I. Principe du Lasso

1. Principe

Variables explicative : $X \in \mathbb{R}^{n\times p}$

Variable expliquée : $y \in \mathbb{R}^n$

Modèle de régression linéaire : Trouver β tel que $y = X\beta + \varepsilon$

$$\min_{\beta\in\mathbb{R}^p} \sum \varepsilon_i^2 = \|\varepsilon\|^2 = \|y - X\beta\|^2$$

Problème quand $p \gg n \Rightarrow$ Hypothèse : Il y a beaucoup de $\beta_j = 0 \Leftrightarrow \|\beta\|_0 \le k$[1]. Cependant, $\|\beta\|_0 \le k$ est une contrainte non-convexe qu'il est difficile d'optimiser, on choisit donc la norme $\|\beta\|_1 \le k$ qui est la plus petite norme entrainant une contrainte convexe.

$$\boxed{\begin{cases} \min\limits_{\beta\in\mathbb{R}^p} & \|y - X\beta\|^2 \\ \text{s.c.} & \|\beta\|_1 \le k \end{cases}} \Leftrightarrow \begin{cases} \min\limits_{\beta\in\mathbb{R}^p} & \frac{1}{2}\beta^\top X^\top X\beta - y^\top X\beta \\ \text{s.c.} & \sum\limits_{j=1}^{p} |\beta_j| \le k \end{cases} \Leftrightarrow \begin{cases} \min\limits_{(\beta^+,\beta^-)\in\mathbb{R}^p} & \frac{1}{2}\beta^\top X^\top X\beta - y^\top X\beta \\ \text{s.c.} & (\beta^+ + \beta^-)^\top e \le k \\ & \beta^+, \beta^- \ge 0 \\ \text{avec} & \beta = \beta^+ - \beta^- \end{cases}$$

Le problème est équivalent pour λ fixé à :

$$\boxed{\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1}$$

2. Front de Pareto

Tirer au hasard des valeurs de β, calculer les normes de $\|y - X\beta\|^2 + \lambda\|\beta\|_1$ (optim multi-critère)

II. Chemin de régularisation du Lasso

L'évolution de β est linéaire par morceaux, entre 2 apparitions d'une composante non nulle.

$$J(\beta) = \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$

$$\operatorname*{argmin}_{\beta} J(\beta) \Leftrightarrow \beta^* \ t.q.\ \nabla J(\beta^*) = 0$$

1. Gradient de $\frac{1}{2}\|y - X\beta\|^2$

$$\nabla_\beta\left(\frac{1}{2}\|y - X\beta\|^2\right) = \nabla_\beta(\|y\|^2 - 2\beta^\top X^\top y + \beta^\top X^\top X\beta) = X^\top X\beta - X^\top y$$

Si $\lambda = 0$, $\hat{\beta}_{MC}(\lambda = 0) = (X^\top X)^{-1}(X^\top y)$ (au sens des moindres carrés)

2. Gradient de $\|\beta\|_1$

On veut calculer un gradient d'une forme contenant des valeurs absolues, mais la dérivée de la valeur absolue n'est pas définie en 0.

On va donc utiliser la notion de sous-différentielle.

[1] $\|\cdot\|_0$ vaut le nombre de termes non nuls

a. Sous-gradient

$g \in \mathbb{R}^n$ est un sous-gradient de J en x_0 si $\forall x \in \mathcal{V}(x_0)\ J(x) \geq J(x_0) + g^\top(x - x_0)$

Interprétation : On peut prendre la dérivée de n'importe quelle droite comprise sous la courbe initiale au voisinage de x_0.

b. Sous-différentielle $\partial_x J(x^*)$

Ensemble des sous gradients possibles

c. Solution x^* d'une minimisation d'un cout convexe J

SI J est différentiable alors $x^* = \operatorname{argmin}_x J(x) \Leftrightarrow \nabla J(x^*) = 0$

Si J est non-différentiable alors x^* est solution si $0 \in \partial_x J(x^*)$ $\quad (\Leftrightarrow \exists \alpha \in \partial_x J(\alpha) | g_\alpha = 0)$

d. Application à J

- *Sous différentielle de $J_1(x) = |x|$*

$$\partial_x J_1(x = 0) = \{\gamma | J_1(x) \geq J_1(0) + \gamma x\} = \{\gamma |\ |x| \geq \gamma x\}$$

Si $x > 0, \gamma \leq 1$. Si $x < 0, \gamma \geq -1. \Rightarrow \gamma \in \left[-1\,;1\right]$

- *Sous différentielle de J*

$$\partial_\beta J(\beta) = X^\top X\beta - X^\top y + \lambda v \text{ avec } v_i = \begin{cases} sign(\beta_j) & \text{si } \beta_j \neq 0 \\ \alpha_j \in [-1;1] & \text{si } \beta_j = 0 \end{cases}$$

Soit $I_\beta = \{j | \beta_j \neq 0\}, I_0 = \{j | \beta_j = 0\}$

On veut : $\exists \alpha_j$ t.q. $\partial_\beta J(\beta_\lambda^*) = 0 \Leftrightarrow 0 = X^\top X\beta - X^\top y + \lambda v$

Pour $j \in I_\beta$:

$$\left[X(:,I_\beta)^\top X(:,I_\beta)\right]\beta(I_\beta) + \lambda \operatorname{sign} \beta(I_\beta) - X(:,I_\beta)^\top y = 0$$

$$\Leftrightarrow \boxed{\left[X(:,I_\beta)^\top X(:,I_\beta)\right]\beta(I_\beta) = X(:,I_\beta)^\top y - \lambda \operatorname{sign} \beta(I_\beta)}$$

3. Calcul du chemin de régression

a. Initialisation

- $\lambda = 0$

$$J(\beta) = \frac{1}{2}\|y - X\beta\|^2 \Rightarrow X^\top X\widehat{\beta_0} = X^\top y$$

- $\lambda \in V(0)$

$$X^\top X\beta_\lambda = X^\top y - \lambda sign\left(\widehat{\beta_0}\right)$$

On a donc quand on fait varier λ :

$$\beta_\lambda = \widehat{\beta_0} - \lambda \underbrace{(X^\top X)^{-1} sign(\widehat{\beta_0})}_{v}$$

b. Process intératif pour l'étape k

On part des MC et on cherche le λ qui va annuler un β_j

Soit $\lambda_k, \widehat{\beta_k}$

Pour $\lambda \in V(\lambda_k) > \lambda_k$,

$\left[X(:,I_\beta)^\top X(:,I_\beta)\right]\beta_\lambda = X(:,I_\beta)^\top y - \lambda \operatorname{sign}\widehat{\beta_k}$ (1)

Vrai pour λ_k

$\left[X(:,I_\beta)^\top X(:,I_\beta)\right]\beta_k = X(:,I_\beta)^\top y - \lambda_k \operatorname{sign}\widehat{\beta_k}$ (2)

(1) – (2) :

$$\boxed{\beta_\lambda = \beta_k - (\lambda - \lambda_k)\left(\left[X(:,I_\beta)^\top X(:,I_\beta)\right]^{-1}\operatorname{sign}\widehat{\beta_k}\right)} \qquad \boxed{\lambda_{k+1} = \beta_{k\,j} + \frac{\lambda_k v_i}{v_i} = \lambda_k + \frac{\beta_{k_j}}{v_j}}$$

III. Lasso « component wise » (à la Gauss-Seidel)

1. Introduction

Pour p = 1

$$\min_{\beta\in\mathbb{R}} \frac{1}{2}\sum_{i=1}^{n}(x_i\beta - y_i)^2 + \lambda|\beta|$$

Si $\lambda = 0$, on a le problème des moindres carrés :

$$\min_{\beta\in\mathbb{R}} J_{MC} = \frac{1}{2}\sum_{i=1}^{n}(x_i\beta - y_i)^2$$

$$\frac{dJ_{MC}}{d\beta} = \sum_{i=1}^{n}(x_i\hat\beta - y_i)x_i = 0 \Rightarrow \hat\beta\sum_{i=1}^{n}x_i^2 + \sum_{i=1}^{n}x_i y_i = 0 \Rightarrow \hat\beta_{MC} = \frac{\sum_{i=1}^{n}x_i y_i}{\sum_{i=1}^{n}x_i^2} = \frac{x^\top y}{\|x\|^2}$$

Pour tout λ, on a le Lasso a une variable :

$$\min_{\beta\in\mathbb{R}} J_{MC} = \frac{1}{2}\sum_{i=1}^{n}(x_i\beta - y_i)^2 + \lambda|\beta|$$

$$\partial_\beta J_{MC} = \sum_{i=1}^{n}(x_i\hat\beta_\lambda - y_i)x_i + \lambda\begin{Bmatrix} \operatorname{sign}\hat\beta_\lambda & \text{si } \beta \neq 0 \\ \alpha \in [-1;1] & \text{si } \beta = 0 \end{Bmatrix} = 0$$

Si $\hat\beta_\lambda \neq 0$:

$$\partial_\beta J_{MC} = \sum_{i=1}^{n}(x_i\hat\beta_\lambda - y_i)x_i + \lambda \operatorname{sign}\hat\beta_\lambda = 0$$

$$\hat\beta_\lambda = \frac{x^\top y}{\|x\|^2} - \lambda\frac{\operatorname{sign}\hat\beta_\lambda}{\|x\|^2} = \hat\beta_{MC} - \lambda\frac{\operatorname{sign}\hat\beta_\lambda}{\|x\|^2} = \begin{Bmatrix} \hat\beta_{MC} - \dfrac{\lambda}{\|x\|^2} & \text{si } \hat\beta_{MC} > 0 \\ \hat\beta_{MC} + \dfrac{\lambda}{\|x\|^2} & \text{si } \hat\beta_{MC} < 0 \end{Bmatrix} = \operatorname{sign}\hat\beta_\lambda\left(|\hat\beta_{MC}| - \frac{\lambda}{\|x\|^2}\right)$$

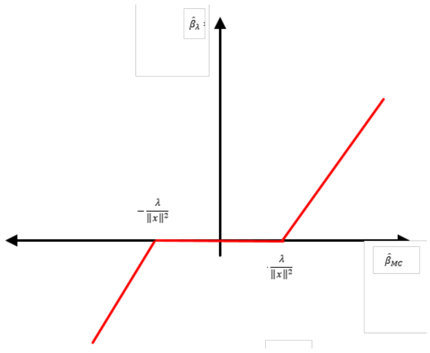
On considère $\operatorname{sign}\hat\beta_\lambda = \operatorname{sign}\hat\beta_{MC}$

Si $\hat{\beta}_\lambda = 0$:

$$\partial_\beta J_{MC} = \sum_{i=1}^{n} (x_i \hat{\beta}_\lambda - y_i) x_i + \lambda\alpha = 0 \qquad \alpha \in [-1;1]$$

$$\hat{\beta}_\lambda = \frac{x^\top y}{\|x\|^2} - \lambda \frac{\alpha}{\|x\|^2} = 0$$

$$-\frac{\lambda}{\|x\|^2} \leq \frac{\lambda\alpha}{\|x\|^2} \leq \frac{\lambda}{\|x\|^2} \Leftrightarrow -\frac{\lambda}{\|x\|^2} \leq \underbrace{\frac{x^\top y}{\|x\|^2}}_{\hat{\beta}_{MC}} \leq \frac{\lambda}{\|x\|^2}$$



Donc :

$$\hat{\beta}_\lambda = \operatorname{sign} \hat{\beta}_{MC} \max\left(0, |\hat{\beta}_{MC}| - \frac{\lambda}{\|x\|^2}\right)$$

2. Algorithme component wise

a. Algorithme

Tant que (non convergé)

Choisir p

$\beta(p) = \operatorname{sign} \hat{\beta}_{MC} \max\left(0, |\hat{\beta}_{MC}| - \frac{\lambda}{\|x\|^2}\right)$

Fin tant que

b. Correction du coût

Si on connaît tous les β sauf β_j

$$J = \frac{1}{2} \sum_{i=1}^{n} \left(x_{ij}\beta_j + \sum_{\substack{k=1 \\ k\neq j}}^{p} x_{ik}\beta_k - y_i \right) + \lambda|\beta_j| + \sum_{\substack{k=1 \\ k\neq j}}^{p} |\beta_k|$$

$$\min_{\beta_j} \frac{1}{2} \sum_{i=1}^{n} (x_{ij}\beta_j + z_i)^2 + \lambda|\beta_j| \qquad z_i = \sum_{\substack{k=1 \\ k\neq j}}^{p} x_{ik}\beta_k - y_i$$

3. Autres rétrécisseur

a. Hard threshold

$$\hat{\beta}_{H\,j} = \begin{cases} 0 & \text{si } \left|\hat{\beta}_{MC\,j}\right| < \dfrac{\lambda}{\|x\|^2} \\ \hat{\beta}_{MC\,j} & \text{sinon} \end{cases}$$

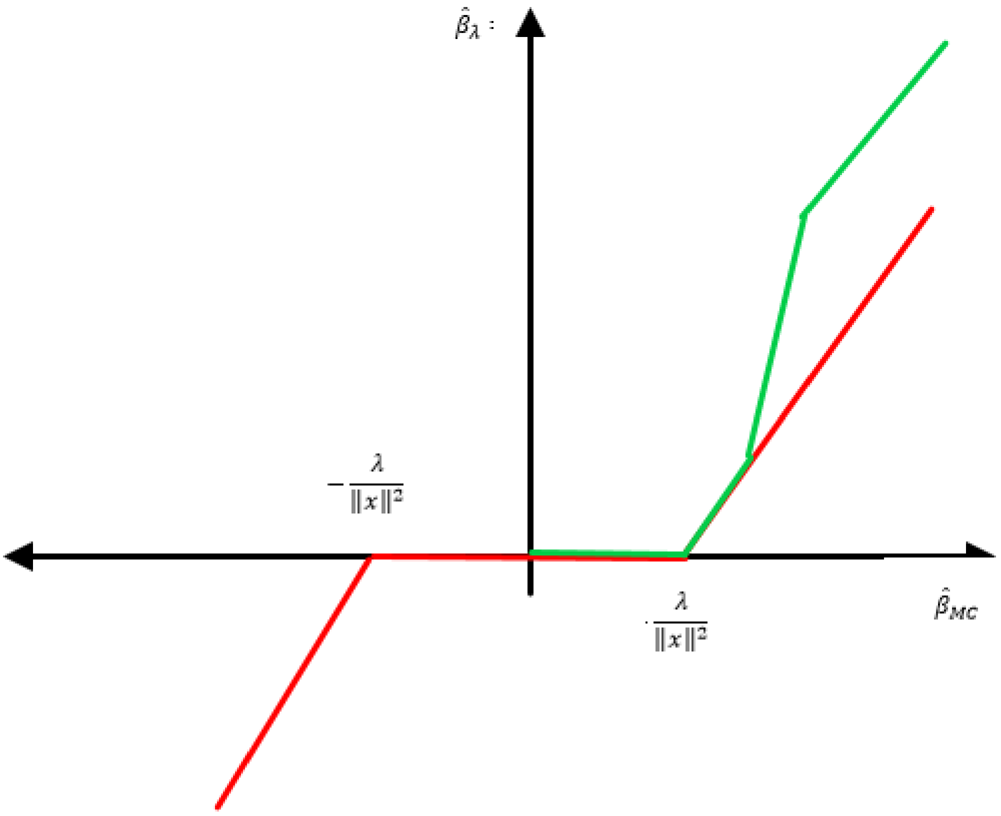
Débiaisé mais discontinu

b. Minimax Concave Penalty

$$\hat{\beta}_{H\,j} = \begin{cases} 0 & \text{si } \left|\hat{\beta}_{MC\,j}\right| < \dfrac{\lambda}{\|x\|^2} \\ \left(2\left|\hat{\beta}_{MC\,j}\right| - \dfrac{2\lambda}{\|x\|^2}\right) \operatorname{sign} \hat{\beta}_{MC\,j} & \text{si } \dfrac{\lambda}{\|x\|^2} < \left|\hat{\beta}_{MC\,j}\right| < \dfrac{2\lambda}{\|x\|^2} \\ \hat{\beta}_{MC\,j} & \text{si } \left|\hat{\beta}_{MC\,j}\right| > \dfrac{2\lambda}{\|x\|^2} \end{cases}$$

Débiaisé et continu

c. Smoothly Cliped Absolute Deviation



On peut les réécrire :

$$\min_{\beta_j} \frac{1}{2} \sum_{i=1}^{n} (x_i\beta - y_i)^2 + \lambda \operatorname{pen} \beta$$

LASSO : $\text{pen}\,\beta = |\beta|$

HARD : $\text{pen}\,\beta = \begin{cases} |\beta| & |\hat{\beta}_{MC}| < \lambda \\ cst & |\hat{\beta}_{MC}| \geq \lambda \end{cases}$

MCP : $\text{pen}\,\beta = \begin{cases} |\beta| & |\hat{\beta}_{MC}| < \lambda \\ a|\beta|^2 + b|\beta| + c & \lambda < |\beta| < 2\lambda \\ cst & |\hat{\beta}_{MC}| \geq 2\lambda \end{cases}$

IV. Adaptative Lasso

1. Etude d'une loi binomiale (HS)

$B \in \{-1;1\} \sim \mathcal{B}\left(\frac{1}{2}\right)$ $\qquad \sum_{i=1}^{n} B_i \underset{n\to\infty}{\longrightarrow} \pm\infty$ $\qquad \frac{1}{n}\sum_{i=1}^{n} B_i \underset{n\to\infty}{\longrightarrow} 0$ $\qquad \frac{1}{\sqrt{n}}\sum_{i=1}^{n} B_i \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,\sigma^2)$

2. Modèle Lasso

$y = X\beta^* + \varepsilon$ et $\mathcal{A}^* = \left\{j \middle| \beta_j^* \neq 0\right\}$ (ensemble des variables actives)

avec $\varepsilon \sim \mathcal{N}(0,\Sigma)$

3. Estimateur Lasso

$$\hat{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1 \qquad \hat{\mathcal{A}}_\lambda = \left\{j \middle| \hat{\beta}_{\lambda_j} \neq 0\right\}$$

On veut que $\left(\hat{\beta}_\lambda, \hat{\mathcal{A}}_\lambda\right) \to (\beta^*, \mathcal{A}^*)$, c'est-à-dire :

$$\sqrt{n}\left(\hat{\beta}_\lambda - \beta^*\right) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0,\Sigma) \qquad \mathbb{P}\left(\hat{\mathcal{A}}_\lambda - \mathcal{A}^*\right) \underset{n\to\infty}{\longrightarrow} 1$$

Théorème : $\exists\, c \in [0,1[\ t.q.\ \mathbb{P}\left(\hat{\mathcal{A}}_\lambda = \mathcal{A}^*\right) \leq c < 1$

On n'est donc pas sûr de converger vers la bonne valeur.

4. Problème Adaptative Lasso

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j|$$

$$w_j = \frac{1}{|\hat{\beta}_{MC\,j}|}$$

5. Problème Garote

$$\min_{c\in\mathbb{R}^p} \frac{1}{2}\left\|y - X\,\text{diag}\left(\hat{\beta}_{MC}\right)c\right\|^2 + \lambda\|c\|_1 \qquad \hat{\beta}_{G_j} = \hat{\beta}_{MC_j}\,c_j$$

6. Etude Adaptative Lasso

a. Reformulation

$$\begin{cases} \underset{\beta\in\mathbb{R}^p}{\min} & \frac{1}{2}\|y - z\|^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j| \\ \text{s.c.} & z = X\beta \end{cases}$$

b. Lagrangien

$$\mathcal{L}(\beta, z, \alpha) = \frac{1}{2}\|y - z\|^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j| + \alpha^\top(z - X\beta)$$

$$\nabla_\beta \mathcal{L} = -X^\top\alpha + \lambda\gamma w \qquad \gamma \in [-1; 1] \qquad\qquad \nabla_z \mathcal{L} = z - y + \alpha \Rightarrow \underline{z = y - \alpha}$$

Dual

$$\begin{cases} \max\limits_{\alpha} & -\frac{1}{2}\|\alpha\|^2 + \alpha^\top z \\ \text{s.c.} & |X^\top a| \leq \lambda w \end{cases}$$

car $\gamma \in [-1; 1] \Rightarrow -\lambda w \leq \lambda\gamma w \leq \lambda w$ et $X^\top\alpha = \lambda\gamma w \Rightarrow -\lambda w \leq X^\top\alpha \leq \lambda w$

$$\Leftrightarrow \begin{cases} \min\limits_{\alpha} & \frac{1}{2}\|\alpha\|^2 - \alpha^\top y \\ \text{s.c.} & |X^\top a| \leq \lambda w \end{cases}$$

$$a = y - z = y - X\beta$$

$$\Leftrightarrow \begin{cases} \min\limits_{\beta} & \frac{1}{2}\|y - X\beta\|^2 - (y - X\beta)^\top y \\ \text{s.c.} & |X^\top(y - X\beta)| \leq \lambda w \end{cases} \Leftrightarrow \begin{cases} \min\limits_{\beta} & \frac{1}{2}\|X\beta\|^2 \\ \text{s.c.} & |X^\top(y - X\beta)| \leq \lambda w \end{cases}$$

V. Elastic Net et gradient proximal

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1 + \frac{\mu}{2}\|\beta\|^2$$

1. Régression Ridge

$$\min_{\beta\in\mathbb{R}^p} \begin{aligned} J_R &= \frac{1}{2}\|y - X\beta\|^2 + \frac{\mu}{2}\|\beta\|^2 = \frac{1}{2}(y - X\beta)^\top(y - X\beta) + \frac{\mu}{2}\beta^\top\beta \\ &= \frac{1}{2}y^\top y - y^\top X\beta + \frac{1}{2}\beta^\top X^\top X\beta + \frac{\mu}{2}\beta^\top\beta \end{aligned}$$

$$\nabla_\beta J_R = -X^\top y + (X^\top X + \mu I)\beta = 0$$

$$\beta = (X^\top X + \mu I)^{-1}X^\top y$$

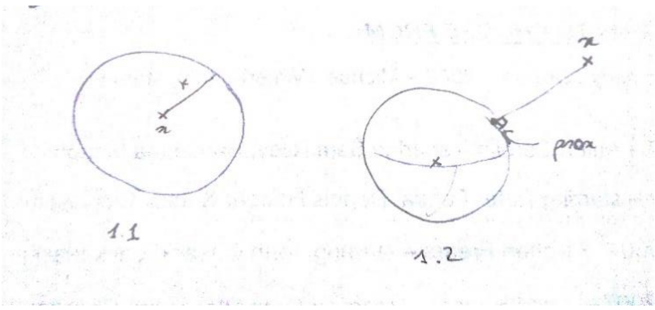
2. Opérateur proximal

a. Définition

$$\Omega \text{ convexe} : \begin{array}{ccc} \mathbb{R}^p & \to & \mathbb{R} \\ x & \to & \Omega(x) \end{array} \qquad \operatorname*{prox}_\Omega x : \begin{array}{ccc} \mathbb{R}^p & \to & \mathbb{R}^p \\ x & \to & \operatorname*{prox}_\Omega x \end{array}$$

$$\boxed{\operatorname*{prox}_\Omega x = \operatorname*{argmin}_{u\in\mathbb{R}^p}\left(\Omega(u) + \frac{1}{2}\|x - u\|^2\right) = \underbrace{\begin{cases} \operatorname*{argmin}_u & \Omega(u) \\ \text{s.c.} & \|x - u\|^2 \leq k \end{cases}}_{\substack{\text{recherche dans une} \\ \text{région de confiance 1.1}}} \Leftrightarrow \underbrace{\begin{cases} \operatorname*{argmin}_u & \frac{1}{2}\|x - u\|^2 \\ \text{s.c.} & \Omega(u) \leq k \end{cases}}_{1.2}}$$

$$\nabla_u J = \nabla_u \Omega(u) - x + u$$



b. Exemples

- $\Omega_1(x) = 0$

$$\operatorname*{prox}_{\Omega_1} x = u^* = \operatorname*{argmin}_{u\in\mathbb{R}^p}\left(0 + \frac{1}{2}\|x - u\|^2\right)$$

$$\begin{aligned} &\nabla_u J_1 = -x + u^* = 0 \\ &\Rightarrow \operatorname*{prox}_{\Omega_1} x = x \end{aligned}$$

- $\Omega_2(x) = \frac{\mu}{2}\|x\|^2$

$$\operatorname*{prox}_{\Omega_2} x = u^* = \operatorname*{argmin}_{u\in\mathbb{R}^p}\left(\frac{\mu}{2}\|u\|^2 + \frac{1}{2}\|x - u\|^2\right)$$

$$\begin{aligned} &\nabla_u J_2 = \mu u^* - x + u^* = 0 \\ &\Rightarrow \operatorname*{prox}_{\Omega_2} x = \frac{1}{1+\mu}x \end{aligned}$$

- $\Omega_3(x) = \lambda\|x\|_1$

$$\operatorname{prox}_{\Omega_3} x = u^* = \underset{u\in\mathbb{R}^p}{\operatorname{argmin}}\left(\lambda\|u\|_1 + \frac{1}{2}\|x-u\|^2\right)$$

$$\partial_u J_3 = \begin{Bmatrix} -\lambda & \text{si } u<0 \\ \alpha\lambda & \text{si } u=0 \\ \lambda & \text{si } u>0 \end{Bmatrix} + u^* - x = 0 \qquad \alpha\in[-1,1]$$

$$\Leftrightarrow \begin{Bmatrix} -\lambda+u-x=0 & \text{si } u<0 & \Rightarrow u = x+\lambda \\ \alpha\lambda+u-x=0 & \text{si } u=0 & \Rightarrow x\le|\lambda| \\ \lambda+u-x=0 & \text{si } u>0 & \Rightarrow u=x-\lambda \end{Bmatrix}$$

$$\Rightarrow \operatorname{prox}_{\Omega_3} x = \begin{Bmatrix} x-\lambda & \text{si } x>\lambda \\ 0 & \text{si } x\le|\lambda| \\ x+\lambda & \text{si } x<-\lambda \end{Bmatrix} \Leftrightarrow \operatorname{sign} x \max(0,|x|-\lambda)$$

- $\Omega_4(x) = \begin{cases} 0 & si\ x\in C \\ \infty & sinon \end{cases}$

$$\operatorname{prox}_{\Omega_4} x = \underset{u\in\mathbb{R}^p}{\operatorname{argmin}}\left(1_{\{u\in C\}} + \frac{1}{2}\|x-u\|^2\right) \quad \text{avec } 1_{\{u\in C\}} = \begin{cases} 0 & \text{si } x\in C \\ \infty & \text{sinon} \end{cases}$$

$$\operatorname{prox}_{\Omega_4} x = \underset{u\in C}{\operatorname{argmin}}\frac{1}{2}\|x-u\|^2 = \text{projection orthogonale de } x \text{ sur } C$$

3. Méthode du gradient proximal

a. Principe

On veut $\min_{x\in\mathbb{R}^p} C(x) + \lambda\Omega(x)$, C et Ω sont convexe. C est différentiable.

b. Algo général

On construit 2 suites $x^{(k)}$ et $\tilde{x}^{(k)}$

$$\tilde{x}^{(k+1)} = x^{(k)} - \rho^{(k)}\nabla_x C\left(x^{(k)}\right)$$
$$x^{(k+1)} = \operatorname{prox}_{\rho^{(k)}\lambda\Omega} \tilde{x}^{(k+1)} = \operatorname{prox}_{\rho^{(k)}\lambda\Omega}\left(x^{(k)} - \rho^{(k)}\nabla_x C\left(x^{(k)}\right)\right)$$

c. Application à Elastic Net

$$\min_{\beta\in\mathbb{R}^p} \underbrace{\frac{1}{2}\|y-X\beta\|^2 + \frac{\mu}{2}\|\beta\|^2}_{C(\beta)} + \lambda\underbrace{\|\beta\|_1}_{\Omega(\beta)}$$

$$\tilde{\beta}^{(k+1)} = \beta^{(k)} - \rho^{(k)}\left((X^\top X + \mu I)\beta^{(k)} - X^\top y\right)$$
$$\beta^{(k+1)} = \operatorname{prox}_{\rho^{(k)}\lambda\Omega} \tilde{\beta}^{(k+1)}$$
$$\beta^{(k+1)} = \operatorname{sign} \tilde{\beta}^{k+1} .* \max\left(0, \left|\tilde{\beta}^{(k+1)}\right| - \rho^{(k)}\lambda\right)$$

$$\operatorname{prox}_{\rho^{(k)}\lambda\Omega}\left(x^{(k)} - \rho^{(k)}\nabla_x C\left(x^{(k)}\right)\right)$$
$$= \underset{u\in\mathbb{R}^p}{\operatorname{argmin}}\left(\rho^{(k)}\lambda\Omega(u) + \frac{1}{2}\left\|x^{(k)} - \rho^{(k)}\nabla_x C\left(x^{(k)}\right) - u\right\|^2\right)$$
$$= \underset{u\in\mathbb{R}^p}{\operatorname{argmin}}\left(\lambda\Omega(u) + \frac{1}{2\rho^{(k)}}\left\|x^{(k)}-u\right\|^2 + {\rho^{(k)}}^2\left\|\nabla_x C\left(x^{(k)}\right)\right\|^2 - 2\rho^{(k)}\left(x^{(k)}-u\right)^\top\nabla_x C\left(x^{(k)}\right)\right)$$
$$= \underset{u\in\mathbb{R}^p}{\operatorname{argmin}}\left(\lambda\Omega(u) + \underbrace{C\left(x^{(k)}\right) + \left(u-x^{(k)}\right)^\top\nabla_x C\left(x^{(k)}\right) + \frac{1}{2\rho^{(k)}}\left\|x^{(k)}-u\right\|^2}_{\text{approximation locale de } C(u)}\right)$$

VI. Gradient proximal

1. Principe

On cherche $\min_w L(w) + \Omega(w)$

Hypothèses : L est convexe et différentiable, Ω est convexe et différentiable ou non.

Exemple :

$$\ell(w) = \frac{1}{2}\|y - Xw\|_2^2 \qquad \Omega(w) = \frac{\lambda}{2}\|w\|_1$$

2. L'algo du gradient proximal est

$$w_{k+1} = \operatorname*{prox}_{\gamma\Omega} w_k - \gamma\nabla\ell(w_k)$$

3. Avec le prox

$$\operatorname*{prox}_{\Omega}(x) = \operatorname*{argmin}_u \frac{1}{2}\|x - u\|_2^2 + \Omega(u)$$

a. Caractérisation du problème

On a $u^* = \operatorname*{prox}_{\Omega}(x)$. u^* est optimal si $0 \in \{\partial\Omega(u^*) - (x - u^*)\} \Leftrightarrow \boxed{(x - u^*) \in \partial\Omega(u^*)}$

b. Corollaire

Si $u^* = \operatorname*{prox}_{\Omega}(u^*)$ alors u^* minimise $\Omega(\cdot)$

c. Algo du point fixe

Si on pose $u_{k+1} = \operatorname*{prox}_{\Omega}(u_k)$, cet algorithme converge vers le min de $\operatorname*{prox}_{\Omega}$ $\forall\,\Omega$ car l'opérateur proximal est contractant.

4. Théorème

$$w^* = \min_w \ell(w) + \Omega(w) \Leftrightarrow w^* = \operatorname*{prox}_{\nu\Omega}\big(w^* - \nu\nabla\ell(w^*)\big) \qquad \nu > 0$$

a. Preuve

On veut minimiser $\ell(w) + \Omega(w)$. w^* est minimum si :

$$\begin{aligned}
&0 \in \{\nabla\ell(w^*) + \partial\Omega(w^*)\}\\
&\Leftrightarrow -\nabla\ell(w^*) \in \partial\Omega(w^*)\\
&\Leftrightarrow -\nu\nabla\ell(w^*) \in \partial\nu\Omega(w^*)\\
&\Leftrightarrow w^* - \nu\nabla\ell(w^*) - w^* \in \partial\nu\Omega(w^*)
\end{aligned}$$

Or w^* minimise $\Omega(\cdot) \Leftrightarrow (x - w^*) \in \partial\Omega(w^*)$

On a donc $w^* = \operatorname*{prox}_{\nu\Omega}\big(w^* - \nu\nabla\ell(w^*)\big)$

5. Approximation quadratique du coût

L'approx quadratique est : $Q(v, w) = \ell(w) + \nabla\ell(w)^\top(v - w) + \frac{L}{2}\|v - w\|^2 + \Omega(v)$

Soit à w_k fixé :

$$Q(v, w_k) = \ell(w_k) + \nabla\ell(w_k)^\top(v - w_k) + \frac{L}{2}\|v - w_k\|^2 + \Omega(v)$$

a. Propriété

$$v^* = \operatorname*{argmin}_v Q(v, w_k)$$

On a $v^* = \operatorname{prox}_{\frac{1}{L}\Omega}\left(w_k - \frac{1}{L}\nabla\ell(w_k)\right)$

A chaque itération du proximal, on réalise en fait une approximation quadratique du coût que l'on minimise.

- *Preuve*

On cherche

$$\begin{aligned}
\operatorname{argmin} Q &= \operatorname{argmin} \ell(w_k) + \nabla\ell(w_k)^\top v - (\nabla\ell(w_k)^\top w_k) + \frac{L}{2}v^\top v - \frac{L}{2}v^\top w + \frac{L}{2}w^\top w + \Omega(v) \\
&= \operatorname{argmin} \nabla\ell(w_k)^\top v + \frac{L}{2}v^\top v - \frac{L}{2}v^\top w_k + \Omega(v) + \Bbbk \\
&= \operatorname{argmin} L\left[\frac{1}{2}v^\top v - v^\top\left(w_k - \frac{1}{2}\nabla\ell(w_k)\right)\right] + \Omega(v) + \Bbbk \\
&= \operatorname{argmin} \frac{L}{2}\left\|v - \left(w_k - \frac{1}{2}\nabla\ell(w_k)\right)\right\|^2 + \Omega(v) + \Bbbk \quad (*) \\
&= \operatorname{argmin} \frac{1}{2}\left\|v - \left(w_k - \frac{1}{L}\nabla\ell(w_k)\right)\right\|^2 + \frac{1}{L}\Omega(v)
\end{aligned}$$

(*) On rajoute un terme constant en v pour pouvoir ajouter la norme.

b. Propriété

$$w_{k+1} = \operatorname{prox}_{\frac{1}{L_k}\Omega}\left(w_k - \frac{1}{L_k}\nabla\ell(w_k)\right)$$

Si $\forall\, k$, L_k est t.q. $\forall\, v \;\; l(v) + \Omega(v) \leq Q(v, w_k)$

Alors on a : $l(w_{k+1}) + \Omega(w_{k+1}) \leq l(w_k) + \Omega(w_k)$

- *Preuve*

$$l(w_k) + \Omega(w_k) = Q(w_k, w_k) \geq Q(w_{k+1}, w_k) \geq (w_{k+1}) + \Omega(w_{k+1})$$

Car $w_{k+1} = \min Q(v, w_k)$

c. Majorant de ℓ

Soit $\ell(v)$ une fonction différentiable telle que $\exists\, L_f \qquad \forall v, w\ \|\nabla\ell(w) - \nabla\ell(v)\| \leq L_f\|w - v\|$

(Càd que la fonction est régulière)

Alors $\forall\, L \geq L_f \qquad \ell(v) \leq \underbrace{\ell(w_k) + \nabla\ell(w)^\top(v - w) + \frac{L}{2}\|v - w\|^2}_{Q_\ell}$

Q_ℓ est majorante de ℓ.

d. Majorant de la Hesienne

Si $\ell(v)$ est doublement différentiable et $\forall\, w_k, \exists\, L_f, \|H(w_k)\| \leq L_f$

Alors $\|\nabla\ell(w) - \nabla\ell(v)\| \leq L_f \|w - v\|$

6. TP

$$l = \frac{1}{2}\left(1 - y(Xw + w_0)\right)_+^\top \left(1 - y(Xw + w_0)\right)_+$$

$$\Omega = \|w\|_1$$

$$(\cdot)_+ = \max(0,\cdot)$$

$$\nabla_w l = -(YX)^\top (1 - Y(Xw + w_0)_+)$$
$$\nabla_{w_0} l = -y(1 - Y(Xw + w_0)_+)$$

VII. Factorisation non négative (Non negative matrix factorization)

1. Introduction

Soit $X \in \mathbb{R}^{n\times p}$, on cherche $\hat{X} = UV^\top \quad U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{k\times p}$ tel que $X \approx \hat{X}$

$$\hat{x_\iota} = VU_{i\bullet}^\top$$

On a donc V dictionnaire d'information et U représente l'information pour chaque point, la contribution de chaque composante du dictionnaire à l'observation.

On peut alors contraindre U, par exemple $U \geq 0$.

2. Cas de la régression non négative

a. Le problème

$$\begin{array}{ll} \min\limits_{\beta\in\mathbb{R}^k} & \frac{1}{2}\|X\beta - y\|^2 \\ \text{s.c.} & \beta \geq 0 \end{array}$$

b. Problème dual

$$\mathcal{L} = \frac{1}{2}\|X\beta - y\|^2 - \gamma^\top\beta = \frac{1}{2}\beta^\top X^\top X\beta + \beta^\top(X^\top X\beta - X^\top y - \gamma) + y^\top y$$

$$\nabla_\beta \mathcal{L} = X^\top(X\beta - y) - \gamma = 0$$

$$\mathcal{L} = \frac{1}{2}\beta^\top X^\top X\beta + \beta^\top \underbrace{(X^\top X\beta - X^\top y - \gamma)}_{=0} + y^\top y = \frac{1}{2}\beta^\top X^\top X\beta + y^\top y$$

$$\begin{array}{ll} \min\limits_{\beta\in\mathbb{R}^k} & \frac{1}{2}\gamma^\top(X^\top X)^{-1}\gamma + y^\top X(X^\top X)^{-1}\gamma \\ \text{s.c.} & \gamma \geq 0 \end{array}$$

Aussi difficile que le primal, voire plus car matrice à inverser.

c. Solution

- Problème dual (lagrangien)
- CVX
- MonQP
- Proximal

- *Proximal*

$$g = X^\top(X\beta - y)$$
$$\hat{\beta} = \beta - \rho g$$
$$\beta = \max(0, \hat{\beta})$$

3. Cas matriciel

$$\begin{aligned} \min_U \quad & \|UV - Y\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p \left(y_{ij} - u_{i\bullet} v_{j\bullet}^\top\right)^2 = \sum_{i=1}^n \left\| \underbrace{y_{i\bullet}^\top}_{y} - \underbrace{v}_{X} \underbrace{u_{i\bullet}^\top}_{\beta} \right\|^2 \\ \text{s.c.} \quad & \beta \geq 0 \end{aligned}$$

Algorithme des moindres carrés alternés :

Connaissant V, $u_{i\bullet}$ est la solution de $\begin{aligned} \min_{\beta \in \mathbb{R}^k} \quad & \frac{1}{2}\|X\beta - y\|^2 \\ \text{s.c.} \quad & \beta \geq 0 \end{aligned}$ avec $X = V$, $y = Y(i,:)^\top$

Connaissant U, $v_{j\bullet}^\top = (U^\top U)^{-1} U^\top y_{\bullet i}$

$$v_{j\bullet}^\top = v_{j\bullet}^\top - \rho\left(U^\top\left(U v_{j\bullet} - y_{\bullet j}\right)\right)$$

4. Rappels de SVD

a. Rappel sur les valeurs propres

Soit $M \in \mathbb{R}^{n \times n}$, λ est valeur propre si $\underbrace{\det M - \lambda I}_{\substack{\text{polynome} \\ \text{d'ordre } n}} = 0$

- M a n valeurs propres complexes λ. Elles sont réelles positives si M est définie positive.
- v est le vecteur propre associé à λ tel que $Mv = \lambda v$
  Si M est définie positive, $LL^\top v = \lambda v \Rightarrow LU = \sqrt{\lambda} V \quad L^\top V = \sqrt{\lambda} U$
  $\sqrt{\lambda}$ est la valeur singulière de L et U, V sont les vecteurs singuliers.

b. Reconstruction à l'ordre 1

Soit X une matrice $n \times p$, on cherche sa meilleure approximation à l'ordre 1.

$$\min_{\hat{X}} \|X - uv^\top\|_F^2 \text{ avec rang } \hat{X} = 1, \|u\| = \|v\| = 1$$

$$\|X - uv^\top\|_F^2 = \|X\|_F^2 - 2\langle X, uv^\top\rangle + \|u\|^2\|v\|^2$$
$$\nabla_u J = 2Xv - 2\|v\|^2 u$$
$$\nabla_v J = 2X^\top u - 2\|u\|^2 v$$

La solution est :

$$Xv = \mu u$$
$$X^\top u = \mu v$$

c. Théorème de décomposition

Soit X une matrice $n \times p$, (μ_k, u_k, v_k) la décomposition en valeurs singulières, on a $X = \sum_{k=1}^{\min(n,p)} \mu_k u_k v_k^\top$

Le $\hat{X}$ de rang K est $\hat{X} = \sum_{k=1}^{K} \mu_k u_k v_k^\top$ avec les K plus grand μ_k.

5. Séparation de sources

On veut approximer au mieux la matrice de données par le produit d'un dictionnaire D et d'une matrice d'activation A, soumis à des contraintes de régularisation tel que :

$$\min_{D,A} \|X - DA\|_F^2 + \Omega_D(D) + \Omega_A(A)$$

6. Factorisation non-négative

$$\min_{D,A} \|X - DA\|_F^2$$
$$\text{sc.} d_{ij} \geq 0 \text{ et } a_{ij} \geq 0$$
$$\|d_{ij}\|^2 \leq 1$$

a. Algorithme

Optimisation alternée

$$A_{k+1} = \text{argmin} \frac{1}{2} \|X - D_k A\|_F^2 + \Omega_A(A)$$
$$D_{k+1} = \text{argmin} \frac{1}{2} \|X - D A_{k+1}\|_F^2 + \Omega_D(D)$$

$t = 1$
$A_t = A_k$
$\nu = \frac{1}{\|D^\top D\|_*}$ (1/1^e v.p)
Répéter
$\nabla_A J = D_k^\top (X - DA_t)$
$A_{t+1} = \underset{\nu\Omega_A}{\text{prox}}(A_t - \nu \nabla_A J)$
Jusqu'à convergence

b. Proximaux

- *1^er^*

$$prox(u) = argmin_x \frac{1}{2} \|x - u\|^2 = \begin{cases} 0 & \text{si } u_i \leq 0 \\ u_i & \text{sinon} \end{cases}$$
$$s.c. x_i \geq 0$$

$$L = \frac{1}{2} \|x - u\|^2 - \sum \alpha_i x_i$$

$$\nabla_{x_i} L = x_i - u_i - \alpha_i = 0 \Rightarrow x_i = u_i + \alpha_i$$

a l'optimalité:

$$\alpha_i - x_i = 0$$

$$x_i \neq 0, \alpha_i = 0 \Rightarrow x_i = u_i$$
$$\alpha_i \neq 0, x_i = 0 \Rightarrow \alpha_i = -u_i$$
$$\alpha_i \geq 0 \; possible \; que \; si \; u_i \leq 0$$

- *2e*

$$\begin{array}{ll} \text{prox} & \text{u} \\ \text{s.c.} & x \geq 0 \\ & \|x\| \leq 1 \end{array} = \begin{cases} 0 & \text{si } u_i < 0 \\ \dfrac{u_i}{\|u_+\|} & \text{si } u_i > 0 \text{ et } \|u_+\| \geq 1 \\ u_i & \text{si } u_i \geq 0 \text{ et } \|u_+\| \leq 1 \end{cases}$$

$$L = \frac{1}{2}\|x-u\|^2 - \sum \alpha_i x_i + \nu(\|x\|^2 - 1)$$

$$\nabla_{x_i} L = x_i - u_i - \alpha_i + 2\nu x_i = 0$$

$$\Leftrightarrow x_i = \frac{u_i + \alpha_i}{1 + 2\nu}$$

A l'optimalité

$$\alpha_i x_i = 0 \Rightarrow$$

$$x_i = \begin{array}{cc} 0 & \text{si } u_i < 0 \\ \dfrac{u_i}{1+2\nu} & \text{sinon} \end{array}$$

$$\nu(\|x\|^2 - 1) = 0 \Rightarrow$$

$$\|x\| = \frac{\|u_+\|}{1+2\nu} \Rightarrow \begin{array}{ll} \nu \neq 0 \Rightarrow & \|x\|^2 = 1 \Rightarrow 1 + 2\nu = \|u_+\| \\ \nu = 0 \Leftarrow & \|x\|^2 < 1 \Leftarrow \|u_+\| < 1 \end{array}$$

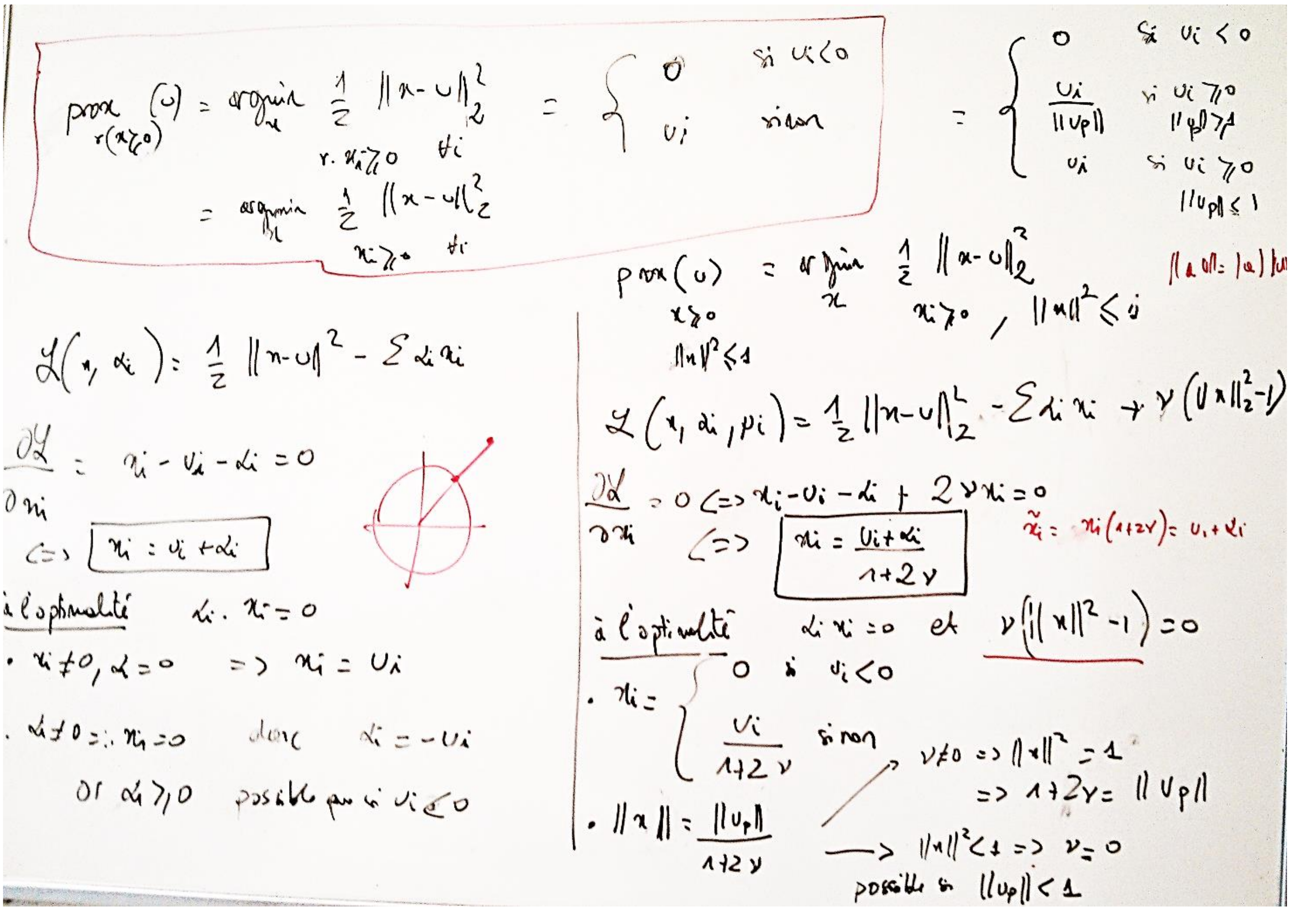
- *3^e^*

$$L = \frac{1}{2}(x_i - u_i)^2 + \lambda\nu|x_i| - \alpha_i x_i$$

$$\nabla_{x_i} L = x_i - u_i + \lambda\nu g - \alpha_i = 0 \qquad g = \begin{array}{cc} 1 & \text{si } x_i > 0 \\ [0;1] & si\ x = 0 \end{array} \text{sous gradient}$$

$$\alpha_i g \,? \Rightarrow x^*$$

$$x_i > 0$$



VIII. Sélection de modèles : choix des hyper-paramètres

1. Exmeple du Lasso

On se donne une grille de M valeurs ordonnées λ_m

Pour chaque λ_m on résout le Lasso et on obtient $\hat{\beta}_m$

Pour chaque $\hat{\beta}_m$ on évalue sa qualité => SURE

2. SURE : Stein Unbiased Risk Estimator

a. Risque de l'estimateur

- *Erreur de prédiction*

$$x_i^{new}, y_i^{new}$$

$$EP(\hat{\beta}_m) = \lim_{l \to \infty} \frac{1}{l} \sum_{i=1}^{l} \left(x_i^{new} \hat{\beta}_m - y_i^{new}\right)^2 = \mathbb{E}\left(x_i^{new} \hat{\beta}_m - y_i^{new}\right)^2 \quad \text{avec} \quad (x_i^{new}, y_i^{new}) \sim \mathbb{P}(x, y)$$

- *Hypothèses*

- X est déterministe
- On pose un modèle $y = X\beta + \varepsilon \quad \varepsilon_i \sim \mathcal{N}(0,\ \sigma_i^2)$

$EP_{simpl.}(\hat{\beta}_m) = \mathbb{E}_y\left(\|X\hat{\beta}_m - y\|\right)$ épérance sur la VA y

$Risque(\hat{\beta}) = \mathbb{E}_\beta\left(EP(\hat{\beta}_m)\right)$ espérance sur β

Simplification de l'erreur de prédiction

$$EP(\hat\beta_m) = \mathbb{E}\left(\|X\hat\beta_m - y^{new}\|^2\right) = \mathbb{E}\left(\|X\hat\beta_m - X\beta + X\beta - y\|^2\right)$$
$$= \mathbb{E}\left(\|X\hat\beta_m - X\beta\|^2\right) + 2\,\underbrace{\mathbb{E}\left((X\hat\beta_m - X\beta)^\top \underbrace{(X\beta - y)}_{\varepsilon}\right)}_{=0} + \mathbb{E}\left(\left\|\underbrace{(X\beta - y)}_{\varepsilon}\right\|^2\right)$$
$$= \|X\hat\beta_m - X\beta^*\|^2 + n\sigma^2$$

b. Réécriture

$$\|X\hat\beta_m - X\beta^*\|^2 = \|X\hat\beta_m - y + y - X\beta^*\|^2 \quad y\ observation$$
$$= \underbrace{\|X\hat\beta_m - y\|^2}_{connu} + 2\,\underbrace{(X\hat\beta_m - y)^\top}_{r^\top}\underbrace{(y - X\beta^*)}_{\varepsilon} + \underbrace{\|y - X\beta^*\|^2}_{indéendant\ de\ m}$$
$$= \underbrace{\|X\hat\beta_m - y\|^2}_{connu} + 2(X\hat\beta_m)^\top(y - X\beta^*) + cst(m)$$

c. SURE

$$SURE(\hat\beta_m) \coloneqq \|X\hat\beta_m - y\|^2 + 2\,\mathrm{div}(X\hat\beta_m)\,\sigma^2 - n\sigma^2$$

Identité de Stein

$$\mathbb{E}\left(SURE(\hat\beta_m)\right) = \mathbb{E}\left(\|X\hat\beta_m - X\beta^*\|^2\right)$$

Preuve

$$\mathbb{E}\left(\|X\hat\beta_m - X\beta^*\|^2\right) = \mathbb{E}\left(\|X\hat\beta_m - y\|^2\right) + 2\mathbb{E}\left((X\hat\beta_m)^\top\varepsilon\right) - \underbrace{2\mathbb{E}((X\beta^*)^\top\varepsilon)}_{0} - \underbrace{\mathbb{E}\|\varepsilon_i\|^2}_{\sigma^2}$$

Soit φ fct de R dans R

$$\mathbb{E}(\varepsilon\varphi(\varepsilon)) = \int_\varepsilon \varepsilon\varphi(\varepsilon)\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\varepsilon^2}{2\sigma^2}}\right)d\varepsilon$$
$$u = \varphi(\varepsilon) \quad u' = \varphi'(\varepsilon)$$
$$v' = \varepsilon e^{-\frac{\varepsilon^2}{2\sigma^2}} \quad v = \sigma^2 e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

IPP :

$$= \sigma^2\int \varphi'(\varepsilon)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{\varepsilon^2}{2\sigma^2}}\,d\varepsilon$$
$$= \sigma^2\mathbb{E}(\varphi'(\varepsilon))$$

$$\mathbb{E}\left(\|X\hat\beta_m - X\beta^*\|^2\right) = \mathbb{E}\left(\|X\hat\beta_m - y\|^2\right) + 2\sigma^2\mathbb{E}\left(\underbrace{\sum\frac{\partial X\hat\beta_m}{\partial\varepsilon_i}}_{\mathrm{div}(X\hat\beta_m)}\right) - \underbrace{2\mathbb{E}((X\beta^*)^\top\varepsilon)}_{0} - \underbrace{\mathbb{E}\|\varepsilon_i\|^2}_{\sigma^2}$$

- *Calcul de la divergence*

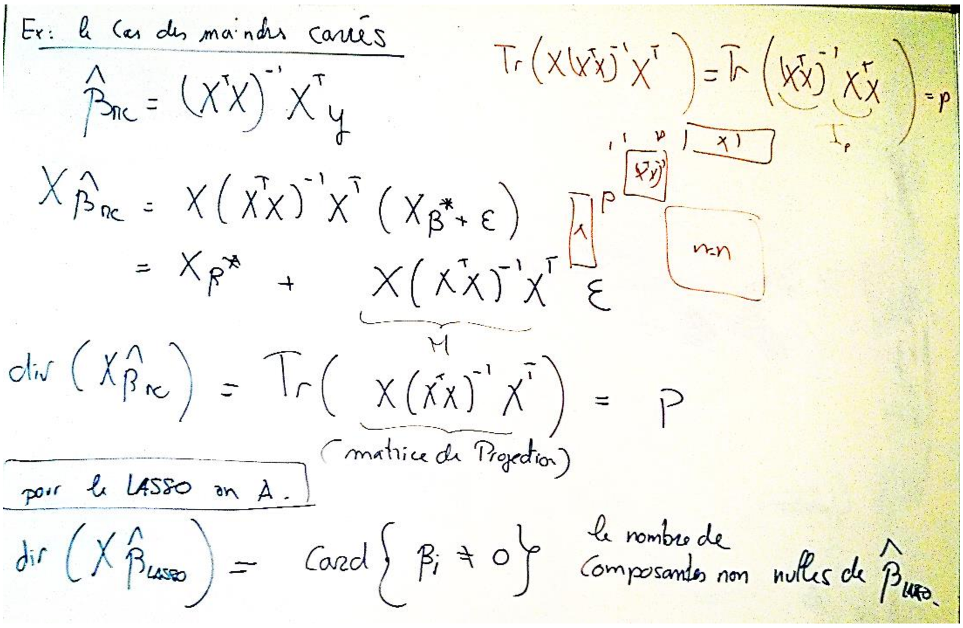
$$\varphi: \mathbb{R}^n \to \mathbb{R}^n \quad \varphi(v) = Mv = w$$

$$\operatorname{div}\varphi(v) = \frac{\partial\varphi(v)}{\partial v_1} + \cdots + \frac{\partial\varphi(v)}{\partial v_n} = M_{11} + M_{22} + M_{nn} = \operatorname{trace} M$$

3. Résumé

$$\begin{aligned}
\|X\hat\beta_m - X\beta^*\|^2 &= \|X\hat\beta_m - y\|^2 + 2(X\hat\beta_m)^\top\varepsilon - 2(X\beta^*)^\top\varepsilon - \|\varepsilon\|^2 \\
SURE(\hat\beta_m) &= \|X\hat\beta_m - y\|^2 + 2\sigma^2\operatorname{div}(X\hat\beta_m) \qquad - n\sigma^2 \\
\operatorname{div}(Mv) &= \operatorname{trace} M
\end{aligned}$$

4. Exemple de div



IX. Optimisation non-convexe & MCP

DC (Difference of Convex) / CCCP (Concave Convex Procedure) / LLA (Local Linear Approximation) / MM (MaxMin) / Iterative reweighted

= Relaxation convexe iterative

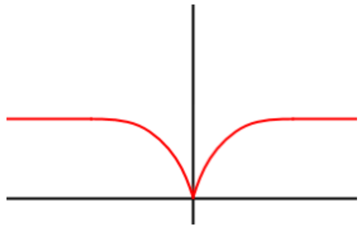
1. Problème du MCP (Minimum Concave Penalty)

$$y = X\beta^* + \varepsilon$$

$$\min_\beta \frac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} pen(|\beta_j|)$$

Exemples de pénalités :

- $pen(t) = t \Rightarrow$ lasso
- $pen(t) = \frac{t}{w} \Rightarrow$ adaptative lasso
- $\boxed{pen_{\lambda,\gamma}(t) = \begin{cases} \lambda t - \frac{t^2}{2\gamma} & \text{si } t < \gamma\lambda \\ \frac{\gamma\lambda^2}{2} & \text{sinon} \end{cases}}$



Le MCP s'écrit aussi :

$$\min_\beta \frac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| - \sum_{j=1}^{p} h(|\beta_j|, \lambda, \gamma)$$

$$h(t, \lambda, \gamma) = \frac{t^2}{2\lambda\gamma}\mathbb{I}_{\{t \le \gamma\lambda\}} + \left(t - \frac{\gamma\lambda}{2}\right)\mathbb{I}_{\{t > \gamma\lambda\}} = \begin{cases} \frac{t^2}{2\lambda\gamma} & \text{si } t \le \gamma\lambda \\ t - \frac{\gamma\lambda}{2} & \text{sinon} \end{cases}$$

2. Algorithme

Soit $\min_x f(x) - h(x)$

Tant que on a pas convergé

$$x^{new} = \text{argmin}_x\, f(x) - \nabla_x h(x^{old})^\top x$$

Fin tant que

3. Application au MCP

$$h'(t) = \frac{t}{\gamma}\mathbb{I}_{\{t \le \gamma\lambda\}} + \lambda\mathbb{I}_{\{t > \gamma\lambda\}}$$

$$\beta^{new} = \text{argmin}_x \frac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| - \sum_{j=1}^{p} \left(\frac{|\beta_j^{old}|}{\gamma}\mathbb{I}_{\{t \le \gamma\lambda\}} + \lambda\mathbb{I}_{\{t > \gamma\lambda\}}\right)|\beta_j^{old}|$$

$$\Leftrightarrow \beta^{new} = \text{argmin}_x \frac{1}{2}\|X\beta - y\|^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|$$

$$w_j = \begin{cases} 1 - \frac{|\beta_j^{old}|}{\lambda\gamma} & \text{si } |\beta_j^{old}| < \gamma\lambda \\ 0 & \text{sinon} \end{cases}$$

$$\partial_\beta J = \underbrace{X^\top(X\beta - y)}_{g} + \left\{ \begin{matrix} \alpha_j & si\ \beta_i = 0 \\ \lambda sign(\beta_j) - \frac{\beta_j}{\gamma} & si\ 0 < |\beta_j| < \gamma\lambda \\ 0 & |\beta_j| \geq \gamma\lambda \end{matrix} \right\}$$

$$0 \in \partial_\beta J \text{ si } \left\{ \begin{matrix} (g-\lambda)(g+\lambda) \leq 0 & si\ \beta_i = 0 \\ g + \lambda sign(\beta_j) - \frac{\beta_j}{\gamma} = 0 & si\ 0 < |\beta_j| < \gamma\lambda \\ g_i = 0 & |\beta_j| \geq \gamma\lambda \end{matrix} \right\}$$