Analysis of the article: False discoveries occur early on the Lasso path

Adnan Asadullah Imad Al Moslli

February 2020

Présentation du contexte et des principaux résultats

1 Contexte

1.1 Optimalité théorique

On suppose que :
$\epsilon_i \sim \mathcal{N}(0, \sigma^2), \beta \in \{\beta_0[k] = \{\beta \in \mathbb{R}^p : |\beta|_0 \leq k\}$ (au plus k-sparse)

Mesure de qualité (risque) :

$$R(\hat{\beta}, \beta) = \mathbb{E}[||X\hat{\beta} - X\beta||^2]$$

Meilleut estimateur de β :

$$\inf_{\hat{\beta}} R(\hat{\beta}, \beta) = 0$$

Problème : estimateur dégénéré (on ne connaît pas β)

Risque minimax : donne une garantie/un contrôle dans le pire des cas

$$\inf_{\hat{\beta}} \sup_{\beta \in \beta_0[k]} R(\hat{\beta}, \beta) = 0$$

Risque bayésien :

$$\inf_{\hat{\beta}} \int_{\beta \in \beta_0[k]} R(\hat{\beta}, \beta) d\mu(\beta)$$

Remarque : Ce dernier dépend de la mesure μ choisie

Si k est inconnu, on a :

$$\hat{\beta} \in arg \min_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2 + \overbrace{2\mu\sigma^2|\beta|_0(1 + \log \frac{p}{|\beta|_0})}^{terme\ de\ pnalisation\ des\ \beta\ qui\ ont\ beaucoup\ de\ coefficients\ non\ nuls}$$

Dans ce cas, on obtient : $\sup_{\beta \in \beta_0[k]} R(\hat{\beta}, \beta) \leq c\sigma^2 k(1 + log\frac{p}{k})$ pour $k \leq \frac{n}{\log(p)}$

Remarque : Impossible à calculer en temps fini (sauf cas particuliers)

Exemple : $X^TX = I_p$

$$||X(z-\beta)||^2 = (z-\beta)^T \overbrace{X^TX}^{I_p}(z-\beta) = ||(z-\beta)||^2$$

$$\hat{\beta} \in arg\min_{\underbrace{\beta}_{=\cup_{k\geq 0}\beta_0[k]} \in \mathbb{R}^p} [\sum_j (\beta_j - z_j)^2 + |\beta|_0 \times \overbrace{2\mu\sigma^2(1+\log\frac{p}{|\beta|_0})}^{\lambda}]$$

$$\min_{\beta\in\beta_0[k]} \sum_{j=1}^p (\beta_j - z_j)^2 = \sum_{j=1}^p z_j^2 - \sum_{j=1}^k z_{(j)}^2$$

Le minimum est atteint par $\hat{\beta}$ avec $\hat{\beta}_j = z_j$ si z_j^2 est une des k plus grandes valeurs.

On obtient : $\min_k ||z||^2 - \sum_{j=1}^k [z_{(j)}^2 - \lambda]$

On prend donc : $\hat{k} = \max_k z_{(k)}^2 \geq \lambda \Longrightarrow \hat{\beta}_j = z_j \mathbb{1}_{|z_j|\geq\sqrt{\lambda}}$

1.2 Lasso (pénalisation L^1)

Le Lasso correspond tout simplement à la solution d'un problème d'optimisation linéaire où le terme de pénalisation est de norme 1.

Relaxation convexe : $|\beta_0| \longrightarrow |\beta_1|$

Si $X^TX = I_p$, on a : $\hat{\beta}_j = [X_j^TY - \frac{\lambda}{2}sgn(X_j^TY)]\mathbb{1}_{|X_j^TY|\geq\lambda/2} = X_j^TY(1 - \frac{\lambda}{2|X_j^TY|})_+$

Remarque : La question est maintenant de savoir si cet estimateur est aussi bon que le précédent.

$$\text{Pour } \beta \in \ \beta_0[k],\ R(\hat{\beta},\beta) \leq \underbrace{C(X,\beta)}_{tend\ diverger\ si\ les\ X_j\ sont\ trs\ corrls} \times\sigma^2 k(1+log\frac{p}{k})\text{`}$$

Remarque : C'est en quelque sorte le prix à payer pour avoir un temps de calcul raisonable.

Exemple : $X_{ij} \sim \mathcal{N}(0, 1/n)$

$||X_j||^2 = \sum_{i=1}^n X_{ij}^2 \approx \frac{n}{n} + o(1/\sqrt{n})$ (TCL)

$< X_j, X_k >_{j\neq k} = \sum_{i=1}^n X_{ij}X_{ik} = o(1/\sqrt{n})$

Remarque : Ainsi, le Lasso a cette particularité de naturellement sélectionner des variables, notamment en grande dimension ($p \geq n$).

On compare ensuite : $f_\beta(z) = \sum_j \beta_j\phi_j(z)$ avec $f_{\hat{\beta}}(z) = \sum_j \hat{\beta}_j\phi_j(z)$

1.3 Feature selection

La question qui se pose maintenant est la suivante : a-t-on une bonne sélection des coefficients non nuls j ?

Exemple : Dans le cas $X_{ij} \sim \mathcal{N}(0, 1/n)$

Si $|\beta|_0 \leq \frac{n}{log(p)}$ alors $supp(\hat{\beta}) \approx supp(\beta)$ avec grande probabilité si $|\beta_j| \geq \sigma\sqrt{log(p)}, \beta_j \neq 0$

$$\hat{\beta}_j = X_j^T Y \overbrace{(1 - \frac{\lambda}{2|X_j^T Y|})_+}^{biais}$$

Remarque : Le lasso tend à sélectionner trop de variables car il choisit un λ trop petit afin de résuire le biais.

Le choix de λ peut se faire par Cross-Validation (CV) :

Répéter :
- Choix de n/2 observations au hasard ;
- Faire tourner le Lasso.

Choisir : Les variables j sélectionnées dans au moins 80% des cas.

1.4 Motivation de l'article

Le Lasso est donc communément utilisé lorque la solution est sparse (beaucoup de features ne sont pas prédictives). On sait que le Lasso est efficace dans un régime asymptotique. Cependant, dans un régime non asymptotique, sa robustesse n'est pas démontrée. Au contraire, en pratique, le Lasso ne sélectionne pas toujours les bonnes variables.

Ce problème du Lasso est générallement attribué au fait qu'il faille atteindre une taille d'effet, notamment du fait de la corrélation entre les variables prédictives.

L'article cherche alors à étudier le Lasso lorsqu'il y a peu voire pas du tout de bruit ni de corrélation pour se rendre compte que le problème est toujours présent et l'explique par un problème de coefficient dû à la norme L^1 qui n'existe pas lorsque l'on utilise la norme L^0.

2 Principaux résultats

2.1 Lasso Trade-off

C'est le principal résultat de l'article. Il montre qu'il est impossible d'avoir un taux important de bonnes sélections de variable combiné à un taux faible de mauvaises sélections. Commençons par définir ces deux termes.

Définitions : True Positive Proportion (TPP) et False Discovery Proportion (FDP)

$FDP(\lambda) = \frac{V(\lambda)}{max(|\{j:\hat{\beta}_j(\lambda)\neq 0\}|,1)}$ avec $V(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0 \, and \, \beta_j(\lambda) = 0\}$ (False Discovery Proportion)

$TPP(\lambda) = \frac{T(\lambda)}{max(k,1)}$ avec $T(\lambda) = \{j : \hat{\beta}_j(\lambda) \neq 0 \, and \, \beta_j(\lambda) \neq 0\}$ (True Positive Proportion)

Théorème 1 : Lasso Trade-off

On note $S = \{j : \beta_j \neq 0\}, k = |S| \approx p\epsilon = \epsilon/\delta n$ with $\delta \in (0, \infty), \epsilon \in (0, 1)$

On considère la fonction $q_{\epsilon,\lambda}(u) = \frac{2(1-\epsilon)\Phi(-t^*(u))}{2(1-\epsilon)\Phi(-t^*(u))+\epsilon u}$

$\forall \sigma \geq, \forall \lambda \geq \lambda_0, FDP(\lambda) \geq q_{\epsilon,\lambda}(TPP(\lambda)) - \eta$ pour n'importe quelles constantes $\lambda_0 > 0$ and $\eta > 0$ pouvant être arbitrairement petites.

2.2 Performance de la norme l_0

Le deuxième grand résultat de l'article a pour objectif d'expliquer pourquoi le Lasso souffre de ces limitations. Pour cela, il compare l'estimation Lasso avec un estimateur de maximum de vraisemblance appliqué avec une norme l_0 :

$\hat{\beta}_0 = arg\min_{b\in\mathbb{R}^p} ||y - Xb||^2 + \lambda||b||_0$

Théorème 2 : l_0 performance

On prend $\epsilon < \delta$ et on considère le point prior suivant :

$$\Pi = \begin{cases} M \, avec \, probabilit \, \epsilon \\ 0 \, avec \, probabilit \, 1 - \epsilon \end{cases}$$

Dans ce cas, on a :

$$\begin{cases} \lim_{M\to\infty} lim_{n,p\to\infty} FDP = 0 \\ \lim_{M\to\infty} lim_{n,p\to\infty} TPP = 1 \end{cases}$$

Remarque : Ce théorème montre qu'avec la norme l_0, on peut avoir 100% de bonne sélection et 0% de mauvaise sélection en même temps. ‘

Le problème du Lasso viendrait donc du terme de régularisation. Lorsque λ est grand, les estimations du Lasso (coefficients β) sont tirés vers le bas, ce qui a pour effet d'ajouter du bruit (appelé "shrinkage noise"). Ainsi, plus le Lasso va sélectionner des variables, plus ce bruit va également augmenter et se propager également dans la direction des variables nulles, ce qui aura pour conséquence que le Lasso va faire plus d'erreur en les sélectionnant.

Partie 2 : Analyse critique du papier

3 Portée des résultats et limitations

Les résultats obtenus dans cet article s'inscrivent dans un cadre bien précis.

3.1 Portée : Régression avec des variables Gaussiennes

Les résultats présentés peuvent êêtre jugés comme plutôôt négatifs. Or, les auteurs se sont placés dans un cadre classique où les variables prédictives $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(0, 1)$.
Ce cadre est considéré comme facilitant pour la sélection de variable puisque les variables ne sont pas corrélées. Ainsi, ces résultats devraient pouvoir se généraliser dans des cadres moins favorables.

Remarque : Il n'y a pas de contrainte sur la distribution des coefficients $\beta_1, ..., \beta_p$ qi ce n'est qu'ils doivent avoir un moment d'ordre 2 borné ($\mathbb{E}[\beta^2] < \infty$) et une distribution empirique qui converge.

3.2 Limitation : Sparsité linéaire

Tout d'abord, on suppose que l'on est dans un régime de sparsité linéaire, c'est-à-dire que l'on suppose que le nombre de coefficients non nuls est égal à ϵp avec $\epsilon > 0$ et p le nombre de features.

Cela exclut de fait des des régimes plus courants dans lesquels le nombre de coeffcient non nuls tend vers 0 lorque la taille du problème tend vers l'infini. Ainsi, on peut peux pas généraliser ces résultats à ce type de cas. Cependant, le régime de sparsité linéaire est également intéressant car il décrit des données qui sont certes en grande dimension mais pas forcément infinie avec un degré de sparsité assez important.

4 Message à retenir

Le message à retenir de cet article peut tenir sur une seule ligne : Si $|\beta|_0$ est proportinel à p, ça se passe mal.

Si on doit résumer en plus d'une ligne, en grande dimension, l'augmentation de l'espace de représentation des données pose des problèmes de comparaison car celles-ci ont tendance à s'écarter, ce qui rend plus difficile les tâches de régression/classification : c'est ce qu'on appelle "the curse of high dimensionality" (le fléau de la grande dimension).

Le LASSO (Least Absolute Shrinkage and Selection Operator) a pour but de solutioner ce problème en ajoutant un terme de régularisation en norme L1, qui, contrairement au Ridge (régularisation en norme L2), va annuler des coefficients et ainsi sélectionner des variables.

C'est très bien mais encore faut-il que le modèle sélectionne les bonnes variables (celles qui sont explicatives). Cela a été prouvé dans le cas où le nombre de d'individus n est inférieur au nombre de variables p par example. Certaines limites du Lasso ont également été mises en lumière, notamment dans le cas où les variables sont corrélées, la consistence de la sélection n'est pas assurée.

La contribution de cet article se situe dans le cas où le nombre de coefficients non nuls $|\beta|_0 = K$ est proportionnel au nombre de variables p. Dans ce cas, il montre que la performance du Lasso est limitée car plus le nombre de variables explicatives sélectionnées sera important, plus le nombre de variables non explicatives sélectionnées le sera également.

Partie 3 : Exploration personnelle (numérique)

5 Problématique et plan d'expérience

Dans cette partie, nous allons essayer de sortir du cadre précis dans lequel se sont placés les auteurs pour voir comment évoluent les résultats qu'ils ont obtenus.

On peut distinguer plusieurs points qui définissent ce cadre :
- Les variables explicatives sont gaussiennes et indépendantes ;
- L'article se place dans un régime de sparsité linéaire ;

Une expérience intéressante pourrait consister à étudier les résultats de l'article avec des variables explicatives corrélées :
- Les résultats de l'auteur sont-ils toujours valables lorsque la matrice de design est composé d'éléments qui sont corrélés ?
- Comment évolue trade-off lorsque l'on fait varier la corrélation ?

Pour répondre à cette problématique, on va suivre le plan d'expérience suivant :
- Tout d'abord, on va cherche à reproduire les résultats qui sont présentés dans l'article pour avoir une base solide de comparaison ;
- Ensuite, on va sortir du régime de sparsité linéaire en considérant une faible sparsité ;
- Puis on va ajouter une corrélation standard pour voir coment le Lasso réagit ;
- Enfin, on va faire varier cette corrélation pour analyser l'évoultion du trade-off lorsque l'on augmente la corrélation des variables explicatives.

Pour l'analyse de la corrélation, nous allons notamment nous référer à l'article "How Correlations Influence Lasso Prediction" publié par Mohamed Hebiri et Johannes C. Lederer en 2012. On reprend leur plan d'expérience pour analyser l'influence de la corrélation sur les résultats du Lasso.

Dans un premier, on va produire une matrice de design X toujours issue d'une distribution Gaussienne de moyenne 0 mais dont la matrice de variance-covariance sera composée de 1 sur la diagonale et de composantes égale à $\rho \in [O, 1[$ en dehors de celle-ci.

Dans un second temps, on va en plus modifier chaque colonne j de la matrice de design tel que : $X^{(j)} \leftarrow X^{(j)} + \eta N$ avec η une constante de contrôle et $N \sim \mathcal{N}(0, 1)$

En marge de cela, nous allons analyser les performances du Lasso avec et sans bruit pour vérifier si, comme le disent les auteurs, leurs résultats s'appliquent bien dans ce cas.

Enfin, nous allons comparer les performances du Lasso avec un autre modèle : Elastic Net.
Il s'agit d'un mixte entre la régression Lasso et la régression Ridge :

$$\hat{\beta} \in arg \min_{\beta \in \mathbb{R}^p} ||Y - X\beta||^2 + \lambda_1 ||\beta||_1 + \lambda_1 ||\beta||^2$$

6 Réalisation des expériences

6.1 Lasso path

```
TPP au premier faux positif :  0.19
FDP lorsque TPP = 50% :  0.038461538461538464
FDP lorsque TPP = 100% :  0.14163090128755365
```

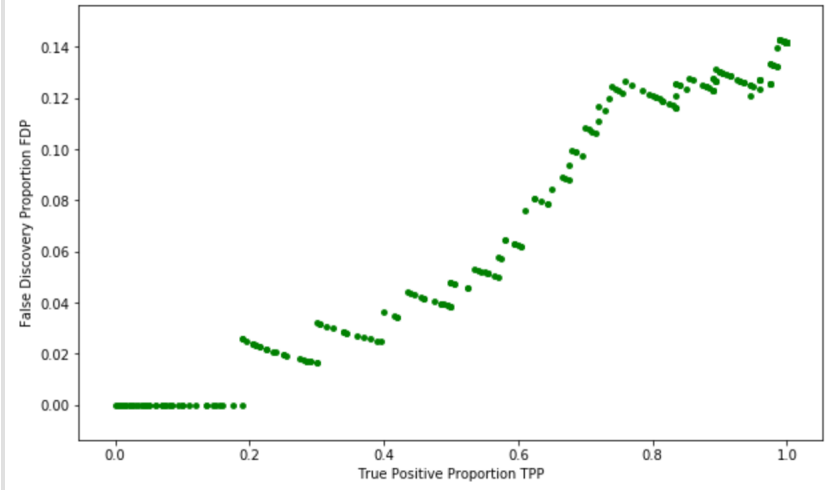

FIGURE 1 – True positive and false positive rates along the Lasso path

On peut voir que plus on augmente la proportion de vrais positifs (en faisant varier le paramètre de régularisation), plus la proportion de faux positifs augmente mécaniquement.

La première erreur intervient très tôt, lorsque le taux de vrai positif n'étant encore que de 19% à ce moment-là.

Pour atteindre 50% de vrai positif, le Lasso a concédé 3,8% de faux positifs

Enfin, lorsque l'on atteint 100% de vrais positifs (tous les coefficients non nuls ont été trouvés), le taux de faux positifs s'élève à près de 14%.

Remarque : Si on regénère le lasso path, ces résultats risquent de varier mais selon l'article, le Lasso fait toujours une erreur au moins avant d'atteindre un taux de vrai positif de 44%. Pour avoir des chiffres plus robuste, on va simuler le Lasso path plusieurs fois.

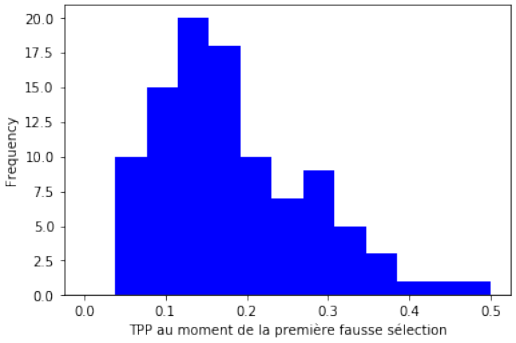

FIGURE 2 – Taux de vrai positif au moment de la première erreur (pour 100 simulations)

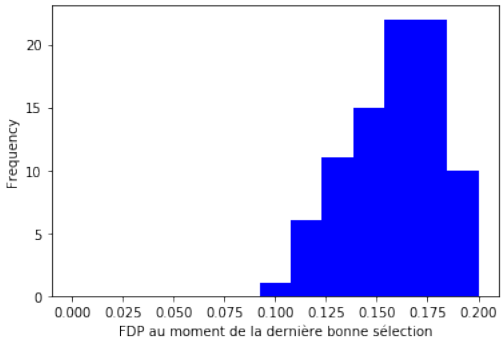

FIGURE 3 – Taux de faux positif lorsque l'on a atteint 100% de vrai positif (pour 100 simulations)

On peut voir que la plupart du temps, le taux de vrai positif se situe entre 10% et 20% au moment de la première erreur. On peut donc confirmer ce que dit l'article dès sont intitulé : les erreurs interviennent tôt sur le "Lasso path".

Le second histogramme permet toutefois de relativiser cette faiblesse du Lasso puisque le taux d'erreur à la fin ne semble pas dépasser le 20%. Ainsi, bien que le Lasso fait des erreurs assez tôt, au final, il n'en fait pas énormément et conserve donc une certaine capacité à sélectionner les variables explicatives.

6.2 Lasso Trade-off Diagram

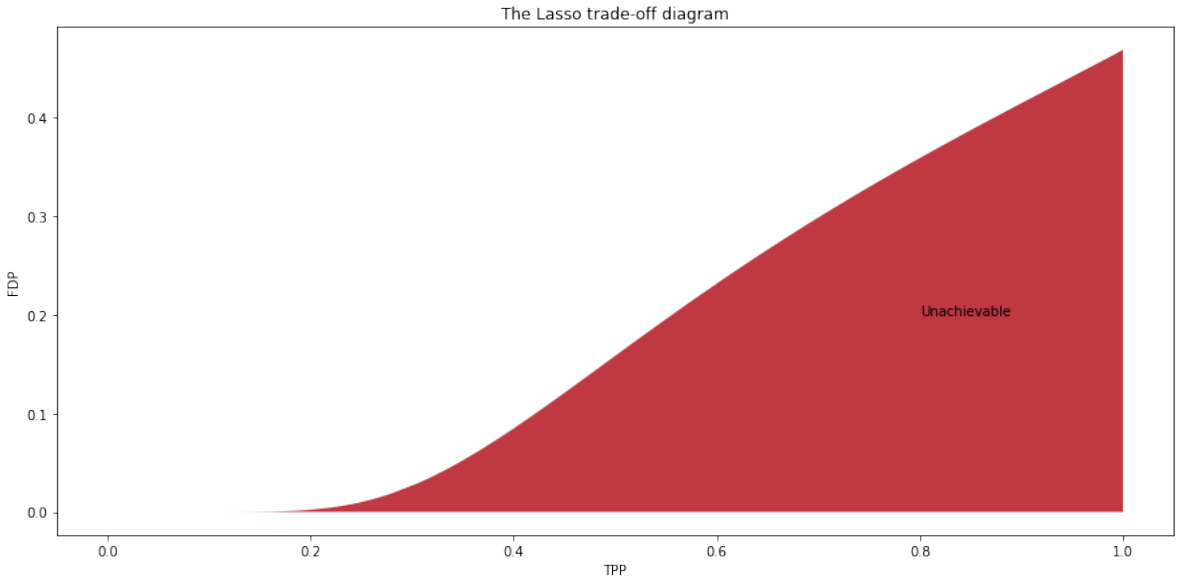

FIGURE 4 – $\delta = 0.5, \epsilon = 0.15$

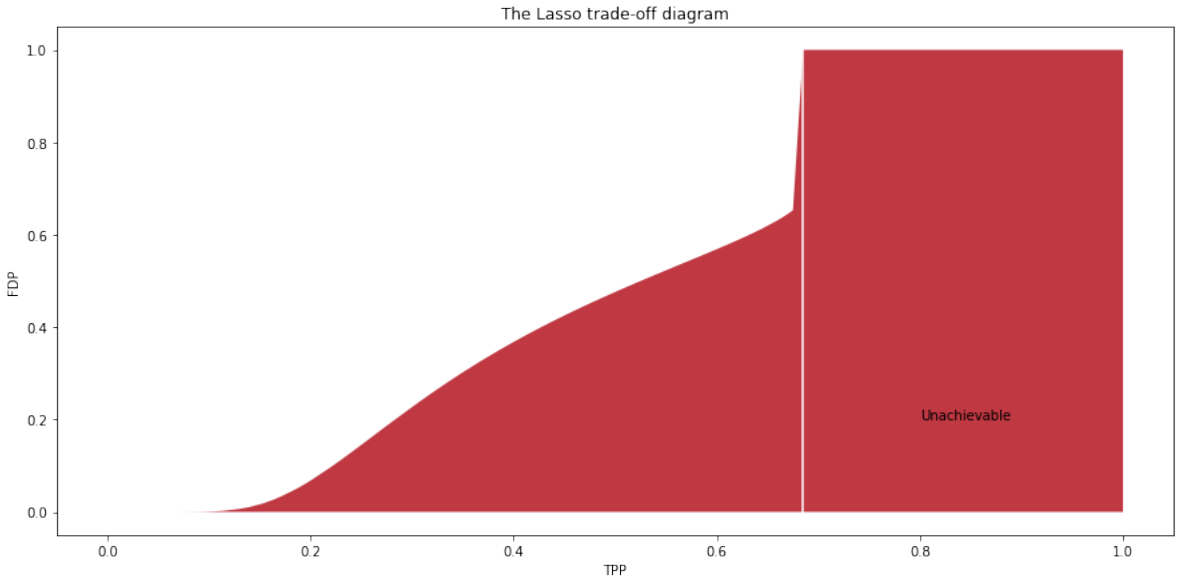

FIGURE 5 – $\delta = 0.3, \epsilon = 0.15$

Ces diagrammes représentent la fameuse fonction q qui décrit la borne inférieure du taux de faux positif. Ainsi, on peut apercevoir visuellement la zone qui est inatteignable pour le Lasso : lorsque l'on augmente le taux de vrais positifs, le taux de faux positif se trouvera forcément au-dessus d'un certain niveau représenté par la zone rouge dans le graphique (voir Théorème 1).

On voit que lorsque l'on augmente δ, c'est-à-dire le ratio du nombre de features sur le nombre de variables, la capacité du Lasso à sélectionner les variables explicatives se retrouvent très clairement limitée puisqu'on ne peut même pas dépasser 70% de vrais positifs. En effet, au-delà de ce seuil, toutes les variables sélectionnées le sont par erreur (ce ne sont pas des variables explicatives).
Les courbes ci-dessous représentent cette borne pour différentes valeurs de δ et de ϵ :

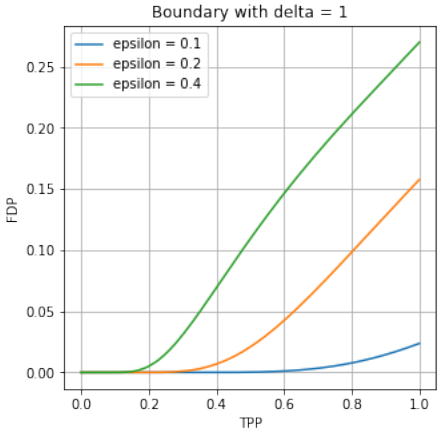

FIGURE 6 – $\delta = 1$

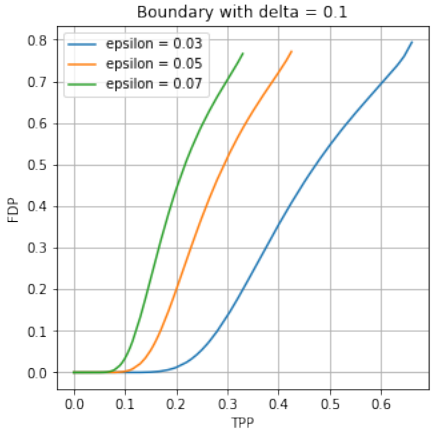

FIGURE 7 – $\delta = 0.1$

Les figures 6 et 7 nous permettent d'observer la "boundary curve" pour un nombre de variables égale au nombre d'observation et pour un nombre de variables dix fois supérieur au nombre d'observations respectivement.

Elles confirment que lorsque l'on augmente le nombre de variables par rapport au nombre d'observation, c'est-à-dire la dimensionalité, les performances du Lasso se dégradent : il a de plus en plus de mal à sélectionner les variables explicatives. Ce phénomène est amplifié lorsque l'on augmente la sparsité, à tel point que le taux de vrais positifs ne dépasse même pas les 50% pour $\delta = n/p = 0.1$ et $\epsilon = k/p = 0.07$

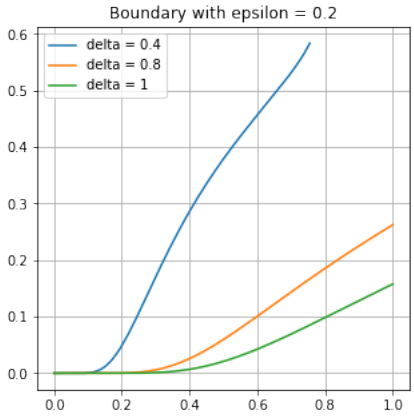

FIGURE 8 – $\epsilon = 0.2$

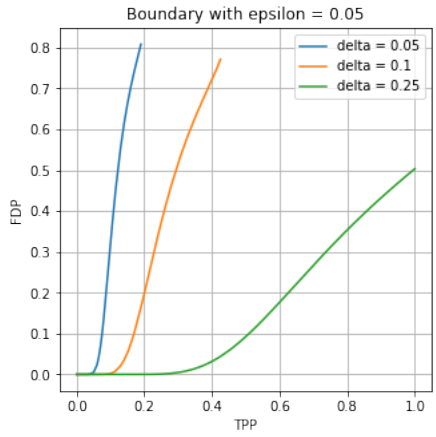

FIGURE 9 – $\epsilon = 0.05$

La figure 10 illustre la capacité de la fonction q à décrire les limites du Lasso. On peut voir qu'elle décrit plus ou moins bien la zone que ne peut pas atteindre le Lasso, qui correspond à un taux de vrai positif proche de 100% en même temps qu'un taux de faux positifs proche de %.

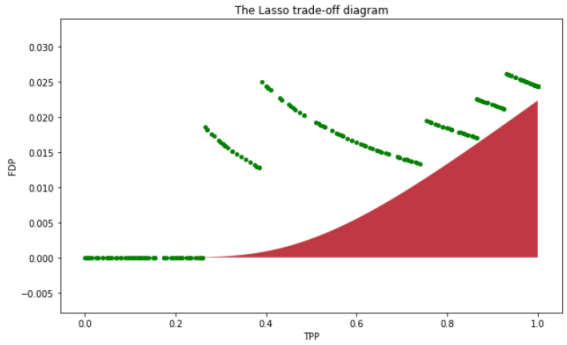

FIGURE 10 – Lasso Trade-off illustré avec le Lasso path

6.3 Performance du Lasso en faible dimension (et sparsité)

La littérature a montré plus d'une fois la capcité du Lasso à sélectionner les bonnes variables dans un régime sous-linéaire, c'est-à-dire avec une faible sparsité. Il a été asymptotiquement démontré que si $n >= (2 + o(1))klog(p)$ alors le Lasso sélectionne parfaitement les variables explicatives.

La figure 11 montre la performance du Lasso pour une matrice de design de taille 250 x 1000 (valeurs finies) et un nombre de coefficients nuls $k = 18$ avec un bruit de variance $\sigma^2 = 1$:

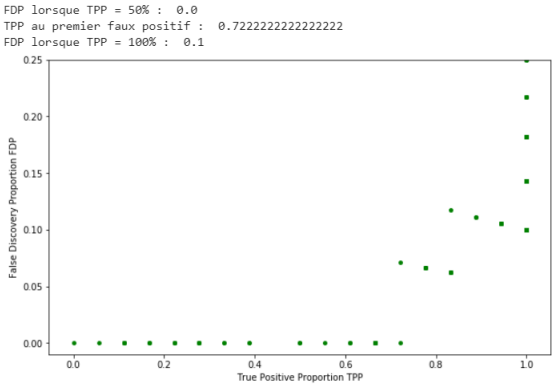

FIGURE 11 – 18 betas non nul

On peut apercevoir que dans ce cas, le Lasso fait quelques erreurs avant d'obtenir toutes les variables explicatives et que l'on a un taux de faux positif de près de 15% à ce moment. Ainsi, en dimension finie, la sélection parfaite n'est pas garantie.

6.4 Analyse du Lasso avec des variables corrélées

Commençons par ajouter de la corrélation standard à partir de la matrive de variance covariance.

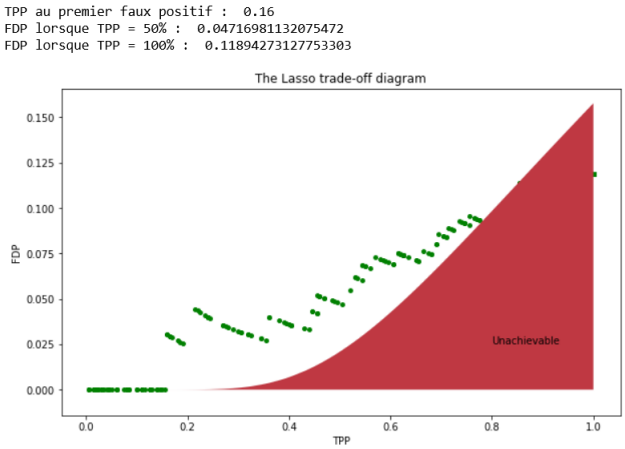

FIGURE 12 – $\rho = 0.001$

```
FDP lorsque TPP = 50% :  0.0
TPP au premier faux positif :  0.605
FDP lorsque TPP = 100% :  0.024390243902439025
```

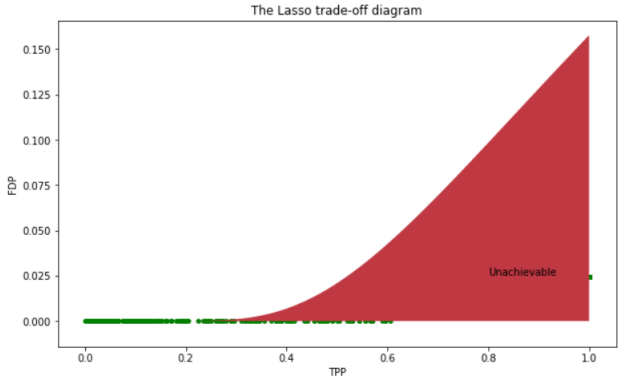

FIGURE 13 – $\rho = 0.1$

Lorsque l'on ajoute un peu de corrélation entre les variables, on voit que la "boundary curve" n'est plus adaptée. En effet, il y a un recouvrement entre les performance de Lasso et la zone inatteignable. C'est plutôt appréciable : on arrive, en ajoutant un peu de corrélation entre les variables, à atteindre un niveau de précision qui semblait inaccessible. En effet, on a un taux de vrais positifs de 100% avec seulemnt 2,5% de faux positifs pour $\rho = 0.1$.

Par contre, on peut voir que le niveau de corrélation choisi est très important puisqu'avec $\rho = 0.3$, on n'atteint même pas 100% de vrais positifs et on a déjà fait 8% d'erreur pour avoir 50% des variables explicatives sélectionnées.

```
TPP au premier faux positif :  0.22
FDP lorsque TPP = 50% :  0.04716981132075472
```

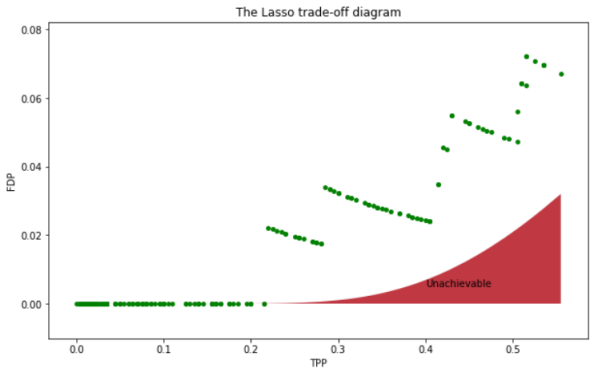

FIGURE 14 – $\rho = 0.3$

Lorsque l'on augmente encore davantage la corrélation, le nombre de sélection se réduit et le Lasso perd sa capacité à sélectionner les variables. Selon l'article sur lequel on s'est basé, cela est dû au fait qu'il faille adapter le coefficient de régularisation λ à cette corrélation. On peut le voir ici : pour la majorité des termes de régularisation, les coeffcients $\hat{\beta}$ sont nuls, ce qui fait que l'on n'ajoute aucun point au diagramme. Par contre, pour le peu de point que l'on a, on a un taux de faux positif de 0%, ce qui va dans le sens de l'article qui dit que la corrélation des variables n'est pas un problème pour le Lasso, si le terme de régularisation est bien choisi.

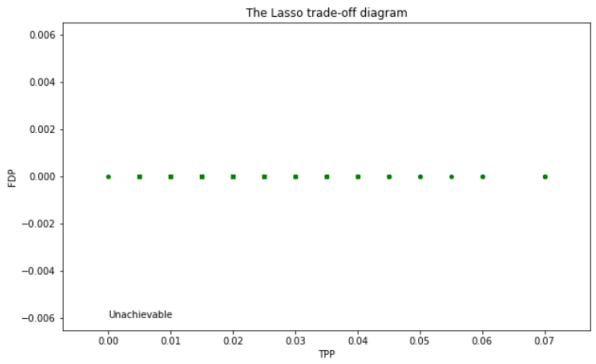

FIGURE 15 – $\rho = 0.5$

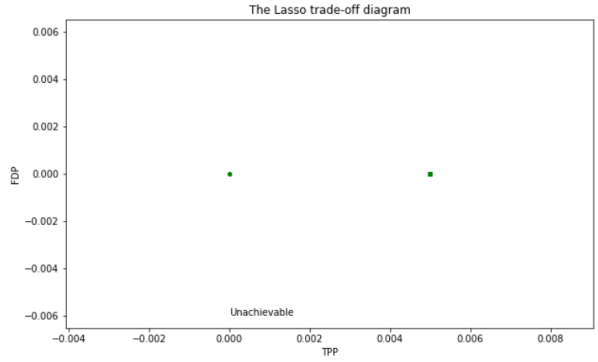

FIGURE 16 – $\rho = 0.9$

On ajoute maintenant à chaque colonne le même terme ηN, quand $\rho = 0.001$.

```
TPP au premier faux positif :  0.16
FDP lorsque TPP = 50% :  0.08256880733944955
FDP lorsque TPP = 100% :  0.13043478260869565
```

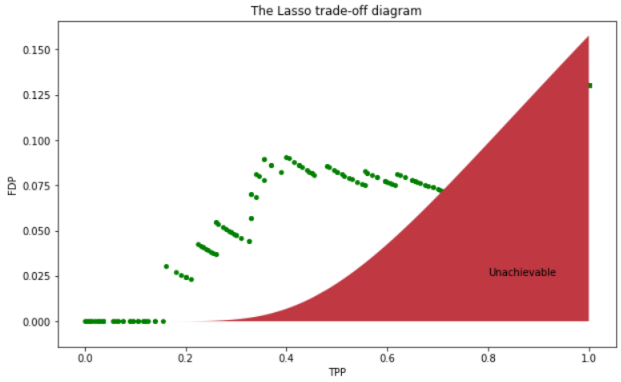

FIGURE 17 – $\eta = 0.001$

```
TPP au premier faux positif :  0.28
FDP lorsque TPP = 50% :  0.05660377358490566
FDP lorsque TPP = 100% :  0.0867579908675799
```

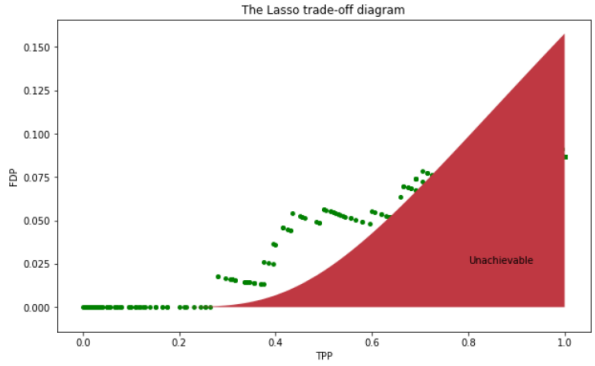

FIGURE 18 – $\eta = 0.01$

```
FDP lorsque TPP = 50% :  0.04672897196261682
```

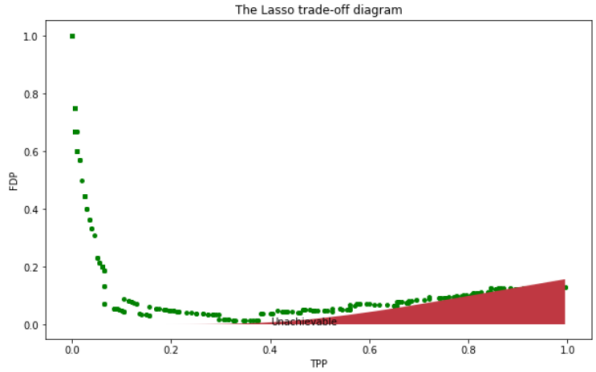

FIGURE 19 – $\eta = 0.1$

On peut voir que cela ne change pas grand chose sauf lorsque $\eta = 0.1$ et dans ce cas, on a un résultat assez étrange avec une courbe qui décroît, c'est-à-dire que le taux de faux positif est de 100% au début, puis décroît dans un premier temps avant de remonter un peu pour retrouver un niveau normal.

6.5 LASSO face au bruit

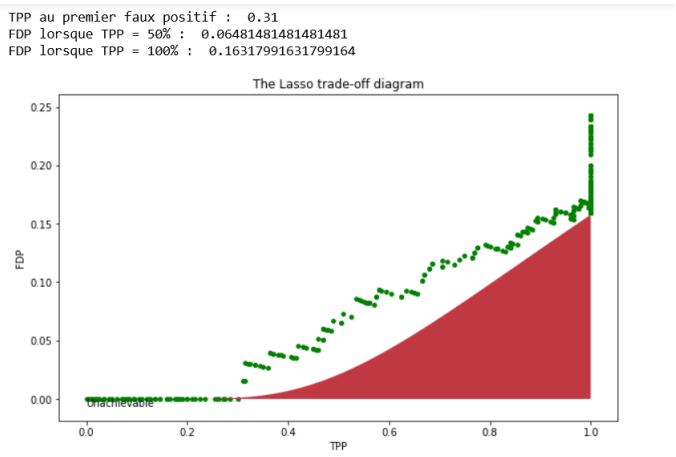

FIGURE 20 – Lasso path et Trade-off (avec du bruit)

Dans l'article, les auteurs rappellent plusieurs fois que leurs résultats s'appliquent aussi bien dans le cas où la variable à prédire est bruitée au non, c'est-à-dire qu'on lui ajoute un terme aléatoire, ici de variance $\sigma = 1$. La figure 20 confirme cela puisqu'on n'entre pas dans la zone intteignable. Cependant, la "boundary curve" décrit moins bien la trajectoire du Lasso que précédemment, l'espace entre la zone rouge et les points étant plus important que dans le cas sans bruit.

6.6 LASSO vs ENET

Le modèle Elastic Net pourrait sembler intéressant à première vue. En effet, on sait maintenant que le Lasso souffre d'un problème de mauvaise sélection intervenant assez tôt.

Dans l'article, le Lasso était comparé au Ridge qui, lui, ne souffrait pas de ce problème puisqu'il arrivait à sélectionner toutes les variables explicatives sans erreur. Toutefois, on sait que le Ridge n'est pas adapté pour sélectionner des variables car il ne va pas facilement mettre des coefficients β à 0. On pourrait donc espérer que Elastic Net combine le meilleur des deux méthodes.

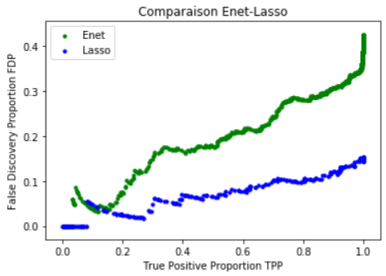

FIGURE 21 – Comparaison enet et lasso

La figure 19 permet de comparer le "Lasso path" avec le "Enet path". On peut voir que le modèle Enet souffre également de ce problème de mauvaise sélection qui intervient aussi tôt que pour le Lasso. En plus de cela, son taux d'erreur croit plus vite que pour le Lasso. Enet n'est donc pas plus performant que le Lasso pour sélectionner les variables explicatives dans un régime de sparsité linéaire pour des variables gaussiennes indépendantes.

7 Conclusion

L'article "False Discoveries occur Early on the Lasso Path" met en évidence les faiblesses et les limites du Lasso pour sélectionner les variables explicatives. Notre travail ne contredit pas les résultats qui y sont présentés mais cherche à nuancer ce constat.

En effet, il faut dire que les résultats ont été dérivés dans un cadre bien précis : dans un régime de sparsité linéaire avec des variables gaussiennes indépendantes en entrée. Ce cadre n'est ni le cadre préférentiel pour le Lasso, ni un cadre que l'on retrouve communément en pratique.

Nous avons donc cherché à nous en éloigner un peu en ajoutant de la corrélation entre les variables, ce qui est souvent le cas en pratique. On a pu constater que le Lasso s'en sortait assez bien et que cela pouvais même améliorer ces performances et lui permettre de dépasser les limites définies dans l'article, bien qu'une forte corrélation nécessite certains ajustements du terme de régularisation.

Enfin, le Lasso reste également assez résistant au bruit et performe mieux que d'autres modèles comme Elastic Net pour sélectionner naturellement les variables explicatives. Les résultats de l'article sont très intéressants car ils mettent en exergue les limites du Lasso, qui sont bien réelles, en les quantifiant et en les expliquant. Cependant, ils ne doivent pas nous détourner de ce modèle qui reste assez pratique et performant pour sélectionner des variables en grande dimension.