# False Discoveries occur Early on the Lasso Path

Ramzi MISSAOUI
ramzi.missaoui@polytechnique.edu

Soufiane MOUTEI
soufiane.moutei@polytechnique.edu

## Abstract

*This report is written in the context of the the course Theoretical guidelines for high-dimensional data analysis, under the supervision of Christophe Giraud. The goal of this report is to present a detailed analysis of the article "False Discoveries occur Early on the Lasso Path (https://arxiv.org/abs/1511.01957), written by Weijie Su, Malgorzata Bogdan and Emmanuel Candes in November 2015, the most recent version being dated September 2016. We will try to grasp the key points, discuss the main results then talk about some limitations of the Lasso model.*

*We'll have a quick reminder about the regression problem and the Lasso Method. After that, we will highlight the Lasso path and variable selection and talk about the theorem and its hypotheses. After that, well highlight the limitations of the lasso and the issues caused by this type of regulation.*

*We also provide an implementation and a numerical analysis in the attached notebook which can be found in the github address provided below [1] The aim was to reproduce some of the paper's experiments to highlight and check the authors's results.*

## Contents



## 1. Introduction

Statisticians have been facing many challenges when dealing with modern datasets in different fields such as finance, genetics, imaging which amount to dealing with linear system of equations with more unknown variables than equations, let alone the fact that the equations are often corrupted by noisy measurements. This is where the lasso model comes to the rescue when dealing with high dimensional data.

[1] https://github.com/rmissaoui/theoretical-guidelines-for-high-dimensional-data-analysis

Lets first have a look at the classical version of the problem. The data is usually represented by the observation couples $(x_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}, i = 1..n;$. The goal is to find a regression function relating the predictors $x_i$ and the response $Y_i$. The regression problem is formulated as follows:

$$Y_i = \sum_{j=1}^{p} \beta_j X_{i,j} + \epsilon_i, \forall i = 1...n$$

where:

- $\epsilon$ is a vector of n independent and identically distributed random variables corresponding to a noise term.
- $X \in \mathbb{R}^n \times \mathbb{R}^p$ is the predicor variable where the $j^{th}$ column $X_j$ corresponds to the n observations of the $j^{th}$ predictor.
- $\beta \in \mathbb{R}^p$ is the unknown vector of coefficients for the regression function to be estimated using the data.

Under this configuration, and assuming that X has a full column rank, $\beta$ is estimated using the ordinary least squares technique, which yields:

$$\hat{\beta} = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - Xb\|_2^2 = \left(X^T X\right)^{-1} X^T Y$$

Then assuming that $\mathbb{E}\epsilon_i = 0$ and $cov(\epsilon_i, \epsilon_j) = \sigma^2 \mathbb{1}_{i=j}$, the model is evaluated using the mean squared prediction error (MSPE) giving:

$$MSPE(\hat{\beta}) = \mathbb{E}\left(\frac{1}{n}\|X\left(\beta - \hat{\beta}\right)\|_2^2\right) = \frac{p}{n}\sigma^2$$

We can see here that the ordinary least squares method works pretty well in the case where $p \ll n$.

However, in high dimensional data, where this particular condition is usually not satisfied, we need to work on a new method to tackle this problem.

### 1.1. How about having a sparse estimator $\beta$ ?

A sparse estimator which leads to select the most important variables and estimate their coefficients appears to be a practical solution, thus bringing us back to the previous case.

To ensure this sparsity, one might think about penalizing the number of non zero coefficients of the estimator in the below fashion:

$$\hat{\beta} = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \left(\frac{1}{n}\|Y - Xb\|_2^2 + \lambda\|b\|_0\right)$$

Where $\|b\|_0 = \sum_{k=1}^{p} \mathbb{1}_{b_k \neq 0}$, the number of nonzero components.

$\lambda$ is chosen in a way to regulate the amount of nonzero elements in $\beta$, small values increase sparsity and vice-versa.

The previous problem is equivalent to:

$$\hat{\beta} = \underset{\beta \in \beta_0[k]}{\operatorname{argmin}} \left(\frac{1}{n}\|Y - Xb\|_2^2\right)$$

where $\beta_0[k] = \{\beta; \|\beta\|_0 < k\}$

This is clearly a non convex problem, proceeding that way would require to evaluate the objective for all the k possible combinations of $\beta$ which includes $\binom{n}{k}$ cases $\forall k = 1..p$. As k is unknown, one would end up testing $2^p = \sum_{k=1}^{p} \binom{n}{k}$ cases which is not feasable given the big values p can take.

**Can we convexify this problem?** The answer is yes! By considering the $l_1$ norm instead of the $l_0$ norm, we happen to exactly deal with the Lasso estimator, which is the topic of the next section.

### 1.2. Lasso presentation

In the context of high dimensional data, where the number of features p is greater than the number of observations n, Lasso (Least absolute shrinkage and selection operator) can replace the ordinary least squares method.

One of the important features of the Lasso, especially compared to other techniques such as Ridge regression, is that it automatically reduces the number of variables. Indeed, the Lasso produces models where a large number of the regression coefficients are estimated to be equal to zero, if not most, thereby selecting the important variables.

Here is the formulation of Lasso estimator:

$$\hat{\beta} = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \left( \frac{1}{n} \|Y - Xb\|_2^2 + \lambda \|b\|_1 \right)$$

where $\|b\|_1 = \sum_{i=1}^{p} |b_i|$ and $\|x\|_2^2 = \sum_{i=1}^{n} x_i^2$

The hyper parameter $\lambda$ controls the strength of the regulation. Let's have a look at the extreme cases:

- $\lambda = 0 \Rightarrow$ ordinary linear regression if $p \leq n$.
- $\lambda = +\infty \Rightarrow$ All the coefficients of $\beta$ are zero.
- The increase of $\lambda$ induces the decrease of some coefficients from $\hat{\beta}$ ending in the previous case scenario when it's big enough.

The previous lasso estimator can be reformulated as follows:

$$\hat{\beta}(t) = \underset{b \in \mathbb{R}^p, \|b\|_1 \leq t}{\operatorname{argmin}} \|Y - Xb\|_2^2$$

In particular, this allows us to visualize in figure 1, in 2 dimensions, the case where Lasso cancels one component ($\beta 1$), thus eliminates the predicor $x_1$, and shrinks the other ($\beta 2$) .

The ellipses represent the counter lines for the function: $\beta \mapsto \|Y - X\beta\|_2^2$. The left green square is nothing but the area where $|\beta_1| + |\beta_2| \leq t$ for a fixed $t$. We can observe that the lasso constraint has corners. Those pointy edges increase the chance of having a sparse solution.

The green circular area illustrates the case where we use the $\mathcal{L}_2$ norm, which corresponds to the Ridge regression problem. The intersection between the counter lines and the green area gives the solutions of our optimization problem.

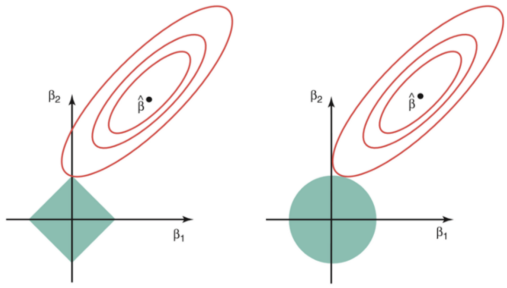


Figure 1. Lasso in 2D

## 2. The model hypothesis and main results

### 2.1. Model Hypothesis

#### 2.1.1 Working conditions

We remind you that the paper put as in the design where $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$, and the errors $\epsilon \in \mathbb{R}^n$ In addition to that, the matrix $\mathcal{X}$ has $i.i.d\, \mathcal{N}(0, \frac{1}{n})$ elements, ensuring the normalization of the columns. As to the errors, $\epsilon_i$ are $i.i.d\, \mathcal{N}(0, \sigma^2)$, where $\sigma$ is arbitrary choisen, yet fixed. The authors of the paper make the assumption that $\sigma$ can be zero, which corresponds to noiseless observation.

#### 2.1.2 Linear sparsity

The article makes the assumption that the coefficients $\beta_1 ... \beta_p$ are independent copies of a random variable $\pi$ such that: $\mathbb{E}\pi < \infty$ and $\forall \epsilon \in [0, 1], \mathbb{P}(\pi \neq 0) = \epsilon$

More limiting conditions were added when dealing with a high dimensional problem by setting:
$\frac{p}{n} \to \delta > 0;\ p, n \to \infty$

In addition to these assumptions, the authors add a condition on the degree of sparsity. In fact, as cited in the paper, the expected number of nonzero regression coefficients is linear in p and equal to $\epsilon.p$ for $\epsilon > 0$. Therefore, this model is opposed to asymptotic discussions. For instance, in large dimension problems the proportion of non-zero coefficients tends towards zero, which is not the case here.

#### 2.1.3 Gaussian Designs

The authors claim that the gaussian designs of the matrix $\mathcal{X}$ with independent columns are assumed to work in favor for model selection thanks to the low correlations between the features which the lasso takes advantage of. Let us now turn to the main result of the article.

### 2.2. Paper Stipulation

This paper shows that, in a regime of linear sparsity and under the regression framework, the lasso does not perform as expected, which consists in, to cite the paper, finding the important variables with few errors. The authors show that its impossible to increase the True Positive Proportion (TPP) [2], which reflects finding the significant variables, without increasing the False Discovery Proportion (FDP) [3] at the same time, which reflects the false discoveries . More formally, there is an asymptotic trade off between the False discovery Proportion and the True Positive Proportion.

The authors claim that this result holds even in the case of the noiseless observation, $\sigma = 0$, which is a weak assumption.

Let's recall the the lasso regression formulation:

$$\hat{\beta} = \underset{b\in\mathbb{R}^p}{\operatorname{argmin}} \left( \frac{1}{n} \|Y - Xb\|_2^2 + \lambda \|b\|_1 \right)$$

Then the FDP and the TPP are defined as follows:

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

$$TPP = \frac{|\left\{ j : \hat{\beta}_j(\lambda) \neq 0, \beta_j \neq 0 \right\}|}{max\left( |\left\{ j : \hat{\beta}_j(\lambda) \neq 0 \right\}|, 1 \right)}$$

The authors claim that this result holds even in the case of the noiseless observation, $\sigma = 0$, which is a weak assumption.

**Let's move on and have a look at the theorm.**

[2] **TPP** is defined as the ratio between the number of true discoveries and that of potential true discoveries to be made.

[3] **FDP** the ratio between the number of false discoveries and the total number of discoveries, along the Lasso path.

## 2.3. Main Results

### 2.3.1 The Lasso Trade-off Diagram

**Theorem 1** *Fix $\delta \in (0, \infty)$ and $\epsilon \in (0, 1)$ and consider the function $q^*(.) = q^*(.; \delta, \epsilon) > 0$; Then under the working hypothesis and for any arbitrary small constants $\lambda_0 > 0$ and $\eta > 0$ then the event:*

$$\cap_{\lambda > \lambda_0} \{FDP(\lambda) \geq q^* (TPP(\lambda)) - \eta\}$$

*holds with a probabilty tending to one in both noisy and noiseless cases.*

The function $q^*$ is called the boundary curve defined by:

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

The fact that $q^*$ is strictly increasing illustrates the asymptotic trade-off between FDP and TPP. In fact, when TPP increases the lower bound of FDP gets higher, which leads to the increase of FDP with a probabilty tending to one.

To better understand the behaviour of the function, you can find a display of some examples of it with different values of sparsity ($\epsilon$) and dimensionality ($\delta$) in figure 2.

An other interesting way to understand this trade-off, is by considering the FDP as a measure of the type I error, and 1 - TPP as a measure of type II error. Therefore, on the lasso path, both types of error rates cannot be simultaneously low.

Figure 3 shows the Lasso trade-off diagram for two different cases. The boundary between the red and white area is assured by the function $q^*$. The red area is where both types of errors are small (ie high TPP and low FDP). This illustrates the paper's theorem as it says that the red region is not reachable on the Lasso path.

Please refer to the joint notebook for the detailed implementations, we tried to reproduce the results in the paper and we can say that, except figure 4 which is quite a mess, the experiments that we put into place were satisfying.

According to the last point of the theorem as cited in the paper, the boundary curve $q^*$ is tight, which means that if we take an other function $q$ such that there exists a point $u$ where $q(u) > q^*(u)$, q will will not satisfy the first two points of the theorem for some prior distribution $\Pi$ on the regressor coefficients.

This comes from the following fact: For any point $(u, q(u))$ on the curve, we can approach it by setting $\epsilon\prime \in (0, 1)$ and setting the prior to be:

$$\Pi = \begin{cases} M, & w.p.\epsilon\epsilon\prime \\ M^{-1}, & w.p.\epsilon(1-\epsilon\prime) \\ 0, & w.p.1-\epsilon \end{cases} \tag{1}$$

And for any $u$ between 0 and 1 there is exists $\epsilon\prime > 0$ such that

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

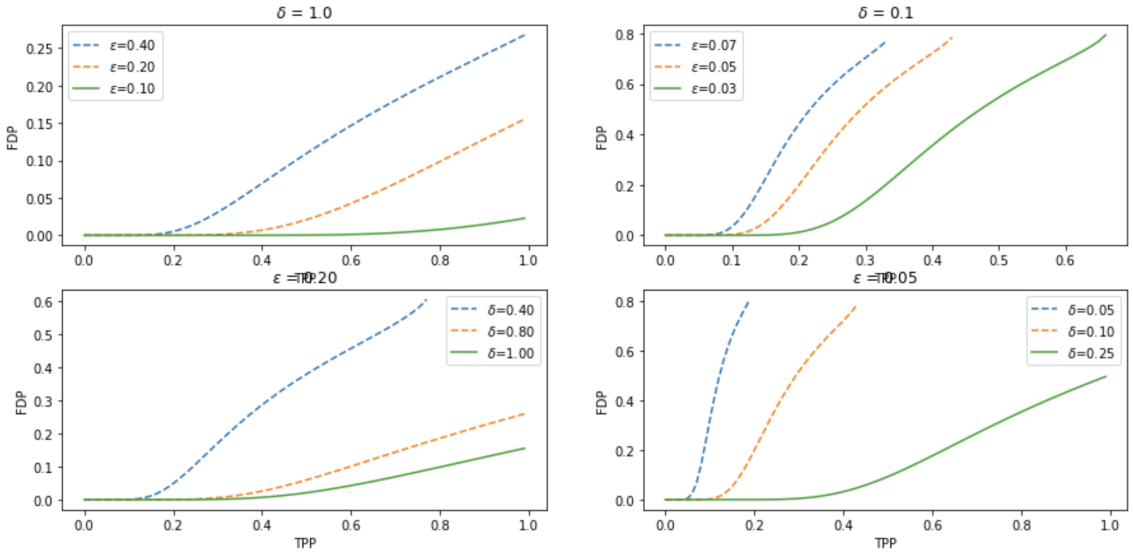


Figure 2. Top-left: $\delta = 1$; top-right: $\epsilon = 0.2$; bottom-left: $\delta = 0.1$; bottom-right: $\epsilon = 0.05$

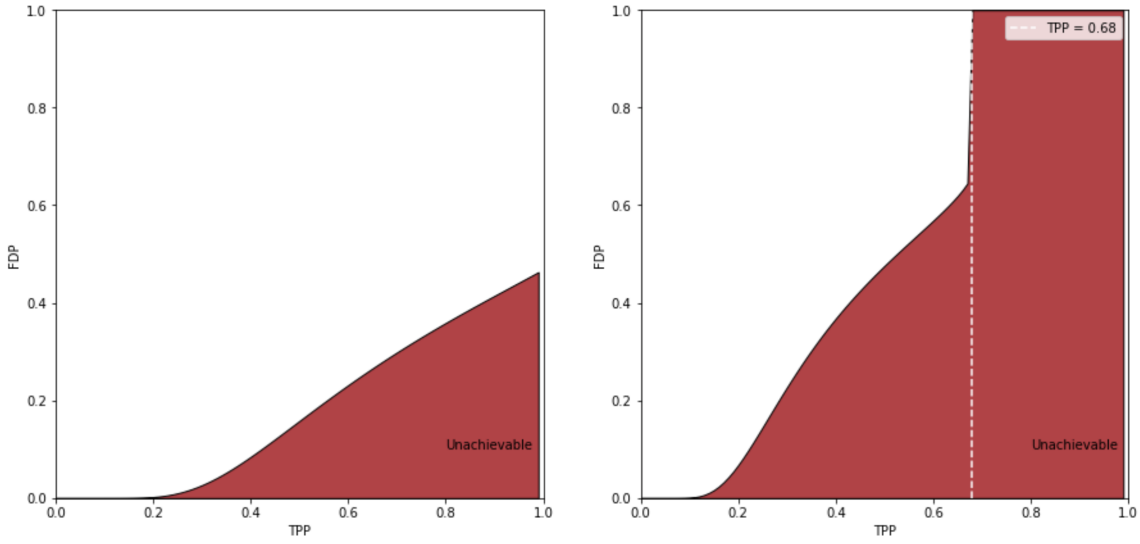


Figure 3. The Lasso trade-off diagram: left: $\delta = 0.5$ and $\epsilon = 0.15$; right:$\delta = 0.3$ and $\epsilon = 0.15$. The vertical truncation occurs at 0.6791).

Figure 4 illustrates the results of the our simulation for finite values of n and p in the noiseless condition ie $\sigma = 0$.

We plot all pairs (TPP, FDP) of 10 Lasso paths for each of the cases n=p=1000 and n=p=5000. Despite the bad quality of the graph, we can confirm the fact that when TPP ¡ 0.8, the large majority of pairs (TPP,FDP) along these 10 paths are above the boundary.

Also, the average FDP gets closer to the boundary as TPP gets closer to one, a fraction of the paths fall below the line as well.

For the technical details, we adopted the same configuration as the paper in order to be able to compare both results. For both graphs $n/p = \delta = 1$, $\epsilon = 0.2$ and noise is zero ($\sigma = 0$); **(a)** FDP vs. TPP along 10 independent Lasso paths with $P(\Pi = 50) = 1 - P(\Pi = 0) = \epsilon$; **(b)** Mean FDP vs. mean TPP averaged at different values of $\lambda$ over 100 replicates for $n = p = 1000, P(\Pi = 0) = 1 - \epsilon$ as before, and $P(\Pi = 50|\Pi \neq 0) = 1 - P(\Pi = 0.1|\Pi \neq 0) = \epsilon\prime$

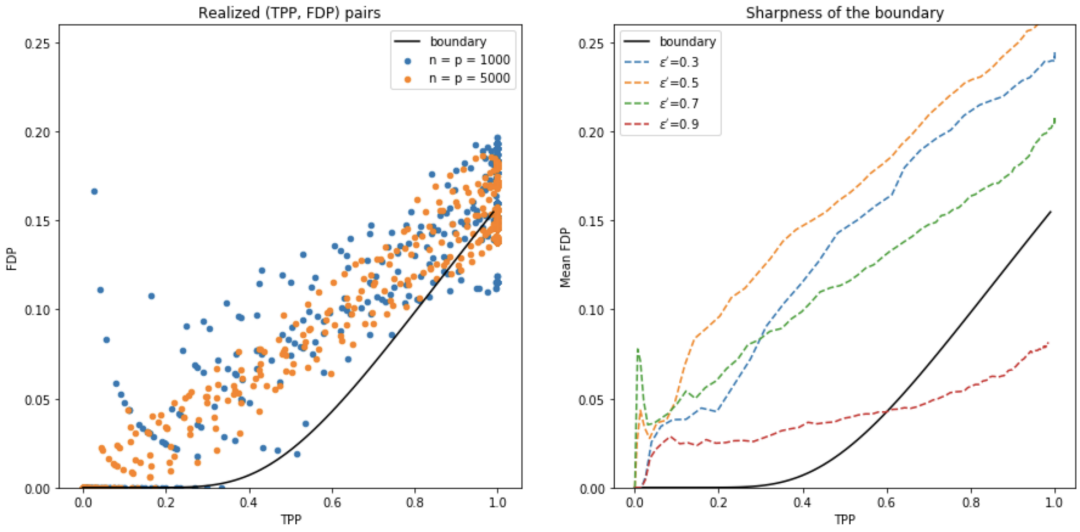


Figure 4. Numerical illustration

### 2.3.2 Lasso Shrinkage Noise

We briefly mentioned earlier, when talking about the $\mathcal{L}_0$ regularization, that there are other methods to have a good model selection performance even under exponentional computations, thus sparing them from the difficulties in identifying variables in the model.

Let's recall the $\mathcal{L}_0$-penalized maximum likelyhood estimate:

$$\hat{\beta} = \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \left( \frac{1}{n} \|Y - Xb\|_2^2 + \lambda \|b\|_0 \right)$$

Let's cite the second theorem from the paper:

**Theorem 2** *Under our working hypothesis, take $\epsilon < \delta$ for identifiability, and consider the two- point prior:*

$$\Pi = \begin{cases} M, & w.p.\ \epsilon \\ 0, & w.p.\ 1 - \epsilon \end{cases} \tag{2}$$

*Then we can find $\lambda(M)$ such that in probability, the discoveries of the $\mathcal{L}_0$ estimator obey:*

$$\lim_{M\to+\infty}\lim_{n,p\to+\infty} TPP(\lambda) = 1 \lim_{M\to+\infty}\lim_{n,p\to+\infty} FDP(\lambda) = 0$$

The theorem 2 highlights the absence of the asymptotic trade-off between TPP and FDP in the Ridge regression problem making it a better solution putting aside the computational limitations of the problem as it's non convex.

Lasso limitations is mainly caused by the pseudo-noise introduced by the shrinkage. In fact, when the regularization factor is large, the lasso estimates get dwarfed. In other terms, if strong variables are selected, the penalty noise is inflated, and projecting it along the directions of some zero variables may actually eclipse the signal associated with the strong regression coefficients. It is for this reason that false discoveries enter the Lasso path.

Let's consider a reduced Lasso problem under the assumption that the true support $\mathcal{T}$ is a deterministic subset of size $\epsilon p$ in which each non-zero coefficient takes a value $M > 0$. We also assume $\delta > \epsilon$. To make things more challenging, we place ourselves in the noiseless case where $\sigma = 0$.

$$\hat{\beta}_{\mathcal{T}}(\lambda) = \underset{b_{\mathcal{T}}\in\mathbb{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2}\|y - X_{\mathcal{T}} b_{\mathcal{T}}\|^2 + \lambda\|b_{\mathcal{T}}\|_1$$

The article then suggests to take $\lambda$ to have the same magnitude as $M$. The solution of the reduced lasso problem $\hat{\beta}_{\mathcal{T}}$ must verify the KKT conditions given by:

$$-\lambda\mathbf{1} < \mathcal{X}_{\mathcal{T}}^T(y - \mathcal{X}_{\mathcal{T}}\hat{\beta}_{\mathcal{T}}) \leq \lambda\mathbf{1}$$

- When $|\mathcal{X}_j^T(y - \mathcal{X}_{\mathcal{T}}\hat{\beta}_{\mathcal{T}})| \leq \lambda, \forall j \in \bar{\mathcal{T}}$, then completing $\hat{\beta}_{\mathcal{T}}(\lambda)$ whith zeros with give the solution to the full Lasso problem.

- If for $j \in \bar{\mathcal{T}}$, $|\mathcal{X}_j^T(y - \mathcal{X}_{\mathcal{T}}\hat{\beta}_{\mathcal{T}})| \leq \lambda$ then $X_j$ is selected by the incremental Lasso with design variables indexed by $\mathcal{T} \cup \{j\}$.

Formally, this means that if we take $j \in \bar{\mathcal{T}}$ and define:

$$\hat{\beta}_{\mathcal{T}\cup\{j\}}(\lambda) = \underset{b\in\mathbb{R}^{\epsilon p+1}}{\operatorname{argmin}} \frac{1}{2}\|y - X_{\mathcal{T}\cup\{j\}} b\|^2 + \lambda\|b\|_1$$

Then we have: $\hat{\beta}_j = 0$

All in All, two main points are worth mentionning:

First, if the full Lasso selects very few variables in $\bar{\mathcal{T}}$, then the value of the prediction $X\hat{\beta}$ will be very close to the one obtained in the reduced model.

However, if the lasso selects only a small variable proportion from $\mathcal{T}$, hence a large proportion from $\bar{\mathcal{T}}$, then there would be a significant number of false discoveries.

## 3. Conclusion

Under suitable conditions and assumptions, among which, weakly correlated variables and linear regime sparsity, the lasso did not always perform the task of selecting important vavriable with minimal error rate.

We showed the relationship between the False Discovery Proportion (FDP) and the False Negative Rate (1 - TPP), we highlighted the trade-off which consists in the fact that we can't increase TPP without increasing FDP. The figure 3 points out the unreachable regions where one would ideally have a high TPP and a low false discoveries ie FDP.

We focussed our analysis on lasso and talked about ridge regression and the subset selection estimator with the $\mathcal{L}_0$ regularization.

We Would like to mention that there are other variants of Lasso, including Elastic net which combines lasso and ridge regularization, fused lasso which penalizes differences instead of the individual coefficients, adaptive lasso which benefits from the variable selection of Lasso yet re-estimates parameters of selected variables by means of OLS, ... They all aim at overcoming the different limitations including shrinkage noise and correlation concerns.

Analysing this article was of a great award, it allowed us to better understand and cease the lasso in comparasion to the other optimization problems.