Theoretical guidelines for high-dimensional data analysis: False Discoveries occur Early on the Lasso Path

Auteurs: BALL Alhousseynou et Thi Hai Yen VU
February 14, 2020

Abstract

Ce rapport est rédigé dans le cadre du cours de Theoretical guidelines for high-dimensional data analysis, sous la direction de Christophe Giraud. Le but de ce rapport est de montrer les forces et faiblesses du Lasso en s'appuyant sur l'article "False Discoveries occur Early on the Lasso Path". La version pdf du papier est accessible via ce lien (https://statweb.stanford.edu/ candes/publications/downloads/LassoFDR.pdf).

Introduction

En statistiques, le modèle Lasso est connu comme une méthode de contraction des coefficients de la régression. Il est souvent utilisé pour retrouver des coefficients importants dans un modèle de régression. Cependant, on n'a aucune garantie que le modèle Lasso trouverait parfaitement des bons coefficients du modèle. Ce rapport, qui est basé sur l'article "False Discoveries occur Early on the Lasso Path", vise à montrer que le modèle Lasso a tendance à explorer à la fois des bon et des faux coefficients très tôt tant que le modèle de régression est dans un régime de sparsité linéaire, c'est à dire que le taux de variables non nuls est une petite constante.

On va commencer par donner le contexte et les notions de bases comme le modèle de régression, la méthode Lasso, le chemin du Lasso, l'erreur du type I et II, etc. Puis on va présenter les principaux résultats de l'article et quelques intuitions concernant ces résultats. Finalement, on va explorer ces résultats avec nos propres expériences numériques, puis conclure.

1 Contexte

1.1 Régression linéaire

Globalement, une régression linéaire consiste à chercher des vrais coefficients d'un modèle linéaire sachant des vecteurs d'entrée et les observations correspondantes. Formellement, étant donné une matrice de features $\mathbf{X} \in \mathbb{R}^{n \times p}$ et un vecteur d'observation $\mathbf{y} \in \mathbb{R}^n$ satisfaisant :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{z}, \tag{1}$$

où $\boldsymbol{\beta} \in \mathbb{R}^p$ est le vrai vecteur de coefficient et $\mathbf{z} \in \mathbb{R}^n$ est un vecteur de bruit. Le but du régression linéaire est de trouver les coefficients $\widehat{\boldsymbol{\beta}}$ minimisant une fonction de coût $\mathcal{L}(\mathbf{Xb}, \mathbf{y}, \mathbf{b})$:

$$\widehat{\boldsymbol{\beta}} \in \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \, \mathcal{L}(\mathbf{Xb}, \mathbf{y}, \mathbf{b}). \tag{2}$$

La choix de fonction de coût est important et peut beaucoup affecter la solution du modèle. Par exemple, une fonction fréquemment utilisé est la perte de la norme au carré :

$$\mathcal{L}(\mathbf{Xb}, \mathbf{y}, \mathbf{b}) = \frac{1}{2} \left\| \mathbf{y} - \mathbf{Xb} \right\|^2$$

qui conduisent à la *méthode des moindres carrés ordinaire* avec la solution $\widehat{\boldsymbol{\beta}}$ donnée par (dans le cas où $\mathbf{X}$ est une matrice de plein rang) :

$$\widehat{\boldsymbol{\beta}} \in \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \, \frac{1}{2} \left\| \mathbf{y} - \mathbf{Xb} \right\|^2 = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Cette méthode marche bien si le nombre de données est beaucoup plus que le nombre de features ($n \gg p$). Cependant, quand il y a moins de données, elle marche réellement moins bien et de plus elle n'est pas robuste aux valeurs aberrantes.

1.2 Méthode Lasso

Comme on en a discuté avant, quand $p > n$, la méthode des moindres carrés ordinaire marche moins bien. Dans ce cas, on va se concentre à la méthode Lasso, qui consiste à utiliser la perte de la norme au carré plus un terme de régularisation :

$$\mathcal{L}_\lambda(\mathbf{Xb}, \mathbf{y}, \mathbf{b}) = \frac{1}{2} \|\mathbf{y} - \mathbf{Xb}\|^2 + \lambda \|\mathbf{b}\|_1 \tag{3}$$

qui entraîne la solution du Lasso :

$$\widehat{\boldsymbol{\beta}}_\lambda \in \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{Xb}\|^2 + \lambda \|\mathbf{b}\|_1 . \tag{4}$$

Le terme au norme ℓ_1 aide à pénaliser les grands coefficients dans $\mathbf{b}$ et a pour but de contracter les coefficients.

Pour mieux comprendre cet effet, on va illustrer les graphes des normes ℓ_1 et ℓ_2 au carré :

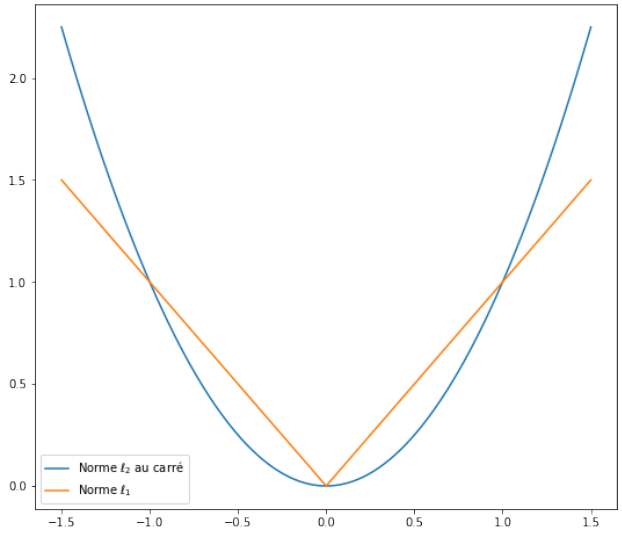
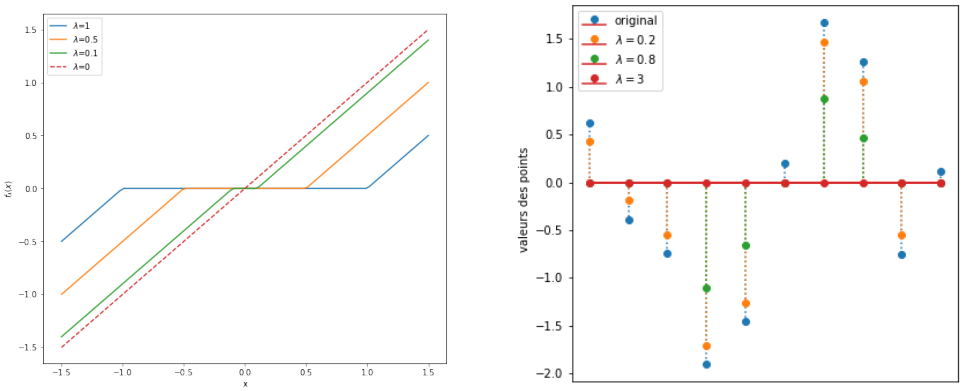


Figure 1: Graphes des normes ℓ_1 et ℓ_2 au carré

On voit que la norme ℓ_2 au carré $\|\mathbf{b}\|_2^2$ s'accroît beaucoup plus vite lors que les valeurs absolues de ses coordonnées sont plus grandes que 1, alors que elles sont très petites lors que ses coordonnées sont assez proche de 0. Cela s'explique le fait que si l'on veut minimiser une fonction en utilisant la régularisation de norme ℓ_2 au carré, les variables vont aller très vite au dessous de 1 mais ils ne tendent nécessairement pas vers 0. La norme ℓ_1, par contre, est beaucoup plus grande que la norme ℓ_2 au carré dans un voisinage de 0, ce qui implique le fait que les variables doivent se contractent à 0 pour que la fonction de perte soit plus petite possible.



(a) Les graphes des fonctions de rétrécissement

(b) L'application des fonctions de rétrécissement sur un ensemble de points aléatoires

Figure 2: L'effet de la fonction de rétrécissement pour différentes valeurs de λ

L'effet de contraction peut être clairement illustré dans le cas où la matrice $\mathbf{X}$ est composé de colonnes orthogonales. En effet, dans ce cas là on peut donner la solution exacte de la méthode Lasso :

$$\left[\widehat{\boldsymbol{\beta}}_\lambda\right]_j = \operatorname{sgn}\left(\mathbf{X}_j^\top \mathbf{y}\right)\left(|\mathbf{X}_j^\top \mathbf{y}| - \lambda\right)_+ \quad \forall j = 1, \ldots, p, \tag{5}$$

où $\mathbf{X}_j$ est la j-ème colonne de $\mathbf{X}$, et $(\cdot)_+ := \max(\cdot, 0)$. Notons de plus que si les colonnes de $\mathbf{X}$ sont orthogonales, on a alors $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$, donc la solution de moindres carrés nous donne :

$$\widehat{\boldsymbol{\beta}}^{LS} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{y}, \tag{6}$$

ce qui implique que $\left[\widehat{\boldsymbol{\beta}}^{LS}\right]_j = \mathbf{X}_j^\top \mathbf{y}$. On en déduit de l'équation (5) que :

$$\left[\widehat{\boldsymbol{\beta}}_\lambda\right]_j = f_\lambda\left(\left[\widehat{\boldsymbol{\beta}}^{LS}\right]_j\right) \quad \forall j = 1, \ldots, p, \tag{7}$$

où $f_\lambda(u) = \operatorname{sgn}(u)\left(|u| - \lambda\right)_+$ est la fonction de rétrécissement (shrinkage, ou soft threshold function). La figure 2 montre les graphes de cette fonction pour différentes valeurs de λ.

L'équation (7) montre en fait que la solution de Lasso est juste un rétrécissement (shrinkage) de λ appliqué à chaque coordonnée de la solution de moindres carrés. On déduit de la figure 2 que tant que λ agrandit, $\widehat{\boldsymbol{\beta}}_\lambda$ a de plus en plus de coefficients nuls, et dès que λ est assez grand, $\widehat{\boldsymbol{\beta}}_\lambda$ devient zéro.

Finalement, comme Lasso donne souvent les solution creuses (spare), il est surtout utilisé dans les situations où les vrais coefficients sont censés être creuses.

1.3 Chemin de Lasso

On vient de voir dans la partie précédente que la valeur de λ qui détermine l'intensité du terme de régularisation peut contrôler la sparsité (sparsity) de la solution de Lasso. Plus précisément, on voit que comme λ accroît, la solution de Lasso devient plus creuse. On va définir alors cette évolution de solutions comme le chemin de Lasso.

Formellement, le chemin de Lasso est défini comme la famille de solutions de Lasso $\widehat{\boldsymbol{\beta}}(\lambda) := \widehat{\boldsymbol{\beta}}_\lambda$ défini en (4), quand λ varie de 0 à $+\infty$. On considère les cas extrêmes de λ:

- $\lambda = 0$: Dans ce cas, le Lasso devient le problème des moindres carrés ordinaire. On a donc $\widehat{\boldsymbol{\beta}}(0) = \widehat{\boldsymbol{\beta}}^{LS}$.
- $\lambda \to +\infty$: Tous les coefficients deviennent zéros.

Le chemin de Lasso peut être vu sous la formulation suivante qui est induite par la dualité lagrangienne :

$$\widehat{\boldsymbol{\beta}}(\lambda) \in \underset{\mathbf{b} \in B_{\ell_1}(\hat{R}_\lambda)}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \quad \text{avec } \hat{R}_\lambda = |\widehat{\boldsymbol{\beta}}(\lambda)|_1. \tag{8}$$

Cette équation veut dire que la solution de Lasso $\widehat{\boldsymbol{\beta}}(\lambda)$ est en fait le point tangent entre la boule $B_{\ell_1}(\hat{R}_\lambda)$ et un contour de la perte quadratique. La figure 3 illustre ces points de solution de Lasso pour différentes valeurs de λ, formant un chemin de Lasso en rouge qui part de la solution de moindres carrés ordinaire. On peut voir que lors que λ agrandit, les points tendent vers zéro, et à partir de $\lambda = 0.3$, un coefficient de $\widehat{\boldsymbol{\beta}}(\lambda)$ devient zéro.

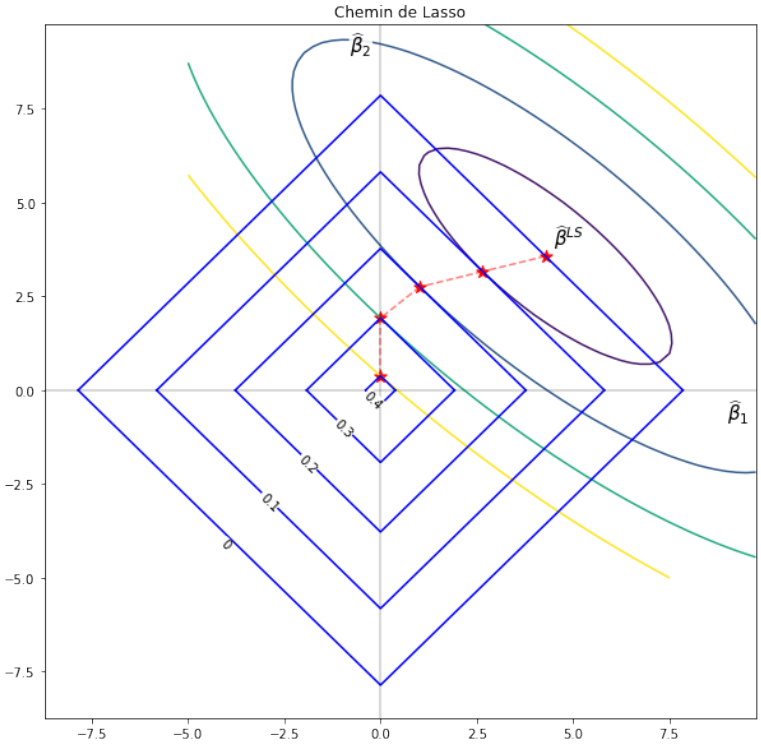


Figure 3: Solution de Lasso pour différentes valeurs de λ

On dit que un variable j est sélectionné à λ si $\widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0$. On dit également que j entre dans le chemin à λ_0 si il y a $\varepsilon > 0$ tel que $\widehat{\boldsymbol{\beta}}_j(\lambda) = 0$ pour $\lambda \in [\lambda_0 - \varepsilon, \lambda_0]$ et $\widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0$ pour $\lambda \in (\lambda_0, \lambda_0 + \varepsilon]$. De même un variable j sort du chemin à λ_0 si $\widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0$ pour $\lambda \in [\lambda_0 - \varepsilon, \lambda_0]$ et $\widehat{\boldsymbol{\beta}}_j(\lambda) = 0$ pour $\lambda \in (\lambda_0, \lambda_0 + \varepsilon]$. Par exemple, dans la figure 3 le variable 2 est sélectionné à $\lambda < 0.3$, et il sort du chemin à $\lambda_0 = 0.3$ (suppose que $\widehat{\boldsymbol{\beta}}_2(\lambda) \neq 0$ pour $\lambda < 0.3$).

1.4 TPP et FDP du Lasso

On a dit avant que le Lasso est souvent utilisé pour les problèmes où les vrais coefficients sont creuses. Le Lasso, dans ce cas, vise à trouver tous les bons coefficients du problème (c'est à dire les coefficients non nuls) sans faire beaucoup d'erreurs. On peut dire alors que le Lasso fait une vraie découverte en $\widehat{\boldsymbol{\beta}}_j(\lambda)$ si $\widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0$ et $\boldsymbol{\beta}_j \neq 0$, et fait une fausse découverte si $\widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0$ et $\boldsymbol{\beta}_j = 0$.

Maintenant, on peut définir le TPP du Lasso (True Positive Proportion ou Proportion de vrai positive) comme la proportion entre le nombre de vraies découvertes et le nombre maximum de vraies découvertes qui peuvent être faites. De même, le FDP du Lasso (False Discovery Proportion ou Proportion de fausses découvertes) est défini comme le taux entre le nombre de fausses découvertes et le nombre total de découvertes du Lasso. Formellement, on peut écrire les formulas pour le TPP et FPR comme suit :

$$\text{TPP}(\lambda) = \frac{\left|\left\{j : \widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0 \text{ et } \beta_j \neq 0\right\}\right|}{\max\left(\left|\{j : \beta_j \neq 0\}\right|, 1\right)} \quad \text{et} \quad \text{FDP}(\lambda) = \frac{\left|\left\{j : \widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0 \text{ et } \beta_j = 0\right\}\right|}{\max\left(\left|\left\{j : \widehat{\boldsymbol{\beta}}_j(\lambda) \neq 0\right\}\right|, 1\right)}. \tag{9}$$

À ce point de vue, on peut voir le FDP comme une mesure d'erreur de type I et $1 -$ TPP peut être considéré comme un erreur de type II.

La figure 4 montre un exemple du graphe des paires (TPP, FDP) le long du chemin de Lasso (en bleue). On observe que tant que λ augmente, le nombre de coefficients non nuls dans la solution de Lasso diminue à zéro, ce qui entraîne les FDP et TPP à zéro. Par contre, quand λ diminue, la solution de Lasso tend vers la solution de moindres carrées, qui contient majoritairement de coefficients non nuls. En conséquence, le TPP tend graduellement vers 1. On constate que le Lasso ne découvre que parfaitement les vrais coefficients des premiers coefficients. Puis jusqu'à TPP ≈ 0.18, il commence à découvrir de faux coefficients à la fois avec les vrais coefficients, en augmentant le FDP lors que le TPP croît à 1.

2 Présentation des principaux résultats

2.1 Résultat principal

On va maintenant présenter le résultat principal de l'article.

Théorème 1 *Soient $\delta \in (0, \infty)$ et $\epsilon \in (0, 1)$ et considérons la fonction $q^*(.) = q^*(.; \delta, \epsilon) > 0$ définie à l'équation (11). Alors, sous quelques hypothèses spécifiques qui seront données dans la section 3, et pour toutes petites constantes arbitraires $\lambda_0 > 0$ et $\eta > 0$, les conclusions suivantes tiennent :*

(a) L'évènement :

$$\bigcap_{\lambda > \lambda_0} \{\text{FDP}(\lambda) \geq q^*(\text{TPP}(\lambda)) - \eta\} \tag{10}$$

a une probabilité tendant à un dans les deux conditions $\sigma = 0$ (sans bruit) et $\sigma > 0$ (bruité).

*(b) La courbe limite q^* est **serrée**: pour toute courbe $q(u) \geq q^*(u)$ tel qu'il existe $q(u_0) > q^*(u_0)$, la conclusion (a) échouera pour une certaine distribution a priori Π sur les vrais coefficients de régression.*

La fonction q^* est appelée la courbe limite et est donnée par:

$$q^*(u, \delta, \epsilon) = \frac{2(1-\epsilon)\Phi\left(-t^*(u)\right)}{2(1-\epsilon)\Phi\left(-t^*(u)\right) + \epsilon u} \tag{11}$$

où $t^*(u)$ est la plus grande racine positive de l'équation suivante (dans la variable t) :

$$\frac{2(1-\epsilon)\left[(1+t^2)\Phi(-t) - t\phi(t)\right] + \epsilon(1+t^2) - \delta}{\epsilon\left[(1+t^2)(1-2\Phi(-t)) + 2\phi(t)\right]} = \frac{1-u}{1-2\Phi(-t)}, \tag{12}$$

et $\phi(\cdot), \Phi(\cdot)$ sont respectivement la densité et la fonction de répartition de la loi normale standard. On peut montrer que la fonction q^* est en fait lisse (de plus infiniment différentiable) sur son domaine, strictement croissante et s'annule à $u = 0$. Les exemples de la fonction q^* sont illustrés dans la figure 10.

On se concentre maintenant sur les deux affirmations du Théorème 1.

La première conclusion affirme que, pour n'importe quelle valeur de $\lambda > 0$, pour toutes les paires $(\text{TPP}(\lambda), \text{FDP}(\lambda))$, presque sûrement les points $(\text{TPP}(\lambda), \text{FDP}(\lambda))$ le long du chemin de Lasso vont être toujours situés au dessus de la courbe limite q^*. Celui-ci combiné avec les propriétés de q^* ci-dessus ($q^*(1) > q^*(0) = 0$) indique que l'on ne pourrait jamais trouver un $\lambda > 0$ pour que le Lasso trouve parfaitement tous les vrais coefficients (cela correspond au cas où $\text{TPP}(\lambda) = 1$ et $\text{FDP}(\lambda) = 0$).

La deuxième conclusion affirme la étroitesse ou *la sharpness* de la courbe limite. Dans le sens où pour tous les points $(u, q^*(u))$ sur la courbe, on peut approcher ce point aussi près que l'on veut en utilisant une distribution a priori Π sur les vrais coefficients de régression donnée par :

$$\Pi = \begin{cases} M, & \text{avec probabilité } \epsilon \cdot \epsilon' \\ M^{-1}, & \text{avec probabilité } \epsilon \cdot (1 - \epsilon') \\ 0, & \text{avec probabilité } 1 - \epsilon \end{cases} \tag{13}$$

pour un $\epsilon' = \epsilon'(u) > 0$ fixé. Notons que ϵ' est dépendant de u, et dans ce cas on peut montrer que :

$$\lim_{M \to +\infty} \lim_{n,p \to +\infty} (\text{TPP}(\lambda), \text{FDP}(\lambda)) \to (u, q^*(u)), \tag{14}$$

avec la convergence en probabilité.

2.2 Performance de la régularisation ℓ_0

Comme on a vu précédemment, les solutions de Lasso sont minorées par une fonction croissant dont la courbe ne permet pas les points $(\text{TPP}(\lambda), \text{FDP}(\lambda))$ de toucher le point $(1, 0)$. On va voir que ce n'est pas le cas pour un régularisation de norme ℓ_0. Formellement, on rappelle la formulation de cette méthode :

$$\widehat{\boldsymbol{\beta}}_0(\lambda) \in \underset{\mathbf{b} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|_0 . \tag{15}$$

Le théorème 2 s'annonce alors :

Théorème 2 *Sous quelques hypothèses spécifiques qui sont données dans la section 3, pour $\epsilon < \delta$, considérons la distribution a priori :*

$$\Pi = \begin{cases} M, & \textit{avec probabilité } \epsilon \\ 0, & \textit{avec probabilité } 1 - \epsilon \end{cases} . \tag{16}$$

Alors, on peut trouver $\lambda(M)$ tel que les découvertes de la méthode ℓ_0 données à l'équation (15) satisfait :

$$\lim_{M \to +\infty} \lim_{n,p \to +\infty} \text{FDP} = 0 \quad \textit{et} \quad \lim_{M \to +\infty} \lim_{n,p \to +\infty} \text{TPP} = 1. \tag{17}$$

Ce théorème nous dit que, pour tout ϵ et λ, on peut toujours approcher le point $(1, 0)$ d'autant proche que l'on veut. C'est à dire que l'on peut s'approcher à une découverte parfaite avec même une méthode qui donne les solutions creuses comme la méthode ℓ_0.

3 Prise de recul

Les résultats présentés dans l'article ne sont valables que sous certains conditions. Cette section discutera de ces hypothèses.

- **Hypothèses de base** :
  Dans tout ce travail on a supposé que $\mathbf{X}$ est une matrice de features dans $\mathbb{R}^{n\times p}$, les coefficients de régression $\boldsymbol{\beta} \in \mathbb{R}^p$ et le vecteur de bruit $\mathbf{z} \in \mathbb{R}^n$. Maintenant on suppose de plus que $\mathbf{X} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/n)$. Cela garantit que les colonnes de $\mathbf{X}$ sont approximativement normalisées. Également, les bruits sont i.i.d. $\mathcal{N}(0, \sigma^2)$ avec un σ fixé. L'article suppose aussi que les coefficients $\beta_1, \cdots, \beta_p$ sont des copies de variables aléatoires indépendantes d'une distribution a priori Π tel que $\mathbb{E}(\Pi^2) < \infty$ et pour une constante $\epsilon \in [0, 1]$, $\mathbb{P}(\Pi \neq 0) = \epsilon$. En gros, $\mathbf{X}, \beta, \mathbf{z}$ sont les variables indépendants. De plus, en grande dimension, on s'intéresse seulement au cas où le rapport $\dfrac{n}{p} \to \delta > 0$ quand $p, n \to \infty$.
- **Linear sparsity** :
  Les auteurs ajoutent une condition sur le degré de sparsity. En fait, le nombre attendu de coefficients de régression non nuls est linéaire en p et égal à $\epsilon \cdot p$ pour $\epsilon > 0$. Par conséquent, ce modèle s'oppose aux discussions asymptotiques. Par exemple, dans les problèmes de grande dimension, la proportion de les coefficients non nuls tendent vers zéro, ce qui n'est pas le cas ici.
- **Gaussian designs** :
  Cette conception gaussienne stipule que les colonnes sont indépendantes. Du coup, ils sont plus faciles à traiter.

4 Simulations numériques

4.1 Comparaison Lasso et Elastic Net

Dans cette partie, nous comparons le comportements du lasso et du elastic net vis à vis des TPP et des FDP.
La figure 4 étudie les performances du lasso et du Elastic net selon un plan gaussien aléatoire avec un nombre d'observation $\mathbf{n} = 1010$ et un nombre de features $\mathbf{p} = 1000$, où les entrées de $\mathbf{X}$ sont indépendants de loi $\mathcal{N}(0, 1)$. On prend $\beta_1 = \cdots = \beta_{200} = 4$, $\beta_{201} = ... = \beta_{1000} = 0$ et les erreurs suivent la loi normale centré réduite.
On remarque que le Lasso sélectionne les variables nulles assez tôt, du moins plus vite que l'Elastic Net. Il est certain que lorsque le Lasso inclut la moitié des vrais prédicteurs de sorte que la proportion de faux négatifs tombe en dessous de 50% ou que la proportion de vrais positifs (TPP) dépasse la barre des 50%.

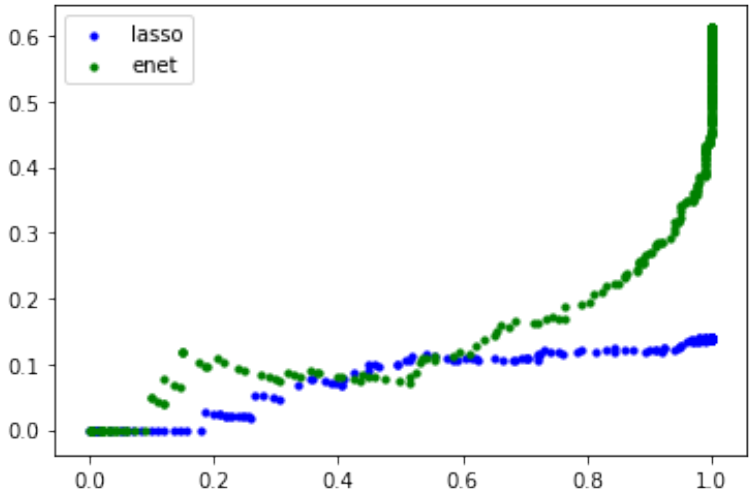


Figure 4: Représentation des FDP en fonction des TPP

Dans l'objectif d'approfondir plus la comparaison, une représentation des FDP et des TPP en fonction du coefficient de régularisation (α) a été faite. Globalement, l'évolution des vais découvertes positives reste le même suivant les deux type de régularisation (Lasso et Elastic Net). Cependant, une différence fondamentale est notée sur l'évolution des fausses découvertes positives.

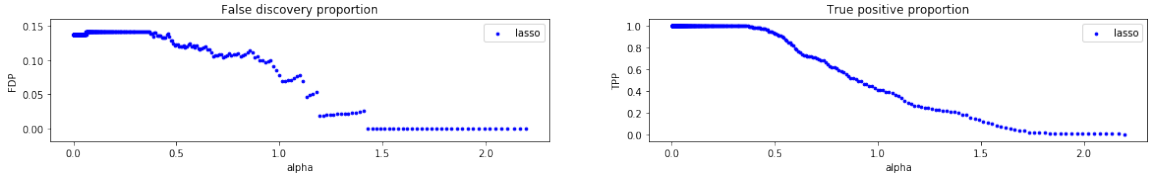


Figure 5: Représentation des FDP et des TPP en fonction de alpha dans le cadre du Lasso

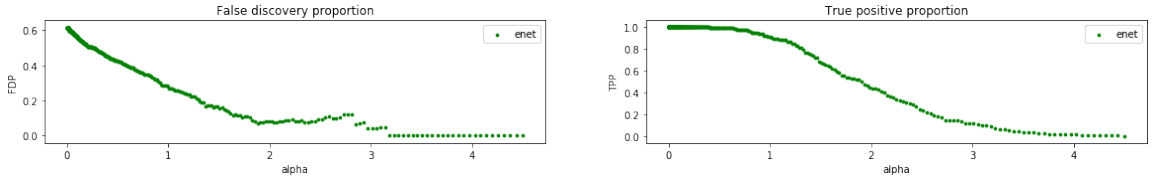


Figure 6: Représentation des FDP et des TPP en fonction de alpha dans le cadre du Elastic Net

La figure 7 et 8 présentent un examen plus approfondi de ce phénomène et résume les résultats de 100 des expériences indépendantes dans le cadre du même plan aléatoire gaussien. Nous pouvons confirmer, dans le cadre du lasso, que la première fausse découverte se produit relativement tôt, et que la proportion de fausses découvertes lorsque la proportion de vrais positifs atteint 1 est relativement élevée. Ce qui n'est pas le cas c'est l'Elastic Net.

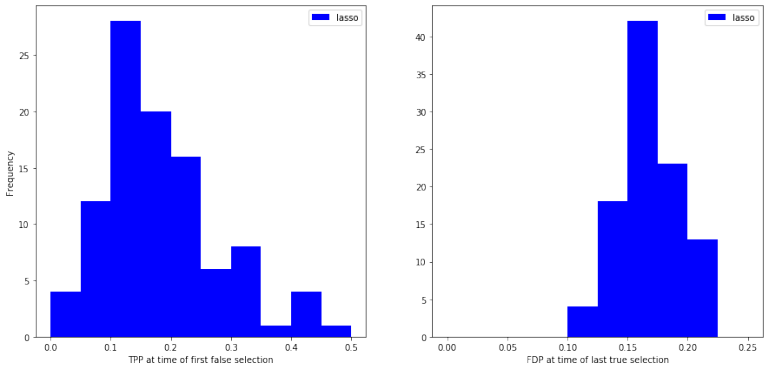


Figure 7: Lasso: TPP au moment de la première fausse sélection et FDP au moment de la dernière vraie sélection

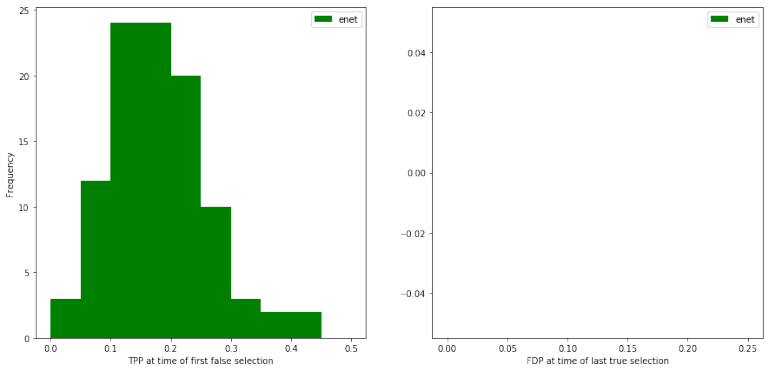


Figure 8: Elastic Net: TPP au moment de la première fausse sélection et FDP au moment de la dernière vraie sélection

4.2 Lasso trade off et Sharpness

La figure 9 montre le Lasso trade-off pour les deux cas différents. La limite entre le rouge et le blanc est assurée par la fonction q^*. La zone rouge est celle où les deux types d'erreurs sont faibles en d'autres terme un TPP élevé et un FDP faible.

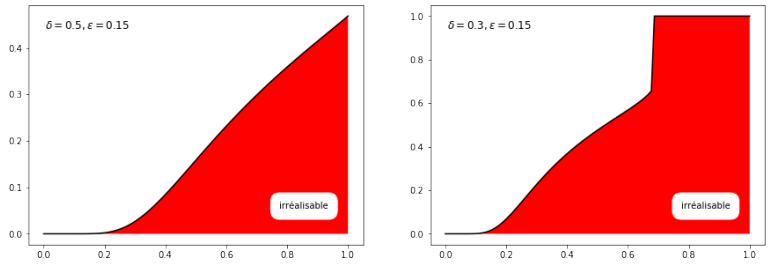


Figure 9: The Lasso trade-off diagram

Le fait que q^* est strictement en augmentation illustre le compromis asymptotique entre le FDP et le TPP. En fait, lorsque le TPP augmente la limite inférieure du FDP devient plus élevée, ce qui conduit à l'augmentation du FDP avec une probabilité tendant vers un. Pour mieux comprendre le comportement de la fonction, la figure 10 fournit une représentation du phénomène avec les différentes valeurs de sparity (ϵ) et de dimensionalité (δ). Une autre façon intéressante de comprendre ce compromis est de considérer le FDP comme une mesure de type I erreur, et 1 - TPP comme mesure de l'erreur de type II. Par conséquent, sur le chemin du lasso, les deux types de taux d'erreur ne peuvent pas être simultanément faible.

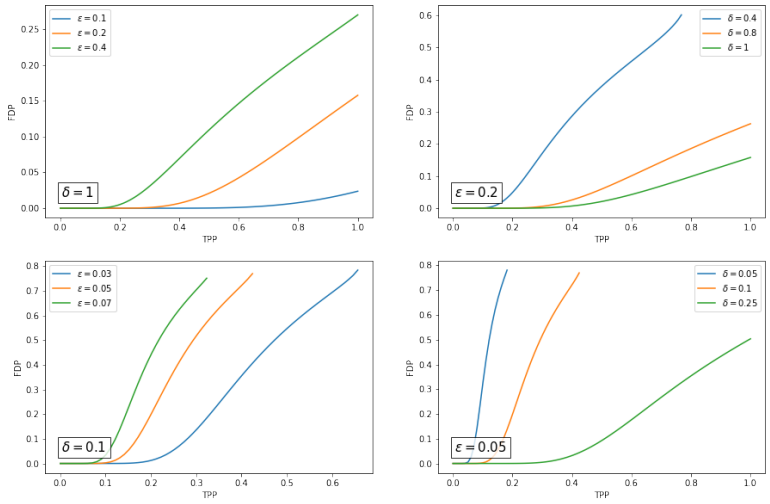


Figure 10: La fonction q^* pour les différentes valeurs de ϵ et δ

La figure 11 illustre les résultats de notre simulation pour les valeurs de n et p dans des conditions où $\sigma = 0$ et $\epsilon = 0.2$, avec $\mathbb{P}(\Pi = 1) = 1 - \mathbb{P}(\Pi = 0) = \epsilon$. Nous avons tracer toutes les paires (TPP, FDP) de 10 chemins de Lasso pour chacun des cas $n = p = 1000$ et $n = p = 5000$. On note que la grande majorité des paires (TPP,FDP) le long de ces 10 chemins se trouvent au-dessus de la frontière. De plus, le FDP moyen se rapproche de la limite alors que le TPP se rapproche de 1, une fraction des trajectoires se situe en dessous de la ligne également. On note aussi que quand n et p sont grands, les points se concentrent et se dispersent moins.

Les figures 12 - 13 montrent les propriétés de la sharpness de la courbe limite et illustrent l'équation (14). On est toujours dans des conditions où $\sigma = 0$, $n = p = 1000$ et $\epsilon = 0.2$, et on trace les courbes comme la moyenne de FDP par rapport à la moyenne de TPP sur 10 simulations chacun. La figure 12 montre les courbes pour les différentes valeurs de ϵ', où $\mathbb{P}(\Pi = 0) = 1 - \epsilon$ et $\mathbb{P}(\Pi = 50 | \Pi \neq 0) = 1 - \mathbb{P}(\Pi = 0.1 | \Pi \neq 0) = \epsilon'$. On note que pour différentes valeurs de ϵ', les courbes moyennes touchent la courbe limite aux différents endroits. La figure 13 illustre les courbes pour les différentes valeurs de M, avec $\epsilon' = 0.5$. On note que lors que M augment, la courbe moyenne touche la courbe limite de plus en plus.

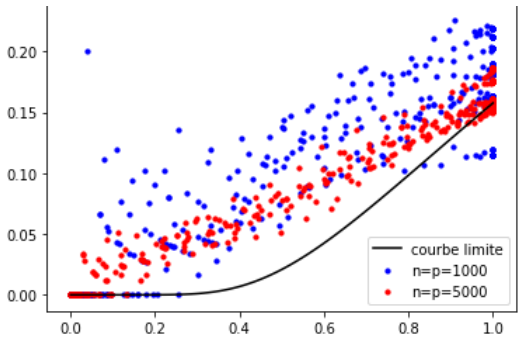


Figure 11: Paires réalisées (TPP, FDP)

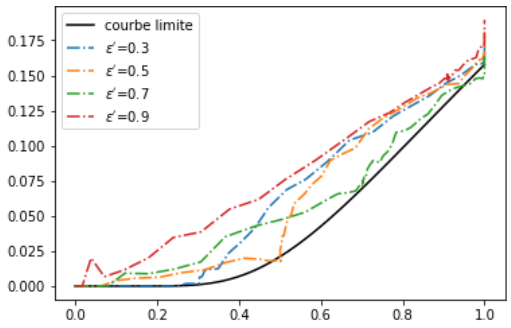


Figure 12: Sharpness de la courbe limite pour différents valeurs de ε'

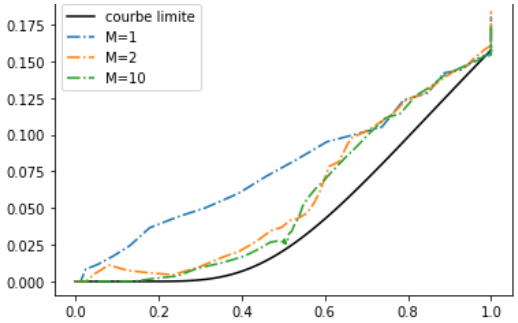


Figure 13: Sharpness de la courbe limite pour différents valeurs de M

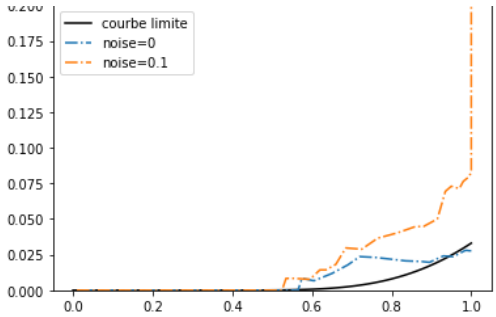


Figure 14: Sharpness de la courbe limite pour différents niveaux de bruit

La dernière figure 14 montrent les cas avec et sans bruit. On prend également $n = p = 1000$, et $\epsilon = 0.2$, avec $\mathbb{P}(\Pi = 9.3) = 1 - \mathbb{P}(\Pi = 0) = \epsilon$. On voit que tous les deux courbes moyennes sont au-dessus de la courbe limite. On voit également que la courbe avec bruit est plus loin de la courbe limite que le cas sans bruit, indiquant que c'est plus difficile à apprendre avec les bruits. En fait, on a trouvé après plusieurs expériences que les bruits donnent parfois très bons résultats mais des fois ils donnent les mauvais résultats comme dans la figure 14, montrant que le Lasso n'est pas très robuste avec bruits.

Conclusion

En définitive, sous les hypothèses sous-jacentes, le lasso n'a pas toujours réussi à sélectionner des variables importants avec un taux d'erreur minimal. Nous avons montré la relation entre la proportion de fausses découvertes (FDP) et le taux de faux négatifs (1-TPP). En effet, nous avons souligné le compromis qui consiste dans le fait que nous ne pouvons pas augmenter le TPP sans augmenter FDP.

Nous avons comparé le lasso et l'Elastic Net qui combine le lasso et la régularisation ridge pour voir comment il parvient à surmonter les différentes limitations, notamment le bruit de rétrécissement et la corrélation.

L'analyse de cet article a été une grande récompense, elle nous a permis de mieux comprendre et de cerner les limites du lasso et de le comparer aux autres problèmes d'optimisation.