Institut de mathématiques de Bourgogne

Stage de Master 1 Mathématiques Appliquées

Le chemin des solutions du LASSO généralisé : Application à l'apprentissage statistique

Thibault Barbazza

Année universitaire 2024–2025

Table des matières

1 Introduction

Le modèle de régression linéaire $Y = X\beta + \epsilon$, où $Y \in \mathbb{R}^n$ est la réponse du modèle, $X \in \mathbb{R}^{n\times p}$ est la matrice de régression, $\beta \in \mathbb{R}^p$ est le paramètre inconnu des coefficients de régression et $\epsilon \in \mathbb{R}^n$ est le résidu aléatoire est l'un des modèle en statistique les plus connus. Le paramètre β peut être estimé via l'estimateur LASSO généralisé qui est défini comme une solution du problème d'optimisation convexe suivant:

$$\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|D\beta\|_1 \right\} \tag{1}$$

où $D \in \mathbb{R}^{n\times p}$. Le choix le plus standard pour la matrice D est l'identité, le terme de pénalité est alors la norme l_1 et l'estimateur obtenu est le LASSO qui est connu pour favoriser la parcimonie. De nombreuse autres matrices D sont également pertinentes pour cet estimateur. L'estimateur LASSO généralisé dépend du paramètre de régularisation λ; on notera $\hat{\beta}(\lambda)$ une solution du problème (1). Une approche classique en apprentissage statistique pour choisir ce paramètre de régularisation est de minimiser la somme des carrés résiduels sur un jeu de données de validation X^{val},Y^{val} :

$$\lambda > 0 \mapsto \|Y^{val} - X^{val}\hat{\beta}(\lambda)\|_2^2$$

La minimisation de cette expression nécessite de calculer le chemin des solutions du LASSO généralisé, c'est-à-dire la fonction $\lambda > 0 \mapsto \hat{\beta}(\lambda)$.

2 Quelques rappels

2.1 Sous-gradient et sous-différentiel

Les quelques notions que nous allons aborder nous seront utiles par la suite.

Définition 2.1. (Sous-gradient)
Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction convexe. On dit que $g \in \mathbb{R}^n$ est un sous-gradient de f en $x^* \in \mathbb{R}^n$ si

$$f(x) \geq f(x^*) + g^\top (x - x^*) \quad \forall x \in \mathbb{R}^n.$$

Remarque 2.1. Les sous-gradients de f au point x^* fournissent toutes les minorantes affines de f et tangentes à f en x^*.

Exemple : Considérons la fonction $f : x \mapsto |x|$ sur $\mathbb{R}$. Il est clair que $\forall x \in \mathbb{R}_+^*$, $g = 1$ et que $\forall x \in \mathbb{R}_-^*$, $g = -1$. Mais en $x = 0$, il existe une infinité de droites passant par l'origine et qui minorent $x \mapsto |x|$, les pentes de ces droites doivent être à valeur dans $[-1; 1]$ En d'autres termes, si $g \in \mathbb{R}$ est un sous-gradient de f au point x^* alors :

$$\begin{cases} g = -1 & \text{si } \ x^* < 0 \\ g = 1 & \text{si } \ x^* > 0 \\ g \in [-1; 1] & \text{si } \ x^* = 0 \end{cases}$$

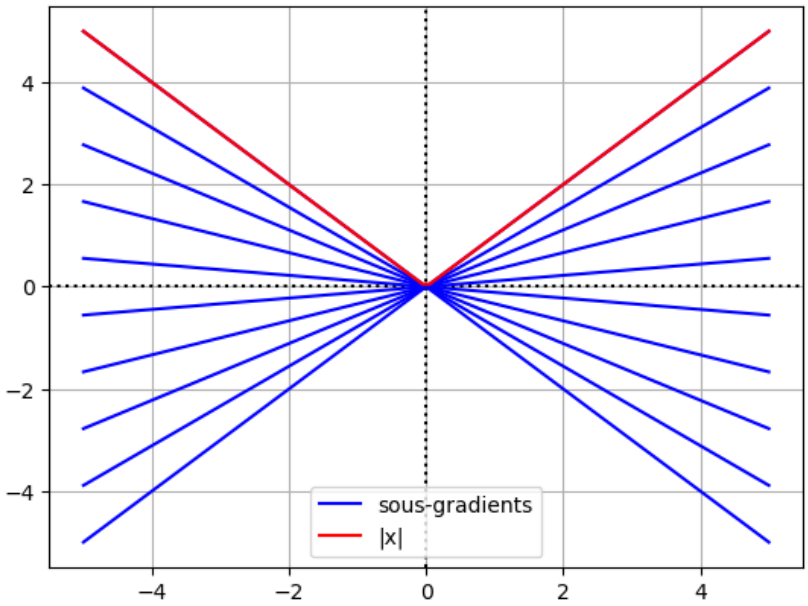

FIG. 1 – *Illustration des minorantes affines de la fonction valeur absolue qui sont tangentes à l'origine*

Définition 2.2. (Sous-différentiel)
Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction convexe. L'ensemble des sous-gradients de f en x^* est appelé le sous-différentiel de f en x^*. Il est noté

$$\partial_f(x^*) = \left\{ g \in \mathbb{R}^n : f(x) \geq f(x^*) + g^\top (x - x^*) \quad \forall x \in \mathbb{R}^n \right\}$$

Exemple : Soit la fonction $f : x \mapsto \|x\|_1 = \sum_{i=1}^n |x_i|$. Alors $g \in \mathbb{R}^n$ est un sous-gradient de f en x^* si g est tel que

$$\begin{cases} g_i = \text{sign}(x_i^*) & \text{si} \quad x_i^* \neq 0 \\ |g_i| \leq 1 & \text{si} \quad x_i^* = 0 \end{cases}$$

Voyons maintenant quelques propriétés sur ces deux objets.

Propriété 2.1. 1. f est dite sous-différentiable en x^* si $\partial_f(x^*) \neq \emptyset$
2. f est dite sous-différentiable si elle l'est pour tout x
3. Si f est convexe et différentiable en x^* alors $\partial_f(x^*) = \{\nabla f(x^*)\}$
4. $\forall \lambda > 0,\ \partial_{\lambda f} = \lambda \partial_f$
5. Soient f_1,$f_2 : \mathbb{R}^n \longrightarrow \mathbb{R}$ deux fonction convexes. Alors

$$\partial_{f_1+f_2} = \partial_{f_1} + \partial_{f_2}$$

Proposition 2.1. (Condition d'optimalité)
Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction convexe. Alors x^* est un minimiseur de f si et seulement si $0 \in \partial_f(x^*)$

Démonstration. Supposons que $x^* \in \mathbb{R}^n$ est tel que $f(x^*) \leq f(x)\ \ \forall x \in \mathbb{R}^n$. C'est équivalent à $f(x^*)+(0\times \mathbb{1}_n)^\top (x-x^*) \leq f(x)\ \ \forall x \in \mathbb{R}^n$. Et par définition, $0 \in \partial_f(x^*)$

□

2.2 Estimateur LASSO

Dans ce cas, le problème d'optimisation convexe (1) s'écrit de cette manière :

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{2}$$

Dans la suite, on notera $S(\lambda)$ l'ensemble des solutions du problème (2) :

$$S(\lambda) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Proposition 2.2. Soient $X \in \mathbb{R}^{n\times p}$,$Y \in \mathbb{R}^n$,$\lambda > 0$ et $S(\lambda)$ l'ensemble des solutions du LASSO. Si $\hat{\beta} \in S(\lambda)$ et $\tilde{\beta} \in S(\lambda)$ alors $X\hat{\beta} = X\tilde{\beta}$ et $\|\hat{\beta}\|_1 = \|\tilde{\beta}\|_1$.

Démonstration. Montrons que $X\hat{\beta} = X\tilde{\beta}$. On suppose que $\hat{\beta} \in S(\lambda)$ et $\tilde{\beta} \in S(\lambda)$. On suppose de plus, par l'absurde, que $X\hat{\beta} \neq X\tilde{\beta}$.
Posons $\beta_0 = \frac{\hat{\beta}+\tilde{\beta}}{2}$, alors on a :

$$\|Y - X\beta_0\|_2^2 = \|\frac{1}{2}(Y - X\hat{\beta}) + \frac{1}{2}(Y - X\tilde{\beta})\|$$

On a de plus, par définition de la norme :

$$\|\beta_0\|_1 \leq \frac{1}{2}\|\hat{\beta}\|_1 + \frac{1}{2}\|\tilde{\beta}\|_1$$

Ainsi, on en déduit par l'inégalité triangulaire que :

$$\frac{1}{2}\|Y - X\beta_0\|_2^2 + \lambda\|\beta_0\|_1 < \frac{1}{2}\left[\frac{1}{2}\|Y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 + \frac{1}{2}\|Y - X\tilde{\beta}\|_2^2 + \lambda\|\tilde{\beta}\|_1\right]$$

Et puisqu'on a supposé que $\hat{\beta} \in S(\lambda)$ et $\tilde{\beta} \in S(\lambda)$, cela implique que :

$$\begin{cases} \hat{\beta} \notin S(\lambda) \\ \text{ou} \\ \tilde{\beta} \notin S(\lambda) \end{cases}$$

Ce qui nous donne notre contradiction. Puisque $X\hat{\beta} = X\tilde{\beta}$, il est maintenant facile de montrer que $\|\hat{\beta}\|_1 = \|\tilde{\beta}\|_1$. □

Voyons maintenant une caractérisation des solutions du LASSO.

Proposition 2.3. (Caractérisation des solutions du LASSO)
Soient $X = [X_1|\dots|X_p] \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$, et $\lambda > 0$.Les assertions suivantes sont équivalentes.

$$\hat{\beta} \in S(\lambda) \tag{3a}$$

$$\begin{cases} \|X^\top(Y - X\hat{\beta})\|_\infty \leq \lambda \\ X_i^\top(Y - X\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_i) \quad \forall i \in \text{Supp}(\hat{\beta}) \end{cases} \tag{3b}$$

$$\begin{cases} \|X^\top(Y - X\hat{\beta})\|_\infty \leq \lambda \\ \hat{\beta}^\top X^\top(Y - X\hat{\beta}) = \lambda\|\hat{\beta}\|_1 \end{cases} \tag{3c}$$

Démonstration. (3a ⇒ 3b)
Notons $F_\lambda(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$. Et notons de plus

$$\begin{cases} F_1(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2 \\ F_{2,\lambda}(\beta) = \lambda\|\beta\|_1 \end{cases}$$

Supposons maintenant que β^* minimise F_λ.
Puisque la fonction F_1 est convexe et différentiable, on a que

$$\partial_{F_1}(\beta^*) = \{\nabla F_1(\beta^*)\} = \left\{-(X^\top(Y - X\beta^*)\right\}$$

Ensuite, on a vu que

$$g \in \partial_{\|.\|_1}(\beta) \text{ si } \begin{cases} g_i = \text{sign}(\beta_i) & \text{si } \beta_i \neq 0 \\ |g_i| \leq 1 & \text{si } \beta_i = 0 \end{cases}$$

Ce qui implique que :

$$g \in \partial_{F_{2,\lambda}}(\beta) \text{ si } \begin{cases} g_i = \lambda\text{sign}(\beta_i) & \text{si } \beta_i \neq 0 \\ |g_i| \leq \lambda & \text{si } \beta_i = 0 \end{cases}$$

Ainsi, on obtient que $\forall \beta \in \mathbb{R}^p$,

$$\partial_{F_\lambda}(\beta) = -X^\top(Y - X\beta) + g \text{ où } g \in \partial_{F_{2,\lambda}}(\beta)$$

Et d'après la condition d'optimalité vue un peu plus haut, on en déduit que :

$$\beta^* \in S(\lambda)$$

$$\Leftrightarrow$$

$$0 \in \partial_{F_\lambda}(\beta^*)$$

$$\Leftrightarrow$$

$$\begin{cases} \|X^\top(Y - X\beta^*)\|_\infty \leq \lambda \\ X_i^\top(Y - X\beta^*) = \lambda\text{sign}(\beta_i^*) \quad \forall i \in \text{Supp}(\beta^*) \end{cases}$$

□

Corollaire 2.1. Soient $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$ et $\lambda > 0$. Alors

$$S(\lambda) = \{0\} \Leftrightarrow \|X^\top Y\|_\infty \leq \lambda$$

Démonstration.

Existence:
D'un sens, supposons que $\|X^\top Y\|_\infty \leq \lambda$. Ce qui implique que

$$\begin{cases} \|X^\top(Y - X0)\|_\infty \leq \lambda \\ 0^\top X^\top(Y - X0) = \lambda\|0\|_1 \quad (0 = 0) \end{cases}$$

Et donc $0 \in S(\lambda)$.
De l'autre sens, supposons que $0 \in S(\lambda)$. D'après la caractérisation des solutions du LASSO, on a que

$$\|X^\top(Y - X0)\|_\infty \leq \lambda \Leftrightarrow \|X^\top Y\|_\infty \leq \lambda$$

Pour l'instant on a montré que $\{0\} \subset S(\lambda)$.
Unicité:
Montrons que $\{0\} = S(\lambda)$.
On a vu plus haut que deux éléments de $S(\lambda)$ sont de même norme l_1. Ce que implique que si $\beta \in S(\lambda)$ alors $\|\beta\|_1 = \|0\|_1 = 0$
Et donc $\beta = 0$. □

Proposition 2.4. (Problème d'optimisation lié au LASSO)
Soient $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$, $\lambda > 0$ et $\hat{\beta} \in S(\lambda)$. On pose $\hat{r} = Y - X\hat{\beta}$. Alors

$$\hat{r} = \underset{r\in\mathbb{R}^n}{\arg\min} \|Y - r\|_2^2 \text{ sous contrainte } \|X^\top r\|_\infty \leq \lambda$$

Démonstration. Montrons tout d'abord que $\hat{r}$ est admissible.
Comme $\hat{\beta} \in S(\lambda)$, alors

$$\lambda \geq \|X^\top (Y - X\hat{\beta})\|_\infty = \|X^\top \hat{r}\|_\infty$$

Donc $\hat{r} \in C = \left\{r \in \mathbb{R}^n : \|X^\top \hat{r}\|_\infty \leq \lambda\right\}$
Montrons maintenant que la projection de Y sur l'ensemble convexe fermé C vaut $\hat{r}$.
Pour n'importe quel $z \in C$ on veut que le produit scalaire entre z et Y soit inférieur ou égal à 0. S'il était positif, la projection de Y sur C ne serait pas $\hat{r}$.

$$\begin{aligned}
(Y - \hat{r})^\top (z - \hat{r}) &= (X\hat{\beta})^\top (z - \hat{r}) \\
&= \hat{\beta}^\top X^\top z - \hat{\beta}^\top X^\top (Y - X\hat{\beta}) \\
\text{car } \hat{\beta} \in S(\lambda) \quad &\leq \hat{\beta}^\top X^\top z - \lambda\|\hat{\beta}\|_1 \\
&\leq \|\hat{\beta}\|_1 \|X^\top z\|_\infty - \lambda\|\hat{\beta}\|_1 \\
\text{car } z \in C \quad &\leq \lambda\|\hat{\beta}\|_1 - \lambda\|\hat{\beta}\|_1 = 0
\end{aligned}$$

□

3 Chemin des solutions du LASSO

3.1 Résultats théoriques et exemple

Dans cette partie, nous allons calculer ce qu'on appelle le chemin des solution du LASSO. En d'autres termes, la fonction $\lambda \mapsto \hat{\beta}(\lambda)$ où $\hat{\beta}(\lambda)$ est solution de (2). Nous allons voir que, sous certaines hypothèses (raisonnables), la solution de ce problème est unique, et que la fonction $\lambda \mapsto \hat{\beta}(\lambda)$ est affine par morceaux.

Proposition 3.1. La fonction $\lambda > 0 \mapsto \hat{\beta}(\lambda)$ (Où $\hat{\beta}(\lambda)$ est l'unique élément de $S(\lambda)$) est:

i) affine par morceaux

ii) continue

Démonstration. **i)** Notons $I_s = \left\{\lambda \in]0, +\infty[: \exists \hat{\beta} \in S(\lambda) \text{ tel que } \text{sign}(\hat{\beta}) = s\right\}$ Tout d'abord, il est clair que

$$\bigcup_{s\in\{-1;0;1\}^p} I_s =]0, +\infty[\,.$$

Soient $s^0 \in \{-1; 0; 1\}^p$ et I_{s^0}. On se donne λ_1,$\lambda_2 \in I_{s^0}$ tels que $\hat{\beta}(\lambda_1) \in S(\lambda_1)$ et $\hat{\beta}(\lambda_2) \in S(\lambda_2)$. On suppose de plus que

$$\text{sign}(\hat{\beta}(\lambda_1)) = \text{sign}(\hat{\beta}(\lambda_2)) = s^0$$

On pose $\lambda = \alpha\lambda_1 + (1-\alpha)\lambda_2, \ \ \forall \alpha \in [0,1]$. Montrons que

$$\hat{\beta}(\lambda) = \alpha\hat{\beta}(\lambda_1) + (1-\alpha)\hat{\beta}(\lambda_2) \in S(\lambda)$$

$$\Leftrightarrow$$

$$\begin{cases} \|X^\top(Y - X\hat{\beta}(\lambda))\|_\infty \leq \lambda \\ X_i^\top(Y - X\hat{\beta}(\lambda)) = \lambda \text{sign}(\hat{\beta}_i(\lambda)) \ \ \forall i \in \text{Supp}(\hat{\beta}(\lambda)) \end{cases}$$

Premièrement,

$$\begin{aligned} \|X^\top(Y - X\hat{\beta})\|_\infty &= \|X^\top \left[\alpha Y + (1-\alpha)Y - \alpha X\hat{\beta}(\lambda_1) - (1-\alpha)X\hat{\beta}(\lambda_2)\right] \|_\infty \\ &= \|\alpha X^\top(Y - X\hat{\beta}(\lambda_1)) + (1-\alpha)X^\top(Y - X\hat{\beta}(\lambda_2))\|_\infty \\ &\leq \alpha\|X^\top(Y - X\hat{\beta}(\lambda_1))\|_\infty + (1-\alpha)\|X^\top(Y - X\hat{\beta}(\lambda_2))\|_\infty \\ &\leq \alpha\lambda_1 + (1-\alpha)\lambda_2 \quad \text{car} \ \ \hat{\beta}(\lambda_1) \in S(\lambda_1),\, \hat{\beta}(\lambda_2) \in S(\lambda_2) \\ &= \lambda \end{aligned}$$

Deuxièmement, puisque par construction on a

$$sign(\hat{\beta}(\lambda)) = sign(\hat{\beta}(\lambda_1)) = sign(\hat{\beta}(\lambda_2)) = s^0$$

Alors pour tout $i \in supp(\hat{\beta})$ les égalités suivantes sont satisfaites

$$\begin{aligned} X_i^\top (Y - X\hat{\beta}(\lambda) &= X_i^\top \left[\alpha(Y - X\hat{\beta}(\lambda_1)) + (1-\alpha)(Y - X\hat{\beta}(\lambda_2))\right] \\ &= \alpha X_i^\top (Y - X\hat{\beta}(\lambda_1)) + (1-\alpha) X_i^\top (Y - X\hat{\beta}(\lambda_2)) \\ &= \alpha\lambda_1 s_i^0 + (1-\alpha)\lambda_2 s_i^0 \\ &= (\alpha\lambda_1 + (1-\alpha)\lambda_2) s_i^0 \\ &= \lambda s_i^0 \\ &= \lambda \text{sign}(\hat{\beta}_i(\lambda)) \end{aligned}$$

Ainsi, $\hat{\beta}(\lambda) \in S(\lambda)$.
Donc:

- $\rightarrow$ I_s est un intervalle car λ, qui est une combinaison convexe de λ_1 et λ_2, appartient à I_s
- $\rightarrow$ Puisque $\hat{\beta}(\lambda) = \alpha\hat{\beta}(\lambda_1) + (1-\alpha)\hat{\beta}(\lambda_2) \in S(\lambda)$, on en déduit que $\lambda > 0 \mapsto \hat{\beta}(\lambda)$ est affine par morceaux.

ii) Soient $\lambda > 0$, et $(\lambda_n)_{n\in\mathbb{N}}$ une suite convergeant vers λ. Puisque $\hat{\beta}(\lambda_n)$ est bornée, à une extraction près, on peut supposer que

$$\left(\hat{\beta}(\lambda_n)\right)_{n\in\mathbb{N}} \underset{n\to\infty}{\longrightarrow} L \in \mathbb{R}^p$$

Montrons que $L = \hat{\beta}(\lambda)$.
Puisque $\hat{\beta}(\lambda_n)$ est un minimiseur LASSO (solution de (2)), on a :

$$\frac{1}{2}\|Y - X\hat{\beta}(\lambda_n)\|_2^2 + \lambda_n\|\hat{\beta}(\lambda_n)\|_1 \leq \frac{1}{2}\|Y - X\hat{\beta}(\lambda)\|_2^2 + \lambda_n\|\hat{\beta}(\lambda)\|_1$$

Puis en passant à la limite

$$\frac{1}{2}\|Y - XL\|_2^2 + \lambda\|L\|_1 \leq \frac{1}{2}\|Y - X\hat{\beta}(\lambda)\|_2^2 + \lambda\|\hat{\beta}(\lambda)\|_1$$

Ce qui implique que $L \in S(\lambda)$. Et puisque $S(\lambda)$ admet un unique élément, alors

$$L = \hat{\beta}(\lambda).$$

Donc $\lim\limits_{n\to\infty} \hat{\beta}(\lambda_n) = \hat{\beta}(\lambda)$ et $\lambda > 0 \mapsto \hat{\beta}(\lambda)$ est continue. $\square$

Proposition 3.2. ([2])
$\hat{\beta}(\lambda) \in \mathbb{R}^p$ est solution de (2) si et seulement si

$$\forall j \in \{1,\dots,p\}, \hat{\beta}(\lambda) \text{ vérifie } \begin{cases} |X_j^\top (Y - X\hat{\beta}(\lambda))| \leq \lambda \\ X_j^\top (Y - X\hat{\beta}(\lambda)) = \lambda \text{sign}(\hat{\beta}_j(\lambda)), \ \ \forall j \in \text{Supp}(\hat{\beta}(\lambda)) \end{cases}$$

Définissons $J = \left\{ j \in \{1,\dots,p\} : |X_j^\top (Y - X\hat{\beta}(\lambda))| = \lambda \right\}$.
On suppose ensuite que la matrice $X_J = [X_j]_{j \in J}$ est de rang plein. Alors la solution de (2) est unique est on a :

$$\hat{\beta}_j(\lambda) = (X_j^\top X_j)^{-1}(X_j^\top Y - \lambda \eta_j) \ \ \forall j \in J.$$

Où $\eta = \text{sign}(X^\top (Y - X\hat{\beta}(\lambda))) \in \{-1; 0; 1\}^p$.

Exemple :(Calcul du chemin des solutions)
On se donne $X = \begin{pmatrix} 2 & 0 & 1 \\ 1 & 2 & 3 \end{pmatrix}$, $Y = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$
Calculons $\|X^\top Y\|_\infty = \max_{i \in \{1,\dots,p\}} |X_i^\top y| = \max\{7; 6; 11\} = 11$.
Donc $\lambda_1 = 11$. Cela implique $s = \text{sign}(\hat{\beta}(\lambda_1)) = (0; 0; 1)$
Calculons l'équation de la droite $\hat{\beta}_3(\lambda)$ au voisinage inférieur de λ_1:
Puisque $\text{sign}(\hat{\beta}_3(\lambda)) = 1$, on a que

$$X_3^\top (Y - X_3 \hat{\beta}_3(\lambda)) = \lambda$$

$$\Leftrightarrow$$

$$\hat{\beta}_3(\lambda) = \frac{-1}{10}\lambda + \frac{11}{10}$$

Ensuite, on calcule le plus petit $\tau > 0$ tel que

$$\begin{cases} |X_1^\top (Y - X\hat{\beta}(\lambda_1 - \tau))| = \lambda_1 - \tau \\ |X_2^\top (Y - X\hat{\beta}(\lambda_1 - \tau))| = \lambda_1 - \tau \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} |X_1^\top (Y - X_3\hat{\beta}_3(\lambda_1 - \tau))| = \lambda_1 - \tau \\ |X_2^\top (Y - X_3\hat{\beta}_3(\lambda_1 - \tau))| = \lambda_1 - \tau \end{cases}$$

Il faut ensuite considérer les deux cas où l'expression dans la valeur absolue est positive ou négative:

Cas positif:

$$\begin{cases} X_1^\top(Y - X_3\hat{\beta}_3(\lambda_1 - \tau)) = \lambda_1 - \tau \\ X_2^\top(Y - X_3\hat{\beta}_3(\lambda_1 - \tau)) = \lambda_1 - \tau \end{cases}$$
$$\Leftrightarrow$$
$$\begin{cases} 7 - 5\left(-\frac{11-\tau}{10} + \frac{11}{10}\right) = 11 - \tau \\ 6 - 6\left(-\frac{11-\tau}{10} + \frac{11}{10}\right) = 11 - \tau \end{cases}$$
$$\Leftrightarrow$$
$$\begin{cases} \tau = 8 \\ \tau = 12.5 \end{cases}$$

Cas négatif:

$$\begin{cases} X_1^\top(-Y + X_3\hat{\beta}_3(\lambda_1 - \tau)) = \lambda_1 - \tau \\ X_2^\top(-Y + X_3\hat{\beta}_3(\lambda_1 - \tau)) = \lambda_1 - \tau \end{cases}$$
$$\Leftrightarrow$$
$$\begin{cases} -7 + 5\left(-\frac{11-\tau}{10} + \frac{11}{10}\right) = 11 - \tau \\ -6 + 6\left(-\frac{11-\tau}{10} + \frac{11}{10}\right) = 11 - \tau \end{cases}$$
$$\Leftrightarrow$$
$$\begin{cases} \tau = 12 \\ \tau = 10.625 \end{cases}$$

Ainsi, on retient $\tau = 8$ et donc $\lambda_2 = 11 - 8 = 3$. Notons qu'il est atteint dans le cas positif et pour X_1. Cela nous indique que $s = \text{sign}(\hat{\beta}(\lambda_2)) = (1; 0; 1)$ On peut donc, encore une fois, calculer les équation des droites $\hat{\beta}_1(\lambda_2)$, $\hat{\beta}_3(\lambda_2)$ au voisinage inférieur de λ_2.
Puisque $s = \text{sign}(\hat{\beta}(\lambda_2)) = (1; 0; 1)$ on a

$$\begin{cases} X_1^\top \left(Y - X_1\hat{\beta}_1(\lambda) - X_3\hat{\beta}_3(\lambda)\right) = \lambda \\ X_3^\top \left(Y - X_1\hat{\beta}_1(\lambda) - X_3\hat{\beta}_3(\lambda)\right) = \lambda \end{cases}$$
$$\Leftrightarrow$$
$$\begin{cases} \hat{\beta}_1(\lambda) = -\frac{1}{5}\lambda + \frac{3}{5} \\ \hat{\beta}_3(\lambda) = \frac{4}{5} \end{cases}$$

Ensuite, une fois encore, on peut calculer le plus petit $\tau > 0$ tel que

$$|X_2^\top \left(Y - X_1\hat{\beta}_1(\lambda_2 - \tau) - X_3\hat{\beta}_3(\lambda_2 - \tau)\right)| = \lambda_2 - \tau$$

(On peut se restreindre à cette seule équation car $s = \text{sign}(\hat{\beta}(\lambda_2)) = (1; 0; 1)$ et que les pentes des deux droites $\hat{\beta}_1(\lambda)$, $\hat{\beta}_3(\lambda)$ sont négatives).

$$\Leftrightarrow$$

$$\begin{cases} X_2^\top Y - X_2^\top X_1 \hat{\beta}_1(\lambda_2 - \tau) - X_2^\top X_3 \hat{\beta}_3(\lambda_2 - \tau) = \lambda_2 - \tau \\ \text{Ou} \\ -X_2^\top Y + X_2^\top X_1 \hat{\beta}_1(\lambda_2 - \tau) + X_2^\top X_3 \hat{\beta}_3(\lambda_2 - \tau) = \lambda_2 - \tau \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} \tau = 3 \\ \text{Ou} \\ \tau = 3 \end{cases}$$

Or $\lambda_2 - \tau = 3 - 3 = 0$. Et on a terminé.
Voyons ce que donne le chemin des solutions:

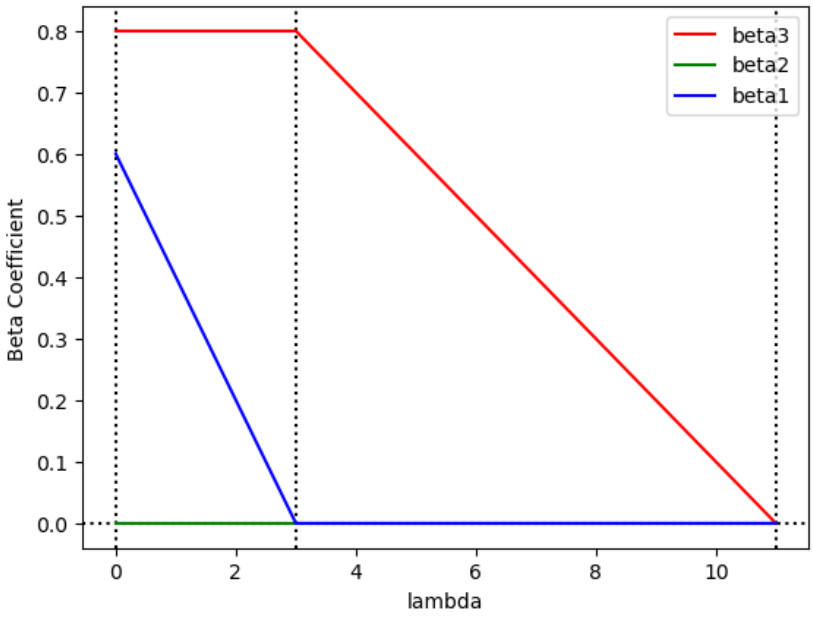

FIG. 2 – *Chemin des solutions:* $\lambda \mapsto \hat{\beta}(\lambda)$

3.2 Algorithme

Algorithme fortement inspiré par [2]. Cet algorithme formalise l'exemple que l'on a vu précédemment.

Algorithm 1: Algorithme d'homotopie pour le LASSO

Input: X, y

Output: La liste $\left(\lambda; \hat{\beta}(\lambda)\right)$

$\rightarrow$ $\lambda = \|X^\top y\|_\infty$.

$\rightarrow$ $J = \{j_0\}$ tel que $|X_{j_0}^\top y| = \lambda$. ($i \in J$ si $\hat{\beta}_i(\lambda) \neq 0$)

while $\lambda > 0$ **do**

- $\rightarrow$ Calculer a,$b \in \mathbb{R}^p$ les pentes et ordonnées à l'origine des composantes affines.
- $\rightarrow$ Trouver le plus petit pas $\tau > 0$ tel que :
  - $\rightarrow$ $\exists j \in J^c$ tel que : $|X_j^\top (y - X\hat{\beta}(\lambda - \tau))| = \lambda - \tau$. Ajouter j à J.
  - $\rightarrow$ $\exists j \in J$ tel que : $\hat{\beta}(\lambda) \neq 0$ et $\hat{\beta}(\lambda - \tau) = 0$. Enlever j de J.
- $\rightarrow$ $\lambda = \lambda - \tau$.
- $\rightarrow$ $\hat{\beta}_i(\lambda) = a_i\lambda + b_i \quad \forall i \in \{1, \ldots, p\}$.
- $\rightarrow$ Garder λ et $\hat{\beta}(\lambda)$ en mémoire.

end

On a vu précédemment que pour que cet algorithme fonctionne correctement il faut que la matrice $(X_J^\top X_J)^{-1}$ soit inversible. Ce qui est une hypothèse raisonnable lorsqu'on travaille avec des données réelles. Cependant, des problèmes peuvent survenir dans les rares cas où le τ trouvé est inférieur à la précision numérique.

Voici un exemple du résultat de l'algorithme sur un jeu de données réelles. ($X \in \mathbb{R}^{71\times 4088}$,$Y \in \mathbb{R}^{71}$)

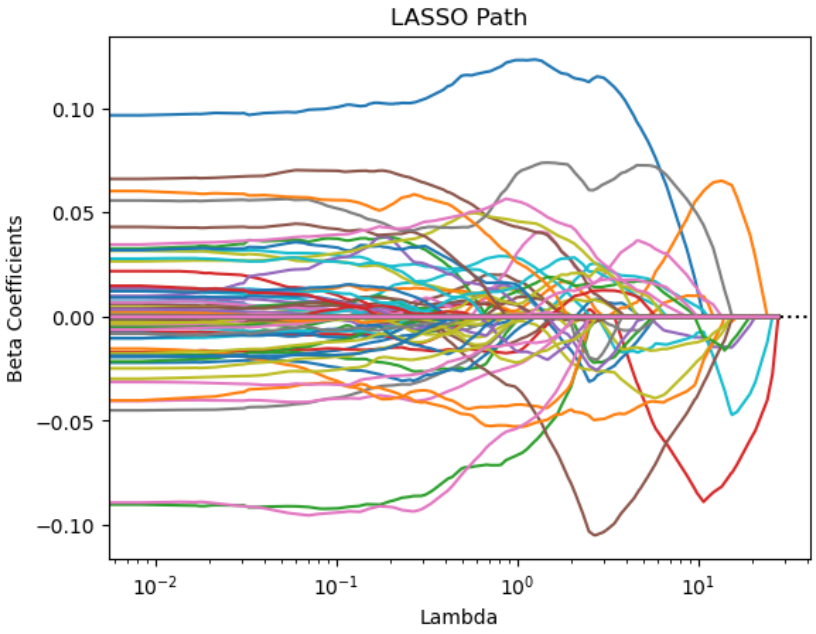

FIG. 3 – *Chemin des solutions:* $\lambda \mapsto \hat{\beta}(\lambda)$

3.3 Application

Désormais, on connaît le chemin des solution du LASSO pour l'exemple considéré. C'est-à-dire la fonction $\lambda \mapsto \hat{\beta}(\lambda)$. Mais il faut encore choisir ce qu'on appelle le coefficient de régularisation λ. Pour cela, on peut utiliser une de ces deux méthodes.

3.3.1 Somme des carrés résiduel

L'idée de cette première méthode est de calculer le chemin des solutions sur une partie de l'échantillon (en pratique 70%) dit d'entraînement (X^{train} et Y^{train}). Puis de minimiser la somme des carrés résiduels sur l'autre partie de l'échantillon dit de test (X^{test} et Y^{test}). C'est-à-dire minimiser la fonction

$$\lambda > 0 \mapsto \|Y^{test} - X^{test}\hat{\beta}(\lambda)\|_2^2.$$

Cela nous permet d'avoir λ_{opt}.
Voyons tout cela:
On sépare le jeu de données en deux parties (entraînement et test)
$X^{train} \in \mathbb{R}^{49\times4088}$ et $X^{train} \in \mathbb{R}^{22\times4088}$, $Y^{train} \in \mathbb{R}^{49}$ et $Y^{test} \in \mathbb{R}^{22}$
On calcule le chemin des solutions sur l'échantillon d'entraînement.

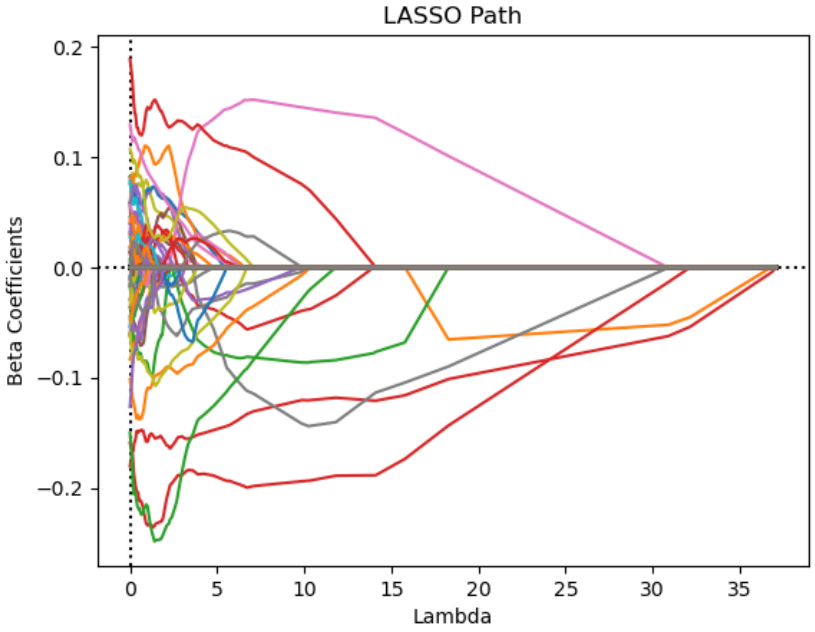

FIG. 4 – *Chemin des solutions:* $\lambda \mapsto \hat{\beta}(\lambda)$ *sur l'échantillon d'entraînement*

Ensuite on minimise la somme des carrés résiduel sur l'échantillon de test.

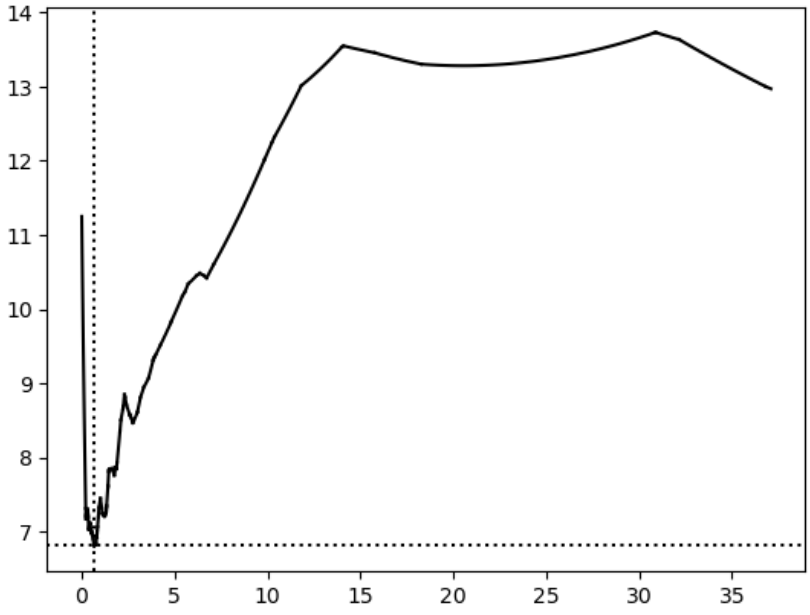

FIG. 5 – $\lambda > 0 \mapsto \|Y^{test} - X^{test}\hat{\beta}(\lambda)\|_2^2$.

On obtient $\lambda_{opt,SCR} = 0.6786279422588005$

Proposition 3.3. La fonction $\lambda > 0 \mapsto \|Y - X\hat{\beta}(\lambda)\|_2^2$ est continue, et quadratique par morceaux.

Démonstration. Montrons que cette fonction est continue. On a vu précédemment que la fonction $\lambda > 0 \mapsto \hat{\beta}(\lambda)$ est affine par morceaux et continue.
Donc $\lambda > 0 \mapsto Y - X\hat{\beta}(\lambda)$ est clairement de même nature que la fonction précédente. De plus, $x \mapsto \|x\|_2$ est une fonction continue (cela se prouve en montrant qu'elle est 1-Lipschitz), de même pour $x \mapsto x^2$. Or la composée de deux fonctions continue est aussi une fonction continue. Et donc

$$\lambda > 0 \mapsto \|Y - X\hat{\beta}(\lambda)\|_2^2 \quad \text{est continue.}$$

Montrons maintenant qu'elle est quadratique par morceaux. Soit $(\lambda_l)_l$ la suite (finie, décroissante) des noeuds du chemin des solution($\lambda_1 = \|X^\top Y\|_\infty$). On prend $\lambda \in [\lambda_{i+1}; \lambda_i]$, pour $i \in \{1, \ldots, L-1\}$ avec $\lambda_L = 0$. Montrons que $\forall \lambda \in [\lambda_{i+1}; \lambda_i]$, $\|Y - X\hat{\beta}(\lambda)\|_2^2$ est quadratique en λ.

On sait qu'entre λ_i et λ_{i+1}, $\lambda \mapsto \hat{\beta}(\lambda)$ est affine. On peut donc écrire

$$\lambda \mapsto \hat{\beta}(\lambda) = \begin{pmatrix} a_{i,1}\lambda + b_{i,1} \\ \vdots \\ a_{i,p}\lambda + b_{i,p} \end{pmatrix}$$

Calcul préliminaire:

$$\begin{aligned} X\hat{\beta}(\lambda) = X \begin{pmatrix} a_{i,1}\lambda + b_{i,1} \\ \vdots \\ a_{i,p}\lambda + b_{i,p} \end{pmatrix} &= \begin{pmatrix} \sum_{k=1}^{p} X_{1,k}(a_{i,k}\lambda + b_{i,k}) \\ \vdots \\ \sum_{k=1}^{p} X_{n,k}(a_{i,k}\lambda + b_{i,k}) \end{pmatrix} \\ &= \lambda \underbrace{\begin{pmatrix} \sum_{k=1}^{p} X_{1,k}a_{i,k} \\ \vdots \\ \sum_{k=1}^{p} X_{n,k}a_{i,k} \end{pmatrix}}_{v_1} + \underbrace{\begin{pmatrix} \sum_{k=1}^{p} X_{1,k}b_{i,k} \\ \vdots \\ \sum_{k=1}^{p} X_{n,k}b_{i,k} \end{pmatrix}}_{v_2} \end{aligned}$$

Ainsi, on a

$$\begin{aligned} \|Y - X\hat{\beta}(\lambda)\|_2^2 &= Y^\top Y - (\lambda v_1 + v_2)^\top(\lambda v_1 + v_2) - 2Y^\top(\lambda v_1 + v_2) \\ &= \lambda^2 \left[\|v_1\|_2^2\right] + \lambda \left[2v_1^\top v_2 - 2Y^\top v_1\right] + Y^\top Y + \|v_2\|_2^2 - 2Y^\top v_2 \end{aligned}$$

D'où, $\forall \lambda \in [\lambda_{i+1}; \lambda_i]$, $\|Y - X\hat{\beta}(\lambda)\|_2^2$ est quadratique en λ. □

3.3.2 Validation croisée

Principe de la méthode: (K-folds) Pour $X \in \mathbb{R}^{n\times p}$ et $Y \in \mathbb{R}^n$

→ On divise l'ensemble $\{1,\dots,n\}$ en K sous-ensembles $F_1,\dots,F_K$ de tailles à peu près égales.

→ Pour $k = 1,\dots,K$

→ On calcule $\lambda > 0 \mapsto \beta(\lambda)$ sur $[X_i]_{i\notin F_k}$ où X_i est la i-ème ligne de X, et $[Y_i]_{i\notin F_k}$ où Y_i est la i-ème composante de Y.

→ On calcule $\epsilon_k(\lambda) = \sum_{i\in F_k}(Y_i - X_i\hat{\beta}(\lambda))^2$

→ Enfin, on minimise $CV(\lambda) = \frac{1}{n}\sum_{k=1}^{K}\epsilon_k(\lambda) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in F_k}(Y_i - X_i\hat{\beta}(\lambda))^2$ pour obtenir λ_{opt}.

On applique cela au même jeu de données que pour la première méthode.

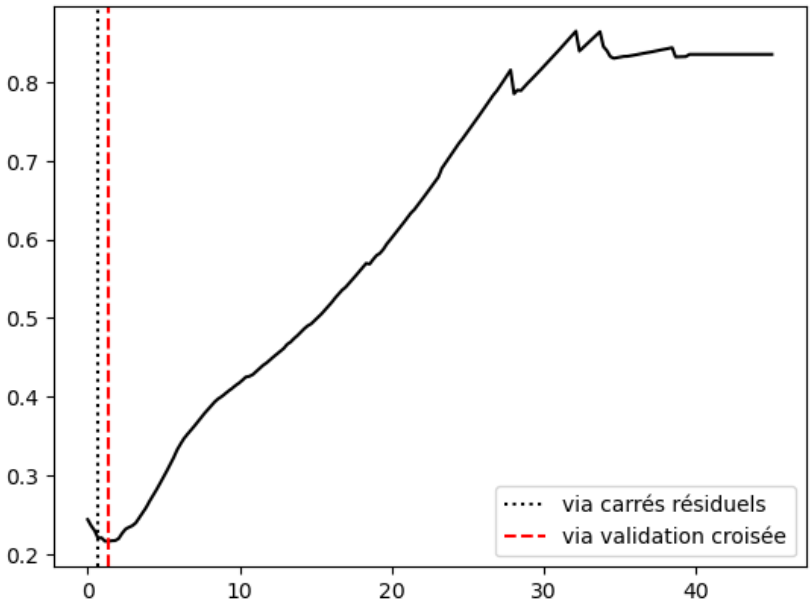

FIG. 6 – $\lambda > 0 \mapsto CV(\lambda)$

On obtient $\lambda_{opt,CV} = 1.3664824120603$, un écart non-négligeable avec la première méthode.
Avec ces deux méthodes, on a obtenu 2 valeurs de λ différentes. On trouve les $\hat{\beta}(\lambda)$ associés via le chemin des solutions calculé sur l'ensemble des données.
On note:

$\rightarrow$ $E_{SCR} = \left\{ i \in \{1, \dots, p\} : \hat{\beta}_i(\lambda_{opt,SCR}) \neq 0 \right\}$

$\rightarrow$ $E_{CV} = \left\{ i \in \{1, \dots, p\} : \hat{\beta}_i(\lambda_{opt,CV}) \neq 0 \right\}$

Et on a:

$\rightarrow$ $Card\,(E_{SCR}) = 62$

$\rightarrow$ $Card\,(E_{CV}) = 52$

Ce qui est normal car avec la validation croisée on a un paramètre de régularisation plus fort que pour l'autre méthode et donc moins de variables sélectionnées.

Remarque 3.1. $E_{CV} \not\subset E_{SCR}$.

Autrement dit, les 52 variables sélectionnées via CV ne sont pas toutes sélectionnées parmi les 62 variables sélectionnées via SCR. En réalité, il y a 44 variables sélectionnées par les deux méthodes.

4 Chemin des solutions du LASSO généralisé

Dans cette partie, on se concentre sur le problème du LASSO généralisé qui est rappelons le :

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \|D\beta\|_1 \right\}$$

Le contenu de cette partie sera très fortement inspiré de [3].
Tout d'abord, parlons des cas où il est possible de se rapporter à un problème du LASSO.

$\rightarrow$ Si $D \in \mathbb{R}^{p \times p}$ carrée et inversible, alors il suffit de poser $\theta = D\beta$ et on obtient

$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - XD^{-1}\theta\|_2^2 + \lambda \|\theta\|_1 \right\}$$

$\rightarrow$ De manière plus générale, si $D \in \mathbb{R}^{m \times p}$ avec $\mathrm{rg}(D) = m$ on peut encore se rapporter à un problème LASSO:

$\rightarrow$ On construit $\tilde{D} = \begin{bmatrix} D \\ A \end{bmatrix} \in \mathbb{R}^{p \times p}$, où $A \in \mathbb{R}^{(p-m) \times p}$ dont les lignes sont orthogonales à celles de D.

→ On pose $\theta = (\theta_1, \theta_2)^\top = \tilde{D}\beta$.

$$\min_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - X\tilde{D}^{-1}\theta\|_2^2 + \lambda \|\theta_1\|_1 \right\} \tag{4}$$

ceci est presque le problème LASSO à l'exception près que la pénalité n'est que sur θ_1.

→ $X\tilde{D}^{-1}\theta = X_1\theta_1 + X_2\theta_2$. Il est clair que la solution du 2ème bloc des coefficients est donné par l'estimateur des moindres carrés

$$\tilde{\theta}_2 = \left(X_2^\top X_2\right)^{-1} X_2^\top (Y - X_1\theta_1) \in \mathbb{R}^{(p-m)}$$

→ Enfin, (4) devient

$$\min_{\theta_1 \in \mathbb{R}^m} \left\{ \frac{1}{2} \|(I - P)Y - (I - P)X_1\theta_1\|_2^2 + \lambda \|\theta_1\|_1 \right\} \tag{5}$$

où $P = X_2 \left(X_2^\top X_2\right)^{-1} X_2^\top$ est la projection sur l'espace engendré par les colonnes de X_2.

→ Cependant, si $D \in \mathbb{R}^{m \times p}$ avec $\mathrm{rg}(D) < m$, de telles transformations ne sont plus possibles et cela suggère le besoin d'un nouvel algorithme.

4.1 Cas $X = I$

Dans cette partie, on se concentre sur le problème suivant:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \|Y - \beta\|_2^2 + \lambda \|D\beta\|_1 \right\} \quad \text{avec} \quad D \in \mathbb{R}^{m \times p} \tag{6}$$

Problème dual:
Le problème (6) est clairement équivalent au problème

$$\min_{\beta \in \mathbb{R}^p, z \in \mathbb{R}^m} \left\{ \frac{1}{2} \|Y - \beta\|_2^2 + \lambda \|z\|_1 \right\} \quad \text{s.c} \quad D\beta = z \tag{7}$$

Afin d'obtenir le problème dual de (7), on doit doit minimiser le Lagrangien de la fonction objectif.

$$L(\beta,z,u) = \frac{1}{2} \|Y - \beta\|_2^2 + \lambda \|z\|_1 + u^\top (D\beta - z)$$

Notons

→ $L_1(\beta) = \frac{1}{2}\|Y - \beta\|_2^2 + u^\top D\beta$

→ $L_2(z) = \lambda\|z\|_1 - u^\top z$

On doit donc minimiser L_1 et L_2 par rapport à β et z respectivement. Premièrement, $\nabla L_1(\beta) = \beta - Y + D^\top u$ donc

$$\nabla L_1(\beta) = 0 \Leftrightarrow \beta = Y - D^\top u$$

Puisque L_1 est convexe, ce β minimise L_1.

$$\begin{aligned} L_1(Y - D^\top u) &= \frac{1}{2}\|Y - Y + D^\top u\|_2^2 + u^\top D(Y - D^\top u) \\ &= \frac{1}{2}\langle D^\top u, D^\top u\rangle + u^\top D(Y - D^\top u) \\ &= \frac{1}{2}\langle D^\top u, D^\top u\rangle + \langle u, DY\rangle - \langle u, DD^\top u\rangle \\ &= -\frac{1}{2}\|D^\top u\|_2^2 + u^\top DY \end{aligned}$$

On remarque que $-\frac{1}{2}\|D^\top u\|_2^2 + u^\top DY$ et $-\frac{1}{2}\|Y - D^\top u\|_2^2$ sont égales, au terme constant (par rapport à u) $-\frac{1}{2}Y^\top Y$ près. Or lorsqu'on minimisera le problème dual, ce terme n'aura pas d'importance ainsi on a:

$$\min_\beta L_1(\beta) = -\frac{1}{2}\|Y - D^\top u\|_2^2$$

Deuxièmement, pour L_2, composante par composante, on note:

$$f(z_i) = \lambda|z_i| - u_i z_i$$

→ Si $z_i \geq 0$, alors $f(z_i) = \lambda z_i - u_i z_i = z_i(\lambda - u_i)$

  → Si $\lambda - u_i \geq 0 \Leftrightarrow \lambda \geq u_i$, alors f est croissante et $\min_{z_i \geq 0} f(z_i) = 0$

  → Si $\lambda - u_i \leq 0 \Leftrightarrow \lambda \leq u_i$, alors f est décroissante et $\min_{z_i \geq 0} f(z_i) = -\infty$

$\rightarrow$ Si $z_i \leq 0$, alors $f(z_i) = -\lambda z_i - u_i z_i = z_i(-\lambda - u_i)$

- $\rightarrow$ Si $-\lambda - u_i \geq 0 \Leftrightarrow \lambda \leq -u_i$, alors f est croissante et $\min_{z_i \leq 0} f(z_i) = -\infty$
- $\rightarrow$ Si $-\lambda - u_i \leq 0 \Leftrightarrow -\lambda \leq u_i$, alors f est décroissante et $\min_{z_i \leq 0} f(z_i) = -\infty$

Cela implique que:

$$\min_{z_i \in \mathbb{R}} \lambda |z_i| - u_i z_i = \begin{cases} 0 & \text{si } |u_i| \leq \lambda \\ -\infty & \text{sinon} \end{cases}$$

Et de manière générale:

$$\min_{z \in \mathbb{R}^m} \lambda \|z\|_1 - u^\top z = \begin{cases} 0 & \text{si } \|u\|_\infty \leq \lambda \\ -\infty & \text{sinon} \end{cases}$$

Ainsi, le problème dual de (7) est :

$$\min_{u \in \mathbb{R}^m} \frac{1}{2} \|Y - D^\top u\|_2^2 \text{ s.c } \|u\|_\infty \leq \lambda \tag{8}$$

Dans la suite, on notera $\hat{\beta}_\lambda$ et $\hat{u}_\lambda$ les solutions respectives du problème primal et dual.

Remarque 4.1. $\hat{\beta}_\lambda$ et $\hat{u}_\lambda$ ne sont pas de même dimension ($m \leq p$)

Enfin, puisqu'on dispose de la relation

$$\hat{\beta}_\lambda = Y - D^\top \hat{u}_\lambda \tag{9}$$

alors si on résout le problème dual on aura résolu le problème primal.

4.1.1 Cas 1d fused-LASSO ($D = D_{1d}$)

Dans ce cas on s'intéresse au problème (8) où

$$D = D_{1d} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ & & & \cdots & & \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{pmatrix} \in \mathbb{R}^{(p-1)\times p}$$

Le problème (6) peut se reécrire de la façon suivante

$$\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\|Y-\beta\|_2^2+\lambda\sum_{i=1}^{p-1}|\beta_{i+1}-\beta_i|\right\}$$

Comme $\mathrm{rg}(D_{1d})=p-1$ le problème (8) est strictement convexe (car $DD^\top$ est définie positive) et admet une unique solution.
On va maintenant développer un algorithme pour calculer le chemin des solution de (8), qui se base sur la proposition suivante :

Proposition 4.1. ([3])
On suppose que $D=D_{1d}$, alors pour tout $i\in\{1,\cdots,p-1\}$ la solution $\hat{u}_\lambda$ du problème (8) satisfait :

$$\begin{cases}\hat{u}_{\lambda_0,i}=\lambda_0\Rightarrow\hat{u}_{\lambda,i}=\lambda\ \ \forall\lambda\in[0,\lambda_0]\\ \text{et}\\ \hat{u}_{\lambda_0,i}=-\lambda_0\Rightarrow\hat{u}_{\lambda,i}=-\lambda\ \ \forall\lambda\in[0,\lambda_0]\end{cases}$$

Démonstration. Pour prouver cette proposition on a besoin du lemme suivant:

Lemme 4.1. On considère la fonction suivante:

$$T_\lambda(x)=\begin{cases}-\lambda & \text{si}\ \ x<-\lambda\\ x & \text{si}\ \ |x|\leq\lambda\\ \lambda & \text{si}\ \ x>\lambda\end{cases}$$

Alors, $\forall\lambda_0,\lambda\in\mathbb{R}$ et $\forall x,y\in\mathbb{R}$ on a

$$|T_{\lambda_0}(x)-T_\lambda(y)|\leq\max\{|x-y|,|\lambda_0-\lambda|\}$$

Démonstration. Simplement en distinguant les cas. □

De manière équivalente, montrons que

$$\|\hat{u}_{\lambda_0}-\hat{u}_\lambda\|_\infty\leq\lambda_0-\lambda$$

Pour cette preuve, on considère l'utilisation de la descente de coordonnées pour trouver la solution $\hat{u}_\lambda$ en commençant au point $\hat{u}_{\lambda_0}$ comme estimation initiale. On va pouvoir suivre simplement comment les itérés changent du fait de la simplicité de la matrice D.

On prend $u^{(0)} = \hat{u}_{\lambda_0}$ et on parcours les coordonnées dans l'ordre $i = 1, \cdots, p-1$. Pour dériver la i-ème mise à jour, on fixe $u_j \ \forall j \neq i$ et on minimise sur u_i.
On va donc minimiser la fonction objectif de (8) seulement par rapport à u_i.
On a donc :

$$\min_{u_i \in \mathbb{R}} \frac{1}{2} \left(Y_i - (u_{i-1} - u_i)\right)^2 + \frac{1}{2} \left(Y_{i+1} - (u_i - u_{i+1})\right)^2 \quad \text{s.c} \quad |u_i| \leq \lambda$$

C'est juste une fonction quadratique contrainte à rester dans un intervalle.
En notant cette fonction $g(u_i)$, on a

$$g(u_i) = u_i^2 + u_i \left[-u_{i+1} + Y_i - Y_{i+1} - u_{i-1}\right] + \frac{1}{2}\left(Y_i^2 + Y_{i+1}^2\right) + \frac{1}{2}\left(u_{i-1}^2 + u_{i+1}^2\right) - Y_i u_{i-1} + Y_{i+1} u_{i+1}$$

Et donc, sans la contrainte le minimum est atteint en

$$\hat{u}_i = \frac{Y_{i+1} - Y_i + u_{i-1} + u_{i+1}}{2}$$

On a plus qu'à prendre en compte la contrainte et on obtient finalement que la mise à jour de la i-ème coordonnée est

$$u_i = T_\lambda \left(\frac{Y_{i+1} - Y_i + u_{i-1} + u_{i+1}}{2} \right)$$

Donc dans la première itération de descente de coordonnées on a :

$$u_i^{(1)} = T_\lambda \left(\frac{Y_{i+1} - Y_i + u_{i-1}^{(1)} + u_{i+1}^{(0)}}{2} \right)$$

Ensuite, en utilisant le fait que $\hat{u}_{\lambda_0}$ est lui-même la solution correspondant à λ_0 on a

$$\begin{aligned}
|\hat{u}_{\lambda_0,i} - u_i^{(1)}| &= |T_{\lambda_0} \left(\frac{Y_{i+1} - Y_i + \hat{u}_{\lambda_0,i-1} + \hat{u}_{\lambda_0,i+1}}{2} \right) - T_\lambda \left(\frac{Y_{i+1} - Y_i + u_{i-1}^{(1)} + u_{i+1}^{(0)}}{2} \right)| \\
&\leq \max \left\{ \left| \frac{\hat{u}_{\lambda_0,i+1} + \hat{u}_{\lambda_0,i-1} - u_{i+1}^{(0)} - u_{i-1}^{(1)}}{2} \right| ; \lambda_0 - \lambda \right\} \quad \text{(Par le lemme)} \\
&= \max \left\{ \left| \frac{\hat{u}_{\lambda_0,i-1} - u_{i-1}^{(1)}}{2} \right| ; \lambda_0 - \lambda \right\}
\end{aligned}$$

Il s'ensuit que $\|\hat{u}_{\lambda_0} - u^{(1)}\|_\infty \leq \lambda_0 - \lambda$. En poursuivant le même raisonnement, on montre que :

$$\|\hat{u}_{\lambda_0} - u^{(k)}\|_\infty \leq \lambda_0 - \lambda \quad \text{Pour toute itération k}$$

Et lorsque $k \longrightarrow \infty$, on a $\|\hat{u}_{\lambda_0} - \hat{u}_\lambda\|_\infty \leq \lambda_0 - \lambda$ □

Remarque 4.2. Plus simplement, cette proposition dit que si le chemin d'une des coordonnées $\hat{u}_i$ heurte les droites $y = \lambda$ ou $y = -\lambda$ alors elles y restent tant que $\lambda \neq 0$

Remarque 4.3. Cette proposition est vraie car la matrice D considérée est bien spéciale, elle n'est pas vraie en général. Cependant, elle reste vraie si la matrice $DD^\top$ est à diagonale dominante

Voici maintenant, en quelques mots, comment va fonctionner l'algorithme qui calcule le chemin des solutions du problème dual (8) :

- → On verra que le chemin des solutions est continu et affine par morceaux.
- → Comme la proposition précédente l'indique, lorsqu'un coefficient atteint la "limite", on sait qu'il y restera jusqu'à $\lambda = 0$. Cela nous permet de ne plus le considérer.
- → A chaque fois qu'un coefficient heurte la limite, le chemin des autres coefficients change de pente.
- → Il suffit donc, à chaque fois qu'un coefficient heurte la limite, de recalculer les pentes des autres coefficients puis de regarder le prochain coefficient qui heurte la limite.
- → Au fil des itérations, on maintiendra deux listes:
  - → $B(\lambda)$ qui contient les indices des coefficients qui sont sur la limite enλ
  - → $s(\lambda)$ qui contient leur signe ($\pm\lambda$)
  - → Exemple: $B(\lambda) = (5,2)$ et $s(\lambda) = (-1,1)$ alors:
    - → $\hat{u}_{\lambda,5} = -\lambda$
    - → $\hat{u}_{\lambda,2} = \lambda$
- → On aura aussi connaissance du chemin des solution du problème primal (6) via la relation (9).

Algorithm 2: Dual path alogrithm for the 1d fused-LASSO

Input: Y

Output: La liste $(\lambda; \hat{u}(\lambda))$

$\rightarrow$ $\lambda_0 = \infty$.

$\rightarrow$ $B = \emptyset$

$\rightarrow$ $s = \emptyset$

for $k \leftarrow 0$ **to** $n-2$ **do**

$\rightarrow$ Calculer la solution en λ_k par les moindres carrés.

$\rightarrow$ Calculer les pentes pour $\lambda \leq \lambda_k$

$\rightarrow$ Calculer λ_{k+1} qui correspond au premier coefficient qui heurte la limite.

$\rightarrow$ Ajouter l'indice de ce coefficient à B et son signe à s

end

Remarque 4.4. On a au maximum $n-1$ itérations car $\hat{u}_\lambda \in \mathbb{R}^{(p-1)}$ et qu'à chaque itération un coefficient heurte la limite. Et lorsqu'ils sont tous sur la limite, c'est fini car on sait qu'ils y resterons.

Donnons quelques éléments de notation avant de poursuivre.
Si $A \in \mathbb{R}^{(p-1)\times p}$ une matrice, et $B \subset \{1, \cdots, p-1\}$ alors A_B désigne la matrice A à laquelle on a seulement garder les lignes indexées par B. De même pour A_{-B}, elle désigne la matrice A à laquelle on à retirer les lignes indexées par $B^{\complement}$.
Voyons maintenant les détails de cet algorithme.
Supposons que l'on soit à l'itération k. On a alors $B = B(\lambda_k)$ et $s = s(\lambda_k)$. Par la proposition 4.1. on a

$$\hat{u}_{\lambda,B} = \lambda s \ \forall \lambda \in [0, \lambda_k]$$

Ainsi, pour $\lambda \leq \lambda_k$ on peut réduire le problème (8) :

$$\min_{u_{-B} \in \mathbb{R}^{p-1-Card(B)}} \frac{1}{2} \|Y - \lambda D_B^\top s - D_{-B}^\top u_{-B}\|_2^2 \ \text{ s.c } \ \|u_{-B}\|_\infty \leq \lambda \qquad (10)$$

Par construction, $\hat{u}_{\lambda_k,-B}$ se situe strictement entre $-\lambda_k$ et λ_k pour chaque coordonnée. Ainsi la contrainte est forcément vérifiée et on résoud (10) sans la contrainte et on obtient l'estimation des moindres carrés :

$$\hat{u}_{\lambda_k,-B} = \left(D_{-B}D_{-B}^\top\right)^{-1} D_{-B}\left(Y - \lambda_k D_B^\top s\right)$$

Or on sait que $\hat{u}_{\lambda_k,-B} = a\lambda - b$ jusqu'à qu'une des coordonnées intérieures heurte la limite. Ce moment peut être trouvé en résolvant, pour chaque $i \in -B$:

$$a_i\lambda - b_i = \pm\lambda \Leftrightarrow \lambda = \frac{a_i}{b_i \pm 1}$$

Après calculs on a

$$t_i = \frac{a_i}{b_i \pm 1} = \frac{\left[\left(D_{-B}D_{-B}^\top\right)^{-1} D_{-B}Y\right]_i}{\left[\left(D_{-B}D_{-B}^\top\right)^{-1} D_{-B}D_B^\top s\right]_i \pm 1}$$

Ainsi, $\lambda_{k+1} = \max_i t_i$ est le premier λ pour lequel un coefficient intérieur heurte la limite. Puis $i_{k+1} = \arg\max_i t_i$ et $s_{k+1} = sign(\hat{u}_{\lambda_{k+1},i_{k+1}})$. Enfin, ajouter i_{k+1} à B et s_{k+1} à s.

Voici un exemple de cet algorithme dans un cas très simple.
Exemple: Pour $Y = \begin{pmatrix}1 & -5 & 15 & -10 & 4 & 7 & -2 & -3\end{pmatrix} \in \mathbb{R}^8$

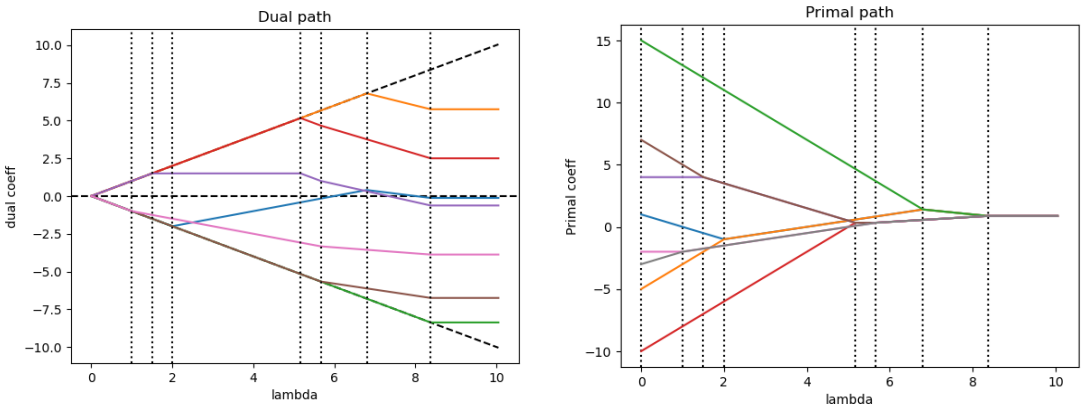

FIG. 7 – *A gauche le chemin des solution du problème dual* (6)*, et à droite le chemin des solutions du problème primal* (8)

On remarque que, pour le chemin du problème primal, lorsque λ est supérieur à une certaine valeur alors on a $\hat{\beta}_1(\lambda) = \hat{\beta}_2(\lambda) = \cdots = \hat{\beta}_p(\lambda)$

Proposition 4.2. Si $\lambda \geq \| \left(DD^\top\right)^{-1} DY\|_\infty$ alors $\hat{\beta}(\lambda) = \overline{Y}\mathbb{1}_p$

Démonstration. Supposons que $\lambda \geq \| \left(DD^\top\right)^{-1} DY\|_\infty$.
Soit $\hat{u}(\lambda)$ la solution du problème dual, par construction, la contrainte est forcément vérifiée. Donc $\hat{u}(\lambda)$ doit être tel que

$$D^\top \hat{u}(\lambda) = P_{\mathrm{Im}(D^\top)}(Y) = D^\top \left(DD^\top\right)^{-1} DY$$

D'où

$$\hat{u}(\lambda) = \left(DD^\top\right)^{-1} DY.$$

Et d'après la relation $\hat{\beta}(\lambda) = Y - D^\top \hat{u}(\lambda)$ on a

$$\begin{aligned}\hat{\beta}(\lambda) &= Y - D^\top \left(DD^\top\right)^{-1} DY \\ &= \left(I_p - D^\top \left(DD^\top\right)^{-1} D\right) Y\end{aligned}$$

Il reste à montrer que :

$$I_p - \underbrace{D^\top \left(DD^\top\right)^{-1} D}_{=A} = \frac{1}{p}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$$

Tout d'abord, $A^\top = A$ et $A^2 = A$, ce qui suffit pour dire que A est une matrice de projection orthogonale. On a donc $Sp(A) = \{0; 1\}$, et l'espace propre associé à la valeur propre 0 est

$$\mathrm{Ker}(A) = \left\{x \in \mathbb{R}^p : D^\top \left(DD^\top\right)^{-1} Dx = 0\right\} = \left\{x \in \mathbb{R}^p : DD^\top \left(DD^\top\right)^{-1} Dx = 0\right\} = \mathrm{Ker}(D).$$

L'espace propre associé à la valeur propre 1 est donc $\mathrm{Ker}(D)^\perp$. Donc $I_p - A$ est le projecteur orthogonal sur $\left(\mathrm{Ker}(D)^\perp\right)^\perp = \mathrm{Ker}(D)$

D'autre part, $\mathrm{Ker}(D) = \mathrm{vect}(\mathbb{1}_p)$. On sait que l'expression matricielle d'une projection orthogonale sur une droite vectorielle est:

$$P_{\mathrm{vect}(\mathbb{1}_p)} = \frac{\mathbb{1}_p(\mathbb{1}_p)^\top}{(\mathbb{1}_p)^\top \mathbb{1}_p} = \frac{1}{p}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$$

Autrement dit, lorsque $\lambda \geq \| \left(DD^\top\right)^{-1} DY\|_\infty$ les $\hat{\beta}_i(\lambda)$ sont égaux et valent $\frac{1}{p}\sum_{i=1}^p Y_i$, la moyenne de Y. □

Références

[1] Taylor B Arnold et Ryan J Tibshirani : Efficient implementations of the generalized lasso dual path algorithm. *Journal of Computational and Graphical Statistics*, 25(1):1–27, 2016.

[2] Julien Mairal et Bin Yu : Complexity analysis of the lasso regularization path. *In Proceedings of the 29th International Conference on Machine Learning*, pages 353–360, 2012.

[3] Ryan J. Tibshirani et Jonathan Taylor : The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335 – 1371, 2011.