MS BGD
MDI 720 : Lasso

Joseph Salmon
http://josephsalmon.eu
Télécom Paristech, Institut Mines-Télécom

Plan

Rappels

Sélection de variables et parcimonie

La pénalisation ℓ_0 et ses limites
La pénalisation ℓ_1
Sous-gradient / sous-différentielle

Améliorations et extensions du Lasso

LSLasso / Elastic-Net
Pénalités non-convexes / Adaptive Lasso
Structure sur le support
Stabilisation
Extensions des moindres carrés / Lasso

Retour sur le modèle

$$\mathbf{y} = X\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

$$X = [\mathbf{x}_1, \dots, \mathbf{x}_p] = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n\times p}, \boldsymbol{\theta}^\star \in \mathbb{R}^p$$

Motivation

Utilité des estimateurs $\hat{\boldsymbol{\theta}}$ 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 $\boldsymbol{\theta}^\star$ a peu de coefficients non nuls

Méthodes de sélection de variables

- Méthodes de **dépistage par corrélation** (🇬🇧 : *correlation screening*) : supprimer les $\mathbf{x}_j$ de faible corrélation avec $\mathbf{y}$
  • avantages : rapide (+++), coût : p produits scalaires de taille n, intuitive (+++)
  • défauts : néglige les interactions entre variables $\mathbf{x}_j$, 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 $\mathbf{x}_j$ de faible corrélation avec $\mathbf{y}$
  • avantages : rapide ($+++$), coût : p produits scalaires de taille n, intuitive ($+++$)
  • défauts : néglige les interactions entre variables $\mathbf{x}_j$, 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 $\mathbf{x}_j$ de faible corrélation avec $\mathbf{y}$
  • avantages : rapide (+++), coût : p produits scalaires de taille n, intuitive (+++)
  • défauts : néglige les interactions entre variables $\mathbf{x}_j$, 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 ℓ_0

Définitions

Le **support** du vecteur $\boldsymbol{\theta}$ est l'ensemble des indices des coordonnées non nulles :

$$\operatorname{supp}(\boldsymbol{\theta}) = \{j \in [\![1, p]\!], \theta_j \neq 0\}$$

La **pseudo-norme** ℓ_0 d'un vecteur $\boldsymbol{\theta} \in \mathbb{R}^p$ est son nombre de coordonnées non-nulles :

$$\|\boldsymbol{\theta}\|_0 = \operatorname{card}\{j \in [\![1, p]\!], \theta_j \neq 0\}$$

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

Rem: $\|\cdot\|_0$ n'est pas non plus convexe, $\boldsymbol{\theta}_1 = (1, 0, 1, \dots, 0)$ $\boldsymbol{\theta}_2 = (0, 1, 1, \dots, 0)$ et $3 = \|\frac{\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2}{2}\|_0 \geqslant \frac{\|\boldsymbol{\theta}_1\|_0 + \|\boldsymbol{\theta}_2\|_0}{2} = 2$

Sommaire

La pénalisation ℓ_0

Première tentative de méthode pénalisée pour introduire de la parcimonie : utiliser ℓ_0 pour la pénalisation / régularisation

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

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

Le Lasso : la définition pénalisée

Lasso : *Least Absolute Shrinkage and Selection Operator*

Tibshirani (1996)

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

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

- On retrouve de nouveau les cas limites :

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

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

Interprétation contrainte

Un problème de la forme :

$$\boxed{\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\lambda\|\boldsymbol{\theta}\|_1}_{\text{régularisation}} \right)}$$

admet la même solution qu'une version contrainte :

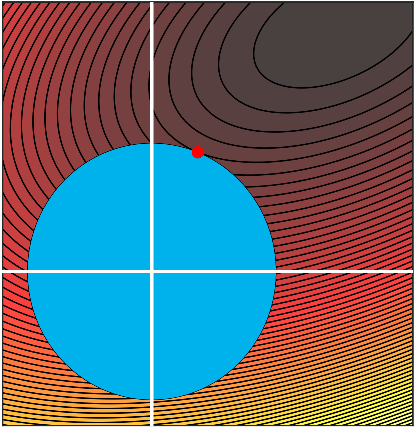
$$\begin{cases} \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2 \\ \text{t.q. } \|\boldsymbol{\theta}\|_1 \leqslant T \end{cases}$$

pour un certain $T > 0$.

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

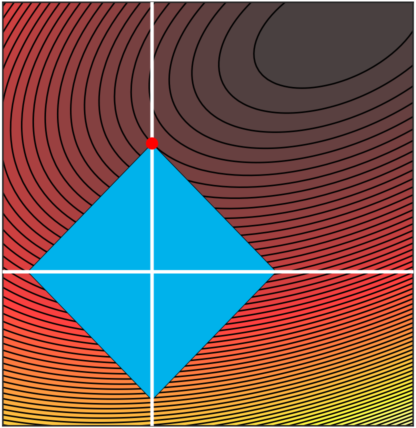
- Si $T \to 0$ on retrouve comme solution le vecteur nul : $0 \in \mathbb{R}^p$
- Si $T \to \infty$ on retrouve $\hat{\boldsymbol{\theta}}^{\mathrm{MCO}}$ (non contraint)

Mise à zéro de certains coefficients



Optimisation sous contrainte ℓ_2 : solution non parcimonieuse

Mise à zéro de certains coefficients



Optimisation sous contrainte ℓ_1 : solution parcimonieuse

Sommaire

Sous-gradients / sous-différentielles

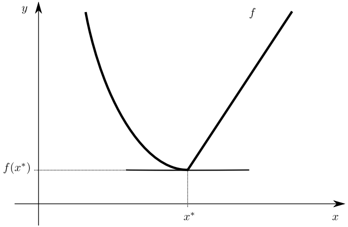
Définitions

Pour $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

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

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



Sous-gradients / sous-différentielles

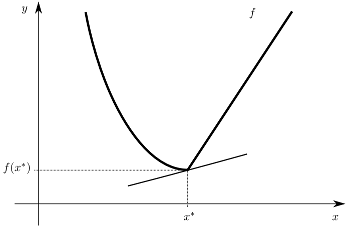
Définitions

Pour $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

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

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



Sous-gradients / sous-différentielles

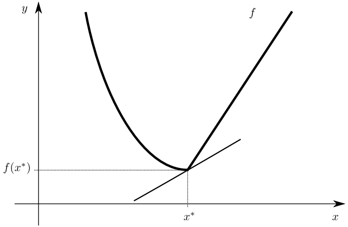
Définitions

Pour $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

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

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



Sous-gradients / sous-différentielles

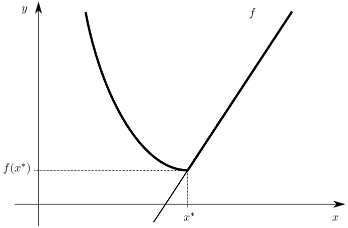
Définitions

Pour $f : \mathbb{R}^n \to \mathbb{R}$ 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

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

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

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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $\forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle$

Règle de Fermat

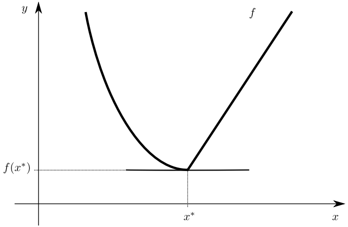
Théorème

Un point x^* est un minimum d'une fonction convexe $f : \mathbb{R}^n \to \mathbb{R}$ 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 $\forall x \in \mathbb{R}^n, f(x) \geqslant f(x^*) + \langle 0, x - x^* \rangle$

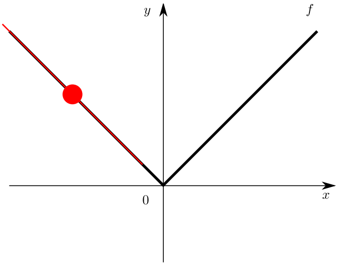
Rem: visuellement cela correspond à une tangente horizontale



Sous-différentielle de la valeur absolue

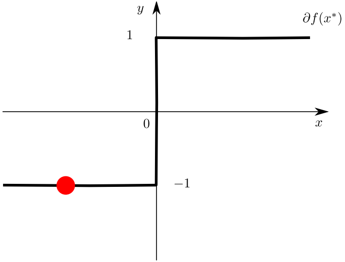
Fonction (abs) :

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



Sous-différentielle (sign)

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



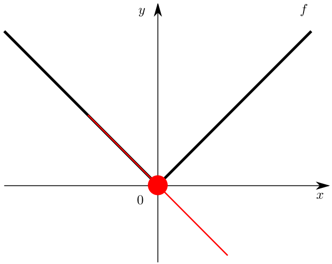
Sous-différentielle de la valeur absolue

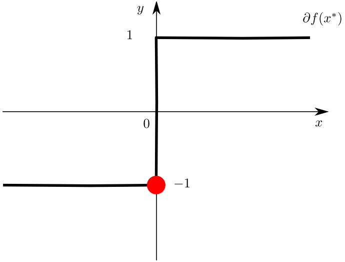
Fonction (abs) :

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

Sous-différentielle (sign)

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





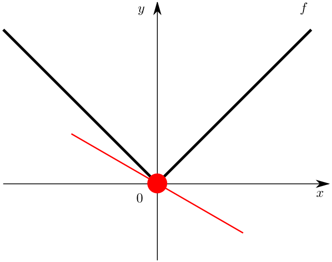
Sous-différentielle de la valeur absolue

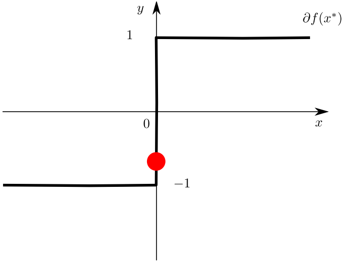
Fonction (abs) :

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

Sous-différentielle (sign)

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





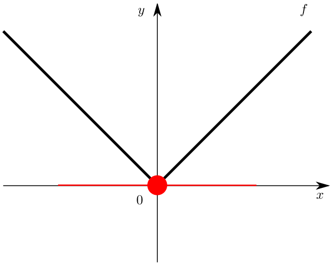
Sous-différentielle de la valeur absolue

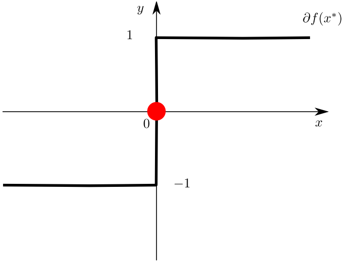
Fonction (abs) :

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

Sous-différentielle (sign)

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





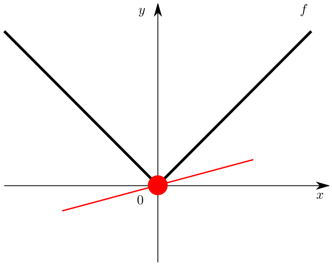
Sous-différentielle de la valeur absolue

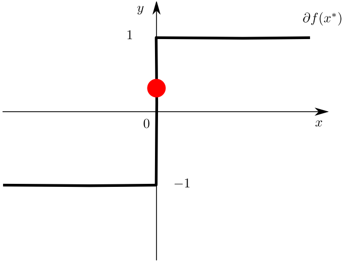
Fonction (abs) :

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

Sous-différentielle (sign)

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





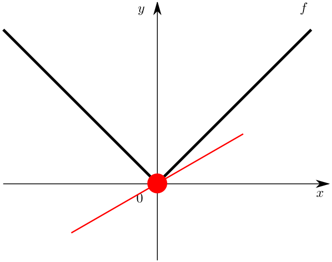
Sous-différentielle de la valeur absolue

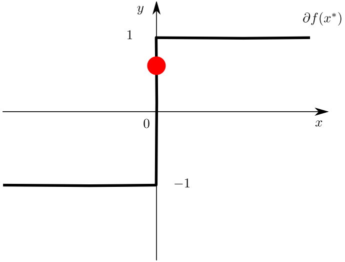
Fonction (abs) :

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

Sous-différentielle (sign)

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





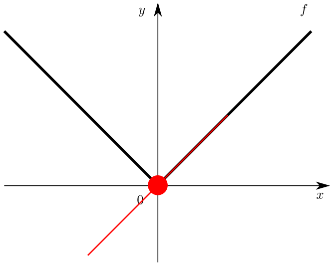
Sous-différentielle de la valeur absolue

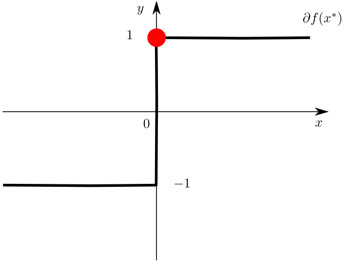
Fonction (abs) :

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

Sous-différentielle (sign)

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





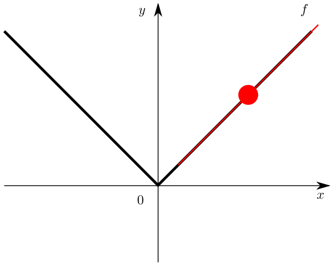
Sous-différentielle de la valeur absolue

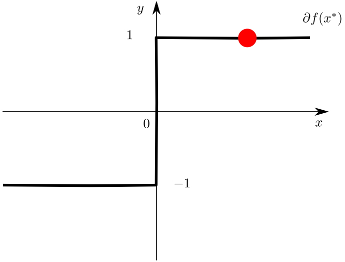
Fonction (abs) :

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

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

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

Conditions nécessaires et suffisantes d'optimalité (Fermat) :

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

<u>Rem</u>: si $\lambda > \lambda_{\max} := \max_{j\in[\![1,p]\!]} |\langle \mathbf{x}_j, \mathbf{y}\rangle|$, alors $\hat{\boldsymbol{\theta}}_\lambda^{\text{Lasso}} = 0$.

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

Le cas orthogonal : le seuillage doux

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

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

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

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

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

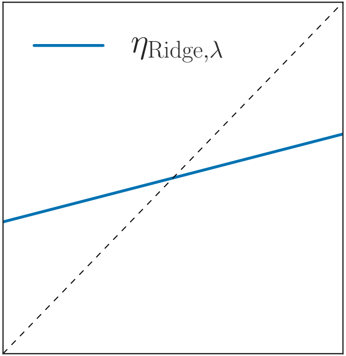
Il faut donc minimiser : $x \mapsto \frac{1}{2}(z - x)^2 + \lambda|x|$ pour $z = \mathbf{x}_j^\top \mathbf{y}$

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

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

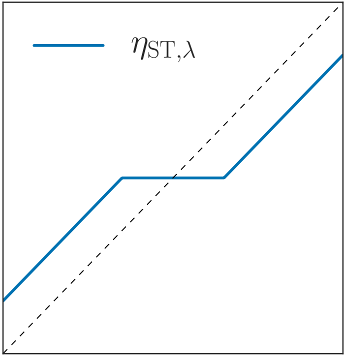


Contraction ℓ_2 : Ridge

Régularisation en 1D : Lasso

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

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

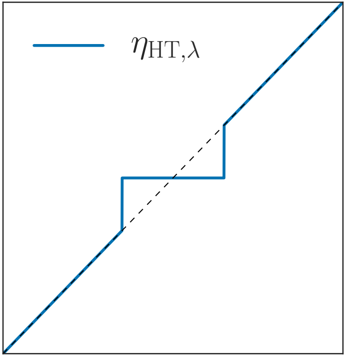


Contraction ℓ_1 : Seuillage doux (🇬🇧 : *soft thresholding*)

Régularisation en 1D : ℓ_0

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

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

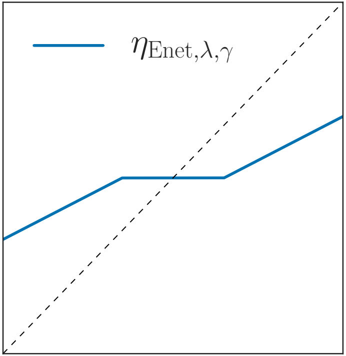


Contraction ℓ_0 : Seuillage dur (🇬🇧 : *hard thresholding*)

Régularisation en 1D : Elastic Net

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

$$\eta_\lambda(z) = \text{Exercice}$$



Contraction ℓ_1/ℓ_2

Seuillage doux : forme explicite

$$\eta_{\mathrm{Lasso},\lambda}(z) = \begin{cases} z+\lambda & \text{si } \ z \ \leqslant -\lambda \\ 0 & \text{si } |z| \leqslant \lambda \\ z-\lambda & \text{si } \ z \ \geqslant \lambda \end{cases}$$

Exo: Prouver le résultat précédent en utilisant les sous-gradients

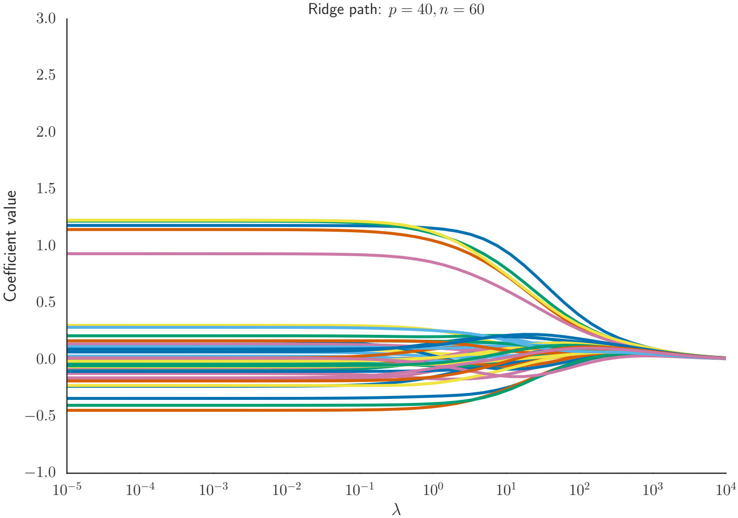
---

Exemple numérique : simulation

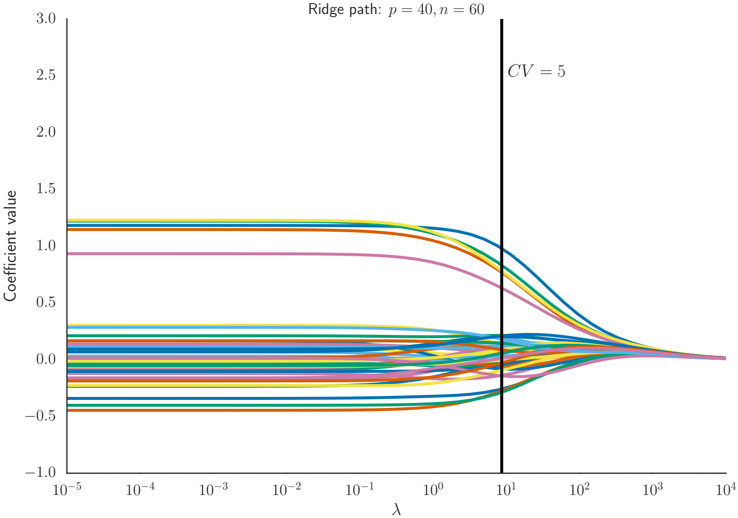
- $\boldsymbol{\theta}^{\star} = (1, 1, 1, 1, 1, 0, \ldots, 0) \in \mathbb{R}^p$ (5 coefficients non-nuls)
- $X \in \mathbb{R}^{n \times p}$ a des colonnes tirées selon une loi gaussienne
- $y = X\boldsymbol{\theta}^{\star} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$ avec $\boldsymbol{\varepsilon} \sim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

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

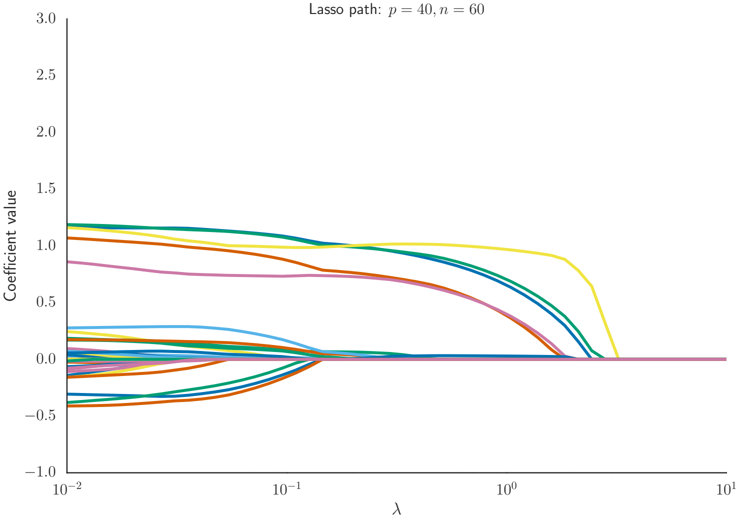
Lasso vs Ridge



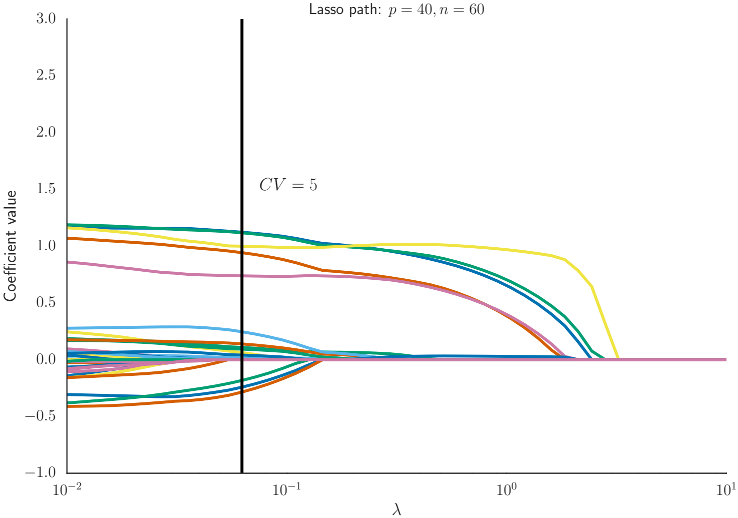
Lasso vs Ridge



Lasso vs Ridge



Lasso vs Ridge



Intérêt du Lasso

- Enjeu numérique : le Lasso est un problème **convexe**
- Sélection de variables/ solutions parcimonieuses (sparse) : $\hat{\boldsymbol{\theta}}_{\lambda}^{\mathrm{Lasso}}$ a potentiellement de nombreux coefficients nuls. Le paramètre λ contrôle le niveau de parcimonie : si λ 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 carrées 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

Le biais du Lasso

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

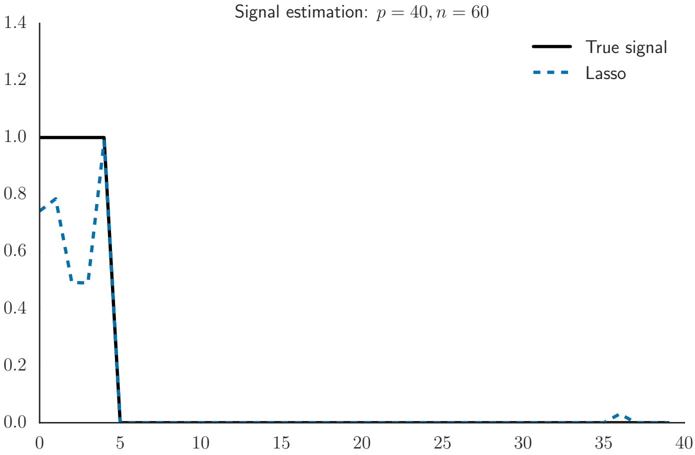


Illustration sur l'exemple

Le biais du Lasso

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

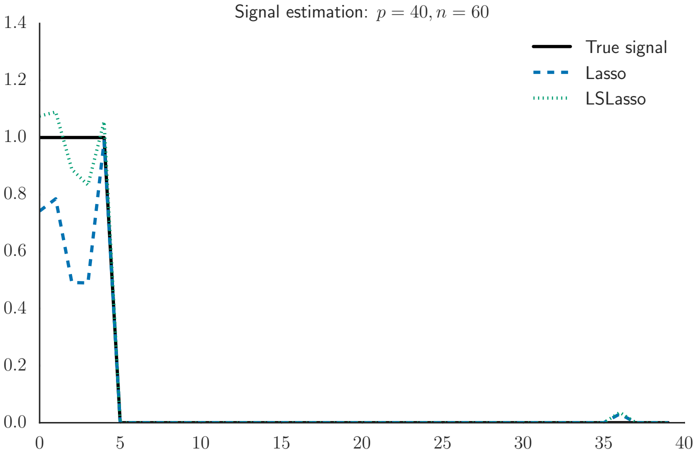


Illustration sur l'exemple

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

LSLasso (Least Square Lasso)

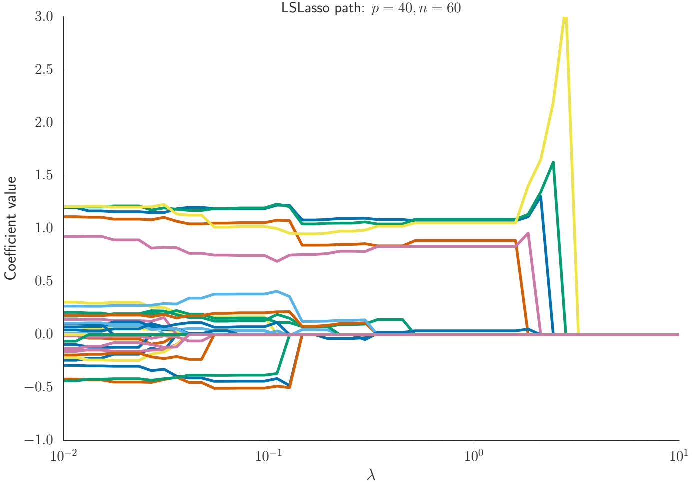
1. Lasso : obtenir $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}}$
2. Moindres-carrés sur les variables actives $\mathrm{supp}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})$

$$\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{LSLasso}} = \mathop{\arg\min}_{\substack{\boldsymbol{\theta}\in\mathbb{R}^p \\ \mathrm{supp}(\boldsymbol{\theta})=\mathrm{supp}(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso}})}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$$

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

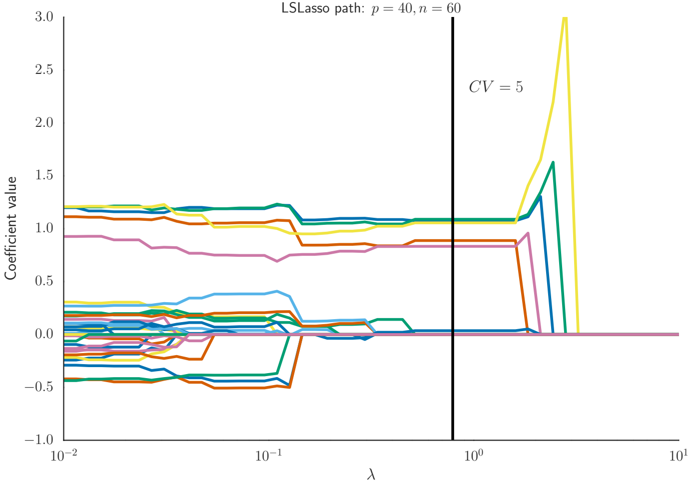
Rem: LSLasso pas forcément codé dans les packages usuels

Débiasage



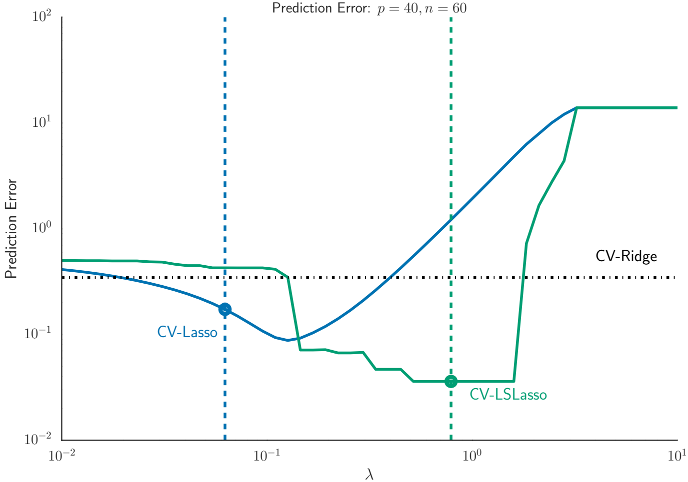
LSLasso path: $p = 40, n = 60$
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Coefficient value
10^{-2}
10^{-1}
10^{0}
10^{1}
λ

Débiasage



LSLasso path: $p = 40, n = 60$
$CV = 5$
Coefficient value
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
10^{-2}
10^{-1}
10^{0}
10^{1}
λ

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 <u>l'estimation</u>

Limites

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

Elastic Net : régularisation ℓ_1/ℓ_2

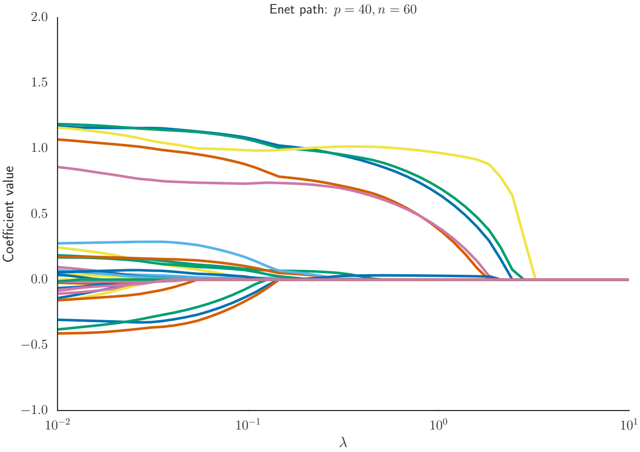
L'Elastic Net introduit par Zou et Hastie (2005) est solution de

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

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

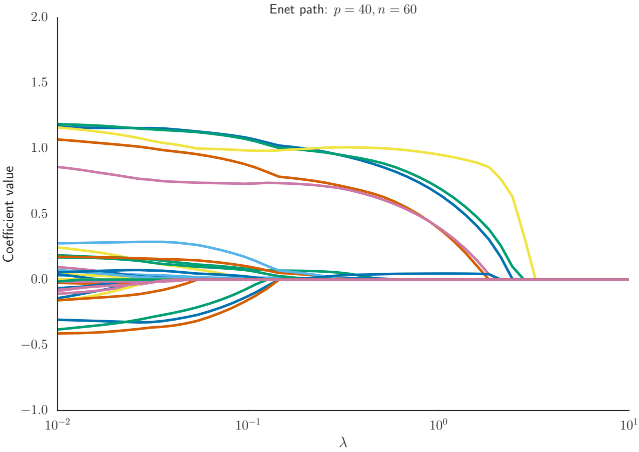
<u>Rem</u>: la solution est unique et la taille du support de l'Elastic Net est plus petite que $\min(n, p)$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



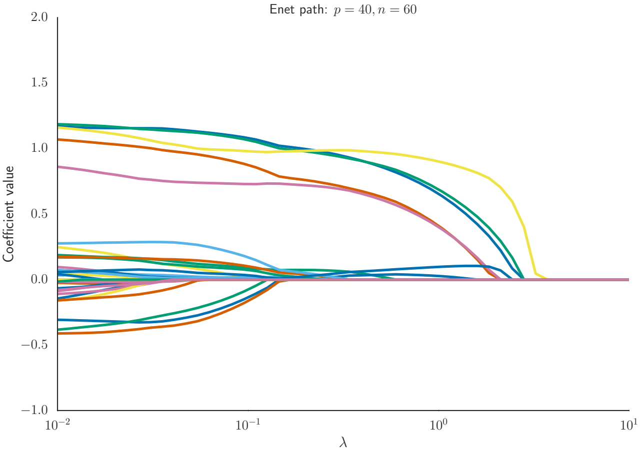
$$\gamma = 1.00$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



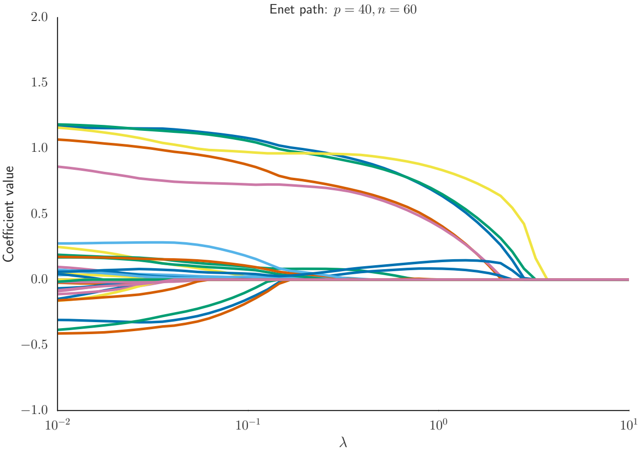
$$\gamma = 0.99$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



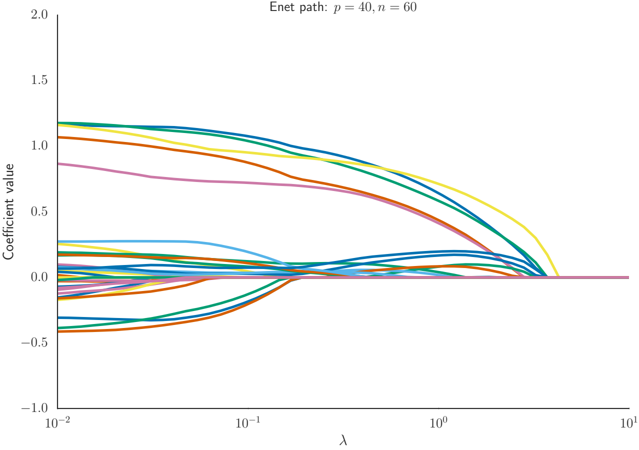
$$\gamma = 0.95$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



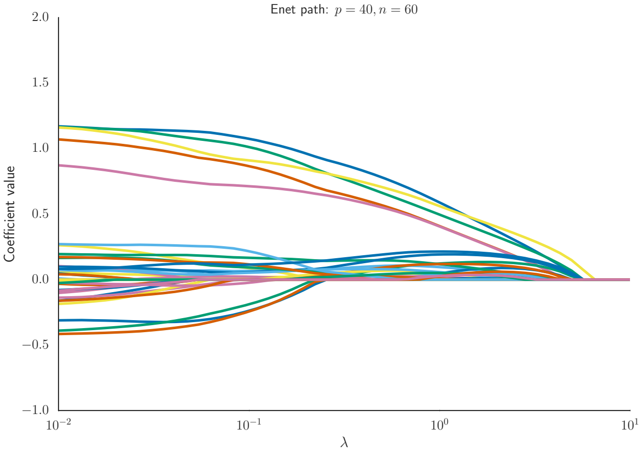
$$\gamma = 0.90$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



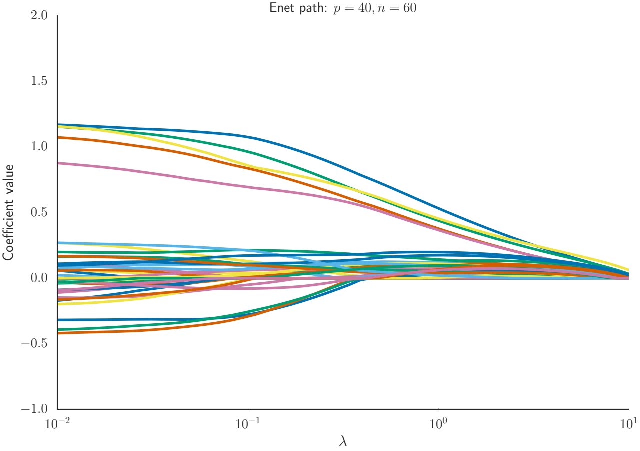
$$\gamma = 0.75$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



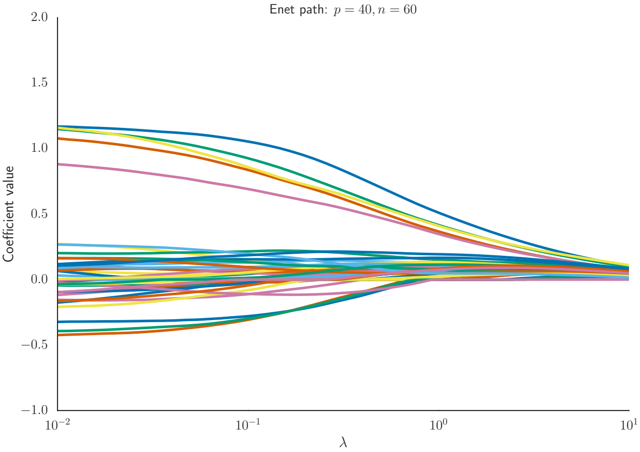
$$\gamma = 0.50$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



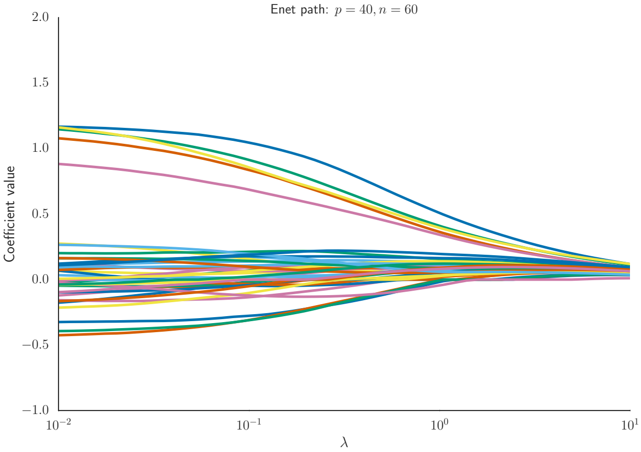
$$\gamma = 0.25$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



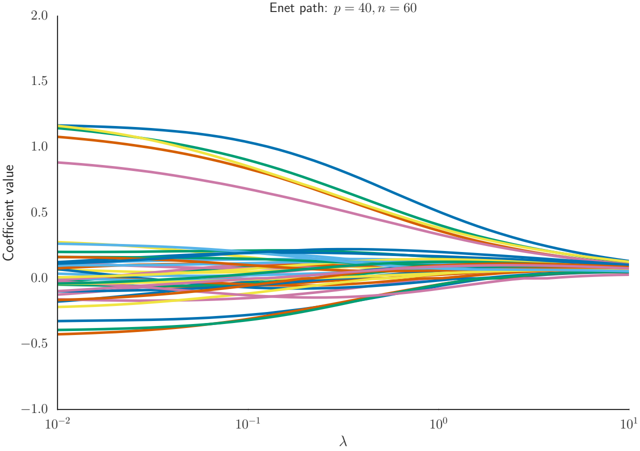
$\gamma = 0.1$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



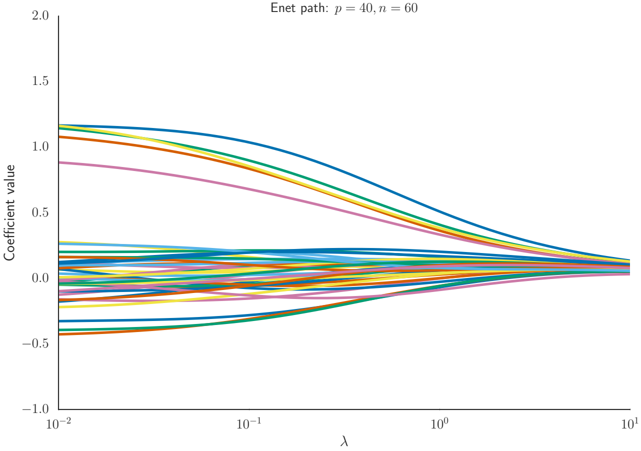
$$\gamma = 0.05$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



$$\gamma = 0.01$$

Elastic-Net : $\gamma\|\boldsymbol{\theta}\|_1 + (1-\gamma)\|\boldsymbol{\theta}\|_2^2/2$



$$\gamma = 0.00$$

Sommaire

Pénalités non-convexes

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

$$\boxed{\hat{\boldsymbol{\theta}}^{\mathrm{pen}}_{\lambda,\gamma} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\textbf{attache aux données}} + \underbrace{\sum_{j=1}^{p} \mathrm{pen}_{\lambda,\gamma}(|\theta_j|)}_{\textbf{régularisation}} \right)}$$

Pénalités non-convexes

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

$$\boxed{\hat{\boldsymbol{\theta}}^{\mathrm{pen}}_{\lambda,\gamma} = \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^{p} \mathrm{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)}$$

- Adaptive-Lasso Zou (2006) / ℓ_1 re-pondérés Candès *et al.* (2008)

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

Pénalités non-convexes

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

$$\boxed{\hat{\boldsymbol{\theta}}^{\mathrm{pen}}_{\lambda,\gamma} = \mathop{\arg\min}_{\boldsymbol{\theta}\in\mathbb{R}^p} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^{p} \mathrm{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)}$$

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

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

Pénalités non-convexes

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

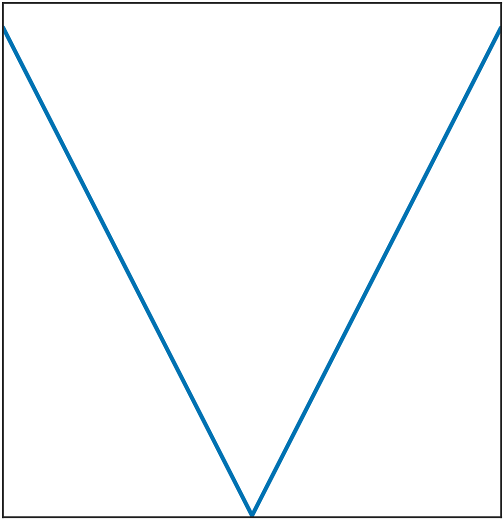
$$\boxed{\hat{\boldsymbol{\theta}}^{\mathrm{pen}}_{\lambda,\gamma} = \underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\arg\min} \left(\underbrace{\frac{1}{2}\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}_{\text{attache aux données}} + \underbrace{\sum_{j=1}^{p} \mathrm{pen}_{\lambda,\gamma}(|\theta_j|)}_{\text{régularisation}} \right)}$$

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

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

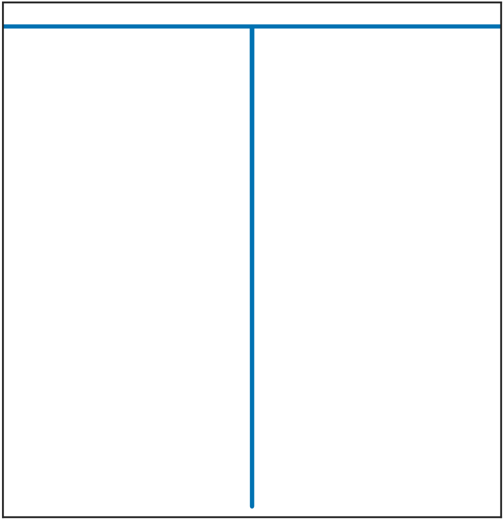
Rem: difficultés algorithmiques (arrêt, minima locaux, etc.)

Forme des pénalités classiques



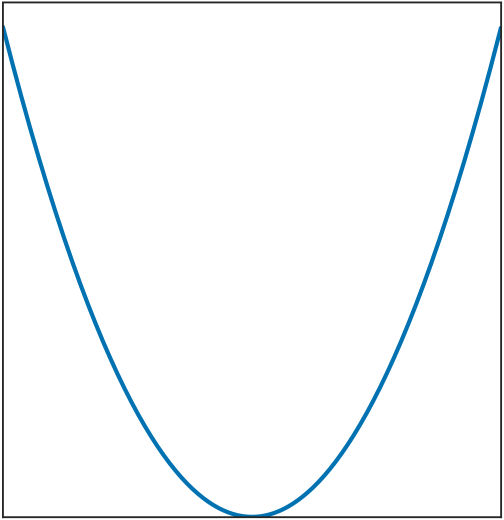
l1

Forme des pénalités classiques



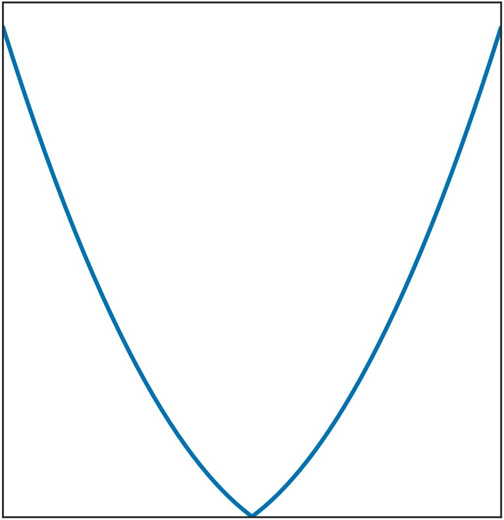
l0

Forme des pénalités classiques



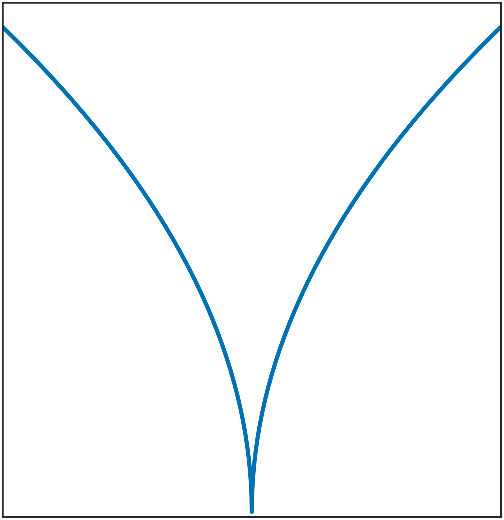
l22

Forme des pénalités classiques



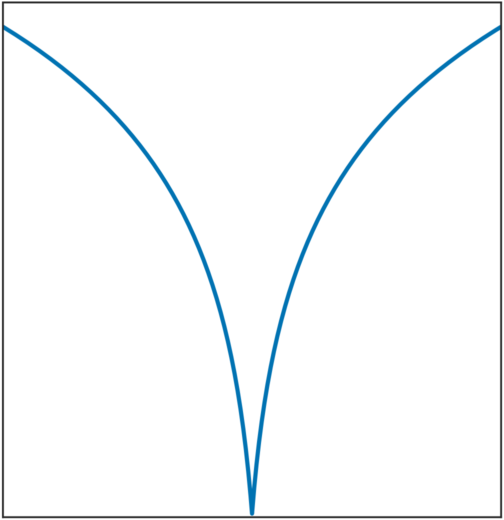
enet

Forme des pénalités classiques



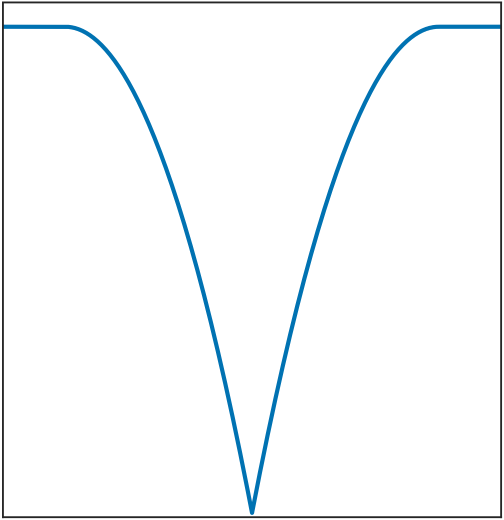
sqrt

Forme des pénalités classiques



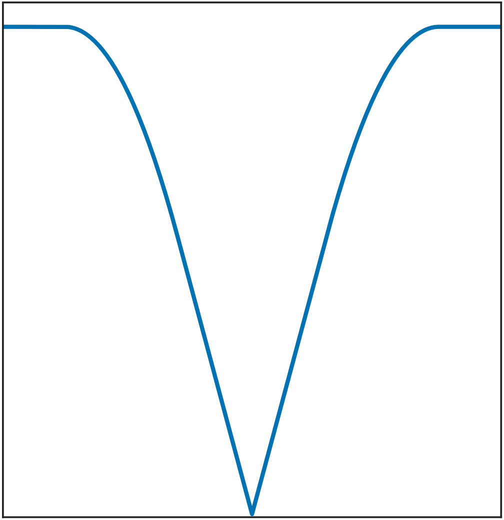
log

Forme des pénalités classiques



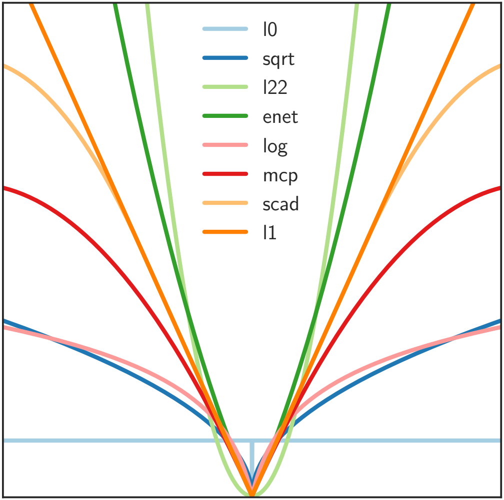
mcp

Forme des pénalités classiques



scad

Forme des pénalités classiques



Adaptive-Lasso

Plusieurs noms pour une même idée :

- Adaptive-Lasso Zou (2006)
- ℓ_1 re-pondérés Candès *et al.* (2008)
- approche DC-programming (pour *Difference of Convex Programming*) Gasso *et al.* (2008)

Adaptive-Lasso

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

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

Entrées : $X, \mathbf{y}$, nombre d'itérations K, régularisation λ
Initialisation : $\hat{w} \leftarrow (1, \ldots, 1)^\top$

Adaptive-Lasso

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

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

Entrées : $X, \mathbf{y}$, nombre d'itérations K, régularisation λ
Initialisation : $\hat{w} \leftarrow (1, \ldots, 1)^\top$
pour $k = 1, \ldots, K$ **faire**

Adaptive-Lasso

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

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

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

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

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

$$\hat{\boldsymbol{\theta}} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} \left(\frac{\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2}{2} + \lambda \sum_{j=1}^{p} \hat{w}_j |\theta_j| \right)$$

Adaptive-Lasso

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

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

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

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

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

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

$$\hat{w}_j \leftarrow \frac{1}{|\hat{\theta}_j|^{\frac{1}{2}}}, \ \forall j \in [\![1, p]\!]$$

Adaptive-Lasso

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

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

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

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

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

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

$$\hat{w}_j \leftarrow \frac{1}{|\hat{\theta}_j|^{\frac{1}{2}}}, \ \forall j \in [\![1, p]\!]$$

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

Adaptive-Lasso

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

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

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

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

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

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

$$\hat{w}_j \leftarrow \frac{1}{|\hat{\theta}_j|^{\frac{1}{2}}}, \ \forall j \in [\![1, p]\!]$$

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

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

Sommaire

Structure du support

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

Vecteur et ses coordonnées actives (en orange) :

Support creux : quelconque

Pénalité envisagée : Lasso

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

Structure du support

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

Vecteur et ses coordonnées actives (en orange) :

Support creux : groupes

Pénalité envisagée : Groupe-Lasso

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

Structure du support

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

Vecteur et ses coordonnées actives (en orange) :

Support creux : groupes + sous groupes

Pénalité envisagée : Sparse-Groupe-Lasso

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

Groupe-Lasso

La pénalisation par la norme ℓ_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

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 Buhlmann (2010)

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

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

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

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

Calculer le Lasso sur cet échantillon : $\hat{\boldsymbol{\theta}}_{\lambda}^{\mathrm{Lasso},(k)}$

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

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

Calculer le Lasso sur cet échantillon : $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso},(k)}$

Calculer le support associé : $S_k = \operatorname{supp}\left(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso},(k)}\right)$

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

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

Calculer le Lasso sur cet échantillon : $\hat{\boldsymbol{\theta}}_{\lambda}^{\mathrm{Lasso},(k)}$

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

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

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

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

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

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

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

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

Exo: coder le Bolasso avec `Python` et `sklearn`

Bolasso Bach (2008)

Algorithme : Bootstrap Lasso

Entrées : $X, \mathbf{y}$, nombre de réplications B, régularisation λ

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

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

Calculer le Lasso sur cet échantillon : $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso},(k)}$

Calculer le support associé : $S_k = \operatorname{supp}\left(\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Lasso},(k)}\right)$

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

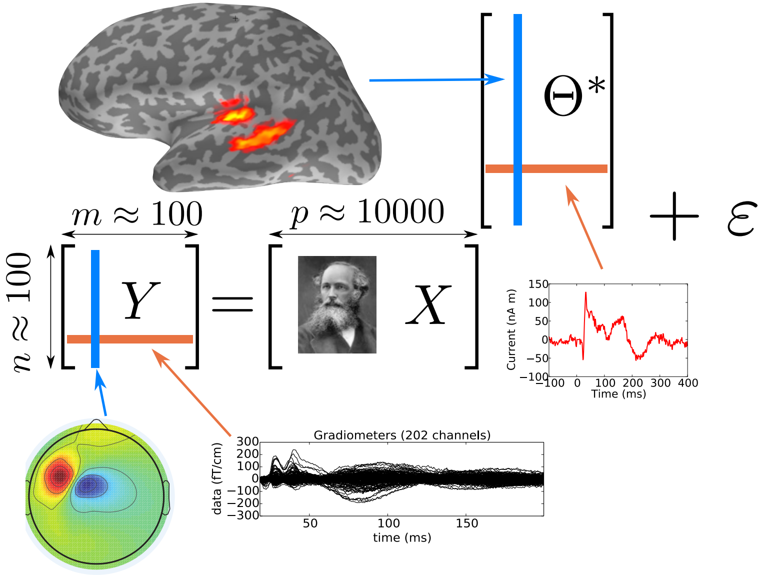
Calculer : $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Bolasso}} \in \underset{\substack{\boldsymbol{\theta} \in \mathbb{R}^p \\ \operatorname{supp}(\boldsymbol{\theta}) = S}}{\arg\min} \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}\|_2^2$

Sorties : un support S, et un vecteur $\hat{\boldsymbol{\theta}}_\lambda^{\mathrm{Bolasso}}$

Exo: coder le Bolasso avec `Python` et `sklearn`

Sommaire

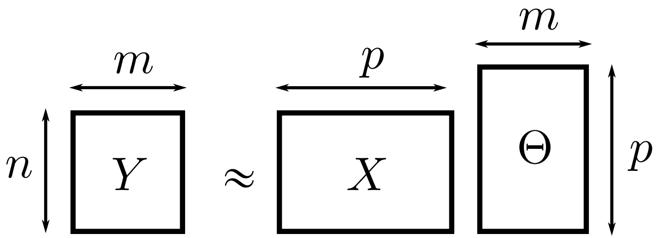
Exemple



$Y = X \Theta^* + \varepsilon$
$m \approx 100$
$p \approx 10000$
$n \approx 100$
Current (nA m)
150
100
50
0
−50
−100
−100
0
100
200
300
400
Time (ms)
Gradiometers (202 channels)
data (fT/cm)
300
200
100
0
−100
−200
−300
50
100
150
time (ms)

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 carres pénalisées

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

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

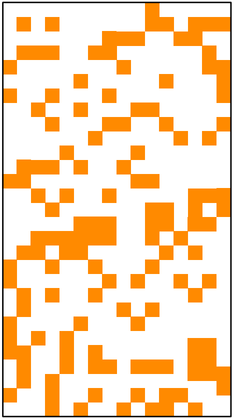
où Ω est une pénalité / régularisation à préciser

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

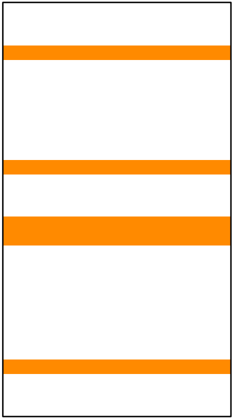
Support creux :
quelconque

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

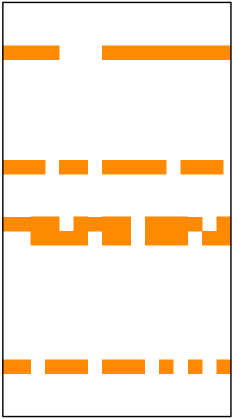
Pénalité Groupe-Lasso :

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

Rem: on note $\Theta_{j,:}$ la j^{e} ligne de Θ

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 :

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

Références I

- F. Bach.
  Bolasso : model consistent Lasso estimation through the bootstrap.
  In *ICML*, 2008.
- D. Bertsimas, A. King, and R. Mazumder.
  Best subset selection via a modern optimization lens.
  Ann. Statist., 44(2) :813–852, 2016.
- P. Bühlmann and S. van de Geer.
  Statistics for high-dimensional data.
  Springer Series in Statistics. Springer, Heidelberg, 2011.
  Methods, theory and applications.
- E. J. Candès, M. B. Wakin, and S. P. Boyd.
  Enhancing sparsity by reweighted l_1 minimization.
  J. Fourier Anal. Applicat., 14(5-6) :877–905, 2008.

Références II

- O. Fercoq, A. Gramfort, and J. Salmon.
  Mind the duality gap : safer rules for the lasso.
  In *ICML*, pages 333–342, 2015.
- J. Fan and R. Li.
  Variable selection via nonconcave penalized likelihood and its oracle properties.
  J. Amer. Statist. Assoc., 96(456) :1348–1360, 2001.
- G. Gasso, A. Rakotomamonjy, and S. Canu.
  Recovering sparse signals with non-convex penalties and DC programming.
  IEEE Trans. Sig. Process., 57(12) :4686–4698, 2009.
- Bien J, J. Taylor, and R. Tibshirani.
  A lasso for hierarchical interactions.
  Ann. Statist., 41(3) :1111–1141, 2013.

Références III

- N. Meinshausen and P. Bühlmann.
  Stability selection.
  Journal of the Royal Statistical Society : Series B (Statistical Methodology), 72(4) :417–473, 2010.
- N. Parikh, S. Boyd, E. Chu, B. Peleato, and J. Eckstein.
  Proximal algorithms.
  Foundations and Trends in Machine Learning, 1(3) :1–108, 2013.
- N. Simon, J. Friedman, T. Hastie, and R. Tibshirani.
  A sparse-group lasso.
  J. Comput. Graph. Statist., 22(2) :231–245, 2013.
- R. Tibshirani.
  Regression shrinkage and selection via the lasso.
  J. Roy. Statist. Soc. Ser. B, 58(1) :267–288, 1996.
- M. Yuan and Y. Lin.
  Model selection and estimation in regression with grouped variables.
  J. Roy. Statist. Soc. Ser. B, 68(1) :49–67, 2006.

Références IV

- H. Zou and T. Hastie.
  Regularization and variable selection via the elastic net.
  J. Roy. Statist. Soc. Ser. B, 67(2) :301–320, 2005.
- C.-H Zhang.
  Nearly unbiased variable selection under minimax concave penalty.
  Ann. Statist., 38(2) :894–942, 2010.
- H. Zou.
  The adaptive lasso and its oracle properties.
  J. Am. Statist. Assoc., 101(476) :1418–1429, 2006.