False Discoveries occur Early on the Lasso Path

Weiji Su, Malgorzat Bogdan, Emmnuel J. Candès

Chloe Baraille, Othmane Sebbouh

February 15, 2018

Abstract

In this report, we will analyze the article "False Discoveries occur Early on the Lasso Path", which will be referred to as "the article" in the body of the report. In the framework of the Lasso regression, the article establishes that, in a regime of linear sparsity, there's an asymptotic trade-off between the false discovery proportion (FDP) and the true positive proportion (TPP). Heuristically, this means that we can't find all the significant variables (thereby increasing the TPP) without including many weak ones (thereby increasing FDP). The results of the article are demonstrated under weak assumptions, namely independent Gaussian designs and a possibly null noise, and hence will likely extend to easier problems.

We will first discuss the support recovery properties of the Lasso in an asymptotic framework, and we will show that under some assumptions, the Lasso estimator recovers the true support with probability tending to one. This result does not contradict the FDP-TPP trade-off exhibited by the article in a regime of linear sparsity. We will first pause on some definitions and hypotheses before presenting the main theorem and heuristics. Lastly, we will study the performance of the Bolasso, which essentially relies on bootstrap, in terms of support recovery to tackle the FDP-TPP trade-off.

Contents

1 Introduction

1.1 Why not the l_0 norm ?

Consider the standard linear model:

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

where X is an $n * p$ matrix of features (supposed to be deterministic in this initial step), $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1...n$, and observations are *i.i.d.*. This problem rewrites:

$$Y = X\beta + \epsilon \tag{1}$$

where $Y \in \mathbf{R}^n$, $\epsilon \in \mathbf{R}^n$, $\beta \in \mathbf{R}^p$. We suppose that β is "structured" i.e. $\beta \in \beta_0[k] = \{\beta : \|\beta\|_0 \leq k\}$. We recall that $\|\beta\|_0 = \sum_{j=1}^{p} \{\beta_j \neq 0\}$. A natural estimator of β could be the subset selection estimator, defined by:

$$\hat{\beta}_0 = argmin_{\beta \in \beta_0[k]} \|Y - X\beta\|_2^2$$

The subset selection estimator directly penalizes the presence of non zero components in β and so, will induce sparsity in the subset selection solution $\hat{\beta}$. We have some theoretical results on this estimator. If we consider the estimation risk $R_X[\hat{\beta}, \beta] = \mathbf{E}(\|X\hat{\beta} - X\beta\|_2^2)$, then we have, for $k \leq c\frac{n}{log(p)}$:

$$min_{\hat{\beta}} max_{\beta \in \beta_0[k]} R_X[\hat{\beta}_0, \beta] \geq c_X \sigma^2 k log\left(\frac{p}{k}\right)$$

Furthermore, for $\hat{\beta}_0 = argmin_{\beta \in \beta_0[k]} \|Y - X\beta\|_2^2$,

$$Sup_X Sup_{\beta \in \beta_0[k]} R_X[\hat{\beta}_0, \beta] \leq c\sigma^2 k log\left(\frac{p}{k}\right)$$

We can expect the subset selection estimator to recover the true support of β, which is the issue we are dealing with here. The problem is that $\beta_0[k]$ is not a convex set. Solving $argmin_{\beta \in \beta_0[k]} \|Y - X\beta\|_2^2$ hence implies to look at all the possible supports of β i.e. to explore $\{j \in \{1...p\} : \beta_j \neq 0\}$, whose cardinal is $\binom{k}{p} \approx \left(\frac{p}{k}\right)^k$. This problem is known to be NP-hard, and hence impossible to solve computationally. Furthermore, the number k of non zero coefficients is generally unknown.

1.2 The Lasso

One way to deal with this issue is to convexify the l_0 norm:

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - X\beta\|^2 + \lambda\|\beta\|_1 \tag{2}$$

which is exactly the definition of the Lasso estimator. We can also exhibit some convergence rates for the Lasso. More precisely, if we assume $\|X_j\|_2^2 = n$ for $j = 1...n$, for $\lambda = \mathcal{O}(\sigma\sqrt{nlog(p)})$, the Lasso fit satisfies:

$$\|X\beta - X\hat{\beta}\|_2^2 = \mathcal{O}_{\mathbf{P}}\left(\sigma\sqrt{\frac{log(p)}{n}.\|\beta_0\|_1}\right)$$

1.2.1 Support recovery for deterministic designs

The Lasso estimator is well-kown for its shrinkage and variable selection properties. As the article we will discuss later deals with the latter, we focus on the Lasso variable selection and support recovery properties. The results we describe are due to Mainwright (2009), who introduced a proof technique called the *primal-dual witness method*.

Again we assume the standard linear model defined in (1) for fixed X and β, with the scaling $\|X_j\| = n$ for $j = 1...n$. Let $S = Supp(\beta)$. We want to show that for some λ, $Supp(\hat{\beta}) = S$ with high probability. We also want that $sign(\hat{\beta_S}) = sign(\beta_S)$ with high probability. The *primal-dual witness method* plugs in the support S into the KKT conditions. The subgradient optimality for the Lasso problem defined in (2) writes:

$$X^T(Y - X\hat{\beta}) = \lambda s$$

where $s \in \partial\|\hat{\beta}\|_1$, subgradient of the l_1 norm evaluated at $\hat{\beta}$. Suppose that $Supp(\hat{\beta}) = S$ at $\hat{\beta}$. Then the KKT conditions become:

$$\begin{aligned} X_S^T(Y - X_S\hat{\beta}_S) &= \lambda s_S \\ X_{-S}^T(Y - X_S\hat{\beta}_S) &= \lambda s_{-S} \end{aligned} \tag{3}$$

Then, if we can satisfy the two conditions above with a proper subgradient s such that $s_S = sign(\hat{\beta}_S)$ and $\|s_{-S}\|_\infty = max_{i \notin S}|s_i| < 1$, we have met our goal (we recall that β_j is selected iff $s_j < -1$ or $s_j > 1$, see Figure 2 in 3.3). We first solve $\hat{\beta}_S$ in the first equation of (3):

$$X_S^T(Y - X_S\hat{\beta}_S) = \lambda s_S \iff \hat{\beta}_S = (X_S^T X_S)^{-1}[X_S^T Y - \lambda sign(\beta_S)] \tag{4}$$

Solving the second equation of (3) in s_{-S} implies that:

$$X_{-S}^T(Y - X_S\hat{\beta}_S) = \lambda X_{-S} \iff \frac{1}{\lambda}X_{-S}^T[(I - X_S(X_S^T X_S)^{-1}X_S^T)Y] + X_{-S}^T X_S sign(\beta_S) \tag{5}$$

We can derive the *primal-dual witness method* from the above. It proceeds as follows:

- Solve for the Lasso solution over the S components, get $\hat{\beta}_S$ as in (4) and set $\hat{\beta}_{-S} = 0$
- Solve for the subgradient over the S^c components, get s_{-S} as in (5)
- Check that $sign(\hat{\beta}_S) = sign(\beta_S)$ and that $\|s_{-S}\|_\infty < 1$. If these two conditions are verified, we have the existence of a unique Lasso solution that exactly recovers the true support and signs.

The conditions (4) and (5) can be re-expressed using that $Y = X\beta + \epsilon$:

$$\begin{aligned} sign(\hat{\beta}_S) &= sign[\beta_S + (X_S^T X_S)^{-1}X_S^T\epsilon_S - \lambda sign(\beta_S)] \\ \|\frac{1}{\lambda}X_{-S}^T(I - X_S(X_S^T X_S)^{-1}X_S^T)\epsilon + X_{-S}^T(X_S^T X_S)^{-1}sign(\beta)\|_\infty &< 1 \end{aligned} \tag{6}$$

The two required conditions reduce to statements about Gaussian processes. To make things easier, we need to make assumptions on X and β:

- Mutual incoherence : for some $\gamma > 0$, $\|(X_S^T X_S)^{-1}X_S^T X_j\|_1 \leq 1 - \gamma$ for $j \notin S$
- Minimum eigenvalue : for some $C > 0$, $\Lambda_{min}(\frac{1}{n}X_S^T X_S) \geq C$, where $\Lambda_{min}(A)$ is the minimum eigenvalue of A.
- Minimum signal: $|\beta_{0,i}| > \lambda\left(\|(X_S^T X_S)^{-1}\|_\infty + \frac{4\sigma}{c}\right)$ for $i \in S$

Under these assumptions on X and β, and with $\lambda \geq 2\sigma\frac{\sqrt{2nlog(p)}}{\gamma}$, (6) holds with high probability, i.e., the Lasso solution is unique and recovers rhe exact support and signs with high probability.

We can make some remarks on the assumptions above. Mutual incoherence says that a variable in S^C cannot be too correlated with a variable in S. Minimum eigenvalue condition says that variables in S cannot be too correlated with each other. Minimal signal condition ensures that the non zero entries of β are big enough to detect.

1.2.2 Support recovery for Gaussian designs

A common theoretical pursuit is to show that these hold with high probability under a random choice of X, where the entries of X are Gaussian. Thes conditions are not verified in the regime of linear sparsity. More precisely, under the standard Gaussian design, if $n = \nu p$ for some $\nu \in [0, 1]$, the Lasso can only recover vectors β with support $k \leq (1 + o(1))\frac{\nu p}{2log(p)}$ (where k denotes the number of non zero coefficients of β). It fails with probability converging to one for any vector

$\beta \in \mathbf{R}^p$ with $k = \mathcal{O}(p)$.

This is the case the article we discuss below deals with. The article shows that in the most adverse situations, $\hat{\beta}$ cannot recover the full suppport of β. Nevertheless, even in the most favourable cases, the support can only be recovered at the price of making "false discoveries", i.e., selecting variables that do not belong to the true support of β, and we will never have $Supp(\hat{\beta}) = Supp(\beta)$.

2 Ratios' definitions and the model's hypothesis

The article's claim is that, in a regime of linear sparsity, there is an asymptotic trade-off between FDP and TPP. We consider the general formulation of the Lasso regression as in (2):

$$\hat{\beta}(\lambda) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - Xb\|^2 + \lambda\|b\|_1$$

and we suppose that the regression coefficients are: $\beta = [\beta_1, ..., \beta_p]^T$. Then the False Discoveries Proportion (FDP) and True Positives Proportions (TPP) are defined as follows:

$$FDP(\lambda) = \frac{|j : \hat{\beta}_j(\lambda) \neq 0,\ \beta j = 0|}{|j : \hat{\beta}_j(\lambda) \neq 0|}$$

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

Note that, from a statistical hypothesis testing point of view, FDP can be seen as the proportion of type I error errors: noting $H_0 : \beta_j = 0$, FDP counts the number of times the hypothesis H_0 is rejected while it shouldn't have been, on the number of times it was rejected. Similarly, TPP can be seen as the proportion of (1 - Type II error).

We now develop the hypotheses of the model described in the article. The main hypothesis that determines the framework of the article is the asymptotic linear sparsity.

2.1 Linear sparsity

First, the article imposes restricting conditions on the asymptotic dimensions of the problem, by setting $\frac{n}{p} \to \delta$, where $\delta > 0$. The article also makes the assumption that $\beta_1, ..., \beta_p$ are copies of a variable $\prod$ which verifies, for an $\epsilon \in]0, 1[$, $\mathbf{P}(\prod \neq 0) = \epsilon$, where we can think of ϵ as a small constant. Then, the expected number of non-zero coefficients is $\mathbf{E}(\sum_{i=1}^p \mathbf{1}_{\beta_i \neq 0}) = \sum_{i=1}^p \mathbf{P}(\beta_i \neq 0) = \epsilon.p$

The reason behind this hypothesis is the fact that it covers very practical settings, often encountered in practice, where the dimensions are moderately large, but where we still can consider very sparse signals. This hypothesis will later be identified as the source of the bad performance of the Lasso.

2.2 Gaussian Designs

The article considers Gaussian designs, i.e. the columns of the design $X \in \mathbf{R}^{n \times p}$ are drawn from i.i.d. $\mathcal{N}(0, \frac{1}{n})$. This design is known to be favorable for model selection. Indeed, low correlation between the features makes it easier for the Lasso to select the variables.

2.3 The noise

The article makes the hypothesis that the noise variables z_i are i.i.d. $\mathcal{N}(0, \sigma)$, where σ can be null. This is a very weak assumption, as the article will show that even in the noiseless case, the trade-off between decreasing FDP and increasing TPP that is demonstrated will hold. This in turn shows that there must be some other source of noise which is inherent to the Lasso.

3 The main theorem and heuristics

After presenting the hypotheses, we now turn to the theorem itself which states the existing a trade-off between increasing the TPP and decreasing the FDP. We only give the (first) most important result for the clarity of this report.

3.1 The trade-off theorem

Fix $\delta \in]0, \infty[$ and $\epsilon \in]0, \infty[$ and consider the function $q^*(.; \delta; \epsilon) > 0$ defined in the article. Then under the hypothesis of Section II and for any arbitrary small constants $\lambda_0 > 0$ and $\delta > 0$, the following conclusion holds:

The event $\bigcap_{\lambda \geq \lambda_0} \{FDP(\lambda) \geq q^*(TPP(\lambda)) - \eta\}$ holds with probability tending to one. This is true even if the noise is null i.e. even if $\sigma = 0$.

Note that the function q^* defined in the article is strictly increasing. Hence increasing TPP(λ) means increasing FDP(λ). This is true for all the values of λ, even if it is chosen adaptively, i.e. for example, even if it is chosen after cross-validation.

Rather than demonstrating this result, for which the proof can be found in the article, we chose to move directly to the heuristic explanations and further expand some points that didn't appear so trivial to us. The explanation below aims at clarifying why, in the best conditions (i.e. absence of noise, Gaussian designs), the trade-off still holds. It highlights the fact that there is some noise induced by the Lasso's characteristics.

3.2 Analysis of the heuristic explanation of the theorem

Not all methods share the same difficulties in identifying the variables of the model. For example, the l_0-penalized maximum likelihood estimate given by:

$$\hat{\beta}_0(\lambda) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - Xb\|^2 + \lambda \|b\|_0$$

can lead to perfect support recovery, i.e. a 100% TPP and 0% FDP. This observation strengthens the article's belief that the Lasso's own characteristics explain the existence of the trade-off between FDP and TPP. The article explains that what they call "shrinkage noise" causes the Lasso path to be full of false discoveries when increasing the number of true positives. To understand the heuristic concept, we recall the formulation of the Lasso regression as in (1):

$$\hat{\beta}(\lambda) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - Xb\|^2 + \lambda \|b\|_1$$

The difference between the Lasso regression and the l_0 regression is the shrinkage induced by the l_1-norm. In fact, where the l_0-norm operates hard thresholding, by selecting only the variables for which the coefficient β_j is above a certain level, the l_1 norm shrinks some coefficients in addition to putting some others to 0. But the shrinkage creates some pseudo-noise for the model selection.

First, we explain why Lasso selects variables and shrinks some coefficients. To do so, we consider an equivalent formulation of the Lasso regression , which more prone to illustration:

$$\hat{\beta}(t) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\|y - Xb\|^2 \quad \text{s.t. } \|b\|_1 \leq t \tag{7}$$

The problem's constraints and solution can be illustrated in Figure 1 with $p = 2$. It clearly appears that the Lasso will give a the couple $(\hat{\beta}_1, \hat{\beta}_2)$ where $\hat{\beta}_1 = 0$ and $\hat{\beta}_2 < \beta_2$, where (β_1, β_2) is the solution the OLS problem, i.e. the Lasso selects only the second variable, but shrinks it.

To connect the two formulations of the Lasso (1) and (7), we can observe that increasing λ is equivalent to decreasing t, as it means more regularization. Hence, if λ is very large, t will be very low, which means the coefficients will be even more shrunk by the Lasso. If there are many variables with a large effect, the noise induced by the shrinkage property of the Lasso will be even

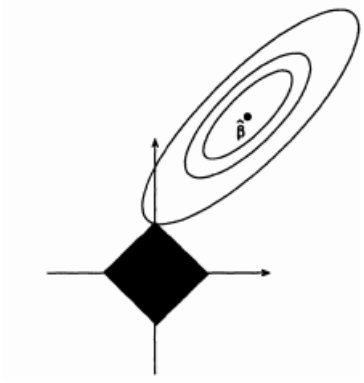

Figure 1: Illustration of the Lasso problem in the case p=2

bigger. Indeed, the Lasso makes the important variables weaker, which makes strong variables and weak variables close. Hence, some weak variables are not eliminated by the Lasso.

What is remarkable here is the fact that for this phenomenon to occur, we must have an important number of strong variables. This is where the hypothesis of linear sparsity is decisive. In the following, we outlay the formal explanation of the article's heuristics and clarify some of the points that required some prior knowledge of the Lasso regression.

3.3 Formal explanation of the heuristic

For simplicity, the article considers $\mathcal{T}$ a deterministic support of size $\epsilon.p$, and for every $j \in \mathcal{T}, \beta_j = M > 0$.

It also assumes that:

- $\delta > \epsilon$ i.e. $\frac{n}{p} > \epsilon$ i.e. $n > \epsilon.p$, which means that the design X has more rows than columns.
- the noise is null, i.e. $\sigma = 0$

Under these assumptions, the article considers the reduced Lasso problem:

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

Then, the article states: "Now take λ to be of the same magnitude as M so that roughly half of the signal variables are selected". To understand this statement, we rewrite the Lasso problem, noting β the true coefficients vector:

$$\begin{aligned}
\hat{\beta}_{\mathcal{T}}(\lambda) &= \underset{b \in \mathbf{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2}\|y - X_{\mathcal{T}} b\|^2 + \lambda \|b\|_1 \\
&= \underset{b \in \mathbf{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2}\|X_{\mathcal{T}} b_0 - X_{\mathcal{T}} b\|^2 + \lambda \|b\|_1 \\
&= \underset{b \in \mathbf{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2}\|X_{\mathcal{T}} (b_0 - b)\|^2 + \lambda \|b\|_1 \\
&= \underset{b \in \mathbf{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2}(b_0 - b)^T X_{\mathcal{T}}^T X_{\mathcal{T}} (b_0 - b) + \lambda \|b\|_1
\end{aligned}$$

Since the columns of X are approximately normalized and that they are drawn independently, $X^T X = I_{\epsilon.p}$. Hence:

$$\hat{\beta}_{\mathcal{T}}(\lambda) = \underset{b \in \mathbf{R}^{\epsilon p}}{\operatorname{argmin}} \frac{1}{2}\|b_0 - b\|_2 + \lambda \|b\|_1$$

which is exactly the expression of the proximal operator taken at b_0.

As we know that for $x \in \mathbf{R}^{\epsilon p}, prox_{\lambda|.|}(x) = (prox_{\lambda|.|}(x_1), ...prox_{\lambda|.|}(x_{\epsilon p}))$, for every $j \in \mathcal{T}$:

$$\beta_j = prox_{\lambda|.|}(b_{0j})$$

where:

$$prox_{\lambda|.|}(v) = \begin{cases} v-\lambda, & v > \lambda \\ 0 \quad ,v \in [-\lambda, \lambda] \\ v+\lambda, & v < \lambda \end{cases}$$

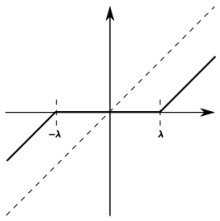

Figure 2: The proximal operator for $\lambda|.|$

As we assumed that $b_{0j} = M$ for all non zero-coefficients and since $M > 0$, β_j will be equal to $M - \lambda$ when $M > \lambda$.

But since we assumed that M and λ are of the same magnitude, i.e. $\lambda \approx M$, we can say that approximately half the time $\lambda > M$, and the signal variables get selected. In other terms, half the signal variables get selected.

Back to the reduced Lasso problem: if $\hat{\beta_{\mathcal{T}}}$ is a solution, it must verify the KKT conditions, including the stationarity condition, which translates, after bounding each of the coordinates of the subgradient of $\|.\|_1$ in $b_{\mathcal{T}}$ by 1 and -1:

$$-\lambda \mathbf{1} \leq X_{\mathcal{T}}^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) \leq \lambda \mathbf{1}$$

The article then suggests: "A first simple fact is this: for $j \in \bar{\mathcal{T}}$, if $X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) > \lambda$, then X_j must be selected by the incremental Lasso with design variables indexed by $\mathcal{T} \cup \{j\}$."

Proof. We want to prove that, for $j \in \bar{\mathcal{T}}$, if we note $\mathcal{T}' = \mathcal{T} \cup \{j\}$ and :

$$\hat{\beta}_{\mathcal{T}'}(\lambda) = \underset{b \in \mathbf{R}^{\epsilon p+1}}{\operatorname{argmin}} \frac{1}{2}\|y - X_{\mathcal{T}'}b\|^2 + \lambda\|b\|_1$$

we will have $\hat{\beta}_j \neq 0$.

First note that since $\hat{\beta_{\mathcal{T}}}$ is a solution to the reduced Lasso to $\mathcal{T}$, we can say that the solution to the incremental Lasso will be of the form: $\hat{\beta_{\mathcal{T}'}} = \left(\hat{\beta_{\mathcal{T}}}|\hat{\beta}_j\right)^T \in \mathbf{R}^{\epsilon p+1}$. Also, we can write: $X_{\mathcal{T}'} = (X_{\mathcal{T}}|X_j) \in \mathbf{R}^{n,\epsilon p+1}$. Then:

$$\begin{aligned} X_{\mathcal{T}'}^T(y - X_{\mathcal{T}'}\hat{\beta_{\mathcal{T}'}}) &= X_{\mathcal{T}'}^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}} - X_j\hat{\beta}_j) \\ &= X_{\mathcal{T}'}^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) - \hat{\beta}_j X_{\mathcal{T}'}X_j \end{aligned}$$

the last coordinate of the vector $X_{\mathcal{T}'}^T(y - X_{\mathcal{T}}) - \hat{\beta}_j X_{\mathcal{T}'}X_j$ is $X_j^T(y - X_{\mathcal{T}}) - \hat{\beta}_j X_j^T X_j$. Since $\hat{\beta_{\mathcal{T}'}}$ must verify the KKT conditions, we will have:

$$-\lambda \leq X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) - \hat{\beta}_j X_j^T X_j \leq \lambda$$

which means that $\hat{\beta}_j X_j^T X_j \geq X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) - \lambda$. But since we assumed that $X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) > \lambda$, we have, noticing that, $X_j^T X_j = 1 : \hat{\beta}_j > 0$ □

A decisive argument in the formal explanation of the impact of the so-called shrinkage noise on the performance of the Lasso is that any $j \in \bar{\mathcal{T}}$ that verifies $|X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}})|$ will be selected by the full Lasso (i.e. the Lasso on $\mathbf{R}^p$). The article justifies this heuristic by stating that there are two alternatives:

- if the full Lasso selects only a few variables in $\bar{\mathcal{T}}$, then the residuals $y - X\hat{\beta}$ are close from the reduced Lasso's residuals: $y - X_{\mathcal{T}}$. Hence $|X_j^T(y - X\hat{\beta})| \approx |X_j^T(y - X\hat{\beta_{\mathcal{T}}})| > \lambda$, which means that X_j is highly correlated with $y - X\hat{\beta}$. But since $y - X\hat{\beta}$ is close to λ, X_j is close to λ, which means it is likely to be selected by the full Lasso.

- in the second case, if the full Lasso selects many variables in $\bar{\mathcal{T}}$, the heuristic isn't necessary, because the fact that the full Lasso already selects a considerable number of variables in $\bar{\mathcal{T}}$ confirms the trade-off FDP/TPP.

After stating this, the goal is to show that there is a high probability for any $j \in \bar{\mathcal{T}}$ to verify $|X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}})|$. This would imply that there is a high probability for a wrong variable to be selected. And since we have chosen $\lambda \approx M$ such that roughly half the variables are selected in the true support, this will mean that once half the variables are selected by the Lasso, we start having a non-vanishing number of wrong variables selected.

Noting $g_{\mathcal{T}}$ a subgradient of $\|.\|_1$ at $\hat{\beta_{\mathcal{T}}}$ and that $y = X_{\mathcal{T}}\beta_{\mathcal{T}}$ (because $\sigma = 0$) :

$$X_{\mathcal{T}}^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) = X_{\mathcal{T}}^T(X_{\mathcal{T}}\beta_{\mathcal{T}} - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}}) = \lambda g_{\mathcal{T}}$$

Since the columns of $X_{\mathcal{T}}^T X_{\mathcal{T}}$ are orthonormal, this matrix is invertible. Hence: $X_{\mathcal{T}}(\beta_{\mathcal{T}} - \hat{\beta_{\mathcal{T}}}) = \lambda X_{\mathcal{T}}(X_{\mathcal{T}}^T X_{\mathcal{T}})^{-1} g_{\mathcal{T}}$

As explained before, $\delta > \epsilon$ implies that $X_{\mathcal{T}}$ has more rows than columns, which implies that $X_{\mathcal{T}}(X_{\mathcal{T}}^T X_{\mathcal{T}})^{-1}$ has a smallest singular value > 0 because it is a fixed random matrix. Since for any matrix $A \in \mathcal{M}_{np}$, we have: $\|Ax\| \geq \mu_{min}\|x\|$, where μ_{min} is the smallest singular value of A: $\|X_{\mathcal{T}}(X_{\mathcal{T}}^T X_{\mathcal{T}})^{-1} g_{\mathcal{T}}\| \geq \lambda.\mu_{min}.\|g_{\mathcal{T}}\|$

The subgradient takes values either in [-1, 1] or 1. Each time we make a discovery, the subgradient takes the value 1. Since we make about half discoveries and the size of the support is $\epsilon.p$, the subgradient will take the value 1 $\frac{\epsilon.p}{2}$ times. Hence, noting c_1 a constant depending on ϵ and δ, we have with high probability that:

$$\|X_{\mathcal{T}}(\beta_{\mathcal{T}} - \hat{\beta_{\mathcal{T}}})\| \geq \lambda.c_1.p$$

Since $\hat{\beta_{\mathcal{T}}}$ and $X_{\bar{\mathcal{T}}}$ are independent, and the coordinates of X_j^T are independently drawn of a $\mathcal{N}(0, \frac{1}{n})$, $X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}})$ is a $\mathcal{N}(0, \frac{\|X_{\mathcal{T}}(\beta_{\mathcal{T}} - \hat{\beta_{\mathcal{T}}})\|^2}{n})$. Then the lower bound of the variance is given by: $\frac{c_1\lambda^2 p}{n}$, which is strictly positive. Since the last result holds for any $j \in \bar{\mathcal{T}}$ and $|\bar{\mathcal{T}}| = (1 - \epsilon).p$, it holds for an expected number of variables linear in p, which achieves the proof. This reinforces the necessity of the linear sparsity assumption: if it was not the case, i.e. if the support (assumed to be of size $\epsilon.p$) was much smaller, the variance of $X_j^T(y - X_{\mathcal{T}}\hat{\beta_{\mathcal{T}}})$ would also have been small. Then, the expected FDP would have been very low.

3.4 Experiments

Now that we have established the assumptions of the theorem and stated the theorem, we now verify it with a numerical implementation. We replicate the results of the article for fixed values of $n, p, \delta and \epsilon$. The implementations and the results can be found in the joined python notebook.

Figure 3: The boundary curve (green) Illustration of TPP FDP trade-off for n=p=1000 (blue) and n=p=5000 (yellow), and the boundary curve

4 Extension: the Bolasso

After studying the Lasso's characteristics developed in the article, we wanted to see if there were any variants of the Lasso that allow to go around the FDP/TPP trade-off developed in the article. Indeed, many variants of the Lasso exists, and they tackle different flaws of the Lasso, among them:

- the Gauss-Lasso estimator: used to find a solution to the bias induced by the shrinkage,
- the Group-Lasso: used in the specific case of grouped sparse coordinates of β.

We now study empirically the impact of another variant of the Lasso, the "Bolasso", introduced by Bach (2008). We first develop the model assumptions of the Bolasso and compare them to those of our article (by our article, we mean "False Discoveries Occur Early on the Lasso Path"). Second, we report Bach (2008)'s proposition that gives an upper bound to the probability of wrong support recovery. Finally, we numerically test the Bolasso's algorithm and verify if its result outperform the ordinary Lasso in terms of the trade-off developed in our article.

4.1 The assumptions of the Bolasso

As we are not studying the Lasso asymptotically in our article, we will restrain our analysis to of Bach (2008) to the non-asymptotic results. First, we recall the assumptions of the Bolasso. Keeping the notations of our article, Bach (2008) makes the following assumptions:

- **(A1)** : for a column X_j of X and an element y_i of y, the cumulant generating functions $\mathbf{E}(e^{s\|X_j\|_2^2})$ and $\mathbf{E}(e^{s.y_j^2})$ are finite for some s > 0.
- **(A2)** : The joint matrix of second order moments $\mathbf{E}(X^TX)$ is invertible
- **(A3)** : $\mathbf{E}(Y|X) = X^T\beta$ and $\text{var}(Y|X) = \sigma^2$

We now prove that our model obeys the previous assumptions.

Proof.

- **(A1)** : With the assumptions of our article, we have that for any i, j : $X_{ij} \sim \mathcal{N}(0, \frac{1}{n}) \implies \sqrt{n}X_{ij} \sim \mathcal{N}(0,1) \implies nX_{ij}^2 \sim \chi_2^1 \implies n\|X_j\|_2^2 = \sum_{i=1}^n X_{ij}^2 \sim \chi_2^n$

  For $Z \sim \chi_2^n$, we have: $\mathbf{E}(e^{sZ}) = (1-2s)^{-\frac{n}{2}}$ if $s < \frac{1}{2}$. Hence, with $s = \frac{1}{2}$ (then $\frac{s}{n} < \frac{1}{2}$ (assuming $n \geq 2$) and $Z = s\|X_j\|_2^2 = \frac{s}{n}(n\sum_{i=1}^n X_{ij}^2)$, we have:

$$\begin{aligned}\mathbf{E}(e^{s\|X_j\|_2^2}) &= (1-\frac{2s}{n})^{-\frac{n}{2}} = (1-\frac{1}{n})^{-\frac{n}{2}} \\ &= \exp(-\frac{n}{2}\log(1-\frac{1}{n})) = \exp(\frac{1}{2}\frac{\log(1-\frac{1}{n})}{-\frac{1}{n}})) \quad \to \exp(0\times 1) = 1\end{aligned}$$

  Hence, we have indeed that $\mathbf{E}(e^{s\|X_j\|_2^2})$ is finite for $s = \frac{1}{2}$. Moreover, we make the reasonable assumption that the second part of **(A1)** holds in our framework.

- **(A2)** : This is a direct result of the construction of the design matrix in our article
- **(A3)** : The model of our article is :

$$Y = X\beta + z$$

  where $z \sim \mathcal{N}(0, \sigma^2)$. **(A3)** is then verified because $\mathbf{E}(X\beta|X) = X\beta$, $var(X\beta|X) = 0$ and z and X are independent.

□

One precision is in order to make sure that the assumptions of our model and those of Bach (2008) aren't contradictory: we restrain ourselves, in our experimentations, to the case where linear sparsity holds, in order to compare as closely as possible the ordinary Lasso and the Bolasso.

4.2 The Bolasso algorithm

In this subsection, we develop the algorithm of the Bolasso, which essentially relies on using bootstrap in order to recover the true support of the Lasso. The idea is to randomly chose samples of the available dataset and run the Lasso several times on resampled datasets, to save the found supports, and then to define the true support to be the intersection of all the supports previously found. As we are only interested in the support recovery of the Lasso, we will restrain ourselves to the results on variable selection, not predictions. Furthermore, since we are interested in the Lasso path, we run this algorithm for all the λ's along the Lasso path.

Given data $(X_i^T, y_i)_{i=1}^n$, we choose K subsamples with replacement in a random manner. The subsamples are hence not independent, but the observations in each subsample are. On each subsample, we will estimate the support $\mathcal{T}^k, k \in 1, \dots, m$. Then the Bolasso's estimate of the support will be $\hat{\mathcal{T}} = \cap_{k=1}^m \hat{\mathcal{T}}^k$.

Bach (2008) proves that this algorithm gives a consistent model selection by giving an upper bound to the probability of wrong model selection. This however, is constrained to the choice of $\mu_n = \mu_0 \sqrt{n}$ with μ_0 in the following Lasso problem:

$$\hat{\beta}(\lambda) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2} \|y - Xb\|^2 + \mu_n \|b\|_1$$

This choice is justified by the fact that it leads, for each Lasso, to finding all the relevant variables with probability tending to 1 exponentially fast – although it also implies that the Lasso picks the irrelevant variables with a strictly positive probability (See Bach (2008)).

Proposition 1. *Assume **(A1-3)**. Then, under all the previously stated hypotheses:*

$$\mathbf{P}(\hat{\mathcal{T}} \neq \mathcal{T}) \leq m.A_1.e^{-A_2.n} + A_3.e^{\frac{\log(n)}{\sqrt{n}}} + A_4 \frac{\log(m)}{m}$$

The goal of this extension to the article is now to verify empirically if the trade-off theorem of our article still holds for the Bolasso in the conditions of the trade-off theorem. This is the object of the following experiments section.

4.3 Experiments on the Bolasso

Please refer to the notebook for details on the implementation of the algorithm. We run the Bolasso algorithm for different numbers m of resamples of the data. Below, we plot the performance of the Bolasso for $m = 1, \dots, 6$.

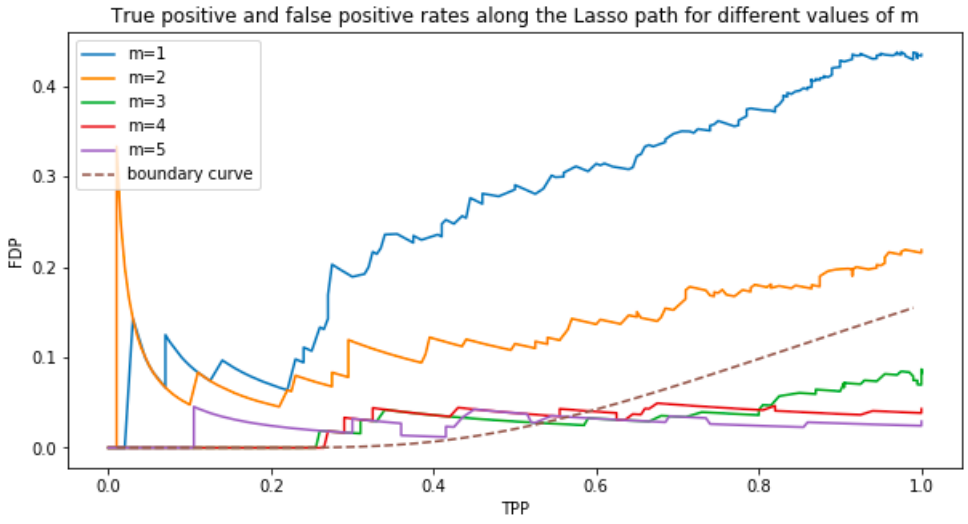

Figure 4: The Bolasso for m=1,...,6 and the voundary curve (dashed line)

We can see that as the number m increases, the Bolasso succeeds to make an almost perfect support recovery, without making too many false discoveries.

Documentation and sources

[1] Weiji Su, Malgorzat Bogdan, Emmnuel J. Candès — *False Discoveries occur Early on the Lasso Path*, Technical report (Stanford University. Department of Statistics), no. 2015-20, November 2015.

[2] Ryan Tibshirani and Larry Wasserman — *Sparsity and the Lasso*, Statistical Machine Learning course at Stanford, Spring 2015.

[3] Martin J. Wainwright — *Sharp threshold for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (Lasso)*, IEEE Transactions on Information Theory 55(5):2183–2202, 2009.

[4] Francis Bach — *Bolasso: Model Consistent Lasso Estimation through the Bootstrap.*