Université de Lille 1

Sciences et Technologies

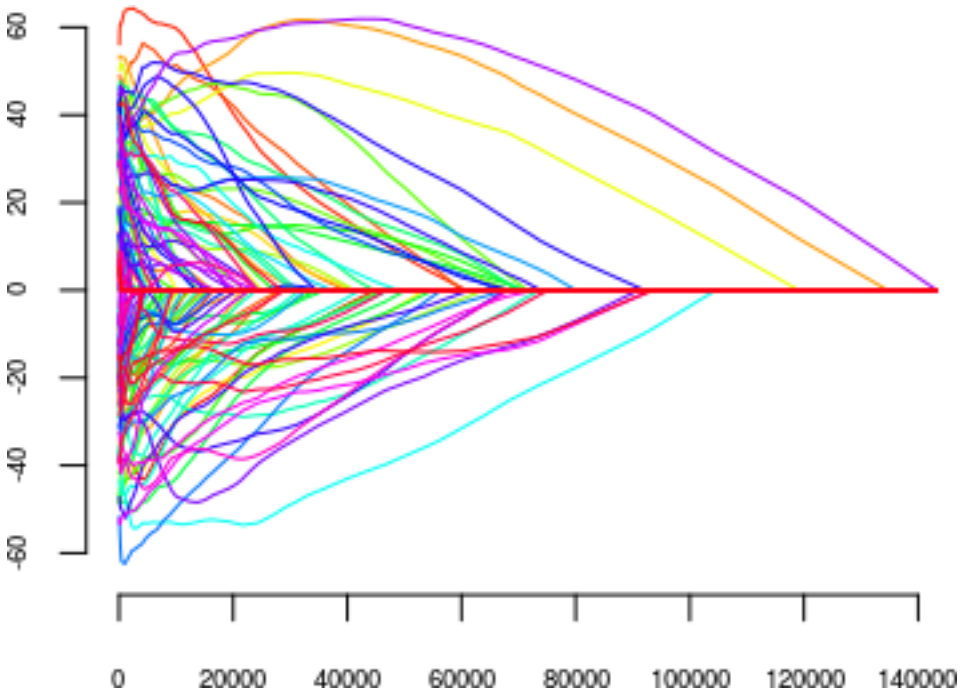
Master 1 ISN 2018

Le LASSO

Least Absolute Shrinkage and Selection Operator

Auteurs :
Nicolas BERNARD
Ahmed FETI

Encadrant :
M. HARDY



Remerciments :

Nous souhaitons remercier notre tuteur M. Adrien Hardy et toute l'équipe pédagogique du Master 1 "Mathématiques Appliquées, Statistiques" qui ,tout au long de cette année, nous ont aidés à maitriser les différents outils permettant la réalisation de ce travail.

Nous remercions aussi nos familles pour leur soutient et notre ami Axel pour son aide avec LaTeX.

Résumé

Le problème général étudié dans ces parties est celui de la régression linéaire en grande dimension. Une méthode populaire pour estimer le paramètre inconnu de la régression dans ce contexte est l'estimateur des moindres carrés pénalisé par la norme ℓ_1 du coefficient, connu sous le nom de LASSO (Tibshirani, 1996).

L'introduction d'une pénalisation réduit la variabilité de l'estimation, améliorant ainsi la précision de prédiction. En outre, la pénalisation de type ℓ_1 rétrécit certains coefficients, alors que les autres sont annulés exactement, aboutissant ainsi à des modèles parcimonieux.

On montrera comment majorer l'erreur estimée de la régression classique et de celle du Lasso, puis on présentera une méthode par homotopie permettant d'obtenir les estimateurs du Lasso. Enfin, à travers des simulations, on mettra en vue les différents estimateurs, les erreurs résiduelles et les limites du Lasso.

Mots-clés : Régression linéaire multiple, Lasso, régression pénalisée, prédiction, sélection de variables, ensemble des variables actives, Inégalité Oracle, méthode homotopique.

Table des matières

Partie 1

Introduction

Dans une régression linéaire multiple, on souhaite expliquer le comportement d'une variable notée $y \in \mathbb{R}$ (variable expliquée) avec comme informations p caractéristiques notées $(x_1, ..., x_p) \in \mathbb{R}^p$ (variables explicatives), c'est-à-dire, de trouver une approximation de y comme combinaison linéaire des caractéristiques :

$$\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p \approx y.$$

Pour cela, on dispose de n observations, de ces p variables explicatives que l'on range dans une matrice $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ et de la variable expliquée rangée dans un vecteur $Y \in \mathbb{R}^n$, et on fait l'hypothèse qu'il existe un vecteur $\beta^0 \in \mathbb{R}^p$ tel que :

$$Y = X\beta^0 + \epsilon, \text{ où le terme d'erreur noté } \epsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Mais β^0 est inconnu, on cherche à l'approcher par $\widehat{\beta}$ en minimisant l'erreur, pour cela on utilise habituellement le critère des moindes carrées :

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}}\{\|Y - X\beta\|_2^2\}.$$

Ce critère n'est valable que lorsque $n \geq p$ (i.e quand on a plus d'individus que de caractéristiques) par exemple : On souhaite estimer l'âge d'un arbre grâce à deux caractéristiques "taille" et "circonférence", on aura facilement accès à plusieurs observations d'arbres dont on connaît l'âge.

Mais quand $p > n$ (i.e quand on a moins d'individus que de caractéristiques) par exemple : Le génome humain comporte plus de 3 milliards de bases que l'on peut voir ici comme des "caractéristiques", or on n'a pas accès au génome de 3 milliards de personnes. Le problème des moindres carrés n'a pas de solution unique du fait

que X^TX n'est pas inversible, d'où l'idée d'introduire une pénalisation en fontion de la norme de β dans la fonction à minimiser.

$$\widehat{\beta}(\lambda) = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}}\{\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\},\ \lambda > 0.$$

En d'autres termes, On cherche l'hyperplan qui passe au mieux entre tous les points. Mais quand on a plus de variables explicatives (dimensions dans laquelles ces points sont exprimés) que de caractéristiques (points), il y a une infinité d'hyperplans qui minimise cette somme.

L'idée du Lasso, notamment en introduisant la norme 1 pour la pénalisation, force certains coefficients de β à être nuls et donc de réduire la dimension des points. Ce qui revient à faire de la sélection de variables explicatives. On remarque que $\widehat{\beta}(\lambda)$ dépend du paramètre λ appelé paramètre de pénalisation.

Dans la Partie 1, nous verrons que dans le cadre standard ($p \leq n$) où $\widehat{\beta}$ désigne la solution optimale du problème des moindres carrés, le contrôle de l'erreur estimée est aisée. Et dans la partie 2, (cadre de la grande dimension $p > n$), nous allons établir l'inégalité Oracle (contrôle de l'erreur estimée pour l'estimateur LASSO) où $\widehat{\beta} = \widehat{\beta}(\lambda)$ représentera cette fois ci, la solution optimale du LASSO. Dans la Partie 3, nous verrons comment obtenir la solution optimale $\widehat{\beta}$ en fonction de λ par la méthode homotopique. Et enfin dans la Partie 4, nous présenterons l'algorithme qui découle de la partie précédente ainsi que quelques exemples.

1.1 Notations

Dans ce document on utilisera les normes vectorielles 0, 1, 2 (euclidienne) et infinie définies ci-dessous :

$$\begin{aligned}
&\forall \beta \in \mathbb{R}^n, n \in \mathbb{N}_+^*,\\
&\|\beta\|_0 = \sum_{i=1}^{n} 1_{\{\beta_i \neq 0\}},\\
&\|\beta\|_1 = \sum_{i=1}^{n} |\beta_i|,\\
&\|\beta\|_2 = (\sum_{i=1}^{n} \beta_i^2)^{\frac{1}{2}},\\
&\|\beta\|_\infty = \max_{1\leq i\leq n} |\beta_i|.
\end{aligned}$$

La matrice $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ contient les données de n individus et de p caractéristiques sur chaque individu. Chaque ligne représente un individu et les différentes caractéristiques qui lui sont associées. On notera l'individu $i \in \{1, 2, ..., n\}$:

$$X_i = (X_{i,1}, X_{i,2}, ..., X_{i,p}) \in \mathbb{R}^{1\times p}.$$

Chaque colonne représente une caractéristique sur tout les individus. On notera l'observation $j \in \{1, 2, ..., p\}$:

$$X^{(j)} = (X_{1,j}, X_{2,j}, ..., X_{n,j}) \in \mathbb{R}^n.$$

Cette section introduit le modèle linéaire multidimensionnel dans lequel une variable quantitative est expliquée, modélisée, par plusieurs variables quantitatives.

1.2 Modélisation

Une variable quantitative Y dite à expliquer est mise en relation avec p variables quantitatives $X_1, ..., X_p$ dite explicatives. Les données sont supposées provenir d'une observation d'un échantillon statistique de taille n de $\mathbb{R}^p$,

$$(x_{i1}, Y_1), ..., (x_{ip}, Y_p) \text{ où } i = 1, ..., n.$$

L'écriture matricielle du modèle linéaire dans cette situation conduit à supposer que l'espérance de Y appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{X_1, ..., X_p\}$, c'est-à-dire, les variables aléatoires vérifient :

$$Y_i = \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i, \ i = 1, ..., n.$$

avec les hypothèses suivantes :

Hypothèse 1. Les ε_i sont des termes d'erreur, non observés, indépendants et identiquement distribués (i.i.d) : $\mathbb{E}(\varepsilon_i) = 0$, $\mathbf{Var}(\varepsilon) = \sigma^2 I_n$. On considère la normalité de la variable d'erreur $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Les ε_i sont alors i.i.d de loi $\mathcal{N}(0, \sigma^2)$.

Hypothèse 2. Les termes $X^{(j)}$ sont supposés déterministes (facteur contrôlé) ou bien, l'erreur est indépendante de la distribution conjointe de $\{X_1, ..., X_p\}$. On écrit dans ce dernier cas de figure que :

$$\begin{cases} \mathbb{E}[Y|X_1, ..., X_p] = \beta_1 X_1 + ... + \beta_p X_p, \\ \mathbf{Var}[Y|X_1, ..., X_p] = \sigma^2. \end{cases}$$

Hypothèse 3. Les paramètres $\beta_1, ..., \beta_p$ sont inconnus et supposés constants. Les données sont rangées dans une matrice $X \in \mathcal{M}_{n\times p}(\mathbb{R})$ de terme général x_{ij} et dans un vecteur Y de terme général Y_i.

Définition 1.1. Modèle de la Régression : Un modèle de la régression linéaire est défini par une équation de la forme :

$$Y = X\beta + \varepsilon,$$

où :

- $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{pmatrix}$ est un vecteur aléatoire de dimension n,
- $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}$ est le vecteur de dimension n des erreurs,
- $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_p \end{pmatrix}$ est le vecteur de dimension p des paramètres inconnus du modèle,
- $X = \begin{pmatrix} x_{11} & x_{12} & \dots & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & \dots & x_{2p} \\ \vdots & & & & \\ \vdots & & & & \\ & & & & \vdots \\ & & & & \vdots \\ x_{n1} & \dots & \dots & x_{np-1} & x_{np} \end{pmatrix}$ est une matrice de taille $n \times p$ connue,

appelée matrice du plan d'expérience (ou matrice des données).

Les hypothèses concernant le modèle sont :

$$(\mathcal{H}) \begin{cases} (\mathcal{H}1) : rg(X) \leq min(n,p), \\ (\mathcal{H}2) : \mathbb{E}(\varepsilon) = 0, \mathbf{Var}(\varepsilon) = \sigma^2 I_n. \end{cases}$$

L'hypothèse $(\mathcal{H}2)$ signifie que les erreurs sont centrées, de même variance et non corrélées entres elles.

1.3 Cadre standard : estimation des Moindres Carrés Ordinaire

Nous présentons ci-après une méthode populaire pour estimer le paramètre β^0, l'estimateur des moindres carrés (MCO). Elle consiste à chercher une valeur $\widehat{\beta}$ du paramètre qui minimise la somme des carrés des résidus.

1.3.1 Expression de l'estimateur des MCO

Définition 1.2. (Estimateur des MCO)
L'estimateur des Moindres Carrés Ordinaires $\widehat{\beta}$ est défini comme suit :

$$\widehat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \{\|Y - X\beta\|_2^2\}. \tag{1.1}$$

Dans cette partie, on suppose que le rang de la matrice X de l'échantillon est p, et que le nombre d'observations n est supérieur au nombre de caractéristiques p, c'est à dire, $p \leq n$. Sous ces conditions, les colonnes $X^{(1)}, ..., X^{(p)}$ de la matrice X sont linéairement indépendantes. La fonction qui à β associe $\|Y - X\beta\|_2^2$ est strictement convexe, et donc admet une unique solution notée $\widehat{\beta}$.

Proposition 1.1. ***(Expression de** $\widehat{\beta}$**)***
L'estimateur $\widehat{\beta}$ des Moindres Carrés Ordinaires a pour expression :

$$\widehat{\beta} = (X^T X)^{-1} X^T Y,$$

et la matrice de projection H s'écrit : $H = X(X^T X)^{-1} X^T$.

Dans ce cas de figure, l'estimateur de MCO est bien explicite, et linéaire en Y.
Remarque.
L'hypothèse $(\mathcal{H}1)$ ie $rg(X) = p$, assure que la matrice $X^T X$ est bien inversible.

Démonstration. La matrice X^TX est bien symétrique car $(X^TX)^T = X^TX$. De plus, pour tout $\beta \in \mathbb{R}^p$, on a $\beta^T(X^TX)\beta = \|X\beta\|_2^2 \geq 0$, donc la matrice X^TX est positive. En outre, $\beta^T(X^TX)\beta = 0 \Leftrightarrow \|X\beta\|_2^2 = 0 \Leftrightarrow X\beta = 0$, d'où $\beta = 0$ puisque $rg(X) = p$. Autrement dit, la matrice symétrique X^TX est définie positive, donc inversible. □

De plus, en faisant l'hypothèse selon laquelle il existe un vecteur $\beta^0 \in \mathbb{R}^p$, (Foster George (1994) [2]) le vrai paramètre du modèle tel que

$$Y = X\beta^0 + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

alors nous pouvons montrer qu'on peut contrôler l'erreur estimée du modèle, c'est-à-dire, obtenir une majoration de l'erreur de la forme :

$$\frac{\|X(\hat{\beta} - \beta^0)\|_2^2}{n} \leq const.\frac{p}{n}\sigma^2 \;\; \text{où} \;\; const \in \mathbb{R}.$$

1.3.2 Contrôle de l'erreur estimée

Définition 1.3. L'erreur estimée $\widehat{\varepsilon}$ est par définition, l'écart entre des données expérimentales ($X\widehat{\beta}$ le paramètre estimé) et des données calculées à l'aide d'un modèle quantitatif ($X\beta^0$ le vrai paramètre du modèle) en norme 2. C'est une fonction des paramètres de ce modèle, qui peut donc être réduite, minimisée, par l'optimisation de ces derniers. Cela se traduit mathématiquement par :

$$\widehat{\varepsilon} = X(\widehat{\beta} - \beta^0) = H\varepsilon.$$

Les séries des lemmes et propositions qui suivent, vont pouvoir nous aider à faire une étude sur l'estimateur de l'erreur, et enfin obtenir un contrôle de celle-ci.

Proposition 1.2. *La variable aléatoire $\frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{\sigma^2}$, qu'on posera T dans la suite du problème, suit une loi de Khi-deux à p degré de liberté.*

Démonstration. Commencons d'abord par montrer que le vecteur $\frac{X(\hat{\beta}-\beta^0)}{\sigma^2}$ suit une loi normale d'espérance nulle et de variance $H \in \mathcal{M}_n(\mathbb{R})$, où $H = X(X^TX)^{-1}X^T$.

$$\text{i.e } \frac{X(\hat{\beta} - \beta^0)}{\sigma^2} \sim \mathcal{N}(0, H).$$

En effet,

$$\begin{aligned} X(\widehat{\beta}-\beta^0) = X\widehat{\beta} - X\beta^0 &= X(X^TX)^{-1}X^TY - X\beta^0 \\ &= X(X^TX)^{-1}X^T(X\beta^0+\varepsilon) - X\beta^0 \\ &= X(X^TX)^{-1}(X^TX)\beta^0 + X(X^TX)^{-1}X^T\varepsilon - X\beta^0 \\ &= (X^TX)^{-1}X^T\varepsilon \\ &= H\varepsilon \in \mathcal{M}_n(\mathbb{R}). \end{aligned}$$

Par ailleurs, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, on en déduit alors que par combinaison linéaire, le vecteur $H\varepsilon$ suit une loi normale d'espérance nulle et de variance $\sigma^2(H^T I_n H) = \sigma^2 H$ car $H^T H = H$. Par conséquent, le vecteur $X(\hat{\beta} - \beta^0)$ est bien un vecteur aléatoire suivant une loi normale d'espérance nulle et de variance $\sigma^2 H$. □

Rappel sur les projecteurs : *Soit P une matrice carré de taille n. On dit que P est une matrice de projection si $P^2 = P$. Si en plus de vérifier $P^2 = P$, la matrice P est symétrique ie $P^T = P$, alors P est la projection orthogonale de $x \in \mathbb{R}^n$ sur* $\mathrm{Im}(P)$ *parallèlement à* $\mathrm{Ker}(P)$*, c'est à dire que dans la décomposition $x = Px + (x - Px)$, les vecteurs Px et $x - Px$ sont orthogonaux. Par ailleurs, selon le théorème spectral, toute matrice symétrique réelle étant diagonalisable en base orthonormée, il existe alors une matrice orthogonale Q (i.e $QQ^T = I_n$, ce qui signifie que les colonnes de Q forment une base orthonormée de $\mathbb{R}^n$) et une matrice diagonale Δ telles que $P = Q\Delta Q^T$. On voit alors que les termes diagonaux de Δ sont composés de p "1" et de $(n-p)$ "0", où p est la dimension de* $\mathrm{Im}(P)$*, espace sur lequel on projette.*

Démonstration. Revenons à la preuve de la proposition 1.1.

La matrice H vérifie les deux propriétés qu'on vient de rappeler :

• $H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$. Donc la matrice H est une matrice de projection et $\mathrm{Sp}(H) = \{0, 1\}$ où la notation $\mathrm{Sp}(H)$ désigne le spectre de H.

• H vérifie également, $H^T = H$, c'est à dire, H est une matrice symétrique. Autrement dit, H est la matrice de projection orthogonale de $x \in \mathbb{R}^n$ sur $\mathrm{Im}(H)$ parallèlement à $\mathrm{Ker}(H)$. Ce qui précède assure également que : $\dim(\mathrm{Im}(H)) = \mathrm{Tr}(H) = p$ et $\mathrm{Tr}(I_n - H) = n - p$. En d'autres termes, les vecteurs Hx et $(I_n - H)x$ sont orthogonaux. Dès lors, en vertu du théorème de Cochran :

• Les vecteurs aléatoires $H\varepsilon$ et $(I_n - H)\varepsilon$ sont indépendants et de lois respectives $\mathcal{N}(0, \sigma^2 H)$ et $\mathcal{N}(0, \sigma^2(I_n - H))$.

• Les variables aléatoires réelles $\frac{\|H\varepsilon\|^2}{\sigma^2}$ et $\frac{\|(I_n - H)\varepsilon\|^2}{\sigma^2}$ sont indépendantes et de lois respectives $\chi^2(p)$ et $\chi^2(n-p)$.

Ainsi, en posant $T = \frac{\|H\varepsilon\|^2}{\sigma^2} = \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{\sigma^2}$ alors $\boxed{T \sim \mathcal{X}^2(p)}$. □

Lemme 1.4. *L'erreur résiduelle* $\tilde{T_n} = \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{n}$ *est une variable aléatoire d'espérance* $\mathbb{E}(\tilde{T_n}) = \frac{\sigma^2}{n}p$ *et de variance* $\mathbf{Var}(\tilde{T_n}) = 2p\frac{\sigma^4}{n^2}$.

Démonstration. D'après ce qui prècède, on vient de montrer que la variable aléatoire $T = \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{\sigma^2} \sim \mathcal{X}^2(p)$ et $\tilde{T_n} = \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{n} = T\frac{\sigma^2}{n}$, alors $\mathbb{E}(\tilde{T_n}) = \mathbb{E}(T) \times \frac{\sigma^2}{n} = \frac{\sigma^2}{n}p$ et $\mathbf{Var}(\tilde{T_n}) = \mathbf{Var}(T) \times \frac{\sigma^4}{n^2} = 2\frac{\sigma^4}{n^2}p$.

Ainsi : $\boxed{\mathbb{E}(\tilde{T_n}) = p\frac{\sigma^2}{n}}$ et $\boxed{\mathbf{Var}(\tilde{T_n}) = 2p\frac{\sigma^4}{n^2}}$. □

Après avoir donné quelques propriétés de l'erreur estimée sur son espérance et sa variance, nous allons pouvoir les exploiter afin d'obtenir une majoration de celle-ci.

Corollaire 1.1. *L'erreur résiduelle* $\tilde{T_n} = \frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{n} = \mathcal{O}_{\mathbb{P}}(\frac{p}{n})$,

où on utilisera la notation suivante : $Z_n = \mathcal{O}_{\mathbb{P}}(a_n)$ *signifie que* $\forall \varepsilon > 0, \exists M :$ $\mathbb{P}(|Z_n/a_n| > M) \leq \varepsilon$.

Démonstration. Soit $\varepsilon > 0$, on pose $Z_n = \tilde{T_n}\mathbb{E}(\tilde{T_n}) = \tilde{T_n} - p\frac{\sigma^2}{n}$. Donc par application de l'inégalité de Bienaymé-Tchebychev, on a que :

$$\forall \alpha > 0, \mathbb{P}(|Z_n| > \alpha) = \mathbb{P}(|\tilde{T_n} - \mathbb{E}(\tilde{T_n})| > \alpha) \leq \frac{\mathbf{Var}(\tilde{T_n})}{\alpha^2} = 2p\frac{\sigma^4}{n^2\alpha^2}.$$

Par ailleurs, pour tout $\eta > 0$ et avec $\alpha = \eta\frac{p}{n}$, on a que :

$$\tilde{T_n} \geq (\sigma^2 + \eta)\frac{p}{n} \Longrightarrow |\tilde{T_n} - \sigma^2\frac{p}{n}| \geq \eta\frac{p}{n}$$

Donc $\forall \eta > 0, \mathbb{P}(\tilde{T_n} \geq (\sigma^2 + \eta)\frac{p}{n}) \leq \mathbb{P}(|\tilde{T_n} - \sigma^2\frac{p}{n}| \geq \eta\frac{p}{n}) \leq 2\frac{\sigma^4}{p\eta^2}$,

$$\Longrightarrow \forall \eta > 0, \mathbb{P}(\tilde{T_n} \geq (\sigma^2 + \eta)\frac{p}{n}) \leq 2\frac{\sigma^4}{p\eta^2}.$$

En particulier, pour tout $\varepsilon > 0$ et avec $\eta = \sigma^2\sqrt{\frac{2}{p\varepsilon}} > 0$, on a alors :

$$\mathbb{P}(\tilde{T_n} \geq (1 + \sqrt{\frac{2}{p\varepsilon}})\sigma^2) \leq \varepsilon.$$

Ainsi en posant $M = (1 + \sqrt{\frac{2}{p\varepsilon}})$ et $a_n = \frac{p}{n}\sigma^2$, nous en concluons que :

$$\forall \varepsilon > 0, \exists M > 0 \ ; \ \mathbb{P}(|\tilde{T_n}/a_n| > M) \leq \varepsilon.$$

□

Dès lors, il existe une constant $const \in \mathbb{R}$ tel que l'erreur estimée est majorée par $const \times \frac{p}{n}\sigma^2$:

$$\boxed{\frac{\|X(\hat{\beta}-\beta^0)\|_2^2}{n} \leq const.\frac{p}{n}\sigma^2.}$$

1.4 Conclusion

Nous venons de voir dans cette première partie que lorsque $rg(X) = p$ (cas où $p \leq n$), la fonction $\|Y - X\beta\|_2^2\|$ est strictement convexe, donc la matrice $X^T X$ est inversible. Ainsi le problème (1.1) admet une unique solution $\widehat{\beta}$, et celle-ci est bien explicite. Cela nous a permis d'approcher les paramètres estimés $X\widehat{\beta}$ par les vrais paramètres du modèle $X\beta^0$. Autrement dit, l'erreur estimée a pu être contrôlée par le terme $const \times \frac{p}{n}\sigma^2$, c'est-à-dire, l'erreur de prédiction est : $\mathcal{O}_{\mathbb{P}}(\frac{p}{n}\sigma^2)$.

Cependant, quand $rg(X) < p$ (ie. $p > n$), c'est-à-dire, quand il y a peu d'individus mais beaucoup d'informations sur chaque individu, la matrice $X^T X$ n'est plus inversible. Le problème des MCO admet alors une infinité de solutions. Par exemple, pour une solution $\widehat{\beta}$, la quantité $\widehat{\beta} + \eta$ est aussi solution pour tout $\eta \in null(X)$.

En outre, la non-unicité de la solution $\widehat{\beta}$ induit une réelle difficulté sur l'analyse et l'interprétation des solutions. Pour y remédier, des nombreux estimateurs à la fois stables, fiables dont l'interprétation (notamment sur l'erreur estimée) est aisée, ont été mise en place. L'estimateur répondant à ces critères est l'estimateur Lasso (Least Absolute Shrinkage and Selection Operator).

Partie 2

Contrôle de l'erreur estimée du LASSO (oracle inequality)

Dans le cadre de la grande dimension où le nombre d'individus (ou variable) p est supérieur au nombre d'observation n, i.e $n < p$ (contrairement au cadre Standard $n > p$), le rang de la matrice X d'échantillon est strictement inférieur à p ie $rg(X) < p$. Par conséquent, la matrice $X^T X$ n'est plus inversible. Dès lors l'estimateur des MCO n'est plus unique.

2.1 Principe général de la régression pénalisée

Les variables X_i n'étant pas toutes pertinentes, l'objectif est d'éliminer les variables inutiles et uniquement celles-ci. L'idée du LASSO est donc non pas de faire une régression linéaire classique (multiple), mais une régression régularisée qui rend nuls certains coefficients de l'estimation de β.

2.1.1 Estimation

Définition 2.1. Un estimateur par minimisation du risque empirique régularisé (pour la perte quadratique) est, dans le cadre de la régression linéaire, défini par :

$$\widehat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \{ \frac{1}{n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_q^q \}.$$

avec λ un paramètre positif, appelé paramètre de régularisation et $1 \leq q \leq +\infty$.

Remarque :
Selon la norme choisie, on a :
• Pour $q = 2$, on a la régression de Ridge.
• Pour $q = 1$, on a la régression LASSO.

En d'autres termes, (cas où $q = 1$), cette méthode d'estimation introduite par Tibshirani (1996) est définie comme étant un minimiseur du critère des moindres carrés pénalisés par la norme ℓ_1 du vecteur β. On parle de pénalité ℓ_1 ou de pénalité LASSO.

Définition 2.2. L'estimateur **LASSO** (Least Absolute Selection and Shrinkage Operator) est défini pour $\lambda \geq 0$ par :

$$\widehat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}}\{\frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1\}. \tag{2.1}$$

On pose l'application, $\mathcal{F}_\lambda : \beta \mapsto \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$. L'objectif de cette partie sera donc d'étudier la fonction $\mathcal{F}_\lambda$, qui va nous permettre, en amont, de donner quelques propriètés de l'estimateur LASSO "$\widehat{\beta}(\lambda)$", et en aval, d'établir l'inégalité Oracle (Oracl Inequality),

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq \text{const.}\frac{\sigma^2 \log(p)}{n} s_0.$$

qui fait intervenir le cardinal s_0 d'un nouvel ensemble S_0 qu'on appelera "Activ Set"(l'ensemble des variables actives). Ainsi, à la connaissance de S_0, on aura une erreur de prédiction d'ordre : $\mathcal{O}_{\mathbb{P}}(\frac{s_0}{n}) \longrightarrow_{\mathbb{P}} 0$.

Les différentes parties qui suivent, vont nous aider à établir cette inégalité, et ainsi prouver l'ordre de l'erreur estimée.

Proposition 2.1. *La fonction $\mathcal{F}_\lambda$ est convexe, non différentiable.*

On montrera également que la solution du problème (2.1) de minimisation ne peut pas être unique.

Démonstration. Soient $\beta_1, \beta_2 \in \mathbb{R}^p$ et $t \in [0, 1]$,

$$\mathcal{F}_\lambda(t\beta_1 + (1-t)\beta_2) = \frac{1}{n}\|Y - X(t\beta_1 + (1-t)\beta_2)\|_2^2 + \lambda\|t\beta_1 + (1-t)\beta_2\|_1. \tag{2.2}$$

Alors par application de l'inégalité triangulaire pour le deuxième terme de (2.2), on a que :

$$\|t\beta_1 + (1-t)\beta_2\|_1 \leq t\|\beta_1\| + (1-t)\|\beta_2\|_1. \tag{2.3}$$

Et d'autre part,

$$\begin{aligned}
&\|Y - X(t\beta_1 + (1-t)\beta_2)\|_2^2 \\
=&\|t(Y - X\beta_1) + (1-t)(Y - X\beta_2)\|_2 \\
=&t^2\|Y - X\beta_1\|_2^2 + (1-t)^2\|Y - X\beta_2\|_2^2 + 2t(1-t) < Y - X\beta_1, Y - X\beta_2 > \\
\leq&t^2\|Y - X\beta_1\|_2^2 + (1-t)^2\|Y - X\beta_2\|_2^2 + 2t(1-t)\|Y - X\beta_1\|_2\|Y - X\beta_2\|_2 \qquad (2.4)\\
\leq&t^2\|Y - X\beta_1\|_2^2 + (1-t)^2\|Y - X\beta_2\|_2^2 + t(1-t)(\|Y - X\beta_1\|_2^2 + \|Y - X\beta_2\|_2^2) \qquad (2.5)\\
=&t\|Y - X\beta_1\|_2^2 + (1-t)\|Y - X\beta_2\|_2^2. \qquad (2.6)
\end{aligned}$$

Dans les inégalités (2.4) et (2.5), nous avons utilisé respectivement l'inégalité de Cauchy-Schwarz ainsi que l'inéglaité suivante : $uv \leq \frac{1}{2}(u^2 + v^2)$ pour tout $u, v \in \mathbb{R}$. Ainsi :

$$\|Y - X(t\beta_1 + (1-t)\beta_2)\|_2^2 \leq t\|Y - X\beta_1\|_2^2 + (1-t)\|Y - X\beta_2\|_2^2. \tag{2.7}$$

Il en vient qu'en exploitant les inégalités (2.3) et (2.6) :

$$\mathcal{F}_\lambda(t\beta_1 + (1-t)\beta_2) \leq t\mathcal{F}_\lambda(\beta_1) + (1-t)\mathcal{F}_\lambda(\beta_2).$$

$\Longrightarrow$ la fonction $\mathcal{F}_\lambda$ est bien convexe. Puisque $p > n$ alors $rg(X) < p$ (et donc la matrice X^TX est non inversible). En outre, la matrice Hessienne de $\mathcal{F}_\lambda$ qui est X^TX, n'est pas définie-positive, donc on en déduit que la fonction $\mathcal{F}_\lambda$ n'est pas strictement convexe. Par consèquent, le problème (2.1) n'admet pas une unique solution optimale. De plus, étant donné que pour tout $x \in \mathbb{R}^p$ la fonction norme 1, $x \mapsto \|x\|_1$, n'est pas différentiable sur $\mathbb{R}^p$ alors on en déduit que $\mathcal{F}_\lambda$ ne l'est également pas. □

On vient ainsi de prouver que le Lasso n'a pas nécessairement une solution unique car le problème (2.1) n'est pas strictement convexe. En révanche, la valeur de $X\widehat{\beta}(\lambda)$ qui minimise la fonction objective $\mathcal{F}_\lambda$, elle, est unique ; car la fonction objective $\mathcal{F}_\lambda$ est strictement convexe en $X\beta$. Enfin, l'estimateur Lasso n'a pas de forme explicite.

2.1.2 Propriétés de l'estimateur Lasso

Il se trouve que, pour une valeur de $\lambda \in \mathbb{R}_+$ donnée, il existe un $t \in \mathbb{R}_+$ unique tel que le Lasso est équivalent à résoudre le problème suivant :

$$\underset{\beta\in\mathbb{R}^p}{\text{Argmin}}\{\tfrac{1}{n}\|Y-X\beta\|_2^2\} \textit{ tel que } \|\beta\|_1 \leq t.$$

Proposition 2.2. *Minimiser la fonction $\mathcal{F}_\lambda(\beta) = \frac{1}{n}\|Y-X\beta\|_2^2 + \lambda\|\beta\|_1$, où $\beta \in \mathbb{R}^p$, est équivalent à minimiser $\|Y-X\beta\|_2^2$ sous la contrainte de la forme $\|\beta\|_1 \leq t$ pour un t convenablement choisi, c'est à dire,*

$$\widehat{\beta}(\lambda) = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}}\{\frac{1}{n}\|Y-X\beta\|_2^2 + \lambda\|\beta\|_1\} \Longleftrightarrow \widehat{\beta}(t) = \underset{\beta\in\mathbb{R}^p, \|\beta\|_1\leq t}{\text{Argmin}}\{\frac{1}{n}\|Y-X\beta\|_2^2\}.$$

Démonstration. Il suffit de prendre $t = \|\widehat{\beta}(\lambda)\|_1$ car pour tout $\beta \in \mathbb{R}^p$ tel que $\|\beta\|_1 \leq t$, on a alors pour $\lambda \geq 0$, $\lambda\|\beta\|_1 \leq \|\widehat{\beta}(\lambda)\|_1$. Et par définition de $\widehat{\beta}(\lambda)$, $\frac{1}{n}\|Y - X\widehat{\beta}(\lambda)\|_2^2 \leq \frac{1}{n}\|Y-X\beta\|_2^2$ [1] □

Le paramètre λ contrôle la puissance de la régularisation.

Propriété 2.1. *Si $\lambda = 0$, le Lasso correspond à une régression linéaire classique ; on retrouve l'estimateur des Moindres Carrés (si $p \leq n$), ie $\widehat{\beta}(\lambda) = \widehat{\beta}$. La méthode du Lasso sélectionne les variables sans exception.*

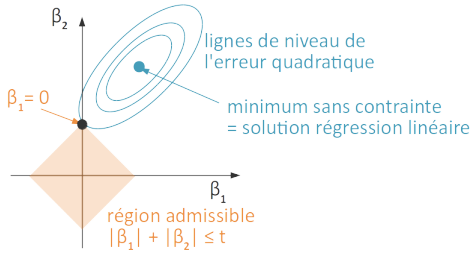
Propriété 2.2. *Si λ tend vers l'infini, tous les coefficients de $\widehat{\beta}$ sont nuls : pour $j = 1, ..., p$, $\widehat{\beta}_j = 0$, c'est-à-dire, $\widehat{\beta}_j(\lambda) = 0 \Longrightarrow \widehat{\beta}(\lambda) = 0$. Dans ce cas, le Lasso ne sélectionne aucune variable explicative.*

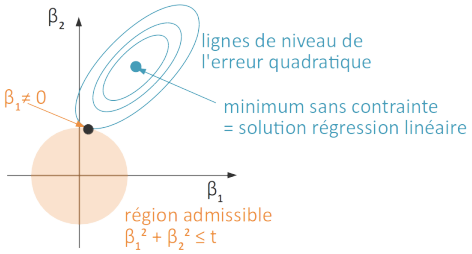
Propriété 2.3. *Si $\lambda \in]0, +\infty[$, le nombre de variable sélectionnées par le Lasso diminue lorsque devient grand, c'est-à-dire, si λ est grand, la contrainte exercé sur le vecteur β l'est également. En d'autres termes, l'augmentation du paramètre λ induit la diminution de certains coefficients de $\widehat{\beta}(\lambda)$ vers 0 jusqu'à ce qu'ils soient exactement nuls.*

La solution obtenue est dite parcimonieuse (creuse) car elle comporte un grand nombre de coefficients nuls. Instinctivement, Cela nous permet de comprendre pourquoi le Lasso, dans la plupart des cas, rend exactement nuls certains coefficients de $\widehat{\beta}$. La figure du gauche illustre en dimension 2 le cas où le Lasso annule une coordonnée.

2.1.3 Interprétation géométrique

Figure 1 : les deux figures ci-dessous donnent une comparaison de la régression Lasso et celle du Ridge avec les différentes contraintes exercées.





Géométriquement, cela signifie que la solution du Lasso est un point situé à l'intersection d'une ligne de niveau du terme d'erreur $\frac{1}{n}\|Y - X\beta X\|_2$ et de la région $\|\beta\|_1 \leq t$ dite "admissible", c'est-à-dire, la région de $\mathbb{R}^p$ où la contrainte est verifiée.

De plus, puisque le terme d'erreur sus-mentionné est quadratique en β alors la ligne de niveau est une ellipsoide. Par ailleurs, la région admissible $\|\beta\|_1 \leq t$ est une boule de l_1 de rayon t, autrement dit, un hypercube. Et enfin, comme cet hypercube a des sommets alors l'ellipsoide est susceptible de la rencontrer sur un de ces sommets, là où une ou plusieurs coordonnées sont nulles. En guise de comparaison avec la régression de Ridge :

$$\underset{\beta \in \mathbb{R}^p}{\text{Argmin}}\{\tfrac{1}{n}\|Y - X\beta\|_2^2\} \text{ tel que } \|\beta\|_2^2 \leq t.$$

La solution de ce problème se situe, comme précédemment, sur l'intersection d'une courbe de niveau de l'erreur $\frac{1}{n}\|Y - X\beta X\|_2$ avec la boule l_2 de rayon t, c'est-à-dire, la boule Euclidienne "ronde" $\|\beta\|_2^2 \leq t$. Cette fois-ci, il n'y a ici aucune raison que cette intersection se fasse à un endroit où une ou plusieurs coordonnées s'annulent.

Contrairement aux estimateurs de MCO, on a une grande difficulté pour l'analyse des propriétés des estimateurs LASSO, du fait que ces derniers ne sont pas explicite. Néanmoins, nous pouvons tout de même utiliser un outil capable de les caractériser et en donner une meilleure prédiction. Cet outil est connu sous le nom de l'inégalité Oracle (Oracle Inequality).

2.2 Inégalité Oracle (ou majoration de l'erreur estimée)

Dans cette section, nous allons donner une série de lemmes et corollaire dans l'optique d'établir l'inégalité Oracle, qui permet de donner une meilleure approximation de $X\beta^0$. Autrement dit, d'en donner une meilleure prédiction.

2.2.1 Inégalité fondamentale

Étant donnée que $\widehat{\beta}$ minimise la fonction $\mathcal{F}_\lambda$, alors par définition :

$$\forall \beta \in \mathbb{R}^p, \text{ on a : } \mathcal{F}_\lambda(\widehat{\beta}) \leq \mathcal{F}_\lambda(\beta).$$

En particulier, pour $\beta = \beta^0$ on a : $\mathcal{F}_\lambda(\widehat{\beta}) \leq \mathcal{F}_\lambda(\beta^0)$. Cela conduit au lemme suivant qu'on nomme l'inégalité fondamentale (Basic Inequality).

Lemme 2.3. *(Inégalité fondamentale)*
Soit $\widehat{\beta}$ une solution du Lasso (problème (2.1)), alors la deuxième condition d'optimalité du LASSO est donné par l'inégalité suivante :

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}\|_1 \leq 2\frac{\varepsilon^T X(\widehat{\beta} - \beta^0)}{n} + \lambda\|\beta^0\|_1. \tag{2.8}$$

[4]

Démonstration. Par définition de $\widehat{\beta}$, on a l'inégalité suivante :

$$\begin{aligned}
&\frac{\|Y - X\widehat{\beta}\|_2^2}{n} + \lambda\|\hat{\beta}\|_1 \leq \frac{\|Y - X\beta^0\|_2^2}{n} + \lambda\|\beta^0\|_1 \\
\Leftrightarrow &\frac{\|Y - X\widehat{\beta}\|_2^2 - \|Y - X\beta^0\|_2^2}{n} + \lambda\|\widehat{\beta}\|_1 \leq \lambda\|\beta^0\|_1.
\end{aligned}$$

Or on a supposé que $Y = X\beta^0 + \varepsilon$ donc :

$$\begin{aligned}
\|Y - X\widehat{\beta}\|_2^2 - \|Y - X\beta^0\|_2^2 =& \|X\beta^0 + \varepsilon - X\widehat{\beta}\|_2^2 - \|X\beta^0 + \varepsilon - X\beta^0\|_2^2 \\
=& \|X(\beta^0 - \widehat{\beta}) + \varepsilon\|_2^2 - \|\varepsilon\|_2^2 \\
=& \|X(\beta^0 - \widehat{\beta})\|_2^2 + 2\langle X(\beta^0 - \widehat{\beta}), \varepsilon\rangle + \|\varepsilon\|_2^2 - \|\varepsilon\|_2^2 \\
=& \|X(\widehat{\beta} - \beta^0)\|_2^2 - 2\varepsilon^T X(\widehat{\beta} - \beta^0).
\end{aligned}$$

En remplaçant dans l'inégalité précédente, on obtient :

$$\begin{aligned}
&\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2 - 2\varepsilon^T X(\widehat{\beta} - \beta^0)}{n} + \lambda\|\widehat{\beta}\|_1 \leq \lambda\|\beta^0\|_1 \\
\Leftrightarrow &\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}\|_1 \leq 2\frac{\varepsilon^T X(\widehat{\beta} - \beta^0)}{n} + \lambda\|\beta^0\|_1.
\end{aligned}$$

□

Cette caractérisation nous sera utile pour étudier les propriétés des estimations et d'erreur de prédiction du Lasso.

Comme ε est aléatoire alors le terme $2\frac{\varepsilon^T X(\widehat{\beta}-\beta^0)}{n}$ l'est également. On appellera ce terme "processus empirique" dans le reste du problème. C'est le terme où l'erreur joue un rôle. Par ailleurs, en utilisant la norme ℓ_1, on peut facilement majorer le processus empirique, et cela donne :

$$2|\varepsilon^T X(\widehat{\beta}-\beta^0)| \leq \|2\varepsilon^T X^{(j)}\|_\infty \|\widehat{\beta}-\beta^0\|_1.$$

L'idée de la pénalité serait de négliger le terme aléatoire "le processus empirique", et pour se faire, introduisons l'événement $\mathcal{S}$ défini par :

$$\mathcal{S} := \{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty \leq \lambda_0\},$$

où λ_0 est choisi arbitrairement de sorte que nous ayons $\lambda \geq \lambda_0$, et enfin qu'on puisse nous débarrasser de la partie aléatoire de (2.8), c'est-à-dire, du processus empirique. Pour une valeur appropriée de λ_0, nous allons montrer que l'événement $\mathcal{S} = \{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty \leq \lambda_0\}$ a une grande probabilité. Pour cela, on définit la matrice de Gram normalisée comme suit $\widehat{\Sigma} := \frac{X^T X}{n}$ dont les termes diagonaux sont :

$$\widehat{\sigma}_j^2 := \widehat{\Sigma}_{jj} \ , \ j \in \{1, ..., p\},$$

et on définit la variance de ε comme étant la matrice de la variance-covariance i.e

$$\mathbf{Var}(\varepsilon) = (\mathbf{Cov}(\varepsilon_\mathbf{i}, \varepsilon_\mathbf{j}))_{1 \leq i,j \leq n},$$

en l'occurrence,

$$\mathbf{Var}(\varepsilon) = \sigma^2 I_n.$$

Cela conduit au lemme suivant.

Lemme 2.4. *On suppose que pour $j = 1, ..., p$, $\widehat{\sigma}_j^2 = \widehat{\Sigma}_{jj} = 1$, et pour tout $t \in \mathbb{R}_+^*$, on pose $\lambda_0 = 2\sigma\sqrt{\frac{t^2+2\log(p)}{n}}$. Alors pour tout $t > 0$, $\mathbb{P}(\{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty \leq \lambda_0\}) \geq 1 - 2\exp(-\frac{t^2}{2})$.*

Pour la preuve de ce lemme, nous introduisons une proposition portant sur l'inégalité de la concentration sous-Gaussienne que l'on admettra.

Proposition 2.3. ***Inégalité de la concentration sous Gaussienne (admise) :*** *Soit X une variable aléatoire. Si $X \sim \mathcal{N}(\mu, \tau^2)$ alors pour tout $x > 0$,*

$$\mathbb{P}(|X - \mu| > x) \leq 2\exp(-\frac{x^2}{2\tau^2}).$$

Démonstration. (Lemme 2.5)
En posant $V_j := \frac{\varepsilon^T X^{(j)}}{\sqrt{n\sigma^2}}$, montrons que la variable V_j suit une loi normale centrée réduite i.e $V_j \sim \mathcal{N}(0,1)$.
En effet, on sait que le vecteur $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Dès lors la variable $V_j = \frac{\varepsilon^T X^{(j)}}{\sqrt{n\sigma^2}}$ est gaussienne comme combinaison linéaire de variables Gaussiennes.
Calculons son espérance et sa variance :

$$\begin{aligned}
\mathbb{E}(V_j) &= \frac{1}{\sqrt{n\sigma^2}}\mathbb{E}(\varepsilon^T)X^{(j)} = 0.\\
\mathbf{Var}(V_j) &= \frac{1}{n\sigma^2}\mathbf{Var}(\varepsilon^T X^{(j)}) = \frac{1}{n\sigma^2}X^{(j)^T}\mathbf{Var}(\varepsilon^T)X^{(j)}\\
&= \frac{1}{n\sigma^2}X^{(j)^T}\sigma^2 I_n X^{(j)} = \frac{n\widehat{\sigma}_j^2}{n} = 1.
\end{aligned}$$

Par application de la proposition (2.3) aux variables aléatoires $V_j \sim \mathcal{N}(0,1)$ on a, pour $x > 0$:

$$\mathbb{P}(|V_j| > x) = \mathbb{P}(|V_j - \mathbb{E}(V_j)| > x) \leq 2\exp(-\frac{x^2}{2}), \forall j = 1, ..., p.$$

En utilisant la σ-finie sous additivité, on obtient la majoration suivante :

$$\mathbb{P}(\max_{1\leq j\leq p}|V_j| > x) = \mathbb{P}(\bigcup_{j=1}^{p}\{|V_j| > x\}) \leq \sum_{j=1}^{p}\mathbb{P}(\{|V_j| > x\}) \leq 2p \times \exp(-\frac{x^2}{2}).$$

En particulier, pour $x = \sqrt{t^2 + 2\log(p)}$ strictement positif, on obtient alors :

$$\begin{aligned}
&\mathbb{P}(\{\|\frac{\varepsilon^T X^{(j)}}{\sqrt{n\sigma^2}}\|_\infty > \sqrt{t^2 + 2\log(p)}) \leq 2p \times \exp(-\frac{t^2 + 2\log(p)}{2})\\
\Leftrightarrow &\mathbb{P}(\{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty > 2\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}) \leq 2\exp(-\frac{t^2}{2})\\
\Leftrightarrow &\mathbb{P}(\{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty \leq 2\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}}) \geq 1 - 2\exp(-\frac{t^2}{2})\\
\Leftrightarrow &\mathbb{P}(\{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty \leq \lambda_0\}) \geq 1 - 2\exp(-\frac{t^2}{2}).
\end{aligned}$$

□

2.2.2 Inégalité de la consistance du Lasso

Pour établir l'inégalité de la consistance du LASSO (Consistence of the Lasso), nous allons appliquer les deux lemmes qui précédent (lemmes 2.4 et 2.5). Ce résultat est analogue à celui de Greenshtein (2006).

Corollaire 2.1. *(Consistency of the Lasso)*
On suppose que pour tout $j = 1, ..., p$ *on a* $\widehat{\sigma}_j^2 = 1$. *Pour* $t > 0$, *soit le paramètre de pénalisation* $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2log(p)}{n}}$ *où* $\widehat{\sigma}$ *est un estimateur de* σ.
Alors, avec une probabilité d'au moins $1-\alpha$ *où* $\alpha := 2\exp(\frac{-t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma)$, *on a :*

$$\frac{2\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq 3\lambda\|\beta^0\|_1.$$

Démonstration. Le lemme 2.1 nous informe que pour tout $\lambda > 0$,

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\hat{\beta}\|_1 \leq 2\frac{\varepsilon^T X(\widehat{\beta} - \beta^0)}{n} + \lambda\|\beta^0\|_1.$$

Pour tout λ positive, de sorte que $\lambda \geq 2\lambda_0$ alors par utilisation du lemme 2.2, on a avec une probabilité d'au moins $1 - 2\exp(-\frac{t^2}{2})$:

$$\begin{aligned}
\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\hat{\beta}\|_1 &\leq \|2\tfrac{\varepsilon^T X(\widehat{\beta}-\beta^0)}{n}\|_1 + \lambda\|\beta^0\|_1 \\
&\leq (\max_{1\leq j\leq p}|\tfrac{2\varepsilon^T X^{(j)}}{n}|)\|\widehat{\beta} - \beta^0\|_1 + \lambda\|\beta^0\|_1 \\
&\leq \lambda_0\|\widehat{\beta} - \beta^0\|_1 + \lambda\|\beta^0\|_1 \\
&\leq \lambda_0\|\widehat{\beta}\|_1 + \lambda_0\|\beta^0\|_1 + \lambda\|\beta^0\|_1 \\
&\leq \tfrac{\lambda}{2}\|\widehat{\beta}\|_1 + \tfrac{\lambda}{2}\|\beta^0\|_1 + \lambda\|\beta^0\|_1.
\end{aligned}$$

En multipliant par 2 dans les deux membres, il s'en suit que :

$$2\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + 2\lambda\|\widehat{\beta}\|_1 \leq 2\|\widehat{\beta}\|_1 + \lambda\|\beta^0\|_1 + 2\lambda\|\beta^0\|_1,$$

c'est à dire,

$$\begin{aligned}
2\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} &\leq 3\lambda\|\beta^0\|_1 - \lambda\|\widehat{\beta}\|_1 \\
&\leq 3\lambda\|\beta^0\|_1.
\end{aligned}$$

Finalement, on obtient pour tout λ positive, l'inégalité souhaitée avec au moins une probabilité de $1 - 2\exp(-\frac{t^2}{2})$,

$$2\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq 3\lambda\|\beta^0\|_1.$$

Après avoir établi l'inégalité de la consistance, nous allons maintenant prouver qu'elle satisfait la probabilité large d'au moins $1-\alpha$.
Pour se faire, on pose :

$$Y = \frac{2}{3n}\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{\|\beta^0\|_1}.$$

Alors

$$\begin{aligned}\mathbb{P}(Y \geq \lambda) &= \mathbb{P}(Y \leq \lambda|\lambda > 2\lambda_0).\mathbb{P}(\lambda > 2\lambda_0) + \mathbb{P}(Y \leq \lambda|\lambda \leq 2\lambda_0).\mathbb{P}(\lambda \leq 2\lambda_0)\\ &\leq \mathbb{P}(Y \leq \lambda|\lambda > 2\lambda_0) + \mathbb{P}(\lambda \leq 2\lambda_0)\\ &\leq 2\exp(-\tfrac{t^2}{2}) + \mathbb{P}(\lambda \leq 2\lambda_0)\\ &= 2\exp(-\tfrac{t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma).\end{aligned}$$

Par consèquent,

$$\mathbb{P}(Y \leq \lambda) = 1 - \mathbb{P}(Y \geq \lambda) \geq 1 - 2\exp(-\frac{t^2}{2}) - \mathbb{P}(\widehat{\sigma} \leq \sigma) = 1-\alpha,$$

avec $\alpha := 2\exp(-\frac{t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma)$. □

Ainsi, nous obtenons la probabilité voulue.

Asymptotiquement, pour un estimateur $\widehat{\sigma}$ bien choisi, pas trop petit mais pas trop grand, par exemple $\widehat{\sigma^2} = \frac{Y^TY}{n}$ qui converge vers σ^2 et satisfait $\widehat{\sigma} \leq \sigma \leq const$ où $const \in \mathbb{R}$, alors nous pouvons conclure que l'erreur estimée est de l'ordre de $\frac{\log(p)}{n}\sigma^2$.

2.2.3 La condition de compatibilité

L'estimateur Lasso $\widehat{\beta}$ peut tout à fait être creux : possède des composantes nulles. L'idée serait d'éliminer ces composantes, et d'en garder celles qui ne le sont pas. Pour cela, nous introduisons dans cette section une notion fondamentale : la sparsité.

La sparsité :[2][4] *Soit $S_0 \subseteq \{1,...,p\}$ l'ensemble des variables actives "Active Set" défini comme suit :*

$$S_0 = \{j; \beta_j^0 \neq 0, j = 1,...,p\},$$

et on fait l'hypothèse qu'il existe un entier s_0 qu'on appelle l'indice de sparsité (sparsity index) tel que $Card(S_0) = s_0 \leq p$.
Cette hypothèse traduit un contrôle sur la sparsité du modèle. Elle témoigne du fait qu'il y a peu de variables explicatives pertinentes pour l'étude.
L'ensemble S_0 est le support du vecteur des paramètres inconnus β^0 du modèle. On définit également pour tout vecteur β appartenant à $\mathbb{R}^p$ le vecteur suivant :

$$\beta_{S_0} := (\beta_{j,S_0})_{1\leq j\leq p} \text{ où } \beta_{j,S_0} = \beta_j 1_{\{j\in S_0\}}.$$

Le vecteur β_{S_0} est le vecteur des paramètres β restreint au sous-ensemble S_0. La notation $\beta_j 1_{\{j\in S_0\}}$ signifie que si j n'appartient pas à l'ensemble S_0 alors la j-ième composante du vecteur β_{S_0} est nul. On définit de même pour le complementaire S_0^c de l'ensemble S_0.

$$\beta_{S_0^c} := (\beta_{j,S_0^c})_{1\leq j\leq p} \text{ où } \beta_{j,S_0^c} = \beta_j 1_{\{j\notin S_0\}}.$$

Par consèquent, on peut décomposer pour tout β appartenant à $\mathbb{R}^p$ (en fonctions des vecteurs β_{S_0} et $\beta_{S_0^c}$ définis aux sous-ensembles S_0 et S_0^c), comme suit :

$$\beta = \beta_{S_0} + \beta_{S_0^c}.$$

En d'autres termes, l'ensemble S_0 permet alors de faire une sélection des variables pour le coefficient $\widehat{\beta}$ pour chaque valeur du paramètre de régularisation λ. A la connaissance de S_0, il sera possible d'estimer la vrai valeur de l'estimateur, $\sqrt{n}(\widehat{\beta}_{S_0^c} - \beta_{S^0}) \longrightarrow \mathcal{N}(0, \Sigma_{S_0})$, à la vitesse de convergence raisonnable $\sqrt{n}$. Et enfin, de donner une bonne approximation de $X\beta^0$, on parlera alors de prédiction.

Lemme 2.5. *On suppose que l'inégalité* $\|\frac{\varepsilon^T X}{n}\|_\infty \leq \lambda_0$ *est vérifiée, alors avec une probabilité d'au moins* $1-2\exp(-\frac{t^2}{2})$ *: on obtient pour* $\lambda \geq 2\lambda_0$*, l'inégalité suivante :*

$$\frac{2\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1, \tag{2.9}$$

et on a également,

$$\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1 \leq \sqrt{s_0}\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_2.$$

Démonstration. On suppose que l'inégalité $\|\frac{\varepsilon^T X}{n}\|_\infty \leq \lambda_0$ est vérifiée, alors par application du lemme 2.4 (Basic Inequality), et que pour tout $\lambda \geq 2\lambda_0$, nous avons :

$$2\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + 2\lambda\|\widehat{\beta}\|_1 \leq \lambda\|\widehat{\beta}-\beta^0\|_1 + 2\lambda\|\beta^0\|_1.$$

Par ailleurs, $\beta = \beta_{S_0} + \beta_{S_0^c}$, donc en particulier, $\|\widehat{\beta}\|_1 = \|\widehat{\beta_{S_0}}\|_1 + \|\widehat{\beta_{S_0^c}}\|_1$.
D'une part, comme $\|\widehat{\beta_{S_0^c}}\|_1 \leq \|\beta^0_{S_0}\|_1$ alors par application de l'inégalité triangulaire, on a que :

$$\begin{aligned}
\|\beta^0_{S_0}\|_1 - \|\widehat{\beta_{S_0^c}}\|_1 \leq \|\widehat{\beta_{S_0^c}} - \beta^0_{S_0}\|_1 &\iff \|\beta^0_{S_0}\|_1 - \|\widehat{\beta_{S_0^c}} - \beta^0_{S_0}\|_1 \leq \|\widehat{\beta_{S_0^c}}\|_1 \\
&\iff \|\beta^0_{S_0}\|_1 + \|\widehat{\beta_{S_0^c}}\|_1 - \|\widehat{\beta_{S_0^c}} - \beta^0_{S_0}\|_1 \leq \|\widehat{\beta_{S_0^c}}\|_1 + \|\widehat{\beta_{S_0}}\|_1 \\
&\iff \|\beta^0_{S_0}\|_1 + \|\widehat{\beta_{S_0^c}}\|_1 - \|\widehat{\beta_{S_0}} - \beta^0_{S_0}\|_1 \leq \|\widehat{\beta}\|_1.
\end{aligned}$$

Dès lors, on en déduit que le terme de gauche de (2.9) est minoré par :

$$2\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n}+2\lambda\|\beta^0_{S_0}\|_1+2\lambda\|\widehat{\beta_{S_0^c}}\|_1-2\lambda\|\widehat{\beta_{S_0}}-\beta^0_{S_0}\|_1\leq 2\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n}+2\lambda\|\widehat{\beta}\|_1.$$

Et d'autre part, en remarquant que $\beta^0_{S_0^c}=0$ car β^0 est le vrai estimateur, alors on peut réécrire le terme $\|\widehat{\beta}-\beta_0\|_1$ en fonction de l'ensemble S_0 et S_0^c.

$$\begin{aligned}\|\widehat{\beta}-\beta_0\|_1=\|\widehat{\beta_{S_0}}+\widehat{\beta_{S_0^c}}-(\beta^0_{S_0}+\beta^0_{S_0^c})\|_1 &= \|(\widehat{\beta_{S_0}}-\beta^0_{S_0})+(\widehat{\beta_{S_0^c}}-\beta^0_{S_0^c})\|_1\\ &= \|\widehat{\beta_{S_0}}-\beta^0_{S_0}\|_1+\|\widehat{\beta_{S_0^c}}-\beta^0_{S_0^c}\|_1\\ &= \|\widehat{\beta_{S_0}}-\beta^0_{S_0}\|_1+\|\widehat{\beta_{S_0^c}}\|_1.\end{aligned}$$

En somme, l'inégalité (2.9) devient alors :

$$2\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n}+2\lambda\|\beta^0_{S_0}\|_1+2\lambda\|\widehat{\beta_{S_0^c}}\|_1-2\lambda\|\widehat{\beta_{S_0}}-\beta^0_{S_0}\|_1\leq\lambda\|\widehat{\beta_{S_0}}-\beta^0_{S_0}\|_1+\lambda\|\widehat{\beta_{S_0^c}}\|_1+2\lambda\|\beta^0_{S_0}\|_1.$$

D'où

$$\frac{2\|X(\widehat{\beta}-\beta^0)\|_2^2}{n}+\lambda\|\widehat{\beta}_{S_0^c}\|_1\leq 3\lambda\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1.$$

□

L'inégalité ci-dessus fait intervenir deux normes différentes : la norme ℓ_1 et la norme ℓ_2. Pour nous débarrasser de la norme ℓ_1, nous allons utiliser l'inégalité de Cauchy-Schwarz. Cela fera l'objet de la preuve de la deuxième égalité du lemme 2.6 que nous présentons ci-dessous :

$$\begin{aligned}\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1 &= \textstyle\sum_{j=1}^p|\widehat{\beta}_{j,S_0}-\beta^0_{j,S_0}|\\ &= \textstyle\sum_{j\in S_0}|\widehat{\beta}_j1_{\{j\in S_0\}}-\beta^0_j1_{\{j\in S_0\}}|\\ &= \textstyle\sum_{j\in S_0}|(\widehat{\beta}_j-\beta^0_j)1_{\{j\in S_0\}}|\\ &\leq (\textstyle\sum_{j\in S_0}|\widehat{\beta}_j-\beta^0_j|^2)^{\frac{1}{2}}(\sum_{j\in S_0}1_{\{j\in S_0\}})^{\frac{1}{2}}\\ &\leq \|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_2\sqrt{Card(S_0)}\\ &\leq \sqrt{s_0}\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_2.\end{aligned}$$

Hypothèse sur la matrice de Gram :[3] De nombreux résultats théoriques ont été établis ces dernières années dans la littérature statistique sur le Lasso. Ces résultats ont été obtenus au prix d'hypothèses plus ou moins contraignantes, notamment sur la matrice de Gram en nous restreignant sur l'ensemble S_0.

Compatibility condition. *La condition de compatibilité (qu'on l'admet) de l'ensemble S_0 est satisfaite s'il existe $\phi_0 > 0$ (on appellera ϕ_0^2 la "compatibility constant") tel que pour tout $\beta \in \mathcal{R}^p$ avec $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$ on a :*

$$\|\beta_{S_0}\|_1^2 \leq (\beta^t \widehat{\Sigma} \beta) \frac{s_0}{\phi_0^2}. \tag{2.10}$$

[4]

On remarque que la constante 3 dans cette définition est choisie arbitrairement. Elle peut être remplacer à tout moment en prenant une constante supérieur à 1, et en réajustant certains coefficients (en particulier λ).
L'idée de la condition (2.10) repose sur l'inégalité suivante : pour tout $\beta \in \mathbb{R}_*^p$, $\alpha_{min}\|\beta\|_2^2 \leq \beta^T \widehat{\Sigma} \beta \leq \alpha_{max}\|\beta\|_2^2$ où α_{min} et α_{max} désignent respectivement la plus petite et la plus grande valeur propre de la matrice de Gram $\widehat{\Sigma}$, si on majore le terme de gauche de (2.10), $\|\beta_{S_0^c}\|_1^2$ par $s_0\|\beta_{S_0}\|_2^2$. Néanmoins, l'inégalité $\|\beta_{S_0^c}\|_1 \leq 3\|\beta_{S_0}\|_1$ met quelques restrictions sur l'ensemble S_0. Cela conduit à imposer la stricte positivité des valeurs propres de la matrice de Gram $\widehat{\Sigma}$. Si on connaît l'ensemble S_0, on pourra déterminer son cardinal s_0, l'indice de sparsité, et ainsi définir la condition de compatibilité.[3]

Application à $\beta = \widehat{\beta} - \beta^0$.
On va faire usage de cette condition de compatibilité pour le vecteur $\widehat{\beta} - \beta^0$. Le lemme 2.3 nous permet d'écrire également que pour tout λ positif, satisfaisant la condition $\lambda \geq 2\lambda_0$, on a :

$$\lambda\|\widehat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1, \text{ c'est à dire, } \|\widehat{\beta}_{S_0^c}\|_1 \leq 3\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1.$$

En particulier, en prenant $\beta = \widehat{\beta} - \beta^0$, alors par application de la "compatibility condition,

$$\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1^2 \leq (\widehat{\beta} - \beta^0)\widehat{\Sigma}(\widehat{\beta} - \beta^0)\frac{s_0}{\phi_0^2},$$

i.e

$$\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1^2 \leq \frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n}\frac{s_0}{\phi_0^2}.$$

Par conséquent,

$$\|\widehat{\beta}_{S_0} - \beta^0_{S_0}\|_1 \leq \frac{\|X(\widehat{\beta} - \beta^0)\|_2}{\sqrt{n}}\frac{\sqrt{s_0}}{\phi_0}.$$

2.2.4 Contrôle de l'inégalité Oracle

Théorème 2.1. *Supposons la condition de compatibilité vraie pour l'ensemble des variables actives S_0. Alors sur $\{\|2\frac{\varepsilon^T X^{(j)}}{n}\|_\infty \leq \lambda_0\}$, c'est-à-dire, avec au moins une probabilité de $1-\exp(-\frac{t^2}{2})$ pour tout $t>0$, on a pour tout λ positive vérifiant $\lambda \geq 2\lambda_0$, l'inégalité suivante :*

$$\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}-\beta^0\|_1 \leq 4\lambda^2\frac{s_0}{\phi_0^2}.$$

Démonstration. En partant toujours du lemme 2.3, nous allons établir l'inégalité demandée en exploitant la "compatibility condition" ainsi que l'inégalité suivante : Pour tout $u, v \in \mathbb{R}_+$, $(u-2v)^2 \geq 0 \Longleftrightarrow 4uv \leq u^2+4v^2$. Nous avons avec au moins une probabilité de $1-\exp(-\frac{t^2}{2})$ que :

$$2\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}_{S_0^c}\|_1 \leq 3\lambda\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1,$$

ce qui est équivalent à,

$$2\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}-\beta_0\| \leq \lambda\|\widehat{\beta}-\beta_0\|_1 + 3\lambda\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1 - \lambda\|\widehat{\beta}_{S_0^c}\|_1.$$

Donc,

$$\begin{aligned} 2\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}-\beta_0\| &\leq 3\lambda\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1 + \lambda\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1 + \lambda\|\widehat{\beta}_{S_0^c}\|_1 - \lambda\|\widehat{\beta}_{S_0^c}\|_1 \\ &\leq 4\lambda\|\widehat{\beta}_{S_0}-\beta^0_{S_0}\|_1 \\ &\leq 4\lambda\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2}{\sqrt{n}}\tfrac{\sqrt{s_0}}{\phi_0}. \end{aligned}$$

Dès lors pour $u = \frac{\|X(\widehat{\beta}-\beta^0)\|_2}{\sqrt{n}}$ et $v = \frac{\lambda\sqrt{s_0}}{\phi_0}$,
il s'en suit que :

$$\begin{aligned} 2\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}-\beta_0\| &\leq 4\lambda\tfrac{\|X(\widehat{\beta}-\beta^0)\|_2}{\sqrt{n}}\tfrac{\sqrt{s_0}}{\phi_0} \\ &\leq \tfrac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \tfrac{4\lambda^2 s_0}{\phi_0^2}. \end{aligned}$$

D'où l'inégalité souhaitée :

$$\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta}-\beta^0\|_1 \leq 4\lambda^2\frac{s_0}{\phi_0^2}.$$

□

Ainsi avec une probabilité d'au moins $1-\exp(-\frac{t^2}{2})$, on trouve le résultat annoncé.

Corollaire 2.2. *On suppose que pour tout $j = 1, ..., p$ on a $\widehat{\sigma}_j^2 = 1$ et que la condition de compatibilité de l'ensemble S_0 est satisfaite. Pour $t > 0$, soit le paramètre de pénalisation $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log(p)}{n}}$ où $\widehat{\sigma}$ est un estimateur de σ.*
Alors, avec une probabilité d'au moins $1-\alpha$ où $\alpha := 2\exp(\frac{-t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma)$, on a :

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta} - \beta^0\|_1 \leq 4\lambda^2 \frac{s_0}{\phi_0^2}.$$

Démonstration. En partant du théorème 2.1, on sait qu'avec une probabilité d'au moins $1 - \exp(-\frac{t^2}{2})$, on a $\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta} - \beta^0\|_1 \leq 4\lambda^2 \frac{s_0}{\phi_0^2}$. Il est donc clair que :

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq \frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} + \lambda\|\widehat{\beta} - \beta^0\|_1 \leq 4\lambda^2 \frac{s_0}{\phi_0^2},$$

c'est-à-dire,

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq 4\lambda^2 \frac{s_0}{\phi_0^2}.$$

Par un raisonnement similaire au corollaire 2.1, on a qu'en posant :

$$Y = \sqrt{\frac{ns_0\phi^2}{2}} \frac{\|X(\widehat{\beta} - \beta^0)\|_2}{n},$$

$$\begin{aligned}
\mathbb{P}(Y \geq \lambda) &= \mathbb{P}(Y \leq \lambda | \lambda > 2\lambda_0).\mathbb{P}(\lambda > 2\lambda_0) + \mathbb{P}(Y \leq \lambda | \lambda \leq 2\lambda_0).\mathbb{P}(\lambda \leq 2\lambda_0) \\
&\leq \mathbb{P}(Y \leq \lambda | \lambda > 2\lambda_0) + \mathbb{P}(\lambda \leq 2\lambda_0) \\
&\leq 2\exp(-\tfrac{t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma).
\end{aligned}$$

Par consèquent,

$$\mathbb{P}(Y \leq \lambda) = 1 - \mathbb{P}(Y \geq \lambda) \geq 1 - 2\exp(-\frac{t^2}{2}) - \mathbb{P}(\widehat{\sigma} \leq \sigma) = 1 - \alpha,$$

avec $\alpha := 2\exp(-\frac{t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma)$. □

Corollaire 2.3. ***(Inégalité Oracle)*** *On suppose que pour tout $j = 1, ..., p$ on a $\widehat{\sigma}_j^2 = 1$ et que la condition de compatibilité de l'ensemble S_0 est satisfaite. Pour $t > 0$, soit le paramètre de pénalisation $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log(p)}{n}}$ où $\widehat{\sigma}$ est un estimateur*

de σ. Alors avec une probabilité d'au moins $1-\alpha$ où $\alpha := 2\exp(\frac{-t^2}{2}) + \mathbb{P}(\widehat{\sigma} \leq \sigma)$, il existe une constante positive const $\in \mathbb{R}_+$ tel que :

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq const. \frac{\sigma^2 \log(p)}{n} s_0.$$

[4]

Démonstration. En partant du corollaire 2.2, nous avons :

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq 4\lambda^2 \frac{s_0}{\phi_0^2}.$$

Par ailleurs, $\lambda = 4\widehat{\sigma}\sqrt{\frac{t^2+2\log(p)}{n}}$

$$\Rightarrow \frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq 4\lambda^2 \frac{s_0}{\phi_0^2} = 64 \frac{\widehat{\sigma}^2}{\phi_0^2} \frac{t^2 + 2\log(p)}{n} s_0 \leq 64 \frac{\sigma^2}{\phi_0^2} \frac{t^2 + 2\log(p)}{n} s_0.$$

En particulier, pour $t = \sqrt{2\log(p)}$, on a que :

$$\frac{t^2 + 2\log(p)}{n} s_0 = \frac{\sqrt{2\log(p)}^2 + 2\log(p)}{n} s_0 = \frac{4\log(p)}{n} s_0.$$

Par consèquent,

$$\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq 64 \frac{\sigma^2}{\phi_0^2} \frac{4\log(p)}{n} s_0.$$

Ainsi en posant,

$$const := \frac{256}{\phi_0^2} \in \mathbb{R}.$$

Alors on retrouve l'inégalité attendue :

$$\boxed{\frac{\|X(\widehat{\beta} - \beta^0)\|_2^2}{n} \leq \text{const.} \frac{\sigma^2 \log(p)}{n} s_0.}$$

□

L'inégalité Oracle permet de comparer l'écart entre des données expérimentales ($X\widehat{\beta}$ venant du paramètre estimé) et des données calculées à l'aide d'un modèle quantitatif ($X\beta^0$ venant du paramètre théorique du modèle) pour le Lasso.

Contrairement au cadre standard MCO dont l'erreur de prédiction est de l'ordre de $\mathcal{O}_{\mathbb{P}}(\frac{p}{n}\sigma^2)$, le contrôle de l'erreur estimée pour le Lasso (ou la prédiction) est de

$\mathcal{O}_{\mathbb{P}}(\frac{\sigma^2 \log(p)}{n} s_0)$ avec une large probabilité, sous réserve que l'ensemble des variables actives S_0 soit identifiable. L'inégalité Oracle a la particularité d'améliorer la précision de l'erreur estimée du Lasso. La constante de l'inégalité *cont* est explicite, et dépend de la constante de la condition de compatibilité ϕ_0.

2.3 Conclusion

Nous venons de voir au cours de cette deuxième partie que dans le cadre de la grande dimension ($p > n$), les estimateurs Lasso n'ont pas de forme explicite contrairement à ceux des MCO. Néanmoins, à la connaissance d'un nouvel ensemble S_0 appelé l'ensemble des variables active "Active Set" avec l'hypothèse de sparsité, il a été tout à fait possible de construire un estimateur Lasso qui possède quelques propriétés, et vérifient différentes inégalités notamment celle d'Orale (Oracle Inequality). Ces inégalités ont la particularité de dépendre du nombre de composante non-nulles sur les paramètres estimées. Cela a permis de contrôler, avec une large probabilité, l'erreur d'estimaion.

Par ailleurs, différents algorithmes existent pour approximer l'ensemble des solutions du problème Lasso pour λ variant dans $[0, +\infty[$. En particulier, l'algorithme LARS apporte une réponse à ce problème d'optimisation. Cet algorithme a l'avantage d'être très rapide et a accru l'intérêt porté à la méthode Lasso. Nous verrons dans les prochaines parties une mise en application de cet algorithme.

Partie 3

Résolution du Lasso par la méthode homotopique

3.1 Justification théorique de l'algorithme

On cherche à minimiser F_λ la fonction définie par :

$$F_\lambda(\beta) = \frac{1}{2}\|X\beta - Y\|_2^2 + \lambda\|\beta\|_1,\ \beta \in \mathbb{R}^p.$$

Les minimums de cette fonction sont atteints pour les mêmes valeurs que celles de la fonction $\mathcal{F}_{\lambda\prime}(\beta) = \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\prime\|\beta\|_1$, il suffit de prendre $\lambda = \frac{n\lambda\prime}{2}$.

On note β_λ une solution optimale de $\min_{\beta\in\mathbb{R}^p} F_\lambda$, on montrera par la suite que pour un λ^0 assez grand $\beta_{\lambda^0} = 0$.

L'idée de la méthode homotopique est de calculer toutes les solutions en partant de $\beta_{\lambda^0} = 0$ jusqu'à β_λ pour un λ donné ou jusqu'à ce qu'il n'y ait plus unicité de la solution.

Pour calculer ces solutions on va utiliser des conditions qui caractérisent la solution de F_λ :

$$\begin{aligned} &(X^t(X\beta - Y))_l = -\lambda \ \ \text{sign}(\beta_l), && \text{si } \beta_l \neq 0, \\ &|(X^t(X\beta - Y))_l| \leq \lambda, && \text{si } \beta_l = 0. \end{aligned}$$

On verra que la fonction qui à λ associe β_λ est affine par morceaux, il est donc suffisant de connaître λ et β_λ aux extrémités de chaque partie affine.

3.1.1 Sous-différentiel

On souhaite minimiser la fonction convexe $F_\lambda(\beta)$. Quand une fonction convexe $f : I \subset \mathbb{R}^n \mapsto \mathbb{R}$ est différentiable en un point $x_0 \in I$ on a l'inégalité suivante :

$$f(x) \geq f(x_0) + \mathrm{d}f(x_0)(x - x_0), \forall x \in I,$$

où $\mathrm{d}f(x_0) \in \mathbb{R}^{1\times n}$ est la différentielle de f au point x_0 on cherche alors si cette différentielle s'annule en un point $\widehat{x}$ car on pourra alors écrire :

$$f(x) \geq f(\widehat{x}), \forall x \in I,$$

et en déduire que $\widehat{x}$ est une solution optimale de notre fonction.

Mais la fonction $F_\lambda(\beta)$ n'est pas différentiable à cause de la norme 1, introduisons la notion de sous-différentiel et le théorème qui en découle.

Définition 3.1. Soit f une fonction convexe définie sur I, un ouvert de $\mathbb{R}^n$, à valeur dans $\mathbb{R}$. Un sous-gradient de $f : I \mapsto \mathbb{R}$ en un point x_0 de I est un vecteur $S \in \mathbb{R}^{1\times n}$ tel que, pour tout x appartenant à I :

$$f(x) \geq f(x_0) + S(x - x_0).$$

L'ensemble des tous les sous-gradients est appelé sous-différentiel de la fonction f en x_0, noté $\partial f(x_0)$.

Cas de la norme 1

Pour nous aider à calculer la sous différentielle de F_λ, traitons le cas de la norme 1, $\| \, . \, \|_1 : \begin{array}{ccc} \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ x & \longmapsto & \sum_{i=1}^n |x_i| \end{array}$ qui est une somme de fonctions valeur absolue.

Proposition 3.1. *Le sous-différentiel d'une somme finie de fonctions sous-différentiables est la somme des sous-différentiels de ces fonctions.*

Démonstration. ($\Leftarrow$) Soit f_1 et f_2 deux fonctions définies sur I sous-différentiables en x_0.
Si $S_1 \in \partial f_1(x_0)$ et $S_2 \in \partial f_2(x_0)$, alors :

$$\begin{cases} f_1(x) \geq f_1(x_0) + S_1(x - x_0), \ \forall x \in I \\ f_2(x) \geq f_2(x_0) + S_2(x - x_0), \ \forall x \in I \end{cases}$$

$$\Rightarrow f_1(x) + f_2(x) \geq f_1(x_0) + f_2(x_0) + S_1(x - x_0) + S_2(x - x_0), \ \forall x \in I$$
$$\Rightarrow (f_1 + f_2)(x) \geq (f_1 + f_2)(x_0) + (S_1 + S_2)(x - x_0), \ \forall x \in I$$
$$\Rightarrow S_1 + S_2 \in \partial(f_1 + f_2)(x_0).$$

($\Rightarrow$) Admis □

Cas de la norme 1 (suite)

On peut donc se restreindre à calculer le sous-différentiel des fonctions :
$f_i : \begin{array}{ccc} \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ x & \longmapsto & |x_i| \end{array}$, $\forall i \in \{1, 2, ..., n\}$.
Soit $\tilde{x} \in \mathbb{R}^n$, on cherche $S \in \mathbb{R}^{1\times n}$ tel que :

$$|x_i| \geq |\tilde{x}_i| + S(x - \tilde{x}),\ \forall x \in \mathbb{R}^n.$$

En particulier pour $x := \tilde{x} + e_j$ où e_j est le vecteur canonique et $j \in \{1, 2, ..., n\}$, $j \neq i$, on a donc :

$$\begin{aligned} & |\tilde{x}_i| \geq |\tilde{x}_i| + S(x - \tilde{x}) \\ \Rightarrow\ & 0 \geq S \times e_j \\ \Rightarrow\ & 0 \geq S_j. \end{aligned}$$

Avec le même raisonnement pour $x := \tilde{x} - e_j$, on a aussi :

$$0 \leq S_j$$

Donc $S_j = 0$, $\forall j \in \{1, 2, ..., n\}$, $j \neq i$.

On peut encore se restreindre à calculer le sous différentiel de la fonction valeur absolue : $|\,.\,| : \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}^+ \\ x & \longmapsto & |x| \end{array}$.
Cette fonction est dérivable pour tout $x \neq 0$ et sa dérivé vaut $\text{sign}(x)$. Calculons son sous-différentiel en zéro, on cherche $s \in \mathbb{R}$ tel que :

$$\begin{aligned} & |x| \geq |0| + s(x - 0), && \forall x \in \mathbb{R} \\ \Leftrightarrow\ & |x| \geq sx, && \forall x \in \mathbb{R} \\ \Rightarrow\ & \begin{cases} x \geq sx, & \forall x \in \mathbb{R}^+ \\ -x \geq sx, & \forall x \in \mathbb{R}^- \end{cases} \\ \Rightarrow\ & \begin{cases} s \leq 1, & \forall x \in \mathbb{R}^+ \\ s \geq -1, & \forall x \in \mathbb{R}^- \\ s \in \mathbb{R}, & \forall x \in \mathbb{R}^- \end{cases} \\ \Rightarrow\ & s \in [-1, 1]. \end{aligned}$$

On obtient que le sous-différentiel de $F_\lambda(\beta)$ est :

$$\boxed{\partial F_\lambda(\beta) = X^T(X\beta - Y) + \lambda\{v \in \mathbb{R}^n : v_l = \begin{cases}[-1,1], & \text{si } \beta_l = 0 \\ \text{sign}(\beta_l), & \text{sinon}\end{cases}, l \in \{1,2,...,n\}\}.}$$

3.1.2 Condition d'optimalité

Cette notion de sous-différentiel à été introduite dans le but d'obtenir une caractérisation des solutions optimales. Rappelons que F_λ est strictement convexe et donc que sa solution optimale est unique.

Théorème 3.1. *Soit I un ouvert de $\mathbb{R}^p$ et $f : I \mapsto \mathbb{R}$ une fonction convexe alors $\widehat{\beta} \in I$ minimise f si et seulement si $0 \in \partial f(\widehat{\beta})$.*

Démonstration.

$$\begin{aligned} & \widehat{\beta} \text{ minimise } f \\ \Leftrightarrow\ & f(\beta) \geq f(\widehat{\beta}), \forall \beta \in I \\ \Leftrightarrow\ & f(\beta) \geq f(\widehat{\beta}) + 0(\beta - \widehat{\beta}), \forall \beta \in I \\ \Leftrightarrow\ & 0 \in \partial f(\widehat{\beta}). \end{aligned}$$

□

On a donc :

$$\begin{aligned} & \beta \text{ minimise } F_\lambda(\beta) \\ \Leftrightarrow & 0 \in \partial F_\lambda(\beta) \\ \Leftrightarrow & \forall l \in \{1,2,...,n\}, 0 \in (\partial F_\lambda(\beta))_l \end{aligned}$$

$\Leftrightarrow \forall l \in \{1,2,...,n\}$,

— Si $\beta_l = 0$:
$0 \in (\partial F_\lambda(\beta))_l$
$\Leftrightarrow\ 0 \in (X^t(X\beta - Y))_l + \lambda \times [-1,1]$
$\Leftrightarrow\ (X^t(X\beta - Y))_l \in [-\lambda, \lambda]$
$\Leftrightarrow\ |(X^t(X\beta - Y))_l| \leq \lambda.$

— Sinon $\beta_l \neq 0$:
$0 \in (\partial F_\lambda(\beta))_l$
$\Leftrightarrow\ 0 \in (X^t(X\beta - Y))_l + \lambda \times \text{sign}(\beta_l)$
$\Leftrightarrow\ (X^t(X\beta - Y))_l = -\lambda \times \text{sign}(\beta_l).$

Cette condition va nous permettre de suivre l'évolution de β_λ en fonction de λ.

Condition d'optimalité

$$(X^t(X\beta - Y))_l = -\lambda \quad \text{sign}(\beta_l), \quad \text{si } \beta_l \neq 0, \tag{3.1}$$
$$|(X^t(X\beta - Y))_l| \leq \lambda, \quad \text{si } \beta_l = 0. \tag{3.2}$$

3.1.3 Algorithme de calcul de la solution optimale en fonction de λ

Dans cet algorithme on supposera que les minimums, maximums que l'on va calculer ne sont atteints que pour une seule composante, ce qui permettra de simplifier différents points de la démonstration. On peut retrouver le cas général dans cette référence ([5] Page 480, Remark 15.3.(b)).

A) Initialisation

La méthode démarre à $\beta^{(0)} := 0 \in \mathbb{R}^p$, on a :

$$\begin{aligned} & 0 \in \partial F_\lambda(\beta^{(0)}) \\ \Leftrightarrow\ & |(X^t(X \times 0 - Y))_l| \leq \lambda, \quad \forall l \in \{1, 2, ..., p\} \\ \Leftrightarrow\ & |-(X^T Y)_l| \leq \lambda, \quad \forall l \in \{1, 2, ..., p\} \\ \Leftrightarrow\ & \|X^T Y\|_\infty \leq \lambda. \end{aligned}$$

On pose alors $\lambda^0 := \|X^T Y\|_\infty$, on remarque que pour tout $\lambda \geq \lambda^0$ zéro est bien une solution optimale de F_λ.

B) Première itération

Calcul de la direction : On a supposé qu'il n'y a qu'un seul indice tel que : $l^1 := \underset{l \in \{1,2,...,p\}}{\text{Argmax}} \{|(X^T Y)_l|\}$, alors quand λ va être légèrement plus petit que λ^0 la condition d'optimalité (3.2) ne sera plus satisfaite pour la composante l^1 et donc β_{l^1} va être différent de zéro et sera définit en fonction de la condition d'optimalité (3.1) et ce jusqu'à ce que λ diminue assez pour que la condition d'optimalité (3.2) ne soit plus satisfaite pour une autre composante de β. On notera cette diminution γ^1. i.e $\gamma^1 :=$ suprémum de l'ensemble des γ entiers positifs tel que $|-(X^T Y)_l| \leq (\lambda - \gamma)$, pour tout $l \in \{1, 2, ..., p\} \backslash \{l^1\}$.

Pour tout $\lambda^1(\gamma) := \lambda^0 - \gamma,\ \gamma \in]0, \gamma^1]$, on peut quantifier la variation de β par rapport à γ en posant $\beta^{(1)}(\gamma) := \beta^{(0)} + \gamma d^1$, où d^1 est un vecteur de $\mathbb{R}^p$ dont la seule composante non nulle est la composante l^1.

Utilisons la condition d'optimalité (3.1) pour en déduire d^1 :

$$
\begin{aligned}
& (X^t(X\beta^{(1)}(\gamma) - Y)_{l^1} = -\lambda^1(\gamma) \times \text{sign}(\beta^{(1)}(\gamma)_{l^1}) \\
\Leftrightarrow\ & (X^t(X(\beta^{(0)} + \gamma d^1) - Y)_{l^1} = -(\lambda^0 - \gamma) \times \text{sign}((\beta^{(0)} + \gamma d^1)_{l^1}) \\
\Leftrightarrow\ & (X^t(X\gamma d^1 - Y)_{l^1} = -(\|X^T Y\|_\infty - \gamma) \times \text{sign}(d^1_{l^1}) \\
\Leftrightarrow\ & (X^t X\gamma d^1)_{l^1} - (X^T Y)_{l^1} = -|X^T Y|_{l^1} \times \text{sign}(d^1_{l^1}) + \gamma \times \text{sign}(d^1_{l^1}) \qquad (E)
\end{aligned}
$$

On remarque que l'on peut simplifier cette équation en supposant que :

$\text{sign}(d^1_{l^1}) = \text{sign}((X^T Y)_{l^1})$

$$
\begin{aligned}
(E) \Leftrightarrow\ & \gamma(X^t X d^1)_{l^1} = \gamma \times \text{sign}((X^T Y)_{l^1}) \\
\Leftrightarrow\ & (X^t X d^1)_{l^1} = \text{sign}((X^T Y)_{l^1}) \\
\Leftrightarrow\ & d^1_{l^1} \|X^{(l^1)}\|_2^2 = \text{sign}((X^T Y)_{l^1}) \\
\Leftrightarrow\ & \boxed{d^1_{l^1} = \frac{\text{sign}((X^T Y)_{l^1})}{\|X^{(l^1)}\|_2^2}.}
\end{aligned}
$$

Vérifions maintenant notre supposition, par l'absurde supposons que : $\text{sign}(d^1_{l^1}) = -\text{sign}((X^T Y)_{l^1})$,

$$
\begin{aligned}
(E) \Leftrightarrow\ & \gamma d^1_{l^1} \|X^{(l^1)}\|_2^2 - \lambda^0 \text{sign}((X^T Y)_{l^1}) = \lambda^0 \text{sign}((X^T Y)_{l^1}) - \gamma \text{sign}((X^T Y)_{l^1}) \\
\Leftrightarrow\ & \gamma d^1_{l^1} \|X^{(l^1)}\|_2^2 = (2\lambda^0 - \gamma)\text{sign}((X^T Y)_{l^1}) \\
\Leftrightarrow\ & d^1_{l^1} = \|X^{(l^1)}\|_2^{-2}(\frac{2\lambda^0}{\gamma} - 1)\text{sign}((X^T Y)_{l^1}).
\end{aligned}
$$

On s'intéresse maintenant aux signes des membres de cette équation.

$$
\begin{aligned}
\Rightarrow\ & \text{sign}(d^1_{l^1}) = \text{sign}(\|X^{(l^1)}\|_2^{-2}(\frac{2\lambda^0}{\gamma} - 1)\text{sign}((X^T Y)_{l^1})) \\
\Rightarrow\ & -\text{sign}((X^T Y)_{l^1}) = \text{sign}(\frac{2\lambda^0}{\gamma} - 1) \times \text{sign}((X^T Y)_{l^1}) \\
\Rightarrow\ & -1 = \text{sign}(\frac{2\lambda^0}{\gamma} - 1) \\
\Rightarrow\ & \frac{2\lambda^0}{\gamma} - 1 < 0 \quad \Rightarrow\ 2\lambda^0 < \gamma.
\end{aligned}
$$

Absurde car cela voudrais dire que la condition (3.2) est vraie pour tout $\lambda \in [0, +\infty]$ et pour tout $l \neq l^1$.

Calcul du pas : On va maintenant calculer γ^1 la diminution maximum de λ^0, i.e : le plus grand γ tel que (3.2) est toujours vraie pour tout $l \neq l^1$.

$$\begin{aligned} & |(X^T(X\beta^{(1)}(\gamma) - Y)_l| \leq \lambda^1(\gamma) \\ \Leftrightarrow \quad & |(X^T(X(\beta^{(0)} + \gamma d^1) - Y)_l| \leq (\lambda^0 - \gamma) \\ \Leftrightarrow \quad & |\gamma(X^T X d^1)_l - (X^T Y)_l| \leq \lambda^0 - \gamma. \end{aligned} \tag{3.3}$$

On va traiter les sous-cas selon le signe de $\gamma(X^T X d^1)_l - (X^T Y)_l$.

— Si : $\gamma(X^T X d^1)_l - (X^T Y)_l > 0$
(3.3) $\Leftrightarrow$ $\gamma(X^T X d^1)_l - (X^T Y)_l \leq \lambda^0 - \gamma$
$\Leftrightarrow$ $\gamma + \gamma(X^T X d^1)_l \leq \lambda^0 + (X^T Y)_l$
$\Leftrightarrow$ $\gamma(1 + (X^T X d^1)_l) \leq \lambda^0 + (X^T Y)_l$

Et selon le signe de $(1 + (X^T X d^1)_l)$.

☐ Si $(1 + (X^T X d^1)_l) > 0$
(3.3) $\Leftrightarrow$ $\gamma \leq \frac{\lambda^0 + (X^T Y)_l}{1 + (X^T X d^1)_l}$

☐ Si $(1 + (X^T X d^1)_l) < 0$
(3.3) $\Leftrightarrow$ $\gamma \geq \frac{\lambda^0 + (X^T Y)_l}{1 + (X^T X d^1)_l}$

— Si : $\gamma(X^T X d^1)_l - (X^T Y)_l < 0$
(3.3) $\Leftrightarrow$ $-\gamma(X^T X d^1)_l + (X^T Y)_l \leq \lambda^0 - \gamma$
$\Leftrightarrow$ $\gamma - \gamma(X^T X d^1)_l \leq \lambda^0 - (X^T Y)_l$
$\Leftrightarrow$ $\gamma(1 - (X^T X d^1)_l) \leq \lambda^0 - (X^T Y)_l$

Et selon le signe de $(1 - (X^T X d^1)_l)$.

☐ Si $(1 - (X^T X d^1)_l) > 0$
(3.3) $\Leftrightarrow$ $\gamma \leq \frac{\lambda^0 - (X^T Y)_l}{1 - (X^T X d^1)_l}$

☐ Si $(1 - (X^T X d^1)_l) < 0$
(3.3) $\Leftrightarrow$ $\gamma \geq \frac{\lambda^0 - (X^T Y)_l}{1 - (X^T X d^1)_l}$

Donc :

$$\begin{cases} \gamma \leq \dfrac{\lambda^0 + (X^T Y)_l}{1 + (X^T X d^1)_l}, \ \forall l \neq l^1 \\ \gamma \leq \dfrac{\lambda^0 - (X^T Y)_l}{1 - (X^T X d^1)_l}, \ \forall l \neq l^1 \\ 0 < \gamma. \end{cases}$$

On a donc que :

$$\boxed{\gamma^1 = \min_{l \neq l^1}^{+} \left\{ \frac{\lambda^0 - (X^T Y)_l}{1 - (X^T X d^1)_l}, \frac{\lambda^0 + (X^T Y)_l}{1 + (X^T X d^1)_l} \right\},}$$

où $\min^{+}$ est le min parmi les valeurs positives, qui est bien défini et différent de zéro car on sait que : pour tout $l \in \{1, 2, ..., p\} \backslash \{l^1\}, \lambda^0 + (X^T Y)_l > 0$ et $\lambda^0 - (X^T Y)_l > 0$, parce que l'on a supposé que $l^1 = \underset{l \in \{1,2,...,p\}}{\text{Argmax}} \{|(X^T Y)_l|\}$ est unique. Au moins un des dénominateur est positif donc une des deux expressions est positive pour chaque $l \neq l^1$.

On a supposé qu'il n'y ait qu'un seul indice qui minimise cette expression et le notera l^2.

Calcul de $\beta^{(1)}$, λ^1 et préparation de la prochaine itération : On a maintenant accès à $\beta^{(1)} := \beta^{(1)}(\gamma^1) = \beta^{(0)} + \gamma^1 d^1$ et $\lambda^1 := \lambda^0 - \gamma^1$, il suffit alors de stocker λ^0, λ^1, $\beta^{(0)}$ et $\beta^{(1)}$, on connaîtra donc le segment $[(\lambda^1, \beta^{(1)}), (\lambda^0, \beta^{(0)})]$ qui représente les solutions optimales de F_λ en fonction de λ sur l'intervalle $[\lambda^1, \lambda^0]$.

On sait que si on diminue encore λ alors la composante l^2 de β ne sera plus nulle, on va donc introduire $S_2 := \{l^1, l^2\}$ appelé "Active Set" qui est l'ensemble des composante sur lequel β est différent de zéro.

C) $\mathbf{k^{i\text{-ème}}}$ itération ($k \geq 2$)

On dispose de $\beta^{(k-1)}$, λ^{k-1} et de S_k. On notera $C^k := X^t(X\beta^{(k-1)} - Y)$.
Le calcul de la direction et du pas seront fournis sans les détails, la démarche générale est la même que pour la première itération et ces résultats viennent de [5], Chapitre 15.

Calcul de la direction : La condition d'optimalité (3.1) va nous permettre de trouver le système d'équations suivant :

$$X_{S_k}^T X_{S_k} d_{S_k}^k = -\text{sign}(C_{S_k}^k).$$

X_{S_k} représente la sous-matrice constituée des colonnes de X d'indice appartenant l'ensemble S_k. Pour obtenir d^k il faut que la matrice $X_{S_k}^T X_{S_k}$ soit inversible ce qui est vrai si et seulement si les colonnes de la matrice X_{S_k} sont libres. On ne pourra donc pas faire cette étape si le cardinal de S_k est strictement supérieur à n (nombre de lignes de cette matrice).

Dans le cas où $X_{S_k}^T X_{S_k}$ est inversible on a :

$$\boxed{\begin{cases} d^k_{S_k} = -(X^T_{S_k} X_{S_k})^{-1} \text{sign}(C^k_{S_k}) \\ d^k_{S_k{}^c} = 0. \end{cases}}$$

Calcul du pas : Pour le pas, on se retrouve dans une situation légèrement différente, on cherche le plus grand γ tel que pour tout $l \in \{1, ..., p\}$ la condition d'optimalité est satisfaite, soit :

$$\begin{cases} |(X^T(X(\beta^{(k-1)} + \gamma d^k) - Y)_l| \leq (\lambda^{k-1} - \gamma), & \forall l \notin S_k \\ (X^t(X(\beta^{(k-1)} + \gamma d^k) - Y))_l = -\lambda \text{sign}((\beta^{(k-1)} + \gamma d^k)_l), & \forall l \in S_k. \end{cases}$$

γ^k est alors déterminé par :

$$\boxed{\gamma^k = \min \left\{ \underset{l \notin S_k}{\overset{+}{\min}} \left\{ \frac{\lambda^{k-1} - C^k_l}{1 - (X^T X d^k)_l}, \frac{\lambda^{k-1} + C^k_l}{1 + (X^T X d^k)_l} \right\}, \underset{l \in S_k, d^k_l \neq 0}{\overset{+}{\min}} \left\{ -\beta^{(k-1)}_l / d^k_l \right\} \right\}.}$$

Calcul de $\beta^{(k)}$, λ^k et préparation de la prochaine itération : On calcule et stocke $\beta^{(k)} := \beta^{(k)}(\gamma^k) = \beta^{(k-1)} + \gamma^k d^k$ et $\lambda^k := \lambda^{k-1} - \gamma^k$.

Si le l qui atteint le minimum pour calculer γ^k n'appartient pas à S_k, cela veut dire que quand on reculera λ à la prochaine itération cette composante ne pourra plus être nulle on va donc l'ajouter à "active set" : $S_{k+1} := S_k \cup l$.

Si le l qui atteint le minimum pour calculer γ^k appartient à S_k, on aura : $\beta^{(k)}_l = (\beta^{(k-1)} + \gamma^k d^k)_l = \beta^{(k-1)}_l + (-\beta^{(k-1)}_l / d^k_l) d^k_l = 0$, on va donc l'enlever de "Active Set" : $S_{k+1} := S_k \backslash \{l\}$.

D) Arrêt de l'algorithme

Au moins trois raisons peuvent nous motiver ou nous forcer à arrêter les itérations :

— On a toutes les solutions jusqu'à $\lambda = 0$ ou jusqu'au $\lambda > 0$ que l'on désire.

— Le résidu est nul, c'est à dire : $\|X\beta - Y\|_2 = 0$.

— Le cardinal de S_k est strictement supérieur à n, la matrice $X_{S_k}^T X_{S_k}$ n'est plus inversible.

3.2 Autre application de la methode homotipique

On va voir que cette methode permet aussi sous certaines conditions de résoudre un problème du type : $X\beta = Y$ où β est l'inconnue.

Considérons le problème :

$$\operatorname{argmin}\{\|\beta\|_1 : X\beta = Y\}. \tag{3.4}$$

Proposition 3.2. *Supposons que $X\beta = Y$ a une solution. Si la solution de (3.4), notée $\tilde{\beta}$, est unique alors :*

$$\lim_{\lambda \to 0^+} \beta_\lambda = \tilde{\beta}.$$

Plus généralement, les β_λ sont bornés et tout point adhérent de (β_{λ_n}), où (λ_n) est une suite positive telle que $\lambda_n \underset{n\to+\infty}{\longrightarrow} 0^+$ est solution optimale de (3.4).

Démonstration. $\lambda > 0$, soit β_λ solution optimale de $F_\lambda(\beta) = \frac{1}{2}\|X\beta - Y\|_2^2 + \lambda\|\beta\|_1$ et soit $\tilde{\beta}$ solution optimale de (3.4) alors :

$$\frac{1}{2}\|X\beta_\lambda - Y\|_2^2 + \lambda\|\beta_\lambda\|_1 = F_\lambda(\beta_\lambda) \leq F_\lambda(\tilde{\beta}) = \frac{1}{2}\underbrace{\|X\tilde{\beta} - Y\|_2^2}_{=0} + \lambda\|\tilde{\beta}\|_1$$

$$\Rightarrow \quad \tfrac{1}{2}\|X\beta_\lambda - Y\|_2^2 + \lambda\|\beta_\lambda\|_1 \leq \lambda\|\tilde{\beta}\|_1. \tag{3.5}$$

$$\Rightarrow \quad \|\beta_\lambda\|_1 \leq \|\tilde{\beta}\|_1. \tag{3.6}$$

On déduit de (3.6) que la suite (β_{λ_n}) est bornée. Par le théorème de Bolzano-Weierstrass (β_{λ_n}) admet une sous suite convergente, i.e il existe β' un point adhérent de (β_{λ_n}).
Soit $(\beta_{\lambda'_n})$ une sous suite extraite convergente vers β', alors :

$$\|X\beta_{\lambda'_n} - Y\|_2 \underset{n\to+\infty}{\longrightarrow} \|X\beta' - Y\|_2$$

$$\text{et } \lambda'_n\|\tilde{\beta}\| \underset{n\to+\infty}{\longrightarrow} 0.$$

Avec l'inégalité (3.5) on obtient :

$$\begin{aligned}&\frac{1}{2}\|X\beta' - Y\|_2^2 \leq 0\\ \Rightarrow &\|X\beta' - Y\|_2 = 0\\ \Rightarrow &X\beta' = Y.\end{aligned}$$

Du fait que $X\beta' = Y$, par définition de $\tilde{\beta}$ on a :

$$\|\tilde{\beta}\|_1 \leq \|\beta'\|_1.$$

Avec l'inégalité (3.6) on a aussi :

$$\|\beta'\|_1 \leq \|\tilde{\beta}\|_1$$

Donc

$$\|\tilde{\beta}\|_1 = \|\beta'\|_1.$$

Par conséquent β' minimise (3.4), et par hypothèse $\tilde{\beta}$ est unique on en déduit que :

$$\beta' = \tilde{\beta}.$$

Ceci implique que tout point adhérent de (β_{λ_n}) converge vers β', c-a-d :

$$\lim_{\lambda \to 0^+} \beta_\lambda = \tilde{\beta}.$$

□

Théorème 3.2. *Si la solution de (3.4), notée $\tilde{\beta}$, est unique et si à chaque itération de l'algorithme le l qui minimise l'expression est unique alors l'algorithme de la méthode homotopique va aboutir à $\tilde{\beta}$.*

Démonstration. [5] Theorem 15.2 Page 479. □

Partie 4

Implémentation de la méthode homotopique

4.1 Algorithme

Cet Algorithme privilégie la clarté plutôt que de réduire la complexité. La notation argmax utilisée de manière abusive désigne ici la l'indice de ligne du vecteur sur laquelle la norme est atteinte.

L'algorithme implémenté sur R Studio ainsi que deux fonctions, permettant d'obtenir les normes, min et les indices associés, dont l'algorithme dépend sont consultables en annexe.

Algorithm 1 methode homotopique (X,Y,$\lambda_{\text{limite}} = 0$,tol$= 0$)
X matrice de taille (n,p), Y vecteur de taille n, λ_{limite} et tol des paramètres optionnels positifs permettant un arrêt prématuré de l'algorithme.

Ensure: nb de lignes de X = longueur de Y

$n \leftarrow$ nb de lignes de X
$p \leftarrow$ nb de colonnes de X
$\beta \leftarrow 0 \in \mathbb{R}^p$
$\lambda \leftarrow \|X^T(X\beta - Y)\|_\infty$
Set $\leftarrow$ argmax$\|X^T(X\beta - Y)\|_\infty$
Liste $\leftarrow \lambda \cup \beta$
while $(\lambda > \lambda_{\text{limite}})$ et $(\|X\beta - Y\|_2 >$ tol) et (longueur de Set $\leq n$) **do**
  $c \leftarrow X^T(X\beta - Y)$
  $d \leftarrow 0 \in \mathbb{R}^p$
  $d_{\text{set}} \leftarrow -(X_{\text{set}}^T X_{\text{set}})^{-1} \times \text{signe}(c_{\text{set}})$
  $\gamma = \min \left\{ \underset{l \notin S_k}{\overset{+}{\min}} \left\{ \frac{\lambda^{k-1} - C_l^k}{1-(X^T X d^k)_l}, \frac{\lambda^{k-1} + C_l^k}{1+(X^T X d^k)_l} \right\}, \underset{l \in S_k, d_l^k \neq 0}{\overset{+}{\min}} \left\{ -\beta_l^{(k-1)} / d_l^k \right\} \right\}$
  $l = \text{argmin} \left\{ \underset{l \notin S_k}{\overset{+}{\min}} \left\{ \frac{\lambda^{k-1} - C_l^k}{1-(X^T X d^k)_l}, \frac{\lambda^{k-1} + C_l^k}{1+(X^T X d^k)_l} \right\}, \underset{l \in S_k, d_l^k \neq 0}{\overset{+}{\min}} \left\{ -\beta_l^{(k-1)} / d_l^k \right\} \right\}$
  if $\lambda - \gamma < 0$ **then**
    $\gamma \leftarrow \lambda$
  end if
  if $l \in$ Set **then**
    Set $\leftarrow$ Set$\backslash\{l\}$
  else
    Set $\leftarrow$ Set $\cup \{l\}$
  end if
  $\lambda \leftarrow \lambda - \gamma$
  $\beta \leftarrow \beta + \gamma d$
  Liste $\leftarrow$ Liste $\cup \lambda \cup \beta$
end while
return Liste

4.2 Simulation avec un β^0 creux

Dans chaque exemple on a généré aléatoirement une matrice X un vecteur β^0 creux, puis on a calculé $Y := X\beta^0 + \epsilon$ où ϵ est un vecteur gaussien centré appelé bruit (fonctions consultables en annexe). On regardera alors l'évolution des coefficients de $\beta := \beta(\lambda)$ en fonction de λ, l'erreur résiduelle et l'écart entre β^0 et β en fonction de λ.

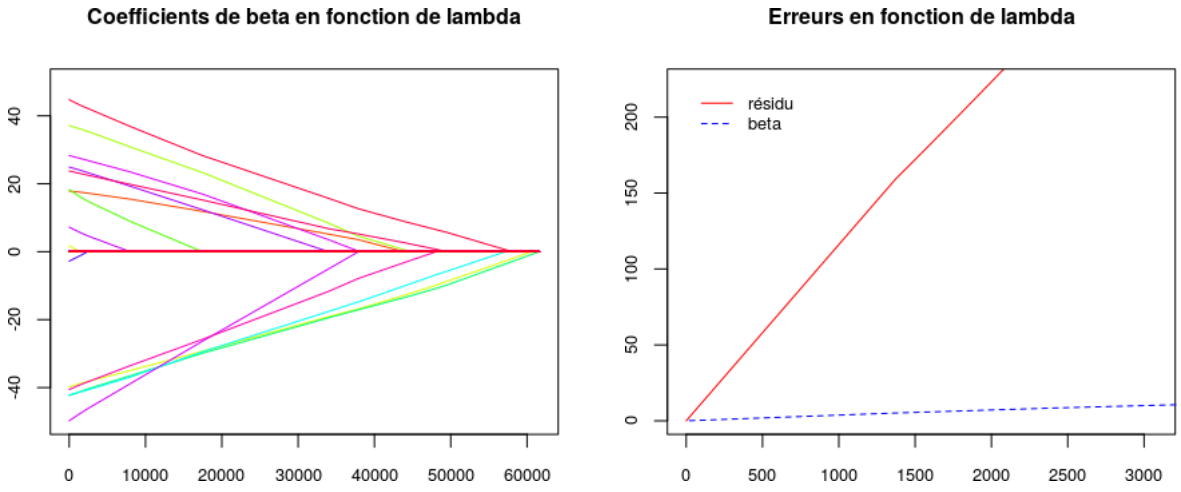
On prend un $\beta^0 \in \mathbb{R}^{150}$ dont 15 coefficients sont non nuls. Les coefficients non nuls de β^0 varient uniformément entre -50 et 50 et les coefficients de X entre -5 et 5.

Pour la variance du bruit, avec une valeur fixe l'erreur aurait parfois eu un impact insignifiant ou démesuré sur les coefficients de y, on a choisi de prendre une valeur qui dépend des coefficients du vecteur $X\beta$: un cinquième de la moyenne en valeur absolue du vecteur $X\beta^0$ i.e $\frac{1}{5n}\sum_{i=1}^{n}|(X\beta^0)_i|$. Ce qui nous semble un bruit non négligeable.

4.2.1 n ≥ p, sans bruit

$X \in \mathcal{M}_{150\times150}(\mathbb{R})$, on est dans le cas où l'équation $X\beta = Y$ a pour solution unique β^0, d'après le Théorème 3.2. l'algorithme va converger vers β^0, (code consultable en annexe).

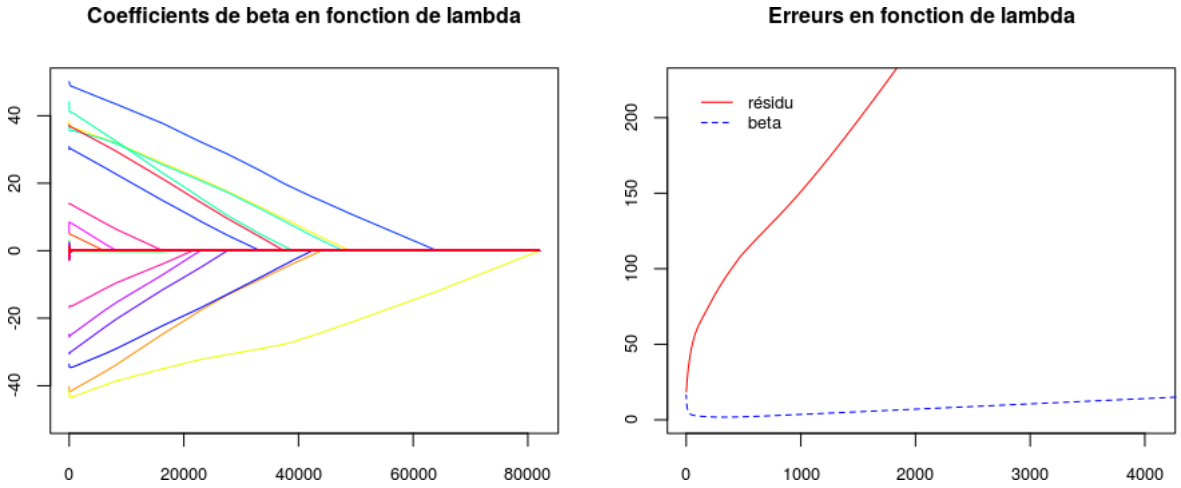
Sur le graphique de gauche on peut observer l'évolution des coefficients de β en fonction de λ, chaque courbe représente un coefficient. Sur le graphique de droite, la courbe en trait continu représente l'erreur résiduelle et celle en petits trait espacés représente $\|\beta - \beta^0\|_2$ en fonction de lambda.



Le $\lambda = 9 \times 10^{-11}$ minimise les deux erreurs, on a $\beta \simeq \beta^0$ à 4×10^{-13} près.

4.2.2 n ≥ p, avec bruit

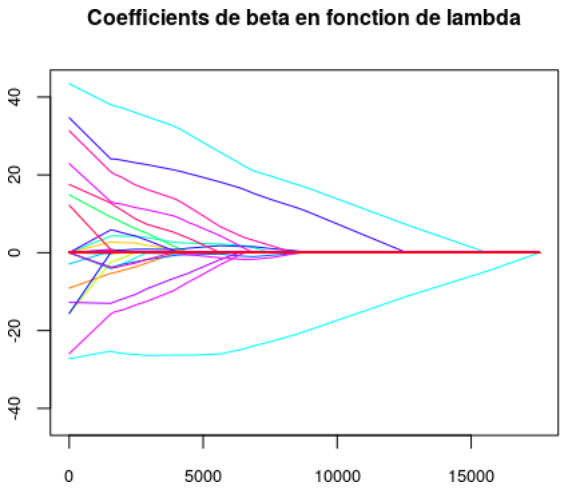
$X \in \mathcal{M}_{150\times150}(\mathbb{R})$, (code consultable en annexe).

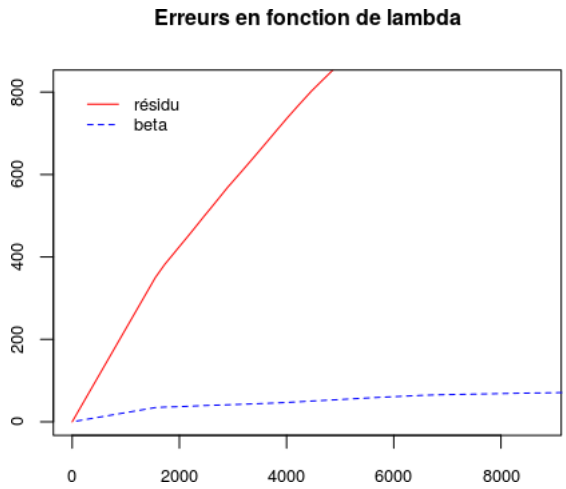


Le λ qui minimise l'erreur résiduelle est : 0 et le λ qui minimise l'écart entre β et β^0 est : 322.7051. Ici choisir $\lambda = 0$ reviens à faire une régression classique M.C.O, mais on peut constater qu'on obtient une valeur de β plus pertinente pour $\lambda = 322.7051$ avec $\|\beta - \beta^0\|_2 = 1.749792$ contre 16.36361 pour $\lambda = 0$. Ce résultat n'est pas surprenant car quand λ est très petit l'algorithme ajuste de plus en plus de coefficients pour diminuer l'erreur résiduelle qui elle dépend du bruit.

4.2.3 n < p, sans bruit

$X \in \mathcal{M}_{50\times150}(\mathbb{R})$, on se place dans le cadre où le critère des M.C.O n'est plus exploitable (code consultable en annexe).

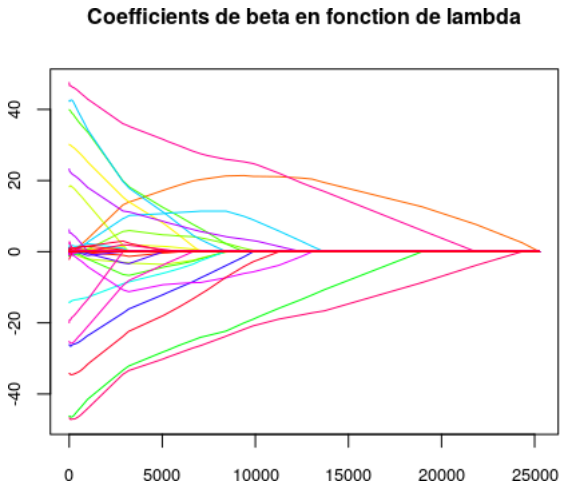


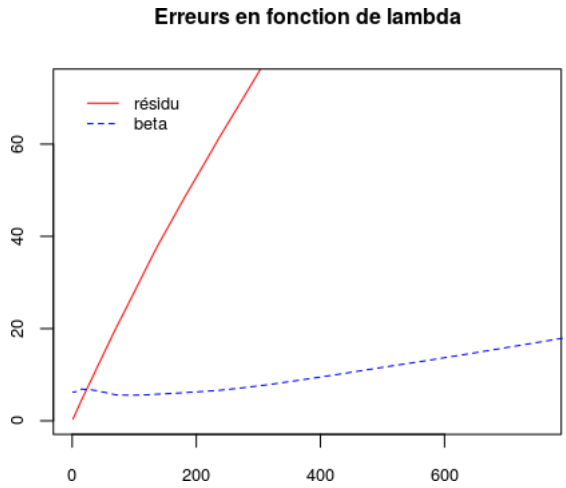


L'erreur résiduelle et l'écart entre β et β^0 sont minimisés pour le même $\lambda = 1.323883 \times 10^{-10}$, respectivement 3.039102×10^{-11} et 3.070182×10^{-12}. les erreurs sont très proches de zéro, on est quasiment à la précision obtenue quand la matrice avait 150 lignes (observations) contre 50 ici.

4.2.4 n < p, avec bruit

$X \in \mathcal{M}_{50\times150}(\mathbb{R})$, on est dans la cas typique où le lasso va être utilisé : M.C.O non utilisable, bruit sur Y, et β^0 creux, (code consultable en annexe).





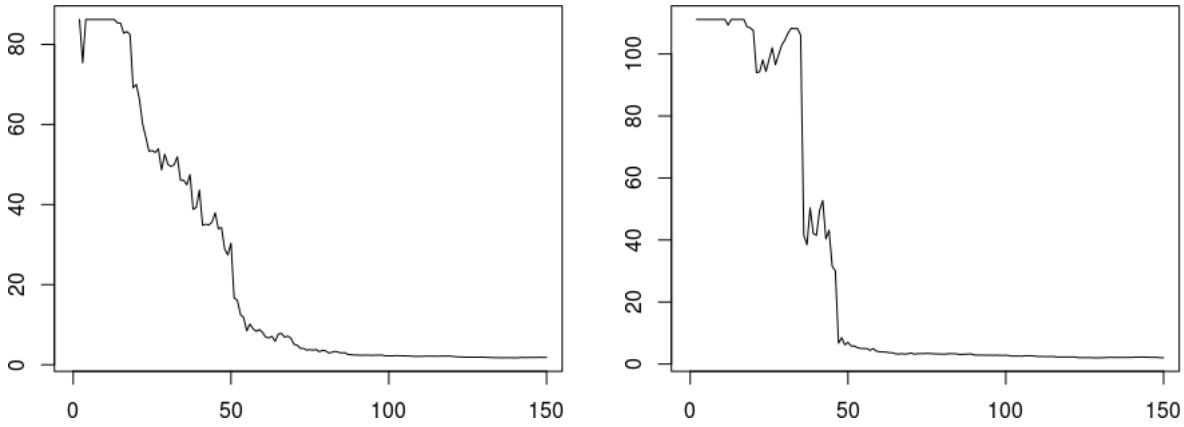
On remarque encore une fois que quand λ est très petit l'algorithme ajuste de plus en plus de coefficients pour diminuer l'erreur résiduelle en s'éloignant de β^0. En

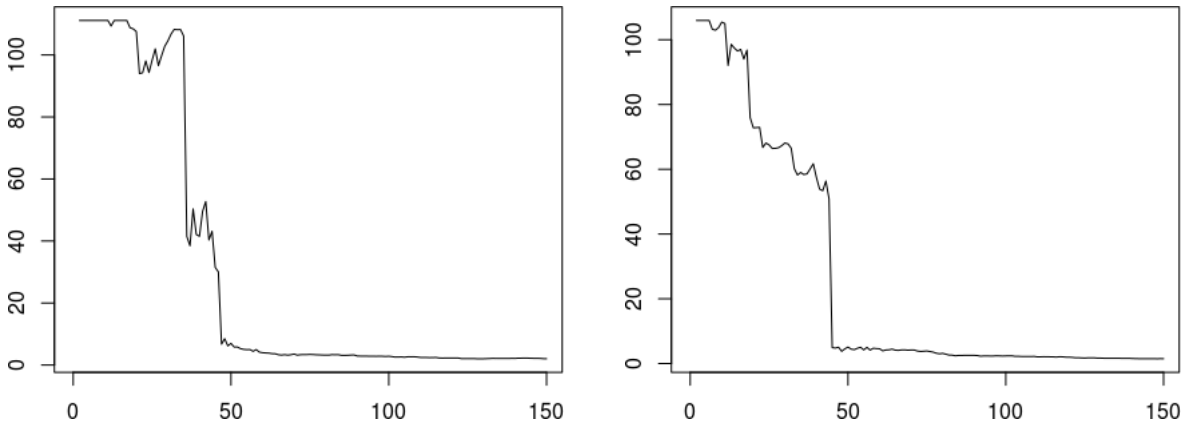
comparant avec l'exemple précédent on peut considérer que ceci est dû au buit non négligeable sur Y.

L'erreur résiduelle est minimisée pour $\lambda = 1.264159$ et vaut : 0.3775168. L'écart entre β et β^0 est minimisé pour $\lambda = 92.61071$ et vaut : 5.564271, cette erreur est plus grande que dans le cas où la matrice X contenait 150 lignes (observations), une question surgit alors : comment évolue cette erreur quand on réduit le nombre d'observations dans la matrice des données ?

4.2.5 Réduction progressive du nombre d'observations

Chacun des graphiques est créé à partir d'une seule matrice $X \in \mathcal{M}_{150\times150}(\mathbb{R})$, d'un vecteur Y contenant les données pour lesquels on ne prend en compte qu'une partie des lignes et d'un seul β^0. On regarde le résultat pour quatre graines différentes (code consultable en annexe). Ces graphiques représentent $\|\beta - \beta^0\|_2$ en fonction de nombre de lignes de la matrice X.



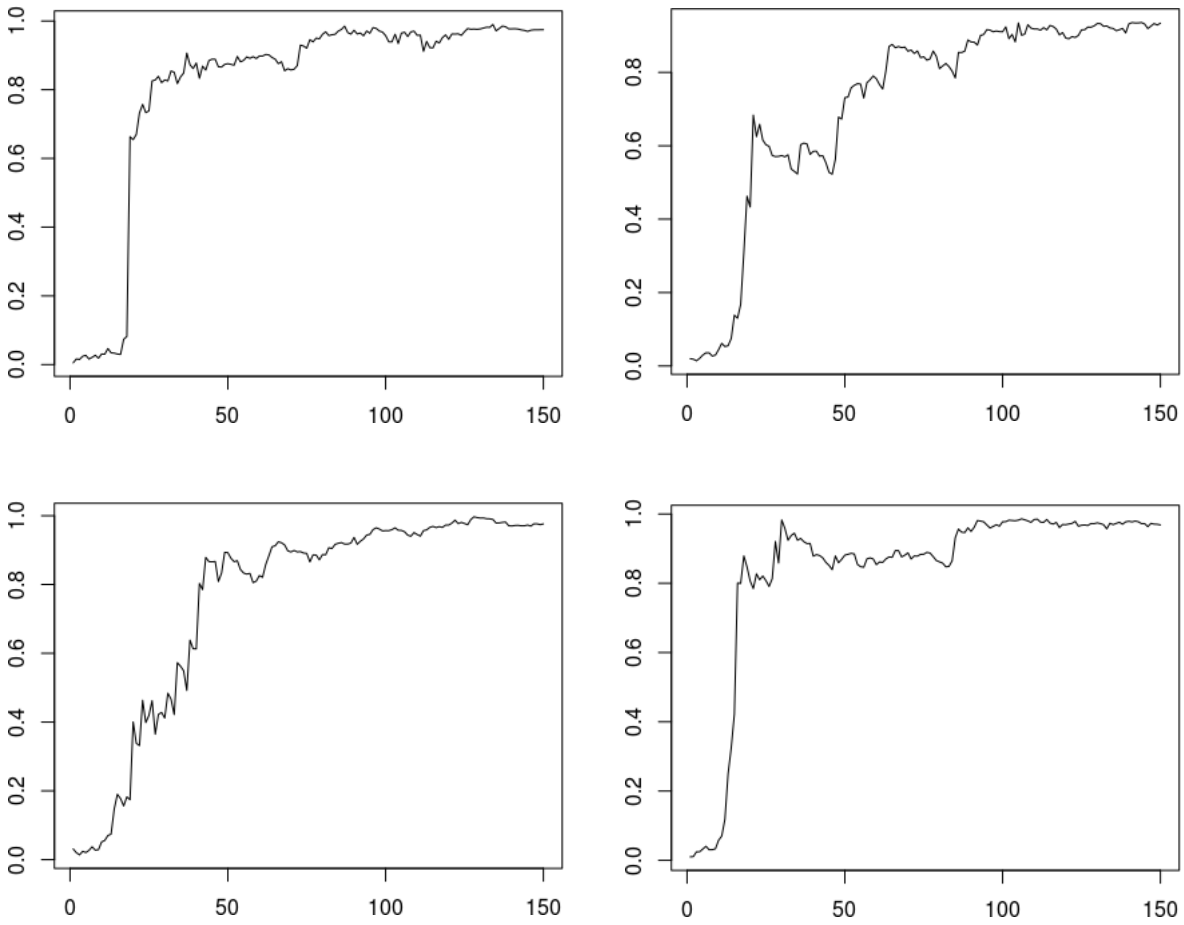


On observe une augmentation significative de $\|\beta - \beta^0\|_2$ quand le nombre d'observations contenues dans X passe en dessous de 50, soit dans notre cas un tiers du nombre de caractéristiques. On aperçoit une limite du Lasso n ne doit pas non plus être trop petit devant p si on veut obtenir un résultat pertinent.

4.3 Simulations avec un β^0 de moins en moins creux

Précédemment on s'est placé dans le cas où β^0 est creux, mais que se passe-t-il dans le cas contraire ? On a choisi de regarder comment évolue l'écart minimum entre β estimé et β^0 quand β^0 est de moins en moins creux. On choisi $X \in \mathcal{M}_{100\times150}(\mathbb{R})$, $\beta^0 \in \mathbb{R}^{150}$ et une variance du bruit de : $\frac{1}{5n}\sum_{i=1}^{n}|(X\beta^0)_i|$.

La matrice X est fixé pour chaque graphique et on ajoute progressivement des coefficients non nul à β^0. Ces graphiques représentent $\frac{\|\beta-\beta^0\|_2}{\|\beta^0\|_2}$ en fonction du nombre de coefficients non nuls de β^0(code consultable en annexe).



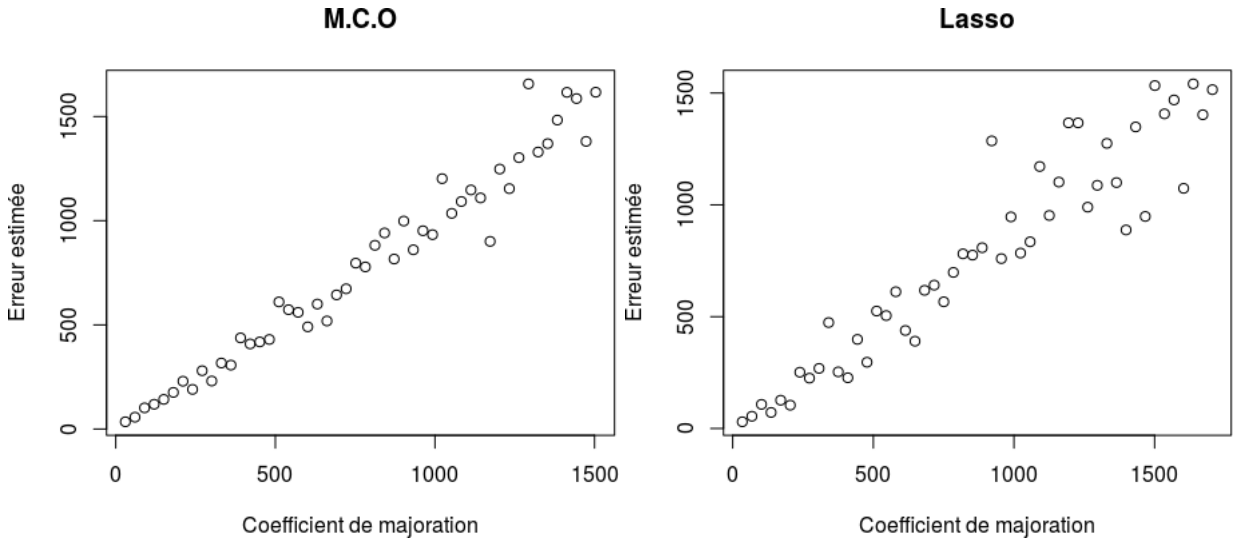
On observe une augmentation très importante à partir d'une vingtaine de coefficients non nuls et une fois cinquante coefficients non nuls l'erreur est comparable à la norme de β^0.

On a ici une autre limite du Lasso que l'on pouvait remarquer dans l'étude théorique car le terme qui majore l'erreur estimée : $const\frac{\sigma^2 \log(p)}{n} s_0$ dépend de $\log(p)$ et de s_0 donc si on est dans les cas où β^0 n'a que peu de coefficients nuls et que $p \approx n$ alors $s_0 \approx n$ et le terme de majoration devient $const \times \sigma^2 \log(p)$ contre $const \times \sigma^2$ dans la cas d'une régression linéaire classique.

Si on veut que le Lasso nous donne un β proche de β^0 il est important que β soit creux.

4.4 Vérifications empirique des majorations de l'erreur estimée

Les figures ci-après illustrent l'évolution de l'erreur estimée $\frac{\|X(\widehat{\beta}-\beta^0)\|_2^2}{n}$ pour les méthodes MCO et Lasso) en fonction de leur majoration ($\sigma^2\frac{p}{n}$, respectivement $\frac{\sigma^2\log(p)}{n}s_0$) vues dans les parties 1 et 2 dont les paramètres sont fixées, hormis la variance σ^2. En d'autres termes, cette erreur varie en fonction de la variance σ^2. Ces figures sont obtenues par application de la méthode homotopique étudiée dans la troisième partie.



On remarque que dans le cadre standard des M.C.O, tout comme dans le cadre du Lasso, la répartition des points se situe approximativement le long d'une droite linéaire, Autrement dit, l'erreur estimée augmente de façon linéaire par rapport à la variance du bruit. Cela est en cohérence avec l'étude théorique menée dans les parties 1 et 2.

Annexe A

Code RStudio

A.1 Fonctions

A.1.1 Méthode homotopique

La fonction "min_positif" est utilisée dans la fonction "methode_homotopique". Elle prend en paramètres la matrice des observations X, le réel $\lambda^{(k-1)}$, le vecteur $c := X^T(X\beta - Y)$ de taille p, le vecteur de direction d de taille p et l'ensemble active set de taille $\leq p$.
min_positif(X,lambda,c,d,set)$min : renvoie le pas γ^k.
min_positif(X,lambda,c,d,set)$indice : renvoie l'indice pour lequel ce minimum est atteint.
min_positif(X,lambda,c,d,set)$position : si le minimum est atteint pour un élément du "set", renvoie sa position dans le set.

```
min_positif <- function (X,lambda,c,d,set,beta){
  non_set <- (1:dim(X)[2])[-set]
  min <- Inf
  indice <- 0
  for (l in non_set){
    produit <- (t(X)[l,]%*%X%*%d)
    temp <- (lambda+c[l])/(1-produit)
    if ((temp>0)&(temp<min)){
      min <- temp
      indice <- l
    }
    temp <- (lambda-c[l])/(1+produit)
```

```
    if ((temp>0)&(temp<min)){
      min <- temp
      indice <- l
    }
  }
  position <- 0
  position_temp <- 0
  for (l in set){
    position_temp <- position_temp + 1
    if (d[l]!=0){
      temp <- -beta[l]/d[l]
      if ((temp>0)&(temp<min)){
        min <- temp
        indice <- l
        position <- position_temp
      }
    }
  }
  return(list("min"=min,"indice"=indice,"position"=position))
}
```

La fonction "norm_inf" prend en paramètres un vecteur renvoie une liste avec la norme infinie et le 1er indice pour lequel cette norme est atteinte.
norm_inf(vecteur)$norme : renvoie la norme infinie de vecteur.
norm_inf(vecteur)$indice : renvoie le 1er indice pour lequel cette norme est atteinte.

```
norm_inf <- function (vecteur){
  norme <- 0
  indice <- 0
  for (i in 1:length(vecteur)){
    if (abs(vecteur[i])>norme){
      norme <- abs(vecteur[i])
      indice <- i
    }
  }
  return(list("norme"=norme,"indice"=indice))
}
```

La fonction "méthode_homotopique" prend en paramètres la matrice des observations X de taille (n, p), le vecteur des réponses Y de taille n et des paramètres

optionnels lambda_limite et tol si on souhaite forcer l'arrêt de la méthode passé un certain λ ou une fois que l'erreur résiduelle est suffisamment petite.
Cette fonction renvoie une liste contenant une suite de $(\lambda, \beta(\lambda))$ coordonnées des points aux extrémités de chaque partie affine de la fonction Lasso.

```
methode_homotopique <- function (X,Y,lambda_limite=0,tol=0){
  n <- dim(X)[1]
  p <- dim(X)[2]
  if (length(Y)!=n){return("Les paramètres d'entrée ne sont
                           pas de dimensions compatibles")}
  beta <- vector(mode = "double",length = p)
  c <- t(X)%*%(X%*%beta-Y)
  norme_c <- norm_inf(c)
  lambda <- norme_c$norme
  retour <- list(lambda,beta)
  set <- norme_c$indice
  while ((lambda > lambda_limite)&(length(set)<=n)
         &(norm(X%*%beta-Y,"2")>tol)){
    c <- t(X)%*%(X%*%beta-Y)
    d <- vector(mode = "double",length = p)
    d[set] <- solve(t(X[,set])%*%X[,set])%*%(-sign(c[set]))
    gamma_indice <- min_positif(X,lambda,c,d,set,beta)
    if ((lambda-gamma_indice$min)<0){
      gamma_indice$min <- lambda
    }
    lambda <- lambda - as.vector(gamma_indice$min)
    beta <- beta + as.vector(gamma_indice$min) * d
    if (gamma_indice$indice%in%set){
      set <- set[-gamma_indice$position]
    }else{
      set <- c(set,gamma_indice$indice)
    }
    retour[[length(retour)+1]] <- lambda
    retour[[length(retour)+1]] <- beta
  }
  return(retour)
}
```

A.1.2 Simulation d'exemples

La fonction "rand_beta" crée un β aléatoire parcimonieux (sparse).
a est le nombre de coefficients non nuls.
p est le nombre de coefficients total.
amplitude (optionnel) est un coefficient multiplicateur sur beta.

```
rand_beta <- function(a,p,amplitude=1){
  beta <- rep(0,p)
  beta_indice <- seq(1:p)
  for (i in 1:a){
    indice <- runif(1,1,length(beta_indice)+1)%/%1
    beta[beta_indice[indice]] <- amplitude*runif(1,-1,1)
    beta_indice <- beta_indice[-indice]
  }
  return (beta)
}
```

La fonction "rand_X" crée une matrice X aléatoirement.
n est le nombre de lignes de X.
p est le nombre de colonnes de X.
amplitude (optionnel) est un coefficient multiplicateur sur X.

```
rand_X <- function(n,p,amplitude=1){
  X <- matrix(nrow=n,ncol=p)
  for (i in 1:n){
    X[i,] <- amplitude*runif(p,-1,1)
  }
  return (X)
}
```

La fonction "rand_Y" génère un $Y := X\beta +$ bruit, où bruit est un vecteur gaussien de loi N(0, sigma$\times$Id).
X est la matrice des observations.
beta est le vecteur des coefficients de la régression.
sigma est la variance du bruit.
Cette fonction utilise le package "mvtnorm"

```
library(mvtnorm)
rand_Y <- function(X,beta,sigma){
```

```
  epsi <- t(rmvnorm(1,rep(0,dim(X)[1]),sigma*diag(dim(X)[1])))
  return (X%*%beta+epsi)
}
```

A.1.3 Affichage des résultats

La fonction "dessin_methode_homotopique" prend en paramètres une liste contenant une suite de $(\lambda, \beta(\lambda))$ et affiche l'évolution des différents coefficients de β en fonction de λ.
Les paramètres optionnels x et y permettent d'ajuster le zoom du graphique sur l'axe des abscisses et des ordonnées.

```
dessin_methode_homotopique <- function(liste,x=1,y=1){
  lambda <- vector()
  beta <- matrix(nrow = length(liste[[2]]),ncol = (length(liste)/2))
  couleur <- rainbow(length(liste[[2]]))
  for (i in 1:(length(liste)/2)){
    lambda <- c(lambda,liste[[2*i-1]])
    beta[,i] <- liste[[2*i]]
  }
  plot(lambda,beta[1,],ylim = c(-y*max(abs(beta)),y*max(abs(beta)))
       ,xlim = c(0,x*liste[[1]]),type = "l",col = couleur[1],
       ylab=' ',xlab=' ', font.axis=1, cex.axis=0.8, cex.lab=0.8
       ,main="Coefficients de beta en fonction de lambda",cex.main=1)
  for (l in 2:length(liste[[2]]))
    lines(lambda,beta[l,],col = couleur[l])
}
```

La fonction "discretisation_beta_lambda" prend en paramètres une liste contenant une suite de $(\lambda, \beta(\lambda))$ et un pas. Elle calcule $\beta(\lambda)$ et λ en discrétisant d'avantage les valeurs de lambda en fonction du pas.

```
discretisation_beta_lambda <- function(liste,pas){
  retour <- list(liste[[1]],liste[[2]])
  lambda <- liste[[1]]
  n <- 0
  last_lambda <- liste[[length(liste)-1]]
  while (lambda - pas > last_lambda){
    beta_nplus1 <- liste[[2*(n+1)+2]]
```

```
    beta_n <- liste[[2*n+2]]
    lambda_nplus1 <-liste[[2*(n+1)+1]]
    lambda_n <- liste[[2*n+1]]
    coeff <- (beta_nplus1-beta_n)/(lambda_nplus1-lambda_n)
    while (lambda - pas > lambda_nplus1){
      lambda <- lambda - pas
      beta <- beta_n + coeff * (lambda - lambda_n)
      retour[[length(retour)+1]] <- lambda
      retour[[length(retour)+1]] <- beta
    }
    retour[[length(retour)+1]] <- lambda_nplus1
    retour[[length(retour)+1]] <- beta_nplus1
    n <- n+1
  }
  return (retour)
}
```

La fonction "lambda_opti_beta" prend en paramètres une liste contenant une suite de $(\lambda, \beta(\lambda))$ et β^0. Elle cherche le λ et donc aussi le $\beta(\lambda)$ qui minimise la différence entre le β^0 et le $\beta(\lambda)$ en norme 2. Cette fonction ne teste les différences qu'aux valeurs dans la liste.

```
lambda_opti_beta <- function(liste,beta){
  nb_beta <- length(liste)/2
  donnee <- matrix(nrow = length(beta),ncol = nb_beta)
  min <- Inf
  for (i in 1:nb_beta){
    temp <- norm(liste[[2*i]]-beta,type = "2")
    if (temp<min){
      min <- temp
      best_beta <- 2*i
    }
  }
  return (list("lambda"=liste[[best_beta-1]]
               ,"best_beta"=liste[[best_beta]]
               ,"indice"=best_beta,"erreur_beta"=min))
}
```

La fonction "lambda_opti_residus" prend en paramètres la matrice des observations X de taille (n, p), le vecteur des réponses Y de taille n, une liste contenant une

suite de $(\lambda, \beta(\lambda))$ et β^0 de taille p. Elle cherche le λ et donc aussi le $\beta(\lambda)$ qui minimise l'erreur résiduelle. Cette fonction ne teste les différences qu'aux valeurs dans la liste.

```
lambda_opti_residus <- function(X,Y,liste,beta){
  nb_beta <- length(liste)/2
  donnee <- matrix(nrow = length(beta),ncol = nb_beta)
  min <- Inf
  for (i in 1:nb_beta){
    temp <- norm(X%*%liste[[2*i]]-Y,type = "2")
    if (temp<min){
      min <- temp
      best_beta <- 2*i
    }
  }
  return (list("lambda"=liste[[best_beta-1]],"indice"=best_beta
               ,"norme2_residus"=min,"best_beta"=liste[[best_beta]]
               ,"erreur_beta"=norm(liste[[best_beta]]-beta,type = "2")))
}
```

La fonction "dessin_erreur" prend en paramètres une liste contenant une suite de $(\lambda, \beta(\lambda))$, la matrice des observations X de taille (n, p), le vecteur des réponses Y de taille n et β^0 de taille p. Elle affiche l'erreur résiduelle et la différence entre le β^0 et le $\beta(\lambda)$ en norme 2 en fonction de λ.
Les paramètres optionnels x et y permettent d'ajuster le zoom du graphique sur l'axe des abscisse et des ordonnées.

```
dessin_erreur <- function(liste,X,Y,beta,x=1,y=1){
  n <- length(liste)/2
  axe_x <- vector()
  axe_y <- vector()
  axe_z <- vector()
  for (i in 1:n){
    axe_x <- c(axe_x,liste[[2*i-1]])
    axe_y <- c(axe_y,norm(X%*%liste[[2*i]]-Y,type = "2"))
    axe_z <- c(axe_z,norm(liste[[2*i]]-beta,type = "2"))
  }
  lim <- max(c(axe_y,axe_z))
  plot(axe_x,axe_y,type = "l",ylab = ' ',xlab = ' ',
       xlim = c(0,x*axe_x[1]), ylim = c(0,y*lim), col="red",lty=1
```

```
      ,font.axis=1,cex.axis=0.8,cex.lab=0.8
      ,main = "Erreurs en fonction de lambda",cex.main=1)
  legend(0,y*lim,legend = c("résidu","beta"),col=c("red","blue")
         , lty=1:2, box.lty=0, cex=0.8)
  lines(axe_x,axe_z,col="blue",lty=2)
}
```

La fonction "dessin_erreur_min" prend en paramètres la matrice des observations X de taille (n, p), le vecteur des réponses Y de taille n et β^0 de taille p. Elle affiche l'écart minimum entre β et β^0 en fonction du nombre d'observations dans la matrice X.
x est un paramètre optionnel permettant d'ajuster le zoom du graphique sur l'axe des abscisse.

```
dessin_erreur_min <- function(X,Y,beta,x=1){
  n <- dim(X)[1]
  axe_x <- vector()
  axe_y <- vector()
  for (i in 2:n){
    liste <- methode_homotopique(X[1:i,],Y[1:i])
    axe_y[i] <- lambda_opti_beta(liste,beta)$erreur_beta
    axe_x[i] <- i
  }
  plot(axe_x,axe_y,type = "l",ylab = ' ',xlab = ' '
       ,xlim = c(0,x*axe_x[n-1]))
}
```

La fonction "dessin_residu_erreur_min" prend en paramètres la matrice des observations X de taille (n, p) et β^0 de taille p. Elle affiche l'écart minimum entre β et β^0 divisé par la norme de β^0 en fonction du nombre de coefficient non nuls de β^0

```
dessin_residu_erreur_min <- function(X,beta){
  p <- length(beta)
  axe_x <- vector()
  axe_y <- vector()
  for (i in 1:p){
    beta_temp <- rep(0,p)
    beta_temp[1:i] <- beta[1:i]
    sigma <- mean(abs(X%*%beta_temp))/5
    Y <- rand_Y(X,beta_temp,sigma)
```

```
    liste <- methode_homotopique(X,Y)
    axe_y[i] <- lambda_opti_beta(liste,beta_temp)$erreur_beta/norm(beta_temp,"2")
    axe_x[i] <- i
  }
  plot(axe_x,axe_y,type = "l",ylab = ' ',xlab = ' ')
}
```

La fonction "maj_erreur_estimée_mco" prend en paramètres la matrice des observations X de taille (n, p) et β^0 de taille p. Elle affiche, pour une variance du bruit qui augmente, l'erreur estimée en fonction de la majoration obtenue dans le cadre des M.C.O.

```
maj_erreur_estimée_mco <- function(X,beta){
  axe_x <- vector()
  axe_y <- vector()
  for (i in 1:50){
    sigma <- mean(abs(X1%*%beta1))*(i*0.1)
    Y1 <- rand_Y(X1,beta1,sigma)
    liste1 <- methode_homotopique(X1,Y1)
    axe_y[i] <- norm(X1%*%(lambda_opti_residus(X1,
        Y1,liste1,beta1)$best_beta-beta1),"2")^2/(dim(X)[1])
    axe_x[i] <- sigma
  }
  print(axe_x)
  print(axe_y)
  plot(axe_x,axe_y,ylab = "Erreur estimée",xlab = "Coefficient de majoration")
}
```

La fonction "maj_erreur_estimée_lasso" prend en paramètres la matrice des observations X de taille (n, p) et β^0 de taille p. Elle affiche, pour une variance du bruit qui augmente, l'erreur estimée en fonction de la majoration obtenue dans le cadre du Lasso.

```
maj_erreur_estimée_lasso <- function(X,beta){
  axe_x <- vector()
  axe_y <- vector()
  for (i in 1:50){
    sigma <- mean(abs(X1%*%beta1))*(i*0.1)
    Y1 <- rand_Y(X1,beta1,sigma)
```

```
    liste1 <- methode_homotopique(X1,Y1)
    axe_y[i] <- norm(X1%*%(lambda_opti_residus(X1,
        Y1,liste1,beta1)$best_beta-beta1),"2")^2/(dim(X)[1])
    axe_x[i] <- sigma*log(dim(X)[1])*15/(dim(X)[1])
    # car la fonction est utilisée pour un
    # beta avec 15 coefficients non nuls
  }
  print(axe_x)
  print(axe_y)
  plot(axe_x,axe_y,ylab = "Erreur estimée",xlab = "Coefficient de majoration")
}
```

A.2 Appel des fonctions

Graphique de la page de couverture

```
set.seed(19648)
beta1 <- rand_beta(150,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/10
Y1 <- rand_Y(X1,beta1,sigma)
liste1 <- methode_homotopique(X1,Y1)
pas <- (liste1[[1]]-liste1[[length(liste1)-1]])/3000
liste1 <- discretisation_beta_lambda(liste1,pas)
dessin_methode_homotopique(liste1)
```

A.2.1 Exemple 4.2.1

```
set.seed(4546)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/10
Y1 <- rand_Y(X1,beta1,0)
liste1 <- methode_homotopique(X1,Y1)
pas <- (liste1[[1]]-liste1[[length(liste1)-1]])/3000
liste1 <- discretisation_beta_lambda(liste1,pas)
dessin_methode_homotopique(liste1)
dessin_erreur(liste1,X1,Y1,beta1,0.05,0.05)
```

```
lambda_opti_residus(X1,Y1,liste1,beta1)
lambda_opti_beta(liste1,beta1)
```

A.2.2 Exemple 4.2.2

```
set.seed(898456)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,sigma)
liste1 <- methode_homotopique(X1,Y1)
pas <- (liste1[[1]]-liste1[[length(liste1)-1]])/3000
liste1 <- discretisation_beta_lambda(liste1,pas)
dessin_methode_homotopique(liste1)
dessin_erreur(liste1,X1,Y1,beta1,0.05,0.05)
lambda_opti_residus(X1,Y1,liste1,beta1)
lambda_opti_beta(liste1,beta1)
```

A.2.3 Exemple 4.2.3

```
set.seed(9854)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(50,150,5)
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,0)
liste1 <- methode_homotopique(X1,Y1)
pas <- (liste1[[1]]-liste1[[length(liste1)-1]])/3000
liste1 <- discretisation_beta_lambda(liste1,pas)
dessin_methode_homotopique(liste1)
dessin_erreur(liste1,X1,Y1,beta1,0.5,0.5)
lambda_opti_residus(X1,Y1,liste1,beta1)
lambda_opti_beta(liste1,beta1)
```

A.2.4 Exemple 4.2.4

```
set.seed(94657)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(50,150,5)
```

```
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,sigma)
liste1 <- methode_homotopique(X1,Y1)
pas <- (liste1[[1]]-liste1[[length(liste1)-1]])/3000
liste1 <- discretisation_beta_lambda(liste1,pas)
dessin_methode_homotopique(liste1)
dessin_erreur(liste1,X1,Y1,beta1,0.03,0.03)
lambda_opti_residus(X1,Y1,liste1,beta1)
lambda_opti_beta(liste1,beta1)
```

A.2.5 Réduction du nombre d'observations 4.2.5

```
set.seed(2654)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,sigma)
dessin_erreur_min(X1,Y1,beta1)
#########
set.seed(166)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,sigma)
dessin_erreur_min(X1,Y1,beta1)
#########
set.seed(154854)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,sigma)
dessin_erreur_min(X1,Y1,beta1)
#########
set.seed(987)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
sigma <- mean(abs(X1%*%beta1))/5
Y1 <- rand_Y(X1,beta1,sigma)
```

```
dessin_erreur_min(X1,Y1,beta1)
```

A.3 β^0 de moins en moins creux 4.3

```
set.seed(498)
beta1 <- rand_beta(150,150,50)
X1 <- rand_X(50,150,5)
dessin_residu_erreur_min(X1,beta1)
#######
set.seed(15647)
beta1 <- rand_beta(150,150,50)
X1 <- rand_X(50,150,5)
dessin_residu_erreur_min(X1,beta1)
#######
set.seed(91199)
beta1 <- rand_beta(150,150,50)
X1 <- rand_X(50,150,5)
dessin_residu_erreur_min(X1,beta1)
#######
set.seed(1984)
beta1 <- rand_beta(150,150,50)
X1 <- rand_X(50,150,5)
dessin_residu_erreur_min(X1,beta1)
```

A.4 Vérifications empirique 4.4

```
set.seed(898456)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(150,150,5)
maj_erreur_estimée_mco(X1,beta1)
#########
set.seed(94657)
beta1 <- rand_beta(15,150,50)
X1 <- rand_X(50,150,5)
maj_erreur_estimée_mco(X1,beta1)
```

Bibliographie

[1] Anisse Ismaili et Pierre Gaillard. "le lasso, ou comment choisir parmi un grand nombre de variables à l'aide de peu d'observations", juin 2009. http ://pierre.gaillard.me/doc/Ga09-report.pdf.

[2] Ryan Tibshirani etLarry Wasserman. "sparsity and the lasso", Spring 2015. http ://www.stat.cmu.edu/ larry/=sml/sparsity.pdf.

[3] Mohamed Hebiri. *"Quelques questions de sélection de variables autour de l'estimateur LASSO"*. PhD thesis, Université Paris Diderot - Paris7 UFR de Mathématiques, jun 2009.

[4] Bühlmann Peter. *Statistics for high-dimensional data : methods, theory and applications*. Springer series in statistics. Springer, Berlin London New York [etc., 2011.

[5] Foucart Simon. *A mathematical introduction to compressive sensing*. Applied and numerical harmonic analysis. Birkhäuser, New York (N.Y.) Heidelberg Dordrecht [etc., 2013.

[6] V. Vialon. "régression pénalisée : le lasso", jun 2009. http ://pbil.univ-lyon1.fr/members/fpicard/franckpicard$_f ichiers/master/coursLasso.pdf$.