

# Lasso et sélection de modèles

**Nicolas Verzelen, Joseph Salmon**

INRA / Université de Montpellier



# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

## Retour sur le modèle

$$\mathbf{y} = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon} \in \mathbb{R}^n \quad (\text{signal observé})$$

$$X = [\mathbf{x}_1, \dots, \mathbf{x}_p] \in \mathbb{R}^{n \times p} \quad (\text{variables explicatives})$$

$$\boldsymbol{\beta}^* \in \mathbb{R}^p \quad (\text{signal/coefficients})$$

Rem: les vecteurs  $\mathbf{x}_1, \dots, \mathbf{x}_p$  sont les colonnes de la matrice

$$X = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix}$$

# Sommaire

## Rappels

### Sélection de variables et parcimonie

- La pénalisation  $\ell_0$  et ses limites

- La pénalisation  $\ell_1$

- Sous-gradient / sous-différentielle

## Améliorations et extensions du Lasso

# Motivation




Utilité des estimateurs  $\hat{\beta}$  avec beaucoup de coefficients nuls :

- ▶ pour l'interprétation
- ▶ pour l'efficacité computationnelle si  $p$  est énorme




Idée sous-jacente : **sélectionner des variables**

Rem: aussi utile si  $\beta^*$  a peu de coefficients non nuls




# Méthodes de sélection de variables

- ▶ Méthodes de **dépistage par corrélation** ( : *correlation screening*) : supprimer les  $x_j$  de faible corrélation avec  $y$ 
  - avantages : rapide (+++), coût :  $p$  produits scalaires de taille  $n$ , intuitive (+++)
  - défauts : néglige les interactions entre variables  $x_j$ , résultats théoriques faibles (- - -)
- ▶ Méthodes **gloutonnes** ( : *greedy*) / **pas à pas** ( : *stage/step-wise*)
  - avantages : rapide (++), coût :  $p$  produits scalaires de taille  $n$  par variable active, intuitive (++)
  - défauts : propagation de mauvaises sélections de variables aux étapes suivantes ; résultats théoriques faibles (-)

# Méthodes de sélection de variables

- ▶ Méthodes de **dépistage par corrélation** ( : *correlation screening*) : supprimer les  $x_j$  de faible corrélation avec  $y$ 
  - avantages : rapide (+++), coût :  $p$  produits scalaires de taille  $n$ , intuitive (+++)
  - défauts : néglige les interactions entre variables  $x_j$ , résultats théoriques faibles (- - -)
- ▶ Méthodes **gloutonnes** ( : *greedy*) / **pas à pas** ( : *stage/step-wise*)
  - avantages : rapide (+++), coût :  $p$  produits scalaires de taille  $n$  par variable active, intuitive (++)
  - défauts : propagation de mauvaises sélections de variables aux étapes suivantes ; résultats théoriques faibles (-)
- ▶ Méthodes **pénalisées** favorisant la parcimonie (e.g., Lasso)
  - avantages : résultats théoriques bons (++)
  - défauts : encore lent (on y travaille *Fercoq et al. (2015)*) (-)

# Méthodes de sélection de variables

- ▶ Méthodes de **dépistage par corrélation** ( : *correlation screening*) : supprimer les  $x_j$  de faible corrélation avec  $y$ 
  - avantages : rapide (+++), coût :  $p$  produits scalaires de taille  $n$ , intuitive (+++)
  - défauts : néglige les interactions entre variables  $x_j$ , résultats théoriques faibles (- - -)
- ▶ Méthodes **gloutonnes** ( : *greedy*) / **pas à pas** ( : *stage/step-wise*)
  - avantages : rapide (++), coût :  $p$  produits scalaires de taille  $n$  par variable active, intuitive (++)
  - défauts : propagation de mauvaises sélections de variables aux étapes suivantes ; résultats théoriques faibles (-)
- ▶ Méthodes **pénalisées** favorisant la parcimonie (e.g., Lasso)
  - avantages : résultats théoriques bons (++)
  - défauts : encore lent (on y travaille [Fercoq et al. \(2015\)](#)) (-)



# La pseudo-norme $\ell_0$

---

## Définitions

---

Le **support** du vecteur  $\beta$  est l'ensemble des indices des coordonnées non nulles :

$$\text{supp}(\beta) = \{j \in \llbracket 1, p \rrbracket, \beta_j \neq 0\}$$

La **pseudo-norme**  $\ell_0$  d'un vecteur  $\beta \in \mathbb{R}^p$  est son nombre de coordonnées non-nulles :

$$\|\beta\|_0 = \text{card}\{j \in \llbracket 1, p \rrbracket, \beta_j \neq 0\}$$

---

Rem:  $\|\cdot\|_0$  n'est pas une norme,  $\forall t \in \mathbb{R}^*, \|t\beta\|_0 = \|\beta\|_0$

Rem:  $\|\cdot\|_0$  n'est pas non plus convexe,  $\beta_1 = (1, 0, 1, 0, \dots, 0)$   
 $\beta_2 = (0, 1, 1, 0, \dots, 0)$  et  $3 = \|\frac{\beta_1 + \beta_2}{2}\|_0 \geq \frac{\|\beta_1\|_0 + \|\beta_2\|_0}{2} = 2$

# Sommaire

## Rappels

### Sélection de variables et parcimonie

La pénalisation  $\ell_0$  et ses limites

La pénalisation  $\ell_1$

Sous-gradient / sous-différentielle

### Améliorations et extensions du Lasso

# La pénalisation $\ell_0$

Première tentative de méthode pénalisée pour introduire de la parcimonie : utiliser  $\ell_0$  pour la pénalisation / régularisation

$$\hat{\beta}_\lambda \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda \|\beta\|_0}_{\text{régularisation}} \right)$$

**Problème combinatoire** !!! (problème “NP-dur”)

Résolution exacte : nécessite de considérer tous les sous-modèles, *i.e.*, calculer les estimateurs pour tous les supports possibles ; il y en a  $2^p$ , ce qui requiert le calcul de  $2^p$  moindres carrés !

Exemples :

$p = 10$  possible :  $\approx 10^3$  moindres carrés

$p = 30$  impossible :  $\approx 10^{10}$  moindres carrés

Rem: avancées récentes en MIP *Bertsimas et al. 16*

# Sommaire

## Rappels

### Sélection de variables et parcimonie

La pénalisation  $\ell_0$  et ses limites

La pénalisation  $\ell_1$

Sous-gradient / sous-différentielle

### Améliorations et extensions du Lasso

# Le Lasso : la définition pénalisée

Lasso : *Least Absolute Shrinkage and Selection Operator*

Tibshirani (1996)

$$\hat{\beta}_{\lambda}^{\text{Lasso}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda \|\beta\|_1}_{\text{régularisation}} \right)$$

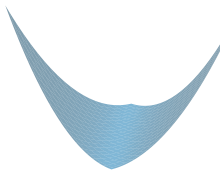
où  $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$  (somme des valeurs absolues des coefficients)

► On retrouve de nouveau les cas limites :

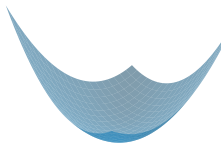
$$\begin{aligned} \lim_{\lambda \rightarrow 0} \hat{\beta}_{\lambda}^{\text{Lasso}} &= \hat{\beta}^{\text{MCO}} \\ \lim_{\lambda \rightarrow +\infty} \hat{\beta}_{\lambda}^{\text{Lasso}} &= 0 \in \mathbb{R}^p \end{aligned}$$

**Attention** : l'estimateur Lasso n'est pas toujours **unique** pour un  $\lambda$  fixé ; prendre par exemple deux colonnes identiques

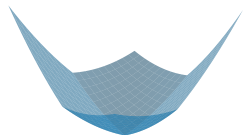
# Moindre carrés / Ridge et Lasso



OLS

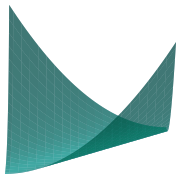


Ridge

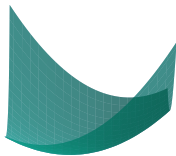


Lasso

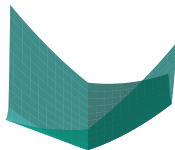
# Moindre carrés / Ridge et Lasso



OLS

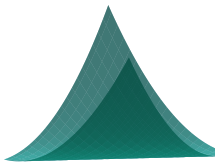


Ridge

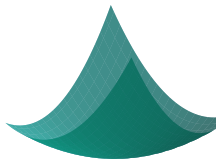


Lasso

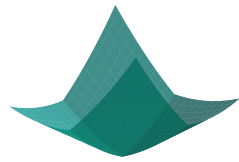
# Moindre carrés / Ridge et Lasso



OLS



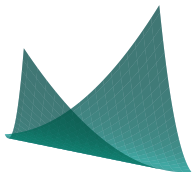
Ridge



Lasso



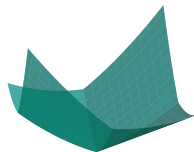
# Moindre carrés / Ridge et Lasso



OLS

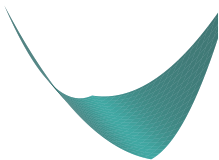


Ridge

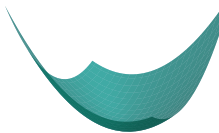


Lasso

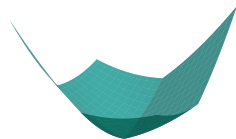
# Moindre carrés / Ridge et Lasso



OLS



Ridge



Lasso

# Interprétation contrainte

Un problème de la forme :

$$\hat{\beta}_{\lambda}^{\text{Lasso}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda \|\beta\|_1}_{\text{régularisation}} \right)$$

admet la même solution qu'une version contrainte :

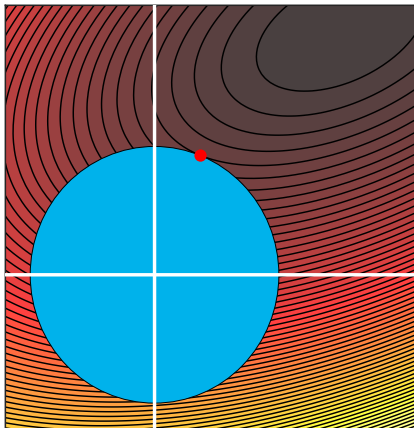
$$\begin{cases} \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - X\beta\|_2^2 \\ \text{t.q. } \|\beta\|_1 \leq T \end{cases}$$

pour un certain  $T > 0$ .

Rem: le lien  $T \leftrightarrow \lambda$  n'est pas explicite

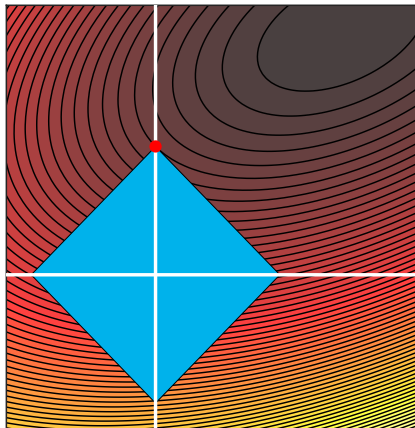
- ▶ Si  $T \rightarrow 0$  on retrouve comme solution le vecteur nul :  $0 \in \mathbb{R}^p$
- ▶ Si  $T \rightarrow \infty$  on retrouve  $\hat{\beta}^{\text{MCO}}$  (non contraint)

## Mise à zéro de certains coefficients



Optimisation sous contrainte  $\ell_2$  : solution non parcimonieuse

## Mise à zéro de certains coefficients



Optimisation sous contrainte  $\ell_1$  : solution parcimonieuse

## Version animée de cette illustration

<https://twitter.com/PierreAblin/status/1107625298936451073>

Crédit : Pierre Ablin

# Sommaire

## Rappels

### Sélection de variables et parcimonie

La pénalisation  $\ell_0$  et ses limites

La pénalisation  $\ell_1$

Sous-gradient / sous-différentielle

### Améliorations et extensions du Lasso

# Sous-gradients / sous-différentielles

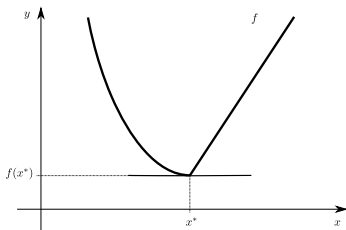
## Définitions

Pour  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction convexe,  $u \in \mathbb{R}^n$  est un **sous-gradient** de  $f$  en  $x^*$ , si pour tout  $x \in \mathbb{R}^n$  on a

$$f(x) \geq f(x^*) + \langle u, x - x^* \rangle$$

La **sous-différentielle** est l'ensemble des sous-gradients :  
 $\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle u, x - x^* \rangle\}.$

Rem: si le sous-gradient est unique, on retrouve le gradient





# Sous-gradients / sous-différentielles

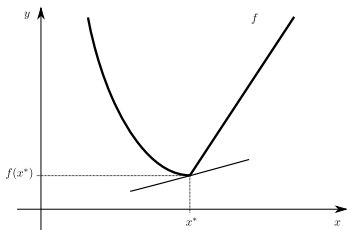
## Définitions

Pour  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction convexe,  $u \in \mathbb{R}^n$  est un **sous-gradient** de  $f$  en  $x^*$ , si pour tout  $x \in \mathbb{R}^n$  on a

$$f(x) \geq f(x^*) + \langle u, x - x^* \rangle$$

La **sous-différentielle** est l'ensemble des sous-gradients :  
 $\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle u, x - x^* \rangle\}.$

Rem: si le sous-gradient est unique, on retrouve le gradient



# Sous-gradients / sous-différentielles

---

## Définitions

---

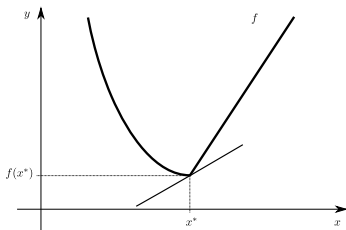
Pour  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction convexe,  $u \in \mathbb{R}^n$  est un **sous-gradient** de  $f$  en  $x^*$ , si pour tout  $x \in \mathbb{R}^n$  on a

$$f(x) \geq f(x^*) + \langle u, x - x^* \rangle$$

La **sous-différentielle** est l'ensemble des sous-gradients :  
 $\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle u, x - x^* \rangle\}.$

---

Rem: si le sous-gradient est unique, on retrouve le gradient



# Sous-gradients / sous-différentielles

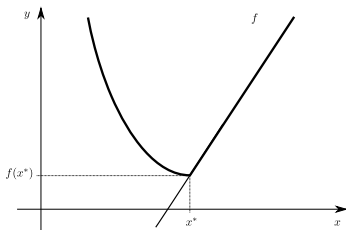
## Définitions

Pour  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction convexe,  $u \in \mathbb{R}^n$  est un **sous-gradient** de  $f$  en  $x^*$ , si pour tout  $x \in \mathbb{R}^n$  on a

$$f(x) \geq f(x^*) + \langle u, x - x^* \rangle$$

La **sous-différentielle** est l'ensemble des sous-gradients :  
 $\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle u, x - x^* \rangle\}.$

Rem: si le sous-gradient est unique, on retrouve le gradient



# Règle de Fermat

---

---

## Théorème

---

---

Un point  $x^*$  est un minimum d'une fonction convexe  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  si et seulement si  $0 \in \partial f(x^*)$

---

---

Preuve : utiliser la définition des sous-gradients :

- 0 est un sous-gradient de  $f$  en  $x^*$  si et seulement si  $\forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$

# Règle de Fermat

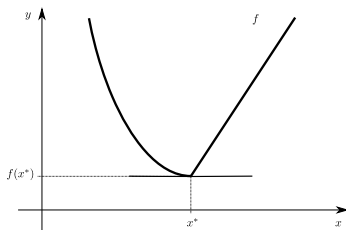
## Théorème

Un point  $x^*$  est un minimum d'une fonction convexe  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  si et seulement si  $0 \in \partial f(x^*)$

Preuve : utiliser la définition des sous-gradients :

- 0 est un sous-gradient de  $f$  en  $x^*$  si et seulement si  
 $\forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle 0, x - x^* \rangle$

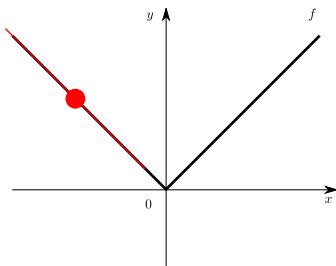
Rem:visuellement cela correspond à une tangente horizontale



# Sous-différentielle de la valeur absolue

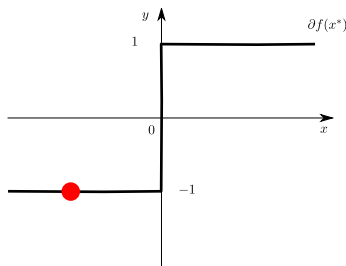
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

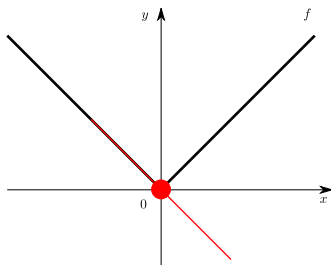
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Sous-différentielle de la valeur absolue

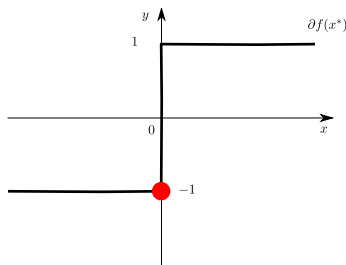
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

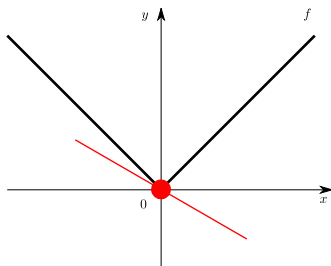
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Sous-différentielle de la valeur absolue

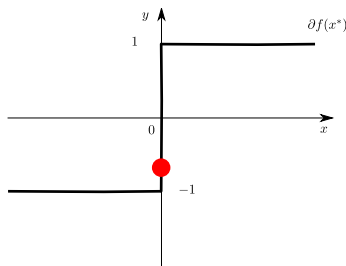
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$

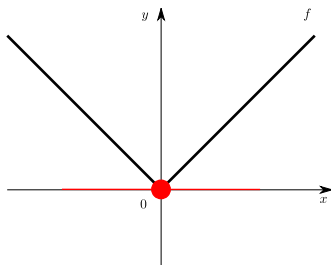




# Sous-différentielle de la valeur absolue

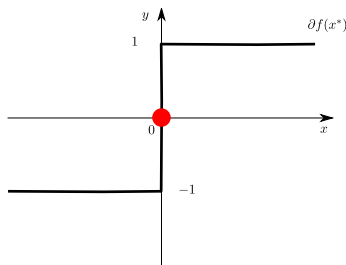
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

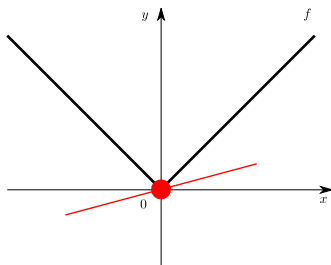
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Sous-différentielle de la valeur absolue

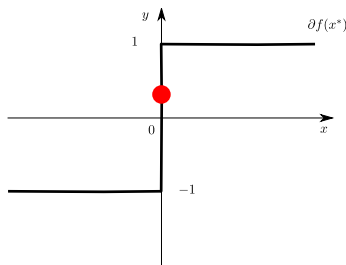
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

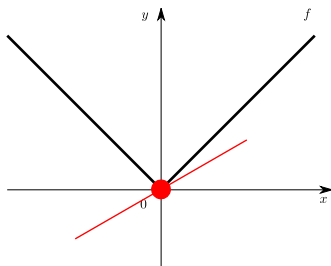
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Sous-différentielle de la valeur absolue

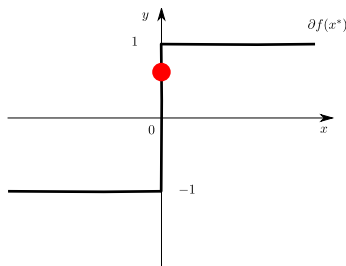
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

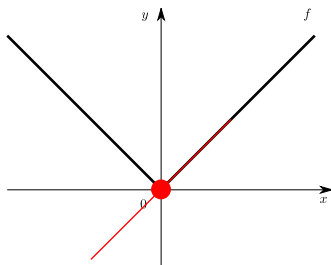
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Sous-différentielle de la valeur absolue

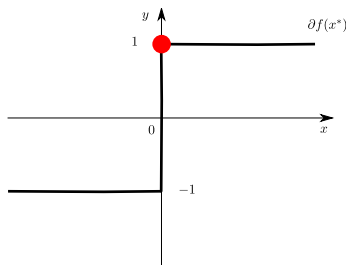
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

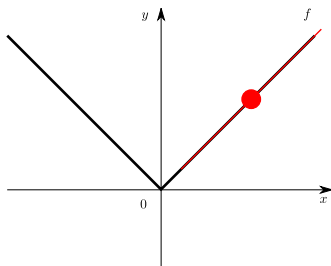
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Sous-différentielle de la valeur absolue

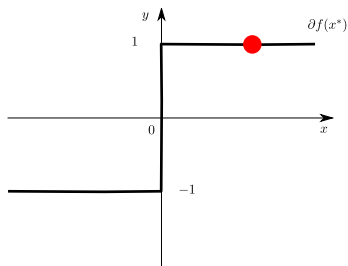
Fonction (abs) :

$$f : \begin{cases} \mathbb{R} & \rightarrow \mathbb{R} \\ x & \mapsto |x| \end{cases}$$



Sous-différentielle (sign)

$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in ]-\infty, 0[ \\ \{1\} & \text{si } x^* \in ]0, \infty[ \\ [-1, 1] & \text{si } x^* = 0 \end{cases}$$



# Condition de Fermat pour le Lasso

$$\hat{\beta}_{\lambda}^{\text{Lasso}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda \|\beta\|_1}_{\text{régularisation}} \right)$$

Rem: on suppose dorénavant la matrice  $X$  normalisée ;  $\|\mathbf{x}_j\| = 1$

Conditions nécessaires et suffisantes d'optimalité (Fermat) :

$$\forall j \in [p], \mathbf{x}_j^{\top} \left( \frac{y - X\hat{\beta}_{\lambda}^{\text{Lasso}}}{\lambda} \right) \in \begin{cases} \{\text{sign}(\hat{\beta}_{\lambda}^{\text{Lasso}})_j\} & \text{si } (\hat{\beta}_{\lambda}^{\text{Lasso}})_j \neq 0, \\ [-1, 1] & \text{si } (\hat{\beta}_{\lambda}^{\text{Lasso}})_j = 0. \end{cases}$$

Rem: si  $\lambda > \lambda_{\max} := \max_{j \in \llbracket 1, p \rrbracket} |\langle \mathbf{x}_j, \mathbf{y} \rangle|$ , alors  $\hat{\beta}_{\lambda}^{\text{Lasso}} = 0$ .

preuve : vérifier les conditions ci-dessus pour 0 et  $\lambda > 0$

## Le cas orthogonal : le seuillage doux

Retour sur un cas simple (*design* orthogonal) :  $X^\top X = \text{Id}_p$

$$\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \|X^\top \mathbf{y} - X^\top X \boldsymbol{\beta}\|_2^2 = \|X^\top \mathbf{y} - \boldsymbol{\beta}\|_2^2$$

car  $X$  est une isométrie dans ce cas, l'objectif du lasso devient :

$$\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p \left( \frac{1}{2}(\mathbf{x}_j^\top \mathbf{y} - \beta_j)^2 + \lambda|\beta_j| \right)$$

**Problème séparable** : problème qui revient à minimiser terme à terme en séparant les termes la somme

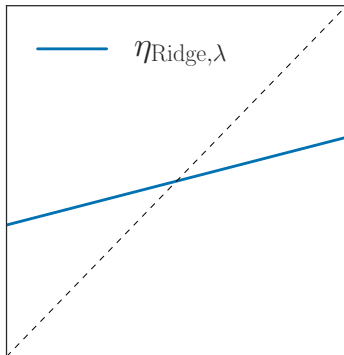
Il faut donc minimiser :  $x \mapsto \frac{1}{2}(z - x)^2 + \lambda|x|$  pour  $z = \mathbf{x}_j^\top \mathbf{y}$

Rem: on parle d'**opérateur proximal** en  $z$  de la fonction  $x \mapsto \lambda|x|$  (cf. Parikh et Boyd (2013), pour les méthodes proximales)

# Régularisation en 1D : Ridge

Résoudre :  $\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z - x)^2 + \frac{\lambda}{2}x^2$

$$\eta_\lambda(z) = \frac{z}{1 + \lambda}$$



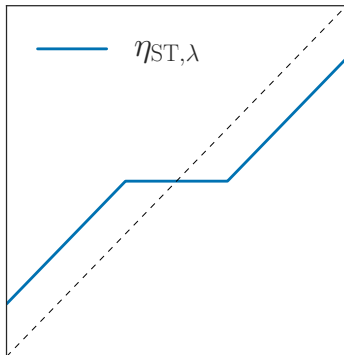
Contraction  $\ell_2$  : Ridge



## Régularisation en 1D : Lasso

Résoudre :  $\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} \frac{1}{2}(z - x)^2 + \lambda|x|$

$$\eta_\lambda(z) = \text{sign}(z)(|z| - \lambda)_+ \text{ (Exercice)}$$

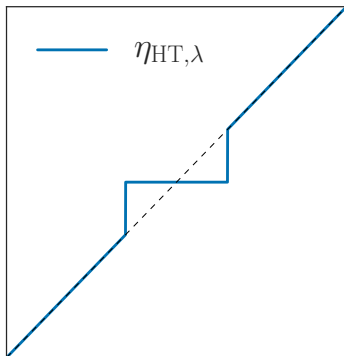



Contraction  $\ell_1$  : Seuillage doux (🇬🇧 : *soft thresholding*)

## Régularisation en 1D : $\ell_0$

Résoudre :  $\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z - x)^2 + \lambda \mathbb{1}_{x \neq 0}$

$$\eta_\lambda(z) = z \mathbb{1}_{|z| \geq \sqrt{2\lambda}}$$

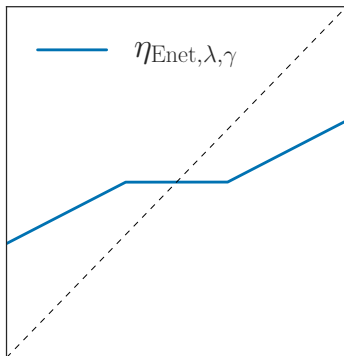


Contraction  $\ell_0$  : Seuillage dur ( : *hard thresholding*)

# Régularisation en 1D : Elastic Net

Résoudre :  $\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} x \mapsto \frac{1}{2}(z - x)^2 + \lambda(\gamma|x| + (1 - \gamma)\frac{x^2}{2})$

$\eta_\lambda(z)$  = Exercice



Contraction  $\ell_1/\ell_2$

## Seuillage doux : forme explicite

$$\eta_{\text{Lasso},\lambda}(z) = \begin{cases} z + \lambda & \text{si } z < -\lambda \\ 0 & \text{si } |z| \leq \lambda \\ z - \lambda & \text{si } z > \lambda \end{cases}$$

---

**Exercice:** Prouver le résultat précédent en utilisant les sous-gradients

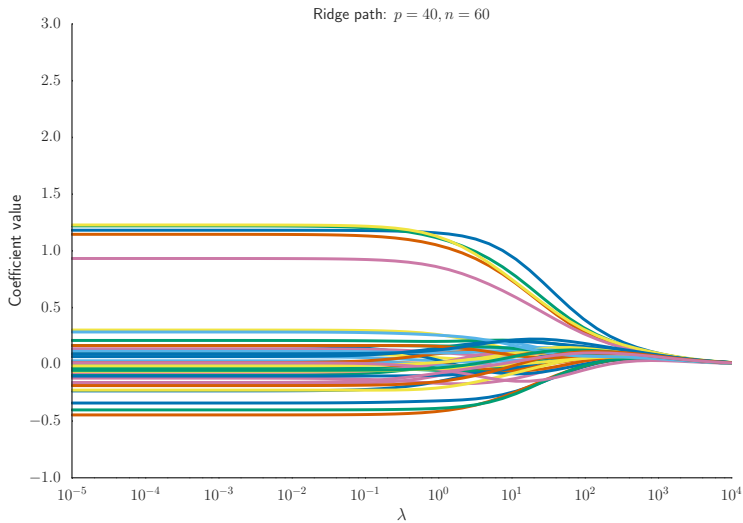
---

## Exemple numérique : simulation

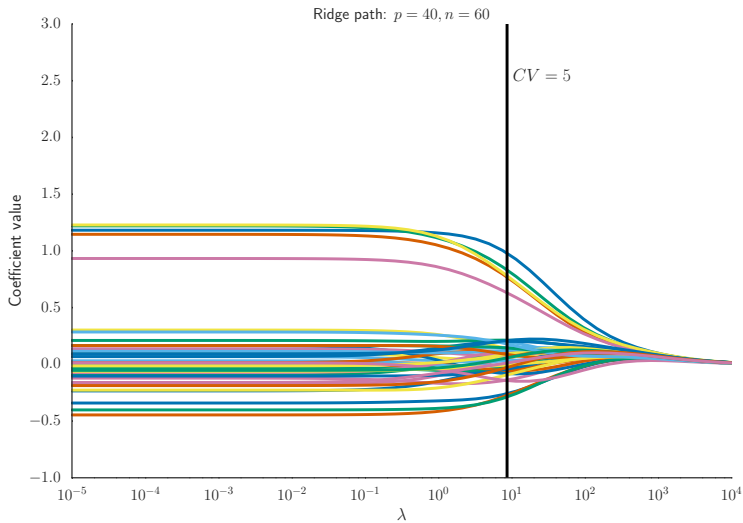
- ▶  $\beta^* = (1, 1, 1, 1, 1, 0, \dots, 0) \in \mathbb{R}^p$  (5 coefficients non-nuls)
- ▶  $X \in \mathbb{R}^{n \times p}$  a des colonnes tirées selon une loi gaussienne
- ▶  $y = X\beta^* + \varepsilon \in \mathbb{R}^n$  avec  $\varepsilon \sim \mathcal{N}(0, \sigma^2 \text{Id}_n)$
- ▶ On utilise une grille de 50 valeurs de  $\lambda$

Pour cet exemple les tailles sont :  $n = 60, p = 40, \sigma = 1$

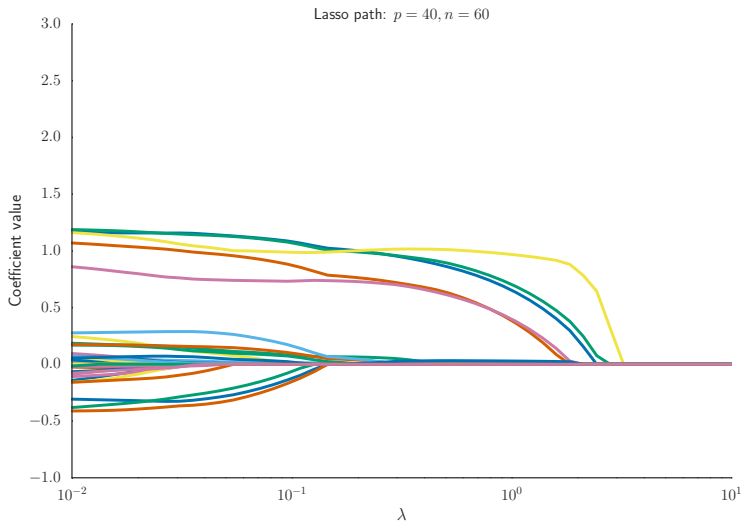
# Lasso vs Ridge



# Lasso vs Ridge

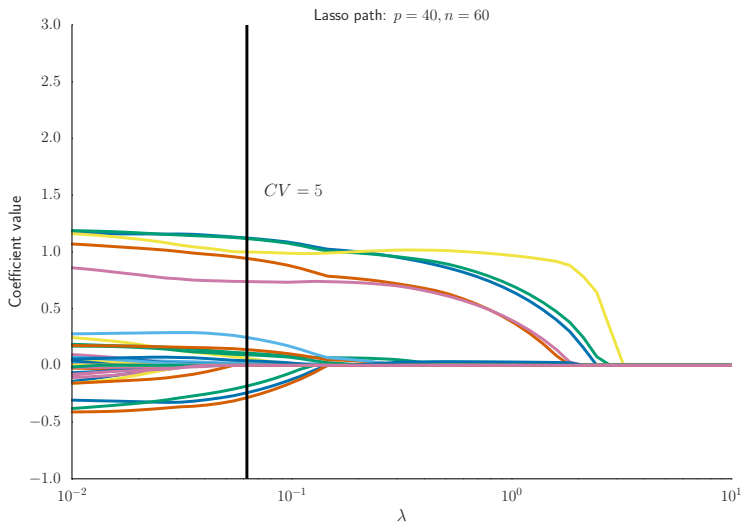


# Lasso vs Ridge





# Lasso vs Ridge



# Intérêt du Lasso

- ▶ Enjeu numérique : le Lasso est un problème **convexe**
- ▶ Sélection de variables/ solutions parcimonieuses (sparse) :  $\hat{\beta}_{\lambda}^{\text{Lasso}}$  a potentiellement de nombreux coefficients nuls. Le paramètre  $\lambda$  contrôle le niveau de parcimonie : si  $\lambda$  est grand, les solutions sont très creuses.

Exemple : on obtient 17 coefficients non nuls pour LassoCV dans la simulation précédente

Rem: RidgeCV n'avait aucun coefficient nul

# Analyse de l'estimateur dans le cas général

Analyse théorique : (nettement) plus poussée que pour les moindres carrés ou que pour Ridge ; peut être trouvée dans des références récentes, cf. [Buhlmann et van de Geer \(2011\)](#) pour des résultats théoriques

En résumé : on biaise l'estimateur des moindres carrés pour réduire la variance

# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

- LSLasso / Elastic-Net

- Pénalités non-convexes / Adaptive Lasso

- Structure sur le support

- Stabilisation

- Extensions des moindres carrés / Lasso

# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

- LSLasso / Elastic-Net

- Pénalités non-convexes / Adaptive Lasso

- Structure sur le support

- Stabilisation

- Extensions des moindres carrés / Lasso

# Le biais du Lasso

Le lasso est biaisé : il contracte les grands coefficients vers 0

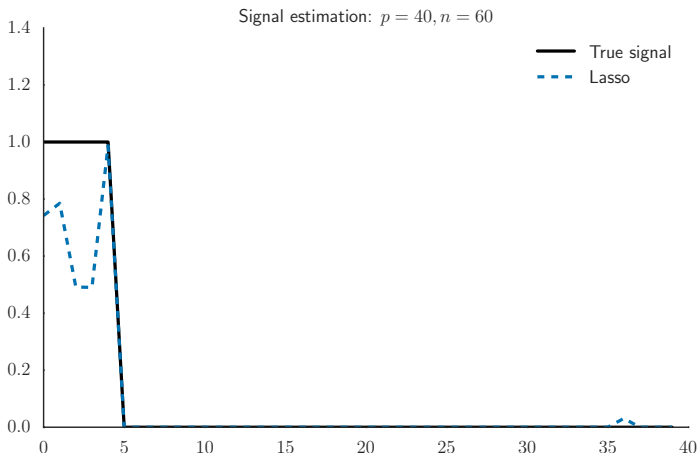


Illustration sur l'exemple

# Le biais du Lasso

Le lasso est biaisé : il contracte les grands coefficients vers 0

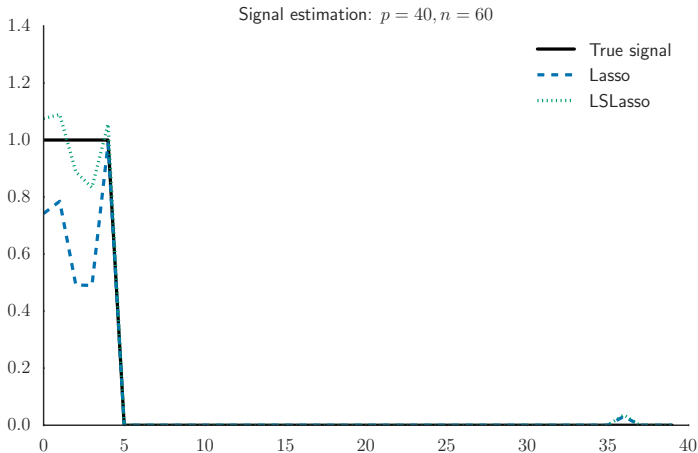


Illustration sur l'exemple

# Le biais du Lasso : un remède simple

Comme les grands coefficients sont parfois contractés vers zéro, il est possible d'utiliser une procédure en deux étapes

## Least Squares Lasso (LSLasso)

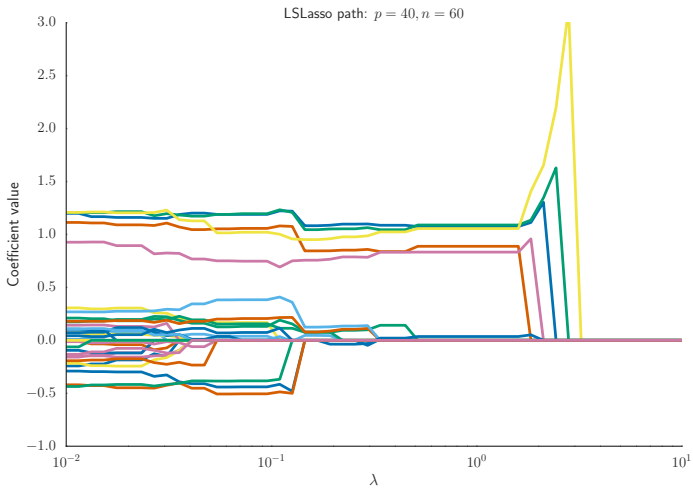
1. Lasso : obtenir  $\hat{\beta}_{\lambda}^{\text{Lasso}}$
2. Moindres-carrés sur les variables actives  $\text{supp}(\hat{\beta}_{\lambda}^{\text{Lasso}})$   
$$\hat{\beta}_{\lambda}^{\text{LSLasso}} \in \arg \min_{\beta \in \mathbb{R}^p \text{ supp}(\beta) = \text{supp}(\hat{\beta}_{\lambda}^{\text{Lasso}})} \frac{1}{2} \|\mathbf{y} - X\beta\|_2^2$$

Attention : il faut faire la CV sur la procédure entière ; choisir  $\lambda$  du Lasso par CV puis faire les moindres carrés garde trop de variables

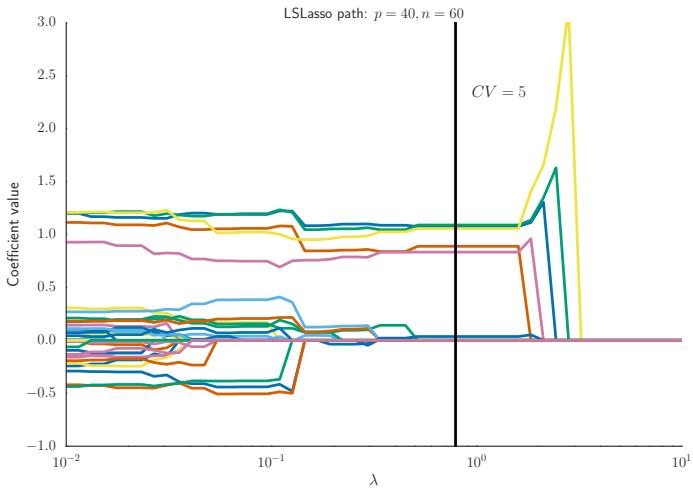
Rem: LSLasso pas forcément codé dans les packages usuels



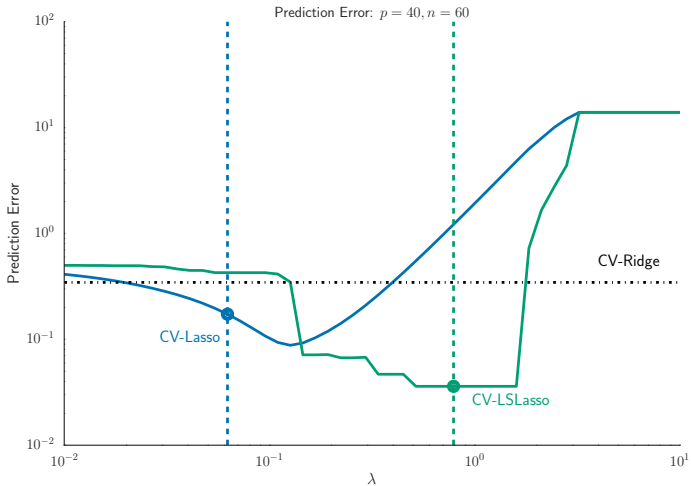
# Débiasage



# Débiasage



# Prédiction : Lasso vs. LSLasso



# Bilan du LSLasso

## Avantages :

- ▶ les “vrais” grands coefficients sont moins atténués
- ▶ en faisant la CV on récupère moins de variables parasites (amélioration de l'interprétabilité)  
e.g., sur l'exemple précédent le LSLassoCV retrouve les 5 “vraies” variables non nulles, et un faux positif

LSLasso : utile pour l'estimation

## Limites :

- ▶ la différence en prédiction n'est pas toujours flagrante
- ▶ nécessite plus de calcul : re-calculer autant de moindres carrés que de paramètres  $\lambda$  (de dimension la taille des supports, car on néglige les autres variables)
- ▶ non packagé

## Elastic Net : régularisation $\ell_1/\ell_2$

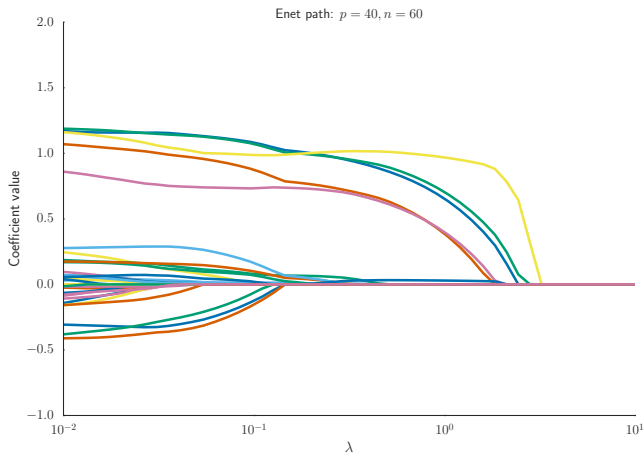
L'Elastic Net introduit par **Zou et Hastie (2005)** est solution de

$$\hat{\beta}_\lambda = \arg \min_{\beta \in \mathbb{R}^p} \left[ \frac{1}{2} \|\mathbf{y} - X\beta\|_2^2 + \lambda \left( \gamma \|\beta\|_1 + (1 - \gamma) \frac{\|\beta\|_2^2}{2} \right) \right]$$

Rem: deux paramètres de régularisation, un pour la régularisation globale, un qui contrôle l'influence Ridge vs. Lasso

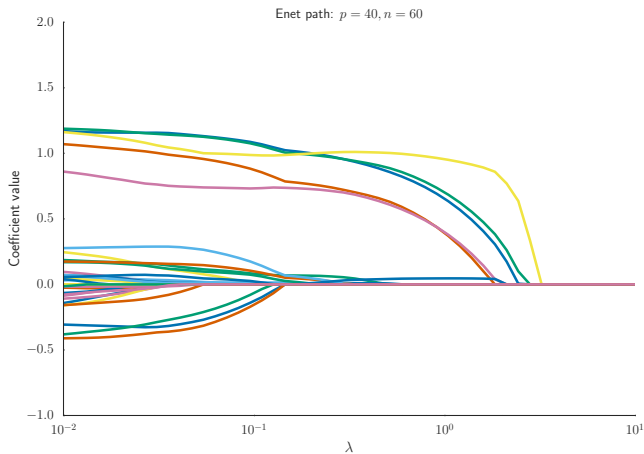
Rem: la solution est unique et la taille du support de l'Elastic Net est plus petite que  $\min(n, p)$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



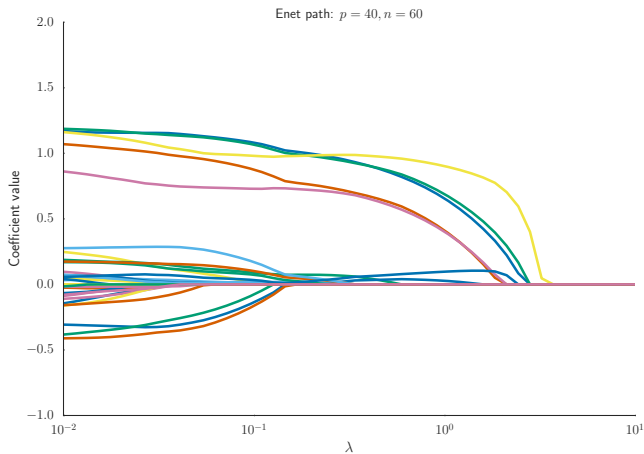
$$\gamma = 1.00$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



$$\gamma = 0.99$$

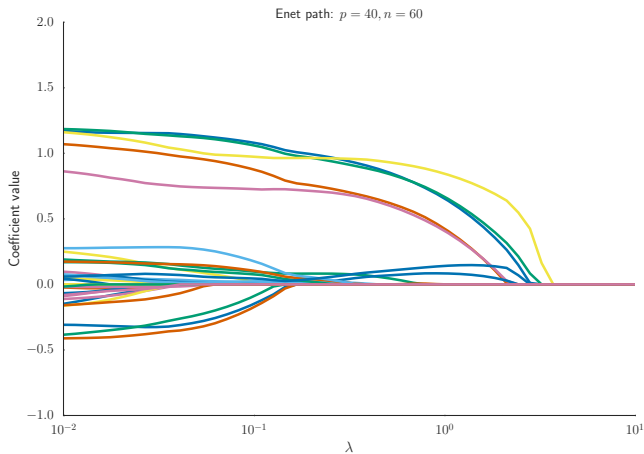
# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



$$\gamma = 0.95$$

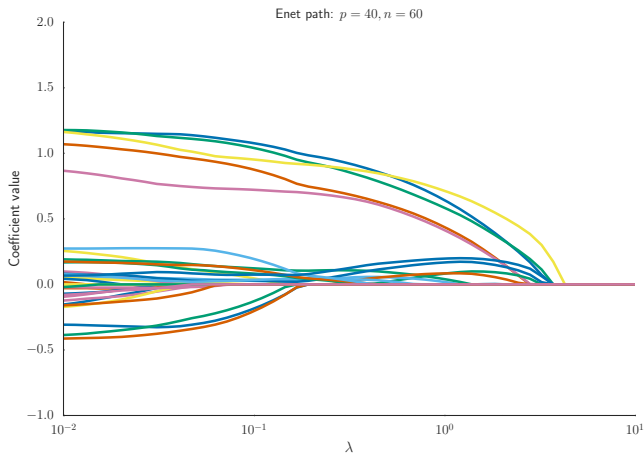


# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



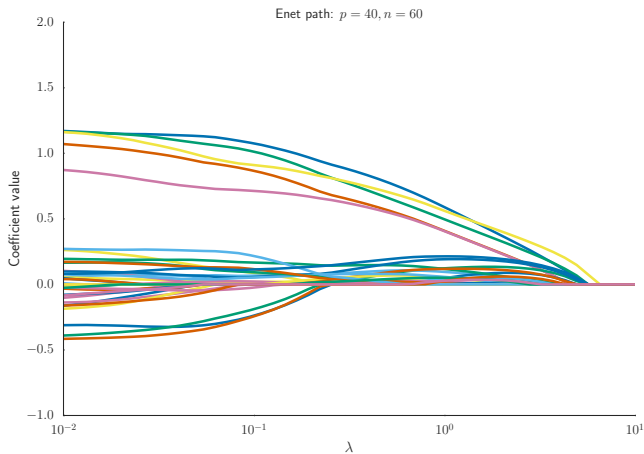
$$\gamma = 0.90$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



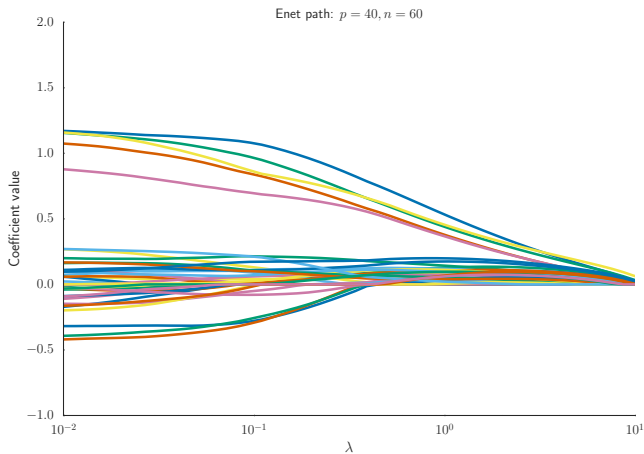
$$\gamma = 0.75$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



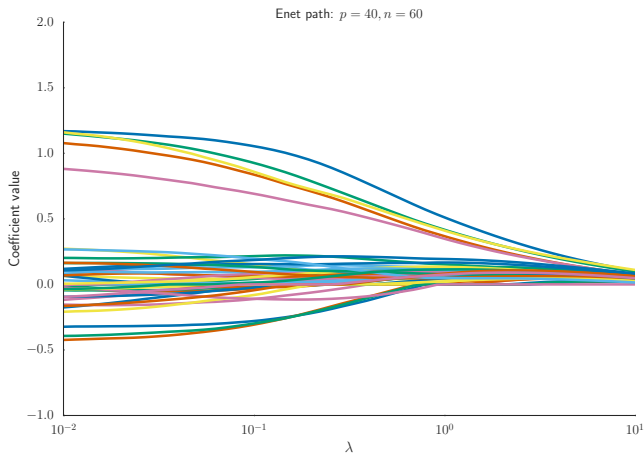
$$\gamma = 0.50$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



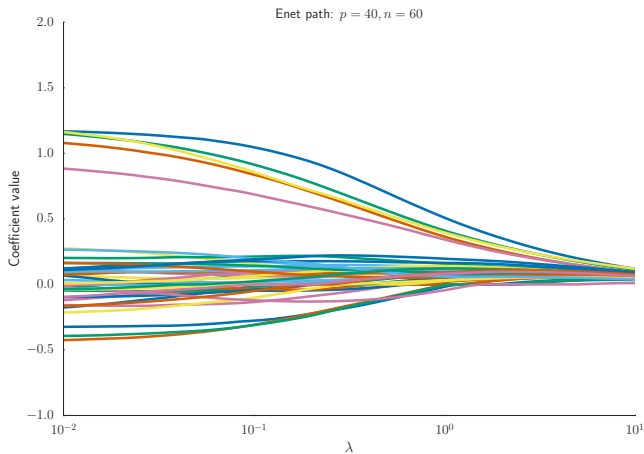
$$\gamma = 0.25$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



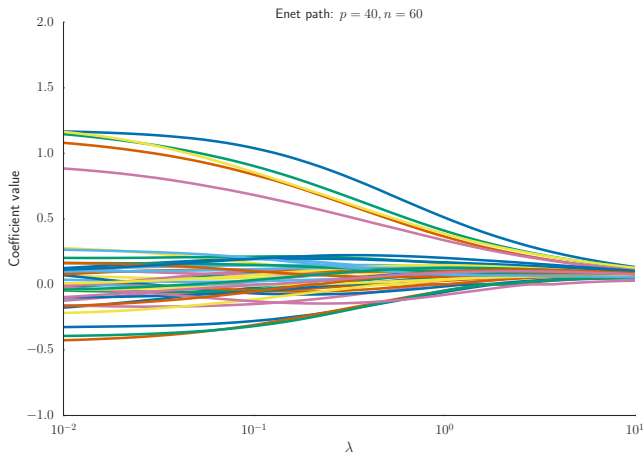
$$\gamma = 0.1$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



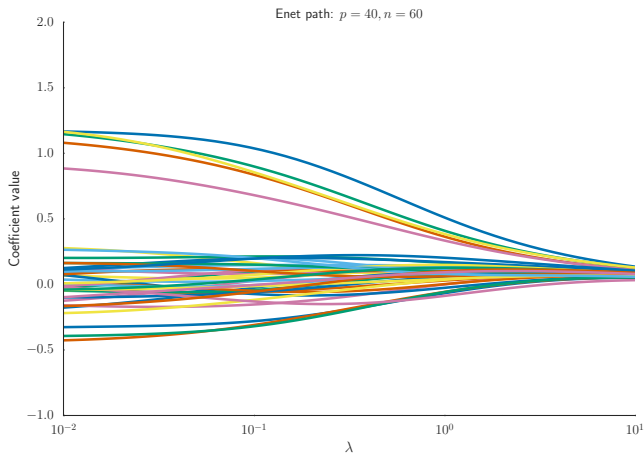
$$\gamma = 0.05$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



$$\gamma = 0.01$$

# Elastic-Net : $\gamma\|\beta\|_1 + (1 - \gamma)\|\beta\|_2^2/2$



$\gamma = 0.00$



# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

LSLasso / Elastic-Net

Pénalités non-convexes / Adaptive Lasso

Structure sur le support

Stabilisation

Extensions des moindres carrés / Lasso

## Pénalités non-convexes

Utiliser une pénalité non-convexe approchant mieux  $\|\cdot\|_0$ , en choisissant  $t \rightarrow \text{pen}_{\lambda,\gamma}(t)$  non-convexe

$$\hat{\beta}_{\lambda,\gamma}^{\text{pen}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\beta_j|)}_{\text{régularisation}} \right)$$

# Pénalités non-convexes

Utiliser une pénalité non-convexe approchant mieux  $\|\cdot\|_0$ , en choisissant  $t \rightarrow \text{pen}_{\lambda,\gamma}(t)$  non-convexe

$$\hat{\beta}_{\lambda,\gamma}^{\text{pen}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\beta_j|)}_{\text{régularisation}} \right)$$

- Adaptive-Lasso Zou (2006),  $\ell_1$  re-pondérés Candès *et al.* (2008)

$$\text{pen}_{\lambda,\gamma}(t) = \lambda |t|^q \text{ avec } 0 < q < 1$$

# Pénalités non-convexes

Utiliser une pénalité non-convexe approchant mieux  $\|\cdot\|_0$ , en choisissant  $t \rightarrow \text{pen}_{\lambda,\gamma}(t)$  non-convexe

$$\hat{\beta}_{\lambda,\gamma}^{\text{pen}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\beta_j|)}_{\text{régularisation}} \right)$$

- MCP (*minimax concave penalty*) Zhang (2010) pour  $\lambda > 0$  et  $\gamma > 1$

$$\text{pen}_{\lambda,\gamma}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\gamma}, & \text{si } |t| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{si } |t| > \gamma\lambda \end{cases}$$

# Pénalités non-convexes

Utiliser une pénalité non-convexe approchant mieux  $\|\cdot\|_0$ , en choisissant  $t \rightarrow \text{pen}_{\lambda,\gamma}(t)$  non-convexe

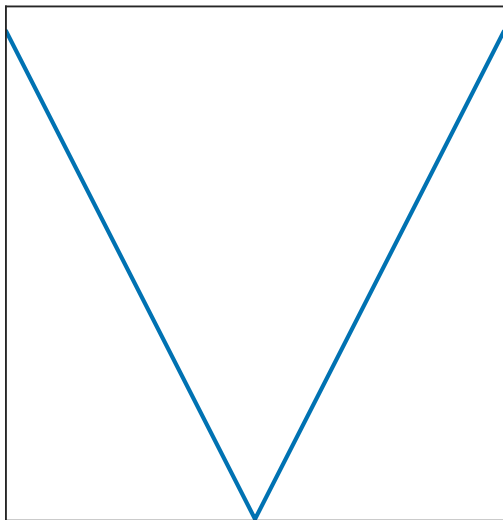
$$\hat{\beta}_{\lambda,\gamma}^{\text{pen}} \in \arg \min_{\beta \in \mathbb{R}^p} \left( \underbrace{\frac{1}{2} \|\mathbf{y} - X\beta\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^p \text{pen}_{\lambda,\gamma}(|\beta_j|)}_{\text{régularisation}} \right)$$

- SCAD (*Smoothly Clipped Absolute Deviation*) Fan et Li (2001) pour  $\lambda > 0$  et  $\gamma > 2$

$$\text{pen}_{\lambda,\gamma}(t) = \begin{cases} \lambda|t|, & \text{si } |t| \leq \lambda \\ \frac{\gamma\lambda|t| - (t^2 + \lambda^2)/2}{\gamma - 1}, & \text{si } \lambda < |t| \leq \gamma\lambda \\ \frac{\lambda^2(\gamma^2 - 1)}{2(\gamma - 1)}, & \text{si } |t| > \gamma\lambda \end{cases}$$

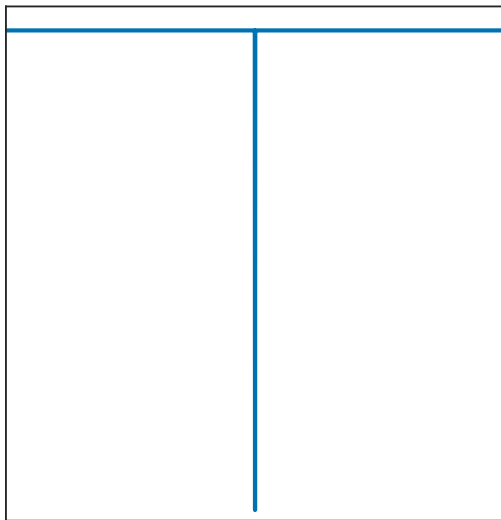
Rem: difficultés algorithmiques (arrêt, minima locaux, etc.)

# Forme des pénalités classiques



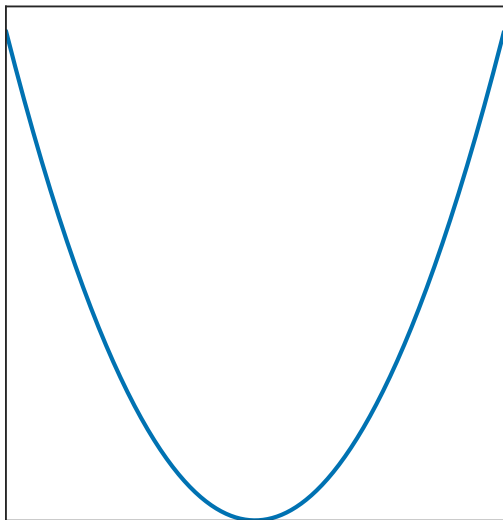
$|x|$

# Forme des pénalités classiques



10

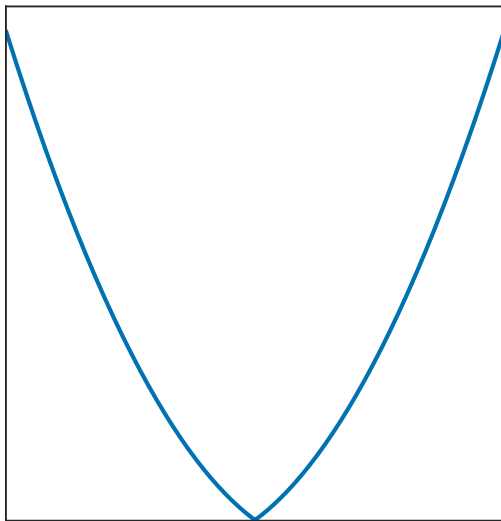
# Forme des pénalités classiques



$l_2$

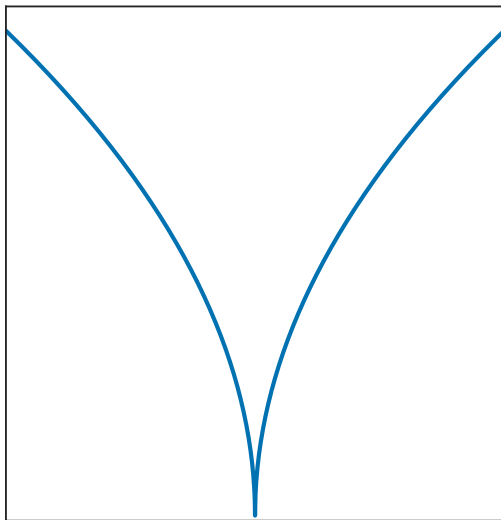


# Forme des pénalités classiques



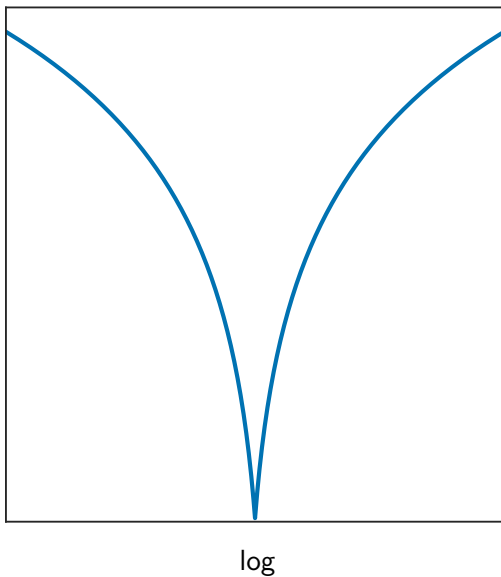
enet

# Forme des pénalités classiques

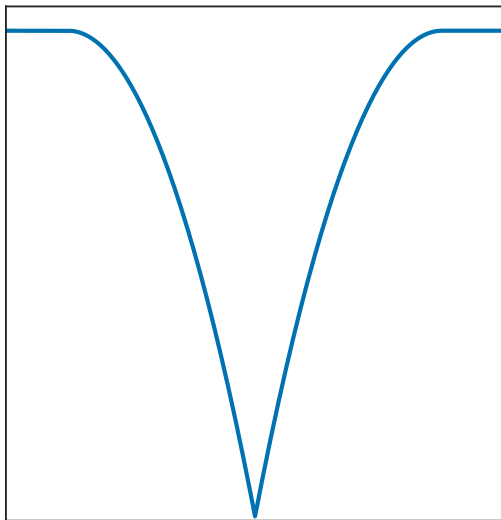


sqrt

## Forme des pénalités classiques

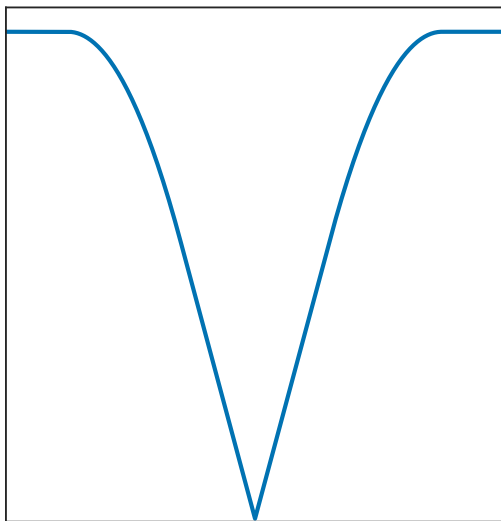


## Forme des pénalités classiques



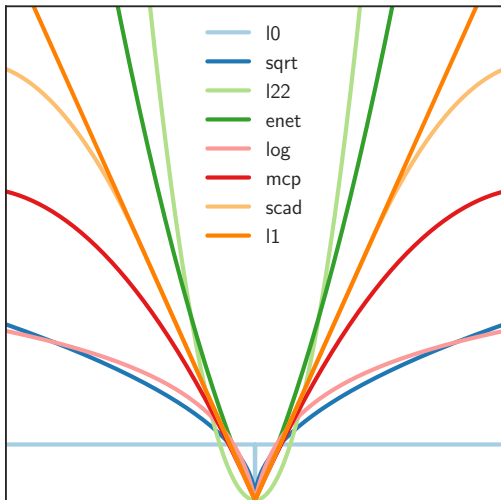
mcp

## Forme des pénalités classiques



scad

# Forme des pénalités classiques



# Adaptive-Lasso

Plusieurs noms pour une même idée :

- ▶ Adaptive-Lasso Zou (2006)
- ▶  $\ell_1$  re-pondérés Candès *et al.* (2008)
- ▶ approche DC-programming (pour *Difference of Convex Programming*) Gasso *et al.* (2008)

# Adaptive-Lasso

Exemple : prendre  $\text{pen}_{\lambda,\gamma}(t) = \lambda|t|^q$  avec  $q = 1/2$

---

**Algorithme** : Adaptive Lasso (cas  $q = 1/2$ )

---

**Entrées** :  $X, y$ , nombre d'itérations  $K$ , régularisation  $\lambda$

Initialisation :  $\hat{w} \leftarrow (1, \dots, 1)^\top$

---



# Adaptive-Lasso

Exemple : prendre  $\text{pen}_{\lambda,\gamma}(t) = \lambda|t|^q$  avec  $q = 1/2$

---

**Algorithme** : Adaptive Lasso (cas  $q = 1/2$ )

---

**Entrées** :  $X, y$ , nombre d'itérations  $K$ , régularisation  $\lambda$

Initialisation :  $\hat{w} \leftarrow (1, \dots, 1)^\top$

**pour**  $k = 1, \dots, K$  **faire**

# Adaptive-Lasso

Exemple : prendre  $\text{pen}_{\lambda,\gamma}(t) = \lambda|t|^q$  avec  $q = 1/2$

---

**Algorithme** : Adaptive Lasso (cas  $q = 1/2$ )

---

**Entrées** :  $X, \mathbf{y}$ , nombre d'itérations  $K$ , régularisation  $\lambda$

Initialisation :  $\hat{w} \leftarrow (1, \dots, 1)^\top$

**pour**  $k = 1, \dots, K$  **faire**

$$\hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \left( \frac{\|\mathbf{y} - X\beta\|_2^2}{2} + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \right)$$

---

# Adaptive-Lasso

Exemple : prendre  $\text{pen}_{\lambda,\gamma}(t) = \lambda|t|^q$  avec  $q = 1/2$

---

**Algorithme** : Adaptive Lasso (cas  $q = 1/2$ )

---

**Entrées** :  $X, y$ , nombre d'itérations  $K$ , régularisation  $\lambda$

Initialisation :  $\hat{w} \leftarrow (1, \dots, 1)^\top$

**pour**  $k = 1, \dots, K$  **faire**

$$\left| \begin{array}{l} \hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \left( \frac{\|y - X\beta\|_2^2}{2} + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \right) \\ \hat{w}_j \leftarrow \frac{1}{|\hat{\beta}_j|^{\frac{1}{2}}}, \forall j \in \llbracket 1, p \rrbracket \end{array} \right.$$

---

# Adaptive-Lasso

Exemple : prendre  $\text{pen}_{\lambda,\gamma}(t) = \lambda|t|^q$  avec  $q = 1/2$

---

**Algorithme** : Adaptive Lasso (cas  $q = 1/2$ )

---

**Entrées** :  $X, y$ , nombre d'itérations  $K$ , régularisation  $\lambda$

Initialisation :  $\hat{w} \leftarrow (1, \dots, 1)^\top$

**pour**  $k = 1, \dots, K$  **faire**

$$\left| \begin{array}{l} \hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \left( \frac{\|y - X\beta\|_2^2}{2} + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \right) \\ \hat{w}_j \leftarrow \frac{1}{|\hat{\beta}_j|^{\frac{1}{2}}}, \forall j \in \llbracket 1, p \rrbracket \end{array} \right.$$

---

Rem: en pratique pas besoin d'itérer beaucoup (5 itérations)

# Adaptive-Lasso

Exemple : prendre  $\text{pen}_{\lambda,\gamma}(t) = \lambda|t|^q$  avec  $q = 1/2$

---

**Algorithme** : Adaptive Lasso (cas  $q = 1/2$ )

---

**Entrées** :  $X, y$ , nombre d'itérations  $K$ , régularisation  $\lambda$

Initialisation :  $\hat{w} \leftarrow (1, \dots, 1)^\top$

**pour**  $k = 1, \dots, K$  **faire**

$$\left| \begin{array}{l} \hat{\beta} \leftarrow \arg \min_{\beta \in \mathbb{R}^p} \left( \frac{\|y - X\beta\|_2^2}{2} + \lambda \sum_{j=1}^p \hat{w}_j |\beta_j| \right) \\ \hat{w}_j \leftarrow \frac{1}{|\hat{\beta}_j|^{\frac{1}{2}}}, \forall j \in \llbracket 1, p \rrbracket \end{array} \right.$$

---

Rem: en pratique pas besoin d'itérer beaucoup (5 itérations)

Rem: utiliser un solveur Lasso pour mettre à jour  $\hat{\beta}$

# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

LSLasso / Elastic-Net

Pénalités non-convexes / Adaptive Lasso

Structure sur le support

Stabilisation

Extensions des moindres carrés / Lasso

# Structure du support

On suppose ici que l'on connaît une structure de groupes sur les variables au préalable de l'étude :  $\llbracket 1, p \rrbracket = \bigcup_{g \in \mathcal{G}} g$

Vecteur et ses coordonnées actives (en orange) :



Support creux : quelconque

Pénalité envisagée : Lasso

$$\|\theta\|_1 = \sum_{j=1}^p |\beta_j|$$

# Structure du support

On suppose ici que l'on connaît une structure de groupes sur les variables au préalable de l'étude :  $\llbracket 1, p \rrbracket = \bigcup_{g \in \mathcal{G}} g$

Vecteur et ses coordonnées actives (en orange) :



Support creux : groupes

Pénalité envisagée : Groupe-Lasso

$$\|\theta\|_{2,1} = \sum_{g \in G} \|\beta_g\|_2$$



# Structure du support

On suppose ici que l'on connaît une structure de groupes sur les variables au préalable de l'étude :  $\llbracket 1, p \rrbracket = \bigcup_{g \in \mathcal{G}} g$

Vecteur et ses coordonnées actives (en orange) :



Support creux : groupes + sous groupes

Pénalité envisagée : Sparse-Groupe-Lasso

$$\alpha \|\theta\|_1 + (1 - \alpha) \|\theta\|_{2,1} = \alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{g \in G} \|\beta_g\|_2$$

# Groupe-Lasso

La pénalisation par la norme  $\ell_1$  assure que peu de coefficients sont actifs, mais aucune autre structure sur le support n'est utilisée

## Structures additionnelles classiques :

- ▶ Parcimonie par groupe/bloc : Groupe-Lasso Yuan et Lin (2006)
- ▶ Parcimonie individuelle et par groupe : Sparse Groupe-Lasso Simon, Friedman, Hastie et Tibshirani (2012)
- ▶ Structures hiérarchiques (par exemple avec les interactions d'ordre supérieur) Bien, Taylor et Tibshirani (2013)
- ▶ Structures sur des graphes, des gradients, etc.

# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

LSLasso / Elastic-Net

Pénalités non-convexes / Adaptive Lasso

Structure sur le support

Stabilisation

Extensions des moindres carrés / Lasso

# Stabilisation du Lasso

Le Lasso peut être **instable** : quand il n'y a pas unicité de la solution (e.g., quand  $p > n$ ) selon le solveur numérique et la précision demandée, les variables sélectionnées peuvent différer.

On peut limiter ce genre de défauts en utilisant des techniques de ré-échantillonnage :

- ▶ Bolasso [Bach \(2008\)](#)
- ▶ Stability Selection [Meinshausen et Bühlmann \(2010\)](#)

# Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, y$ , nombre de réplifications  $B$ , régularisation  $\lambda$

---

---

**Exercice:** coder le Bolasso avec Python et sklearn

---

# Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, y$ , nombre de réplifications  $B$ , régularisation  $\lambda$   
**pour**  $k = 1, \dots, B$  **faire**

|

---

---

**Exercice:** coder le Bolasso avec Python et sklearn

---

## Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, y$ , nombre de réplifications  $B$ , régularisation  $\lambda$

**pour**  $k = 1, \dots, B$  **faire**

    Générer un échantillon *bootstrap* :  $X^{(k)}, y^{(k)}$

---

**Exercice:** coder le Bolasso avec Python et sklearn

---

# Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, y$ , nombre de réplifications  $B$ , régularisation  $\lambda$

**pour**  $k = 1, \dots, B$  **faire**

    Générer un échantillon *bootstrap* :  $X^{(k)}, y^{(k)}$

    Calculer le Lasso sur cet échantillon :  $\hat{\beta}_{\lambda}^{\text{Lasso},(k)}$

---

**Exercice:** coder le Bolasso avec Python et sklearn

---



# Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, y$ , nombre de réplifications  $B$ , régularisation  $\lambda$

**pour**  $k = 1, \dots, B$  **faire**

    Générer un échantillon *bootstrap* :  $X^{(k)}, y^{(k)}$

    Calculer le Lasso sur cet échantillon :  $\hat{\beta}_{\lambda}^{\text{Lasso},(k)}$

    Calculer le support associé :  $S_k = \text{supp} \left( \hat{\beta}_{\lambda}^{\text{Lasso},(k)} \right)$

---

**Exercice:** coder le Bolasso avec Python et sklearn

---

# Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, y$ , nombre de réplifications  $B$ , régularisation  $\lambda$

**pour**  $k = 1, \dots, B$  **faire**

    Générer un échantillon *bootstrap* :  $X^{(k)}, y^{(k)}$

    Calculer le Lasso sur cet échantillon :  $\hat{\beta}_{\lambda}^{\text{Lasso},(k)}$

    Calculer le support associé :  $S_k = \text{supp} \left( \hat{\beta}_{\lambda}^{\text{Lasso},(k)} \right)$

Calculer :  $S := \bigcap_{k=1}^B S_k$

---

**Exercice:** coder le Bolasso avec Python et sklearn

---

# Bolasso Bach (2008)

---

**Algorithme :** Bootstrap Lasso

---

**Entrées :**  $X, \mathbf{y}$ , nombre de réplifications  $B$ , régularisation  $\lambda$   
**pour**  $k = 1, \dots, B$  **faire**

    Générer un échantillon *bootstrap* :  $X^{(k)}, y^{(k)}$

    Calculer le Lasso sur cet échantillon :  $\hat{\beta}_{\lambda}^{\text{Lasso},(k)}$

    Calculer le support associé :  $S_k = \text{supp} \left( \hat{\beta}_{\lambda}^{\text{Lasso},(k)} \right)$

Calculer :  $S := \bigcap_{k=1}^B S_k$

Calculer :  $\hat{\beta}_{\lambda}^{\text{Bolasso}} \in \arg \min_{\substack{\beta \in \mathbb{R}^p \\ \text{supp}(\beta) = S}} \frac{1}{2} \|\mathbf{y} - X\beta\|_2^2$

---

**Exercice:** coder le Bolasso avec Python et sklearn

---

# Bolasso Bach (2008)

---

**Algorithme** : Bootstrap Lasso

---

**Entrées** :  $X, \mathbf{y}$ , nombre de réplifications  $B$ , régularisation  $\lambda$   
**pour**  $k = 1, \dots, B$  **faire**

    Générer un échantillon *bootstrap* :  $X^{(k)}, y^{(k)}$

    Calculer le Lasso sur cet échantillon :  $\hat{\beta}_{\lambda}^{\text{Lasso},(k)}$

    Calculer le support associé :  $S_k = \text{supp} \left( \hat{\beta}_{\lambda}^{\text{Lasso},(k)} \right)$

Calculer :  $S := \bigcap_{k=1}^B S_k$

Calculer :  $\hat{\beta}_{\lambda}^{\text{Bolasso}} \in \arg \min_{\substack{\beta \in \mathbb{R}^p \\ \text{supp}(\beta) = S}} \frac{1}{2} \|\mathbf{y} - X\beta\|_2^2$

**Sorties** : un support  $S$ , et un vecteur  $\hat{\beta}_{\lambda}^{\text{Bolasso}}$

---

---

**Exercice**: coder le Bolasso avec Python et sklearn

---

# Sommaire

Rappels

Sélection de variables et parcimonie

Améliorations et extensions du Lasso

LSLasso / Elastic-Net

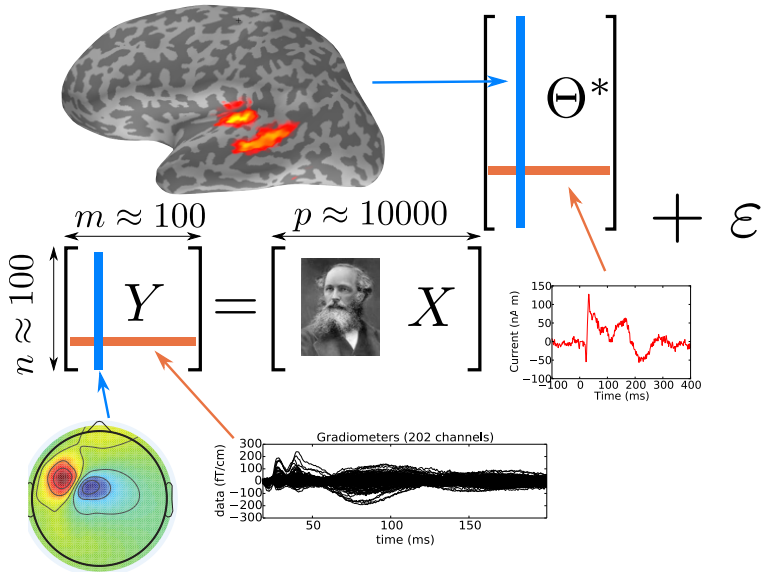
Pénalités non-convexes / Adaptive Lasso

Structure sur le support

Stabilisation

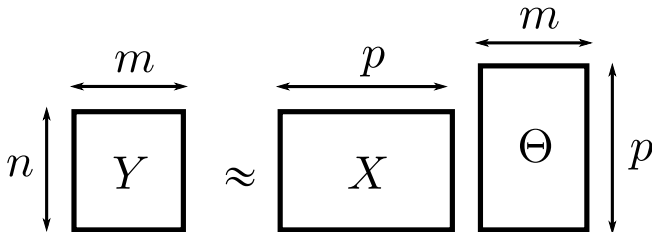
Extensions des moindres carrés / Lasso

# Example



# Régression multi-tâches

On veut résoudre  $m$  régressions linéaires conjointement :  $Y \approx X\Theta$



avec

- ▶  $Y \in \mathbb{R}^{n \times m}$  : matrice des observations
- ▶  $X \in \mathbb{R}^{n \times p}$  : matrice de design (commune)
- ▶  $\Theta \in \mathbb{R}^{p \times m}$  : matrice des coefficients

Exemple : plusieurs signaux sont observés au cours du temps (e.g., divers capteurs d'un même phénomène)

Rem: cf. `MultiTaskLasso` dans `sklearn` pour le numérique

# Moindre carrés pénalisées

Dans le contexte de la régression multi-tâches on peut résoudre les moindres carrés pénalisés :

$$\hat{\Theta}_\lambda \in \arg \min_{\Theta \in \mathbb{R}^{p \times m}} \left( \underbrace{\frac{1}{2} \|Y - X\Theta\|_F^2}_{\text{attache aux données}} + \underbrace{\lambda \Omega(\Theta)}_{\text{régularisation}} \right)$$

où  $\Omega$  est une pénalité / régularisation à préciser

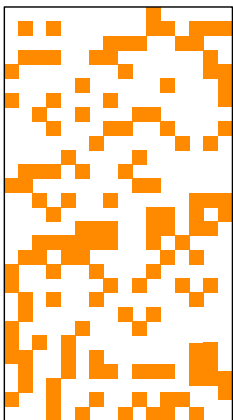
Rem: la norme de Frobenius  $\|\cdot\|_F$  est définie pour toute matrice  $A \in \mathbb{R}^{n_1 \times n_2}$  par

$$\|A\|_F^2 = \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} A_{j_1, j_2}^2$$



# Pénalisation pour le cas multi-tâches

On doit adapter les pénalisations vectorielles rencontrées :



Paramètre  $\Theta \in \mathbb{R}^{p \times m}$

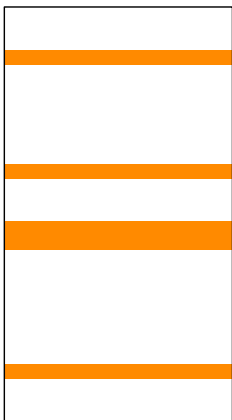
Support creux :  
quelconque

Pénalité Lasso :

$$\|\Theta\|_1 = \sum_{j=1}^p \sum_{k=1}^m |\Theta_{j,k}|$$

# Pénalisation pour le cas multi-tâches

On doit adapter les pénalisations vectorielles rencontrées :



Paramètre  $\Theta \in \mathbb{R}^{p \times m}$

Support creux :  
groupes

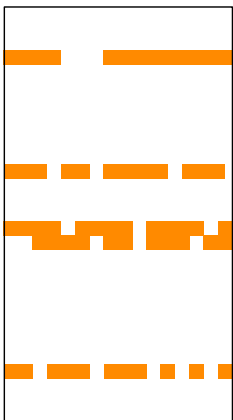
Pénalité Groupe-Lasso :

$$\|\Theta\|_{2,1} = \sum_{j=1}^p \|\Theta_{j:}\|_2$$

Rem: on note  $\Theta_{j:}$  la  $j^{\text{e}}$  ligne de  $\Theta$

# Pénalisation pour le cas multi-tâches

On doit adapter les pénalisations vectorielles rencontrées :



Paramètre  $\Theta \in \mathbb{R}^{p \times m}$

Support creux :  
groupes + sous groupes

Pénalité Sparse-Groupe-Lasso :

$$\alpha \|\Theta\|_1 + (1 - \alpha) \|\Theta\|_{2,1}$$

# Bibliographie I

- ▶ BACH, F. “Bolasso : model consistent Lasso estimation through the bootstrap”. In : *ICML*. 2008.
- ▶ BERTSIMAS, D., A. KING et R. MAZUMDER. “Best subset selection via a modern optimization lens”. In : *Ann. Statist.* 44.2 (2016), p. 813-852.
- ▶ BÜHLMANN, P. et S. VAN DE GEER. *Statistics for high-dimensional data*. Springer Series in Statistics. Methods, theory and applications. Heidelberg : Springer, 2011.
- ▶ FAN, J. et R. LI. “Variable selection via nonconcave penalized likelihood and its oracle properties”. In : *J. Amer. Statist. Assoc.* 96.456 (2001), p. 1348-1360.
- ▶ FERCOQ, O., A. GRAMFORT et J. SALMON. “Mind the duality gap : safer rules for the lasso”. In : *ICML*. 2015, p. 333-342.
- ▶ GASSO, G., A. RAKOTOMAMONJY et S. CANU. “Recovering sparse signals with non-convex penalties and DC programming”. In : *IEEE Trans. Signal Process.* 57.12 (2009), p. 4686-4698.

## Bibliographie II

- ▶ J, Bien, J. TAYLOR et R. TIBSHIRANI. "A lasso for hierarchical interactions". In : *Ann. Statist.* 41.3 (2013), p. 1111-1141.
- ▶ MEINSHAUSEN, N. et P. BÜHLMANN. "Stability selection". In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72.4 (2010), p. 417-473.
- ▶ SIMON, N. et al. "A sparse-group lasso". In : *J. Comput. Graph. Statist.* 22.2 (2013), p. 231-245. ISSN : 1061-8600.
- ▶ TIBSHIRANI, R. "Regression Shrinkage and Selection via the Lasso". In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58.1 (1996), p. 267-288.
- ▶ YUAN, M. et Y. LIN. "Model selection and estimation in regression with grouped variables". In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 68.1 (2006), p. 49-67.
- ▶ ZHANG, C.-H. "Nearly unbiased variable selection under minimax concave penalty". In : *Ann. Statist.* 38.2 (2010), p. 894-942.
- ▶ ZOU, H. "The adaptive lasso and its oracle properties". In : *J. Amer. Statist. Assoc.* 101.476 (2006), p. 1418-1429.

# Bibliographie III

- ▶ ZOU, H. et T. J. HASTIE. "Regularization and variable selection via the elastic net". In : *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.2 (2005), p. 301-320.