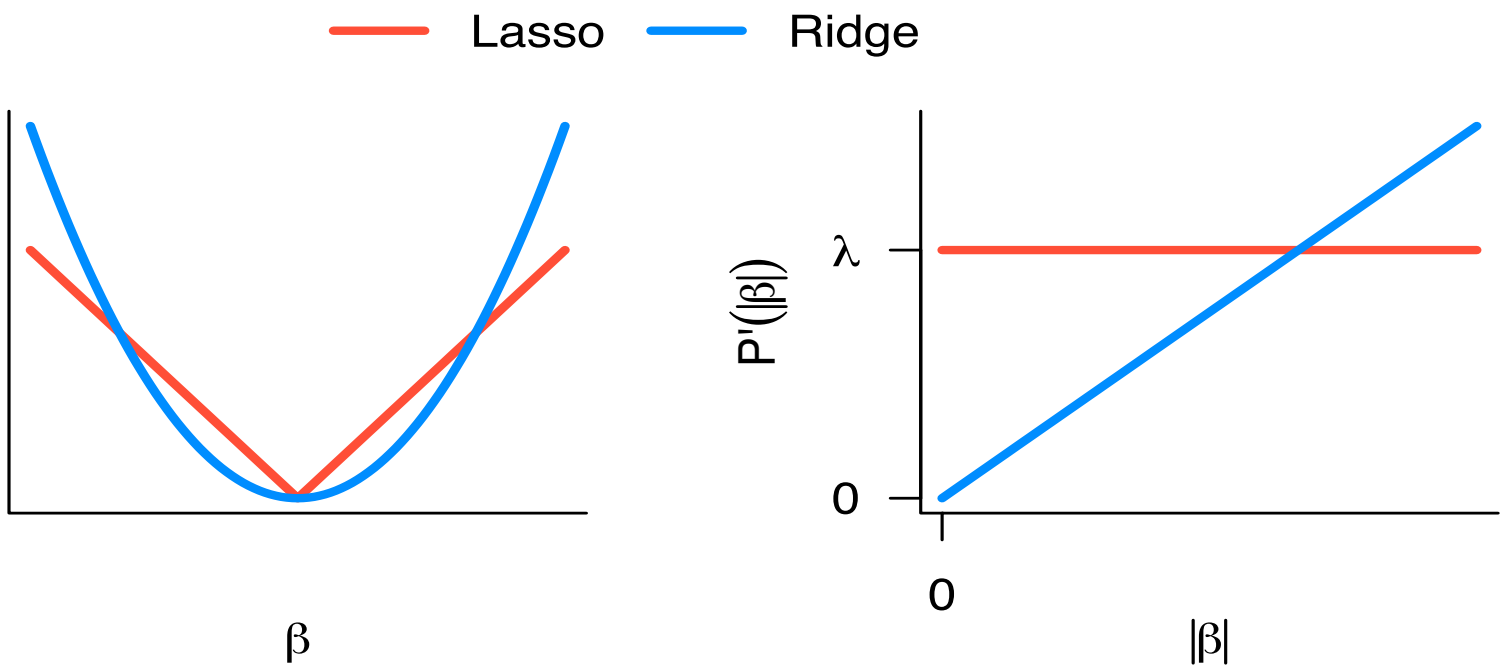
1. Robert Tibshirani (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, **58**(1), 267–288

# Rétrécissement, Sélection et Parcimonie (sparsity)

- Son nom capture l'essence de ce que la peine de lasso accomplit :
  - *Rétrécissement* : comme la régression ridge, le lasso pénalise les grands coefficients de régression et réduit les estimations à zéro
  - *Sélection* : contrairement à la régression ridge, le lasso produit des solutions parcimonieuses : certaines estimations de coefficients sont exactement nulles, ce qui élimine efficacement ces prédicteurs/co-variables du modèle.

  La parcimonie/sparsity a deux propriétés très attrayantes :
  - *Vitesse* : les algorithmes qui tirent parti de la parcimonie peuvent évoluer très efficacement, offrant des avantages informatiques considérables
  - *Interprétabilité* : dans les modèles comportant des centaines, voire des milliers, de prédicteurs, la fragmentation offre une simplification utile du modèle en nous permettant de nous concentrer uniquement sur les prédicteurs avec des estimations de coefficients non nuls.

# Pénalités Ridge et Lasso



# Foctions semi-différentiable

- Un défi évident qui est induite avec le lasso est qu'en introduisant des valeurs absolues, nous ne traitons plus de fonctions différentiables.
- Pour cette raison, nous allons rappeler et étendre quelques résultats de calculs élémentaires au cas de fonctions non différentiables (plus précisément, semi-différentiables)
-

**Definition**

Une fonction $f : \mathbb{R} \longrightarrow \mathbb{R}$ est dite semi-différentiable en un point $x$ si $d_- f(x)$ et $d_+ f(x)$ existent dans $\mathbb{R}$

où $d_- f(x)$ et $d_+ f(x)$ désignent les dérivées à gauche et à droite de $f$ en $x$

- Notons que si une fonction $f$ est semi-différentiable alors elle (i.e., $f$) est continue

# Sous-dérivées et sous-différentielles

## Definition

Soit une fonction semi-différentiable $f : \mathbb{R} \longrightarrow \mathbb{R}$. On dit que $d$ est une sous-dérivée de $f$ en $x$ si $d \in [d_- f(x), d_+ f(x)]$.

L'ensemble $[d_- f(x), d_+ f(x)]$ est appelé la sous-différentielle de $f$ en $x$, et est noté $\partial f(x)$

- Notons que la sous-différentielle est l'ensemble de valeurs d'une fonction : elle peut s'agir d'une valeur unique (si $f$ est différentiable), d'un intervalle de valeurs ou vide (si $d_- f(x) > d_+ f(x)$)
  - Les sous-dérivées sont utiles pour les problèmes de minimisation. Si nous voulons maximiser une fonction, nous nous intéresserions à l'idée miroir des "superdifférentielles"
  - Il s'agit d'une définition un peu plus souple de sous-différentielle que celle utilisée dans la littérature sur l'optimisation convexe, mais nous en avons besoin pour envisager des pénalités "non convexes" plus tard dans le cours.
- Rappelons qu'une fonction est différentiable en $x$ si $df(x) = d_+ f(x)$ i.e., si la sous-différentielle est constitué d'un seul point (singleton)

# Exemple : $|x|$

Dans la majeure partie de ce cours, vous n'avez pas vraiment besoin de connaître les sous-différentielles et les sous-gradients (la version multidimensionnelle des sous-différentielles), mais vous devez connaître la sous-différentielle pour $f(x) = |x|$. La sous-différentielle de $f$ est :

$$\partial f(x) = \begin{cases} -1 & \text{si } x < 0 \\ [-1, 1] & \text{si } x = 0 \\ 1 & \text{si } x > 0 \end{cases}$$

# Optimisation convexe

- Les résultats essentiels de l'optimisation peuvent être étendus à des fonctions semi-différentiables
-

**Théorème**

*Si $f$ est une fonction semi-différentiable et que $x_0$ est un minimum ou un maximum local de $f$, alors* $0 \in \partial f(x_0)$

- Comme avec le calcul normal/classique/habituel, la réciproque n'est pas vraie en général

# Règles de calcul

- Comme pour la différentiabilité habituelle, les règles de base suivantes s'appliquent
-

**Theorem**

*Soit $f$ une fonction semi-différentiable, $a$ et $b$ deux constantes et $g$ une fonction différentiable. Alors :*

$$\begin{aligned} \partial(af(x)+b) &= a\,\partial f(x) \\ \partial(f(x)+g(x)) &= \partial f(x) + g'(x) \end{aligned}$$

- Les notions s'étendent également aux dérivées d'ordres supérieurs. Une fonction $f : \mathbb{R} \longrightarrow \mathbb{R}$ est dite semi-différentiable de second ordre en $x$ si $d^2_- f(x)$ et $d^2_+ f(x)$ existent dans $\mathbb{R}$
- La sous-différentièlle de second ordre est notée :

$$\partial^2 f(x) = [d^2_- f(x), d^2_+ f(x)]$$

# Convexité

- Comme dans le cas de la différentiabilité habituelle, une fonction convexe peut être caractérisée en termes de sa sous-differentielle
-

**Théorème**

*Supposons que $f$ soit semi-différentiable sur $(a, b)$. Alors $f$ est convexe sur $(a, b)$ si, et seulement si, $f$ est croissante sur $(a, b)$.*

-

**Théorème**

*Supposons que $f$ soit semi-différentiable de second ordre sur $(a, b)$. Alors $f$ est convexe sur $(a, b)$ si, et seulement si,*

$$\partial^2 f(x) \geq 0, \ \forall x \in (a, b).$$

# Résultats multidimensionnels

- Les résultats précédents peuvent être étendus (même si nous allons passer sous silence les détails) aux fonctions multidimensionnelles en remplaçant les dérivées à gauche et à droite par des dérivées directionnelles
- Une fonction $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ est dite sous-différentiable si la dérivée partielle/directionnelle $d_u f(x)$ existe dans toutes les directions $u$
-

**Théorème**

*Si $f$ est une fonction semi-différentiable et que $x_0$ est un minimum local de $f$, alors $d_u f(x_0) \geq 0, \forall u$*

-

**Théorème**

*Supposons que $f$ soit une fonction sous-différentiable. Alors $f$ est convexe sur l'ensemble $S$ si, et seulement si, $d_u^2 f(x) \geq 0, \forall x \in S$ et dans toutes les directions $u$*

# Fonctions de score et Fonctions de score pénalisées

- Dans la théorie statistique classique, la dérivée de la fonction log-vraisemblance est appelée fonction de score et les estimateurs du maximum de vraisemblance sont obtenus en annulant cette dérivée, donnant ainsi les *équations de vraisemblance* (ou *équations de score*) :

$$0 = \frac{\partial L}{\partial \theta}(\theta),$$

où $L$ désigne la log-vraisemblance.

- Étendre cette idée aux vraisemblances pénalisées implique de prendre les dérivées de la fonction objective qui est de la forme :

$$Q(\theta) = L(\theta) + P(\theta),$$

donnant la *fonction de score pénalisée*

# Équations de vraisemblance pénalisées

- Pour la régression ridge, la vraisemblance pénalisée est partout différentiable et l'extension aux équations du score pénalisé est simple/directe.
- Pour le lasso et pour les autres pénalités que nous allons considérer dans ce cours, la vraisemblance pénalisée n'est pas différentiable - en particulier, non différentiable en zéro - et des sous-différentielles sont nécessaires pour les caractériser.
- Soit $\partial Q(\theta)$ la sous-différentièlle de $Q$, les *équations de vraisemblance pénalisées* (ou *les équations de score pénalisées*) sont les suivantes :

$$0 \in \partial Q(\theta)$$

# Conditions KKT

- Dans la littérature sur l'optimisation, les équations résultantes sont appelées les conditions de Karush-Kuhn-Tucker (KKT)
- Pour les problèmes d'optimisation convexe tels que le lasso, les conditions de KKT sont à la fois nécessaires et suffisantes pour caractériser la solution.
- Une preuve rigoureuse de cette affirmation en grande dimension impliquerait certains détails que nous avons passé sous silence, mais l'idée est assez simple : pour résoudre le problème pour $\widehat{\beta}$, nous remplaçons simplement la dérivée par le sous-dérivée et la vraisemblance par la vraisemblance pénalisée

# Conditions KKT pour le lasso

- *Résultat* : $\widehat{\beta}$ minimise la fonction objective du lasso si, et seulement si, elle satisfait les conditions KKT :

$$\begin{aligned} \frac{1}{n} x_j^T (y - X\widehat{\beta}) &= \lambda \, sign(\widehat{\beta}_j) && \text{si } \widehat{\beta}_j \neq 0 \\ \frac{1}{n} |x_j^T (y - X\widehat{\beta})| &\leq \lambda && \text{si } \widehat{\beta}_j = 0 \end{aligned}$$

- En d'autres termes, la corrélation entre un prédicteur et les résidus, $x_j^T (y - X\widehat{\beta})/n$, doit dépasser un certain seuil minimal $\lambda$ avant de pouvoir être incluse dans le modèle.
- Lorsque cette corrélation est inférieure à $\lambda$, alors $\widehat{\beta}_j = 0$

# Remarques

- Si on pose :

$$\lambda = \lambda_{\max} \equiv \max_{1 \leq j \leq p} |{}^t x_j y|/n$$

  alors $\widehat{\beta} = 0$ satisfait les conditions de KKT.

- C'est, pour tout $\lambda \geq \lambda_{\max}$, nous avons $\widehat{\beta}(\lambda) = 0$
- D'autre part, si nous fixons $\lambda = 0$, les conditions KKT sont simplement les équations normales pour MCO,

$$^t X(y - X\widehat{\beta}) = 0$$

- Ainsi, le chemin des coefficients pour le lasso commence à $\lambda_{\max}$ et continue jusqu'à $\lambda = 0$ si $X$ est de rang plein. Sinon, la solution ne sera pas unique pour les valeurs $\lambda$ inférieures à un point $\lambda_{\min}$

# Lasso et unicité

- Notons que la fonction objective du lasso est convexe, mais elle n'est pas strictement convexe si ${}^tX\,X$ n'est pas de rang plein
- Par exemple, supposons que $n = 2$ et $p = 2$, avec $(y_1, x_{11}, x_{12}) = (1, 1, 1)$ et $(y_2, x_{21}, x_{22}) = (1, 1, 1)$
- Ensuite, les solutions sont :

$$(\widehat{\beta}_1, \widehat{\beta}_2) = \begin{cases} (0,0) & \text{si } \lambda \geq 1, \\ \in \{(\beta_1, \beta_2) \ : \ \beta_1 + \beta_2 = 1 - \lambda, \beta_1 \geq 0, \beta_2 \geq 0\} & \text{si } 0 \leq \lambda < 1 \end{cases}$$

# Cas particulier : Design/matrice de conception orthogonale

- Comme pour la régression ridge, il est instructif de considérer le cas particulier où la matrice de conception $X$ est orthonormal : $n^{-1}{}^tX\,X = I$
- *Résultat* : dans le cas orthonormal, l'estimation par lasso est :

$$\widehat{\beta}_j(\lambda) = \begin{cases} z_j - \lambda & \text{si } z_j > \lambda, \\ 0 & \text{si } |z_j| \leq \lambda, \\ z_j + \lambda & \text{si } z_j < -\lambda, \end{cases}$$

où $z_j = {}^t x_j y/n$ est la solution MCO

# Seuillage souple

- Le résultat de la diapositive précédente peut être écrit de manière plus compacte :

$$\widehat{\beta}_j(\lambda) = S(z_j|\lambda)$$

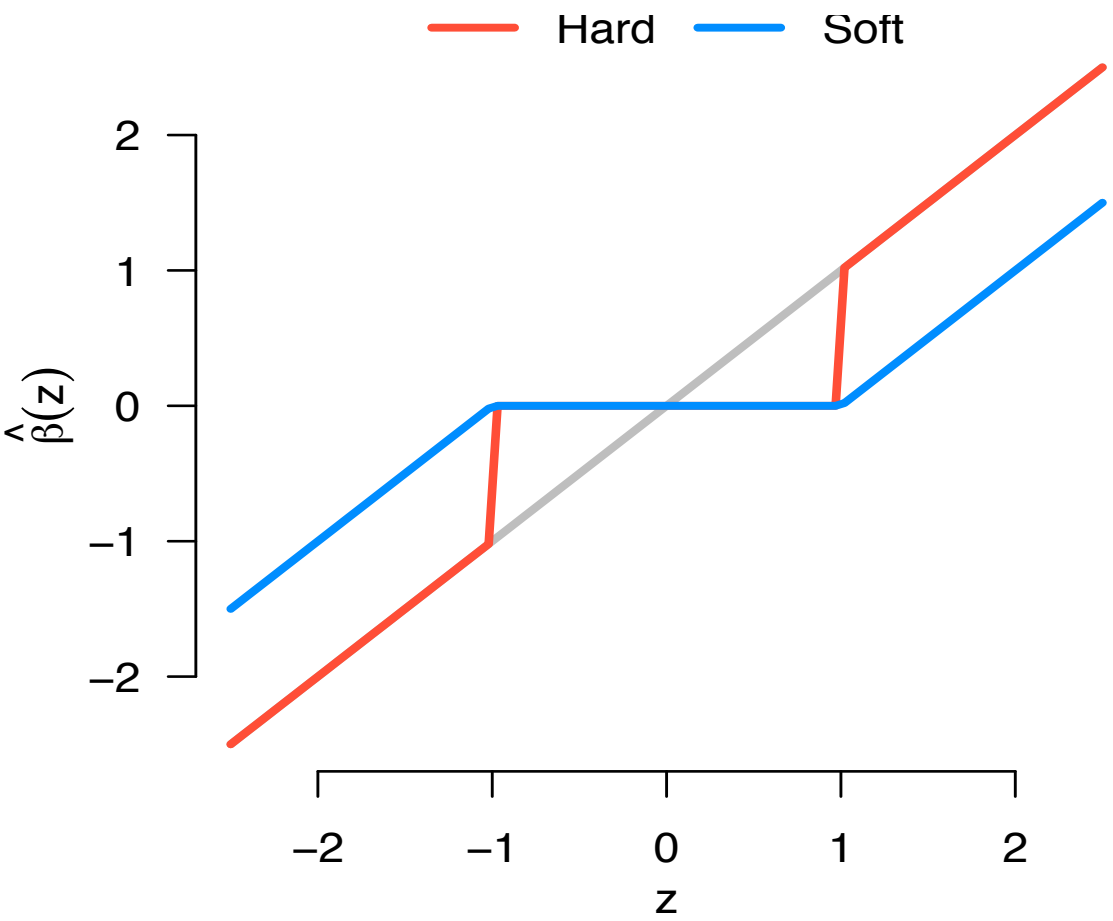
où la fonction $S(.|\lambda)$ est connu sous le nom de l'opérateur de seuillage souple

- Cela avait été proposé à l'origine par Donoho et Johnstone (1994) pour le seuillage souple/doux des coefficients d'ondelettes dans le contexte de la régression non paramétrique.
- En comparaison, l'opérateur de seuillage "rigide" est

$$H(z, \lambda) = z\ I_{\{|z|>\lambda\}},$$

où $I_{(S)}$ est la fonction indicatrice de l'ensemble $S$

# Opérateurs de seuillage : souple (soft) & rigide (hard)



# Probabilité pour que $\widehat{\beta}_j = 0$

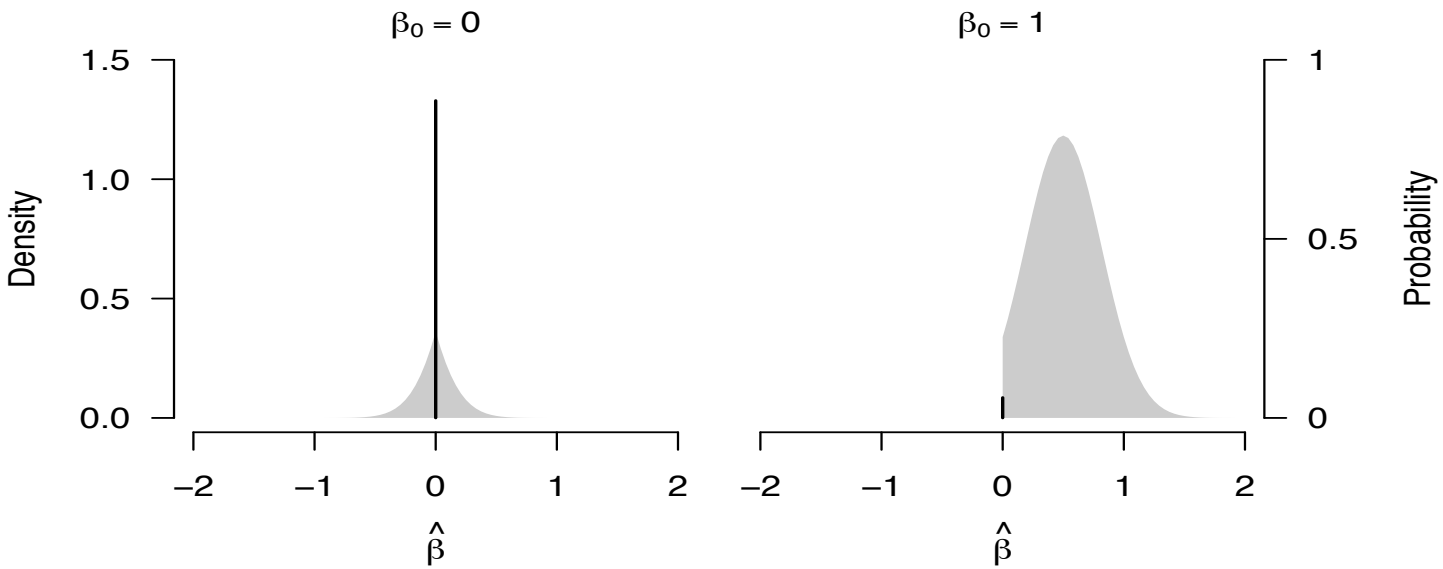
- Avec le seuillage souple, il est clair que le lasso a une probabilité positive de produire une estimation d'exactement 0 - en d'autres termes, de produire une solution sparse/parcimonieuse.
- Plus précisément, la probabilité de supprimer $x_j$ du modèle est $\mathbb{P}(|z_j| \leq \lambda)$
- En supposant que les erreurs $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ sont indépendantes, nous avons

$$z_j \sim \mathcal{N}(\beta, \sigma^2/n) \quad \text{et} \quad \mathbb{P}(\widehat{\beta}_j(\lambda) = 0) = \Phi\left(\frac{\lambda - \beta}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-\lambda - \beta}{\sigma/\sqrt{n}}\right),$$

où $\Phi$ est la CDF de la loi gaussienne standard

# Distribution d'échantillonnage

$\sigma = 1$, $n = 10$ et $\lambda = 1/2$



# Remarques

- Cette distribution d'échantillonnage est très différente de celle d'un MLE classique :
  - La distribution est mixte : une partie de la distribution est continue, mais l'autre a aussi une masse ponctuelle (atome) en zéro
  - La partie continue n'est pas distribuée normalement
  - La distribution est asymétrique (sauf si $\beta = 0$)
  - La distribution n'est pas centrée en la vraie valeur de $\beta$
- Ces faits créent un certain nombre de difficultés pour l'inférence statistique à l'aide du lasso. Nous allons mettre cette question de côté pour le moment, mais nous y reviendrons plus tard