

Elastic Net

Algorithms et Etude de cas

Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France

Bureau provisoire : 008 du BSHM
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr



Introduction

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

Retour sur le modèle linéaire et Motivation

$$y = X\beta^* + \varepsilon \in \mathbb{R}^n$$

$$X = [X_1, \dots, X_p] = \begin{pmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{pmatrix} \in \mathbb{R}^{n \times p}, \beta^* \in \mathbb{R}^p$$

- 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*

Remarque

Il es aussi utile si β^* 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 **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

Définition

- Le support du vecteur β est l'ensemble des indices des coordonnées non nulles :

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

- La pseudo-norme L^0 d'un vecteur $\beta \in \mathbb{R}^p$ est son nombre de coordonnées non-nulles :

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

Remarque

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

$$\beta_1 = (1, 0, 1, \dots, 0), \beta_2 = (0, 1, 1, \dots, 0)$$

et

$$2 = \frac{\|\beta_1\|_0 + \|\beta_2\|_0}{2} \leq \left\| \frac{\beta_1 + \beta_2}{2} \right\|_0 = 3$$

La pénalisation L^0 et ses limites

Première tentative de méthode pénalisée pour introduire de la parcimonie : utiliser L^0 pour la pénalisation/régularisation

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

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 !

Exemple

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

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

Exemple

Il y a des avancées récentes en MIP^a (cf. Bertsimas et al. 2016)

La pénalisation L^1 : Le Lasso

Lasso : *Least Absolute Shrinkage and Selection Operator* (cf. Tibshirani, 1996)

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

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

On retrouve de nouveau les cas limites :

$$\lim_{\lambda \rightarrow 0} \hat{\beta}_\lambda^{Lasso} = \hat{\beta}^{MCO} \quad \text{et} \quad \lim_{\lambda \rightarrow +\infty} \hat{\beta}_\lambda^{Lasso} = \mathbf{0} \in \mathbb{R}^p$$

Remarque

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

Interprétation de la contrainte

Un problème de la forme :

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

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

admet la même solution qu'une version contrainte :

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

pour un certain $T > 0$.

Remarque

le lien entre T et λ 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 $\hat{\beta}^{\text{MCO}}$ (sans contrainte)

Mise à zéro de certains coefficients

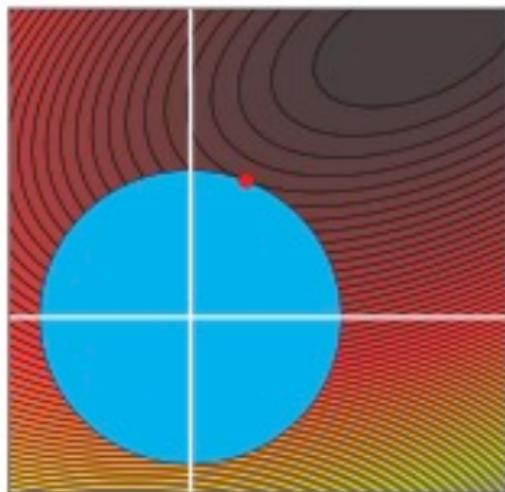


FIGURE – Optimisation sous contrainte L^2 : solution non parcimonieuse

Mise à zéro de certains coefficients

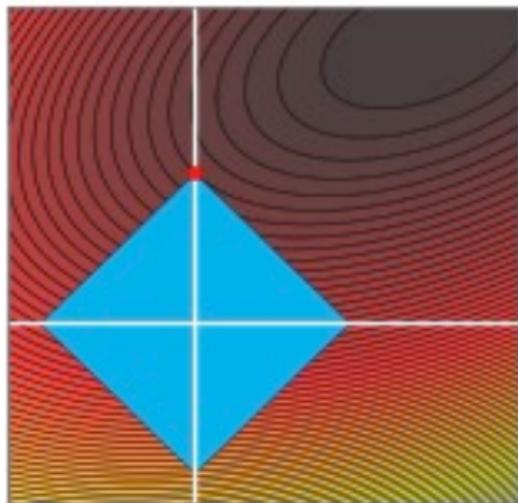


FIGURE – Optimisation sous contrainte L^1 : solution parcimonieuse

Sous-gradients / sous-différentielles

Définition

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 :

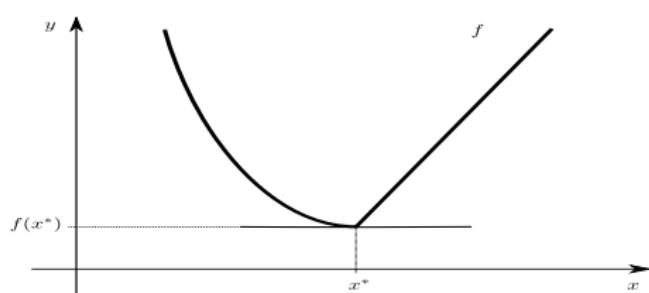
$$f(x) \geq f(x^*) + \langle u, x - x^* \rangle$$

La “sous-différentielle” est l’ensemble des sous-gradients :

$$\partial f(x^*) = \{u \in \mathbb{R}^n : \forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle u, x - x^* \rangle\}.$$

Remarque

Si le sous-gradient est unique, on retrouve le gradient :



Règle de Fermat

Théorème

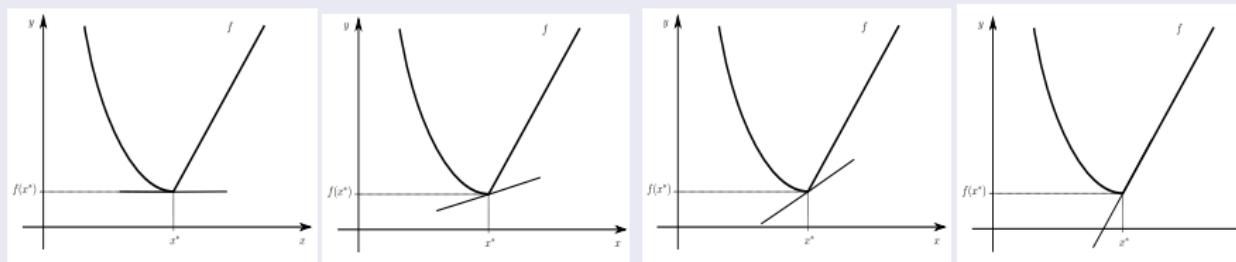
Un point x^* est un minimum d'une fonction convexe $f : \mathbb{R}^n \rightarrow \mathbb{R}$ si, et seulement si, $0 \in \partial f(x^*)$

Pour la preuve, utiliser la définition des sous-gradients :
0 est un sous-gradient de f en x^* si, et seulement si,

$$\forall x \in \mathbb{R}^n, f(x) \geq f(x^*) + \langle u, x - x^* \rangle$$

Remarque

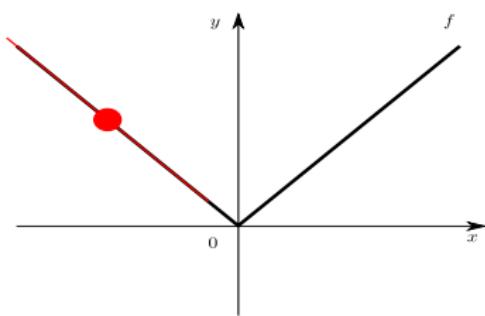
Visuellement cela correspond à une tangente horizontale :



Sous-différentielle de la valeur absolue

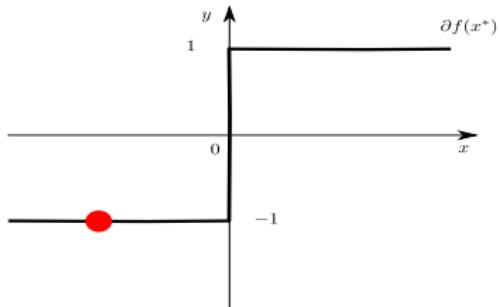
Fonction valeur absolue (abs)

$$f : \begin{cases} \mathbb{R} \longrightarrow \mathbb{R} \\ x \longmapsto |x| \end{cases}$$



Sous-différentielle (sign)

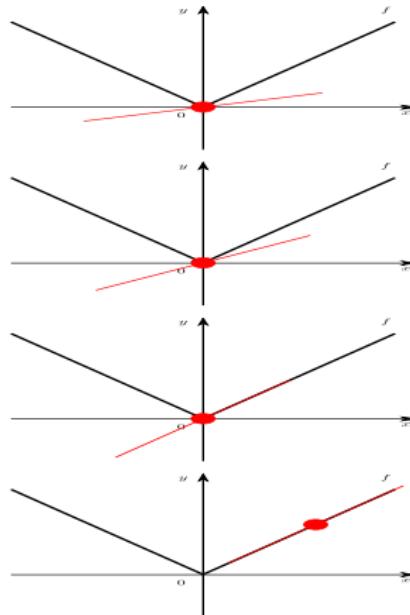
$$\partial f(x^*) = \begin{cases} \{-1\} & \text{si } x^* \in]-\infty, 0[\\ \{1\} & \text{si } x^* \in]0, \infty[\\ [-1,1] & \text{si } x^* = 0 \end{cases}$$



Sous-différentielle de la valeur absolue

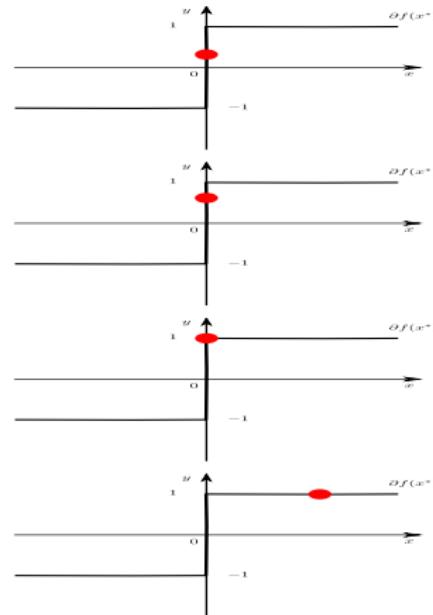
Fonction valeur absolue (abs)

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



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

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

Conditions nécessaires et suffisantes d'optimalité (Fermat) :

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

Remarque

Si $\lambda > \lambda_{\max} := \max_{j=1, \dots, p} | \langle X_j, y \rangle |$, alors $\hat{\beta}_\lambda^{Lasso} = 0$.

La preuve de cette remarque consiste à vérifier les conditions ci-dessus pour 0 et $\lambda > 0$.

Le cas orthogonal : le seuillage doux

Retour sur un cas simple (design orthogonal) : ${}^tXX = Id_p$

$$\|y - X\beta\|_2^2 = \|{}^tXy - {}^tXX\beta\|_2^2 = \|{}^tXy - \beta\|_2^2$$

car X est une isométrie dans ce cas, la fonction objective du Lasso devient :

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

Problème de séparabilité/séparable : problème qui revient à minimiser terme à terme en séparant les termes la somme. Il faut donc minimiser :

$$x \longmapsto \frac{1}{2}(z - x)^2 + \lambda|x| \text{ pour } z = {}^tX_j y$$

Remarque

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

Régularisation en 1D : Ridge

Résoudre :

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

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

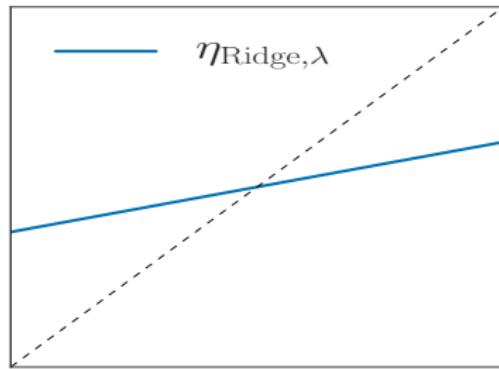


FIGURE – Contraction L^2 : Ridge

Régularisation en 1D : Lasso

Résoudre :

$$\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} \frac{1}{2} (z - x)^2 + \lambda |x|$$
$$\eta_\lambda(z) = sign(z)(|z| - \lambda) \quad (\text{Exercice})$$

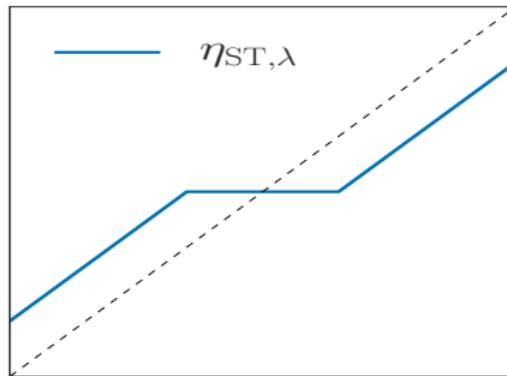


FIGURE – Contraction L^1 : Seuillage doux (Soft thresholding)

Régularisation en 1D : L^0

Résoudre :

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

$$\eta_\lambda(z) = z \mathbf{1}_{|z| \geq \sqrt{2\lambda}} \quad (\text{Exercice})$$

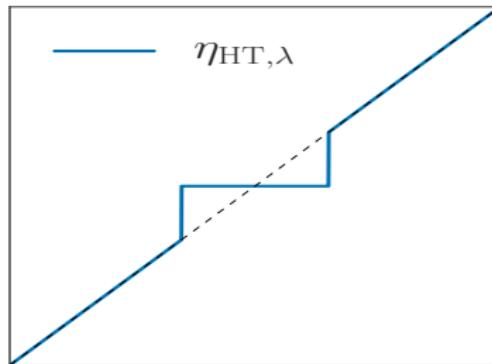


FIGURE – Contraction L^0 : Seuillage dur (Hard thresholding)

Régularisation en 1D : Elastic Net

Résoudre :

$$\eta_\lambda(z) = \arg \min_{x \in \mathbb{R}} \frac{1}{2}(z - x)^2 + \lambda \left(\alpha|x| + (1 - \alpha)\frac{x^2}{2} \right)$$
$$\eta_\lambda(z) = ?? \quad (\text{Exercice})$$

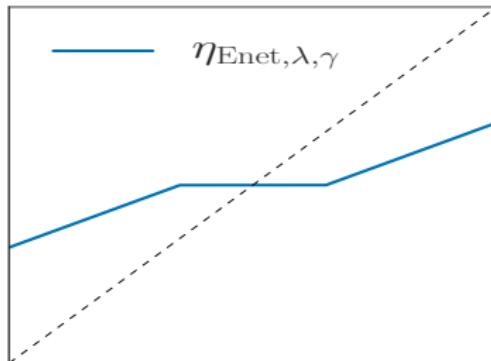


FIGURE – Contraction L^1/L^2

Seuillage doux : forme explicite

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

- On peut prouver ce résultat 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
- $y = {}^t X \beta^* + \varepsilon \in \mathbb{R}^n$ avec $\varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n)$
- On utilise une grille de 50 valeurs de λ

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

Lasso vs Ridge

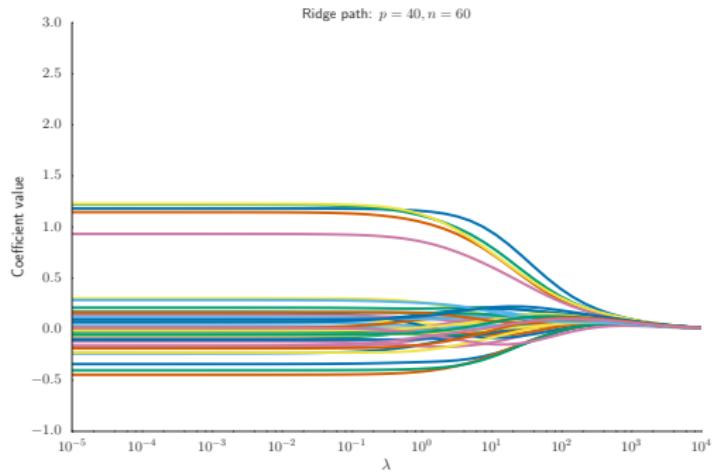


FIGURE – Ridge path : $p = 40, n = 60$

Lasso vs Ridge

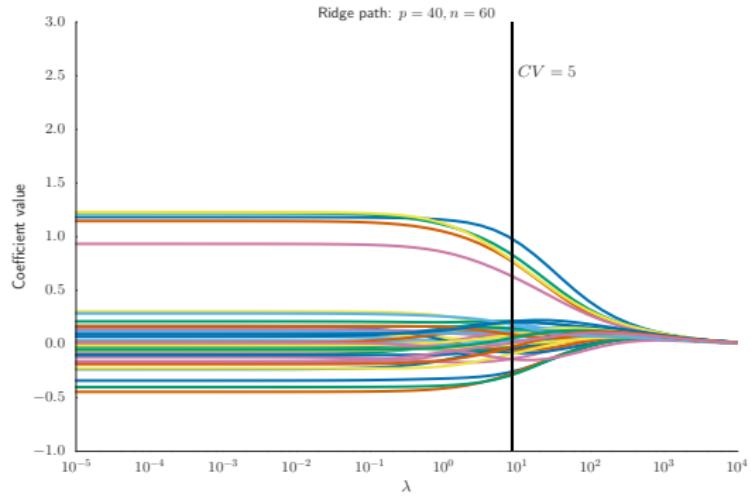


FIGURE – Ridge path : $p = 40, n = 60, CV = 5$

Lasso vs Ridge

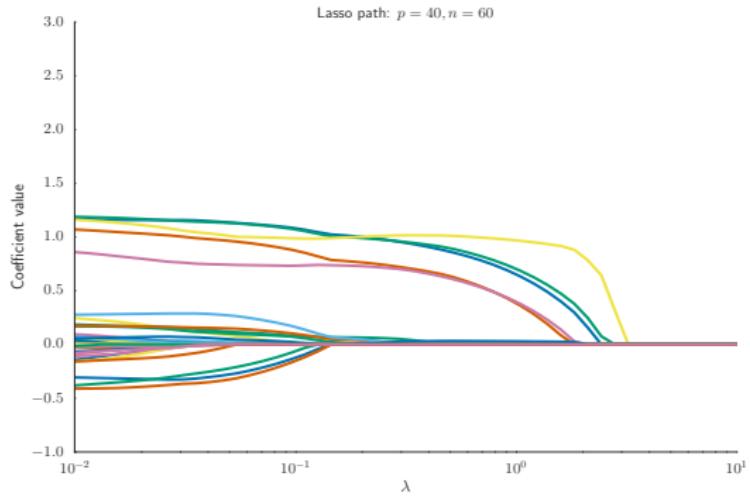


FIGURE – Ridge path : $p = 40, n = 60$

Lasso vs Ridge

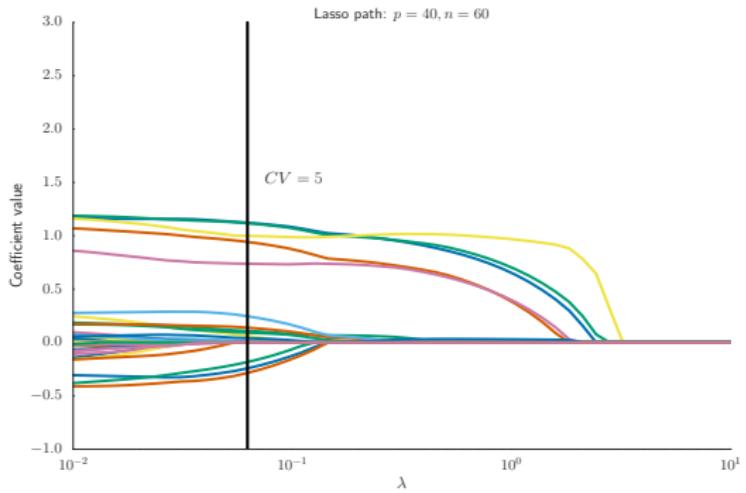


FIGURE – Ridge path : $p = 40, n = 60, CV = 5$

- 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 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

Remarque

RidgeCV n'avait aucun coefficient nul

Analyse de l'estimateur dans le cas général

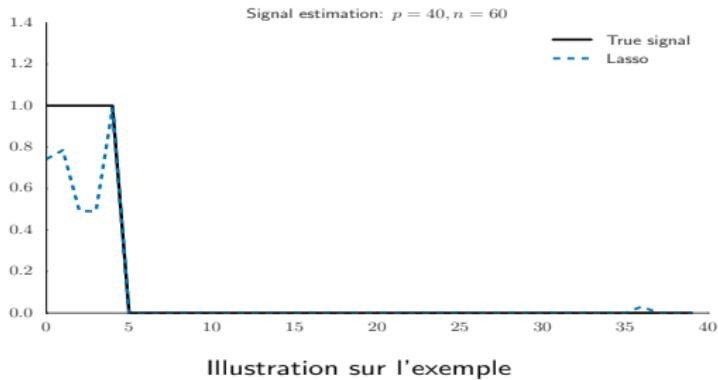
- *Analyse théorique* : (nettement) plus poussée que pour les moindres carrées ou que pour Ridge ; peut être trouvée dans des références récentes (cf. Bühlmann et van de Geer (2011) pour des résultats théoriques)
- *En résumé* : on biaise l'estimateur des moindres carrés pour réduire la variance

Améliorations et extensions du Lasso

LSLasso / Elastic-Net

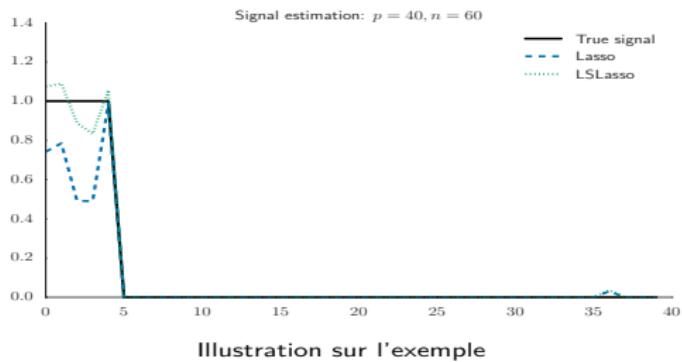
Le biais du Lasso

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



Le biais du Lasso

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



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 $\widehat{\beta}_\lambda^{\text{Lasso}}$*

② *Moindres-carrés sur les variables actives $\text{supp}(\widehat{\beta}_\lambda^{\text{Lasso}})$*

$$\widehat{\beta}_\lambda^{\text{LSLasso}} = \arg \min_{\beta \in \mathbb{R}^p} \frac{1}{2} \|y - X\beta\|_2^2$$

$$\text{supp}(\beta) = \text{supp}(\widehat{\beta}_\lambda^{\text{Lasso}})$$

Remarque

- *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*
- *LSLasso n'est pas forcément codé dans les packages usuels*

Débiasage

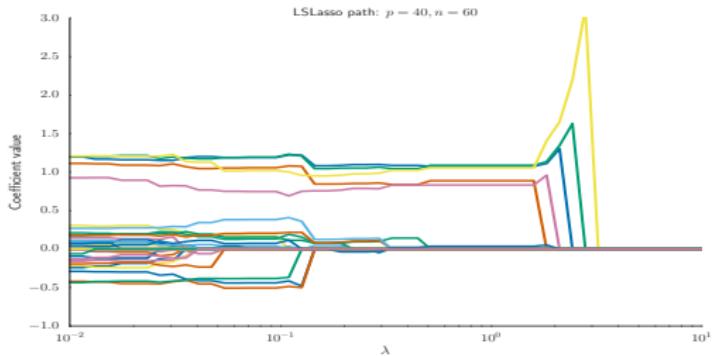


FIGURE – LSLasso pour l'exemple

Débiasage

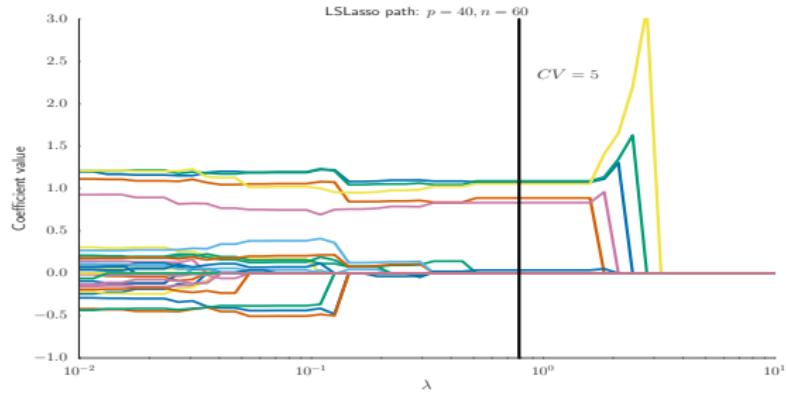


FIGURE – LSLasso pour l'exemple pour $CV = 5$

Prédiction : Lasso vs. LS Lasso

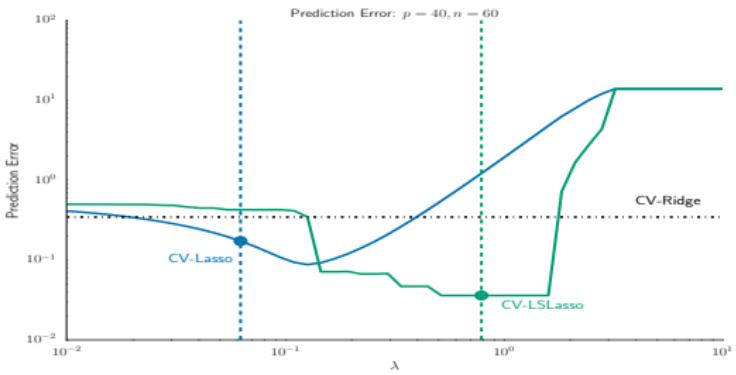


FIGURE – Erreur de prédiction

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 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 :

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

Remarque

- *Deux paramètres de régularisation, un pour la régularisation globale, un qui contrôle l'influence Ridge vs. Lasso*
- *La solution est unique et la taille du support de l'Elastic Net est plus petite que $\min(n, p)$*

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

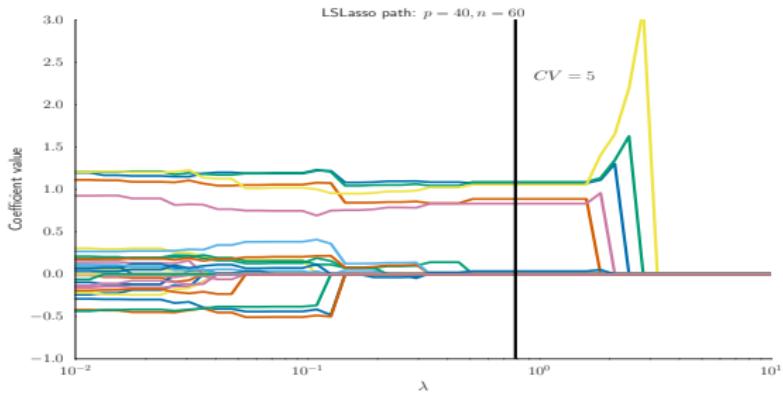


FIGURE — $\alpha = 1.00$

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

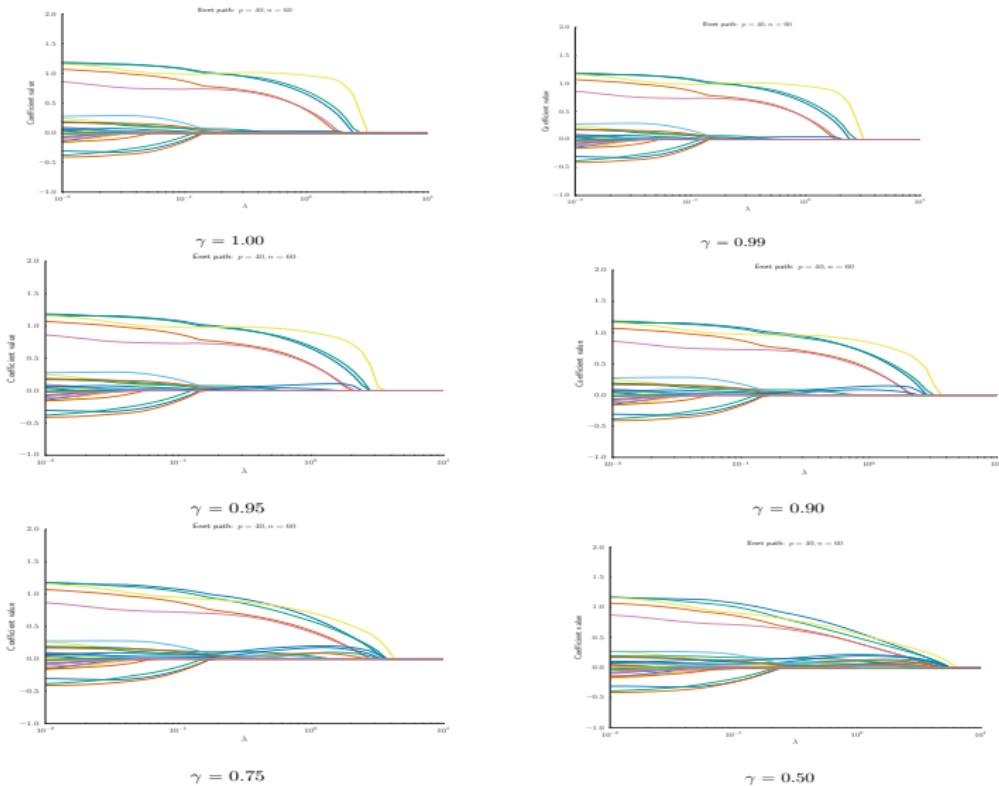


FIGURE – $\alpha = 1.00, 0.99, 0.95, 0.90, 0.75, 0.50, 0.25, 0.1$

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

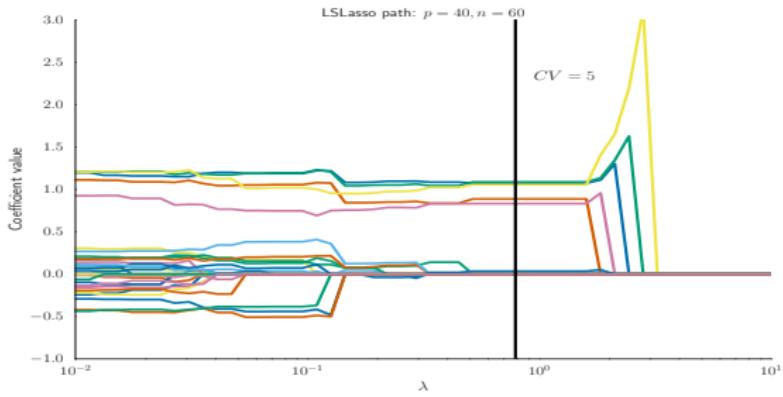


FIGURE — $\alpha = 1.00$

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

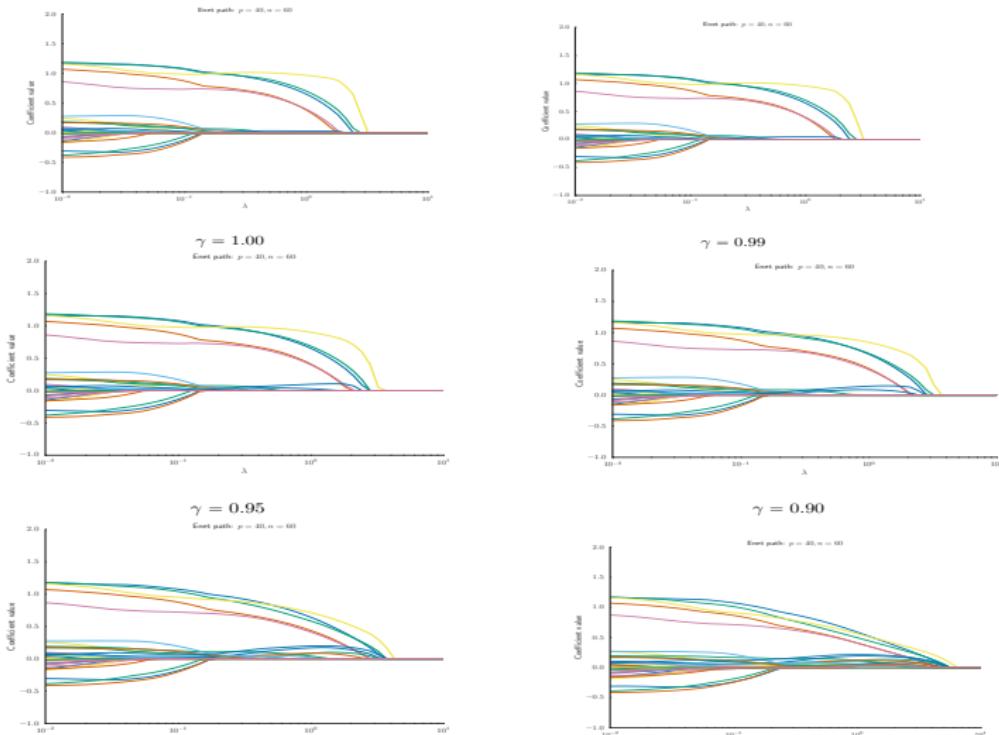


FIGURE – $\alpha = 1.00, 0.99, 0.95, 0.90, 0.75, 0.50$

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

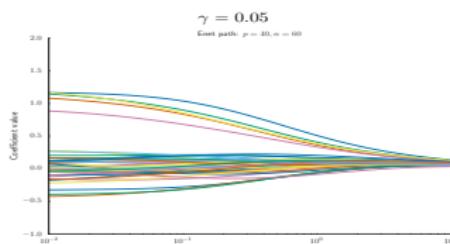
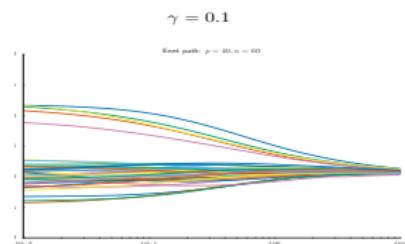
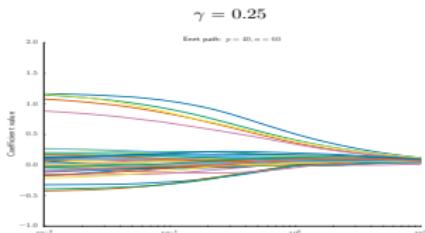
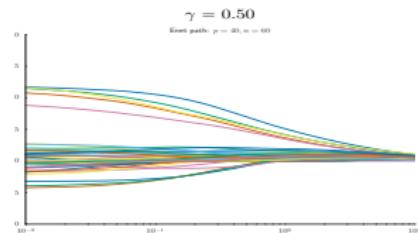
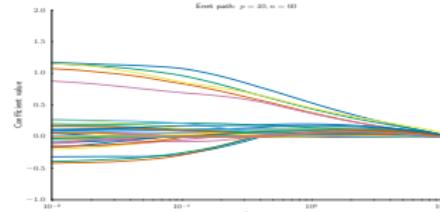
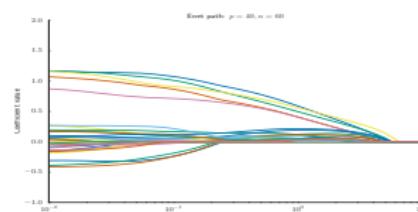


FIGURE – $\alpha = 0.25, 0.1, 0.05, 0.01, 0.00$

Pénalités non-convexes/Adaptive Lasso

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

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

Remarque

- Adaptive-Lasso (cf. Zou, 2006))/ L^1 re-pondérés (cf. Candès et al., 2008)

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

- MCP (minimax concave penalty) (cf. Zhang, 2010) pour $\lambda > 0$ et $\alpha > 1$

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

Pénalités non-convexes/Adaptive Lasso

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

$$\hat{\beta}_{\lambda,\alpha}^{\text{pen}} = \arg \min_{\beta \in \mathbb{R}^p} \left[\frac{1}{2} \|y - X\beta\|_2^2 + \sum_{j=1}^p \text{pen}_{\lambda,\alpha}(|\beta_j|) \right]$$

Remarque

- SCAD (Smoothly Clipped Absolute Deviation) (cf. Fan et Li, 2001) pour $\lambda > 0$ et $\alpha > 2$

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

Ceci présente des difficultés algorithmiques (arrêt, minima locaux, etc.)

Forme des pénalités classiques

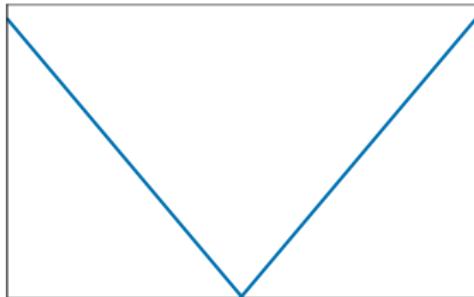


FIGURE – L^1

Forme des pénalités classiques

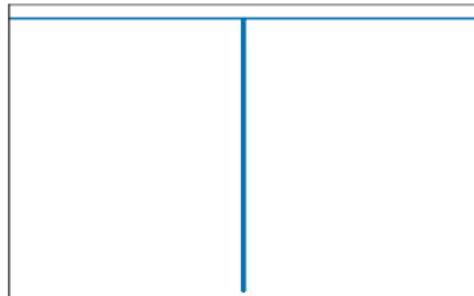


FIGURE – L^0

Forme des pénalités classiques

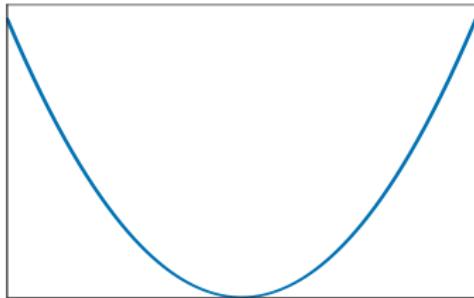


FIGURE – L^2

Forme des pénalités classiques

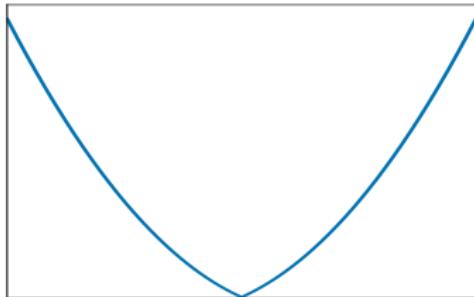


FIGURE – enet

Forme des pénalités classiques

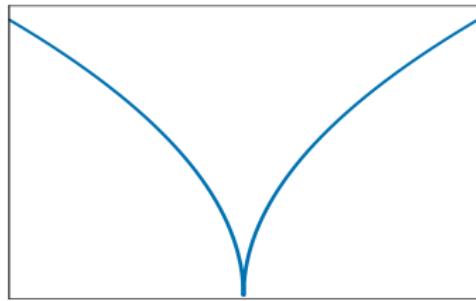


FIGURE – sqrt

Forme des pénalités classiques

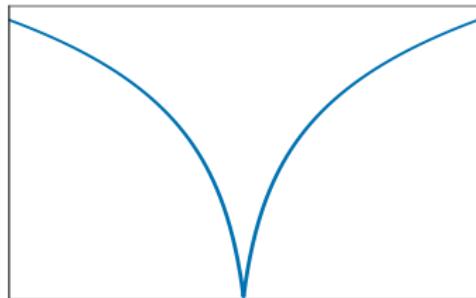


FIGURE – log

Forme des pénalités classiques

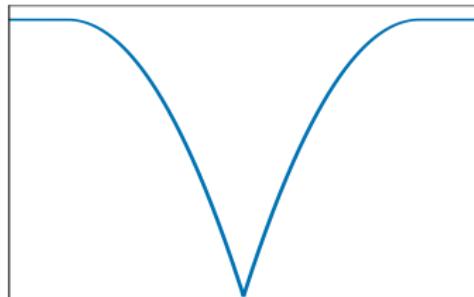


FIGURE – mcp

Forme des pénalités classiques

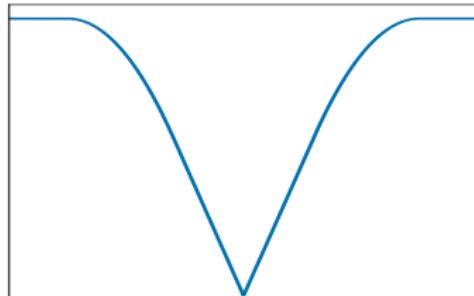
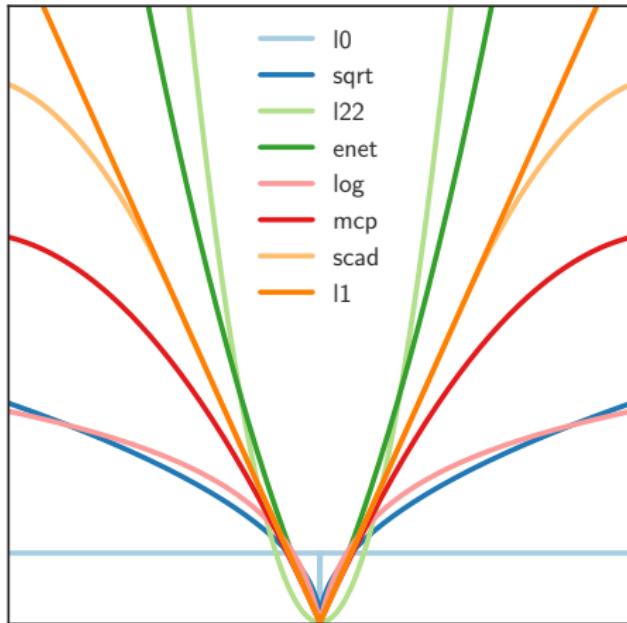


FIGURE – SCAD

Forme des pénalités classiques : Vue globale



Plusieurs noms pour une même idée :

- Adaptive-Lasso (cf. Zou (2006))
- L^1 re-pondérés (cf. Candès et al. (2008))
- Approche DC-programming (pour *Difference of Convex Programming*) (cf. Gasso et al. (2008))

Adaptive-Lasso

Exemple

Prendre $\text{pen}_{\lambda,\alpha}(t) = \lambda |t|^q$ avec $q = 1/2$

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

- Entrée : $X, y, \text{nombre d'itérations } K, \text{régularisation } \lambda$
Initialisation : $\hat{w} \leftarrow t(1, \dots, 1)$
- pour $k = 1, \dots, K$ faire

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

$$\hat{w}_j \leftarrow \frac{1}{|\hat{\beta}_j|^{1/2}}, \forall j = 1, \dots, p$$

Remarque

- En pratique, on n'a pas besoin de beaucoup d'itérations (5 itérations)
- Utiliser un solveur Lasso pour mettre à jour $\hat{\beta}$

Structure du support

On suppose ici que l'on connaît une structure de groupes sur les variables au préalable de l'étude :

$$\{1, \dots, p\} = \cup_{g \in \mathcal{G}} g$$

Vecteur et ses coordonnées actives (en orange) :



FIGURE – Support creux : quelconque

Pénalité envisagée : Lasso

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

Structure du support

On suppose ici que l'on connaît une structure de groupes sur les variables au préalable de l'étude :

$$\{1, \dots, p\} = \cup_{g \in \mathcal{G}} g$$

Vecteur et ses coordonnées actives (en orange) :



FIGURE – Support creux : groupes

Pénalité envisagée : Groupe-Lasso

$$\|\beta\|_{2,1} = \sum_{g \in \mathcal{G}}^p \|\beta_g\|_2$$

Structure du support

On suppose ici que l'on connaît une structure de groupes sur les variables au préalable de l'étude :

$$\{1, \dots, p\} = \cup_{g \in \mathcal{G}} g$$

Vecteur et ses coordonnées actives (en orange) :



FIGURE – Support creux : groupes + sous groupes

Pénalité envisagée : Sparse-Groupe-Lasso

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

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.

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)

- **Algorithme** : Bootstrap Lasso

Entrées : X , 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 : $\hat{\beta}_\lambda^{Lasso,(k)}$

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

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

Calculer :

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

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

D'autres extensions des moindres carrés/ Lasso sont possibles ..

- Régression multi-tâches
- Moindre carres pénalisées
- Pénalisation pour le cas multi-tâches
- ...