# Theoretical Guidelines for High-Dimensional Data Analysis

–

# Projet

Cédric Allain
Clotilde Miura

10 février 2020

## Table des matières

# 1 Résumé de l'article

## 1.1 Contexte

Dans le contexte récent de ces dernières années, à l'heure où les données sont de plus en plus massives, les algorithmes prédictifs mettent en jeu un nombre de variables de plus en plus élevé. On pense par exemple aux applications médicales, dans la prédiction du cancer. Dans ce cas, seul un nombre très faible de gènes dans le génome humain (environ 20 000) sont responsables de l'expression d'un cancer.

Ce sont des problèmes dits *sparses*. Le Lasso, implémenté pour la première fois en 1996 par Robert Tibshirani [3], s'est naturellement imposé dans la résolution de ce type de problèmes. Même s'il est aujourd'hui en passe d'être détrôné par les réseaux de neurones, il est encore largement utilisé, notamment pour son interprétablité par rapport aux « boîtes noires » que représentent actuellement les réseaux de neurones profonds. On lui concède de très bonnes performances dans les problèmes de régression où les variables explicatives sont très peu corrélées et où seul un faible nombre d'entre elles, les *signaux*, ont un coefficient non nul et à forte amplitude.

L'article que nous présentons, Su et al. 2017, *False discoveries occur Early on the Lasso path* [2] met en cause le paradigme selon lequel dans ces conditions, le Lasso est capable de trouver tous les signaux avec très peu d'erreurs voire aucune. En effet, les auteurs démontrent qu'il existe toujours un arbitrage (*trade off* en anglais) entre le taux de vrais positifs (TPP) et le taux de faux positifs/fausses découvertes (FDP), y compris dans le cas de variables indépendantes.

Nous commencerons par exposer dans cette partie le résumé de l'article, puis dans une deuxième partie, nous en établirons une critique avant de proposer une implémentation en Python afin de reproduire les résultats de l'article.

## 1.2 Présentation du cadre de travail

Les auteurs considèrent le cas du modèle linéaire standard où l'on cherche à prédire une réponse $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}$, où $\boldsymbol{y} \in \mathbb{R}^n$ est la variable dépendante à prédire, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ est la matrice de variables, *features*, et où $\boldsymbol{z}$ est un bruit. L'estimateur Lasso est alors la solution du problème de minimisation suivant :

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda \|\boldsymbol{b}\|_1 \tag{1}$$

avec $\lambda$ contrôlant le poids de la pénalisation de la norme $l_1$. Plus $\lambda$ est grand, plus on encourage le modèle à être parcimonieux. Dans le cas où les variables sont indépendantes, i.e. la matrice de covariance $X^T X = I_p$, on a une expression explicite de l'estimateur Lasso :

$$\forall j = 1, 2, ...p \qquad \hat{\beta}_j(\lambda) = (X_j^T Y - \lambda \operatorname{sign}(X_j^T Y)) \mathbb{1}_{\left\{|X_j^T Y| \geq 0\right\}} \tag{2}$$

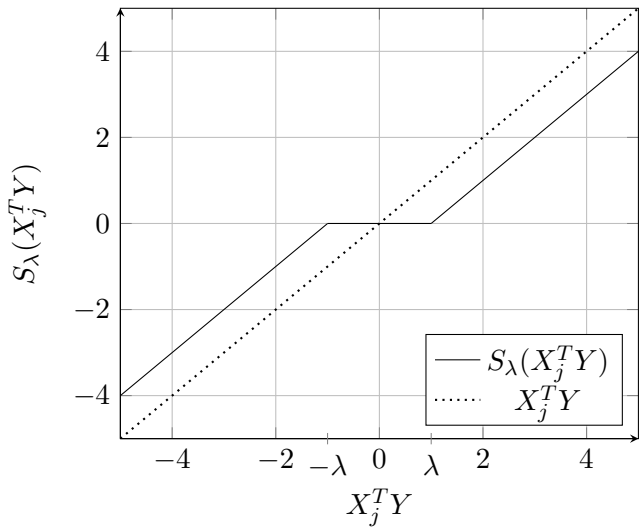



FIGURE 1 – Fonction *soft-thresholding* ($\lambda = 1$)

Des résultats théoriques assurent que le Lasso se comporte bien dans les régimes extrêmement asymptotique [4, 5, 1]. Mais on a souvent tendance à généraliser ce résultat aux applications « pratiques » et à imputer de mauvaises performances à des variables trop corrélées ou aux *small size effects*, c'est-à-dire lorsque $n/p$ est petit.

Les résultats principaux des auteurs démontrent que même dans le cas où le SNR (*signal to noise ratio*) $\to +\infty$ et où les variables sont stochastiquement indépendantes, il existe un arbitrage entre TPP le taux de vrais positifs et FDP la proportion de fausses découvertes. Il convient donc de définir proprement ces deux notions ce que les auteurs font de la manière suivante.

Soit k $= |\{j : \beta_j \neq 0\}|$ le nombre de vrais signaux. Soit $V(\lambda) = |\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ et } \beta_j = 0\}|$ et $T(\lambda) = |\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ et } \beta_j \neq 0\}|$ respectivement le nombre de fausses découvertes et de vraies découvertes du Lasso. Finalement, on définit FDP($\lambda$) et TPP($\lambda$) par :

$$\text{FDP}(\lambda) = \frac{V(\lambda)}{\left|\left\{j : \widehat{\beta}_j(\lambda) \neq 0\right\}\right| \vee 1} \tag{3}$$

$$\text{TPP}(\lambda) = \frac{T(\lambda)}{k \vee 1} \tag{4}$$

Afin de mener leurs expériences, les auteurs choisissent un cadre de travail simple :

— Les features sont **stochastiquement indépendantes** et **gaussiennes** ;
— Le modèle est **linéairement sparse**, i.e. l'espérance du nombre de coefficients non nuls est linéaire en $p$ et égale à $\epsilon \cdot p$ avec $\epsilon > 0$ ;

— Enfin, les coefficients $\beta_1, \ldots, \beta_p$ sont des réalisations indépendantes d'une variable aléatoire Π qui vaut M avec probabilité $\epsilon$ et 0 sinon. $\epsilon$ est donc la proportion de signaux non nuls, $\epsilon \approx k/p$ et M est la valeur des signaux non nuls qu'on suppose tous de même amplitude.

Dans ces conditions largement reconnues favorables au Lasso, les auteurs démontrent que cet estimateur peut néanmoins faire des fausses découvertes. Nous verrons dans la sous partie suivante que ce trade-off peut être formalisé, et qu'il dépend du degré de sparsité du problème $\epsilon$ et de la dimensionnalité du problème $1/\delta$.

## 1.3 Principaux résultats

### Trade off entre TPP et FDP

Avant de formaliser leurs résultats, les auteurs présentent un cas où l'on peut observer le trade off entre le FDP et le TPP, présenté en Figure 2. L'expérience est menée dans le cadre présenté dans la section précédente avec $n = 1010$, $p = 1000$, chaque élément de la matrice $X$ simulé suivant une loi $\mathcal{N}(0, 1)$, $k$ le nombre de vrais signaux égal à 200 et $\beta_1 = \beta_2 = \cdots = \beta_k = 4$ ($M = 4$).

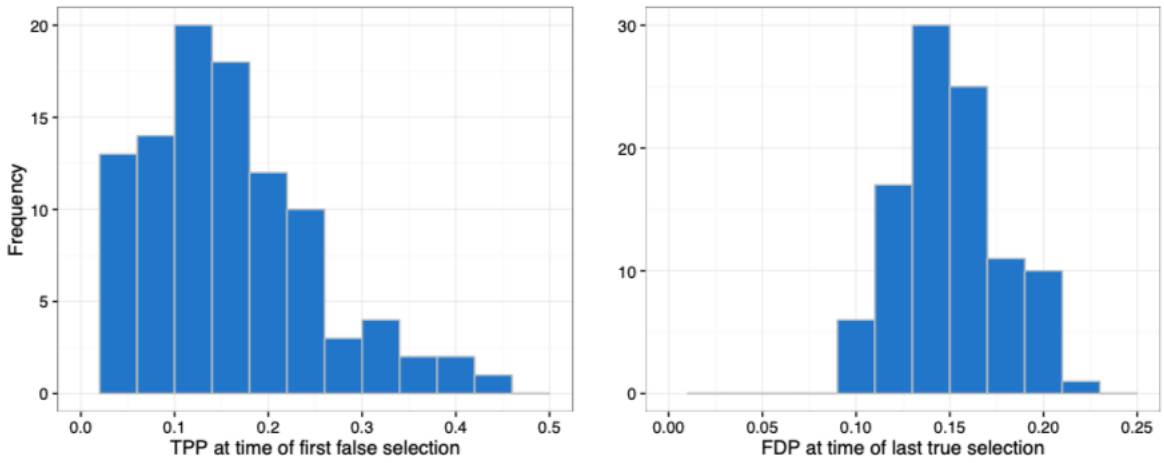



FIGURE 2 – Gauche : taux de vrais positifs au moment de l'apparition de la première fausse découverte. Droite : proportion de fausses découvertes lorsque le taux de vrais positifs atteint 1 (i.e. lorsque la proportion de faux négatifs atteint 0).
Source : Su et al., 2017 [2]

Les auteurs effectuent ainsi 100 simulations de ce modèle et les histogrammes de distribution du taux de vrais positifs au moment de la première fausse découverte et du taux de fausses découvertes au moment où TPP atteint 1 sont présentés respectivement à gauche et à droite de la figure 2. En extrayant la moyenne de ces histogrammes, les auteurs peuvent affirmer qu'en moyenne, pour que le

Lasso sélectionne tous les vrais signaux, il doit « payer » 15 % de fausses découvertes (histogramme de droite dans la Figure 2).

On peut aussi prendre l'approche suivante : au moment où la première fausse découverte est faite, le Lasso n'a trouvé en moyenne que 16 % des vraies variables (histogramme de gauche dans la Figure 2). Ainsi, même dans un cadre a priori idéal et simple (indépendance des variables, problème linéairement sparse), le Lasso échoue à sélectionner toutes les bonnes variables sans erreurs. C'est le trade off entre TPP et FDP ou, d'un point de vue plus statistique, le trade off entre erreur de type I, le taux de faux positifs (FDP), et erreur de type II (1-TPP), le taux de faux négatifs.

Nous verrons ensuite comment les auteurs déterminent une frontière délimitant les couples (TPP, FDP) atteignables des autres, qui correspondraient à de meilleurs scénarios. Il y a donc toujours un compromis à faire entre FDP et TPP qui peut être plus ou moins contraignant suivant le conditionnement du problème en $n$, $p$ et $\epsilon$.

**Frontière du Lasso path**

Nous allons maintenant discuter comment les auteurs modélisent l'arbitrage entre FDP et TPP dans le théorème 1, le résultat central du papier. Le résultat suivant est valable pour les cas avec ou sans bruit ($\sigma \geq 0$). Pour un $\delta = n/p$ et $\epsilon = k/p$ fixés :

$$\mathbb{P}\left(\bigcap_{\lambda \geq \lambda_0} \{\text{FDP}(\lambda) \geq q^\star(\text{TPP}(\lambda)) - \eta\}\right) \xrightarrow{\lambda_0} 1$$

pour tout $\lambda_0$ et $\eta$ arbitrairement petits et pour $q^*(.) = q^*(., \delta, \epsilon)$ fonction définie par les auteurs dans l'équation (2.4) de l'article et appelée *frontière*.

Intuitivement, cette équation montre que peu importe la valeur de $\lambda$, il y a un prix $FDP(\lambda)$ à payer pour avoir un certain $TPP(\lambda)$ et que ce prix croit avec cette valeur du fait de la stricte croissance de $q^*$. Il peut également être plus ou moins élevé suivant la forme de la frontière qui dépend de la dimensionnalité du problème $1/\delta = p/n$ et de la sparsité $\epsilon$ [1] Afin d'étudier l'impact de ces paramètres sur la frontière $q^*$, on étudie les Figures 3 et 4 qui représentent les *trade off diagrams*.

La partie colorée en rouge de la Figure 3 représente les couples (TPP, FDP) qu'aucune valeur de $\lambda$ ne permet d'atteindre, c'est-à-dire les points (TPP, FDP) tels que FDP est inférieur à $q^*(TPP)$ [2] le prix minimum à payer en faux positifs pour atteindre le taux de vrais positifs TPP. (On omet le paramètre $\lambda$ pour alléger les notations). Le graphe de droite présente même un cas où il est impossible pour le Lasso de sélectionner toutes les vraies variables ie d'atteindre un TPP de 1. Le

1. À noter que $\epsilon$ est la proportion de coefficients non nuls, la sparsité peut donc être représenté par $1 - \epsilon$. On décide cependant de garder les mêmes appellations que les auteurs [2].

2. Donc $\eta = 0$

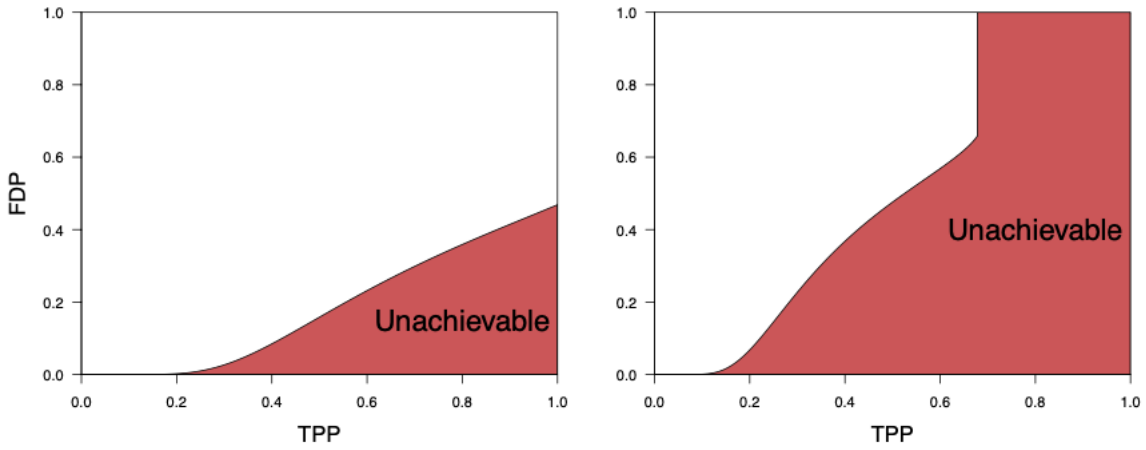



FIGURE 3 – Diagramme du trade off Lasso. Gauche : avec $\delta = 0.5$ et $\epsilon = 0.15$. Droite : avec $\delta = 0.3$ et $\epsilon = 0.15$ (la troncature verticale se produit à 0.6791).
Source : Su et al., 2017 [2]

taux maximal que l'on puisse atteindre est en effet de seulement 0.68 %. Les deux graphes ayant été réalisés sur des données ayant la même sparsité ($\epsilon = 0.15$), on en déduit que c'est la dimensionnalité $1/\delta = p/n$ qui est en cause. Comme elle est plus élevée dans le graphe de droite ($1/\delta = 10/3$) que dans le graphe de gauche ($1/\delta = 2$), il semble que plus la dimensionnalité augmente, plus le compromis à faire entre FDP et TPP est contraignant.

Ceci est confirmé par la Figure 4. Les graphes de gauche montrent que lorsqu'on fixe $\delta$, augmenter la proportion de vrais signaux $\epsilon$ dégrade la frontière $q^*$ au sens où le trade off est moins favorable. Les graphes de droite montrent qu'à $\epsilon$ fixé, augmenter la dimensionnalité du problème, i.e. diminuer $\delta$, dégrade également la frontière.

Ainsi, moins le modèle est sparse (plus le nombre de signaux est important) et plus la dimensionnalité augmente (plus il y a de variables dans le modèle par rapport au nombre d'observations), plus les scénarios s'aggravent et moins l'on peut espérer un bon taux de vrais positifs avec un nombre minimal d'erreurs. Ces paramètres sont vraiment déterminants puisqu'il existe d'énormes disparités entre les frontières. Par exemple dans le cas où $\delta = 1$ et $\epsilon = 0.1$ en haut à gauche, on a un scénario très favorable où il est possible d'atteindre un TPP maximal de 1 tout en concédant un très faible taux de fausses découvertes (0.02 %). A contrario, en regardant la courbe pour $\delta = 0.05$ et $\epsilon = 0.05$ (courbe verte du graphe en bas à droite), on remarque qu'aucune valeur de $\lambda$ ne permet de dépasser un taux de vrais positifs supérieur à 0.2 % tout en payant un coût relativement élevé en termes de FDP.

Étudions un peu plus en détails la frontière $q^*$. Le dernier point du théorème 1 dit que cette frontière est *tigh*, c'est à dire que c'est la plus petite fonction continue qui satisfasse les conditions (a) et (b) du théorème 1. On ne peut donc pas définir une fonction $\tilde{q}$ inférieure à $q^*$ qui augmenterait le

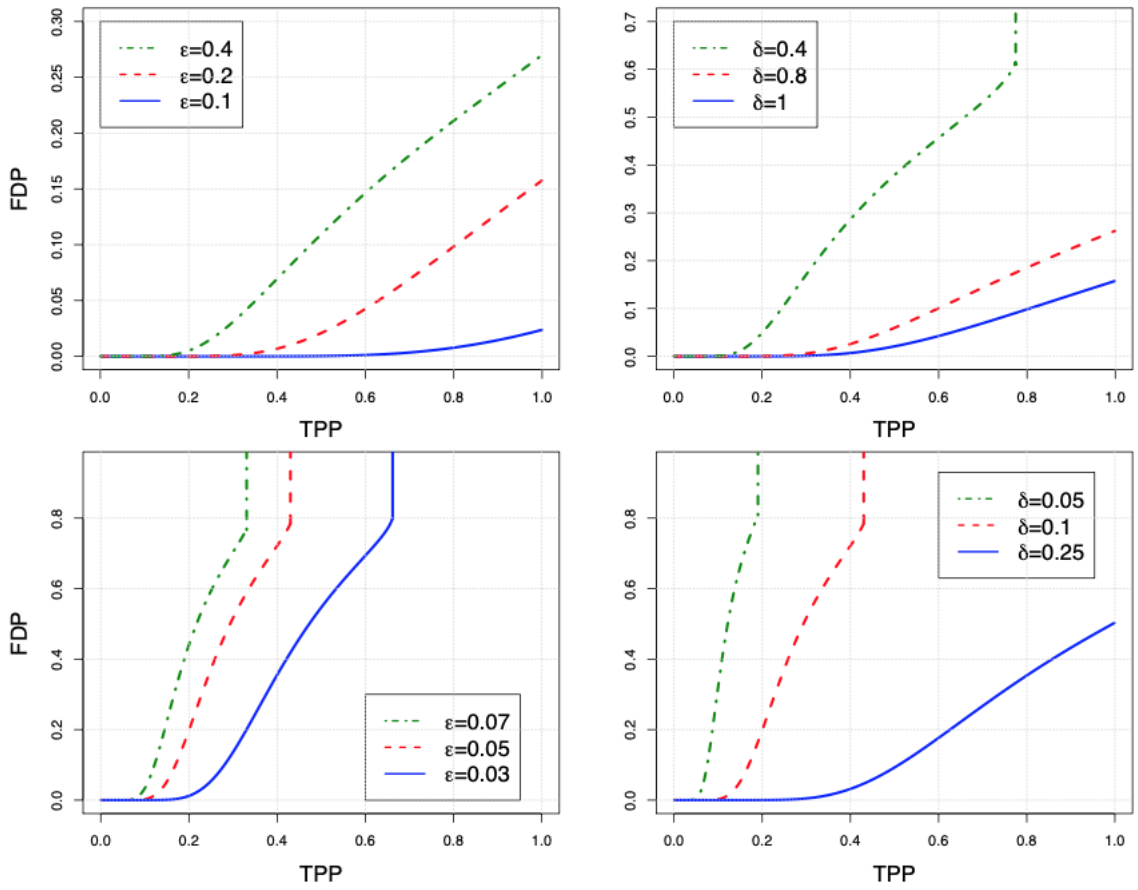



FIGURE 4 – Supérieur gauche correspond à $\delta = 1$ ; Supérieur droit correspond à $\epsilon = 0.2$ ; inférieur gauche correspond à $\delta = 0.1$ ; et inférieur droit à $\epsilon = 0.05$.
Source : Su et al., 2017 [2]

nombre de paires (TPP, FDP) possibles. Les paires les plus favorables étant situées sur la frontière, on peut se demander si il existe des cas où l'on peut la toucher. En affinant la prior des coefficients $\beta_j, j = 1, 2, \ldots, p$ en introduisant $\epsilon'$ la proportion de *strong signals* au sein des coefficients non nuls, on peut écrire :

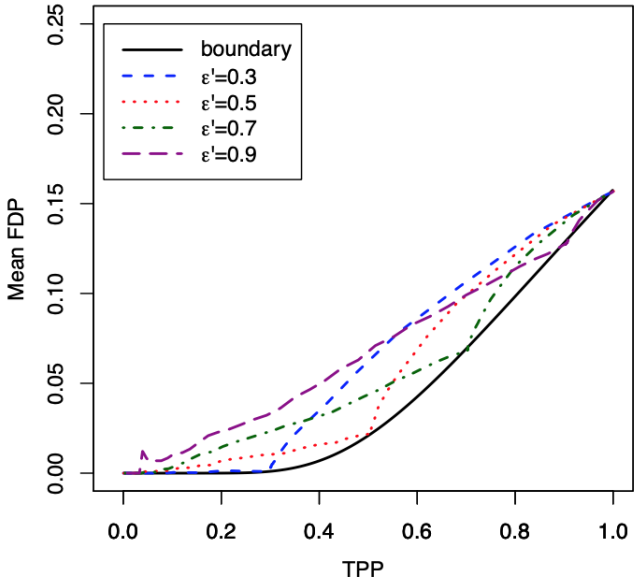
$$\Pi \sim \epsilon\epsilon'\delta_M + \epsilon(1-\epsilon')\delta_{M^{-1}} + (1-\epsilon)\delta_0$$

avec $\delta_x$ la mesure de dirac en $x$.

L'équation (2.5) de l'article montre que chaque point de la courbe $q^*$ peut être approché avec probabilité tendant vers 1 par $\lim_{M\to\infty} \lim_{n,p\to\infty}(\text{TPP}(\lambda), \text{FDP}(\lambda))$ pour un certain $\epsilon' > 0$. Ce résultat est cependant valable sous réserve que $\lambda(M) \underset{M\to\infty}{=} o(M)$ . Cette équation démontre que l'on peut atteindre n'importe quel point de la frontière avec une configuration de mélange entre signaux très forts où $\beta_j = M$ (présents en proportion $\epsilon'$ parmi les signaux non nuls) et signaux très

faibles où $\beta_j = 1/M$. Il faut de plus choisir un $\lambda$ suffisamment élevé par rapport à l'amplitude du signal.

La figure 5 montre, en fonction de $\epsilon'$, le point de la frontière que l'on peut atteindre. Les auteurs ont réalisés 100 simulations identiques où $n = p$ ($\delta = 1$), dans le cas sans bruit et avec un taux de signaux égal à 20 %. Ils ont ensuite calculé les TPP et FDP moyens pour plusieurs valeurs de $\lambda$ et tracé 4 trajectoires de paires de points (TPP($\lambda$), FDP($\lambda$)) pour 4 valeurs de $\epsilon'$ différentes. On remarque que plus la proportion de forts signaux est élevée, plus le scénario optimal à atteindre est situé à un point de la frontière biaisé vers les valeurs élevées de TPP (courbe violette pour $\epsilon' = 0.9$). En revanche, si la proportion de forts signaux est faible, le « meilleur » scénario est un point de la courbe qui arbitre en faveur d'un FDP très faible (courbe bleue pour $\epsilon' = 0.3$). On en déduit donc que la proportion de forts signaux est un paramètre qui détermine le point de la frontière atteignable par le Lasso. Nous allons maintenant étudié dans la partie suivante le *shrinkage noise*, identifié par les auteurs comme étant la principale cause des fausses découvertes du Lasso.




**(b)** Sharpness of the boundary

FIGURE 5 – $n/p = \delta = 1, \epsilon = 0.2$, et le niveau de bruit $\sigma = 0$ (noiseless). FDP moyen en fonction du TPP moyen pour différentes valeurs de $\lambda$ sur 100 simulations pour $n = p = 1000$, $\mathbb{P}(\Pi = 0) = 1 - \epsilon$ comme avant, et $\mathbb{P}(\Pi = 50|\Pi \neq 0) = 1 - \mathbb{P}(\Pi = 0.1|\Pi \neq 0) = \epsilon'$
Source : Su et al., 2017 [2]

## 1.4 Le *shrinkage noise* : principale cause de la sélection de fausses découvertes par le Lasso

.

**Pourquoi y-a-il un shrinkage des coefficients ?**

Pour répondre à cette interrogation, nous allons comparer l'estimateur Lasso à l'estimateur $l_0$ qui est solution du problème de minimisation suivant :

$$\widehat{\boldsymbol{\beta}}_0(\lambda) = \underset{\boldsymbol{b}\in\mathbb{R}^p}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2 + \lambda\|\boldsymbol{b}\|_0 \tag{5}$$

Le problème du Lasso présenté en équation 1 est en réalité une relaxation convexe de ce problème. Dans le cadre de travail adopté par les auteurs, l'estimateur $l_0$ permet de trouver sans aucune erreur tous les vrais signaux du problème et ce peu importe leur amplitude $M$. C'est ce que démontre le théorème 2 du papier. On peut en effet trouver un $\lambda(M)$ tel que :

$$\lim_{M\to\infty}\lim_{n,p\to\infty} \mathrm{FDP}(\lambda) = 0 \quad \text{and} \quad \lim_{M\to\infty}\lim_{n,p\to\infty} \mathrm{TPP}(\lambda) = 1 \tag{6}$$

De plus, on a même une valeur analytique de cet estimateur, aussi appelé estimateur *hard thresholding* vue en cours dans le cas où les variables sont indépendantes. Pour tout $j = 1, 2 \dots p$ :

$$\hat{\beta}_0(\lambda) = X_j^T Y \mathbb{I}_{(|X_j^T Y| \geq \sqrt{\lambda})} \tag{7}$$

En traçant les deux estimateurs sur le même graphe (6), on voit bien l'effet de shrinkage des coefficients du Lasso. La courbe rouge prend des valeurs plus élevées que la courbe noire et plus $\lambda$ est élevé, plus le shrinkage est fort. On note cependant que ceci n'est le cas que lorsque $\sqrt{\lambda} \leq \lambda$, c'est-à-dire lorsque $\lambda \geq 1$. Ceci ne remet pas en cause l'affirmation des auteurs, puisque si l'on suppose que le modèle est suffisamment sparse (i.e. $\epsilon$ suffisamment petit). Il faut alors un $\lambda$ suffisamment grand pour qu'il n'y ait pas trop de fausses découvertes.

**En quoi le shrinkage est-il problématique ?**

Les auteurs soutiennent que le shrinkage des coefficients cause un bruit qui éclipse le signal des vrais signaux et peut conduire l'estimateur à sélectionner de mauvaises variables. Les résidus contiendraient en effet encore de nombreux effets liés aux variables sélectionnées. En effet, soit $j \in \{1, \dots p\}$, et $\hat{\epsilon} = y - X\hat{\beta}_\lambda$ les résidus du modèle. Si l'on projette $X_j$ sur la j$^{\text{e}}$ colonne de X, notée

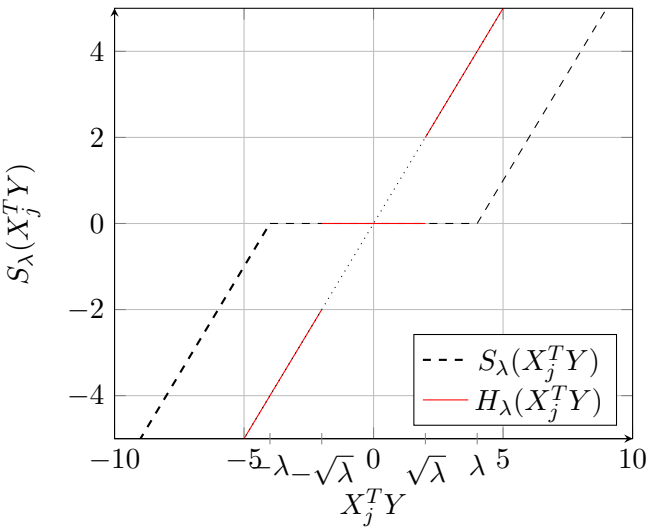



FIGURE 6 – Fonction *soft-thresholding* et *hard thresholding* ($\lambda = 4$)

$X_j$, on obtient :

$$\begin{aligned}
\hat{\epsilon}^T X_j &= X_j^T Y - X_j^T X \hat{\beta}(\lambda) \\
&= X_j^T Y - (0, \dots, \underbrace{||X_j||^2}_{j}, 0, \dots, 0)^T \hat{\beta}(\lambda) \\
&= X_j^T Y - e_j^T \hat{\beta}(\lambda) \text{ car } X^T X = I_p \text{ et donc } ||X_j||^2 = 1 \\
&= X_j^T Y - \hat{\beta}_j(\lambda)
\end{aligned}$$

Si on suppose $X_j^T Y$ positif, alors $\hat{\epsilon}^T X_j = \lambda$ si $\hat{\beta}_j(\lambda) \neq 0$ et $X_j^T Y$ sinon (*cf.* équation 2). Donc, comme expliqué dans l'article, si la variable est sélectionnée, on a bien dans les résidus un effet qui croit avec $\lambda$, la régularisation. C'est ce phénomène que les auteurs définissent comme le shrinkage noise. Plus il y a de variables sélectionnées, plus le bruit est fort et est susceptible d'éclipser les véritables signaux, conduisant à sélectionner à tord de fausses découvertes.

C'est également pour cela que le régime de l'extrême sparsité est plus favorable, puisque moins de variables sont sélectionnées et donc il y a moins de bruit dû au shrinkage des coefficients.

## 2 Quelques critiques possibles

Dans cette sous section, nous allons émettre quelques critiques d'ordre pratique puis théorique.

## 2.1 Critiques pratiques

Les auteurs ne sont pas toujours très clairs quant à l'implémentation de leurs expériences, ce qui peut compliquer la reproductibilité des résultats. Sur le lien du repo github fourni [3], seul le code de la frontière $q^*$ est fourni. De plus, les auteurs n'explicitent pas la plage de valeurs de $\lambda$ utilisée pour générer le Lasso path, ni même l'algorithme utilisé. Sur `scikit-learn`, deux implémentations du Lasso sont en effet possibles, l'une utilisant l'algorithme LARS (*Least Angle Regression*), et l'autre utilisant une descente proximale de coordonnées. Nous avons choisi ce dernier algorithme pour nos expériences dans la partie suivante puisque dans le cas de l'indépendance des variables où $X^T X = I_p$ (en espérance), il mène à la solution exacte du Lasso en une itération, comme nous le démontrons ci-dessous.

On rappelle tout d'abord le problème du Lasso :

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \underbrace{\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2}_{\mathcal{L}(\beta)} + \lambda \left\|\boldsymbol{\beta}\right\|_1$$

Soit $\beta^0(\lambda)$ initialisé par un vecteur positif quelconque. Avec un pas égal à 1, à l'itération suivante, l'algorithme de descente proximale de coordonnées donne pour la variable $j$ :

$$\beta_j^1(\lambda) = S_\lambda\left(\beta_j^0(\lambda) - \nabla_\beta\left(\beta^0(\lambda)_j\right)\right) \tag{8}$$

où $S_\lambda$ est l'opérateur Soft thresholding et $\nabla_\beta(\beta^0(\lambda)_j)$ est la $j^e$ coordonnée du gradient.

$$\nabla_\beta\left(\beta^0(\lambda)\right) = -X^T\left(Y - X\beta^0(\lambda)\right)$$

Donc, la $j^e$ coordonnée du gradient est :

$$\begin{aligned}\nabla_\beta\left(\beta^0(\lambda)_j\right) &= -X_j^T\left(Y - X\beta^0(\lambda)\right) \text{ où } X_j \in \mathbb{R}^n \text{ est la } j^{\text{e}} \text{ colonne de X} \\ &= -X_j^T\left(Y - \sum_{k=1}^{p} X_k\beta_k^0(\lambda)\right)\end{aligned}$$

Or, par indépendance des features, $\forall j \neq k, X_j^T X_k = 0$. Donc :

$$\begin{aligned}\nabla_\beta(\beta^0(\lambda)_j) &= -X_j^T Y + \underbrace{||X_j||^2}_{1}\beta_j^0(\lambda) \\ &= -X_j^T Y + \beta_j^0(\lambda)\end{aligned}$$

3. `https://github.com/wjsu/fdrlasso`

Donc :

$$\begin{aligned}\beta_j^1(\lambda) &= S_\lambda(\cancel{\beta_j^0(\lambda)} + X_j^T Y - \cancel{\beta_j^0(\lambda)}) \\ &= S_\lambda(X_j^T Y)\end{aligned}$$

On retombe bien sur la solution exacte du Lasso avec la descente proximale de coordonnées.

Une deuxième critique peut porter sur la rigueur scientifique des résultats souvent présentés sans *standard deviation*. On donne en général une moyenne sur un ensemble de simulations mais accompagné des standard déviations. C'est par exemple le cas pour la Figure 6 de l'article, dans laquelle, pour le graphique de droite, les auteurs tracent la FDP moyenne en fonction de TPP pour 500 simulations, à la fois avec et sans bruit. Il aurait été intéressant de reporter également les écarts-types afin de pouvoir juger de la significativité des courbes obtenues.

Enfin, en partie 2.6, les auteurs évoquent le cas où $n \geq p$ en affirmant que même dans cette situation, il est impossible de retrouver tous les signaux sans fausses découvertes. Ils n'ont cependant testé en Figure 4 que le cas limite où $n = p$. Dans la partie suivante, nous tentons donc d'aller plus loin et de vérifier leur affirmation pour $n > p$.

## 2.2 Critiques de contenu

Il y a une certaine ambiguïté autour du paramètre $\delta$ qui est en réalité la limite du rapport $n/p$ quand $n, p \to +\infty$. Ainsi, la frontière $q^*$ du théorème 1 est en réalité une frontière asymptotique. Cela explique que dans la Figure 5(a) de l'article, on ait des points en dessous de la frontière dans le régime $n = p = 1000$. On remarque que cette fraction se réduit amplement dans le cas où $n = p = 5000$, autrement dit lorsqu'on augmente la taille du problème. Ainsi la frontière $q^*$ n'est en réalité que théorique, mais l'on peut supposer que lorsque la taille de l'échantillon ainsi que le nombre de variables sont « assez grands », on puisse s'en rapprocher suffisamment pour qu'elle puisse constituer une séparation parfaite entre les paires (TPP, FDP) atteignables et les autres. De plus, l'hypothèse que les signaux forts aient tous la même amplitude notée $M$ est peut être un peu trop réductrice puisqu'en pratique, même après normalisation des données, on peut observer de fortes différences de coefficients entre les variables.

## 2.3 Remarques personnelles et conclusion sur l'article

Enfin, d'un point de vue plus personnel, si nous avons dans l'ensemble trouvé l'article bien écrit et accessible, nous avons cependant eu du mal à comprendre la partie sur le *shrinkage noise*. N'étant pas clairement formalisé mathématiquement, nous n'avons pas clairement saisi comment il induisait le Lasso en erreur. Enfin, l'irrégularité du Lasso path nous semble assez contre intuitive. Il nous semble

étrange, au vu de l'expression de l'estimateur soft thresholding dans le cas indépendant (équation 2), qu'une variable puisse entrer puis ressortir du Lasso path. Néanmoins le principal message de l'article, le *take home message*, est très bien expliqué et a le mérite de briser une croyance bien ancrée mais pourtant erronée : même dans un régime où les variables ne sont pas corrélées et où le signal est très fort par rapport au bruit, le Lasso n'est pas infaillible.

# 3 Implémentation et reproductibilité des résultats

Dans cette section, on se propose dans un premier temps d'essayer de reproduire les résultats obtenus par les auteurs, pour à la fois mettre en lumière les éventuelles manquements à une bonne reproductibilité des résultats, mais surtout pour pouvoir mettre en perspective nos explorations. En effet, on souhaite dans un second temps explorer des pistes de façon empirique, pour éventuellement formuler des hypothèses pour une poursuite de la recherche en ce sens.

## 3.1 Résultats préliminaires

Tout d'abord, on souhaite reproduire la première figure du papier, à savoir celle montrant le trade-off entre proportion de vrais positifs (TPP, *true positive proportion*) et la proportion de fausses découvertes (FDP, *false discovery proportion*). Pour ce faire, les auteurs se placent dans le cadre suivant, que l'on reprend donc pour notre propre expérience, et qui est le même que celui ayant permis d'obtenir la Figure 2 : $p = 1000, n = 1010$, où chaque élément de $X \in \mathbb{R}^{n \times p}$ est simulé suivant une loi $\mathcal{N}(0,1)$, avec $k$ coefficients non nuls valant tous $M = 4$, et où les réponses $y$ sont obtenues de la façon suivante,

$$y = X\beta + z$$

où $z \sim \mathcal{N}_n(0,1)$ est le bruit.

Avec ces paramètres, on obtient le résultat présenté en Figure 7 [4]. Dans cette implémentation, nous avons également tracé le trade-off pour le *positive Lasso*, qui est une version du Lasso où les coefficients sont forcés à être positifs ou nuls.

Pour rappel, le Lasso path est l'ensemble suivant : $\left\{\hat{\beta}(\lambda), \lambda \in (0, \infty)\right\}$, où $\lambda$ est la paramètre de régularisation et où $\hat{\beta}(\lambda)$ est le vecteurs des coefficients estimés par le Lasso avec une régularisation $\lambda$. Les auteurs ne précisent cependant pas pour quelles valeurs de $\lambda$ ils calculent le Lasso path. Nous avons donc décidé de prendre 1 000 valeurs de $\lambda$, avec un ratio $\lambda_{min}/\lambda_{max} = 10^{-3}$, la valeur par défaut de la fonction `lasso_path` de Scikit Learn. Les différentes valeurs de $\lambda$ sont calculées

4. Tous les points ne sont pas représentés. On fait ici le choix de ne pas représenter les points pour lesquels TPP vaut 1, pour des soucis de clarté. Cependant, il faut noter qu'une fois que TPP a atteint le seuil de 1, les points associés ont un FDP qui ne fait que croître avec $\lambda$, jusqu'à atteindre 1.

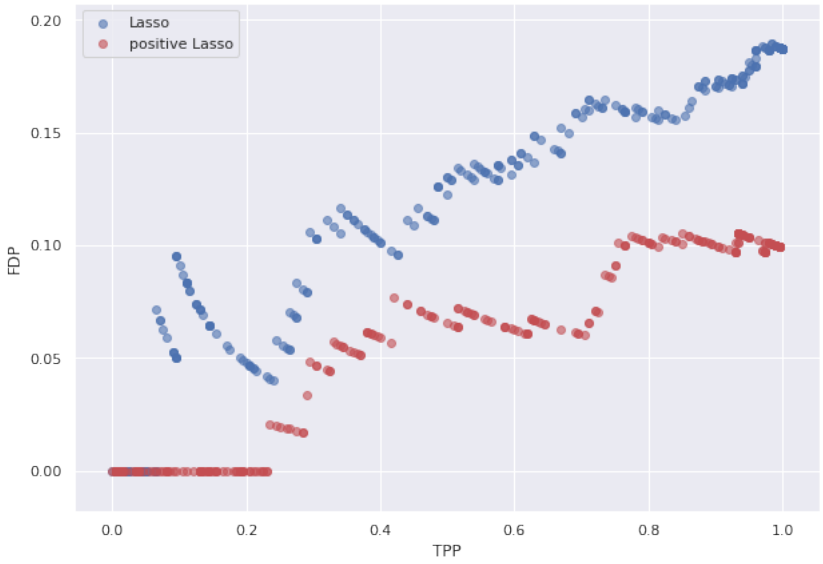



FIGURE 7 – Vrais positifs et fausses découvertes le long du "Lasso path"

automatiquement par la fonction, et sont réparties de façon uniforme sur une échelle logarithmique entre $\log_{10}(\lambda_{min})$ et $\log_{10}(\lambda_{max})$, où $\lambda_{max} = \frac{\max\left(\sqrt{(X^T y)^2}\right)}{n+1}$ avec la notation suivante : soit $x \in \mathbb{R}^p$, alors $\max(x) \coloneqq \max\{x_i, i = 1, \ldots, p\}$, la plus grande coordonnée du vecteur $x$.

On constate que les résultats que l'on obtient pour le Lasso usuel (points bleus dans la Figure 7), sont très semblables à ceux obtenus par les auteurs. Cela nous conforte donc dans notre capacité de reproduire leurs résultats, et nous permet donc de pouvoir comparer séreinement nos prochains résultats à ceux obtenus dans Su et al., 2017 [2]. De plus, on peut constater que le trade-off est plus favorable pour le positive Lasso que pour le Lasso usuel. En effet, on observe que pour avoir au moins 50 % de vrais positifs, il faut « subir » un taux de faux positifs d'environ 6 % pour le positive Lasso alors que pour le Lasso usuel, ce pourcentage s'élève à 13 %, soit plus du double.

Dans la suite de nos implémentations, on se propose entre autres de comparer les résultats obtenus avec le Lasso usuel avec ceux du positive Lasso. De cette façon, on pourra éventuellement formuler des hypothèses quant à la performance du positive Lasso, basé sur nos résultats empiriques.

On souhaite maintenant reproduire la Figure 2, qui s'inscrit dans le même cadre expérimental que celui décrit précédemment. Pour produire la Figure 2, les auteurs ont simulé 100 Lasso paths, nous proposons d'en faire 200, afin d'affiner les résultats. Ainsi, on regarde pour chacune des simulation deux événements : la valeur de TPP au moment de la première fausse découverte, ainsi que la valeur de FDP au moment où TPP atteint 1, c'est-à-dire au moment où le nombre de faux négatifs est nul. On trace ainsi les histogrammes obtenus, et pour faciliter la lecture des résultats, on trace également la *Kernel Density Estimation* pour chacun des histogrammes.

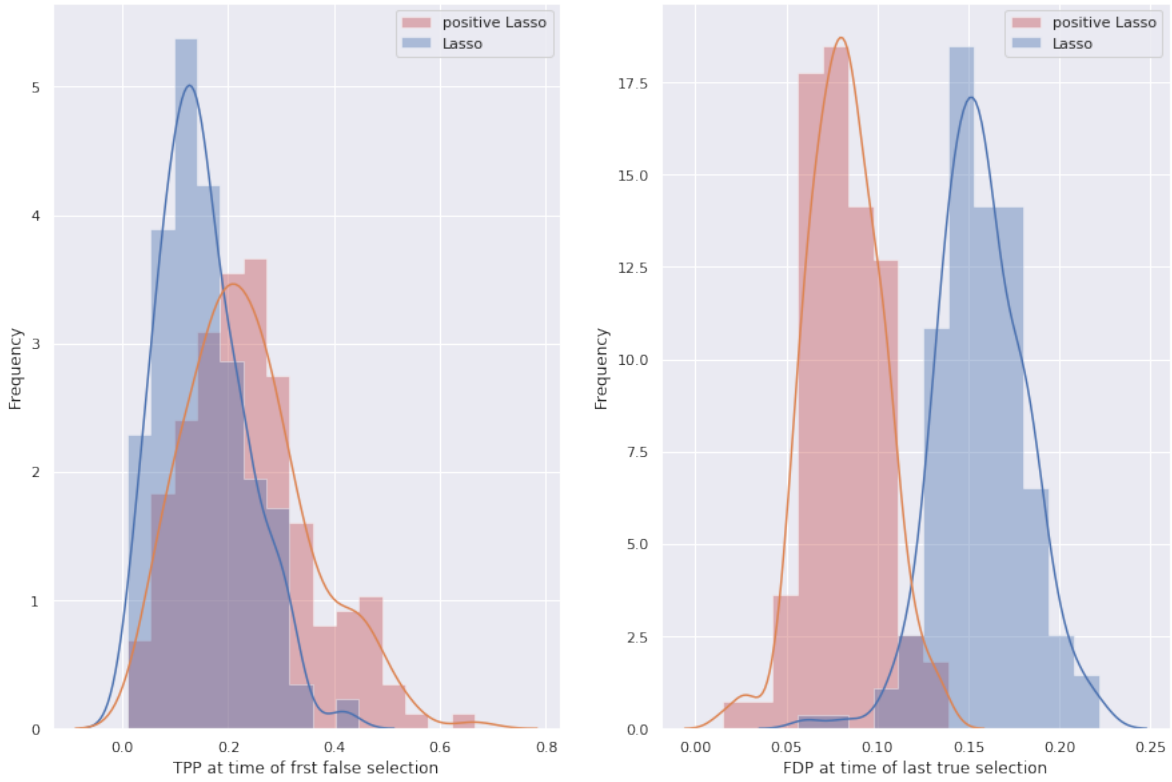



FIGURE 8 – Gauche : taux de vrais positifs au moment de l'apparition de la première fausse découverte. Droite : proportion de fausses découvertes lorsque le taux de vrais positifs atteint 1 (i.e. lorsque la proportion de faux négatifs atteint 0).

De nouveau, on observe des résultats pour le Lasso semblables à ceux obtenus par les auteurs. De plus, on observe également que pour le positive Lasso, les histogrammes sont décalés dans « le bon sens » par rapport aux histogrammes obtenus pour le Lasso. En effet, dans toutes les simulations, pour le positive Lasso, la première fausse découverte entre dans le Lasso path avant que 65 % des vrais signaux soient détectés, alors que cette proportion n'est que de 41 % pour le Lasso usuel (les auteurs trouvent 44 %). Ainsi, le positive Lasso est plus performant dans le sens où il retarde l'apparition des fausses découvertes. Une analyse similaire peut être faite avec l'histogramme de droite de la Figure 8. Ces résultats viennent confirmer ce que l'on pouvait déjà observer avec les résultats de la Figure 7.

### 3.2 Frontière du Lasso path

Dans la suite de leurs expérimentations, les auteurs utilisent la frontière du Lasso path $q^*$, dont ils donnent le code en Matlab pour la calculer, disponible sur le GitHub de Weijie Su : `https://github.com/wjsu/fdrlasso`. Cependant, pour des raisons pratiques, nous avons choisi d'effectuer

nos implémentations en langage Python, nous proposons donc une adaptation du code MatLab fourni en Python : `https://github.com/CedricAllainEnsae/fdrlasso` [5].

Pour s'assurer que notre adaptation Python soit cohérente, on propose donc de reproduire la Figure 3. On obtient ainsi la Figure 9. On peut ainsi observer que pour chacun des cas, $\delta = 0.5$ et $\delta = 0.3$ , on obtient bien les mêmes frontières que dans l'article, on retrouve notamment la troncature à TPP=0.6791 dans le second cas. On peut y voir ici un point positif pour la reproductibilité des résultats.

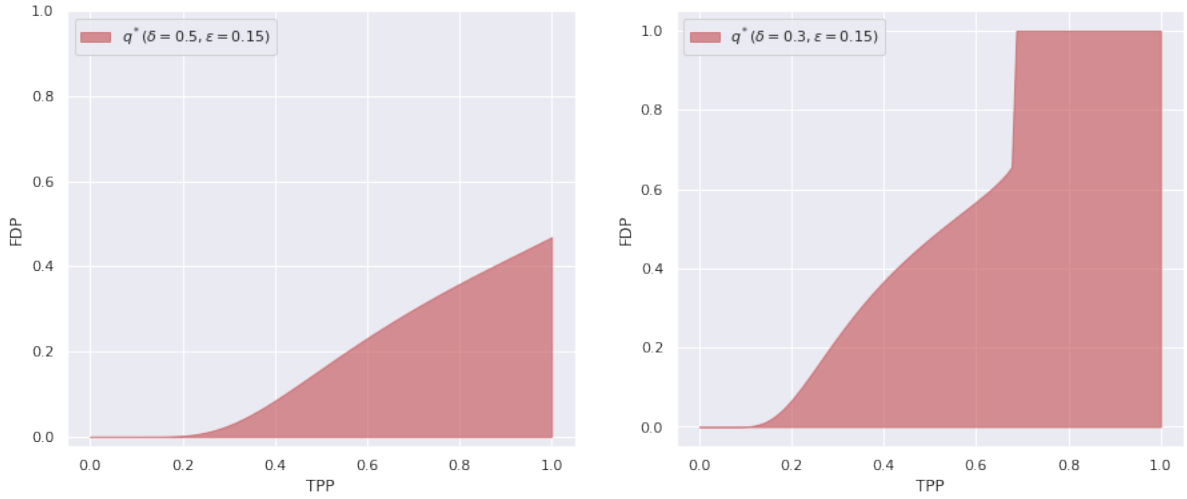



FIGURE 9 – Diagramme du trade-off Lasso. Gauche : avec $\delta = 0.5$ et $\epsilon = 0.15$. Droite : avec $\delta = 0.3$ et $\epsilon = 0.15$

**Remarque** Dans la suite des expérimentations, à la différence des auteurs, la matrice $\boldsymbol{X}$ a des éléments simulés suivant une loi $\mathcal{N}(0, 1)$.

On se propose maintenant d'élargir le cadre d'expérimentation ayant mené à la Figure 4. En effet, comme mentionné dans Section 2, les auteurs ne testent que le cas $\delta \leq 1$. On élargit donc à des valeurs de $\delta$ plus grandes strictement que 1. On conserve cependant les mêmes valeurs de $\epsilon$ de façon à pouvoir comparer nos résultats avec ceux des auteurs. Nos résultats sont présentés en Figure 10.

On observe que pour toutes les valeurs de $\delta$ et de $\epsilon$ testées, aucune n'amène à une situation « pathologique », dans le sens où, comme dans le graphique de droite de la Figure 9 [6], il y aurait une troncature verticale, qui empêcherait d'atteindre des points avec un taux de vrais positifs égal à 1. Ceci s'explique par le fait que tester des valeurs de $\delta$ supérieures à 1 revient à des cas où $n > p$, ce qui est un cas favorable en statistique, et donc pour la régression Lasso en particulier. Ainsi, dans chaque cas, on atteint les points qui ont un TPP de 1. De plus, on peut tirer les mêmes conclusions

5. Le Jupyter Notebook ayant servi à obtenir nos résultats y est également disponible.
6. Dans l'hypothèse où, comme chez les auteurs, une telle troncature existe bien.

que les auteurs, dans le sens où l'on constate que le contrôle de FDP est plus compliqué lorsque la sparsité $\epsilon$ augmente ou lorsque $1/\delta$ augmente.

On peut également observer qu'avec une valeur de $\epsilon$ de 0.2 (valeur plutôt élevée dans notre contexte), augmenter $\delta$ permet de diminuer de façon drastique la proportion de faux positifs. Ce résultat peut être intéressant dans le cas où, pour une expérience donnée, il est possible d'augmenter le nombre d'observations, car comme $\delta \approx n/p$, augmenter $n$ fait augmenter $\delta$.

### 3.3 Une frontière pour le positive Lasso path ?

On reprend maintenant la comparaison entre le Lasso et le positive Lasso, que l'on avait commencée à faire avec les Figures 7 et 8, et on souhaite voir empiriquement si le positive Lasso possède une frontière, et si cette dernière diffère de celle pour le Lasso usuel.

Pour ce faire, on se place dans le cadre utilisé par les auteurs pour produire leur Figure 5 (a) [2], à savoir le cadre suivant : $n/p = \delta = 1, \epsilon = 0.2, M = 50$ et les données sont simulées à partir d'une loi $\mathcal{N}(0, 1)$. On simule ensuite 10 positive Lasso paths pour $p = 1000, 5000$, et dans le cadre bruité ($\sigma = 1$) et non bruité ($\sigma = 0$). Les résultats sont présentés en Figure 11.

Pour chaque cas, on trace les trade-off du positive Lasso obtenus ainsi que la frontière de Lasso usuel pour notre cadre d'expérience. De cette façon, il nous est possible de comparer les performances des deux types de régressions Lasso. Ainsi, on observe que le positive Lasso est plus performant que la Lasso usuel, dans le sens où, pour chaque cas, il existe une valeur de TPP pour laquelle l'ensemble des points passent en dessous de la frontière du Lasso. En effet, dans le cas bruité avec $n = p = 5000$, lorsque TPP$> 0.65$, l'ensemble des paires (TPP, FDP) sont en dessous de la frontière du Lasso, cette valeur est de 0.85 pour le cas avec $n = p = 1000$.

Dans les graphiques de droite de la Figure 11, on trace en plus de la frontière du Lasso usuel la courbe moyenne du nuage de points associé, ainsi que l'écart-type [7]. Avec ces résultats, on peut finalement émettre l'hypothèse qu'au même titre que le Lasso usuel, le positive Lasso admet une frontière, qui se trouve être plus avantageuse pour le contrôle de la proportion de faux positifs. Ainsi, une ouverture possible serait l'exploration théorique de notre hypothèse. Cependant, il faut garder en tête que le positve Lasso ne peut être utilisé que dans des cas précis, où l'on est certain que nos coefficients sont positifs ou nuls.

Finalement, on souhaite reproduire la Figure 7 de l'article mais pour le positive Lasso. Pour ce faire, on se place dans le même cadre que les auteurs, à savoir $n = p = 1000$ ($\delta = 1$), avec du bruit ($\sigma = 1$), et on procède comme suit : pour chaque valeur de $k$ entre 5 et 150 (en prenant un pas de 5), on simule 100 positive lasso paths en mettant les $k$ premiers coefficients à 50 (les $1000 - k$ autres sont nuls), puis on trace le rang moyen de la première fausse découverte, ainsi que le rang minimal

7. On obtient cela à l'aide de la fonction `lineplot` de seaborn.

et maximal atteint, et le rang médian accompagné du premier et troisième quartile [8]. Nos résultats sont présentés dans la Figure 12.

Contrairement aux résultats précédents où le positive Lasso permettait d'obtenir de meilleurs résultats que le Lasso usuel, on peut observer ici que le graphique obtenu est très similaire à celui obtenu par les auteurs. À noter cependant qu'il subsiste une différence avec le résultat des auteurs : pour les grandes valeurs de $k$, supérieures à 120, le rang maximum atteint continu d'être égal à $k$, se situant ainsi sur la première diagonale. Cela illustre le fait que le positive Lasso est parfois capable de ne pas faire de fausses découvertes même pour un $k$ élevé, bien qu'en moyenne il ait une performance équivalente à celle du Lasso usuel.

On peut conclure en stipulant que pour cette expérience, on ne peut que tirer les mêmes conclusions que les auteurs, à savoir que lorsque $k$ est très faible, le modèle retrouve la totalité des signaux, mais lorsque $k$ augmente, on observe des fausses découvertes précoces, et ce même pour des valeurs de $k$ inférieures à $n/(2\log p)$. De plus, lorsque que $k$ dépasse $n/(2\log p)$, le rang moyen de la première fausse découverte diminue de plus en plus, s'éloignant ainsi de $k$ (la première diagonale). On ne peut donc que confirmer les observations des auteurs mais dans le cas du positive Lasso.

## Références


[1] Galen Reeves and Michael C Gastpar. Approximate sparsity pattern recovery : Information-theoretic lower bounds. *IEEE Transactions on Information Theory*, 59(6) :3451–3465, 2013.

[2] Weijie Su, Małgorzata Bogdan, Emmanuel Candes, et al. False discoveries occur early on the lasso path. *The Annals of statistics*, 45(5) :2133–2150, 2017.

[3] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58 :267–288, 1996.

[4] Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12) :5728–5741, 2009.

[5] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5) :2183–2202, 2009.


8. Les auteurs se sont contentés de tracer le rang moyen, minimum et maximum, mais cela ne permet pas d'appréhender la distribution globale des rangs pour une valeur de $k$ donnée.

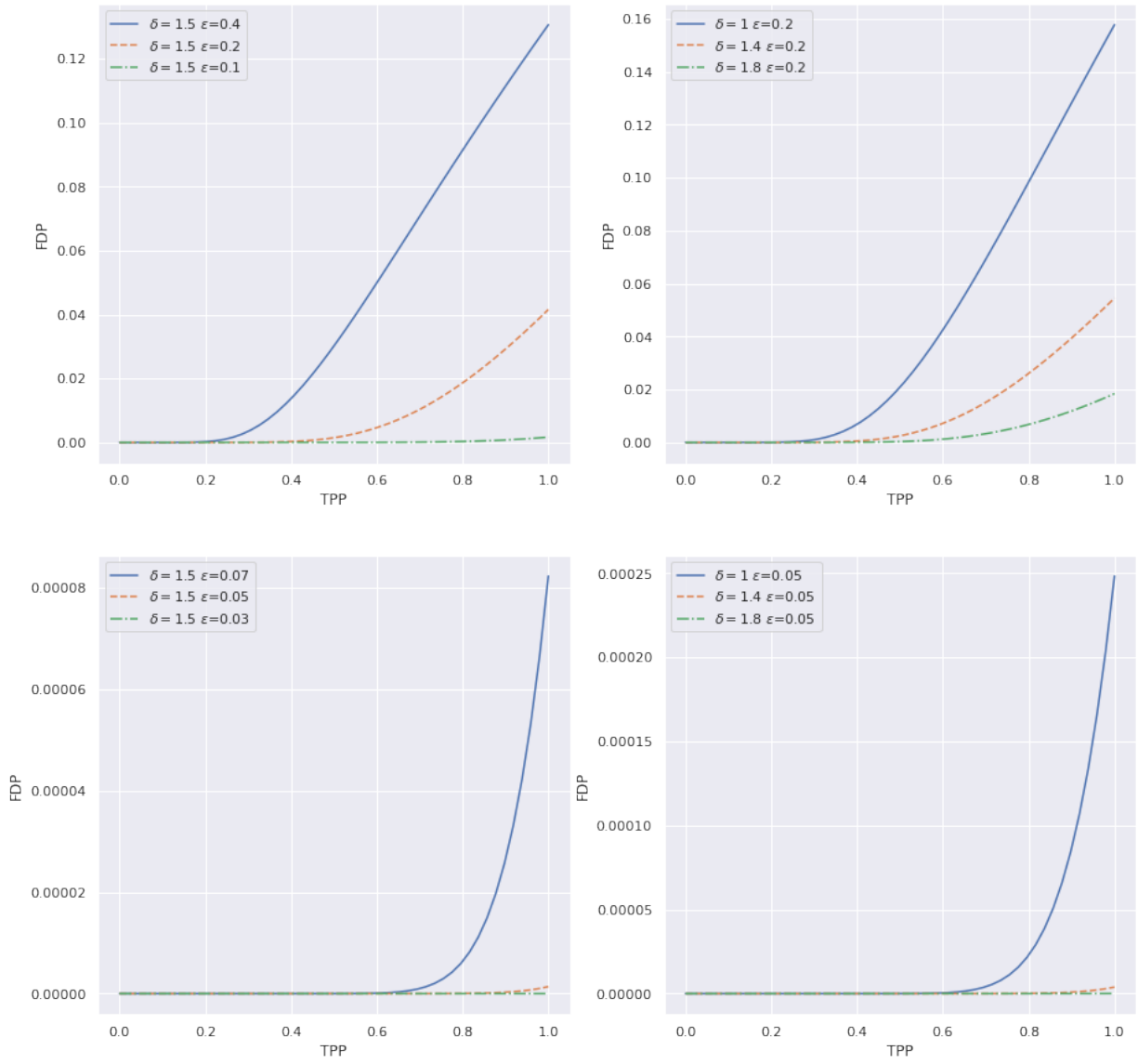



FIGURE 10 – Diagramme du trade-off Lasso. Supérieur et inférieur gauche : avec $\delta = 1.5$ ; supérieur droit : avec $\epsilon = 0.2$ ; inférieur droit : avec $\epsilon = 0.05$.
Remarque : par soucis de visibilité, on ne force pas l'axe des ordonnées à aller jusqu'à 1.

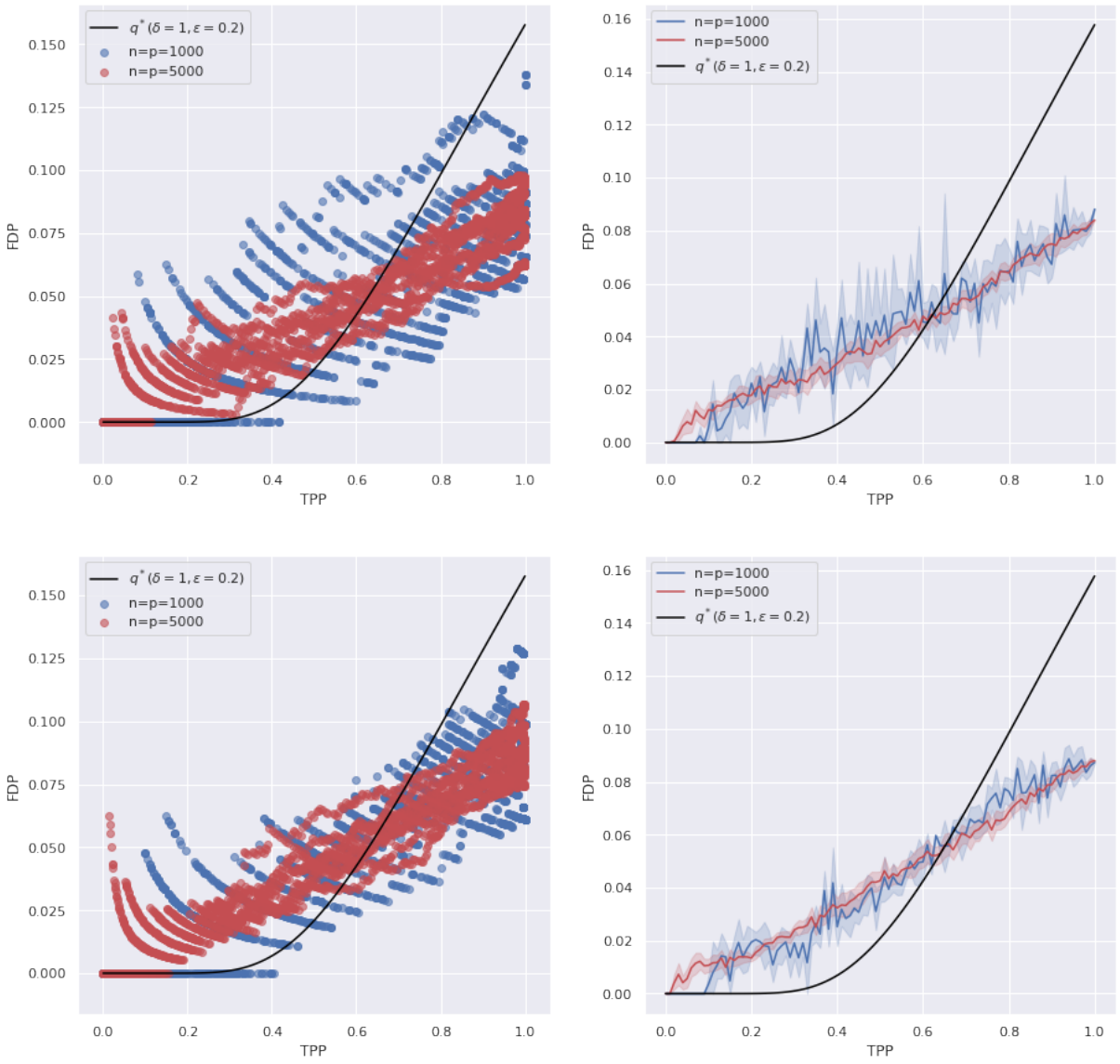



FIGURE 11 – $n/p = \delta = 1, \epsilon = 0.2, M = 50$ pour 10 positive Lasso paths indépendants, avec bruit, i.e. $\sigma = 1$ (graphes supérieurs) et sans , i.e. $\sigma = 0$ (graphes inférieurs). À gauche les points tels quels, à droite une courbe passant par la moyenne des points, avec un intervalle de confiance en clair.

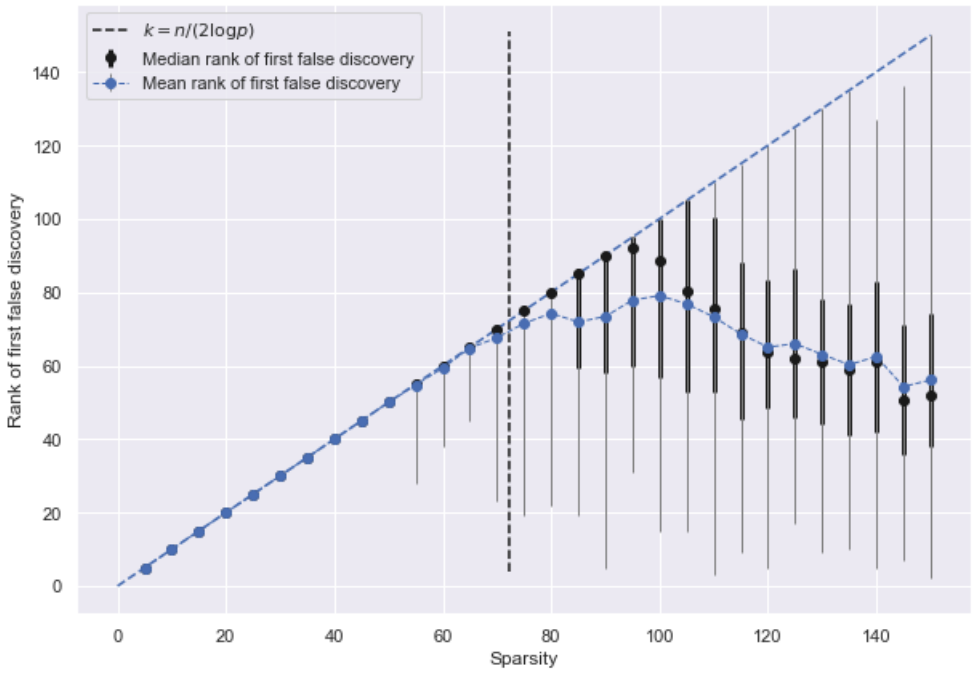



FIGURE 12 – Rang de la première fausse découverte. Ici, $n = p = 1000$ et $\beta_1, \ldots, \beta_k = 50$ pour $k$ allant de 5 à 150 ($\beta_i = 0$ pour $i > k$). Pour chaque valeur de $k$, on trace le rang moyen pour 100 simulations indépendantes du positive Lasso path, ainsi que le rang minimal et maximal atteint, et le rang médian accompagné du premier et troisième quartile. La ligne verticale est placée à $k = n/(2\log p)$, et la première diagonale est tracée pour faciliter la lecture.