Master Data Science
Theoretical guidelines for high-dimensional data analysis

Chosen article : FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH

Exploratory study on false discoveries along the Lasso path

Authors :
Pierre ADEIKALAM
Mohamed ABDEL WEDOUD

Lecturer :
Christophe GIRAUD

February 16, 2020

Contents

1 Introduction

This report will discuss the results of the paper titled "False Discoveries occur Early on the Lasso Path" written by Weijie Sue, Malgorzata Bogdan and Emmanuel J. Candès in 2015.

The main result of this paper allows us to analytically predict the Lasso estimator's underperformance in feature selection in sparse settings. The motivation behind this study was to clear misconceptions about theoretical properties of the Lasso estimator that could have mislead practitioners who regularly use the Lasso estimator to perform prediction when the number of features exceeds the number of observations.

Indeed, the Lasso estimator is known for automatically weeding out irrelevant variables by setting their regression coefficients exactly to 0, therefore removing them completely from the prediction. Moreover, theoretical results from (Wainwright, 2009), (Meinshausen N. and Bühlmann P., 2006), (Zhao P. and Bin Y., 2006) and many others have shown that with high probability the Lasso estimator should exactly recover the sparsity pattern of any regression problem that satisfies specific conditions, even with an extreme amount of redundant features.

The authors highlight the weakness of these theoretical results by providing a simple and intuitive setting where with high probability, this property will not hold even when the regression problem is noiseless. In this setting, there is actually an explicit trade-off between the number of relevant variables detected by the Lasso estimator and the number of irrelevant variables that will be wrongly picked up. The highlight of the authors' results is that the dynamic of this trade-off is easily visualized by what is called the Lasso Path which we will clearly explain in the next section.

Our goal in this report is to (i) intuitively explain the authors' results, (ii) formulate a critique of their impact and their limitations, and (iii) showcase our personal exploration around the problem of variable selection by using the Lasso Estimator and its variants.

We wanted the context and key results section to be understandable by anyone who is slightly familiar with the mathematical formulation of the Lasso Estimator which is why, for ease of read, some terminology such as *design matrix* or *explanatory variable* has been simplified to "observation matrix" and "feature".

2 Context and key results

In this section we will intuitively explain and formalize the setting and the key results of the paper.

2.1 Context

2.1.1 The Setting

The model in this paper is the standard linear model:

$$y = X\beta + z$$

where $X \in \mathbb{R}^{n \times p}$ is an observation matrix such that n is the number of observations and p is the number of observed features, $\beta \in \mathbb{R}^p$ is the true vector of regression coefficients and $z \in \mathbb{R}^n$ is the observation noise.

The assumptions made on X, β and z are:

1. **Independent Gaussian Design** : The observations are independent normalized Gaussian variables, i.e.

$$X_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{1}{n}) \qquad \forall i, j \in \{1, .., n\} \times \{1, ..., p\}$$

2. **Linear sparsity** : We let $\epsilon > 0$ be the proportion of non-null regression coefficients, i.e.

$$\mathbb{P}(\beta_j \neq 0) = \epsilon \qquad \forall j \in \{1, ..., p\}$$

3. **Homoscedasticity** : The observation noise is a vector of independent Gaussian variables of constant variance $\sigma^2 \geq 0$, i.e.

$$z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \qquad \forall i \in \{1, ..., n\}$$

4. **Asymptotic regime**: We suppose that $p, n \to \infty$, with $\frac{n}{p} \to \delta$ for some $\delta > 0$.

For numerical experiments, we just set $n = \delta p$ and $|\{j : \beta_j \neq 0\}| = \epsilon p$ for given values of p, δ and ϵ.

In order to estimate β, we use the **Lasso Estimator** $\hat{\beta}(\lambda)$ given by:

$$\hat{B}(\lambda) \in \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|X\beta - y\|_2^2 + \lambda \|\beta\|_1 \qquad \lambda \geq 0$$

and we say that a variable of index j is **selected** if $\hat{\beta}_j(\lambda) \neq 0$.

2.1.2 The Lasso Path

For a given $\lambda_0 \ggg 1$, the Lasso Path denotes the sequence of Lasso estimators $(\hat{\beta}(\lambda_m))_{m \geq 0}$ such that $\lambda_m \searrow 0$.

In this paper, the purpose of the Lasso Path is to analyse the behavior of the Lasso Estimator for different values of λ thanks to the following metrics:

$$\text{FDP}(\lambda) := \frac{|\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j = 0\}|}{\max(|\{j : \hat{\beta}_j(\lambda) \neq 0\}|, 1)}$$

$$\text{TPP}(\lambda) := \frac{|\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ and } \beta_j \neq 0\}|}{|\{j : \beta_j \neq 0\}|}$$

FDP(λ) denotes the *False Discovery Proportion* and is the ratio between the number of **wrongly** selected variables and the **total** number of selected variables, which is **different** from the *False Positive Rate* practitioners and students are used to in the context of classification.

TPP(λ) denotes the *True Positive Proportion* and is the ratio between the number of **correctly** selected variables and the number of **non-null variables**, which intuitively represents the proportion of non-null variables that were discovered. This metric **does not depend on the False Positive Rate** which differs from the usual *True Positive Rate* in the context of classification.

We believe this clarification was necessary mostly for students who could be mislead by the name of these metrics.

Before presenting the key results, we would like to give some intuition about how these metrics should behave along the Lasso Path.

2.1.3 Intuition

Anyone that has used the Lasso Estimator knows that for large values of λ, the estimated coefficients will all be exactly 0. This is due to the fact that the regularisation term $\lambda\|\beta\|_1$ promotes sparsity of the solution and heavily outweighs the squared term for large values of λ.

Now, if for some $\lambda_0 \ggg 1$, $\hat{\beta}_j(\lambda_0) = 0$ for all j, then necessarily $\text{TPP}(\lambda_0) = \text{FPP}(\lambda_0) = 0$ since no variable will be selected.

At the other end of the spectrum, if we set $\lambda_M = 0$, then $\hat{\beta}(\lambda_M)$ will be the ordinary least squares estimator. In our Gaussian setting, it can be proven that the probability for any $\hat{\beta}_j(\lambda_M)$ to be equal to 0 is 0, meaning that all variables will be selected. Therefore, almost surely $\text{TPP}(0) = 1$.

Hence, as λ decreases, we can expect the *True Positive Proportion* to start at 0 and steadily increase until all variables are selected and the *True Positive Proportion* is equal to 1.

How does the *False Discovery Proportion* evolve as λ decreases? Practitioners and students could expect the Lasso Estimator to be robust enough to first select **most** of the relevant variables and then as we decrease λ further we might start to wrongly select some variables. This kind of behaviour would suggest that *False Discoveries occur **Late** on the Lasso Path*.

However, the name of the paper "*False Discoveries occur Early on the Lasso Path* suggests that this kind of behavior **does not** happen and instead false discoveries begin to appear **before** we have selected most of the relevant variables.

2.2 Key results

2.2.1 An Intuitive Illustration

In Figure 1 we will look at the values of TPP and FDP along the Lasso Path on a simulated dataset. To compute this, we simulate a dataset X with the parameters $p = 1000$, $\delta = 1$, $\epsilon = 0.2$ and $\sigma^2 = 1$.

We remind that $\delta = 1$ means that $n = p\delta = p$, $\epsilon = 0.2$ means that $\epsilon p = 200$ coefficients of β are non-zero and σ^2 is the variance of the observation noise z.

Then, we fit many Lasso estimators for values of λ ranging between 10^{-5} and 10^2 and for each of these estimators we compute their TPP and FDP.

From this plot, we can see the trade-off the authors were talking about. Indeed, for the TPP to grow and reach 1, it is necessary for FDP to grow as well.

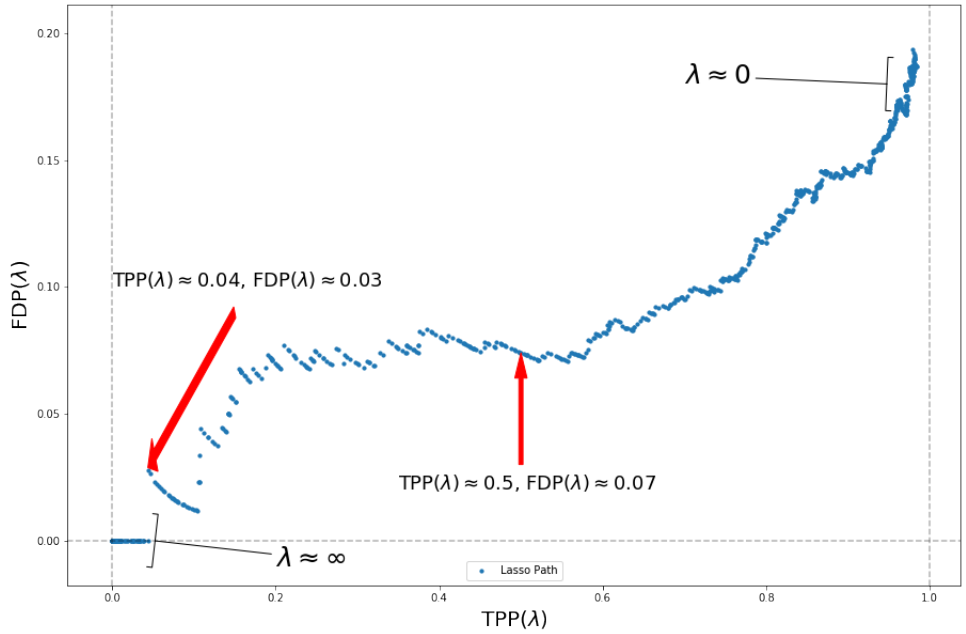

Figure 1: Lasso Path obtained with $p = 1000$, $\delta = 1$, $\epsilon = 0.2$ and $\sigma^2 = 1$.

We can see that the first false discoveries happen as soon as TPP reaches 0.04, meaning we have only detected 4% of the relevant variables. When the TPP reaches 0.5, the FDP has grown to 0.07, and when the TPP is about to reach 1, the FDP is almost at 0.20.

The Lasso Estimator does indeed select relevant variables but at the cost of making false discoveries.

From this observation, the authors have been able to prove that the FDP can actually be analytically bounded from below with high probability, therefore proving that false discoveries **must** happen for the TPP to grow.

In Figure 2, we show the same path along with the authors' theoretical lower bound. The code to obtain this figure is also available on the Jupyter Notebook associated with this report.

This lower bound shows that even in the best case scenario, with high probability false discoveries will start to happen before TPP reaches 0.4. There is almost no scenario in which the TPP reaches 1 and the FDP is kept under control.

We will now look at the theorem of the paper that formalizes this observation.

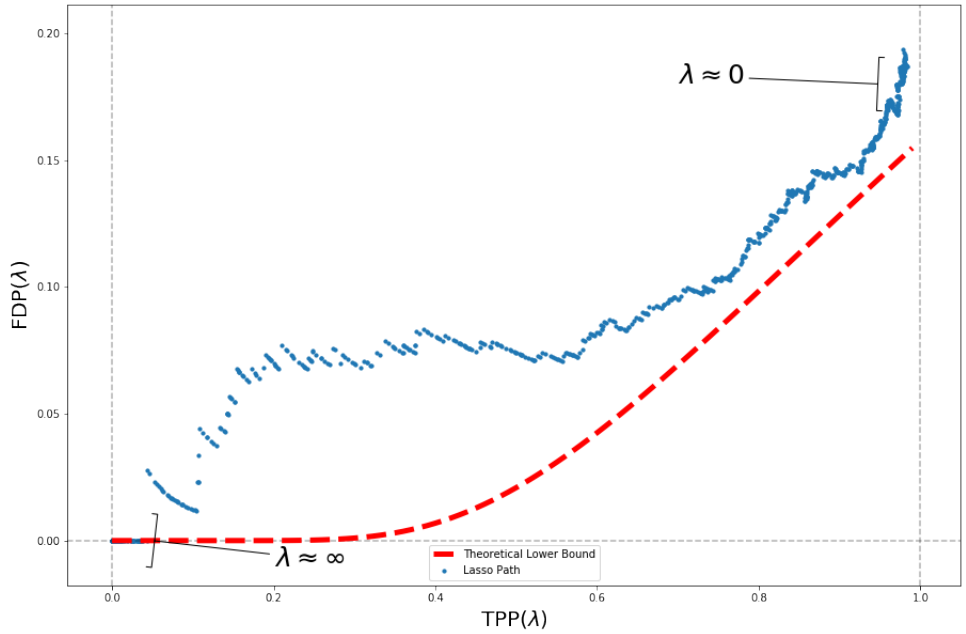

Figure 2: Lasso Path obtained with $p = 1000$, $\delta = 1$, $\epsilon = 0.2$ and $\sigma^2 = 1$. The lower bound is analytical and has been computed thanks to the authors' main result.

2.2.2 Main Result and Interpretation

We will start this subsection by writing the analytical formulation of the boundary curve we have shown in Figure 2.

Fix $u \in [0,1], \delta \in (0,\infty), \epsilon \in (0,1)$. The lower bound is actually a function of u, δ and ϵ that the authors denote by q^* and its computation requires 2 steps.

Step 1: We denote Φ the cumulative distribution function of a standard Gaussian and ϕ its probability density function. The first step of computing $q^*(u,\delta,\epsilon)$ consists in finding $t^*(u,\delta,\epsilon)$, the largest positive root of the equation in t given by:

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

While finding the root of this equation might look scary on paper, it is possible to compute it numerically with a standard *Bracketing Method* and code for doing

so has been provided by the authors at `https://github.com/wjsu/fdrlasso`.

Step 2: Now that we have computed $t^*(u)$, computing $q^*(u)$ is straightforward as its formula is given by:

$$q^*(u) = \frac{2(1-\epsilon)\Phi(-t^*(u)}{2(1-\epsilon)\Phi(-t^*(u)) + \epsilon u}$$

Plotting $q^*(\text{TPP}(\lambda))$ for TPP(λ) between 0 and 1 gives us the lower bound in Figure 2.

Now the main result:

Theorem: Let $\delta \in (0, \infty)$, $\epsilon \in (0, 1)$, $\sigma^2 \geq 0$ and consider the function q^* as defined above. Then, under the assumptions made in Section 2.1.1 and for any arbitrary constants $\lambda_0 > 0$ and $\eta > 0$, we have:

$$\mathbb{P}(\forall \lambda \geq \lambda_0, \text{FDP}(\lambda) \geq q^*(\text{TPP}(\lambda)) - \eta) \overset{\substack{n,p\to\infty \\ n/p\to\delta}}{\longrightarrow} 1$$

This theorem confirms that it is very unlikely that a Lasso Estimator can beat this lower bound and the probability for it to happen decreases to 0 as n and p grow larger, **regardless of the variance of the noise or the magnitude of the non-null true coefficients**.

3 Critical article analysis

In this section we will try to understand the true reach of these results and understand the impact they would have on our daily use of the Lasso.

3.1 Independence Assumption

The assumption that the $X_{i,j}$ were *i.i.d*. seemed strange at first because considering variables that are completely independent from each other would sometimes be kind of odd in real life. Obviously, if one were to predict the temperature, they would not add features like the price of bitcoin, the distance with the closest galaxy, etc... However, those features, do have some hidden interactions (No matter how small they are). For example: bitcoin issuing consumes a lot of energy conducting some minor effects on the climate.

3.2 Gaussian Design

This assumption is one that seems very odd as it does not make practical sense to have $Xi, j \sim \mathcal{N}(0, \frac{1}{n})$. Indeed, this kind of design would mean that $X_{i,j} \xrightarrow{L^2} 0$, thus destroying the signal that was present in the data.

Furthermore, in practice we would hope that after standardization $X_{i,j} \approx \mathcal{N}(0,1)$, not $\mathcal{N}(0, \frac{1}{n})$, so there is no reason a priori that this framework should generalize to real problems. We will show by proof in Section 5 that this design might completely explain the authors' main result.

3.3 Expressivity of the lower bound

When looking at other theoretical results from the litterature that imply that the Lasso has perfect support recovery capabilities, we see that one of the main hypothesis is that (i) the sparsity pattern s of the true coefficients is deterministic and of fixed size. Furthermore, one type of assumption is the (ii) *Mutual Coherence Condition* (Lounici, 2008) that implies that $\frac{1}{n}(X^T X)_{j,j} = 1$ and $\frac{1}{n}(X^T X)_{i,j} < \rho$ for some $0 < \rho < \frac{1}{cs}$ for some $c > 0$. This kind of assumption could not work if we were to simulate data with $Xi, j \sim \mathcal{N}(0, \frac{1}{n})$.

Now, if we try to compare these assumptions with the authors' framework, we find that if we simulate X with $X_{i,j} \sim \mathcal{N}(0,1)$ then $\frac{1}{n}(X^T X) \approx I_n$. Then, if we set ϵ very small, we get a perfect recovery of the support.

We wanted to better understand this lower bound given by q^*. For this we compute q^* between 0 and 1 for many ϵ and δ, and we look at the median of q^* over the segments where $q^* > 0$. The results are displayed in Figure 3.

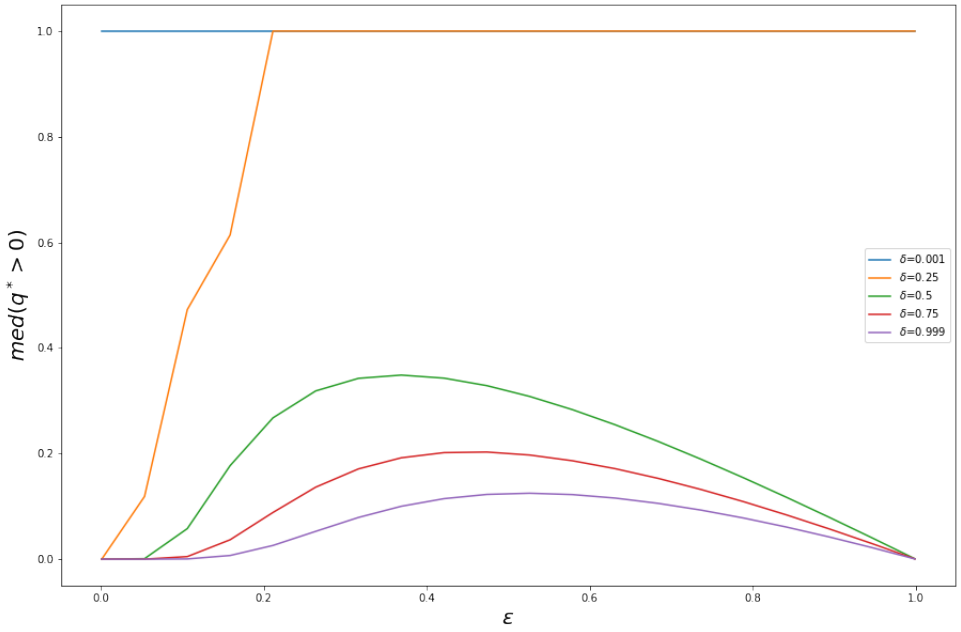

Figure 3: Computation of the median of the lower bound q^* when $q^* > 0$ for different values of ϵ and δ.

What we see is that for small values of $\epsilon < 0.1$ and $\delta > 0.5$, q^* is equal to 0 **everywhere**, meaning that this lower bound has no expressivity for small values of epsilon. When we compute the Lasso path for these values it gets a perfect recovery.

3.4 Convergence notion

The statements a) and b) of Theorem 1 of the article could be somehow confusing about the type of convergence. Indeed, if someone does not care about the details, he or she may understand that the events in a) and b) converge almost surely, but sadly the statement, as it is, stipulates only a convergence in probability which is a lot weaker.

for a fixed λ_0 and a random path (FDP,TPP) at step n set :
$A_n = \{\forall \lambda \geq \lambda_0, \eta > 0, \text{FDP}(\lambda) \geq q^*(\text{TPP}(\lambda)) - \eta\}$

The theorem says that $\mathbb{P}(A_n) \xrightarrow{n\to\infty} 1$, which is different from saying that $\mathbb{P}(lim\,inf\ A_n) = 1$.
$lim\,sup\ A_n$ would be the set of singular events $\{\omega\}$ that occur infinitely often and $lim\,inf\ A_n$ is the set of singular events that start at a certain rank and then never stop occurring. It is clear that $lim\,inf\ A_n \subset lim\,sup\ A_n$, but we could not prove if there is an almost sure event at infinity or not, because the proofs in the article don't provide enough details about probability values for a given intermediary n. However a classical scheme consist of using Borel-Cantelli theorem : if $\mathbb{P}(lim\,sup\ A_n^c) = 0$ then $\mathbb{P}(lim\,inf\ A_n) = 1$ and $\mathbb{P}(lim\,sup\ A_n) = 1$. The Borel-Cantelli theorem tells that if we manage to demonstrate that $\sum \mathbb{P}(A_n^c) < +\infty$ then we have the wanted result. This observation has partially motivated both our numerical and theoretical exploration in Sections 4 and 5.

4 Numerical exploration

4.1 Almost sure event numerical check

As we could not mathematically prove the almost sure event, we tried to approximate the convergence curve of the evolution of $P(A_n^c)$ with respect to n, for given values of ϵ and δ. This approximation is done using Monte Carlo methods : for a fixed n, we simulate N Lasso paths and then compute the proportion of paths that are not included in A_n. We assumed that as $\frac{n}{p} \to \delta$, taking $\frac{n}{p} = \delta$ would not drastically change the results.

ALGORITHM:

- Input = p, δ, ϵ, λ_0,N_λ : number of paths to simulate, σ^2: variance of observation noise , N_λ: number of λ's for which we compute the Lasso path.
- Lasso path : returns a simulation of the ($\lambda \to$ (TPP(λ),FDP(λ)) curve.
- $q^\star$: return the values y of $q^\star$ evaluated on a set of values x given δ and ϵ.
- fdr* : takes as argument a TPP list and computes $q^\star(TPP_i)$ for every TPP_i in the list. Since TPP is random we approximate $q^*(\text{TPP}_\text{i})$ with: for j such that $x_j \leq \text{TPP}_\text{i} \leq x_{j+1}$:

$$q^\star(\text{TPP}_\text{i}) \approx \frac{(\text{TPP}_\text{i} - x_j)(y_{j+1} - y_j)}{(x_{j+1} - x_j)} + y_j \tag{1}$$

- Output = Approximation of $\mathbb{P}(A_n^c)$, $n = \delta \times p$

The computation of $q^\star$ is very expensive which is why we compute it over an array x first and then use Formula 1.

In order to check whether there is an almost sure event or not, we applied the algorithm above, for given values of arguments and for n ranging in {800,1000,1500,1800,2000,2300,2500,2800,3000,3200,3500}. We repeated the experience twice, as we compute a Monte-Carlo approximation to see if the result is qualitatively stable (see Figure 4).

We then repeated the same experience for other parameter values and it

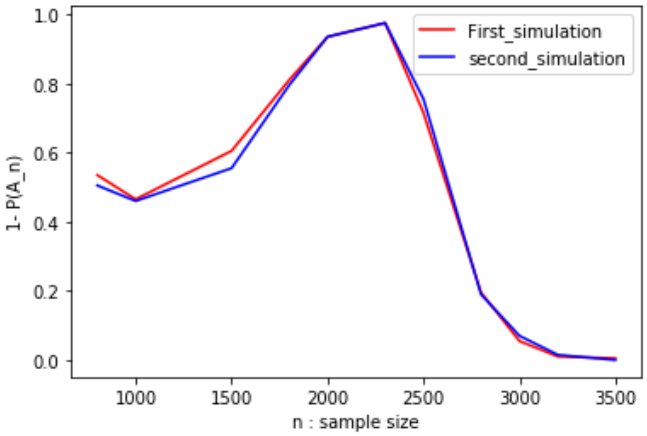

Figure 4: Approximation of $\mathbb{P}(A_n^c)$ for different values of n and parameter values $\lambda_0 = 10^{-3}$ $N_\lambda = 200$, $\epsilon = 0.2$, $\delta = 1$

seems that a bump appears each time. The probability of the event that TPP goes below the bound increases for smaller values of n but starts to decrease very rapidly at after a certain value n_0 of n. We also noticed that the Monte-Carlo approximations seem to be tight (The first and second simulations are very close to each other, and get even closer as n get bigger).
To understand the behavior of the sum as in the end of subsection 3.4, we decided to fit the curve to some asymptotic speed functions $\frac{1}{n^\gamma}$ with different values of γ to see which dynamics the curve most likely obeys to. That fitting was done for values of $n \geq n_0$, as the function starts decreasing fast. The results we obtain are the following:

γ	F-statistic	P-value
1.0	0.6797	0.4472
1.25	0.7573	0.424
1.5	0.8187	0.407
1.75	0.8655	0.3949
2.0	0.9	0.3863

From the table we deduce that the larger γ gets the better both the F-statistic and the P-value become. This means that the speed of convergence is likely to be stronger than a certain value $\gamma_0 > 1$ of γ and as a consequence, numerically $\sum \mathbb{P}(A_n^c)$ seems to be convergent. The strength of such result lays in the fact that some events will start at a certain rank and never stop occurring.

4.2 The case of dependent features

Until now we were working under the assumption that all the variables in our data were independent. One may ask: is it possible to have the same result for correlated X features? As in the previous section, we did not feel too comfortable with the mathematics involved in this problem so we wanted to confirm any intuition numerically. Therefore, we will once more be making some visual experiments to see the asymptotic behavior of the Lasso Path. Moreover, to make things slightly easier we will be assuming linear dependence between the features.

To do so, we will define a dependence degree θ, such that θ is a divider of p ($\frac{p}{\theta} = \kappa \in \mathbb{N}$). Then, we consider the same setting as in the article but instead of generating $X \in \mathbb{R}^{n \times p}$, we will only simulate a smaller matrix $X^{'} \in \mathbb{R}^{n \times \kappa}$. Finally, we make θ copies of $X^{'}$, and set X as the concatenation of these copies.

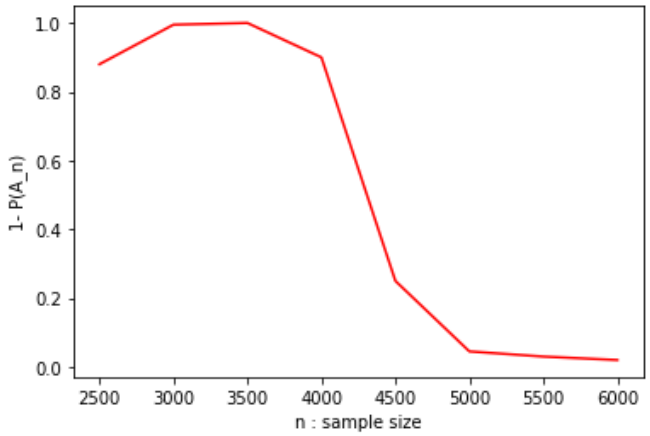

Figure 5: Estimation of $\mathbb{P}(A_n^c)$ with parameter values $\lambda_0 = 10^{-3}$, $N_\lambda = 200$, $\epsilon = 0.2, \delta = 1.0, \theta = 2$

We drew in Figure 6 the evolution of the probability of the event A_n^c, substituting δ by $\delta \times \theta$ in $q^\star$: that would be for a fixed λ_0, and a path (FDP,TPP) a step n,
$A_n = \{\forall \lambda \geq \lambda_0, \eta > 0, \text{FDP}(\lambda) \geq q^*(\text{TPP}(\lambda), \delta \times \theta, \epsilon) - \eta\}$

We did the same for the case where δ remains unchanged and we noticed that for big values of n, the probability of A_n^c is almost 1 for every sampled n. These experiments could guide us into finding $q^\star$ such that points (a) and (b)

of Theorem 1 remain valid. However, these numerical results are still far from concluding on the existence of a tight function (part (d) of Theorem 1).

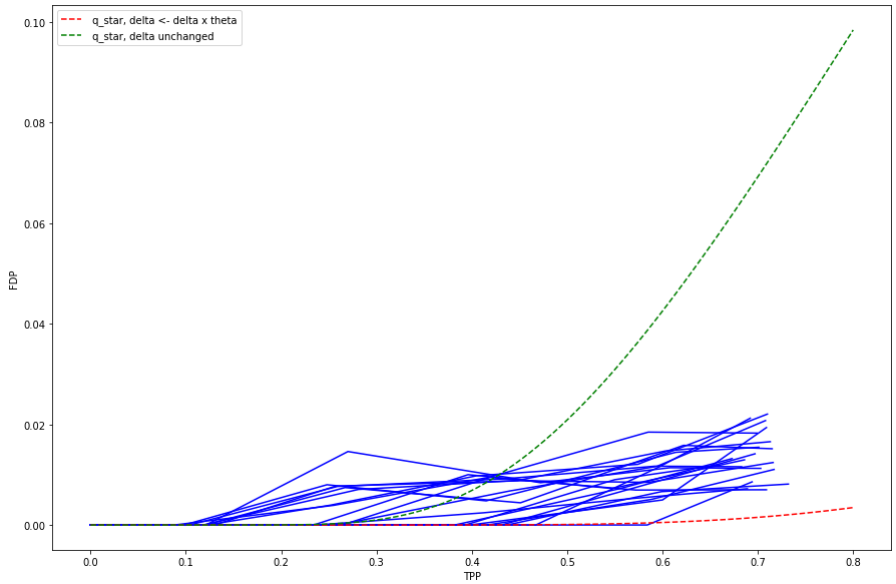

Figure 6: Simulation of 20 paths with parameter values $\lambda_0 = 10^{-3}$, $\epsilon = 0.2$, $\delta = 1.0$, $\theta = 2$, $n = 5000$, in red $q^\star(., \theta \times \delta, .)$, and in green $q^\star(., \delta, .)$

As we can see in Figure 6, all generated paths get at some point under the function $q^\star(., \delta, \epsilon)$, while very few of them do with the function $q^\star(., \delta \times \theta, \epsilon)$. This would suggest that linear dependency of the features benefits the Lasso Estimator.

5 Theoretical exploration

5.1 Theoretical issue with the design

In this subsection, we prove that having $X_{i,j} \sim \mathcal{N}(0, \frac{1}{n})$ may completely explain the authors' result.

Let $\lambda_0 > 0$. From the KKT conditions, we have that

$$0 = \hat{\beta}(\lambda_0) \iff \forall j, |\frac{1}{n} X_j^T y| \leq \lambda_0$$

We assume $\sigma^2 = 0$ (noiseless case of the main result), i.e $z = 0$ and β is bounded.

We have:

$$\frac{1}{n} X_j^T X_j = \frac{1}{n} \sum_{i=1}^{n} X_{j,i}^2 \sim \frac{1}{n^2} \sum_{i=1}^{n} Z_i^2$$

where $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Therefore, $\frac{1}{n} X_j^T X_j \sim \frac{1}{n^2} \chi^2(n)$. Hence,

$$E[\frac{1}{n} X_j^T X_j] = \frac{1}{n^2} n = \frac{1}{n}$$

$$\text{Var}(\frac{1}{n} X_j^T X_j) = \text{Var}(\frac{1}{n^2} \chi^2(n)) = \frac{1}{n^4} 2n = \frac{2}{n^3}$$

Let $i \neq j$, we have:

$$\frac{1}{n} X_j^T X_i = \frac{1}{n} \sum_{k=1}^{n} X_{j,k} X_{i,k} \sim \frac{1}{n^2} \sum_{k=1}^{n} Z_k \tilde{Z}_k$$

where $Z_k, \tilde{Z}_k \overset{i.i.d}{\sim} \mathcal{N}(0,1)$. Moreover, we have:

$$\begin{aligned} \frac{1}{n^2} \sum_{k=1}^{n} Z_k \tilde{Z}_k &\sim \frac{1}{n^2} \sum_{k=1}^{n} \frac{1}{4} (Z_k + \tilde{Z}_k)^2 - \frac{1}{4} (Z_k - \tilde{Z}_k)^2 \\ &\sim \frac{1}{2n^2} \sum_{k=1}^{n} W_k^2 - \tilde{W_k}^2 \end{aligned}$$

where $W_k, \tilde{W}_k \sim \mathcal{N}(0,1)$. W_k and $\tilde{W}_k$ are independent because they are Gaussian and their covariance is 0. Therefore,

$$\frac{1}{n^2} \sum_{k=1}^{n} Z_k \tilde{Z}_k \sim \frac{1}{2n^2} \chi^2(n) - \frac{1}{2n^2} \tilde{\chi}^2(n)$$

Hence,

$$\mathbb{E}[\frac{1}{n}X_j^T X_i] = 0$$

$$\text{Var}(\frac{1}{n}X_j^T X_i) = 2\frac{1}{4n^4}2n = \frac{1}{n^3}$$

Now, we also note that:

$$\forall i \neq j, \mathbb{E}[X_j^T X_j X_j^T X_i] = \mathbb{E}[X_j^T X_j X_j^T]\mathbb{E}[X_i] = 0$$

.

In conclusion:

$$\mathbb{E}[\frac{1}{n}X_j^T X\beta] = \frac{1}{n}\beta_j$$

$$\begin{aligned}
\text{Var}(\frac{1}{n}X_j^T X\beta) &= \text{Var}(\frac{1}{n}X_j^T X_j\beta_j) + \sum_{i\neq j}\text{Var}(X_j^T X_i\beta_i) \\
&= \frac{2}{n^3}\beta_j^2 + \sum_{i\neq j}\frac{2}{n^3}\beta_i \\
&\leq \frac{2}{n^3}\|\beta\|_\infty^2 + 2\frac{p-1}{n^3}\|\beta\|_\infty^2 \\
&\leq \frac{2}{\delta n^2}\|\beta\|_\infty^2
\end{aligned}$$

with $\delta = \frac{n}{p}$. From Bienaymé-Tchebychev inequality we have that:

$$\mathbb{P}(|\frac{1}{n}X_j^T X\beta - \frac{1}{n}\beta_j| > \lambda_0) \leq \frac{2}{\lambda_0\delta n^2}\|\beta\|_\infty^2$$

Furthermore, if we assume without loss of generality that $\beta_j \geq 0$, we have that:

$$\begin{aligned}
\mathbb{P}(|\frac{1}{n}X_j^T X\beta - \frac{1}{n}\beta_j| > \lambda_0) &\geq \mathbb{P}(\left||\frac{1}{n}X_j^T X\beta| - |\frac{1}{n}\beta_j|\right| > \lambda_0) \\
&\geq \mathbb{P}(|\frac{1}{n}X_j^T X\beta| > \lambda_0 + \frac{1}{n}\beta_j)
\end{aligned}$$

Hence,

$$\mathbb{P}(|\frac{1}{n}X_j^T X\beta| > \lambda_0 + \frac{1}{n}\beta_j) \leq \frac{2}{\lambda_0\delta n^2}\|\beta\|_\infty^2 \overset{n\to\infty}{\longrightarrow} 0$$

From an union bound argument we have:

$$\mathbb{P}(\cup_{1\leq j\leq p}|\frac{1}{n}X_j^T X\beta| > \lambda_0 + \frac{1}{n}\beta_j) \leq \frac{2\epsilon p}{\lambda_0 \delta n^2}\|\beta\|_\infty^2 \leq \frac{2\epsilon}{\lambda_0\delta^2 n}\|\beta\|_\infty^2 \overset{n\to\infty}{\longrightarrow} 0$$

Thus, from the Monotone Convergence Theorem (as $n = \delta p$) we deduce that:

$$\forall \lambda_0 > 0,\ \ \mathbb{P}(\lim_{n\to\infty} \hat{\beta}(\lambda_0) = 0_p) = 1$$

If we let $\tilde{\beta} = 0_p$, we have that FDP($\tilde{\beta}$) = 0 and TPP($\tilde{\beta}$) = 0.

Since, $q^*(0) = 0$ (p.7 of the paper), we have that $\text{FDP}(\tilde{\beta}) > q^*(\text{TPP}(\tilde{\beta})) - \eta_0$, for all $\eta_0 > 0$, which gives the first result of the paper's Theorem 1. We want to emphasize that this was only possible because $X_{i,j} \sim \mathcal{N}(0, \frac{1}{n})$. If $X_{i,j} \sim \mathcal{N}(0,1)$, we would not have been able to conclude this way.

Even though we may have proved this only for the noiseless case, then if we did no mistake we feel like this kind of proof might be generalized to the case with noise as the observation noise only increases the variance of the estimator but not its expectation.

5.2 Comparison with the standard linear model

There is a detail that make the result of the article work : the fact that the variance of $X_{i,j}$ is $\frac{1}{n}$. In order to see the effect of this assumption on the Lasso estimator with a fixed n (not at the limit), we will be trying to find some analogies with classical results.

Consider y, X, β, σ^2 and z to be the same as in the setting of the article (see section 2.1.1). Now set $X' = \sqrt{n}X$ and $y' = \sqrt{n}y$. Notice that $X'_{i,j} \sim N(0,1)$ and that $y' = X'\beta + \sigma\sqrt{n}z$.

Let $\hat{\beta}(\lambda)$ be the Lasso Estimator on (X, y) and let $\hat{\theta}(\lambda)$ be the Lasso Estimator on (X', y').

We have that

$$\begin{aligned}
\hat{\beta}(\lambda) &= \underset{\beta}{\text{argmin}} \quad \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \\
&= \underset{\beta}{\text{argmin}} \quad n\|y - X\beta\|_2^2 + n\lambda\|\beta\|_1 \\
&= \underset{\beta}{\text{argmin}} \quad \|y' - X'\beta\|_2^2 + n\lambda\|\beta\|_1
\end{aligned}$$

$$\Rightarrow \hat{\beta}(\lambda) = \hat{\theta}(n\lambda) \tag{2}$$

As X' complies with the *Mutual Coherence Condition* with high probability (classical result), this observation would give us an intuition as to why the behavior of the Lasso Estimator in this paper is so different. However, unlike for the previous section, we have made no assumptions on either β or σ^2. Therefore, analyzing the behavior of $\hat{\theta}(n\lambda)$ would allow us to better understand it in the context of the papers supporting the Lasso's full support recovery capabilities.

Fortunately, this is a simple formalism that reinforces our intuition from the previous subsection (5.1) that we should be able to generalize our result to the case where $\sigma^2 > 0$ and β only has a finite second moment as we feel that $\hat{\theta}(n\lambda)$ should converge to 0 as n goes to infinity.

6 Conclusion

The argument of a possible trade-off between detecting the true coefficients of a model and making false discoveries is compelling due to the ever imperfect nature of machine learning and statistical estimators.

In this paper, the authors give a scenario where the Lasso Estimator cannot be relied upon to perfectly recover the support of a model. Their conclusion will always be of value as it forces us to be more careful with the assumptions we make every day when working with high dimensional problems. Indeed, since at first this result seems to contradict others, upon closer look of the assumptions we saw numerically that these results are actually compatible.

Moreover, the fact that their bound is computable and easily visualized is also very satisfying as it gives us a quick intuition of their result.

However, making us check our assumptions also made us realize that the author's model is different from the one we use regularly, especially when it comes to the variance of the simulated data. This detail has put us in a situation where we had to do our own mathematical investigation to understand the optimality of the Lasso Estimator in this context.

References

[1] By Weijie Su and Ma lgorzata Bogdan and Emmanuel Candes : *FALSE DISCOVERIES OCCUR EARLY ON THE LASSO PATH*, November 2015

[2] Peng Zhao and Bin Yu : *On Model Selection Consistency of Lasso*, April 2006

[3] Karim Lounici : *Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators*, January 2008

[4] Nicolai Meinshausen and Peter Buhlmann : *HIGH-DIMENSIONAL GRAPHS AND VARIABLE SELECTION WITH THE LASSO*, 2006

[5] Martin J. Wainwright: *Sharp thresholds for high-dimensional and noisy sparsity recovery using $\ell 1$-constrained quadratic programming (Lasso)*

[6] Alexandre Tsybakov : *Apprentissage Statistique et Estimation Non-Parametrique (In French)*,2013