Le Lasso, ou comment choisir parmi un grand nombre de variables à l'aide de peu d'observations *

Anisse Ismaili et Pierre Gaillard

Juin 2009

Résumé

Nous disposons d'un ensemble de variables explicatives $(X_i)_{1\leq i\leq p}$ pour expliquer une variable Y de manière linéaire et rien ne nous assure que toutes les variables interviennent dans l'explication. Nous avons donc à notre disposition un ensemble de variables potentiellement explicatives ou variables candidates. Notre objectif est de déterminer quelles sont les variables réellement explicatives. Il s'agit donc de choisir un modèle parmi les 2^p possibles. Comment choisir le bon modèle ? Tous les étudier n'est tout simplement pas envisageable dès que p est grand, qui plus est, il faudrait savoir déterminer si un modèle est meilleur que les autres. La méthode du Lasso propose dans certains cas une solution à ce problème. Nous allons premièrement en expliquer le principe, puis en donner l'algorithme avant de voir quand est-ce qu'elle est efficace. Enfin, nous allons faire des simulations pour illustrer nos résultats.

Table des matières

*Sujet d'exposé de maîtrise proposé et encadré par Sylvain Arlot.

1 La méthode du Lasso

1.1 L'idée

Dans cette partie nous allons expliquer l'idée de la méthode. Nous cherchons à expliquer de manière linéaire une variable Y, par p variables potentiellement explicatives X_i. Pour cela nous faisons n observations. Nous modèlisons la variable Y de la manière suivante :

$$Y^n = X^n\beta^n + \varepsilon^n$$

où $\varepsilon^n = (\varepsilon_1, ..., \varepsilon_n)^T$ est un vecteur de n variables aléatoires i.i.d. de moyenne 0 et de variance σ^2 correspond au bruit lors des observations (qui peut contenir toutes les variables explicatives non prises en compte dans le modèle) ; $Y^n \in \mathbb{R}^n$ correspond aux n observations de la variable Y et $X^n = (X^n_{.,1}, ..., X^n_{.,p}) = ((X^n_{1,.})^T, ..., (X^n_{n,.})^T)^T$ est une matrice $n \times p$, où $X^n_{.,j}$ est la j^{eme} colonne, qui correspond à j^{eme} prédicteur X_j et $X^n_{i,.}$ la i^{eme} ligne qui correspond à la i^{eme} observation. $\beta^n \in \mathbb{R}^p$ est le paramètre à estimer, on l'indexe par n pour permettre à ses coefficients et à sa taille de varier lorsque n augmente (p peut a priori dépendre de n).

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 un régression linéaire classique mais une régression régularisée qui rend nuls certains coefficients de l'estimation de β. Cela consiste à estimer pour $\lambda \in \overline{\mathbb{R}}_+$:

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

où $\|x\|_2^2 = \sum_{i=1}^n x_i^2$ et $\|x\|_1 = \sum_{i=1}^p |x_i|$.

Le paramètre $\lambda \geq 0$ contrôle la puissance de la régularisation. Si on prend $\lambda = 0$, le Lasso correspond à une régression linéaire classique (si $p \leq n$). Si par contre, on prend $\lambda = \infty$, tous les coefficients de $\widehat{\beta}^n(\infty)$ sont nuls. L'augmentation de λ induit la diminution de certains coefficients de $\widehat{\beta}^n(\lambda)$ vers 0 jusqu'à ce qu'ils soient exactement nuls. On peut montrer que ce modèle est équivalent au suivant :

$$\widetilde{\beta}^n(t) = \underset{\beta,\ \|\beta\|_1 \leq t}{\arg\min} \left\|Y^n - X^n\beta\right\|_2^2$$

dans le sens où pour tout $\lambda \in \overline{\mathbb{R}}_+$, il existe $t \geq 0$ tel que : $\widetilde{\beta}^n(t) = \widehat{\beta}^n(\lambda)$. En effet, il suffit de prendre $t = \|\widehat{\beta}^n(\lambda)\|_1$ car alors pour tout β tel que $\|\beta\|_1 \leq t$, $\lambda\|\beta\|_1 \leq \lambda\|\widehat{\beta}^n(\lambda)\|_1$ donc par définition de $\widehat{\beta}^n(\lambda)$, $\|Y^n - X^n\beta\|_2^2 \geq \|Y^n - X^n\widehat{\beta}^n(\lambda)\|_2^2$.

Cela nous permet de comprendre intuitivement pourquoi le Lasso, dans la plupart des cas, rend exactement nuls certains coefficients de $\widehat{\beta}^n(\lambda)$. La figure 1 illustre en dimension 2 le cas où le Lasso annule une coordonnée. Cela dépend de la position de $A = \arg\min_\beta \|Y^n - X^n\beta\|_2^2$ dans $\mathbb{R}^p$ et de t. Plus t est petit plus le Lasso annule de coordonnées. Lorsque la dimension augmente, la zone dans laquelle A annule au

moins une coordonnée augmente au point de devenir presque l'espace $\mathbb{R}^p$ tout entier.

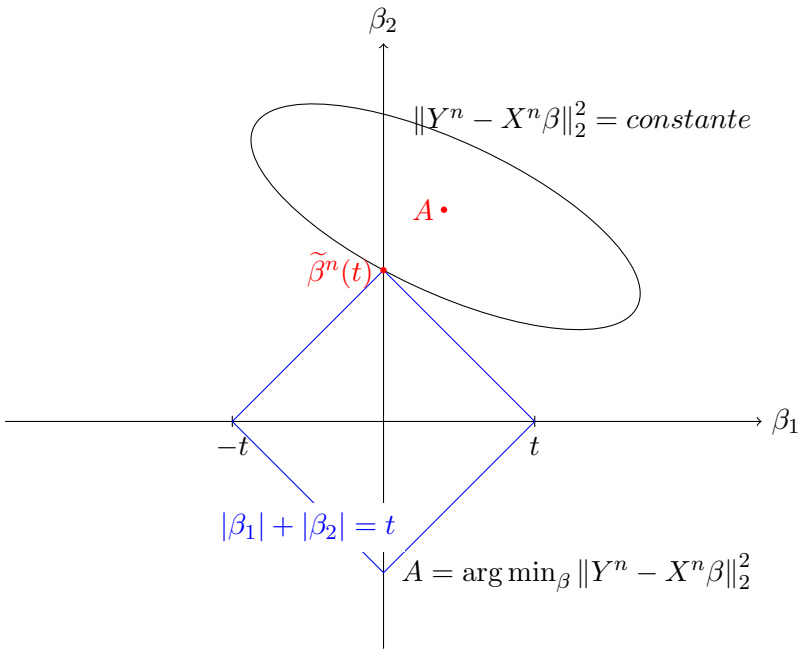

FIGURE 1 – Cas où A se trouve dans la zone qui annule le coefficient β_2.

L'algorithme du Lasso consiste donc à déterminer $\widehat{\beta}^n(\lambda)$ pour tout $\lambda \geq 0$, nous exposerons la méthode dans la partie suivante. Il s'agit ensuite de déterminer le bon λ qui permet de garder uniquement les vraies variables explicatives et d'éliminer les autres. Tout d'abord, l'existence d'un tel λ n'est pas immédiate et est même fausse en générale. Nous étudierons dans la deuxième partie quand est-ce qu'un tel λ existe et nous verrons que si tel est le cas, il est de l'ordre de $\sqrt{n}$.

1.2 L'algorithme

Dans cette partie, on s'intéresse au problème du Lasso suivant. Déterminer pour tout $\lambda > 0$,

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

Un tel problème peut a priori admettre plusieurs solutions.

Lemme 1 (Unicité)**.** *Si X^TX est définie positive, (1) admet une unique solution.*

Démonstration. Soient Y dans $\mathbb{R}^n$ et X dans $\mathbb{R}^{n\times p}$ tel que X^TX est définie positive. Définissons pour tout $\lambda \geq 0$ la fonction f_λ sur $\mathbb{R}^p$ par

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

Soit $\lambda \geq 0$. Alors f_λ est strictement convexe. En effet, $f_\lambda = g_\lambda + h_\lambda$ où pour tout $\beta \in \mathbb{R}^p$,

$$\begin{aligned} g_\lambda(\beta) &= \frac{1}{2}||Y - X\beta||_2^2 \\ h_\lambda(\beta) &= \lambda||\beta||_1 \end{aligned}$$

g_λ et h_λ sont des fonctions convexes de β. De plus, g_λ est strictement convexe car sa matrice Hessienne est justement $X^T X$ qui est définie positive. Finalement, f_λ admet un unique minimum sur $\mathbb{R}^p$ pour tout $\lambda \geq 0$. □

Nous nous placerons à partir de maintenant dans le cas particulier où $X^T X$ est définie positive ce qui permettra d'affirmer que la solution au problème du Lasso est unique. Une version modifiée de l'algorithme LARS (cf. [2] et [3]) permet alors d'en calculer la solution. Celui-ci repose sur le Lemme de « la condition d'optimalité ».

Lemme 2 (Condition d'optimalité)**.** *β est solution de (1) si et seulement si :*

$$\forall j = 1...p, \left\{ \begin{array}{rcll} |X_j^T(Y - X\beta)| & \leq & \lambda & \textit{si } \beta_j = 0 \quad \textit{(L1)} \\ X_j^T(Y - X\beta) & = & \lambda \operatorname{sign}(\beta_j) & \textit{si } \beta_j \neq 0 \quad \textit{(L2)} \end{array} \right.$$

Où X_j désigne la j-ième colonne de X.

Démonstration. On reprend les notations de la preuve du lemme précédent. Comme f_λ est convexe, β est une solution de (1), si et seulement si la dérivée directionnelle $D_u f(\beta)$ est positive dans tout les directions $u \in \mathbb{R}^n$. Cette condition nécessaire et suffisante peut s'écrire :

$$\forall u \in \mathbb{R}^n, \left(-u^T X^T (Y - X\beta) + \lambda \sum_{j=1}^{p} \left\{ \begin{array}{ll} |u_j| & \text{si } \beta_j = 0 \\ u_j \operatorname{sign}(\beta_j) & \text{si } \beta_j \neq 0 \end{array} \right. \right) \geq 0$$

On remarque que cette écriture est vérifiée si et seulement si $D_{e_j} f(\beta)$ et $D_{-e_j} f(\beta)$ sont positifs pour tout $j = 1, ..., p$, où les e_j sont les vecteurs unités. L'insertion des e_j dans l'écriture précédente, nous mène directement au Lemme de la condition d'Optimalité du Lasso. □

Nous allons essayer de chercher une solution potentielle au problème du Lasso à partir de ce lemme. Pour $s \in \{-1, 0, 1\}^p$, on pose Γ_s l'ensemble d'indices j tels que $s_j \neq 0$, X_{Γ_s} la matrice composée des colonnes de X correspondant à l'ensemble Γ_s, ainsi que s_Γ le vecteur s auquel on a retiré les coordonnées nulles. On peut alors définir pour tout $\lambda > 0$,

$$\left\{ \begin{array}{rcll} \beta_i^s(\lambda) & = & 0 & \text{si } s_i = 0 \\ \beta_{\Gamma_s}^s(\lambda) & = & (X_{\Gamma_s}^T X_{\Gamma_s})^{-1}(X_{\Gamma_s}^T Y - \lambda s_\Gamma) & \text{sinon} \end{array} \right.$$

Lemme 3. *Si $s = \operatorname{sign}(\beta^s(\lambda))$ alors $\beta^s(\lambda)$ vérifie (L2) pour tout j tel que $s_j \neq 0$.*

Démonstration. À partir de la définition de $\beta^s(\lambda)$, on a

$$\begin{array}{rrcl} & \beta^s_{\Gamma_s}(\lambda) & = & (X^T_{\Gamma_s} X_{\Gamma_s})^{-1}(X^T_{\Gamma_s} Y - \lambda s_\Gamma) \\ \Rightarrow & X^T_\Gamma X_\Gamma \beta_\Gamma & = & X^T_\Gamma Y - \lambda \operatorname{sign}(\beta_\Gamma) \\ \Rightarrow & X^T_\Gamma Y - X^T_\Gamma X_\Gamma \beta_\Gamma & = & \lambda \operatorname{sign}(\beta_\Gamma) \\ \Rightarrow & X^T_\Gamma (Y - X_\Gamma \beta_\Gamma) & = & \lambda \operatorname{sign}(\beta_\Gamma). \end{array}$$

□

Nous connaissons donc une solution potentielle au problème du Lasso en supposant son signe s connu. On peut remarquer que cette solution est linéaire. L'idée de l'algorithme de LARS repose dessus. On fait en sorte qu'on ait toujours $s = \operatorname{sign}(\beta^s(\lambda))$ et que la fonction $\beta^s(\lambda)$ vérifie la condition (L1), on obtient ainsi une fonction linéaire par morceaux β_{LARS}. Notons $(\lambda_i)_{i\geq 0}$ la famille décroissante avec $\lambda_0 = \infty$, telle que β_{LARS} est linéaire exactement (on ne peut pas les agrandir) sur les intervalles $[\lambda_{i+1}, \lambda_i]$. Pour tout $i \geq 0$, $\beta_{LARS} = \beta^{s_i}$. La difficulté consiste donc à déterminer la famille (λ_i) et les vecteurs de signes s_i associés.

Soit $i \geq 1$. Supposons connu λ_i et s_{i-1}. Donnons la méthode permettant de calculer $s_i \in \{-1, 0, 1\}^p$. Admettons que pour $\lambda < \lambda_i$, la fonction $\beta^{s_{i-1}}(\lambda)$ ne convient plus. Il existe donc deux possibilités.

- Soit $\operatorname{sign}(\beta^{s_{i-1}}(\lambda)) \neq s_{i-1}$. Cela signifie qu'une des coordonnée disons la j-ème de $\beta^{s_{i-1}}(\lambda)$ change de signe, donc s'annule en $\lambda = \lambda_i$. On pose donc $\varepsilon_i = 0$.
- Soit (L1) n'est plus vérifiée pour un j tel que $s_j = 0$. On pose $\varepsilon_i = \operatorname{sign}(X^T_{\Gamma_{s_{i-1}}}(Y - X\beta^{s_{i-1}}(\lambda_i)))$.

On définit alors $s_i = s_{i-1}$ où l'on remplace la j-ième coordonnée par ε_i.

Proposons maintenant l'idée de l'algorithme de LARS pour calculer une fonction β_{LARS} sur $\overline{\mathbb{R}}_+$ qui serait solution du problème du Lasso. Pour $\lambda = \infty$, la solution du problème du Lasso étant trivialement le vecteur nul, on commence par là puis on fait décroître λ au cours de l'algorithme.

- *Initialisation :* on pose $\lambda_0 = \infty$, $s_0 = 0_{\mathbb{R}^p}$ et $\beta_{LARS}(\lambda_0) = \beta^{s_0}(\lambda_0) = 0_{\mathbb{R}^p}$.
- *Étape 1 :* on fait décroître λ en partant de λ_0 jusqu'à ce qu'une coordonnée j de $\beta_{LARS}(\lambda) = \beta^{s_0}(\lambda)$ ne vérifie plus (L1). On définit λ_1 comme cet instant. La coordonnée j de $\beta_{LARS}(\lambda)$ n'est donc plus nulle pour $\lambda < \lambda_1$, on calcule son signe $\varepsilon_1 \in \{-1, 1\}$ comme vu précédemment et on pose $s_1 = s_0$ dans lequel on a remplacé la j-ème coordonnée par ε_1.
- *Étape $i \geq 2$:* on fait décroître λ à partir de λ_i, en calculant $\beta_{LARS}(\lambda) = \beta^{s_i}(\lambda)$ et on s'arrête soit quand une coordonnée j telle que $(s_i)_j \neq 0$ s'annule, car alors $\operatorname{sign}(\beta^{s_i}) \neq s_i$ (en λ_4 sur la figure 2), soit quand la condition (L1) n'est plus satisfaite pour une coordonnée telle que $(s_i)_j = 0$ (en λ_2 sur la figure 2). On note λ_{i+1} cet instant et on définit le nouveau vecteur de signes s_{i+1}.

On calcule ainsi la suite finie $(\lambda_i)_{i\geq 0}$ jusqu'à ce que λ soit nul. De cette manière, nous avons calculé une fonction β_{LARS} qui devrait être solution du problème du Lasso.

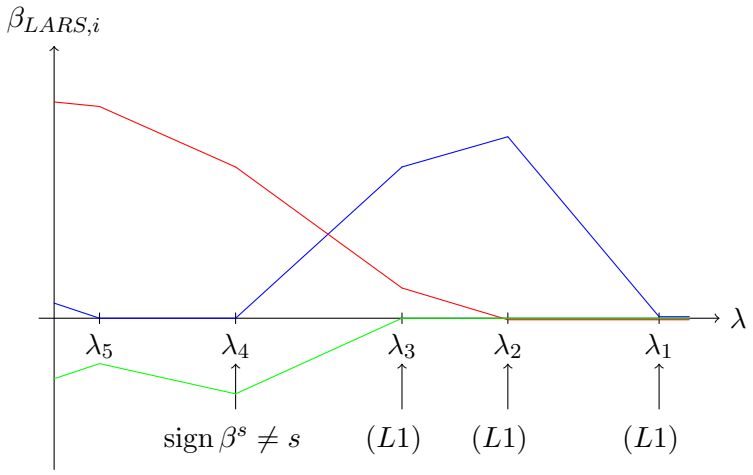

FIGURE 2 – Calcul de la fonction β_{LARS}.

Montrons maintenant que β_{LARS} est en effet solution du problème. Commençons par donner la méthode de calcul des λ_i. Définissons par récurrence la suite (λ_i) de la manière suivante. On pose $\lambda_0 = \infty$. Soit $i \geq 0$, supposons connus λ_i et s_i. Alors pour tout $j = 0, ..., p$ on pose,

$$\lambda_{\max}(j) = \begin{cases} \max_{a\in\{+1,-1\}} \{\lambda_a(j) \mid 0 < \lambda_a(j) < \lambda_i\} & \text{si} \quad s_i(j) = 0 \\ \left(\frac{((X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1}(X_{\Gamma_{s_i}}^T Y))_{j'}}{((X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1} s_{i,\Gamma})_{j'}} \right) & \text{si} \quad s_i(j) \neq 0 \end{cases}$$

où $\lambda_a(j) = \frac{X_j^T [X_{\Gamma_{s_i}} (X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1} (X_{\Gamma_{s_i}}^T Y) - Y]}{[X_j^T X_{\Gamma_{s_i}} (X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1} s_{i,\Gamma}] + a}$, $\max \emptyset = 0$ et j' est l'indice de la j-ème variable dans Γ_{s_i}. On définit alors,

$$\lambda_{i+1} = \max_{j\in\{1,..,p\}} \lambda_{\max}(j).$$

Nous avons donc défini deux suites $(\lambda_i)_{i\geq 0}$ et $(s_i)_{i\geq 0}$ par récurrence simultanée sur i. Montrons maintenant qu'elles conviennent.

Lemme 4 (Validité de β_{LARS})**.** *Tant que $\lambda \in]\lambda_{i+1}, \lambda_i[$, alors (L1) est vérifiée pour tout j tel que $(s_i)_j = 0$ et $\text{sign}(\beta^{s_i}(\lambda)) = s_i$. De plus, la suite (s_i) est bien définie.*

Démonstration. On raisonne par récurrence sur $i \geq 0$. En $i = 0$, le résultat est immédiat. Soit $i \geq 0$, supposons le résultat vrai au rang i et montrons le au rang $i + 1$.

Soit $j \in \{1, .., p\}$ tel que $(s_i)_j \neq 0$. Montrons que

$$\lambda_{max}(j) = \sup \left\{ 0 \leq \lambda < \lambda_i \mid \text{sign}\, \beta_j^{s_i}(\lambda) \neq (s_i)_j \right\}.$$

Juste après λ_i, $\text{sign}\,\beta_j^{s_i}(\lambda_i) = (s_i)_j$ d'après la définition de s_i. Pour que $\beta_j^{s_i}(t)$ change de signe en λ, il faut qu'il s'annule en λ par continuité de $\beta_j^{s_i}$. Cela équivaut à $(X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1}(X_{\Gamma_{s_i}}^T Y - \lambda(s_i)_j)_{j'} = 0$ La résolution donne,

$$\lambda = \frac{((X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1}(X_{\Gamma_{s_i}}^T Y))_{j'}}{((X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1}(\text{sign}(\beta_{\Gamma_{s_i}}(\lambda))))_{j'}}$$

D'où le résultat.

Soit $j \in \{1, .., p\}$ tel que $(s_i)_j = 0$. Nous avons par définition de β^{s_i} immédiatement que $\text{sign}(\beta_j^{s_i}(\lambda)) = (s_i)_j$ pour tout $\lambda \geq 0$. Montrons que

$$\lambda_{max}(j) = \inf \left\{ 0 \leq \lambda \leq \lambda_i \mid \beta_j^{s_i}(\lambda) \text{ vérifie (L1)} \right\}.$$

Juste après λ_i, $\beta_j^{s_i}(\lambda)$ vérifie (L1) d'après la définition de s_i. Pour que $\beta_j^{s_i}(t)$ ne vérifie plus (L1) à partir d'un certain λ, il faut donc par continuité que $|X_j^T(Y - X\beta^{s_i}(\lambda))| = \lambda$. La résolution de cette équation en λ nous donne,

$$\lambda \in \left\{ \frac{X_j^T[X_{\Gamma_{s_i}}(X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1}(X_{\Gamma_{s_i}}^T Y) - Y]}{[X_j^T X_{\Gamma_{s_i}}(X_{\Gamma_{s_i}}^T X_{\Gamma_{s_i}})^{-1}((s_i)_j)] + a} \right\}$$

où $a \in \{+1, -1\}$.

Finalement, par définition de λ_{i+1} le résultat est vrai au rang $i+1$. De plus,

$$\lambda_{i+1} = \inf \left\{ 0 \leq \lambda \leq \lambda_i \;\middle|\; \begin{array}{c} \text{sign}(\beta^{s_i}(\lambda)) = s_i \\ \text{(L1) est vérifiée pour } j \text{ tel que } (s_i)_j = 0 \end{array} \right\}$$

s_{i+1} est donc correctement défini. □

On déduit de ce lemme et du précédent que la fonction β_{LARS} calculée par l'algorithme de LARS-Lasso est bien la solution du problème du Lasso. On peut cependant faire deux remarques. La première est que l'algorithme termine. C'est à dire qu'à partir d'un certain rang $\lambda_i = 0$. Cela provient du fait qu'il n'existe que 3^p fonctions β^s différentes et chacune d'entre elles ne vérifie les conditions du lemme précédent que sur un intervalle. L'algorithme termine donc en au plus 3^p itérations. Remarquons que 3^p peut-être gigantesque dès que p est grand, en pratique l'algorithme termine bien plus rapidement. Ce résultat est donc purement théorique. La deuxième est que dans la définition de s_i, nous avons supposé qu'une seule des coordonnées était insatisfaisante en λ_i. Cependant cela arrive en général avec probabilité 1, nous restreignons donc ici notre étude à ce cas.

Nous pouvons maintenant énoncer l'algorithme LARS-Lasso, qui calcule la fonction β_{LARS} décrite ci-dessus.

Nous pouvons éviter les redondances de certain calculs : $(X_\Gamma^T X_\Gamma)^{-1}$, $X_\Gamma^T Y$, unifier les calculs des vecteurs de $((X_\Gamma^T X_\Gamma)^{-1}(X_\Gamma^T Y))_{j'}$ et de $((X_\Gamma^T X_\Gamma)^{-1}(\text{sign}(\beta_\Gamma(\lambda))))_{j'}$. En pratique, nous implémenterons donc l'algorithme 2.

Algorithme 1 LARS-Lasso

Entrées: X, Y, n, p

$i \longleftarrow 1$
$\lambda \longleftarrow +\infty$
$\beta \longleftarrow 0$
$\text{sign}(\beta) \longleftarrow 0$
$\Gamma \longleftarrow \emptyset$
Resultats $\longleftarrow \emptyset$
Tant que $\lambda > 0$ **faire**
« On commence par chercher la première variable insatisfaisante »
$\lambda_{\max} \longleftarrow 0$
$j_{\max} \longleftarrow 0$
Pour $j = 1, ..., p$ **faire**
Si $\text{sign}(\beta)_j = 0$ **alors**
$\lambda_{\text{test}} \longleftarrow \max \left\{ \frac{X_j^T [X_\Gamma (X_\Gamma^T X_\Gamma)^{-1} (X_\Gamma^T Y) - Y]}{[X_j^T X_\Gamma (X_\Gamma^T X_\Gamma)^{-1} (\text{sign}(\beta_\Gamma(\lambda)))] \pm 1} \right\}$
Sinon
$\lambda_{\text{test}} \longleftarrow \left(\frac{((X_\Gamma^T X_\Gamma)^{-1} (X_\Gamma^T Y))_{j'}}{((X_\Gamma^T X_\Gamma)^{-1} (\text{sign}(\beta_\Gamma(\lambda))))_{j'}} \right)$
Fin si
Si $\lambda_{\text{test}} < \lambda$ **et** $\lambda_{\text{test}} > \lambda_{\max}$ **alors**
$\lambda_{\max} \longleftarrow \lambda_{\text{test}}$
$j_{\max} \longleftarrow j$
Fin si
Fin pour
« On a dans $(\lambda_{\max}, j_{\max})$ le résultat d'insatisfaction cherché »
$\beta \longleftarrow \beta_\Gamma \longleftarrow (X_\Gamma^T X_\Gamma)^{-1} (X_\Gamma^T Y - \lambda_{\max} \text{sign}(\beta_\Gamma))$
$\lambda \longleftarrow \lambda_{\max}$
Si $j_{\max} > 0$ **alors**
Si $\text{sign}(\beta)_{j_{\max}} = 0$ **alors**
$\text{sign}(\beta)_{j_{\max}} \longleftarrow \text{sign}(X_{j_{\max}}^T (Y - X\beta))$
Sinon
$\text{sign}(\beta)_{j_{\max}} \longleftarrow 0$
$\beta_{j_{\max}} \longleftarrow 0$
Fin si
$\Gamma \longleftarrow$ non-nuls$(\text{sign}(\beta))$
Resultats $\longleftarrow$ Resultats $\cup \{(j_{\max}, \lambda_{\max}, \beta)\}$
Sinon
Resultats $\longleftarrow$ Resultats $\cup \{(\lambda_{\max}, \beta)\}$
Fin si
$i \longleftarrow i + 1$
Fin tant que
Retourner Resultats

Algorithme 2 LARS-Lasso Amélioré

Entrées: X, Y, n, p

$\lambda \longleftarrow +\infty$
$\beta \longleftarrow 0$
$\text{sign}(\beta) \longleftarrow 0$
$\Gamma \longleftarrow \emptyset$
Resultats $\longleftarrow \emptyset$
Tant que $\lambda > 0$ **faire**
« On commence par chercher la première variable insatisfaisante ».
$\lambda_{\max} \longleftarrow 0$
$j_{\max} \longleftarrow 0$
$U \longleftarrow (X_\Gamma^T X_\Gamma)^{-1}$
$V \longleftarrow X_\Gamma^T Y$
$W \longleftarrow X_\Gamma U$
$N1 \longleftarrow UV$
$D1 \longleftarrow U \,\text{sign}(\beta_\Gamma)$
$F1 \longleftarrow F1_\Gamma \longleftarrow N1./D1$
Pour $j = 1, ..., p$ **faire**
Si $\text{sign}(\beta)_j = 0$ **alors**
$N0 \longleftarrow X_j^T(WV - Y)$
$D0 \longleftarrow X_j^T W \,\text{sign}(\beta_\Gamma)$
$\lambda_{\text{test}} \longleftarrow \max\left\{\frac{N0}{D0+1}, \frac{N0}{D0-1}\right\}$
Sinon
$\lambda_{\text{test}} \longleftarrow F1_j$
Fin si
Si $\lambda_{\text{test}} < \lambda$ **et** $\lambda_{\text{test}} > \lambda_{\max}$ **alors**
$\lambda_{\max} \longleftarrow \lambda_{\text{test}}$
$j_{\max} \longleftarrow j$
Fin si
Fin pour
« On a dans $(\lambda_{\max}, j_{\max})$ le résultat d'insatisfaction cherché ».
$\beta \longleftarrow \beta_\Gamma \longleftarrow U(V - \lambda_{\max} \,\text{sign}(\beta_\Gamma))$
$\lambda \longleftarrow \lambda_{\max}$
Si $j_{\max} > 0$ **alors**
Si $\text{sign}(\beta)_{j_{\max}} = 0$ **alors**
$\text{sign}(\beta)_{j_{\max}} \longleftarrow \text{sign}(X_{j_{\max}}^T(Y - X\beta))$
Sinon
$\text{sign}(\beta)_{j_{\max}} \longleftarrow 0$
$\beta_{j_{\max}} \longleftarrow 0$
Fin si
$\Gamma \longleftarrow$ non-nuls$(\text{sign}(\beta))$
Resultats $\longleftarrow$ Resultats $\cup \{(j_{\max}, \lambda_{\max}, \beta)\}$
Sinon
Resultats $\longleftarrow$ Resultats $\cup \{(\lambda_{\max}, \beta)\}$
Fin si
Fin tant que
Retourner Resultats

2 Quand le Lasso fonctionne-t-il ?

Les résultats de cette partie provienne d'un article de Zhao et Yu [1].

2.1 Notion de consistance et condition d'irreprésentabilité

Pour des raisons techniques, mais aussi car nous gagnons en précision, nous voulons non seulement que le Lasso sélectionne les vraies variables explicatives mais en plus qu'il en donne les bons signes. Pour simplifier, nous adoptons la notation suivante.

Définition 1. Une estimation $\widehat{\beta}^n$ est *égale en signes* avec le vrai β^n, ce que l'on note $\widehat{\beta}^n =_s \beta^n$, si et seulement si :

$$\text{sign}(\widehat{\beta}^n) = \text{sign}(\beta^n)\,.$$

Définissons maintenant deux consistances en signes de l'estimateur Lasso.

Définition 2 $(CS+)$**.** Le Lasso est *fortement consistant en signes* s'il existe une suite $(\lambda_n)_{n\geq 0}$ de réels positifs indépendante de Y_n et de X_n telle que,

$$\lim_{n\to\infty} P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) = 1\,.$$

Définition 3 $(CS-)$**.** Le Lasso est *généralement consistant en signes* si

$$\lim_{n\to\infty} P(\exists \lambda \geq 0, \widehat{\beta}^n(\lambda) =_s \beta^n) = 1\,.$$

La consistance forte implique que l'on peut utiliser une suite prédéfinie de $(\lambda_n)_{n\geq 0}$ pour obtenir le bon modèle à partir de la méthode du Lasso. La consistance générale quand à elle signifie que pour une réalisation X^n et Y^n, lorsque l'on a fait suffisament d'observations on peut sélectionner le bon modèle mais avec un λ_n qui dépend de X_n et Y_n. Il est immédiat que la consistance forte implique la consistance générale. Nous verrons que de manière surprenante, la réciproque est plus ou moins vraie sous certaines hypothèses.

Il s'agit maintenant de savoir quand le Lasso est consistant en signes, c'est à dire quand fonctionne-t-il ? Nous allons donc donner deux conditions qui porteront sur les variables candidates $(X_{.,i})_{1\leq i\leq p}$ qui assureront les consistances précédentes. De cette façon, lors de l'étude d'un modèle, nous regarderons les variables potentielles $(X_{.,i})_{1\leq i\leq p}$ et nous vérifierons si elles vérifient ces conditions pour savoir si il est pertinent d'utiliser la méthode du Lasso.

Commençons par introduire quelques notations. Sans perte de généralité, nous pouvons supposer que le vrai $\beta^n \in \mathbb{R}^p$ est de la forme $(\beta_1^n, ..., \beta_q^n, 0, ..., 0)$ où $\forall j = 1, ..., q\ \beta_j^n \neq 0$. Les vraies variables explicatives sont donc $(X_i)_{1\leq i\leq q}$. On pose alors :

- $X^n(1) = (X^n_{.,1}, ..., X^n_{.,q}) \in M_{n,q}$ le vecteur des observations des vraies variables explicatives.

- $X^n(2) = (X^n_{.,q+1}, ..., X^n_{.,p}) \in M_{n,p-q}$ le vecteur des observations des variables non explicatives.
- $C^n = \frac{1}{n} X^{nT} X^n$, puis pour $i, j \in \{1, 2\}$ $C^n_{ij} = \frac{1}{n} X^n(i)^T X^n(j)$.

On a donc en représentation par blocs :

$$\mathcal{C}^n = \begin{pmatrix} C^n_{11} & C^n_{12} \\ C^n_{21} & C^n_{22} \end{pmatrix}$$

Il semble naturel de se placer sous l'hypothèse d'identifiabilité du modèle

$$\exists! \beta \text{ tel que } Y^n = X^n \beta + \epsilon^n \, .$$

Cela revient à supposer C^n_{11} inversible. Nous pouvons maintenant donner les deux définitions suivantes.

Définition 4 $(CI+)$**.** On dit que le Lasso vérifie la *condition d'irreprésentabilité forte* si il existe $\eta \in {\mathbb{R}^*_+}^{p-q}$ fixé tel que à partir d'un certain rang,

$$\left| C^n_{21} (C^n_{11})^{-1} \operatorname{sign}(\beta^n_{(1)}) \right| \leq \mathbf{1}_{\mathbb{R}^{p-q}} - \eta$$

où $\mathbf{1}_{\mathbb{R}^{p-q}} \in \mathbb{R}^{p-q}$ et on note pour $x, y \in \mathbb{R}^{p-q}$ $x < y$ si pour tout $i \in \{1, .., p-q\}$ $x_i < y_i$.

Définition 5 $(CI-)$**.** On dit que le Lasso vérifie la *condition d'irreprésentabilité faible* si à partir d'un certain rang,

$$\left| C^n_{21} (C^n_{11})^{-1} \operatorname{sign}(\beta^n_{(1)}) \right| < \mathbf{1}_{\mathbb{R}^{p-q}} \, .$$

Ces conditions semblent au premier abord peu intuitives, mais ce n'est pas le cas. Regardons en effet le cas simple où $p = 2$ et $q = 1$. Notons $\operatorname{Var}(x) = \frac{1}{n} \sum_{i=1}^n x_i^2$ et $\operatorname{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^n x_i y_i$ respectivement la variance empirique de x et la covariance empirique de x et y au cours des différentes observations. Alors on a, $C^n_{11} = \operatorname{Var}(X^n_1)$, $C^n_{22} = \operatorname{Var}(X^n_2)$ et $C^n_{12} = \operatorname{Cov}(X^n_1, X^n_2)$. La condition de faible irreprésentabilité se reformule donc, $|\operatorname{Cov}(X^n_1, X^n_2)| \leq |\operatorname{Var}(X^n_1)|$. Le Lasso pourra donc détecter que X^n_1 est la bonne variable explicative, si elle n'est pas trop liée avec X^n_2 au point que l'on confond ses variations avec celles de X^n_2. Cela semble assez cohérent. Dans le cas où le nombre de variables candidates est plus grand, l'idée est que les variables significatives doivent suffisament varier et ne pas être trop liées au mauvaises variables, faute de quoi nous les confondrions.

On peut encore une fois remarquer que la condition de faible irreprésentabilité est effectivement plus faible que la condition de forte irreprésentabilité qui permet à $\left| C^n_{21} (C^n_{11})^{-1} \operatorname{sign}(\beta^n_{(1)}) \right|$ de se rapprocher de 1 aussi près que possible lorque n tend vers ∞. Enfin, le η de la condition forte détermine la vitesse avec laquelle le Lasso trouve le bon modèle lorsqu'on augmente le nombre d'observations n. Plus η est grand, et plus il est aisé de distinguer les vraies variables des fausses.

Notre principal résultat consiste à prouver la quasi-équivalence entre ces deux conditions et les notions de consistances en signes précédentes dans le sens suivant. Sous certaines hypothèses supplémentaires,

$$CI+ \Rightarrow CS+ \Rightarrow CS- \Rightarrow CI- \, .$$

Pour démontrer ce résultat, nous allons procéder en deux étapes. Premièrement, nous allons étudier le cas plus simple où p et q sont fixes et indépendants de n, puis nous étudierons le cas où p et q sont des fonctions de n et sont a priori grands. Commençons par établir un résultat intermédiaire, qui nous sera utile dans les deux cas. Il consiste à minorer la probabilité qu'a le Lasso de choisir le bon modèle.

Proposition 1. *Si le Lasso vérifie la condition d'irreprésentabilité forte avec une constante $\eta > 0$ et si on pose*

$$A_n = \left\{ \left|(C_{11}^n)^{-1} W^n(1)\right| < \sqrt{n}\left(\left|\beta_{(1)}^n\right| - \tfrac{\lambda_n}{n}\left|(C_{11}^n)^{-1}\operatorname{sign}(\beta_{(1)}^n)\right|\right)\right\}$$
$$B_n = \left\{ \left|C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)\right| \leq \tfrac{\lambda_n}{\sqrt{n}}\eta \right\}$$

où $W^n(1) = \frac{1}{\sqrt{n}} X_n(1)^T \varepsilon_n$ et $W^n(2) = \frac{1}{\sqrt{n}} X_n(2)^T \varepsilon_n$ alors,

$$P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) \geq P(A_n \cap B_n)\,.$$

Nous allons voir dans la preuve que A_n implique que les signes de $\beta_{(1)}^n$ sont correctement estimés et que $A_n \cap B_n$ implique que les coefficients de $\beta_{(2)}^n$ sont diminués jusqu'à zéro. La taille des deux événements dépend du paramètre λ_n. Si λ_n est faible, A_n est grand et les bonnes variables ont tendance à avoir les bons signes mais B_n est petit et les variables non explicatives ne sont pas éliminées du modèle. Plus on choisit un λ_n faible et plus le Lasso retient des mauvaises variables. Ceci correspond bien au fait que lorsque $\lambda_n = 0$, on fait une simple régression linéaire et lorsque $\lambda_n = \infty$, tous les coefficients de $\widehat{\beta}^n$ sont nuls. Il s'agira donc de choisir un λ_n ni trop grand, ni trop petit, on verra qu'il devra être de l'ordre de $\sqrt{n}$. D'un autre côté, plus η est grand et plus B_n l'est aussi sans avoir d'impact sur A_n. Tout ceci correspond à l'intuition précédente, qui dit que si la condition d'irreprésentabilité est vérifiée avec un grand η, il sera facile pour le Lasso de choisir le bon modèle et de distinguer les variables réellement explicatives des autres.

Démonstration. Par définition $\widehat{\beta}^n = \arg\min_\beta(\frac{1}{2}\sum_{i=1}^n (Y_i - X_i\beta)^2 + \lambda_n \|\beta\|_1)$. Posons $\widehat{u}^n = \widehat{\beta}^n - \beta^n$. Nous avons alors,

$$\begin{aligned}
\widehat{u}^n &= \underset{u^n\in\mathbb{R}^p}{\arg\min}\left(\frac{1}{2}\sum_{i=1}^n (\varepsilon_i - X_i u^n)^2 + \lambda_n \|u^n + \beta^n\|_1\right) \\
&= \underset{u^n\in\mathbb{R}^p}{\arg\min}\left(\frac{1}{2}\sum_{i=1}^n \left[(\varepsilon_i - X_i u^n)^2 - \varepsilon_i^2\right] + \lambda_n \|u^n + \beta^n\|_1\right).
\end{aligned}$$

Or,

$$\begin{aligned}
\sum_{i=1}^n \left[(\varepsilon_i - X_i u^n)^2 - \varepsilon_i^2\right] &= \sum_{i=1}^n \left[-2\varepsilon_i X_i u^n + (u^n)^T X_i^T X_i u^n\right] \\
&= -2W^n(\sqrt{n}u^n) + (\sqrt{n}u^n)^T C^n(\sqrt{n}u^n)
\end{aligned}$$

où $W^n = \frac{1}{\sqrt{n}}(X^n)^T\varepsilon^n$.

On peut remarquer que cette somme est différentiable par rapport à u^n. On a alors,

$$\begin{aligned}\frac{d\left[\sum_{i=1}^n (\varepsilon_i - X_i u^n)^2\right]}{du^n} &= \frac{d\left[-2W^n(\sqrt{n}u^n) + (\sqrt{n}u^n)^T C^n(\sqrt{n}u^n)\right]}{du^n} \\ &= 2\sqrt{n}(C^n(\sqrt{n}u^n) - W^n). \qquad (2)\end{aligned}$$

Notons alors $\widehat{u}^n(1)$,$W^n(1)$ et $\widehat{u}^n(2)$,$W^n(2)$ respectivement les q premières et $p-q$ dernières entrées de $\widehat{u}^n$ et W^n. Si il existe $\widehat{u}^n$ tel que $\widehat{u}^n(2) = \mathbf{0}_{\mathbb{R}^{p-q}}$ et :

$$C_{11}^n(\sqrt{n}\widehat{u}^n(1)) - W^n(1) = -\frac{\lambda_n}{\sqrt{n}}\operatorname{sign}(\beta_{(1)}^n)$$

$$|\widehat{u}^n(1)| < \left|\beta_{(1)}^n\right|$$

$$-\frac{\lambda_n}{\sqrt{n}}\mathbf{1}_{\mathbb{R}^{p-q}} \leq C_{21}^n(\sqrt{n}\widehat{u}^n(1)) - W^n(2) \leq \frac{\lambda_n}{\sqrt{n}}\mathbf{1}_{\mathbb{R}^{p-q}}\,.$$

Alors par le lemme 2 sur la condition d'optimalité, et l'équation (2), $\widehat{\beta}^n$ est l'estimateur Lasso et a les bons signes. En substituant $\widehat{u}^n(1)$, $\widehat{u}^n(2)$ et en faisant une inégalité triangulaire, l'existence d'un tel $\widehat{u}^n$ est induite par :

$$\left|(C_{11}^n)^{-1}W^n(1)\right| < \sqrt{n}\left(\left|\beta_{(1)}^n\right| - \frac{\lambda_n}{n}\left|(C_{11}^n)^{-1}\operatorname{sign}(\beta_{(1)}^n)\right|\right)$$

$$\left|C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2)\right| \leq \frac{\lambda_n}{\sqrt{n}}\left(\mathbf{1}_{\mathbb{R}^{p-q}} - \left|C_{21}^n(C_{11}^n)^{-1}\operatorname{sign}(\beta_{(1)}^n\right|\right)$$

donc par $A_n \cap B_n$. Par unicité de l'estimateur Lasso, $A_n \cap B_n$ implique donc que le Lasso choisis les bons signes pour les différentes variables candidates. Finalement, $P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) \geq P(A_n \cap B_n)$. □

2.2 Le cas de petits p et q

Commençons par traiter le cas plus facile où p, q et $\beta^n \in \mathbb{R}^q$ sont fixes quand n tend vers l'infini. Dans cette partie, nous supposons vérifiées les conditions de régularité suivantes :

$$C^n \underset{n\to\infty}{\longrightarrow} C \qquad (3)$$

où $C \in M_p(\mathbb{R})$ est une matrice définie positive, et

$$\frac{1}{n}\max_{1\leq i\leq n}((X_{i,.}^n)^T X_{i,.}^n) \underset{n\to\infty}{\longrightarrow} 0. \qquad (4)$$

Les convergences ci-dessus sont ici déterministes mais les résultats que nous allons établir peuvent s'étendre au cas de motifs aléatoires. Si la première condition de convergence est naturelle, la deuxième est

moins intuitive au premier regard. Nous verrons en détail leur légitimité dans les preuves des résultats. L'idée sous-jacente à la condition (4) est qu'il faut éviter que certaines observations exceptionnelles écrasent toutes les autres. Nous avons maintenant tous les outils à notre disposition pour établir les résultats tant attendus. Commençons par le théorème 1 qui donne : $CI+ \Rightarrow CS+$.

Théorème 1 ($CI+ \Rightarrow CS+$)**.** *Si p, q et $\beta^n = \beta$ sont fixes et les conditions de régularité (3) et (4) vérifiées. Si le Lasso vérifie la condition de forte irreprésentabilité, alors*

- *le Lasso est fortement consistant en signes.*
- *pour tout λ_n tel que $\lambda_n/n \to 0$ et $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ avec $0 \leq c < 1$, $P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c})$.*

Démonstration. D'après la proposition 1 nous avons,

$$P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) \geq P(A_n \cap B_n)\,.$$

Nous allons voir que $P(A_n \cap B_n)$ tend vers 1 exponentiellement vite. Pour cela considérons leurs complémentaires. Par définition de A_n et B_n, si l'on pose $z^n = (C^n_{11})^{-1} W^n(1) \in \mathbb{R}^q$, $\zeta^n = C^n_{21}(C^n_{11})^{-1} W^n(1) - W^n(2) \in \mathbb{R}^{p-q}$ et $b^n = (C^n_{11})^{-1} \operatorname{sign}(\beta^n_{(1)}) \in \mathbb{R}^q$, on a

$$\begin{aligned} P(A_n^c) &\leq \sum_{i=1}^{q} P\left(|z_i^n| \geq \sqrt{n}\left(|\beta_i^n| - \frac{\lambda_n}{n} b_i^n\right)\right) \\ P(B_n^c) &\leq \sum_{i=1}^{p-q} P\left(|\zeta_i^n| \geq \frac{\lambda_n}{\sqrt{n}} \eta_i\right). \end{aligned}$$

Il nous reste donc à montrer que les termes de droite tendent de manière exponentielle vers 0. Pour cela observons les termes z^n et ζ^n sont asymptotiquement des vecteurs Gaussiens, ce qui est un résultat classique montré par exemple dans [6]. Donnons en ici simplement l'intuition.

Commençons par $z^n = (C^n_{11})^{-1} W^n(1)$. D'après la condition de régulatité (3), $C^n \underset{n\to\infty}{\longrightarrow} C$, il suffit donc de montrer que $W^n(1)$ est asymptotiquement un vecteur Gaussien. Or,

$$\begin{aligned} W^n(1) &= (X^n(1))^T \varepsilon^n / \sqrt{n} \\ &= \sum_{i=1}^{n} (X^n_{i,j})_{j=1..q} \varepsilon_i^n / \sqrt{n} \end{aligned}$$

où $X^n_{i,j}$ désigne l'élément (i,j) de la matrice X^n. Pour que cette somme converge comme dans le théoreme central limite, il est important qu'aucun terme de la somme n'ait une contribution significative. Sinon, la valeur de l'un des ε_i^n influencerait directement la valeur de $W^n(1)$, qui ne pourrait pas être Gaussien si ce ε_i^n ne l'est pas également.

La contribution de ε_i^n à la somme peut être mesurée par le carré de la norme L^2 du vecteur par lequel ε_i^n est multiplié dans la somme, c'est-à-dire $\sum_{j=1}^{q} (X^n_{i,j})^2 / n$.

Or, précisement, la condition (4) s'écrit : $\max_i \sum_{j=1}^{p} \left(X_{i,j}^n\right)^2 / n \to 0$. Autrement dit, (4) garantit que dans le pire des cas ($q = p$), aucun des ε_i^n ne contribue trop à $W^n(1)$, qui est donc effectivement une moyenne d'un grand nombre de variables iid, raison pour laquelle il est asymptotiquement Gaussien.

De même, $W^n(2)$ tend vers un vecteur Gaussien et donc ζ^n aussi.

Tous les z_i^n et ζ_i^n convergent donc vers des variables Gaussiennes centrées et de variances finies. Il existe donc une constante $S > 0$ telle que pour tout $i \in \{1, .., q\}$, $j \in \{1, .., p-q\}$ et $n \geq 1$ on a $E\left[\left(z_i^n\right)^2\right] < S^2$ et $E\left[\left(\zeta_i^n\right)^2\right] < S^2$.

Or la fonction de répartition Φ d'une distribution Gaussienne centrée réduite vérifie

$$\forall t \in \mathbb{R} \qquad 1 - \Phi(t) < \frac{e^{-\frac{t^2}{2}}}{t} .$$

De plus comme p, q et β^n sont fixes et $\lambda_n / n \to 0$,

$$t_i^n := \sqrt{n}\left(|\beta_i^n| - \frac{\lambda_n}{n} b_i^n\right) = (1 + o(1))\sqrt{n}\,|\beta_i| .$$

On obtient donc,

$$\begin{aligned}
P(A_n^c) &\leq \sum_{i=1}^{q} P\left(|z_i^n| \geq t_i^n\right) \\
&\leq \sum_{i=1}^{q} \left[1 - P\left(z_i^n \leq t_i^n\right)\right] \\
&\leq (1 + o(1)) \sum_{i=1}^{q} \left[1 - \Phi\left(\frac{t_i^n}{S}\right)\right] \\
&\leq (1 + o(1)) \sum_{i=1}^{q} \frac{S e^{-\frac{(t_i^n)^2}{2S^2}}}{t_i^n} \\
&= o\left(e^{-n^c}\right).
\end{aligned}$$

La dernière égalité découlant de $\lambda_n / n^{\frac{1+c}{2}} \to \infty$ avec $0 \leq c < 1$. On montre de même

$$\begin{aligned}
P(B_n^c) &\leq \sum_{i=1}^{p-q} P\left(|\zeta_i^n| \geq \frac{\lambda_n}{\sqrt{n}} \eta_i\right) \\
&\leq \sum_{i=1}^{p-q} \left[1 - \Phi\left(\frac{\lambda_n}{S\sqrt{n}} \eta_i\right)\right] \\
&= o\left(e^{-n^c}\right).
\end{aligned}$$

Et finalement,

$$\begin{aligned} P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) &\geq P(A_n \cap B_n) \\ &\geq 1 - P(A_n^c) - P(B_n^c) \\ &= 1 - o\left(e^{-n^c}\right). \end{aligned}$$

Donc le Lasso est fortement consistant en signes avec les égalités désirées. □

La condition d'irreprésentabilité forte assure donc non seulement que le Lasso choisit le bon modèle mais en plus avec les bons signes pour les paramètres. Le théorème 2, donne la réciproque de l'équivalence.

Théorème 2 ($CS- \Rightarrow CI-$)**.** *Si p, q et $\beta^n = \beta$ sont fixes et sous les conditions de régularité (3) et (4), si le Lasso est généralement consistant en signes, alors il existe N tel que le Lasso vérifie la condition de faible irreprésentabilité pour tout $n \geq N$.*

Démonstration. Considérons l'ensemble F^n, sur lequel il existe λ_n tel que,

$$\widehat{\beta}^n(\lambda_n) =_s \beta^n .$$

Comme le Lasso est généralement consistant en signes,

$$P\left(F^n\right) \underset{n\to\infty}{\longrightarrow} 1 .$$

Par définition de F^n, $\widehat{\beta}^n_{(1)} \neq 0$ et $\widehat{\beta}^n_{(2)} = 0$. On peut donc appliquer le lemme 2. En utilisant l'égalité (2) de la proposition 1, on obtient sur tout F^n avec les notations de la preuve de la proposition 1,

$$\begin{aligned} C^n_{11}(\sqrt{n}\widehat{u}^n(1)) - W^n(1) &= -\frac{\lambda_n}{\sqrt{n}} \operatorname{sign}(\widehat{\beta}^n_{(1)}) \\ &= -\frac{\lambda_n}{\sqrt{n}} \operatorname{sign}(\beta^n_{(1)}) \qquad (5) \\ \left|C^n_{21}(\sqrt{n}\widehat{u}^n(1)) - W^n(2)\right| &\leq \frac{\lambda_n}{\sqrt{n}} \mathbf{1}_{\mathbb{R}^{p-q}}. \qquad (6) \end{aligned}$$

On peut alors réécrire (6) en remplaçant $\widehat{u}^n(1)$ par sa valeur donnée en (5). On en déduit,

$$F^n \subseteq G^n := \left\{ \frac{\lambda_n}{\sqrt{n}} L^n \leq C^n_{21}(C^n_{11})^{-1} W^n(1) - W^n(2) \leq \frac{\lambda_n}{\sqrt{n}} U^n \right\}$$

où

$$\begin{aligned} L^n &= -1 + C^n_{21}(C^n_{11})^{-1} \operatorname{sign}(\beta^n_{(1)}) \\ U^n &= 1 + C^n_{21}(C^n_{11})^{-1} \operatorname{sign}(\beta^n_{(1)}). \end{aligned}$$

Nous allons maintenant supposer que le Lasso ne vérifie pas la condition $CI-$ et aboutir à une contradiction. Pour tout $N \geq 1$, il existe toujours $n \geq N$ tel que au moins l'un des éléments du vecteur

$C_{21}^n (C_{11}^n)^{-1} \operatorname{sign}(\beta_{(1)}^n)$ soit supérieur à $\mathbf{1}_{\mathbb{R}^{p-q}}$ en valeur absolue. Sans perte de généralité, comme p est fini et fixe, on peut supposer qu'il s'agit toujours du premier élément qui est supérieur à 1. Alors pour tout $\lambda_n \geq 0$,

$$\frac{\lambda_n}{\sqrt{n}} L_1^n \geq 0\,.$$

Or sous les conditions de régularité (3) et (4), $C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)$ est asymptotiquement un vecteur gaussien centré d'après la preuve du théorème 1. Il existe donc une probabilité non nulle telle que son premier élément soit négatif. Donc $P(G^n)$ ne tend pas vers 1. Il vient,

$$\liminf P(F^n) \leq \liminf P(G^n) < 1\,.$$

D'où la contradiction. □

Nous avons donc une équivalence entre la consistance en signes du Lasso et les conditions d'irreprésentabilité. Avant d'utiliser le Lasso, nous regarderons donc si les variables potentiellement explicatives vérifient les conditions d'irreprésentabilité. Si c'est le cas, le Lasso devrait choisir le bon modèle. D'un autre côté, si ce n'est pas le cas il ne pourra le trouver qu'avec probabilité strictement inférieure à 1. Nous testerons cela dans la dernière partie en créant des modèles qui vérifient ou ne vérifient pas les conditions d'irreprésentabilité et en observant le modèle évalué par le Lasso.

2.3 Le cas de grands p et q

Penchons nous maintenant sur le cas plus intéressant où p et q ne sont plus fixes mais peuvent varier avec n. On autorise donc aux tailles de la matrice C^n et du paramètre β^n de grandir avec n. Les hypothèses de régularité (3) et (4) sont donc inapropriées car C^n ne peut plus converger. De plus, β^n peut varier avec n. On devra donc essayer de contrôler la taille de la plus petite entrée de $\beta^n(1)$ et minorer les valeurs propres de C_{11}^n pour toujours pouvoir l'inverser sans difficultés. Nous allons donc proposer de nouvelles hypothèses de régularité. On suppose qu'il existe $0 \leq c_1 < c_2 \leq 1$ et M_1, M_2, M_3, $M_4 > 0$ tels que :

$$\forall j \in \{1, ..., p\} \qquad \tfrac{1}{n} (X_{\cdot,j}^n)^T X_{\cdot,j}^n \leq M_1, \tag{7}$$

$$\forall \|\alpha\|_2^2 = 1 \qquad \alpha^T C_{11}^n \alpha \geq M_2, \tag{8}$$

$$q_n = O(n^{c_1}), \tag{9}$$

$$n^{\frac{1-c_2}{2}} \min_{i=1,...,q} |\beta_i^n| \geq M_3. \tag{10}$$

La condition (7) peut s'obtenir en normalisant la variance des variables potentiellement explicatives. Elle les oblige à bouger un minimum. La condition (8) s'obtient grâce à une minoration des valeurs propres de C_{11}^n afin de pouvoir l'inverser sans rencontrer de problèmes. Les réelles conditions sont (9) et (10). La condition (9) empêche le nombre de vraies variables explicatives d'augmenter trop vite, sans quoi le Lasso n'aurait pas le temps de toutes les détecter. La condition

(10) quant à elle, interdit aux paramètres des variables explicatives de décroître trop vite vers 0. Leur impact serait alors caché par le bruit et ne pourrait pas être perçu. Le bruit joue ici un rôle bien plus important que dans le cas où p et q étaient fixes. En effet, le Lasso pouvait alors déterminer le bon modèle quel que soit le bruit, du moment que nous faisions suffisamment d'observations. Nous verrons que ce n'est plus ici le cas.

Théorème 3. $(CI+ \Rightarrow CS+)$**.**
Si (ε_i^n) est une famille de variables aléatoires i.i.d. avec un $2k^{eme}$ moment fini ($\mathbb{E}\left[(\varepsilon_1^n)^{2k}\right] < \infty$) pour un entier $k \geq 1$ et sous les conditions (7), (8), (9) et (10), si la condition d'irreprésentabilité forte est vérifiée alors,

1. *le Lasso est fortement consistant en signes pour $p = o(n^{(c_2-c_1)k})$.*
2. *$\forall \lambda_n$ tel que $\frac{\lambda_n}{\sqrt{n}} = o\left(n^{\frac{c_2-c_1}{2}}\right)$ et $\frac{1}{p}\left(\frac{\lambda_n}{\sqrt{n}}\right)^{2k} \to \infty$, on a*

$$P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - O\left(\frac{pn^k}{\lambda_n^{2k}}\right) \underset{n\to\infty}{\longrightarrow} 1\,.$$

Démonstration. Nous commençons la preuve de la même manière que celle du théorème 1, en étudiant les convergences de $P(A_n^c)$ et $P(B_n^c)$. Nous voulons montrer qu'elles tendent vers 0 de manière polynomiale. On pose donc $z^n = (C_{11}^n)^{-1}W^n(1) = \frac{1}{\sqrt{n}}(C_{11}^n)^{-1}X_n(1)^T\varepsilon_n \in \mathbb{R}^q$, $\zeta^n = C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2) \in \mathbb{R}^{p-q}$ et $b^n = (C_{11}^n)^{-1}\operatorname{sign}(\beta_{(1)}^n) \in \mathbb{R}^q$ et on a

$$\begin{aligned} P(A_n^c) &\leq \sum_{i=1}^{q} P\left(|z_i^n| \geq \sqrt{n}\left(|\beta_i^n| - \frac{\lambda_n}{n}b_i^n\right)\right) \\ P(B_n^c) &\leq \sum_{i=1}^{p-q} P\left(|\zeta_i^n| \geq \frac{\lambda_n}{\sqrt{n}}\eta_i\right). \end{aligned}$$

Commençons par $P(A_n^c)$ et tentons de la majorer. Pour cela nous allons montrer que les $2k^{emes}$ moments de z_i^n sont finis. Écrivons $z^n = H\varepsilon_n$ avec $H = (H_1, ..., H_q)^T = \frac{1}{\sqrt{n}}(C_{11}^n)^{-1}X_n(1)^T \in M_{q,n}(\mathbb{R})$. Nous avons donc $z_i^n = H_i^T\varepsilon_n$. L'idée est de majorer H_i afin que z_i^n ait les mêmes moments que ε_n. Pour cela, remarquons que

$$\begin{aligned} HH^T &= \frac{1}{n}(C_{11}^n)^{-1}X_n(1)^T((C_{11}^n)^{-1}X_n(1)^T)^T \\ &= (C_{11}^n)^{-1}\frac{1}{n}X_n(1)^TX_n(1)((C_{11}^n)^{-1})^T \\ &= (C_{11}^n)^{-1}(C_{11}^n)^T((C_{11}^n)^T)^{-1} \\ &= (C_{11}^n)^{-1}. \end{aligned}$$

On a alors pour tout $i = 1, ..., q$,

$$\begin{aligned} \|H_i\|_2^2 &= \left[HH^T\right]_{i,i} \\ &= \left[(C_{11}^n)^{-1}\right]_{i,i} \\ &\leq \frac{1}{M_2}. \end{aligned} \tag{11}$$

La dernière inégalité s'obtient en appliquant la condition de régularité (8) avec $\alpha = (0, .., 0, 1, 0, .., 0)^T \in \mathbb{R}^n$ où le 1 est en i^{eme} position. Or, on peut montrer que pour tout $i = 1, ..., q$,

$$\mathbb{E}\left[\left(H_i^T \varepsilon^n\right)^{2k}\right] \leq (2k-1)! \, \|H_i\|_2^2 \, \mathbb{E}\left[\left(\varepsilon_1^n\right)^{2k}\right]$$

Donc pour tout $i = 1, ..., q$, $\mathbb{E}\left[\left(H_i^T \varepsilon^n\right)^{2k}\right] < \infty$.

Si Z est une variable aléatoire dont le $2k^{eme}$ moment est fini alors par l'inégalité de Markov sa distribution de probabilité est majorée de la manière suivante,

$$P(Z > t) = O(t^{-2k}).$$

Nous allons appliquer cette propriété à $|z_i^n|$ qui admet aussi un $2k^{eme}$ moment. Pour cela commençons par remarquer,

$$\begin{aligned} \left|\frac{\lambda_n}{n} b_n\right| &= \frac{\lambda_n}{n} \left|(C_{11}^n)^{-1} \operatorname{sign}(\beta_{(1)}^n)\right| \\ &\leq \frac{\lambda_n}{nM_2} \left\|\operatorname{sign}(\beta_{(1)}^n)\right\|_2 \\ &= \frac{\lambda_n}{nM_2} \sqrt{q} \end{aligned} \tag{12}$$

où l'inégalité est obtenue grâce à la condition de régularité (8). On a alors,

$$\begin{aligned} P\left(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{n} b_i^n)\right) &\leq P\left(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{nM_2}\sqrt{q})\right) \\ &= O\left(n^{-k}\left(|\beta_i^n| - \frac{\lambda_n}{nM_2}\sqrt{q}\right)^{-2k}\right) \\ &= O\left(\frac{n^{-k}}{|\beta_i^n|^{2k}}\left(1 - \frac{\lambda_n\sqrt{q}}{nM_2|\beta_i^n|}\right)^{-2k}\right) \end{aligned}$$

Or, pour λ_n tel que $\frac{\lambda_n}{\sqrt{n}} = o\left(n^{\frac{c_2 - c_1}{2}}\right)$, d'après les conditions (9) et (10),

$$\frac{\lambda_n\sqrt{q}}{nM_2|\beta_i^n|} \underset{n\to\infty}{\longrightarrow} 0.$$

On en déduit,

$$\begin{aligned} P\left(|z_i^n| \geq \sqrt{n}\left(|\beta_i^n| - \frac{\lambda_n}{n} b_i^n\right)\right) &= O\left(\frac{n^{-k}}{|\beta_i^n|^{2k}}\right) \\ &= O\left(n^{-kc_2}\right). \end{aligned}$$

Donc, grâce aux conditions de régularité (9) et (10),

$$\begin{aligned} P(A_n^c) &\leq \sum_{i=1}^{q} P\left(|z_i^n| \geq \sqrt{n}\left(|\beta_i^n| - \frac{\lambda_n}{n} b_i^n\right)\right) \\ &= q \cdot O\left(n^{-kc_2}\right) \\ &= o\left(\frac{pn^k}{\lambda_n^{2k}}\right). \end{aligned}$$

De manière analogue en utilisant la condition de régulartité (7), $\mathbb{E}\left[(\zeta_i^n)^{2k}\right] < \infty$ et on obtient,

$$P(B_n^c) = O\left(\frac{pn^k}{\lambda_n^{2k}}\right).$$

On remarque finalement que pour $p = o\left(n^{c_2 - c_1}\right)$, on peut choisir (λ_n) tel que $\frac{\lambda_n}{\sqrt{n}} = o\left(n^{\frac{c_2 - c_1}{2}}\right)$ et $\frac{1}{p}\left(\frac{\lambda_n}{\sqrt{n}}\right)^{2k} \to \infty$, ce qui conclut la preuve du théorème. □

Le théorème précédent donne à quelle vitesse peut croître p pour que le Lasso choisisse le bon modèle si $CI+$ est vérifiée et si le bruit a un moment fini. Par exemple, si seul le second moment du bruit est fini, p doit croître moins vite que $n^{c_2 - c_1}$. Si tous les moments du bruit existent alors p peut croître aussi vite que n'importe quel polynôme en n. La probabilité que le Lasso choisisse le bon modèle converge vers 1 plus vite que tout polynôme. Dans la nature, le bruit Gaussien occupe une place importante, aussi énonçons nous le corollaire suivant.

Théorème 4. (Bruit Gaussien)**.**
Si (ε_i^n) est une famille de variables aléatoires i.i.d. Gaussiennes, sous les conditions (7), (8), (9) et (10) et s'il existe $0 \leq c_3 < c_2 - c_1$ tels que $p = O(e^{n^{c_3}})$, alors $CI+$ implique $CS+$. En particulier, $\forall \lambda_n$ tel que $\lambda_n \propto n^{\frac{1+c_4}{2}}$ avec $c_3 < c_4 < c_2 - c_1$,

$$P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - O(e^{-n^{c_3}}) \underset{n \to \infty}{\longrightarrow} 1.$$

Démonstration. Nous reprenons les notations de la preuve du théorème précédent. Comme (ε_i^n) est une famille de variables aléatoires i.i.d. Gaussiennes, z_i^n et ζ_i^n sont Gaussiennes. D'après (11), $\mathbb{E}\left[(z_i^n)^2\right]$ est borné, il en est de même pour $\mathbb{E}\left[(\zeta_i^n)^2\right]$. Comme lors des preuves précédentes, il nous reste à borner la fonction de répartition de nos variables afin de majorer $P(A_n^c)$ et $P(B_n^c)$. Cependant, nous n'allons pas utiliser la borne du théorème précédent mais une plus précise vue dans la preuve du théorème 1. La fonction de répartition Φ d'une distribution Gaussienne centrée réduite vérifie

$$\forall t \in \mathbb{R} \qquad 1 - \Phi(t) < \frac{e^{-\frac{t^2}{2}}}{t}.$$

On obtient alors en utilisant les différentes conditions de régularité par (12) pour $\lambda_n \propto n^{\frac{1+c_4}{2}}$,

$$\begin{aligned} P(A_n^c) &= q \cdot O\left(1 - \Phi((1+o(1))M_3 M_2 n^{c_2/2})\right) \\ &= o(e^{-n^{c_3}}) \\ P(B_n^c) &= o(e^{-n^{c_3}}). \end{aligned}$$

Ce qui complète la preuve du théorème. □

Il est enthousiaste de pouvoir permettre au nombre de variables potentiellement explicatives d'augmenter plus rapidement que le nombre d'observations par la méthode du Lasso, allant jusqu'à un rapport exponentiel. Cela montre sa supériorité à une régression linéaire classique. Cependant, nous avons vu que cela n'est pas forcément vrai pour n'importe quelle distribution du bruit. En général, le théorème 3 est fin dans le sens où si le bruit n'admet pas de grands moments, alors l'erreur dû au bruit ne décroitra pas suffisament vite vers 0 pour permettre à p d'augmenter polynomialement à de grands degrés en n. Nous ne montrons pas ici la nécessité des conditions d'irreprésentabilité pour que le Lasso soit consistant. Zhao et Yu ne le montrent pas non plus dans ([1]) pour des raisons techniques.

Maintenant que nous savons quand le Lasso fonctionne ou ne fonctionne pas, il nous reste à tester nos résultats sur des exemples. Ceci est l'objet de la partie suivante.

3 La mise en pratique

L'objectif de cette partie est d'illustrer nos résultats à l'aide de simulations en *Scilab*. L'algorithme implémenté est l'algorithme 2 présenté dans la première partie.

3.1 Deux exemples simples

Commençons par deux exemples simples, où p et q sont petits. Regardons s'ils vérifient les conditions d'irreprésentabilité puis vérifions que le Lasso se comporte comme prévu.

On commence par générer pour $i = 0, ..., 1000$, des variables aléatoires Gaussiennes $x_{i,1}$, $x_{i,2}$, e_i et ε_i i.i.d. de variance 1 et de moyenne 0. Une troisième variable potentiellement explicative $x_{i,3}$ est aussi générée de manière à être corrélée avec $x_{i,1}$ et $x_{i,2}$ de la manière suivante,

$$x_{i,3} = \frac{2}{3}x_{i,1} + \frac{2}{3}x_{i,2} + \frac{1}{3}e_i \,.$$

Par construction, $(x_{i,3})$ est aussi une famille de variables aléatoires i.i.d. de variance 1 et de moyenne 0. La variable que nous cherchons à expliquer est Y_i générée par,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \varepsilon_i \,.$$

Nous allons étudier deux modèles :

1. $\beta_1 = -2$ et $\beta_2 = 3$.
2. $\beta_1 = 2$ et $\beta_2 = 3$.

Dans ces deux modèles, $X(1) = (X_1, X_2)$ et $X(2) = X_3$. Nous sommes dans le cas où p et q sont fixes et indépendants de n. Commençons par vérifier que les conditions de régularité (3) et (4) sont vérifiées. Commençons par la condition (4). Soient $1 \leq j \leq p$ et $\eta > 0$,

$$\begin{aligned}
P\left(\frac{1}{n}\max_{1\leq i\leq n} x_{i,j}^2 > \frac{\eta}{p}\right) &= 1 - P\left(\forall i = 1,...,n \;\; x_{i,j}^2 \leq n\frac{\eta}{p}\right) \\
&= 1 - P\left(x_{1,j}^2 \leq n\frac{\eta}{p}\right)^n \\
&= 1 - \left(\int_0^{\sqrt{n\frac{\eta}{p}}} \sqrt{\frac{2}{\pi}} e^{-\frac{t^2}{2}} dt\right)^n \\
&= 1 - \left(1 - \int_{\sqrt{n\frac{\eta}{p}}}^{\infty} \sqrt{\frac{2}{\pi}} e^{-\frac{t^2}{2}} dt\right)^n \\
&= 1 - \left(1 - \sqrt{\frac{2}{\pi}}\frac{e^{-n\frac{\eta}{p}}}{\sqrt{n\frac{\eta}{p}}} + o\left(\frac{e^{-n\frac{\eta}{p}}}{\sqrt{n\frac{\eta}{p}}}\right)\right)^n \\
&= 1 - \exp\left(-\sqrt{\frac{2n}{\pi\frac{\eta}{p}}} e^{-n\frac{\eta}{p}} + o\left(\sqrt{\frac{pn}{\eta}} e^{-n\frac{\eta}{p}}\right)\right) \\
&\underset{n\to\infty}{\sim} \sqrt{\frac{2pn}{\pi\eta}} e^{-n\frac{\eta}{p}}\,.
\end{aligned}$$

Comme $p = 3$ est fixe, on obtient,

$$\begin{aligned}
P\left(\frac{1}{n}\max_{1\leq i\leq n}\sum_{j=1}^{p} x_{i,j}^2 > \eta\right) &\leq P\left(\exists j \in \{1,..,p\} \;\; \frac{1}{n}\max_{1\leq i\leq n} x_{i,j}^2 > \frac{\eta}{p}\right) \\
&\leq \sum_{j=1}^{p} P\left(\frac{1}{n}\max_{1\leq i\leq n} x_{i,j}^2 > \frac{\eta}{p}\right) \\
&= pO\left(\sqrt{\frac{2pn}{\pi\eta}} e^{-n\frac{\eta}{p}}\right) \\
&\underset{n\to\infty}{\longrightarrow} 0\,.
\end{aligned}$$

Nous en déduisons une convergence en probabilité. Finalement, par Borel-Cantelli, presque sûrement, la condition (4) est vraie à partir d'un certain rang.

De plus,

$$\begin{aligned}
\mathcal{C}^n \xrightarrow{p.s.} \mathcal{C} &= \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_1, X_3) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}(X_2) & \mathrm{Cov}(X_3, X_2) \\ \mathrm{Cov}(X_1, X_3) & \mathrm{Cov}(X_2, X_3) & \mathrm{Var}(X_3) \end{pmatrix} \\
&= \begin{pmatrix} 1 & 0 & \frac{2}{3} \\ 0 & 1 & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & 1 \end{pmatrix}
\end{aligned}$$

La condition (3) est donc aussi vraie p.s. et on a $C_{21}C_{11}^{-1} = \left(\frac{2}{3}, \frac{2}{3}\right)$. On en déduit que (CI+) est vérifiée pour le modèle (1) mais pas pour le modèle (2). En figure 3 et 4, nous donnons le chemin de régularisation du Lasso pour un échantillon de taille $n = 1000$, respectivement pour les modèles (1) et (2). D'après le théorème 1 pour le modèle (2), pour tout λ_n tel que $\lambda_n/n \to 0$ et $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ avec $0 \leq c < 1$, $P(\widehat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c})$. 1000 observations devrait donc être amplement suffisant. Le Lasso est effectivement consistant en signes pour le modèle (1) et pas pour le modèle (2).

En effet, sur la figure 4 nous voyons que le Lasso ne fait pas rétrécir β_3 exactement jusqu'à zéro. Au contraire, la régularisation commence par préférer X_3 à X_1 et X_2 et le Lasso choisit donc en premier X_3 et ne la réduit jamais de nouveau à zéro. Dans le modèle 1 en revanche, (CI+) est vérifiée et si on choisit un bon λ, le Lasso réduit bien β_3 à zéro et donne les bons signes à β_1 et β_2. Notons que β_3 n'est plus exactement nul pour λ très proche de zéro, bien que cela ne se voit pas sur la figure.

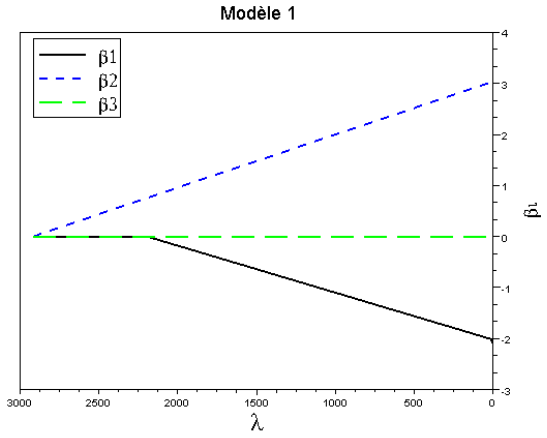

FIGURE 3 – $\beta_1 = -2$, $\beta_2 = 3$, un exemple pour illustrer la consistance en signes du Lasso

3.2 Un exemple un peu plus compliqué

Les deux exemples précédents manquaient un peu d'intérêt pratique. En effet, on peut remarquer qu'une régression linéaire classique (le cas $\lambda = 0$) aurait donné de très bon résultats, ce qui n'est pas très étonnant avec 3 variables et 1000 observations. Qui plus est, notre titre annonçait « comment choisir parmi un grand nombres de variables à l'aide de peu d'observations », nos exemples en étaient plutôt loin ! Cependant, ils avaient l'avantage d'être clairs et simples. Nous allons maintenant nous intéresser au cas plus intéressant où nous disposons de moins d'observations que de variables. On prend $q = 10$ et $p_n = 10n$. Nous générons pour $i = 0, ..., n$, des variables aléatoires $e_{i,1}$,...,e_{i,p_n}, x_i

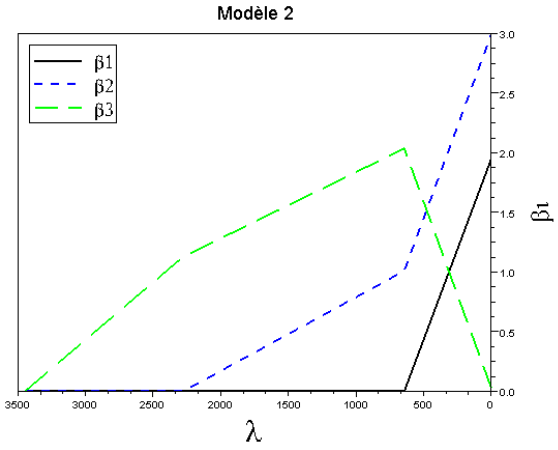

FIGURE 4 – $\beta_1 = 2$, $\beta_2 = 3$, un exemple pour illustrer l'inconsistance en signes du Lasso

et ε_i i.i.d. Gaussiennes centrées réduites. Pour $1 \leq i \leq n$ et $1 \leq j \leq p_n$, on pose

$$x_{i,j} = \frac{1}{5} x_i + \frac{2\sqrt{6}}{5} e_{i,j}$$

Les $X_j^n = (x_{1,j}, ..., x_{n,j})^T$ sont les observations des variables potentiellement explicatives. Ce sont bien des variables de variance 1 et de moyenne 0. La variable à expliquer Y est définie de la manière suivante,

$$Y_i = \sum_{j=1}^{p} \beta_j^n x_{i,j} + \varepsilon_i$$

où $\beta^n \in \mathbb{R}^{p_n}$ est tel que $\beta_j^n = (-1)^j j n^{-\frac{1}{6}} \neq 0$ pour $1 \leq j \leq q_n$ et $\beta_j = 0$ pour $j \geq q_n + 1$. On a donc $X^n(1) = (X_1^n, ..., X_{q_n}^n)$ et $X^n(2) = (X_{q_n+1}^n, ..., X_{p_n}^n)$.

Commençons par nous assurer que les conditions de régularité sont vérifiées presque sûrement. Comme q est fixe, $C_{11}^n \xrightarrow[n\to\infty]{p.s.} C_{11}$ qui est inversible, la condition (8) est donc vraie p.s. La condition (9) est immédiate avec $c_1 = 0$. Nous avons de plus choisi β_j^n de telle sorte que la dernière condition de régularité soit aussi vérifiée avec $c_2 = \frac{2}{3}$ et $M_3 = 1$.

Montrons maintenant que la condition (7) est vraie presque sûrement. Il s'agit de montrer qu'il existe $M_1 \geq 0$ tel que p.s. on a

$$\forall j \in \{1, ..., p_n\} \quad \frac{1}{n} \sum_{i=1}^{n} x_{i,j}^2 \leq M_1 \,.$$

Pour cela, admettons la proposition suivante (cf. [5]).

Proposition 2 (Théorème de concentration Gaussienne)**.** *Soient P_1 la mesure Gaussienne standard sur l'espace euclidien $\mathbb{R}^n$ et ζ une fonction ℓ-lipschitzienne réelle sur $\mathbb{R}^n$. Alors pour tout $t > 0$*

$$P_1\left(\zeta \geq E\left[\zeta\right] + t\right) \leq \frac{1}{2} e^{-\frac{u^2}{2\ell^2}} .$$

Notons $Z^2 = \sum_{i=1}^n x_i^2 = \|x\|^2$ la norme euclidienne sur $\mathbb{R}^n$, alors on a $E(Z^2) = n$ et le but est de majorer la probabilité d'avoir $Z^2 \geq nM_1$. Comme $Z = \|x\|$ est une fonction 1-lipschitzienne de x dans $\mathbb{R}^n$, on peut appliquer le théorème de concentration Gaussienne,

$$P_1\left(Z \geq E[Z] + t\right) \leq \frac{1}{2} e^{-\frac{t^2}{2}} .$$

Or, $E(Z)^2 \leq E(Z^2) = n$, donc on deduit que $P_1\left(Z^2 \geq (\sqrt{n} + t)^2\right) \leq \frac{1}{2} e^{-\frac{t^2}{2}}$. En prenant $t = \sqrt{2n}$, il vient

$$P_1\left(Z^2 \geq 3n\right) \leq \frac{1}{2} e^{-n} .$$

On obtient donc avec $M_1 = 3$,

$$P\left(\exists j \in \{1, .., p_n\} \text{ tel que } \frac{1}{n}\sum_{i=1}^n x_{i,j}^2 \geq M_1\right)$$

$$\begin{aligned}
&\leq \quad \sum_{j=1}^{p_n} P\left(\frac{1}{n}\sum_{i=1}^n x_{i,j}^2 \geq M_1\right) \\
&= \quad p_n P\left(\frac{1}{n}\sum_{i=1}^n x_{i,1}^2 \geq M_1\right) \\
&\leq \quad p_n \frac{1}{2} e^{-n} \\
&\underset{n\to\infty}{\longrightarrow} \quad 0 .
\end{aligned}$$

Par Borel-Cantelli, on conclut que presque sûrement la condition (7) est vraie à partir d'un certain rang.

Assurons nous maintenant que (CI+) est vraie. Admettons le résultat suivant, sa preuve n'étant pas très compliquée mais fastidieuse et ne faisant pas l'objet de notre étude. Voir [1] pour plus de détails.

Proposition 3. *Supposons que la diagonale de C^n est $\mathbf{1}_{\mathbb{R}^p}$ et que ses coefficients c_{ij} soient tels que $|c_{ij}| \leq \frac{c}{2q-1}$ pour une constante $0 \leq c < 1$ alors le Lasso vérifie la condition d'irreprésentabilité forte* (CI+).

Cette proposition vérifie l'intuition que quand les variables potentiellement explicatives sont peu corrélées alors le Lasso est consistant en signes. Nous nous trouvons ici précisément dans ce cas. En effet, nous pouvons normaliser les données de manière à obtenir des 1 sur la diagonale et $c_{i,j}^n \xrightarrow{p.s.} \mathrm{Cov}(x_{1,i}, x_{1,j}) = \frac{1}{25}$. Admettons ici les $p(p-1)$ coefficients non-diagonaux de C^n sont bien simultanément bornés sur

un événement de grande probabilité. Cela se prouve en utilisant le théorème de concentration Gaussienne.

D'après le théorème 4, comme $p_n = 10n = O\left(e^{n^{c_3}}\right)$ pour tout $0 < c_3 < c_2 - c_1 = \frac{2}{3}$, le Lasso est ici consistant en signes. Pouvus que l'on fasse suffisamment d'observations, si l'on prend $\lambda_n \propto n^{\frac{1+c_4}{2}}$ avec $0 < c_4 < \frac{2}{3}$, $\widehat{\beta}^n(\lambda_n)$ choisira le bon modèle en donnant les bons signes aux différents paramètres.

Après simulations, le Lasso s'approche du bon modèle lorsque n grandit. Cependant, nous n'avons techniquement pas pu simuler l'expérience avec n suffisamment grand pour qu'il trouve le bon modèle exactement. La figure 5 illustre une simulation pour $n = 30$. Nous voyons que le Lasso arrive facilement à donner les bon signes des coordonnées telles que $\beta_i \neq 0$, mais a du mal a déterminer toutes les variables innutiles.

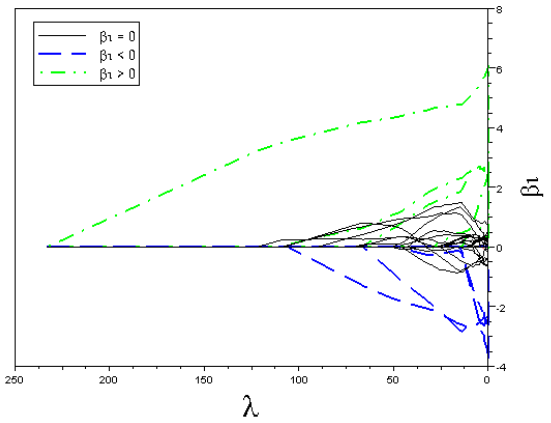

FIGURE 5 – Simulation pour $n = 30$, limites du Lasso.

Ceci représente les limites des résultats établis qui sont asymptotiques, et ne sont pas toujours atteignables dans la vie réelle en un temps raisonnable. Le Lasso trouve le bon modèle à partir de $n = 500$ mais encore qu'avec probabilité strictement inférieure à 1. Il élimine rarement toutes les variables inutiles. Nous pourrions accélérer la convergence en diminuant la corrélation entre les variables ou le bruit.

3.3 Jeu de données réelles et validation croisée

Dans les exemples précédents, nous avions nous-même créé nos modèles. Il était donc aisé de vérifier si le Lasso donnait une bonne estimation. C'était cependant l'objet de ces exemples, illustrer les résultats que nous avions établi. L'enjeu de cette partie est différent. Nous disposons ici d'un jeu de données réelles sur le diabète présent dans [7] et nous voulons déterminer les causes du diabète. Les données ont été centrées et réduites afin de pouvoir appliquer le Lasso correctement.

Nous avons 10 variables potentiellement explicatives (l'age, le sexe,...) d'une autre autres variable qui correspond au diabète, des mesures ont étées faites sur 442 patients. Nous avons donc $p = 10$ et $n = 442$ mais nous ne connaissons pas le modèle. Nous pourrions appliquer la méthode du Lasso sur toutes les observations mais comment choisir le λ qui donne le bon modèle ? Nous allons utiliser la méthode de validation croisée expliquée plus en détails dans [4].

L'idée est d'appliquer la méthode du Lasso sur une partie I_1 des observations. On en déduit un vecteur de paramètres $\beta(\lambda, I_1) \in \mathbb{R}^p$ en appliquant une régression linéaire classique sur les variables sélectionnées par le Lasso en λ,

$$Y_{Lasso}(\lambda, I_1) = \sum_{j=0}^{p} \beta_j(\lambda, I_1) X_j \,.$$

On utilise ensuite la deuxième partie I_2 des observations pour valider le modèle. L'objectif étant de prédire le mieux possibles les différentes observations de Y à partir des observations des variables potentiellement explicatives, c'est à dire minimiser

$$F(\widehat{\beta}(\lambda, I_1), I_2) = \frac{1}{|I_2|} \sum_{i \in I_2} (Y_{Lasso,i}(\lambda, I_1) - Y_i)^2 \,.$$

Nous calculons $F(\widehat{\beta}(\lambda, I_1), I_2)$ pour tout $\lambda \in \overline{\mathbb{R}}_+$, et on choisit

$$\widehat{\lambda} = \underset{\lambda \in \overline{\mathbb{R}}_+}{\arg\min}\, F(\widehat{\beta}(\lambda, I_1), I_2) \,.$$

Nous en déduisons le modèle cherché,

$$\widehat{\beta}(\widehat{\lambda}, I_1), I_2) \,.$$

On peut recommencer cela en faisant varier I_1 et I_2 et choisir le modèle qui nous semble le meilleur.

Après une simulation, nous obtenons en choisissant I_1 comme la première moitié des observations et I_2 la deuxième, le meilleur modèle a lieu en $\widehat{\lambda} = 18$, l'erreur quadratique moyenne sur I_2 est alors

$$F(\widehat{\beta}(\lambda, I_1), I_2) = 2916 \,.$$

Elle est assez importante sachant que Y est de l'ordre de la dizaine. Mais on peut remarquer que l'erreur est tout de même bien moins grosse que celle du modèle $\beta = \mathbf{0}_{\mathbb{R}^p}$, qui suppose qu'aucune de nos variables n'explique réellement le diabète (de manière linéaire), qui vaut plus de 6200.

Il ne faut pas oublier de noter que le Lasso n'est pas toujours le meilleur algorithme a utiliser (notamment parce qu'il suppose une dépendance linéaire en les variables, ce qui est sans doute faux). Nous allons donc tenter une astuce qui va nous peut-être nous permettre d'améliorer notre modèle. Adjoindre aux variables « brutes »leur carré, leur cube, leur inverse, etc. Ceci permet d'avoir une dépendance non-lineaire sans modifier fondamentalement l'algorithme. Cependant, il

ne faut pas oublier une chose. Nous ne vérifions pas ici si les conditions d'irreprésentabilité sont vérifiées mais il faut garder en mémoire que plus les variables potentiellement explicatives sont liées, moins le Lasso a de chances d'être consistant en signes. Nous devons donc tout de même faire attention.

Après simulation, nous obtenons un modèle en $\widehat{\lambda} = 15$ qui n'a plus qu'une erreur quadratique moyenne sur la deuxième moitié des observations

$$F(\widehat{\beta}(\lambda, I_1), I_2) = 2798\,.$$

Nous avons réussi à améliorer légèrement notre modèle, cependant cela reste toujours assez médiocre. L'erreur peut provenir par exemple de variables explicatives que nous n'avons pas prises en compte dans le modèle.

Conclusion

Le Lasso est donc une méthode pour choisir parmis un grand nombre de variables en faisant peu d'observations. C'est une méthode utilisée en pratique dans des domaines comme la biologie ou le nombre de variables potentiellement explicatives est important et où nous ne pouvons pas toujours faire suffisamment d'observations.

Nous avons dans cet article montré que sous certaines hypothèses le Lasso est consistant en signes si et seulement si les conditions d'irreprésentabilité sont vérifiées que p et q soient petits ou grands. Mais en pratique, sont elles souvent vérifiées ? Nous avons vu dans le deuxième exemple que la condition de forte irreprésentabilité est vérifiée dans des cadres assez spéciaux, et nous avons vu dans le premier exemple qu'il arrive qu'elle soit fausse. Nous pourrions donc nous demander à quel point cette condition est forte, afin de savoir si le Lasso est fréquemment ou rarement consistant. C'est à dire, si nous prenons p et q donnés, avec quelle probabilité un modèle choisi aléatoirement vérifie-t-il cette condition ? Nous verrions que pour $q = 1$, $CI+$ est en général vérifiée mais dès que p et q ne sont plus très petits, les conditions d'irreprésentabilité ne sont vérifiées que très rarement.

De plus, même lorsque par chance nous nous trouvons dans ce cas, les résultats montrés précédemment ne sont qu'asymptotiques. Il arrive que le Lasso ne puisse techniquement pas choisir le bon modèle en un temps raisonnable.

Il existe de nombreuses améliorations du Lasso qui permettent de résoudre ces problèmes. Leur étude est un domaine encore très actuel.

Références

[1] P. Zhao et B. Yu. On Model Selection Consistency of Lasso. *Journal of Machine Learning Research* 7, 2541-2563, 2006.

[2] B. Elfron, T. Hastie et R. Tibshirani. Least angle regression. *Annals of Statistics* 32, 407-499, 2004.

[3] J. Mairal. Simple Explanation for LARS. 2008.

[4] P.A. Cornillon et E. Matzner-Løber. *Régression, Théorie et applications*. Springer, 2007.

[5] P. Massart. *Concentration Inequalities and Model Selection*. Springer, 2007.

[6] K. Knight et W.J. Fu. Asymptotics for Lasso-type estimators. *Annals of Statistics* 28, 1356-1378, 2000.

[7] *Jeu de données sur le diabète*. http ://www-stat.stanford.edu/ hastie/Papers/LARS/.

Nous tenons à remercier Sylvain Arlot pour son encadrement et sans lequel cet exposé n'aurait pas vu le jour.