On the sign recovery given by the thresholded LASSO and thresholded Basis Pursuit

Patrick J C Tardivel, Maa Lgorzata Bogdan

On the sign recovery given by the thresholded LASSO and thresholded Basis Pursuit

Patrick J.C. Tardivel[a] *and Małgorzata Bogdan[a,b],

[a] Institute of Mathematics, University of Wrocład, Wrocław, Poland

[b] Lund University, Lund, Sweden

Abstract

We consider the regression model $Y = X\beta^* + \varepsilon$, when the number of observations n is smaller than the number of explicative variables p. It is well known that the popular Least Absolute Shrinkage and Selection Operator (LASSO) can recover the sign of β^* only if a very stringent *irrepresentable* condition is satisfied. In this article, in a first step, we provide a new result about the irrepresentable condition: the probability to recover the sign of β^* with the LASSO is smaller than 1/2 once the irrepresentable condition does not hold. On the other hand, LASSO can consistently estimate β^* under much weaker assumptions than the irrepresentable condition. This implies that appropriately thresholded LASSO can recover the sign of β^* under such weaker assumptions (see e.g. [24] or [34]). In this article we revisit properties of thresholded LASSO and provide new theoretical results in the asymptotic setup under which the design matrix is fixed and the magnitudes of nonzero components of β^* tends to infinity. Apart from LASSO, our results cover also basis pursuit, which can be thought of as a limiting case of LASSO when the tuning parameter tends to 0. Compared to the classical asymptotics with respect to n and p, our approach allows for reduction of the technical burden. In the result our main theorem takes a simple form:

Appropriately thresholded LASSO (with any given value of the tuning parameter) or thresholded basis pursuit can recover the sign of the sufficiently large signal if and only if β^* is *identifiable* **with respect to the l^1 norm, i.e.**

$$\textbf{If} \quad X\gamma = X\beta^* \text{ and } \gamma \neq \beta^* \quad \textbf{then} \quad \|\gamma\|_1 > \|\beta^*\|_1,$$

or in another words, when β^* can be recovered by solving the basis pursuit problem in the noiseless case.

For any given design matrix X, we define the *irrepresentability* and *identifiability* curves. For a given integer r, these curves provide the proportion of β^* having r nonzeros for which respectively the *irrepresentability* and *identifiability* conditions hold. These curves illustrate that the irrepresentable condition is

*Corresponding author: tardivel@math.uni.wroc.pl

much stronger than the identifiability condition (thus highlight our theoretical results) since the gap between the *irrepresentability* and *identifiability* curves is very large.

One notices that the identifiability curves drops very quickly from 1 to 0. These numerical observations are not surprising when X has i.i.d $\mathcal{N}(0,1)$ entries. Indeed, when n and p are both large there exists a value $k_{tr} \in (0,1)$ (given by the asymptotic transition curve [14]) such that the proportion of β^* identifiable with respect to the l^1 norm is close to 1 (resp. close to 0) as soon as $r/n < k_{tr}$ (resp. $r/n > k_{tr}$). Surprisingly, contrarily to classical assumptions (such as the irrepresentability), the identifiability condition does not become a very stringent condition when entries of X are extremely correlated. Indeed, the identifiability curve is the same when entries of X are extremely correlated as when entries of X has i.i.d $\mathcal{N}(0,1)$ entries. In addition, when the entries of X are positively correlated and the components of β^* have the same sign, the identifiability curve is highly above the one associated to i.i.d $\mathcal{N}(0,1)$ entries.

Finally, we illustrate how the knockoff methodology [2, 9] can be used to select the appropriate threshold and that thresholded basis pursuit and LASSO can recover the sign of β^* with a larger probability than adaptive LASSO [38].

Keywords: Active set estimation, Basis pursuit, Identifiability condition, Irrepresentability condition, LASSO, Sign estimation.

1 Introduction

Let us consider the high-dimensional linear Gaussian model

$$Y = X\beta^* + \varepsilon, \tag{1}$$

where $X = (X_1|\dots|X_p)$ is a $n \times p$ with $n < p$, ε is a centered Gaussian vector with $\mathrm{var}(\varepsilon) = \sigma^2 Id_n$ and $\beta^* \in \mathbb{R}^p$ is an unknown parameter. The sign vector of β^* is $\mathrm{sign}(\beta^*) = (\mathrm{sign}(\beta_1^*), \dots \mathrm{sign}(\beta_p^*)) \in \{-1, 0, 1\}^p$ where for $x \in \mathbb{R}$, $\mathrm{sign}(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$. Our main purpose is to recover $\mathrm{sign}(\beta^*)$. This objective is slightly more general than the aim at recovering the active set $\mathcal{A} := \{i \in \{1, \dots, p\} \mid \beta_i^* \neq 0\}$ (because when $\mathrm{sign}(\beta^*)$ is given one obtains $\mathcal{A}$ but the reverse does not hold). The main difficulty to recover $\mathrm{sign}(\beta^*)$ is to discriminate which components of β^* are exactly null. Using a sparse estimator (an estimator for which some components are equal to zero) is thus a natural way to recover $\mathrm{sign}(\beta^*)$. The LASSO estimator [11, 32] defined hereafter is probably the most famous sparse estimator

$$\hat{\beta}(\lambda) := \underset{\beta \in \mathbb{R}^p}{\mathrm{argmin}} \ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2}$$

When $\text{rank}(X) = n$, an other equivalent writing of the LASSO is given hereafter

$$\hat{\beta}_R := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|\beta\|_1 \text{ subject to } \|Y - X\beta\|_2^2 \leq R \tag{3}$$

with a one to one relation between the tuning parameter $\lambda > 0$ and the radius $R > 0$ under which these both problems share the same solution (see *e.g* the chapter 5.3 of the book [4]). The basis pursuit estimator is the solution of (3) when $R = 0$. This estimator is the limit of the LASSO as defined in (2) when the tuning parameter λ tends to 0 [13, 17].

Asymptotic properties of the sign estimator $\text{sign}(\hat{\beta}(\lambda)) := \Big(\text{sign}(\hat{\beta}_1(\lambda)), \ldots, \text{sign}(\hat{\beta}_p(\lambda))\Big)$ (or active set estimator $\mathcal{A}(\hat{\beta}(\lambda)) := \{i \in \{1, \ldots, p\} \mid \hat{\beta}_i(\lambda) \neq 0\}$) have been intensively studied [23, 37, 38].

When n tends to $+\infty$, p is fixed and the Gram matrix $\frac{1}{n}X'X$ converges to an invertible matrix, asymptotic properties are given by Yu et al. [37] for the sign estimator and by Zou [38] for the active set estimator. They proved that the irrepresentable condition is a necessary and "almost" sufficient condition under which, with a tuning parameter λ_n adequately chosen, the following convergence hold

$$\lim_{n \to +\infty} \mathbb{P}\Big(\text{sign}(\hat{\beta}(\lambda_n)) = \text{sign}(\beta^*)\Big) = 1 \text{ and } \lim_{n \to +\infty} \mathbb{P}\Big(\mathcal{A}(\hat{\beta}(\lambda_n)) = \mathcal{A}\Big) = 1.$$

These works had the merit to illuminate the irrepresentable condition as a key condition to recover $\text{sign}(\beta^*)$ with the LASSO but asymptotic results on n are not really interesting. Indeed, when $\frac{1}{n}X'X$ converges to an invertible matrix it is easy to build a consistent sign estimator based on the maximum likelihood estimator which does not require the irrepresentable condition to hold. In the noiseless case and in high dimension, when $n < p$, in their book page 192-194 Bühlmann and van de Geer [6] showed that the irrepresentable condition is a necessary and "almost" sufficient condition so that the non random set $\mathcal{A}(\beta(\lambda))$ of the non random LASSO $\beta(\lambda)$ converges to $\mathcal{A}$ once λ goes to 0. This result illuminates how the irrepresentable condition plays an important role for the active set in high dimension. Because the irrepresentable condition is a necessary and "almost" sufficient condition to recover the active set with the LASSO, this assumption or stronger assumptions are often met in applied and theoretical works [1, 19, 22, 25, 26, 35]. We arg that, one can recover $\text{sign}(\beta^*)$ under a much weaker assumption than the irrepresentable condition.

In this article we introduce a new condition called identifiability condition which is define hereafter.

Definition 1 (Identifiability) *Let X be a $n \times p$ matrix and let $\beta \in \mathbb{R}^p$, β is said to be identifiable with respect to the l^1 norm if the following implication hold*

$$X\gamma = X\beta \text{ and } \gamma \neq \beta \Rightarrow \|\gamma\|_1 > \|\beta\|_1. \tag{4}$$

Under the identifiability assumption, β^* is sparse. Indeed the lemma 3 given in Tardivel et al. [31] shows that $\text{card}\{i \in \{1,\dots,p\} \mid \beta_i^* \neq 0\} \leq n$ consequently, β^* has at least $p-n$ zeros. On the other hand some assumptions on the sparsity on β^* assures that β^* is identifiable with respect to the l^1 norm. For example when $\|X_1\|_2 = \cdots = \|X_p\|_2 = 1$ and the sparsity of β^* satisfies the following inequality (called mutual coherence condition)

$$\text{card}\{i \in \{1,\dots,p\} \mid \beta_i^* \neq 0\} \leq \frac{1}{2}\left(1 + \frac{1}{M}\right), \text{ where } M := \max_{i \neq j} |\langle X_i, X_j \rangle| \tag{5}$$

then β^* is identifiable with respect to the l^1 norm [15, 18, 21]. In the particular case in which the entries of X are i.i.d $\mathcal{N}(0,1)$ and n, p are both very large, the phase transition curve [16] provides, with respect to the ratio n/p, a range of sparsity under which β^* is identifiable with respect to the l^1 norm. To summarize, roughly speaking, β^* is identifiable with respect to the l^1 norm when β^* is sparse enough. Assuming that β^* is identifiable with respect to the l^1 norm is actually weaker than the usual assumptions did on β^* (see *e.g* [34] or [6] page 177). To be sure that the identifiability condition on β^* is very intuitive for a practitioner we have introduced the identifiability curve. Given an arbitrary design X, given an integer r, this curve provides the proportion of vectors β^* having r nonzeros components for which the identifiability condition holds.

Finally, we show that the sign estimator derived from thresholded LASSO and thresholded basis pursuit only need the very weak condition given in (4) to recover $\text{sign}(\beta^*)$.

1.1 organization of the article

In section 2, the theorem 1 provides a new look on the irrepresentable condition as a non-asymptotic necessary condition to recover $\text{sign}(\beta^*)$. The theorem 2 shows that the irrepresentable condition is stronger than the identifiability condition.

In section 3, we show that sign estimators derived from thresholded LASSO, and thresholded basis pursuit only need identifiability condition to recover asymptotically $\text{sign}(\beta^*)$.

The section 4 is devoted to numerical experiments. In this section, we introduce irrepresentability and identifiability curves. These curves provides respectively the maximal number of nonzero for β^* under which the LASSO sign estimator and the thresholded LASSO (resp. basis pursuit) sign estimator recover their target $\text{sign}(\beta^*)$ (as soon as nonzero components of β^* are large enough). When X is a Gaussian matrices with uncorrelated and strongly correlated entries, simulations show that the sign estimators derived from the thresholded LASSO, and thresholded basis pursuit are dramatically better than the LASSO sign estimator.

1.2 Notations and assumption

In this article we always assume that the design matrix X is in general condition (see *e.g* [33] for the definition). This assumption assures that the minimizer of (2) is unique and thus that the LASSO estimator is well defined. This assumption is very weak and generically holds. Indeed, when X is a random matrix such that the entries

$(X_{11}, X_{12}, \ldots, X_{np})$ have a density on $\mathbb{R}^{np}$ then, almost surely, X is in general position [33].

Hereafter the main notations used in this article:

- Let I be a subset of $\{1, \ldots, p\}$, we denote $\overline{I}$ the complement in $\{1, \ldots, p\}$ of I, namely $\overline{I} := \{1, \ldots, p\} \setminus I$.
- The notation X_I denotes for a matrix whose columns are $(X_i)_{i \in I}$.
- Let $\beta \in \mathbb{R}^p$, the notation β_I denotes for the vector $(\beta_i)_{i \in I}$ and when it is useful $[\beta]_i$ denotes the i^{th} component of β and $\text{supp}(\beta)$ denotes for the set $\{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$.

2 The identifiability condition is weaker than the irrepresentability condition

We already said that the irrepresentable condition is a necessary and "almost" sufficient condition to recover asymptotically $\text{sign}(\beta^*)$. Hereafter, in the high-dimensional linear Gaussian model, we have a new look on this well-known condition . The theorem 1 shows that when the irrepresentable condition does not hold namely

$$\|X'_{\overline{\mathcal{A}}} X_{\mathcal{A}} (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} \text{sign}(\beta^*_{\mathcal{A}})\|_\infty > 1,$$

then the probability to recover the sign of β^* cannot be close to 1. Let us point out that the theorem 1 is not asymptotic and deals with the high-dimensional setting contrarily to the theorems given in [37, 38] and there is a noise contrarily to the theorem given by Bühlmann and van de Geer [6].

Theorem 1 *Let X be a $n \times p$ matrix with $n < p$ in general position and let $\beta^* \in \mathbb{R}^p$ and let $\mathcal{A} := \{i \in \{1, \ldots, p\} \mid \beta^*_i \neq 0\}$. If the family $(X_i)_{i \in \mathcal{A}}$ is not linearly independent then whatever $\lambda > 0$, we have $\mathbb{P}(\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)) = 0$. If the family $(X_i)_{i \in \mathcal{A}}$ is linearly independent and the following inequality holds*

$$\|X'_{\overline{\mathcal{A}}} X_{\mathcal{A}} (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} \text{sign}(\beta^*_{\mathcal{A}})\|_\infty > 1$$

then whatever $\lambda > 0$, we have $\mathbb{P}(\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)) \leq 1/2$.

As a consequence of the theorem 1, the inequality $\|X'_{\overline{\mathcal{A}}} X_{\mathcal{A}} (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} \text{sign}(\beta^*_{\mathcal{A}})\|_\infty \leq 1$, called irrepresentable condition, is a necessary condition to recover $\text{sign}(\beta^*)$ with the LASSO (let us just remind that the Gram matrix $X'_{\mathcal{A}} X_{\mathcal{A}}$ is invertible if and only if $(X_i)_{i \in \mathcal{A}}$ is linearly independent). The theorem 2 shows that the irrepresentable condition on β^* is a stronger condition than the identifiability assumption on β^* given in (4).

Theorem 2 *Let X be a $n \times p$ matrix with $n < p$ in general position, let $\beta^* \in \mathbb{R}^p$, let $\mathcal{A} := \{i \in \{1, \ldots, p\} \mid \beta_i^* \neq 0\}$ and let us assume that the family $(X_i)_{i \in \mathcal{A}}$ is linearly independent. If the following inequality holds*

$$\|X'_{\overline{\mathcal{A}}} X_{\mathcal{A}} (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} sign(\beta^*_{\mathcal{A}})\|_\infty \leq 1,$$

then the parameter β^* is identifiable with respect to the l^1 norm, namely $X\beta = X\beta^*$ and $\beta \neq \beta^*$ implies $\|\beta\|_1 > \|\beta^*\|_1$.

Let us notice that when the inequality in the irrepresentable condition is strict instead of large the theorem 2 remains true without assuming that X is in general position. The two theorems 1 and 2 evidenced that when the irrepresentable condition does not hold the LASSO sign estimator does not recover $\text{sign}(\beta^*)$ even if β^* is identifiable with respect to the l^1 norm and the non null component of β^* are very large. The proof of the theorem 2 given in this article is the one reported in the PhD manuscript of Tardivel [30]. More recently, a result close to the theorem 2 was given in the proposition 1 in Descloux and Sardy [13]. The proof of the proposition 1 given in [13] is simple but need more backgrounds than the one given in this article which only need basic linear algebra computations.

Now, let us explain why the identifiability condition is weaker than the usual assumption given in the LASSO literature. Among the conditions reported in the LASSO literature, the compatibility condition is the weakest one [34]. The proposition 1 of Descloux and Sardy [13] shows that the compatibility condition implies the null space property. Finally, it is well known (see *e.g* the lemma 1 in [21]) that the null space property implies that β^* is identifiable with respect to the l^1 norm and that the reverse is not true.

2.1 Sign applications

One notices that the irrepresentable condition just depends from the sign of β^* and not on how large are the non null component of β^*. Given a particular design matrix X, the irrepresentability sign application is defined hereafter.

Irrepresentabity sign application:

$$\Phi_{\text{IC}}^X : s \in \{-1, 0, 1\}^p \mapsto \begin{cases} 1 \text{ if } s = (0, \ldots, 0) \\ 1 \text{ if } \ker(X_I) = \mathbf{0} \text{ and } \|X'_{\overline{I}} X_I (X'_I X_I)^{-1} s_I\|_\infty \leq 1 \text{ where } I := \text{supp}(s) \\ 0 \text{ otherwise} \end{cases} .$$

Given this sign application one determines exactly which are the parameters $\beta^* \in \mathbb{R}^p$ satisfying the irrepresentable condition. Such a sign application provides the limitation of the LASSO sign estimator to recover $\text{sign}(\beta^*)$. Indeed, if $\phi_{\text{IC}}^X(\text{sign}(\beta^*)) = 0$ then the sign of β^* cannot be recovered with the LASSO even if the non null components of β^* are extremely large. We are going to construct such a sign application for the

identifiability condition given in (4) and later we show that this sign application provides the limitation of sign estimators derived from the thresholded LASSO and thresholded basis pursuit to recover $\text{sign}(\beta^*)$. The proposition 1 shows that the identifiability condition just depends from $\text{sign}(\beta^*)$ and not on how large are the non null components of β^*.

Proposition 1 *Let X be a $n \times p$ matrix, let β^* be a parameter identifiable with respect to the l^1 norm and let $\tilde{\beta}$ be a parameter such that $\text{sign}(\beta^*) = \text{sign}(\tilde{\beta})$ then $\tilde{\beta}$ is identifiable with respect to the l^1 norm.*

Given a particular design matrix X, the identifiability sign application is defined hereafter.

Identifiability sign application:

$$\Phi^X_{\text{Idtf}} : s \in \{-1,0,1\}^p \mapsto \begin{cases} 0 \text{ if } s \neq \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|\beta\|_1 \text{ subject to } X\beta = Xs \\ \\ 1 \text{ otherwise} \end{cases} .$$

The restriction of these both sign applications to the set $E := \{s \in \{-1,0,1\}^p \mid \text{card}(\text{supp}(s)) \leq n\}$ is relevant. Indeed, when $\text{card}(\text{supp}(\beta^*)) > n$ the family $(X_i)_{i \in \mathcal{A}}$ is not linearly independent thus $\phi^X_{\text{IC}}(\text{sign}(\beta^*)) = \phi^X_{\text{Idtf}}(\text{sign}(\beta^*)) = 0$ (the proposition given supplementary material shows that $(X_i)_{i \in \mathcal{A}}$ not linearly independent implies that β^* does not satisfy the identifiability condition). Let us provides some basic properties and comments about these sign applications.

1. These two functions are even namely whatever $s \in \{-1,0,1\}^p$ the following equalities holds $\Phi^X_{\text{IC}}(s) = \Phi^X_{\text{IC}}(-s)$ and $\Phi^X_{\text{Idtf}}(s) = \Phi^X_{\text{Idtf}}(-s)$.

2. Due to the theorem 2, whatever $s \in \{-1,0,1\}^p$, $\Phi^X_{\text{IC}}(s) \leq \Phi^X_{\text{Idtf}}(s)$.

3. The computation of Φ^X_{IC} is a straightforward matricial computation; the computation of Φ^X_{Idtf} is no more difficult and need to solve a basis pursuit problem.

The last remark shows that given a parameter $\beta^* \in \mathbb{R}^p$, it is easy to check weather or not β^* is identifiable with respect to the l^1 norm.

Based on simulations, Su [28] already noticed that LASSO does not perform well to recover the active set $\mathcal{A}$ (see eg the figure 1 in [28]). We also aim at illustrating that LASSO does not perform well to recover $\text{sign}(\beta^*)$ by providing a toy example in which computations are easy to handle.

2.2 Example

Let us set $X = (X_1|X_2|X_3)$ where $X_1 = \begin{pmatrix} 2 & 2 \end{pmatrix}', X_2 = \begin{pmatrix} 4 & 2 \end{pmatrix}'$ and $X_3 = \begin{pmatrix} -1/3 & 1/3 \end{pmatrix}'$ and let $\beta^* := (\beta^*_1, 0, 0)$ with $\beta^*_1 \neq 0$. In this toy example, mathematical arguments and the figure 1 illustrates that the active set estimator given by the LASSO cannot recover $\mathcal{A}$ (thus the LASSO sign estimator cannot recover

sign(β^*)). Let us set $\tilde{X} = (X_1|X_2)$ and $\tilde{\beta}^{\text{ols}} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$. We claim that whatever $\beta_1^* \in \mathbb{R}^*$ whatever $\lambda > 0$, if $\mathcal{A}(\hat{\beta}(\lambda)) = \{1\}$ then $\tilde{\beta}_1^{\text{ols}}\tilde{\beta}_2^{\text{ols}} < 0$. It is straightforward that when $|\beta_1^*|$ is very large $\mathbb{P}(\tilde{\beta}_1^{\text{ols}}\tilde{\beta}_2^{\text{ols}} < 0) \approx \mathbb{P}(\text{sign}(\beta_1^*)\tilde{\beta}_2^{\text{ols}} < 0) = 1/2$. Consequently, even in the ideal setting in which β_1^* is extremely large the LASSO could not recover $\mathcal{A} = \{1\}$ with a probability close to 1. On the other hand, whatever $\lambda > 0$, the components $\hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda)$ and $\hat{\beta}_3(\lambda)$ satisfy the following equalities

$$\hat{\beta}_3(\lambda) = 0 \text{ and } (\hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda)) = \underset{\beta \in \mathbb{R}^2}{\text{argmin}} \ \|\tilde{X}(\tilde{\beta}^{\text{ols}} - \beta)\|_2^2 + \lambda\|\beta\|_1.$$

Schneider and Ewald [27] provide the map between the position of the ordinary least square estimator and the active set estimator of the LASSO. This result is useful to provide, in the figure 1, the relation between $\tilde{\beta}^{\text{ols}}$ and $\mathcal{A}(\hat{\beta}(\lambda))$ (let us notice that $\mathcal{A}(\hat{\beta}_1(\lambda), \hat{\beta}_2(\lambda)) = \mathcal{A}(\hat{\beta}(\lambda))$ in this setting).

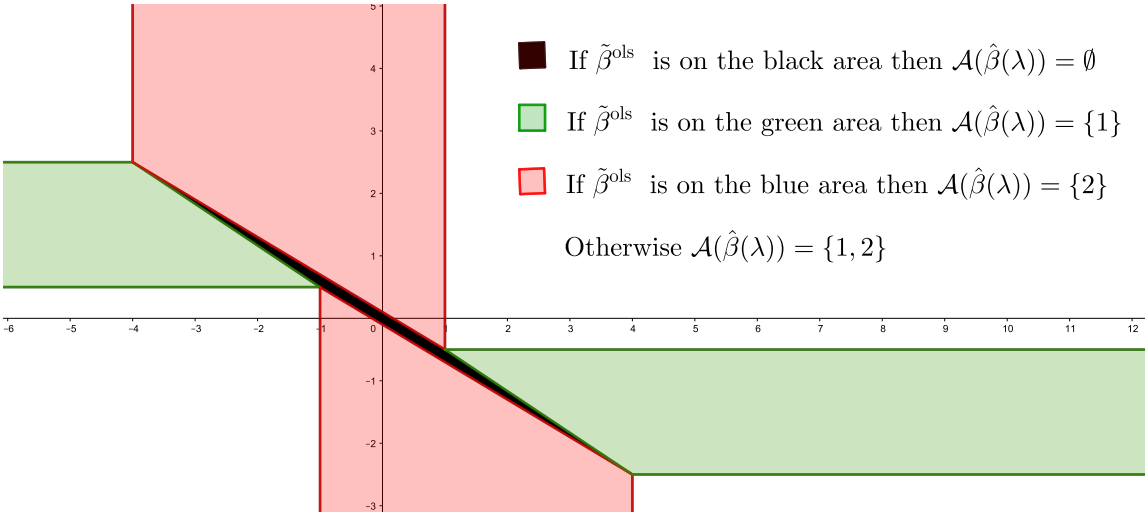

Figure 1: This figure provides $\mathcal{A}(\hat{\beta}(\lambda))$ with respect to the position of $\tilde{\beta}^{\text{ols}} = (\tilde{\beta}_1^{\text{ols}}, \tilde{\beta}_2^{\text{ols}})$ and in the particular case in which $\lambda = 2$. As an illustration that the LASSO does not provide a good active set estimator, let us notice that $\mathcal{A}(\hat{\beta}(\lambda)) = \{1, 2\}$ when $|\tilde{\beta}_1^{\text{ols}}| \geq 1$ and $|\tilde{\beta}_2^{\text{ols}}| \leq 0.5$. Consequently, the LASSO cannot recover $\mathcal{A} = \{1\}$ when $\tilde{\beta}_1^{\text{ols}}$ is too much large and $\tilde{\beta}_2^{\text{ols}}$ is too much small whereas, intuitively, this configuration seems to be the ideal one to recover $\mathcal{A}$.

In the following section we are going to show that sign estimators derived from thresholded LASSO and thresholded basis pursuit do not need that the irrepresentable condition to recover sign(β^*) but only need the weaker irrepresentability condition. Under this last assumption, we show that these sign estimators asymptotically recover exactly sign(β^*). We already said that asymptotic results when n goes to $+\infty$ and p fixed are not really interesting. In the following section we explore a new kind of asymptotic setting in which X and sign(β^*) are fixed and in which the non null components of β^* become arbitrary large.

3 Converging sign estimators

Let us introduce the following family of models

$$Y^k = X(k\beta^*) + \varepsilon \text{ where } k \in \mathbb{N}^*. \tag{6}$$

Obviously, in this family of models, $\text{sign}(k\beta^*)$ does not change with $k \in \mathbb{N}^*$ and the non null components of $k\beta^*$ become large when k increases. The theorem 1 and the previous toy example show that as soon as β^* does not satisfy the irrepresentable condition then, even if k goes to $+\infty$ and whatever $\lambda > 0$, the LASSO sign estimator $\text{sign}(\hat{\beta}(\lambda))$ fails to recover $\text{sign}(\beta^*)$. Fortunately, the irrepresentable condition is not an unsurpassable limitation to recover $\text{sign}(\beta^*)$ and the LASSO is not so bad for sign recovery; this last estimator just need to be a little bit modified. Actually the theorem 3 shows that an appropriately thresholded LASSO (resp. basis pursuit) recover asymptotically $\text{sign}(\beta^*)$ under the identifiability condition on β^* (which is, by the theorem 2, weaker than the irrepresentability condition).

To provide a result in broad a generality we do not assume, in the theorem 3, that β^* is identifiable with respect to the l^1 norm.

Let us denote by $\hat{\beta}^k_R(\epsilon)$ the following estimator

$$\hat{\beta}^k_R(\epsilon) := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|\beta\|_1 \text{ subject to } \|Y^k(\epsilon) - X\beta\|_2^2 \leq R. \tag{7}$$

We now define $\tilde{\beta}$ as the basis pursuit solution in the noiseless case as follows

$$\tilde{\beta} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \|\beta\|_1 \text{ subject to } X\beta = X\beta^*,$$

and denote $\mathcal{B}^- = \{i \in \{1, \dots, p\} \mid \tilde{\beta}_i < 0\}$, $\mathcal{B}^+ = \{i \in \{1, \dots, p\} \mid \tilde{\beta}_i > 0\}$ and $\mathcal{B} = \mathcal{B}^- \cup \mathcal{B}^+$.

Theorem 3 *Let X be a $n \times p$ matrix in general position such that $\text{rank}(X) = n$. Then, for any fixed $\varepsilon \in R^n$, $R \geq 0$ and sufficiently large $k > k_0(R, \varepsilon)$ it holds*

Separation property:

$$\max_{i \notin \mathcal{B}^-} \left\{ [\hat{\beta}^k_R(\epsilon)]_i \right\} < \min_{i \notin \mathcal{B}} \left\{ [\hat{\beta}^k_R(\epsilon)]_i \right\} \leq \max_{i \notin \mathcal{B}} \left\{ [\hat{\beta}^k_R(\epsilon)]_i \right\} < \max_{i \notin \mathcal{B}^+} \left\{ [\hat{\beta}^k_R(\epsilon)]_i \right\}.$$

Sign recovery: *The equality $\text{sign}(\beta^*) = \text{sign}(\tilde{\beta})$ occurs (thus $\mathcal{B}^- = \{i \in \{1, \dots, p\} \mid \beta^*_i < 0\}$ and $\mathcal{B}^+ = \{i \in \{1, \dots, p\} \mid \beta^*_i > 0\}$) if and only if β^* is identifiable with respect to the l^1 norm.*

Let us notice that the assumptions on X are very weak and generically hold when $n \leq p$. The assumption $\text{rank}(X) = n$ assures that whatever $R \geq 0$ the feasible set $\{\beta \in \mathbb{R}^p \mid \|Y^k - X\beta\|_2^2 \leq R\}$ is not empty. The general

position condition assures the uniqueness of $\hat{\beta}_R^k$ (see *e.g* the proposition 1 given in supplementary material for a proof).

The estimator is easy to compute because $\hat{\beta}_R^k$ is the solution of a convex problem. Actually, when $R > 0$ the estimator given in (7) is just an other writing of the standard LASSO estimator as given in (2) (see *e.g* the chapter 5.3 of the book [4]). The expression given in (7) has several advantages. The first one, to our opinion, the theorem 3 is more intuitive with this writing than with the standard LASSO writing as given in (2). The second one, the initial estimator is not restricted to LASSO estimator indeed, when $R = 0$, $\hat{\beta}_0^k$ is a basis pursuit estimator.

The theorem 3 stress that one cannot recover $\text{sign}(\beta^*)$ with a sign estimator derived from (7) when β^* is not identifiable with respect to the l^1 norm since $\text{sign}(\beta^*) \neq \text{sign}(\tilde{\beta})$ (with $\tilde{\beta}$ as defined in the theorem 3). When β^* is identifiable with respect to the l^1 norm (then $\tilde{\beta} = \beta^*$), the theorem 3 does not provide explicitly a converging sign estimator for $\text{sign}(\beta^*)$ but the good properties of the initial estimator suggest many ways to construct one.

Probably the most intuitive way to recover $\text{sign}(\beta^*)$ is to derive sign estimator from the thresholded LASSO estimator ($R > 0$) or thresholded basis pursuit estimator ($R = 0$). The expression of this thresholded estimator is reported in the expression (8) given below. By the the separation property, one knows that it remains to select a good threshold τ to construct a consistent sign estimator (with τ depending from k for the consistency). An alternative way to recover the $\text{sign}(\beta^*)$ is to use the adaptive LASSO. In this case, the keystone is to derive the weights of the adaptive LASSO from the estimator given in (7). Theoretical justifications of the consistency of the sign estimator derived from the adaptive LASSO are given in [38]. However, we point out that the proof some arguments given in the theorem 2 in [38] (dealing with the consistency of the sign estimator derived from adaptive LASSO) are not correct. Indeed, the pointwise convergence of a sequence of convex functions $f_n : \mathbb{R}^p \to \mathbb{R}$ to a function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ does not imply that the sequence $(x_n^*)_{n\in\mathbb{N}}$ (minimizers of f_n) converges to x^* (minimizer of f)[1]. Thus, we do not know if whether or not the result claimed in this theorem is correct.

The theorem 3 confirms recent results given by Bogdan et al. [5]. Indeed, if X has i.i.d $\mathcal{N}(0,1)$ entries, $n/p \to \delta \in (0,1)$ and asymptotically the point $(\text{card}(\mathcal{A})/n, n/p)$ is below the asymptotic phase transition curve [14] (*i.e.* β^* is asymptotically identifiable with respect to the l^1 norm) then the thresholded LASSO almost certainly recovers $\text{sign}(\beta^*)$ (as soon as nonzero components of β^* are large enough).

Obviously, the performance of the sign estimators derived from thresholded LASSO and thresholded basis pursuit depends from the tuning parameter and the threshold. In the numerical experiments, we are going to prescribe values for these parameters.

[1]For example $f_n(x_1, x_2) = |-x_1 + nx_2| + |x_2 - n| - n$ converges pointwise to $f(x_1, x_2) = \begin{cases} |x_1| \text{ if } x_2 = 0 \\ +\infty \text{ if } x_2 \neq 0 \end{cases}$ but $x_n^* = (n^2, n)$ does not converge to $x^* = (0, 0)$.

4 Numerical experiments

We have seen in the theorem 2 that the the irrepresentable condition implies the identifiability condition. The identifiability and irrepresentability curves given in the next section allow to quantify the gap between these conditions. In addition, these two curves provide respectively the maximal number of nonzero for β^* under which sign estimators derived from thresholded LASSO (resp. basis pursuit) and LASSO recover $\text{sign}(\beta^*)$ (as soon as nonzero components of β^* are large enough).

4.1 Identifiability and irrepresentability curves with random Gaussian matrices

We previously define the identifiability and irrepresentability sign functions denoted Φ^X_{Idtf} and Φ^X_{IC}. Given a design matrix X these two sign functions gives *a priori* the limitation of the LASSO sign estimator or the limitation of sign estimators derived from thresholded LASSO and thresholded basis pursuit to recover exactly $\text{sign}(\beta^*)$. Indeed, when $\Phi^X_{\text{IC}}(\text{sign}(\beta^*)) = 0$ (resp. $\Phi^X_{\text{Idtf}}(\text{sign}(\beta^*)) = 0$) the LASSO sign estimator (resp. sign estimators derived from thresholded LASSO and thresholded basis pursuit) cannot recover $\text{sign}(\beta^*)$ with a probability close to 1 even if the non null components of β^* are extremely large. The number of sign vectors is very huge (3^p), that is why we are not going to provide explicitly Φ^X_{Idtf} and Φ^X_{IC} for each sign vector. Instead, for each $r \in \{1, \dots, n\}$, we are going to compute empirically $p^X_{\text{Idtf}}(r) := \mathbb{E}_U(\Phi^X_{\text{Idtf}}(U))$ and $p^X_{\text{IC}}(r) := \mathbb{E}_U(\Phi^X_{IC}(U))$ where U is a uniformly distributed on $\{u \in \{-1, 0, 1\}^p \mid \text{card}(\text{supp}(u)) = r\}$. The identifiability and irrepresentability curves represents respectively the curves of the functions $r \in \{1, \dots, n\} \mapsto p^X_{\text{Idtf}}(r)$ and $r \in \{1, \dots, n\} \mapsto p^X_{\text{IC}}(r)$. In the numerical experiments given in the figure 2, X is a Gaussian matrix described hereafter.

Setting 1: The matrix X is a $n \times p$ matrix with $n = 100$, $p = 300$ and $(X_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ are i.i.d $\mathcal{N}(0, 1)$.

Setting 2: The matrix X is a $n \times p$ matrix with $n = 100$, $p = 300$ and the vectors $(X_{ij})_{1 \leq j \leq p}$ where $i \in \{1, \dots, n\}$ is a family of i.i.d Gaussian vector $\mathcal{N}(\mathbf{0}, \Gamma)$. In this setting Γ is a $p \times p$ matrix where $\Gamma_{ii} = 1$ with $i \in \{1, \dots, p\}$ and $\Gamma_{ij} = \rho$ when $i \neq j$.

In these simulations the curves are obtained from a particular observation of X.

Surprisingly the two identifiability curves given in the setting 1 and 2 are very similar. *A priori*, we expected to recover a curve in the setting 2 much below than the one given in the setting 1. Indeed, classical conditions implying the identifiability of β^* with respect to the l^1 norm are the mutual coherence condition (5) and the restricted isometry property [7, 8]. These conditions are quite weak when the family $(X_i)_{1 \leq i \leq p}$ is almost orthogonal (as in the setting 1 since $\mathbb{E}(X'X) = nId_n$) but are very strong when $(X_i)_{1 \leq i \leq p}$ is far from an orthogonal family (as in the setting 2 since $\mathbb{E}(X'X) = n\Gamma$).

The asymptotic phase transition given in Donoho and Tanner [16] provides an approximation of the identifiability curve in the setting 1. Such an approximation is useful when n and p are too much large so that the

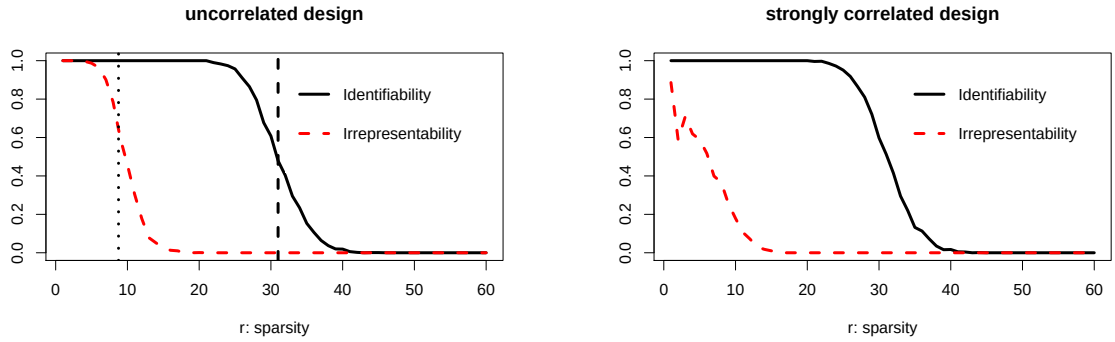

Figure 2: This figure gives the curves of the functions $r \mapsto p^X_{\mathrm{Idtf}}(r)$ and $r \mapsto p^X_{\mathrm{IC}}(r)$ when X is a Gaussian matrix given in the setting 1 (left panel) and setting 2 (right panel). Due to the theorem 2, we already known that whatever the sparsity r, $p^X_{\mathrm{Idtf}}(r) \geq p^X_{\mathrm{IC}}(r)$ thus this figure just emphasizes that the identifiability condition is a much weaker assumption than the irrepresentability condition. The vertical lines in the left panel provides, in the setting 1, an asymptotic approximation of the identifiability and irrepresentability curves. Indeed by the theorem 1 in [16] and the theorem 1 in [35], when p is very large and $n/p = 1/3$ then the identifiability and irrepresentability conditions hold respectively when $r \leq 0.31n$ and $r \leq 0.09n$. To plot these these curves, for a sparsity r the quantities $p^X_{\mathrm{Idtf}}(r)$ and $p^X_{\mathrm{IC}}(r)$ have been computed by simulating 1000 observations of the random vector U.

identifiability curve is too much time expensive to obtain. Unfortunately, to our knowledge, there is not such asymptotic phase transition for Gaussian matrices with correlated entries as in the setting 2.

One notices that in the setting 2, the irrepresentability curve is not monotonic in the neighbourhood of 0; it is not a numerical problem. Actually when r is very small, U has frequently components which are all positive or all negative. Furthermore the figure 3 illustrates that, in the setting 2, when the sign vector s is positive componentwise (resp. negative componentwise), the irrepresentable condition becomes a very strong condition. These both remarks, aim at explaining why, in the setting 2, the irrepresentabilty curve is not monotonic. Hereafter, without any loss of generality, we focus on the particular case in which sign vector is positive componentwise. The figure 3 provides the positive irrepresentability and identifiability curves, which are respectively the curves of the functions $r \mapsto p^X_{\mathrm{Idtf}+}(r) := \mathbb{E}_U(\Phi^X_{\mathrm{Idtf}}(U))$ and $r \mapsto p^X_{\mathrm{IC}+}(r) := \mathbb{E}_U(\Phi^X_{IC}(U))$ where U has uniform distribution over the set $\{u \in \{0,1\}^p \mid \mathrm{card}(\mathrm{supp}(u)) = r\}$.

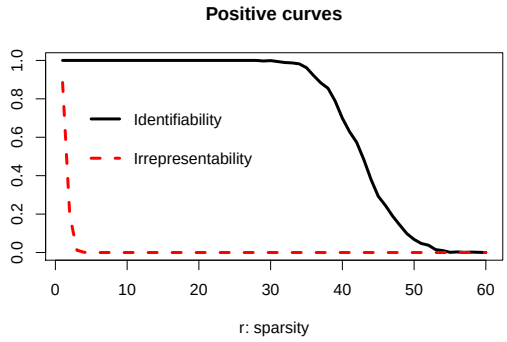

Figure 3: This figure gives the curves of the functions $r \mapsto p^X_{\mathrm{Idtf}+}(r)$ and $r \mapsto p^X_{\mathrm{IC}+}(r)$ when X is a Gaussian matrix given in setting 2. One notices that, with respect to the curves given in the previous figure, the gap between the irrepresentable condition and the identifiability condition becomes larger. In the setting 2, when r is small $p^X_{\mathrm{IC}+}(r) \approx p^X_{\mathrm{IC}}(r)$ (more precisely, $p^X_{\mathrm{IC}+}(r) = p^X_{\mathrm{IC}}(r)$ when $r = 1$) and when r is large enough $p^X_{\mathrm{IC}}(r)$ weakly depends from the correlation of the columns of X. This remark aim at explaining why, in the setting 2, the function $r \mapsto p^X_{\mathrm{IC}}(r)$ is not monotonic in the neighbourhood of 0. To plot these these curves, for a sparsity r, the quantities $p^X_{\mathrm{Idtf}+}(r)$ and $p^X_{\mathrm{IC}+}(r)$ have been computed by simulating 1000 observations of U.

4.2 Numerical comparisons of sign estimators

The theorem 3 suggests many ways to recover $\text{sign}(\beta^*)$, for example, by deriving a sign estimator from adaptive LASSO, thresholded basis pursuit, thresholded LASSO... The purpose of this section is to provide a numerical comparison of these sign estimators under four different simulation scenarios in the high dimensional setup.

4.2.1 Selection of the threshold

We aim at constructing a sign estimator derived from the following thresholded estimator

$$\forall i \in \{1,\dots,p\}, [\hat{\beta}_R^\tau]_i := [\hat{\beta}_R]_i \mathbf{1}_{\left\{|[\hat{\beta}_R]_i|>\tau\right\}} \text{ with } \tau > 0. \tag{8}$$

This estimator is a thresholded basis pursuit when $R = 0$ and a thresholded LASSO when $R > 0$. Given a threshold $\tau > 0$, the probability to recover exactly $\text{sign}(\beta^*)$ with $\text{sign}(\hat{\beta}_R^\tau)$ is described hereafter

$$\mathbb{P}(\text{sign}(\hat{\beta}_R^\tau) = \text{sign}(\beta^*)) = \mathbb{P}\left(\forall i \in \overline{\mathcal{A}}, [\hat{\beta}_R^\tau]_i = 0 \text{ and } \forall i \in \mathcal{A}, \text{sign}([\hat{\beta}_R^\tau]_i) = \text{sign}(\beta_i^*)\right).$$

By choosing a threshold $\tau_{1-\alpha}$ as the $1-\alpha$ quantile of $\max\left\{\left|[\hat{\beta}_R]_i\right|, i \in \overline{\mathcal{A}}\right\}$ then $\mathbb{P}(\forall i \in \overline{\mathcal{A}}, [\hat{\beta}_R^{\tau_{1-\alpha}}]_i = 0) = 1-\alpha$. Consequently when the non null component of β^* are very large then $\mathbb{P}(\text{sign}(\hat{\beta}_R^{\tau_{1-\alpha}}) = \text{sign}(\beta^*))$ becomes arbitrarily close to $1-\alpha$. Obviously $\tau_{1-\alpha}$ cannot be obtained by a straightforward computation since β^* and thus $\mathcal{A}$ are not known.

Let $\tau > 0$ and let us set $\text{FWER} := \mathbb{P}(\exists i \in \overline{\mathcal{A}}, [\hat{\beta}_R^\tau]_i \neq 0)$ (the FWER can be seen as the Family Wise Error Rate in multiple testing procedure). Since $\tau_{1-\alpha}$ cannot be evaluated, the usual way to proceed is to select a threshold τ such that $\tau \geq \tau_{1-\alpha}$ with τ as close as possible to $\tau_{1-\alpha}$ assuring that both $\text{FWER} \leq \alpha$ and that the FWER stays close to α. In order to provide a threshold larger than $\tau_{1-\alpha}$, it could seem appealing to look at the distribution of the initial estimator in the full null model, when $\beta^* = \mathbf{0}$, and to compute $\tau_{1-\alpha}^{\text{fn}}$ as the $1-\alpha$ quantile of $\max\left\{\left|[\hat{\beta}_R^{\text{fn}}]_1\right|, \dots, \left|[\hat{\beta}_R^{\text{fn}}]_p\right|\right\}$ [20]; the random vector $\hat{\beta}_R^{\text{fn}}$ is defined hereafter

$$\hat{\beta}_R^{\text{fn}} := \text{ argmin } \|\beta\|_1 \text{ subject to } \|\varepsilon - X\beta\|_2^2 \leq R.$$

When $R = 0$, Descloux and Sardy [13] suggest this way of proceed to pick a threshold for the basis pursuit estimator. Unfortunately, in the high-dimensional linear model, this intuitive method provides a threshold $\tau_{1-\alpha}^{\text{fn}}$ which is smaller than $\tau_{1-\alpha}$ and thus does not assure that $\text{FWER} \leq \alpha$. Actually, as illustrated in supplementary material when $R = 0$, the distribution of $\max\left\{\left|[\hat{\beta}_0]_i\right|, i \in \overline{\mathcal{A}}\right\}$ depends from β^* and is stochastically larger than the one of $\max\left\{\left|[\hat{\beta}_0^{\text{fn}}]_1\right|, \dots, \left|[\hat{\beta}_0^{\text{fn}}]_p\right|\right\}$ especially when both $\text{card}(\mathcal{A})$ and the non null components of β^* are large.

Indeed, as explained e.g. in [29], the variance of estimates of LASSO regression coefficients increases with

the number and the magnitude of nonzero regression coefficients and this effect is not appropriately taken into account when calculating the threshold proposed in [13]. Instead, one can use recently developed knockoff methodology [2, 9], which allows to predict the magnitude of estimates corresponding to false regressors by creating fake copies of explanatory variables. The copy of a given explanatory variable has the same correlation with the remaining explanatory variables as its original and at the same time is conditionally independent of the response. The knockoff methodology allows to control the false discovery rate by setting the threshold on the difference of the importance statistic (say LASSO regression estimate) between the true explanatory variable and its fake copy. In many practical situations the standard implementation of knockoffs yields high power of detection of true signals. However, the power of this standard implementation is limited when the true number of nonzero regression coefficients is very small or when p is substantially larger than n. While it seems possible to extend the formal knockoff methodology to deal with these situations, in this manuscript we use model free knockoffs proposed in [9] to heuristically approximate the threshold to control the FWER at the assumed level. Specifically, at the first step we use model free knockoffs to generate $30 = p/10$ of fake variables. Then Lasso or BP is run on the matrix supplemented with these additional columns and the maximum of the absolute values of regression coefficients over 30 fake variables is saved. This step is repeated 10 times and the overall maximum of the $p = 300$ absolute values of regression coefficients over fake variables is calculated. The whole procedure is repeated many (here 1000) times and 0.95 quantile of the obtained maxima is used as the threshold to identify important regressors for the LASSO run on the original design matrix X. Similar approach for generating small knockoff matrices was proposed in [36], where a formal knockoff procedure for controlling FDR for gaussian design matrices with independent entries was proposed.

To confirm with the set-up of simulations used to derive the irrepresentability and identifiability curves, in all replicates of our simulation study we used the same fixed design matrix X. In each of the iterations of our experiment we randomly sampled the location of the true signals and the error term. In this situation the threshold proposed in [13] remains constant but the knockoff threshold in principle should differ between different iterations. To reduce the computation burden of our simulations we calculated only one "averaged" knockoff threshold, where in each of 1000 replicates performed to calculate the 0.95 quantile of the maximum of the fake statistics we randomly selected the location of true signals and the error term.

4.2.2 Selection of the tuning parameter

In the simulation study we compared the Basis Pursuit with thresholded LASSO. In case where X is the gaussian matrix with independent entries the tuning parameter was selected with the help of the asymptotic theory of Approximate Message Passing Algorithm (AMP) for LASSO, provided e.g. in [3, 29, 5]. In the setup of this theory the elements of design matrix come from a normal distribution $x_{ij} \sim N(0, 1/\sqrt{n})$, $n/p \to \delta > 0$ and regression coefficients are modeled as iid random variables from a mixture Π of a point mass at zero and

some other distribution $\Pi^\star$: $\Pi = (1-\gamma)\delta_0 + \gamma\Pi^\star$. The sparsity parameter γ defines the mixing proportion of nonzero coefficients. AMP theory allows for evaluation of the asymptotic standard deviation of the noise generated by the shrinkage $\tau = \tau(\lambda, \delta, \gamma, \Pi^\star)$ and selection of $\lambda_{AMP} = \lambda_{AMP}(\delta, \gamma, \Pi^\star)$ for which this noise is minimal. As discussed in [5], this selection of λ allows to maximize the power for any fixed type I error. When calculating the value of the tuning parameter λ_{AMP} corresponding to the minimal noise, we replaced the asymptotic parameters of the AMP theory with their finite sample counterparts

- undersampling $\delta = n/p = 100/300$
- sparsity $\gamma = r/p = r/300$
- signal distribution $\Pi^\star = \delta_t$ where δ_t is a one-point distribution concentrated at t.

The formulas to evaluate the standard deviation of the noise $\tau = \tau(\lambda, \delta, \gamma, \Pi^\star)$ are provided e.g. in [3, 29, 5].

In case of strongly correlated design we additionally use $\lambda_s = 0.5\lambda_{AMP}$.

4.2.3 LASSO and Adaptive LASSO

In our numerical experiments we selected the following values of the tuning parameters for LASSO and adaptive LASSO:

- For LASSO we selected $\lambda_L = 1.5\lambda_{Bon} = 1.5\frac{\Phi^{-1}\left(1-\frac{0.05}{2p}\right)}{\sqrt{n}}$, which is slightly larger than λ_{Bon}, needed to control FWER at the level 0.05 when the design matrix is orthogonal.
- For the adaptive LASSO the weights are derived using initial estimates $\hat{\beta}(\lambda_{AMP})$, where the tuning parameter is selected according to AMP theory, described above. The weight is defined as $w(\beta_i) = \frac{1}{\hat{\beta}_i(\lambda_{AMP})+10^{-7}}$.
- The final decision for adaptive LASSO is based on LASSO with $\lambda = \lambda_L$.

In other words, the adaptive LASSO is given hereafter

$$\hat{\beta}^{\text{adapt}} := \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_L \sum_{i=1}^{p} w(\beta_i)|\beta_i|. \tag{9}$$

In all our simulations LASSO is calculated with *glmnet*, with the convergency diagnostic parameter *thresh* equal to 10^{-12}. Package *default* value $thresh = 10^{-7}$ leads to large errors and misleading estimates of statistical properties of LASSO in case where the design matrix is strongly correlated.

4.2.4 Numerical comparisons

The rows of the design matrix X are sampled as the independent vectors from the multivariate Gaussian distribution, as in setting 1 and 2. All numerical experiments are performed with a particular observation of

X (the same as the one used in the previous subsection). We set $\beta^* \in \mathbb{R}^p$ such that $r := \text{card}\{i \mid \beta_i^* \neq 0\}$ with $r = \{5, 20\}$, $\{i \mid \beta_i^* \neq 0\}$ is a r sample without replacement of $\{1, \ldots, p\}$. The non null components of β^* have a symmetric two point distribution $P(\beta_i = -t) = P(\beta_i = t) = 0.5$ where we consider an increasing sequence of signal magnitudes $t \in \{0.5, 1, \ldots, 15\}$. Additionally, for strongly and positively correlated explanatory variables we consider the setup where all nonzero coefficients are equal to t. In all simulations the error term is generated as $\varepsilon \sim \mathcal{N}(0, Id_n)$.

Figures 4-6 provide the comparison between the following sign estimators.

- **L** is derived from LASSO with $\lambda = \lambda_L$
- **adL** is the adaptive LASSO estimator, described above
- **BPS** is the thresholded Basis Pursuit, with threshold selected as in [13]
- **BPkn** is the thresholded Basis Pursuit, with "knockoff" threshold defined above
- **Lkn** is the thresholded LASSO with $\lambda = \lambda_{AMP}$ and "knockoff" threshold
- **Lkns** is the thresholded LASSO with $\lambda = 0.5\lambda_{AMP}$ and "knockoff" threshold

We report the curves illustrating the following statistical properties as the function of the signal strenght:

- **Probability** is the proportion of 1000 replicates for which the sign was appropriately discovered
- **Power** is the average number of True Positives over all 1000 replicates
- **EFP** is the average number of False Positives over all 1000 replicates

Figure 4-6 illustrate that indeed, LASSO can not deal well with the number of signals which exceed the irrepresentability curve, while thresholded LASSO and Basis Pursuit can appropriately identify the sign if the number of signals is below identifiability curve. Basis Pursuit performs pretty well under our simulated setup, but its performance is worse than the performance of LASSO with appropriately selected λ. In case of the design matrix with independent columns the optimal value of λ selected by AMP theory yields the thresholded LASSO which is superior over all methods compared in our simulations. Maybe surprisingly, this choice of λ performs well also under strongly correlated design and when all signals are of the same sign. In case of the correlated design with the symmetric signal distribution it is better to use $\lambda = 0.5\lambda_{AMP}$. Our simulations show also that the threshold selection provided in [13] does not allow for the sign recovery if the number of true signals is relatively large. Instead, our heuristic application of the knockoff methodology allows for almost perfect control of EFP at the level of 0.05.

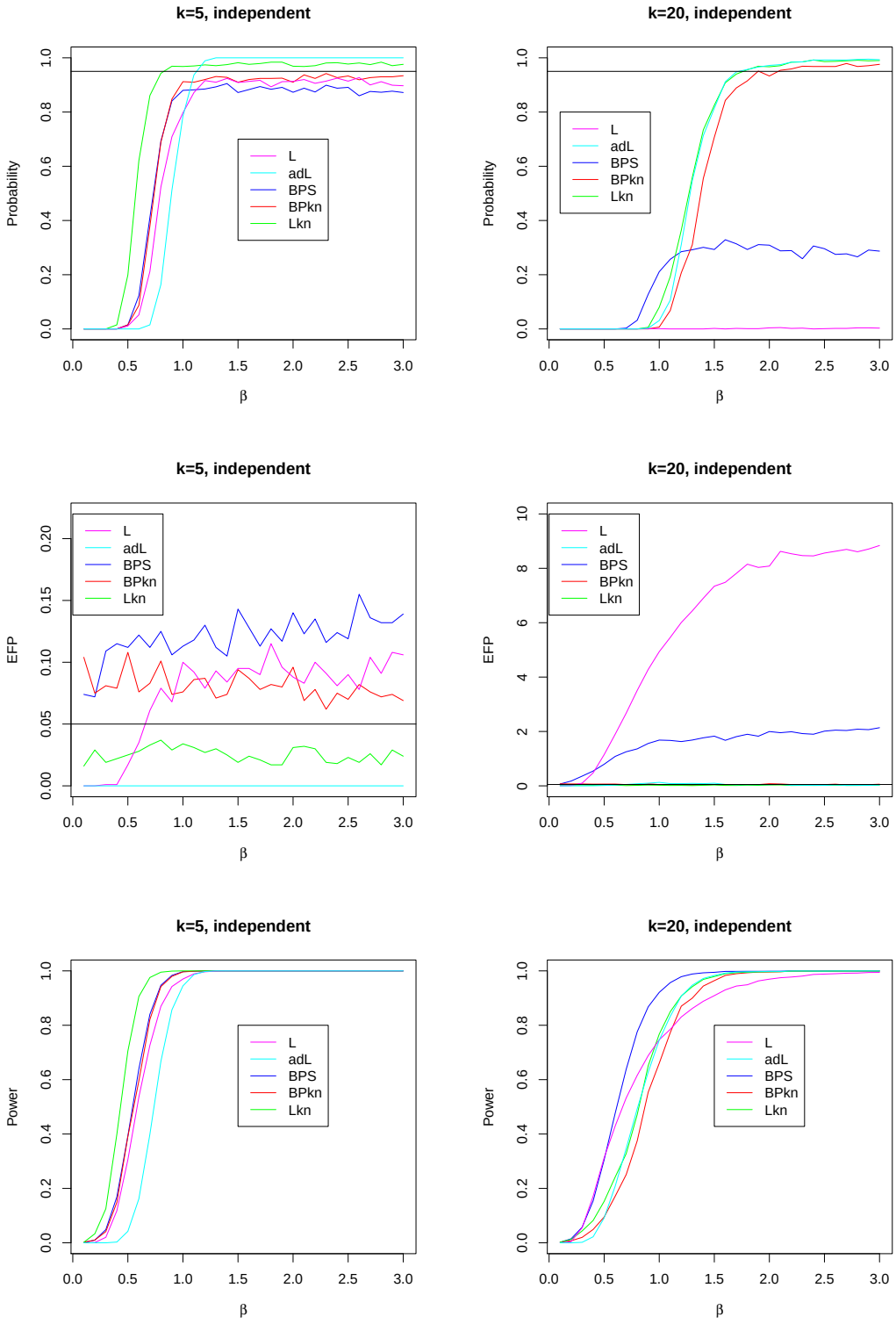

Figure 4: Design with independent columns

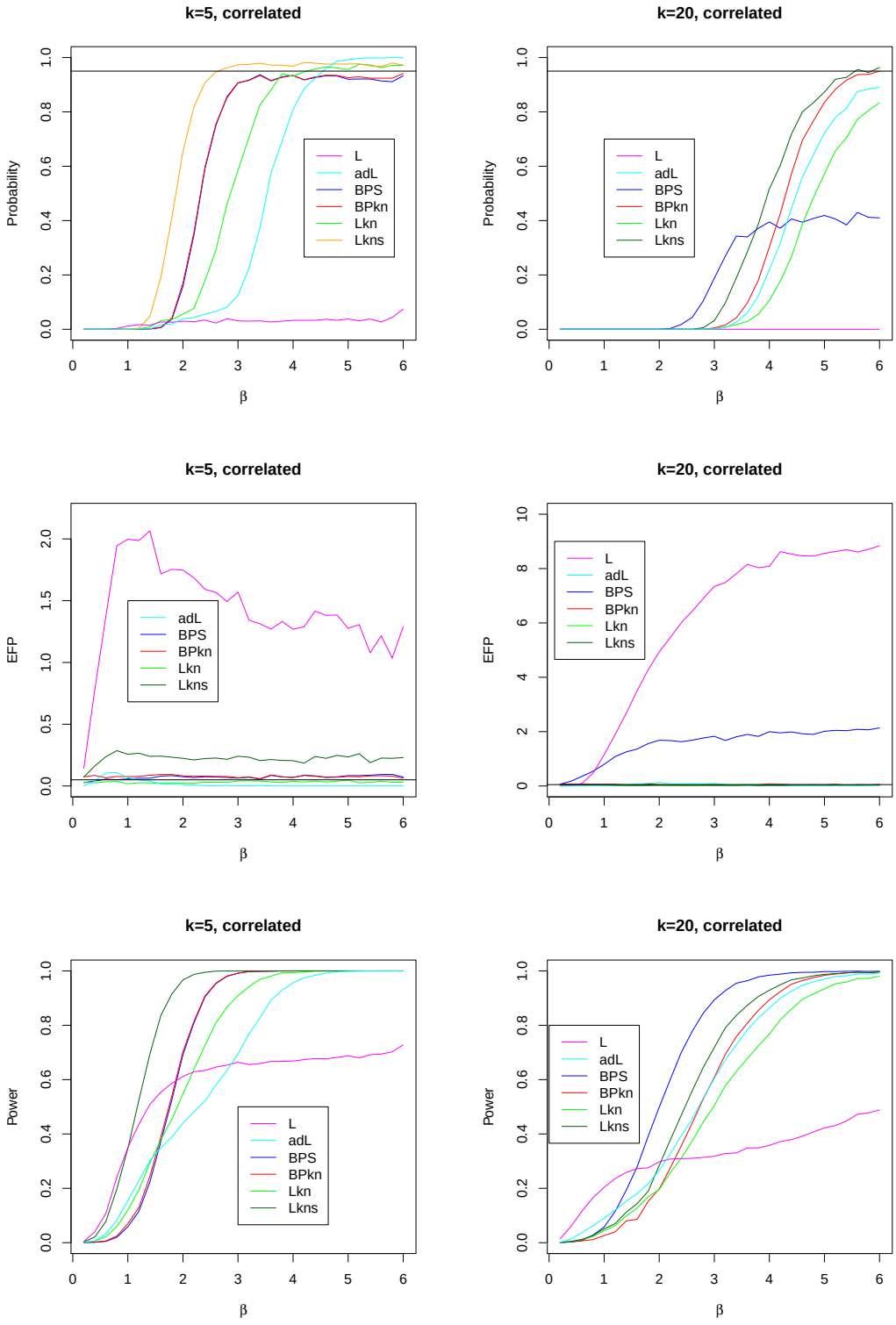

Figure 5: Design with strongly correlated columns and symmetric sygnal distribution

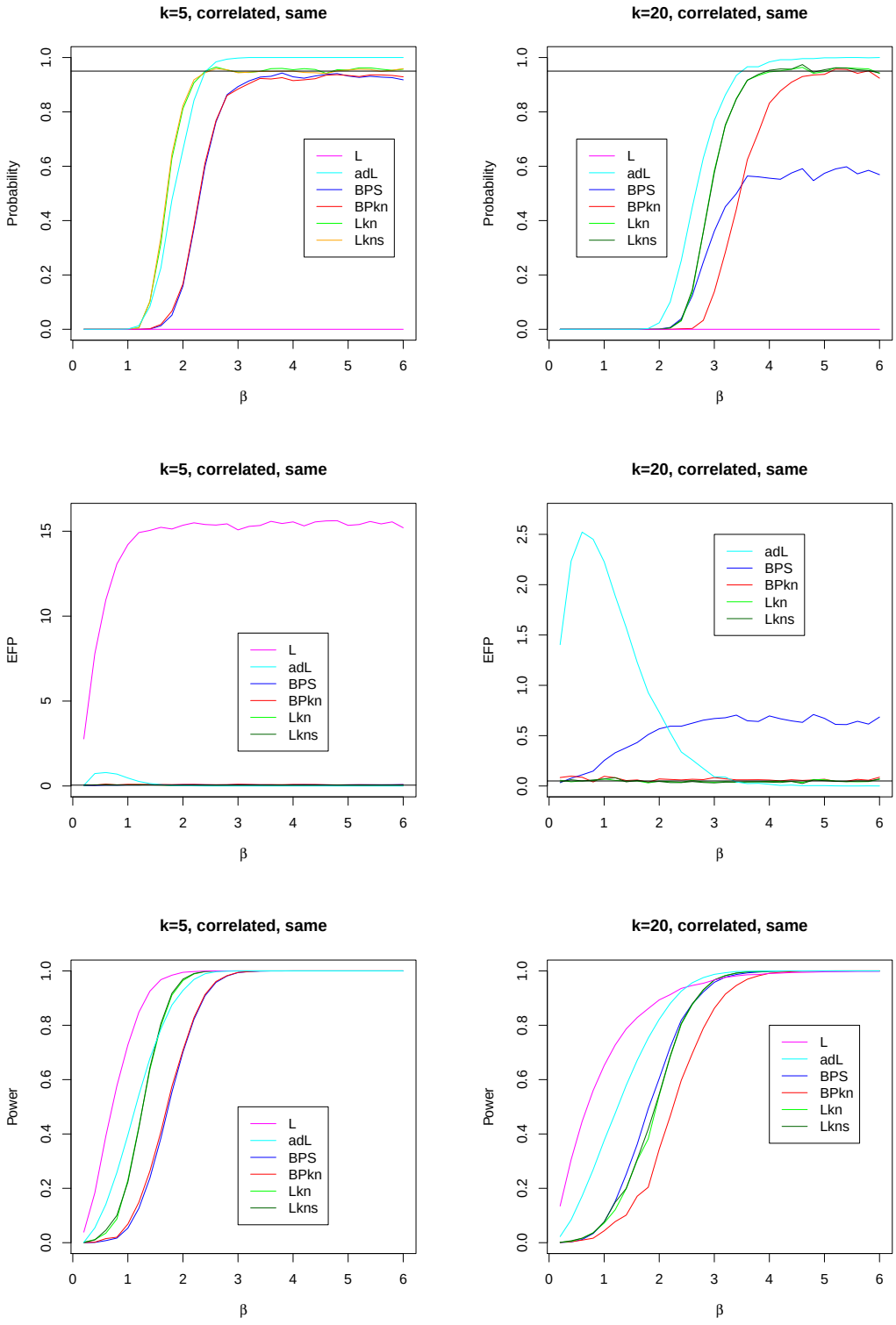

Figure 6: Design with strongly correlated columns and positive signals

5 Conclusion

This article main focus on theoretical properties of sign estimators derived from thresholded LASSO and thresholded basis pursuit. We have shown that the identifiability condition on β^* is the minimal assumption assuring the consistency of these both sign estimators. We have proved that the irrepresentable condition, well known for the consistency of the LASSO sign estimator, is stronger than the identifiability condition. The identifiability curve provides the maximal number of nonzero for β^* under which β^* is identifiable with respect to the l^1 norm.

The performances of sign estimators derived from thresholded LASSO and thresholded basis depend obviously from the threshold. In the numerical experiments, we have prescribed a threshold for the basis pursuit and the LASSO. Our simulations show that thresholded LASSO (resp. basis pursuit) sign estimators outperform adaptive LASSO and LASSO sign estimators.

6 Acknowledgments

We would like to thank E.J.Candès for helpful comments. The research of MB was funded by the NCN grant 2016/23/B/ST1/00454.

7 appendix

In the following subsection, we roughly explain the main arguments of theorems 1, 2 and 3.

7.1 Main ideas for the proofs of theorems

Theorem 1: Let S be the set of non null components of the LASSO $\hat{\beta}(\lambda)$, then the following inequalities hold

$$\begin{aligned} X_S'(Y - X\hat{\beta}(\lambda)) &= \lambda \text{sign}(\hat{\beta}_S(\lambda)), \\ \|X_{\overline{S}}'(Y - X\hat{\beta}(\lambda))\|_\infty &\leq \lambda. \end{aligned}$$

The proof of the theorem 1 is a consequence these above inequalities.

Theorem 2: The proof of theorem 2 shows that when the irrepresentable condition holds then the following inequality occurs

$$\forall h \in \ker(X), \left| \sum_{i \in \mathcal{A}} \text{sign}(\beta_i^*) h_i \right| \leq \sum_{i \notin \mathcal{A}} |h_i|.$$

This inequality implies that the l^1 norm of β^* is minimal. In addition, since X is in general position, β^* is the unique vector having a minimal l^1 norm which concludes the proof.

Theorem 3: The lemma 1 which claims that $\hat{\beta}_R^k(\varepsilon)/k$ converges to $\tilde{\beta}$ is the keystone to prove the theorem 3. This convergence is quite intuitive for the basis pursuit estimator. Indeed, when k is large then $Y \approx X(k\beta^*)$. Thus, intuitively, $\hat{\beta}_0^k$ solution of the problem: minimize $\|\beta\|_1$ subject to $X\beta = Y$ should be close to $k\tilde{\beta}$ solution of the problem: minimize $\|\beta\|_1$ subject to $X\beta = X(k\beta^*)$.

7.2 Proofs

In the proof given by Zou [38] for the small dimensional setting lot of cases are studied (case 1) $\lambda_n/n \to +\infty$, case 2) $\lambda_n/n \to \lambda_0$ and case 3) $\lambda_n/n \to 0$ and $\lambda_n/\sqrt{n} \to +\infty$). The proof given here is thus more straightforward than the one given by Zou and could be easily rewritten for the small dimensional setting.

Proof of the theorem 1: Let S be the set $S := \text{supp}(\hat{\beta}(\lambda))$, according to the lemma 14 given in Tibshirani [33], whatever $\lambda > 0$ the family $(X_i)_{i\in S}$ is linearly independent. Consequently, when $(X_i)_{i\in\mathcal{A}}$ is not linearly independent then it is straightforward that $S \neq \mathcal{A}$ implying thus $\text{sign}(\hat{\beta}(\lambda)) \neq \text{sign}(\beta^*)$.

Now, let us assume that the family $(X_i)_{i\in\mathcal{A}}$ is linearly independent. Let us give two expressions met by the LASSO estimator as defined in (2). The LASSO estimator $\hat{\beta}(\lambda)$ satisfies simultaneously the following two expressions

$$X_S'(Y - X\hat{\beta}(\lambda)) = \lambda \text{sign}(\hat{\beta}_S(\lambda)), \tag{10}$$

$$\|X_{\overline{S}}'(Y - X\hat{\beta}(\lambda))\|_\infty \leq \lambda. \tag{11}$$

These two expressions are given in Bühlmann and van de Geer [6] page 15 or in the proof of the theorem 1 of Zou [38]. In the first step, using the equality (10) and the inequality (11), we are going to show that if $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$ then the following event holds

$$\left\| X_{\overline{\mathcal{A}}}' X_{\mathcal{A}} (X_{\mathcal{A}}' X_{\mathcal{A}})^{-1} \text{sign}(\beta_{\mathcal{A}}^*) + \frac{1}{\lambda} X_{\overline{\mathcal{A}}}' \left(Id - X_{\mathcal{A}} (X_{\mathcal{A}}' X_{\mathcal{A}})^{-1} X_{\mathcal{A}}' \right) \varepsilon \right\|_\infty \leq 1$$

In a second step, to conclude the proof, we are going to show that the event given previously has a probability smaller than 1/2.

Let us assume that $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$ thus $S = \mathcal{A}$. Since $Y = X\beta^* + \varepsilon = X_{\mathcal{A}}\beta_{\mathcal{A}}^* + \varepsilon$ and $X\hat{\beta}(\lambda) = X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda)$ then the equality (10) and the inequality (11) lead to the following expressions

$$X_{\mathcal{A}}' \left(\varepsilon + X_{\mathcal{A}}(\beta_{\mathcal{A}}^* - \hat{\beta}_{\mathcal{A}}(\lambda)) \right) = \lambda \text{sign}(\beta_{\mathcal{A}}^*), \tag{12}$$

$$\left\| X_{\overline{\mathcal{A}}}' \left(\varepsilon + X_{\mathcal{A}}(\beta_{\mathcal{A}}^* - \hat{\beta}_{\mathcal{A}}(\lambda)) \right) \right\|_\infty \leq \lambda. \tag{13}$$

The equality (12) assures that

$$\beta^*_{\mathcal{A}} - \hat{\beta}_{\mathcal{A}}(\lambda) = (X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\left(\lambda \text{sign}(\beta^*_{\mathcal{A}}) - X'_{\mathcal{A}}\varepsilon\right).$$

Let us notice that the assumption $(X_i)_{i\in\mathcal{A}}$ is linearly independent assures that the Gram matrix $X'_{\mathcal{A}}X_{\mathcal{A}}$ is invertible. Using the previous expression in the inequality (13) gives

$$\begin{aligned} \left\|X'_{\overline{\mathcal{A}}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}(\lambda \text{sign}(\beta^*_{\mathcal{A}}) - X'_{\mathcal{A}}\varepsilon) + X'_{\overline{\mathcal{A}}}\varepsilon\right\|_\infty &\leq \lambda \\ \left\|X'_{\overline{\mathcal{A}}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\text{sign}(\beta^*_{\mathcal{A}}) + \frac{1}{\lambda}X'_{\overline{\mathcal{A}}}\left(Id - X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}X'_{\mathcal{A}}\right)\varepsilon\right\|_\infty &\leq 1 \end{aligned}$$

Let us denote ζ be the following Gaussian vector $\zeta \sim \mathcal{N}(u, \Gamma)$ where

$$u := X'_{\overline{\mathcal{A}}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\text{sign}(\beta^*_{\mathcal{A}}) \text{ and where } \Gamma \text{ is the covariance matrix of } \frac{1}{\lambda}X'_{\overline{\mathcal{A}}}\left(Id - X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}X'_{\mathcal{A}}\right)\varepsilon.$$

Because by assumption $\|u\|_\infty > 1$, there is an element $i_0 \in \{1, \ldots, p\}$ for which $|u_{i_0}| > 1$. To conlude the proof, one notices that

$$\mathbb{P}(\|\zeta\|_\infty \leq 1) \leq \mathbb{P}(|\zeta_{i_0}| \leq 1) \leq 1/2.$$

The last inequality occurs because $\zeta_{i_0} \sim \mathcal{N}(u_{i_0}, \Gamma_{i_0,i_0})$ with $|u_{i_0}| > 1$. □

Proof of the theorem 2: From Daubechies et al. [12], β^* is a parameter having a minimal l^1 norm, namely $X\beta^* = X\gamma \Rightarrow \|\gamma\|_1 \geq \|\beta^*\|_1$ holds, if and only if the following inequality occurs

$$\forall h \in \ker(X), \left|\sum_{i\in\mathcal{A}} \text{sign}(\beta^*_i)h_i\right| \leq \sum_{i\notin\mathcal{A}} |h_i|. \tag{14}$$

We are going to show that when the irrepresentable condition holds for β^* then the inequality (14) holds.

For all $h \in \ker(X)$, the following equality holds

$$\sum_{i\in\mathcal{A}} \text{sign}(\beta^*_i)h_i = h'_{\mathcal{A}}\text{sign}(\beta^*_{\mathcal{A}}) = h'_{\mathcal{A}}X'_{\mathcal{A}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\text{sign}(\beta^*_{\mathcal{A}}).$$

Because $\mathbf{0} = Xh = X_{\mathcal{A}}h_{\mathcal{A}} + X_{\overline{\mathcal{A}}}h_{\overline{\mathcal{A}}}$, one deduces the following inequalities

$$\begin{aligned} |h'_{\mathcal{A}}\text{sign}(\beta^*_{\mathcal{A}})| &= \left|h'_{\overline{\mathcal{A}}}X'_{\overline{\mathcal{A}}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\text{sign}(\beta^*_{\mathcal{A}})\right|, \\ &\leq \|h_{\overline{\mathcal{A}}}\|_1 \|X'_{\overline{\mathcal{A}}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\text{sign}(\beta^*_{\mathcal{A}})\|_\infty. \end{aligned} \tag{15}$$

Consequently, when the irrepresentable condition holds for β^* namely, when $\|X'_{\overline{\mathcal{A}}}X_{\mathcal{A}}(X'_{\mathcal{A}}X_{\mathcal{A}})^{-1}\text{sign}(\beta^*_{\mathcal{A}})\|_\infty \leq 1$

then, the inequality (15) gives $|h'_{\mathcal{A}}\text{sign}(\beta^*_{\mathcal{A}})| \leq \|h_{\overline{\mathcal{A}}}\|_1$. Thus, by the equivalence given in (14), β^* is a solution of the following basis pursuit problem

$$\text{minimize } \|\gamma\|_1 \text{ subject to } X\gamma = X\beta^*$$

Because X is in general position the previous optimisation problem has a unique solution (see *e.g.* the proposition 1 in appendix) thus $X\beta^* = X\gamma$ and $\gamma \neq \beta^*$ implies that $\|\gamma\|_1 > \|\beta^*\|_1$ namely β^* is identifiable with respect to the l^1 norm. □

Proof of the proposition 1: According to Daubechies et al. [12], β^* is identifiable with respect to the l^1 norm if and only if the following inequality holds

$$\forall h \in \ker(X) \setminus \{\mathbf{0}\}, \left| \sum_{i \in \text{supp}(\beta^*)} \text{sign}(\beta^*_i) h_i \right| < \sum_{i \notin \text{supp}(\beta^*)} |h_i|.$$

Because $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$, we have $\text{supp}(\tilde{\beta}) = \text{supp}(\beta^*)$ thus the following inequality holds

$$\forall h \in \ker(X) \setminus \{\mathbf{0}\}, \left| \sum_{i \in \text{supp}(\tilde{\beta})} \text{sign}(\tilde{\beta}_i) h_i \right| < \sum_{i \notin \text{supp}(\tilde{\beta})} |h_i|.$$

Consequently, the parameter $\tilde{\beta}$ is identifiable with respect to the l^1 norm. □

Proof of the theorem 3

The lemma 1 given hereafter is the keystone to prove the theorem 3. The proof of this lemma is partially inspired by the one given in Candès et al. [10]

Lemma 1 *Let X be a $n \times p$ matrix in general position such that* $\text{rank}(X) = n$. *Let $\tilde{\beta} \in \mathbb{R}^p$ be the unique solution of the problem:* minimize $\|\beta\|_1$ subject to $X\beta = X\beta^*$ *and let $\hat{\beta}^k_R(\varepsilon)$ be the following estimator*

$$\hat{\beta}^k_R(\varepsilon) := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|\beta\|_1 \text{ subject to } \|Y^k(\varepsilon) - X\beta\|_2^2 \leq R. \tag{16}$$

Then whatever $R \geq 0$, whatever $\varepsilon \in R^n$, the sequence $(\hat{\beta}^{\text{init,k}}_R(\varepsilon)/k)_{k \geq 1}$ converges to $\tilde{\beta}$.

Proof: Let us define $u(\varepsilon) \in \mathbb{R}^p$ as follows

$$u(\varepsilon) := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \|\beta\|_1 \text{ subject to } X\beta = \varepsilon.$$

Because $Y^k(\varepsilon) = X(k\tilde{\beta} + u(\varepsilon))$ and because $\hat{\beta}_R^k(\varepsilon)$ is an admissible point of (16) one deduces the following inequality

$$\left\| \frac{1}{k} X\hat{\beta}_R^k(\varepsilon) - X\tilde{\beta} \right\|_2 \leq \left\| \frac{1}{k} X\hat{\beta}_R^k(\varepsilon) - \frac{1}{k} Y^k(\varepsilon) \right\|_2 + \left\| \frac{1}{k} Y^k(\varepsilon) - X\tilde{\beta} \right\|_2 \leq \frac{\sqrt{R}}{k} + \frac{\|Xu(\varepsilon)\|_2}{k}. \tag{17}$$

Because $k\tilde{\beta} + u(\varepsilon)$ is an admissible point of the problem (16) and because $\hat{\beta}_R^k(\varepsilon)$ is the minimizer of (16), one deduces the following inequalities hold

$$\frac{1}{k}\|\hat{\beta}_R^k(\varepsilon)\|_1 \leq \frac{1}{k}\|k\tilde{\beta} + u(\varepsilon)\|_1 \leq \|\tilde{\beta}\|_1 + \frac{\|u(\varepsilon)\|_1}{k}. \tag{18}$$

Let us notice that the sequence $(\hat{\beta}_R^k(\varepsilon)/k)_{k\in\mathbb{N}^*}$ is bounded (by $\|\beta^*\|_1 + \|u(\varepsilon)\|_1$). Consequently, to prove the convergence of $(\hat{\beta}_R^k(\varepsilon)/k)_{k\in\mathbb{N}^*}$ it is sufficient to show that this sequence has a unique limit point. Let $(\hat{\beta}_R^{\phi(k)}(\varepsilon)/\phi(k))_{k\in\mathbb{N}^*}$ be a converging subsequence to l ($\phi : \mathbb{N}^* \to \mathbb{N}^*$ increasing). By (17) and (18) one deduces that

$$X\tilde{\beta} = Xl \text{ and } \|l\|_1 \leq \|\tilde{\beta}\|_1.$$

By construction of $\tilde{\beta}$ (as a unique solution of a basis pursuit problem), one deduces that $\tilde{\beta} = l$ thus $\tilde{\beta}$ is the unique limit point. Consequently, the following limit holds

$$\forall \varepsilon \in \varepsilon, \lim_{k\to+\infty} \frac{\hat{\beta}_R^k(\varepsilon)}{k} = \tilde{\beta}.$$

□

Proof of the theorem 3:

Separation property: Let us set $\eta_0 > 0$ such that $\eta_0 < \min\{|\tilde{\beta}_i|, i \in \mathcal{B}\}/2$. The convergence of $(\hat{\beta}_R^k(\varepsilon)/k)_{k\in\mathbb{N}^*}$ to $\tilde{\beta}$ implies that there exists $k_0 \in \mathbb{N}^*$ such that

$$\begin{aligned} &\forall k \geq k_0, \|\hat{\beta}_R^k(\varepsilon)/k - \tilde{\beta}\|_\infty \leq \eta_0, \\ &\forall k \geq k_0, \forall i \in \{1,\ldots,p\}, \left|[\hat{\beta}_R^k(\varepsilon)/k]_i - \tilde{\beta}_i\right| \leq \eta_0. \end{aligned}$$

Consequently, when $k \geq k_0$, whatever $i \notin \mathcal{B}$ (thus when $\tilde{\beta}_i = 0$) the following inequalities hold

$$\begin{aligned} &\forall i \notin \mathcal{B}, \left|[\hat{\beta}_R^k(\varepsilon)/k]_i\right| \leq \eta_0, \\ \Rightarrow \quad &-k\eta_0 \leq \min_{i\notin\mathcal{B}} \left\{[\hat{\beta}_R^k(\varepsilon)]_i\right\} \leq \max_{i\notin\mathcal{B}} \left\{[\hat{\beta}_R^k(\varepsilon)]_i\right\} \leq k\eta_0. \end{aligned}$$

Whatever $i \in \mathcal{B}^+$ (thus when $\tilde{\beta}_i > 0$) the following inequalities hold

$$\begin{aligned} & \forall i \in \mathcal{B}^+, [\hat{\beta}^k_R(\varepsilon)/k]_i \geq -\left|[\hat{\beta}^k_R(\varepsilon)/k]_i - \tilde{\beta}_i\right| + \tilde{\beta}_i \\ \Rightarrow \quad & \min_{\mathcal{B}^+}\left\{[\hat{\beta}^k_R(\varepsilon)]_i/k\right\} \geq -\eta_0 + \min\{|\tilde{\beta}_i|, i \in \mathcal{B}\} > \eta_0, \\ \Rightarrow \quad & -\min_{\mathcal{B}^+}\left\{[\hat{\beta}^k_R(\varepsilon)]_i\right\} > k\eta_0 \end{aligned}$$

Whatever $i \in \mathcal{B}^-$ (thus when $\tilde{\beta}_i < 0$) the following inequalities hold

$$\begin{aligned} & \forall i \in \mathcal{B}^-, [\hat{\beta}^k_R(\varepsilon)/k]_i \leq \left|[\hat{\beta}^k_R(\varepsilon)/k]_i - \tilde{\beta}_i\right| + \tilde{\beta}_i \\ \Rightarrow \quad & \min_{\mathcal{B}^-}\left\{[\hat{\beta}^k_R(\varepsilon)]_i/k\right\} \leq \eta_0 - \min\{|\tilde{\beta}_i|, i \in \mathcal{B}\} < -\eta_0, \\ \Rightarrow \quad & -\min_{\mathcal{B}^-}\left\{[\hat{\beta}^k_R(\varepsilon)]_i\right\} < -k\eta_0 \end{aligned}$$

Finally, when $k \geq k_0$ then

$$\max_{i \notin \mathcal{B}^-}\left\{[\hat{\beta}^k_R(\varepsilon)]_i\right\} < \min_{i \notin \mathcal{B}}\left\{[\hat{\beta}^k_R(\varepsilon)]_i\right\} \leq \max_{i \notin \mathcal{B}}\left\{[\hat{\beta}^k_R(\varepsilon)]_i\right\} < \max_{i \notin \mathcal{B}^+}\left\{[\hat{\beta}^k_R(\varepsilon)]_i\right\}.$$

Sign recovery: If β^* is identifiable with respect to the l^1 norm then $\beta^* = \tilde{\beta}$ and consequently, $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$. Reciprocally, let us assume that $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$. Because, by construction, $\tilde{\beta}$ is identifiable with respect to the l^1 norm and because $\text{sign}(\tilde{\beta}) = \text{sign}(\beta^*)$ then, according to the proposition 1, β^* is identifiable with respect to the l^1 norm. □

Supplementary material

We already said that when X is in general position the minimizer of the problem (7) is unique, we also stressed that the estimator derived by minimizing (7) when $R > 0$ is a LASSO. When the LASSO is written in usual way as in (2), a sketch of proof given in Tibshirani [33] shows the uniqueness of the LASSO estimator when X is in general position. In order to provide a self content article, we show that when X is in general position the minimizer of the problem (7) is unique when $R = 0$ as well as when $R > 0$. We already stressed that when β^* is identifiable with respect to the l^1 norm then β^* is sparse. We are going to show that when the identifiability holds for β^* then the family $(X_i)_{i \in \mathcal{A}}$ is linearly independent and thus the number of components of β^* equal to 0 is larger than $p - n$.

References

[1] Francis R Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

[2] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

[3] Mohsen Bayati and Andrea Montanari. The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[4] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[5] M. Bogdan, E. J. Candès, W. Su, and A. Weinstein. O the beaten path: ranking variables with cross-validated lasso. *Technical Report, University of Wroclaw*, 2018.

[6] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, 2011.

[7] T Tony Cai and Anru Zhang. Sharp rip bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1):74–93, 2013.

[8] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[9] Emmanuel J. Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016. To appear in Journal of the Royal Statistical Society Series B.

[10] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[11] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[12] Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Communications on pure and applied mathematics*, 63(1):1–38, 2010.

[13] Pascaline Descloux and Sylvain Sardy. Model selection with lasso-zero: adding straw to the haystack to better find needles. *arXiv preprint arXiv:1805.05133*, 2018.

[14] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Trans. R. Soc. A*, 367(1906):4273–4293, 2009.

[15] David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[16] David L Donoho and Jared Tanner. Precise undersampling theorems. *Proceedings of the IEEE*, 98(6):913–924, 2010.

[17] Charles Dossal. A necessary and sufficient condition for exact sparse recovery by 1 minimization. *Comptes Rendus Mathematique*, 350(1):117–120, 2012.

[18] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Springer, 2013.

[19] Niharika Gauraha, Tatyana Pavlenko, and Swapan K Parui. Post lasso stability selection for high dimensional linear models. In *6th International Conference on Pattern Recognition Applications and Methods (ICPRAM), FEB 24-26, 2017, Porto, Portugal*, pages 638–646. Scitepress, 2017.

[20] Caroline Giacobino, Sylvain Sardy, Jairo Diaz-Rodriguez, Nick Hengartner, et al. Quantile universal threshold. *Electronic Journal of Statistics*, 11(2):4701–4722, 2017.

[21] Rémi Gribonval and Morten Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[22] Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2:90–102, 2008.

[23] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[24] Nicolai Meinshausen and Bin Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270, 2009.

[25] Edouard Ollier and Vivian Viallon. Regression modelling on stratified data with the lasso. *Biometrika*, 104(1):83–96, 2017.

[26] Marie Perrot-Dockès, Céline Lévy-Leduc, Laure Sansonnet, and Julien Chiquet. Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *arXiv preprint arXiv:1707.04145*, 2017.

[27] Ulrike Schneider and Karl Ewald. On the distribution and model selection properties of the lasso estimator in low and high dimensions. *arXiv preprint arXiv:1708.09608*, 2017.

[28] Weijie Su. When does the first spurious variable get selected by sequential regression procedures? *arXiv preprint arXiv:1708.03046*, 2017.

[29] Weijie J Su, Małgorzata Bogdan, and Emmanuel J. Candès. False discoveries occur early on the lasso path. *The Annals of Statistics*, 45(5):2133–2150, 2017.

[30] Patrick Tardivel. *Représentation parcimonieuse et procédures de tests multiples: application à la métabolomique*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2017.

[31] Patrick JC Tardivel, Rémi Servien, and Didier Concordet. Sparsest representations and approximations of an underdetermined linear system. *Inverse Problems*, 34(5):055002, 2018.

[32] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[33] Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

[34] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

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

[36] Asaf Weinstein, Rina Barber, and Emmanuel J. Candès. A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*, 2017.

[37] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[38] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.