On the sign recovery by LASSO, thresholded LASSO and thresholded Basis Pursuit Denoising

Patrick J C Tardivel, Malgorzata Bogdan

▶ **To cite this version:**

Patrick J C Tardivel, Malgorzata Bogdan. On the sign recovery by LASSO, thresholded LASSO and thresholded Basis Pursuit Denoising. 2019. hal-01956603v4

HAL Id: hal-01956603

https://hal.science/hal-01956603v4

Preprint submitted on 26 Jun 2019 (v4), last revised 31 Aug 2021 (v7)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the sign recovery by LASSO, thresholded LASSO and thresholded Basis Pursuit Denoising

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

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

[b] Department of Statistics, Lund University, Lund, Sweden

Abstract

In the high-dimensional regression model $Y = X\beta + \varepsilon$, we provide new theoretical results on the probability of recovering the sign of β by the Least Absolute Selection and Shrinkage Operator (LASSO) and by the thresholded LASSO.

It is well known that "irrepresentability" is a necessary condition for LASSO to recover the sign of β with a large probability. In this article we extend this result by providing a tight upper bound for the probability of LASSO sign recovery. This upper bound is smaller than 1/2 when the irrepresentable condition does not hold and thus generalizes Theorem 2 of Wainwright [27]. The bound depends on the tuning parameter λ and is attained when non-null components of β tend to infinity; its value is equal to the limit of the probability that every null component of β is correctly estimated at 0. Consequently, this bound makes it possible to select λ so as to control the probability of at least one false discovery.

The "irrepresentability" is a stringent necessary condition to recover the sign of β by LASSO which can be substantially relaxed when LASSO estimates are additionally filtered out with an appropriately selected threshold. In this article we provide new theoretical results on thresholded LASSO and thresholded Basis Pursuit DeNoising (BPDN) in the asymptotic setup under which X is fixed and non-null components of β tend to infinity. Compared to the classical asymptotics, where X is a $n \times p$ matrix and both n and p tend to $+\infty$, our approach allows for reduction of the technical burden. Our main Theorem takes a simple form: **When non-null components of β are sufficiently large, appropriately thresholded LASSO or thresholded BPDN can recover the sign of β if and only if β is *identifiable* with respect to the L_1 norm, i.e.**

$$\textbf{If} \quad X\gamma = X\beta \textbf{ and } \gamma \neq \beta \textbf{ then } \|\gamma\|_1 > \|\beta\|_1.$$

To illustrate our results we present examples of *irrepresentability* and *identifiability* curves for some selected design matrices X. These curves provide the proportion of k sparse vectors β for which the *irrep-*

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

resentability and *identifiability* conditions hold. Our examples illustrate that "irrepresentability" is a much stronger condition than "identifiability", especially when the entries in each row of X are strongly correlated. Finally, we illustrate how the knockoff methodology [1, 8] can be used to select an appropriate threshold and that thresholded BPDN and LASSO can recover the sign of β with a larger probability than adaptive LASSO [32].

Keywords: Multiple regression, Basis Pursuit, LASSO, Sparsity, Active set estimation, Sign estimation, Identifiability condition, Irrepresentability condition

1 Introduction

Let us consider the high-dimensional linear model

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

where $X = (X_1|\dots|X_p)$ is a $n \times p$ design matrix with $n \leq p$, ε is a random vector in $\mathbb{R}^n$, and $\beta \in \mathbb{R}^p$ is an unknown vector of regression coefficients. The sign vector of β is $S(\beta) = (S(\beta_1), \dots S(\beta_p)) \in \{-1, 0, 1\}^p$, where for $x \in \mathbb{R}$, $S(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$. Our main purpose is to recover $S(\beta)$. This objective is slightly more general than the aim of recovering the active set, $\text{supp}(\beta) := \{i \in \{1, \dots, p\} \mid \beta_i \neq 0\}$.

The sign of β can be estimated by the sign of the well known LASSO estimator [25]:

$$\widehat{\beta}^{\text{L}} := \underset{b \in \mathbb{R}^p}{\text{argmin}} \ \frac{1}{2}\|Y - Xb\|_2^2 + \lambda\|b\|_1. \tag{2}$$

When $\text{rank}(X) = n$, an alternative formulation of LASSO is provided by the Basis Pursuit DeNoising (BPDN) estimator [9]:

$$\widehat{\beta}^{\text{BPDN}} := \underset{b \in \mathbb{R}^p}{\text{argmin}} \ \|b\|_1 \text{ subject to } \|Y - Xb\|_2^2 \leq R. \tag{3}$$

Given a particular vector $Y \in \mathbb{R}^n$, there is a one-to-one correspondance between the tuning parameter $\lambda > 0$ and the regularization parameter $R > 0$, under which LASSO and BPDN estimates take the same value (see *e.g* page 64 of [17] or the chapter 5.3 of [3]). For example, when $\lambda = \|X'Y\|_\infty$ and when $R = \|Y\|_2^2$ then both LASSO and BPDN estimators are equal to $\mathbf{0}$. However, the relationship between λ and R depends on the specific realization of Y and, in broad generality, given a fixed $\lambda > 0$ for LASSO, we cannot pick a fixed $R > 0$ for BPDN under which these both estimators equal. Thus, BPDN and LASSO are not equivalent estimators. The Basis Pursuit (BP) estimator, solution of (3) when $R = 0$, is a particular case of BPDN. As discussed e.g. in [11, 15], BP can be thought of as the limit of LASSO when the tuning parameter λ tends to 0.

1.1 Sign recovery by LASSO

Properties of the LASSO sign estimator $S(\widehat{\beta}^{\mathrm{L}}(\lambda)) := \left(S(\widehat{\beta}^{\mathrm{L}}_1(\lambda)), \ldots, S(\widehat{\beta}^{\mathrm{L}}_p(\lambda))\right)$ (or properties of the active set estimator $\mathrm{supp}(\widehat{\beta}^{\mathrm{L}}(\lambda)) := \{i \in \{1, \ldots, p\} \mid \widehat{\beta}_i(\lambda) \neq 0\}$) have been intensively studied [16, 21, 27, 31, 32]. Specifically, Zhao and Yu [31] and Zou [32] consider the asymptotic setup under which n tends to $+\infty$ and p is fixed and observe that LASSO can recover $S(\beta)$ only if the restrictive irrepresentable condition is fulfilled. These results were further extended to the case of the fixed design matrix X, where the irrepresentable condition is formulated as follows:

Definition 1 (Irrepresentability condition) *Let $b \in \mathbb{R}^p$, $I := \{i \in \{1, \ldots, p\} \mid b_i \neq 0\}$, and $X_I, X_{\overline{I}}$ be the matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$. Vector b satisfies the irrepresentable condition if* $\ker(X_I) = \mathbf{0}$ *and* $\|X'_{\overline{I}} X_I (X'_I X_I)^{-1} S(b_I)\|_\infty \leq 1$.

According to the Theorem 2 of Wainwright [27], the irrepresentability condition is necessary to recover $S(\beta)$ with high probability. Indeed, when $\ker(X_I) = \mathbf{0}$, $\|X'_{\overline{I}} X_I (X'_I X_I)^{-1} S(\beta_I)\|_\infty > 1$ and both ε and $-\varepsilon$ have the same distribution, then for any selection of the tuning parameter $\lambda > 0$, $\mathbb{P}(S(\widehat{\beta}^{\mathrm{L}}(\lambda)) = S(\beta)) \leq 1/2$. This result holds also in the noiseless case (i.e.when $\varepsilon = \mathbf{0}$), where the probability to recover $S(\beta)$ is equal to zero. Moreover, Bühlmann and van de Geer [5] (page 192-194) showed that, when $\varepsilon = \mathbf{0}$ and the irrepresentability strictly holds (*i.e* when $\|X'_{\overline{I}} X_I (X'_I X_I)^{-1} S(\beta_I)\|_\infty < 1$) then the non-random set $\mathrm{supp}(\beta^L(\lambda))$ recovers $\mathrm{supp}(\beta)$ as soon as non-null components of β are sufficiently large. The proof provided in [5] can be easily adapted for the sign recovery.

In this article we provide a new theoretical result on the sign recovery by LASSO. Specifically, Theorem 1 in Section 2 provides an upper bound for the probability to recover the sign of β which depends on $X, S(\beta), \lambda$ and the distribution of ε. The formula for the bound is not analytic but its value can be well approximated by simple Monte Carlo simulations. We also show that the bound is attained when non-null components of β tend to infinity and its value is the limit of the probability that all null components of β are correctly estimated at 0. Therefore, the bound can be also used to calculate the asymptotic probability that at least one null component of β is not estimated at 0 (the Family Wise Error Rate, FWER). Moreover, as shown in our simulation study, in many examples FWER increases with the magnitude of non-zero elements of β. Consequently, in such cases the bound can be used to select λ so that FWER is controlled independently of the magnitude of the non-null components of β.

1.2 Sign recovery by thresholded LASSO

It is clear that in the noiseless case, the following identifiability condition is necessary and sufficient to recover $S(\beta)$ by the non-random basis pursuit.

Definition 2 (Identifiability condition) *Vector $b \in \mathbb{R}^p$ is identifiable with respect to the design matrix X*

and the L_1 norm (or just identifiable with respect to the L_1 norm) if the following implication holds

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

Proposition 1, proved in the Appendix, shows that the identifiability condition is weaker than the irrepresentability condition.

Proposition 1 *Let X be a $n \times p$ matrix with $p \geq n$ columns in general position. Moreover, let $\beta \in \mathbb{R}^p$, $I := \text{supp}(\beta)$ and assume that $\ker(X_I) = \mathbf{0}$. If the irrepresentability condition holds then the parameter β is identifiable with respect to the L_1 norm.*

Under the identifiability assumption, β is sparse. Indeed, Lemma 3 in Tardivel et al. [24] shows that $k = \text{card}\{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\} \leq n$, i.e. β has at least $p - n$ zeros. On the other hand some assumptions on the sparsity of β guaranty that β is identifiable with respect to the L_1 norm. For example when $\|X_1\|_2 = \cdots = \|X_p\|_2 = 1$ and the number of nonzero elements of β satisfies the following inequality (called mutual coherence condition)

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

then β is identifiable with respect to the L_1 norm [13, 17, 19]. When entries of X are i.i.d $\mathcal{N}(0, 1)$ and n, p are both very large, the phase transition curve of Donoho and Tanner [14] provides, with respect to the ratio n/p, an upper bound on k/n so that β having a sparsity k is identifiable with respect to the L_1 norm.

Theorem 2 in Section 3 highlights the importance of the identifiability condition for the sign recovery by thresholded LASSO and thresholded BPDN. It states that for any value of the tuning parameter λ or the regularization parameter R, the identifiability condition is sufficient and necessary so that LASSO and BPDN estimators appropriately separate negative, null, and positive components of β, if only non-null components of β are sufficiently large. This means that, when non-null components of β are sufficiently large, appropriately thresholded LASSO or BPDN can properly identify the sign of β if and only if the identifiability condition holds for β.

1.3 Graphical illustrations of main results

By definition, the irrepresentability condition depends only on $S(\beta)$ and not on how large the non-null components of β are. Moreover, as claimed in Proposition 2 in Section 4, the identifiability condition also depends only on $S(\beta)$. Thus, the comparison of these two conditions can be performed by considering vectors of parameters such that $\beta = S(\beta)$. In Figure 1, we provide the irrepresentability and the identifiability curves for a selected matrix X of dimensions 100×300, whose elements were independently drawn from the normal $\mathcal{N}(0, 1)$ distribution. These curves provide the proportion of the sign vectors with k nonzero elements which satisfy the identifiability condition or the irrepresentability condition. Figure 1 illustrates that the identifiability curve is

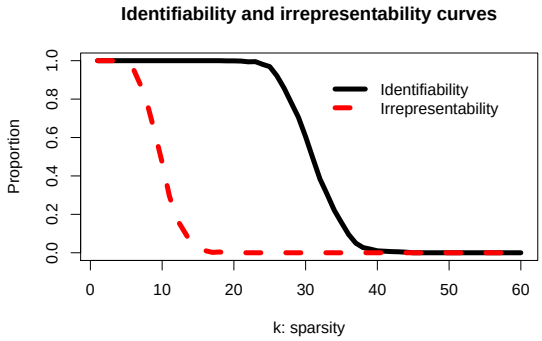

Figure 1: This figure provides the identifiability and irrepresentability curves for the design matrix X of dimensions 100×300, whose entries were independently generated from $\mathcal{N}(0,1)$ distribution. The x-axis represents the sparsity k and the y-axis represents the proportion of sign vectors satisfying the identifiability condition (resp. irrepresentability condition)

substantially shifted to the right with respect to the irrepresentability curve. It can be observed that LASSO can not recover the sign of β when the sparsity k is larger than 10, while thresholded LASSO allows to recover the sign of β when $k \leq 25$ and the non-null components of β are sufficiently large.

Figure 2 illustrates Theorem 1 from Section 2, which provides an upper bound for the probability of LASSO sign recovery. In this Figure we use the same design matrix X as in Figure 1 and the noise ε is a standard Gaussian vector. According to the irrepresentability curve provided in Figure 1, the irrepresentability condition holds when $k \leq 5$. Thus, when $k \leq 5$, one can select the tuning parameter λ in order to obtain any fixed value for this bound. In our experiment the value of λ was selected so that the average value of the bound over 1000 randomly sampled vectors β with $k = 5$ non-zero elements is equal to 0.95. The y axis in Figure 2 represents the probability of recovering $S(\beta)$ by LASSO with the selected λ when β has $k = 5$ non-null components, which are all equal to t. Figure 2 shows that the upper bound for LASSO sign recovery is reached when non-null components of β tend to $+\infty$ and that the selected λ makes it possible to control the FWER below 0.05 for the whole range of the magnitudes of β.

Finally, Figure 3 illustrates Theorem 2 from Section 3 which states that when non-null components of β are large enough, "identifiability" is a sufficient condition under which appropriately thresholded BPDN and thresholded LASSO can recover $S(\beta)$.

In this figure we present results for $k = 20$, for which the irrepresentability condition does not hold (see Figure 1). Consequently, the probability to recover $S(\beta)$ is theoretically smaller than $1/2$ and more precisely, the empirical value of this probability is almost 0. On the other hand, due to a fact that for $k = 20$ the identifiability condition holds, we expect that thresholded LASSO and thresholded BP can recover $S(\beta)$ when the magnitude of β is large enough. In Figure 3 the y axis represents the probability of recovering $S(\beta)$ by thresholded LASSO and thresholded BP, calculated based on 1000 randomly sampled vectors β having $k = 20$ non-null components, which are all equal to t. For both of these procedures, in order to pick a threshold,

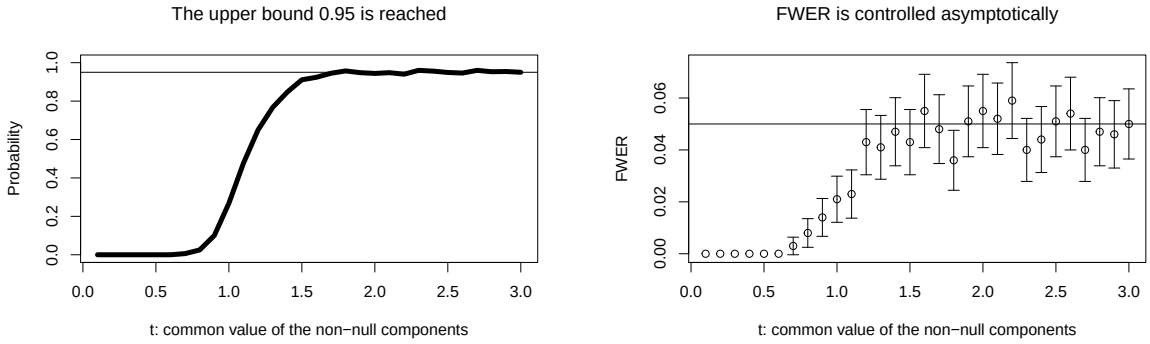

Figure 2: When $k = 5$, the left figure provides the probability to recover $S(\beta)$ with LASSO sign estimator and the right figure provides the FWER: the probability that at least one null component of β is selected by the LASSO estimator. The x-axis represents the value of non-null components of β (in both figures), the y-axis represents the probability of the sign recovery (left figure) and the FWER (right figure). The horizontal lines correspond to $y = 0.95$ (left figure) and $y = 0.05$ (right figure).

we approximate the distribution of LASSO (resp. BP) estimators associated to null components of β using control variables created according to the knockoffs methodology [8] (see Section 5 for details). In our case, a control variable is just a column added to the design matrix and generated according to a standard multivariate Gaussian distribution. The threshold was selected so as to control the FWER at level 0.05. Figure 3 shows that, indeed, both thresholded BP and thresholded LASSO can recover $S(\beta)$ with probability converging to 0.95 when non-null components of β increase in magnitude.

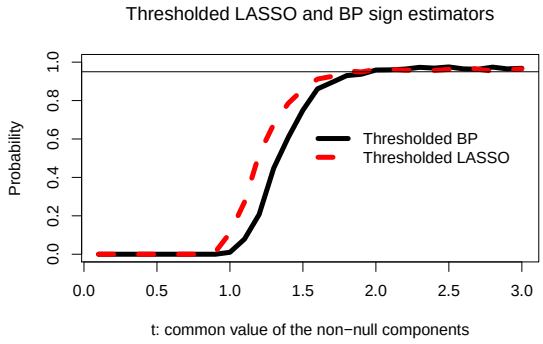

Figure 3: When $k = 20$, this figure provides the probability to recover $S(\beta)$ with thresholded LASSO and thresholded BP. The x-axis represents the value of non-null components of β and the y-axis represents the sign recovery probability.

1.4 Organization of the article

In Section 2 we formulate and discuss Theorem 1, which provides a tight upper bound for LASSO sign estimator to recover $S(\beta)$. In Section 3, Theorem 2 shows that identifiability is a necessary and suficient condition for LASSO to separate the non-null components of β from the noise and to recover asymptotically $S(\beta)$ with thresholded LASSO and thresholded BPDN. In Section 4, Proposition 2 shows that identifiability condition

depends only on $S(\beta)$ and not on the magnitude of non-null components of β. Here we also introduce the irrepresentability and identifiability curves which provide respectively the proportion of sign vectors satisfying the irrepresentability condition and identifiability condition. Section 5 is devoted to numerical experiments which illustrate that sign estimators derived from the thresholded LASSO and thresholded BPDN can be better than sign estimators derived from LASSO and adaptive LASSO and that knockoff methodology allows for the appropriate selection of the threshold for both of these methods.

1.5 Notations and assumptions

In this article we always assume that design matrix X has columns in general position (see *e.g* [26] or the supplementary material for this manuscript). This assumption guarantees that the minimizer of (2) (resp. minimizer of (3)) is unique and thus that the LASSO estimator (resp. BPDN estimator) is well-defined. This assumption is very weak and generically holds. Indeed, when X is a random matrix such that the entries $(X_{11}, X_{12}, \dots, X_{np})$ have a density on $\mathbb{R}^{np}$ then, almost surely, X is in general position [26].

The main notation used in the subsequent sections is as follows:

- Let I be the subset of $\{1, \dots, p\}$. We denote by $\overline{I}$ the complement of I, namely $\overline{I} := \{1, \dots, p\} \setminus I$.
- The notation X_I represents a matrix whose columns are indexed by the elements of I: $(X_i)_{i \in I}$.
- For $b \in \mathbb{R}^p$, b_I denotes the sub-vector containing elements of b with indices in I.
- Symbols $\mathrm{supp}(b), \mathrm{supp}^+(b)$ and $\mathrm{supp}^-(b)$ denote respectively the sets $\{i \in \{1, \dots, p\} \mid b_i \neq 0\}, \{i \in \{1, \dots, p\} \mid b_i > 0\}$ and $\{i \in \{1, \dots, p\} \mid b_i < 0\}$.
- LASSO and BPDN estimators depend on X, β, ε and on the tuning parameter $\lambda > 0$ or the regularization parameter $R \geq 0$. When it is useful, we use the parentheses to recall these dependencies. The estimator $\widehat{\beta}$ represents indistinctly the LASSO estimator or the BPDN estimator.

To formulate our asymptotic results we will often consider a sequence of regression parameters $\beta^{(r)}$, $r \in \mathbb{N}$, for which non-null components tend to infinity in the following way.

Assumption 1

1) *The sign of $\beta^{(r)}$ is invariant namely, there exists a sign vector $s^0 \in \{-1, 0, 1\}^p$ such that for any $r \in \mathbb{N}, S(\beta^{(r)}) = s^0$.*

2) *The following limit holds* $\lim_{r \to +\infty} \min\{|\beta_i^{(r)}|, i \in \mathrm{supp}(s^0)\} = +\infty$

3) *There exists $q > 0$ such that*

$$\forall r \in \mathbb{N}, \frac{\min\{|\beta_i^{(r)}|, i \in \mathrm{supp}(s^0)\}}{\|\beta^{(r)}\|_\infty} \geq q.$$

2 Sign recovery with LASSO sign estimator

In this section we formulate Theorem 1, which provides an upper bound for the probability to recover $S(\beta)$ with LASSO estimator. When β is identifiable with respect to the L_1 norm, this upper bound is reached asymptotically when $\min\{|\beta_i|, i \in \mathrm{supp}(\beta)\}$ tends to $+\infty$.

Theorem 1 *Let $I := \mathrm{supp}(\beta)$ and let $X_I, X_{\overline{I}}$ be matrices whose columns are $(X_i)_{i\in I}$ and $(X_i)_{i\notin I}$, respectively. Let us assume that $\ker(X_I) = \mathbf{0}$ and let $\zeta_{X,\lambda,S(\beta)} := X'_{\overline{I}} X_I (X'_I X_I)^{-1} S(\beta_I) + \frac{1}{\lambda} X'_{\overline{I}} \left(Id - X_I (X'_I X_I)^{-1} X'_I\right) \varepsilon$.*

Upper bound: *The following upper bound for the sign recovery holds.*

$$\mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda)) = S(\beta)\right) \leq \mathbb{P}\left(\left\|\zeta_{X,\lambda,S(\beta)}\right\|_\infty \leq 1\right).$$

Now, let $(\beta^{(r)})$ be a sequence in $\mathbb{R}^p$ satisfying Assumption 1. If s^0 is identifiable with respect to the L_1 norm then the following asymptotic results hold.

Sharpness of the upper bound: *Asymptotically, the upper bound is reached.*

$$\limsup_{r\to+\infty} \mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = s^0\right) \leq \mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty \leq 1\right),$$
$$\liminf_{r\to+\infty} \mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = s^0\right) \geq \mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1\right).$$

Asymptotic control of FWER: *Let us set $\mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1\right) = \gamma$ and $\mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty \leq 1\right) = \bar{\gamma}$. The sign of nonzero elements of $(\beta^{(r)})$ is properly identified with probability converging to 1 and the FWER is controlled at level $1-\gamma$.*

$$\lim_{r\to+\infty} \mathbb{P}\left(\forall i \in I, S(\widehat{\beta}^{\mathrm{L}}_i(\lambda, r)) = s^0_i\right) = 1,$$
$$\limsup_{r\to+\infty} \mathbb{P}\left(\exists i \notin I, \widehat{\beta}^{\mathrm{L}}_i(\lambda, r) \neq 0\right) \leq 1-\gamma,$$
$$\liminf_{r\to+\infty} \mathbb{P}\left(\exists i \notin I, \widehat{\beta}^{\mathrm{L}}_i(\lambda, r) \neq 0\right) \geq 1-\bar{\gamma}.$$

Remark 1 *Results given in Theorem 1 are quite straightforward when X is orthogonal (i.e. when $X'X = I$). Indeed, in this case the upper bound is just the probability that null components of β are simultaneously estimated at 0 namely $\mathbb{P}(\forall i \notin \mathrm{supp}(\beta), \hat{\beta}^{\mathrm{L}}_i(\lambda) = 0)$.*

Remark 2 *Theorem 1 immediately implies Theorem 2 of Wainwright [27] which claims that when ε and $-\varepsilon$ have the same distribution and when $\|X'_{\overline{I}} X_I (X'_I X_I)^{-1} S(\beta_I)\|_\infty > 1$ then, for any $\lambda > 0$, the probability of the sign recovery by LASSO is always smaller than $1/2$.*

The bound for the probability of recovering $S(\beta)$ provided in Theorem 1 is not analytic but can be computed using simple Monte Carlo simulations. When the irrepresentability condition strictly holds for s^0, namely when $\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} s_I^0\|_\infty < 1$, the tuning parameter λ can be selected to fix γ at an arbitrary level in $(0,1)$ (e.g. see Figure 1). Because the irrepresentability condition implies the identifiability condition (as claimed in Proposition 1) such a tuning parameter allows for an asymptotic control of the FWER at level $1-\gamma$ when non-null components tends to $+\infty$. To our knowledge, Theorem 1 is the first theoretical result providing a guide on how to select the tuning parameter in order to control a type I error at a specified level for a given design matrix X.

3 Identifiability is a necessary and sufficient condition for the sign recovery

When β does not satisfy the irrepresentability condition then the LASSO sign estimator $S(\widehat{\beta}^{\mathrm{L}}(\lambda))$ fails to recover $S(\beta)$ with large probability. However, the irrepresentability condition is not an unsurpassable limitation to recover $S(\beta)$. Actually, the following Theorem 2 shows that an appropriately thresholded LASSO (resp. thresholded BPDN) can recover $S(\beta)$ if only the non-zero elements of β are sufficiently large and the identifiability condition holds.

Theorem 2 *Let X be a $n \times p$ matrix with columns in general position and such that $\mathrm{rank}(X) = n$. Moreover, let $\beta^{(r)}$ be a sequence in $\mathbb{R}^p$ satisfying Assumption 1 and let $\widehat{\beta}(\varepsilon, r)$ be the LASSO or BPDN estimator with an arbitrary fixed value of the tuning parameter $\lambda > 0$ or with an arbitrary fixed regularization parameter $R \geq 0$. If s^0 is identifiable with respect to the L_1 norm then for any fixed $\varepsilon \in \mathbb{R}^n$ and sufficiently large $r > r_0(\varepsilon)$ the estimator $\widehat{\beta}(\varepsilon, r)$ separates negative components of $\beta^{(r)}$ (i.e $i \in \mathrm{supp}^-(\beta^{(r)})$), null components of $\beta^{(r)}$ (i.e $i \notin \mathrm{supp}(\beta^{(r)})$) and positive components of $\beta^{(r)}$ (i.e $i \in \mathrm{supp}^+(\beta^{(r)})$):*

i)

$$\mathrm{supp}^-(\beta^{(r)}) \subset \mathrm{supp}^-(\widehat{\beta}_i(\varepsilon, r)) \text{ and } \mathrm{supp}^+(\beta^{(r)}) \subset \mathrm{supp}^+(\widehat{\beta}_i(\varepsilon, r)).$$

ii)

$$\max_{i \in \mathrm{supp}^-(\beta^{(r)})} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} < \min_{i \notin \mathrm{supp}(\beta^{(r)})} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} \leq \max_{i \notin \mathrm{supp}(\beta^{(r)})} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} < \min_{i \in \mathrm{supp}^+(\beta^{(r)})} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\}.$$

If s^0 is not identifiable with respect to the L_1 norm then for any $r \geq 0$ the sign estimator derived from thresholded LASSO or thresholded BPDN cannot recover $S(\beta^{(r)})$;

$$\forall r \in \mathbb{N}, \mathrm{supp}^-(\beta^{(r)}) \not\subset \mathrm{supp}^-(\widehat{\beta}_i(\varepsilon, r)) \text{ or } \mathrm{supp}^+(\beta^{(r)}) \not\subset \mathrm{supp}^+(\widehat{\beta}_i(\varepsilon, r).$$

Let us notice that the assumptions on X are very weak and generically hold when $n \leq p$. The assumption that $\text{rank}(X) = n$ assures that, for any $R \geq 0$, the BPDN estimator is well defined. The general position condition assures the uniqueness of both LASSO and BPDN estimators (see *e.g* Proposition 1 in the supplementary material).

Theorem 2 stresses that one cannot recover $S(\beta)$ with a sign estimator derived from LASSO or BPDN when β is not identifiable with respect to the L_1 norm. When β is identifiable with respect to the L_1 norm, Theorem 2 suggests that $S(\beta)$ can be recovered by deriving sign estimators from the thresholded LASSO or thresholded BPDN. In Section 5 we show how the appropriate thresholds can be obtained with help from control variables constructed according to the knockoff methodology (see e.g. [1, 8]).

Theorem 2 confirms recent results of [29, 4], which describe the asymptotic properties of the thresholded LASSO estimator when the elements of the design matrix X are independent random variables from the Gaussian distribution. Indeed, if X has i.i.d $\mathcal{N}(0,1)$ entries, $n/p \to \delta \in (0,1)$ and if asymptotically the point $(\text{card}(\text{supp}(\beta))/n, n/p)$ is below the asymptotic phase transition curve of Donoho and Tanner [12] (*i.e.* if β is asymptotically identifiable with respect to the L_1 norm) then the thresholded LASSO almost surely recovers $S(\beta)$ (as soon as non-null components of β are large enough).

4 Identifiability and irrepresentability curves

By definition the irrepresentability condition depends only on the sign of β. Given a particular design matrix X, the irrepresentability sign indicator is defined hereafter.

Irrepresentability sign indicator:

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

The irrepresentability indicator indicates if the LASSO sign estimator can recover $S(\beta)$. Indeed, if $\phi_{\text{IC}}^X(S(\beta)) = 0$ then $S(\beta)$ cannot be recovered with the LASSO sign estimator even if non-null components of β are extremely large. The following Proposition 2 shows that the identifiability condition also depends only on $S(\beta)$ and not on the magnitudes of the non-null components of β.

Proposition 2 *Consider two vectors $b \in \mathbb{R}^p$ and $\tilde{b} \in \mathbb{R}^p$ such that $S(b) = S(\tilde{b})$ then $\tilde{b}$ is identifiable with respect to the matrix X and L_1 norm if and only if b is identifiable with respect to the matrix X and L_1 norm.*

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

Identifiability sign indicator:

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

Such an identifiability indicator indicates if the sign estimators obtained by thresholded LASSO and thresholded BPDN can recover $S(\beta)$. Indeed, if $\phi^X_{\text{Idtf}}(S(\beta)) = 0$ then thresholded LASSO (resp. thresholded BPDN) sign estimator cannot recovered $S(\beta)$ even if non-null components of β are extremely large.

According to Proposition 2 in the supplementary material, when columns $(X_i)_{i\in\text{supp}(\beta)}$ are not linearly independent then β does not satisfy the identifiability condition. Consequently, when $\text{card}(\text{supp}(\beta)) > n$ then $\phi^X_{\text{IC}}(S(\beta)) = \phi^X_{\text{Idtf}}(S(\beta)) = 0$. Let us provide some basic properties and comments about the two indicator functions.

1. Both ϕ^X_{IC} and ϕ^X_{Idtf} are even.
2. Due to Proposition 1, for every $s \in \{-1,0,1\}^p$, $\Phi^X_{\text{IC}}(s) \leq \Phi^X_{\text{Idtf}}(s)$.
3. The computation of Φ^X_{IC} requires only the straightforward matricial calculus; the computation of Φ^X_{Idtf} only requires only solving a basic Basis Pursuit problem.

The last remark shows that given a parameter $\beta \in \mathbb{R}^p$, it is easy to check if β is identifiable with respect to the L_1 norm.

4.1 Illustrations of identifiability and irrepresentability curves

The number of sign vectors is very huge (3^p) and therefore we can not provide explicitly Φ^X_{Idtf} and Φ^X_{IC} for each sign vector. Instead, we define the identifiability and irrepresentability curves as the following functions of the sparsity k of the vector β, $k = ||\beta||_0 \in \{1,\ldots,n\}$,

- Identifiability curve is defined as $p^X_{\text{Idtf}}(k) := \mathbb{E}_U(\Phi^X_{\text{Idtf}}(U))$,
- Irrepresentability curve is defined as $p^X_{\text{IC}}(k) := \mathbb{E}_U(\Phi^X_{IC}(U))$,

where U is uniformly distributed on $\{u \in \{-1,0,1\}^p \mid \text{card}(\text{supp}(u)) = k\}$. Additionally, in case when the design matrix X has positively correlated columns, we will also consider a situation when U is uniformly distributed on $\{u \in \{0,1\}^p \mid \text{card}(\text{supp}(u)) = k\}$. More specifically we will consider three following settings:

Setting 1: Matrix X is a fixed $n \times p$ matrix with $n = 100$, $p = 300$, whose elements were generated by independent draws from the standard normal distribution $\mathcal{N}(0,1)$. The distribution of the sign vectors is uniform on $\{u \in \{-1,0,1\}^p \mid \text{card}(\text{supp}(u)) = k\}$.

Setting 2: Matrix X is a fixed design matrix with $n = 100$, $p = 300$, whose rows were generated by independent draws from the multivariate normal distribution $\mathcal{N}(\mathbf{0}, \Gamma)$, with $\Gamma_{ii} = 1$ for $i \in \{1, \ldots, p\}$ and $\Gamma_{ij} = 0.9$ when $i \neq j$. The distribution of the sign vectors is uniform on $\{u \in \{-1, 0, 1\}^p \mid \text{card}(\text{supp}(u)) = k\}$.

Setting 2 with positive components: The matrix X is the same as in Setting 2 but the distribution of the sign vectors is uniform on $\{u \in \{0, 1\}^p \mid \text{card}(\text{supp}(u)) = k\}$.

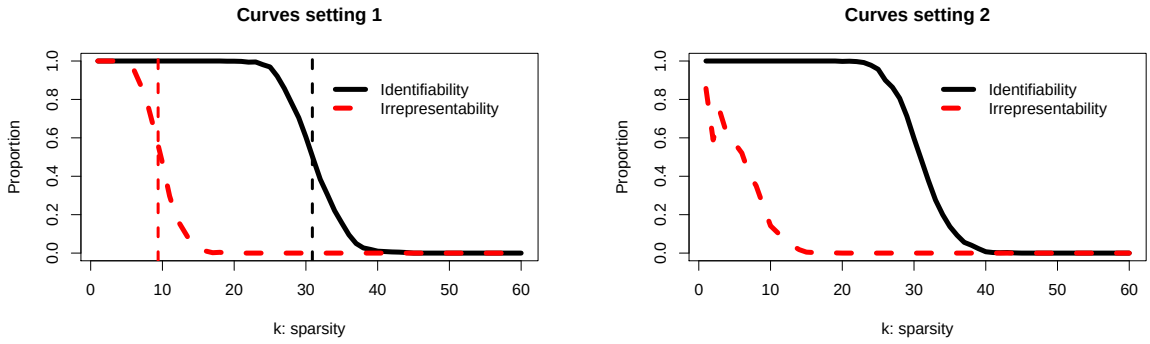

Figure 4: Graphs of the functions $k \mapsto p^X_{\text{Idtf}}(k)$ and $k \mapsto p^X_{\text{IC}}(k)$ for Setting 1 (left panel) and Setting 2 (right panel), calculated based on 1000 random draws of the sign vectors from the assumed symmetric uniform distribution. In the left panel the vertical lines represent values $k = \frac{n}{2 \log p} \approx 0.09n$ and $k = 0.31n$, which correspond to the asymptotic upper limits for k so the irrepresentability and identifiability conditions holds for Gaussian design matrices with independent entries (see [27] and [14]).

Surprisingly, the two identifiability curves obtained for Setting 1 and Setting 2 are very similar. *A priori*, we expected the identifiability curve in Setting 2 to be substantially below the one obtained for Setting 1. This is because the classical conditions known to imply the identifiability of β, like the mutual coherence condition (5) or the restricted isometry property [6, 7], become very stringent when columns of X are strongly correlated. Thus, Figure 4 illustrates that these classical conditions are in fact much stronger than the identifiability condition, particularly with respect to the correlation between columns of the design matrix. For Gaussian random design matrices with independent $\mathcal{N}(0, 1)$ entries, the limit of sparsities for the sign recovery by LASSO and thresholded LASSO can be predicted based on the asymptotic results provided in [27] and [14]. Specifically, Corollary 2 of [27] states that when both n and p tend to $+\infty$ and $n = \nu p$ for some $\nu \in (0, 1)$, LASSO can recover only vectors with support $k \leq (1 + o(1)) \frac{\nu p}{2 \log p}$. In the same asymptotic regime, the asymptotic phase transition curve provided in Donoho and Tanner [14] gives the upper limit at $\epsilon = \frac{k}{n}$, such that the identifiability condition is satisfied. When applied to our selected design matrix X, these asymptotic results predict that the identifiability and irrepresentability conditions should hold respectively when $k \leq 0.31n$ and $k \leq 0.09n$. Figure 4 illustrates very good accuracy of these predictions, despite relatively small dimensions of the design matrix X. Figure 5 illustrates an interesting shape of irrepresentability and identifiability curves in Setting 2 with positive components. Indeed, we can observe that under this scenario the irrepresentability condition becomes much more stringent than in case when the distribution of the elements of the sign vector is symmetric. Interestingly,

the identifiability condition becomes much weaker now, and is satisfied under a substantially larger range of sparsity levels as compared to Setting 2.

The behavior of the irrepresentability curve under Setting 2 with positive components also explains the lack of monotonicity of the irrepresentability curve in Setting 2, which occurs for very small values of k. This is because when $k = 2$ both components of the sign vector are positive or negative with probability of 0.5. Thus, for such a small k, the irrepresentability curve bears some similarity with the one given in Figure 5. Consequently, the probability of the sign recovery for $k = 2$ is much smaller than for $k = 1$. In case of $k = 3$ the probability that all three elements of the sign vector have the same sign is only 0.25 and the probability of the sign recovery increases when compared with the case of $k = 2$.

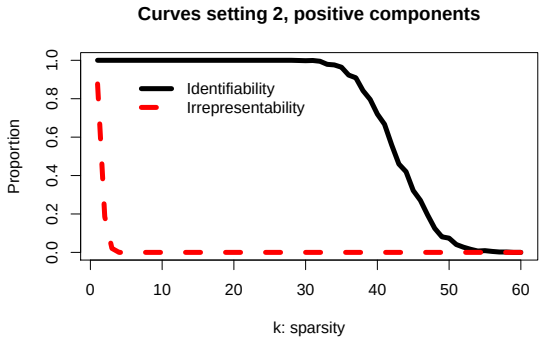

Figure 5: Graphs of the functions $k \mapsto p^X_{\mathrm{Idtf}+}(k)$ and $k \mapsto p^X_{\mathrm{IC}+}(k)$ in Setting 2 with positive components.

5 Numerical comparisons of sign estimators

Theorem 2 states that the sign estimators provided by thresholded LASSO or thresholded BPDN allow to recover $\mathrm{S}(\beta)$ as long as the identifiability condition is satisfied. Another way to recover $\mathrm{S}(\beta)$ is to use a sign estimator provided by adaptive LASSO, proposed in [32]. Indeed, as claimed in [32] or [20], if the weights for adaptive LASSO are based on sufficiently good estimator of β one can obtain a sign estimator which is consistent for $\mathrm{S}(\beta)$ under much weaker assumptions than the irrepresentability condition. The purpose of this section is to provide a numerical comparison of sign estimators derived from LASSO, thresholded LASSO, thresholded BP and adaptive LASSO.

5.1 Selection of the tuning parameter

As explained in [4, 28], a value of the optimal tuning parameter for the sign recovery by thresholded LASSO is substantially smaller than the optimal value of the tuning parameter for LASSO sign estimator. Specifically:

- For LASSO sign estimator, the tuning parameter has to be large enough so that it prevents the inclusion of false discoveries.

- For thresholded LASSO sign estimator the tuning parameter needs to be selected so as to minimize the mean square error of the estimation of β. This tuning parameter does not need to be large, since the threshold will allow to correctly estimate at 0 null-components of β.

5.1.1 Tuning parameter for LASSO sign estimator

When the sign $S(\beta)$ satisfies the irrepresentability condition, then by Theorem 1 one may select a tuning parameter λ_L so that for sufficiently large β, $\mathbb{P}(S(\hat{\beta}(\lambda_L)) = S(\beta))$ is arbitrarily close to any given value (say 0.95). According to the irrepresentability curve associated with the matrix X, applied in Setting 1 in Section 4.1, the irrepresentability condition is satisfied with probability close to 1 when β contains $k = 5$ nonzero elements. Thus, in this setting, we can chose λ_L such that the average value of the upper-bound given in Theorem 1 is equal to 0.95. In other words, λ_L is chosen so that $\mathbb{E}_S(\zeta_{X,\lambda_L,S}) = 0.95$, where S is a random sign vector having a uniform distribution over the set $\{s \in \{-1, 0, 1\}^p \mid \text{card}(\text{supp}(s)) = 5\}$. The computation of this value gives $\lambda_L = 81.18$. Since under the remaining scenarios of our simulation study the irrepresentability condition is typically not satisfied and thus the FWER can not be controlled at a low level, we decided to use the same value $\lambda_L = 81.18$ for all our simulations.

5.1.2 Tuning parameter for thresholded LASSO sign estimator

When X is the gaussian matrix with independent entries the tuning parameter can be selected with the help of the asymptotic theory of Approximate Message Passing (AMP) algorithm for LASSO, provided e.g. in [2, 4, 22]. In the set-up of this theory the elements of the design matrix are i.i.d. Gaussian $\mathcal{N}(0, 1/\sqrt{n})$ variables and components of β are i.i.d random variables having $\Pi = (1 - \gamma)\delta_0 + \gamma\Pi^\star$ mixture distribution, where δ_0 is a point mass distribution concentrated at 0 and $\Pi^\star$ is an arbitrary fixed distribution. The asymptotic characteristics of LASSO, like the asymptotic mean square error, are derived under the assumption that the number of observations n and the number of explanatory variables p tend to infinity and $n/p \to \delta > 0$. Then, the "optimal" value of the tuning parameter λ_{AMP} can be selected so that the asymptotic mean square error is minimal (see e.g. prescription in [4, 28]). As discussed in [4, 28], for any fixed value of the type I error such a tuning parameter allows to maximize the asymptotic power of the thresholded LASSO. In our simulation study we calculated this asymptotic optimal λ_{AMP} using parameter values $\delta = n/p = 100/300$, $\gamma = k/p = k/300$ and $\Pi^\star = 1/2\delta_t + 1/2\delta_{-t}$, where δ_t is a point mass distribution at t. Additionally, we observed that in case when the columns of the design matrix are strongly correlated substantially better results can be obtained by using smaller values of λ. Therefore in our simulation study we additionally use $\lambda_s = 0.5\lambda_{AMP}$.

5.2 Selection of the threshold

We define the thresholded LASSO sign estimator (resp. thresholded BP estimator) as

$$\forall i \in \{1, \dots, p\}, \widehat{\beta}_i^{\tau} := \widehat{\beta}_i \mathbf{1}_{\left\{|\widehat{\beta}_i| > \tau\right\}} . \tag{6}$$

Now, given a threshold $\tau > 0$, we define FWER as

$$\mathrm{FWER}(\tau) := \mathbb{P}\left(\exists i \notin \mathrm{supp}(\beta), \left|\widehat{\beta}_i^{\tau}\right| \neq 0\right).$$

By taking τ_α as the $1-\alpha$ quantile of the distribution of $\max\left\{|\widehat{\beta}_i|, i \in \mathrm{supp}(\beta)\right\}$ we would control FWER exactly at the level α. However, τ_α cannot be obtained by a straightforward computation since β is not known.

In order to provide a threshold larger than τ_α (and thus to control the FWER at level α), it seems appealing to look at the distribution of the supremum norm of the LASSO estimator (resp. BP estimator) in the full null model when $\beta = \mathbf{0}$ [18]. For the BP estimator, Descloux and Sardy [11] suggest the threshold $\tau_\alpha^{\mathrm{fn}}$ defined as the $1-\alpha$ quantile of $\max\left\{\left|\widehat{\beta}_1^{\mathrm{fn}}\right|, \dots, \left|\widehat{\beta}_p^{\mathrm{fn}}\right|\right\}$ where $\widehat{\beta}^{\mathrm{fn}}$ is the following estimator

$$\widehat{\beta}^{\mathrm{fn}} := \ \mathrm{argmin}\ \|\beta\|_1 \text{ subject to } X\beta = \varepsilon, \text{ where } \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I).$$

Unfortunately, when vector β contains some nonzero elements this intuitive method provides a threshold $\tau_\alpha^{\mathrm{fn}}$ which is smaller than τ_α and thus $\mathrm{FWER}(\tau_\alpha^{\mathrm{fn}}) > \alpha$ (see also Su et al. [22] for additional explanations).

The recently developed knockoff methodology [1, 8] allows to control the False Discovery Rate (FDR) This control is achieved by supplementing the design matrix with additional control variables. Originally developed to control FDR, control variables also allow to approximate the distribution of estimators corresponding to null components of β. In this numerical study, we informally use model free knockoffs proposed in [8] to approximate a threshold which controls the FWER at a given level. The approach developed hereafter is suitable for the situation when X is a Gaussian matrix having a distribution invariant to columns' permutation. In this setting, we can generate the knockoff variables individually, instead of generating the full knockoff matrix of $n \times p$ dimensions, as suggested in [8] (see Weinstein et al. [30] for a similar approach). Because adding the controlled variables can change some relevant properties (such as the identifiability condition for β), ideally we should add just one knockoff variable at a time when calculating LASSO estimates. This however would lead to a heavy computational burden of the procedure to estimate the relevant threshold. Therefore, in our simulation study we use model free knockoffs [8, 30] to generate $30 = p/10$ of controlled 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 controlled variables is saved. This step is repeated 10 times and the overall maximum of the $p = 300$ absolute values of regression coefficients over controlled variables is calculated. The whole procedure is

repeated many times (here 1000) and 0.95 quantile of the obtained maxima is used as the threshold to identify null-components of β.

To confirm with the set-up of simulations used to derive the irrepresentability and identifiability curves, in all of 1000 replicates we used the same fixed design matrix X described in settings 1 and 2 of the subsection 4.1, while the locations of k sparse signals and the error terms were randomly generated for each of these replicates. The calculations were performed separately for each value of k and t (magnitude of non-zero elements of β) used in the simulation study.

5.2.1 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 $\lambda_L = 81.18$.
- For the adaptive LASSO the weights are derived using initial estimates $\widehat{\beta}^L(\lambda_{AMP})$, where the tuning parameter is selected according to AMP theory, described above. For $i \in \{1, \ldots, p\}$, weights $w(\beta_i)$ are defined as $w(\beta_i) := 1/(|\widehat{\beta}_i^{\,L}(\lambda_{AMP})| + 10^{-7})$. Using these weights and the tuning parameter λ_L described above, the adaptive LASSO has the following expression

$$\widehat{\beta}^{\text{adapt}} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda_L \sum_{i=1}^{p} w(\beta_i)|\beta_i|. \tag{7}$$

In all our simulations LASSO is calculated with *glmnet*.

5.3 Numerical comparisons

The rows of the design matrix X are sampled as the independent vectors from the multivariate Gaussian distribution, as in settings 1 and 2. All numerical experiments are performed with a particular observation of X (the same as the one used in the previous section). We set $\beta \in \mathbb{R}^p$ such that $k := \text{card}(\text{supp}(\beta))$ where $k = \{5, 20\}$ and $\text{supp}(\beta)$ is a k sample without replacement of $\{1, \ldots, p\}$. The non-null components of β have a uniform distribution $\{-t, t\}$ where $t > 0$. Additionally in setting 2 we consider the set-up where all non-zero 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.

- The sign estimator **L** is derived from LASSO with $\lambda = \lambda_L$.
- The sign estimator **aL** is derived from the adaptive LASSO estimator, described in (7).
- The sign estimator **BP** is derived from the thresholded BP, with threshold selected as in [11].

- The sign estimator **BPk** is derived from the thresholded BP, with a threshold given by the "knockoff" methodology described above.
- The sign estimator **Lk** is derived from the thresholded LASSO with $\lambda = \lambda_{AMP}$ and with a threshold given by the "knockoff" methodology described above.
- The sign estimator **Lks** is derived from the thresholded LASSO with $\lambda = 0.5\lambda_{AMP}$ and with a threshold given by the "knockoff" methodology described above.

In order to recover the sign of β, null components of β have to be estimated simultaneously at zero. This naive remark motivated us to report the curves illustrating the following statistical properties as the function of $t > 0$:

- **FWER** is the proportion of 1000 replicates that at least one null components of β is not estimated at zero.

We report the curve illustrating the probability to recover the sign as the function of $t > 0$:

- **Probability** is the proportion of 1000 replicates for which the sign is recovered.

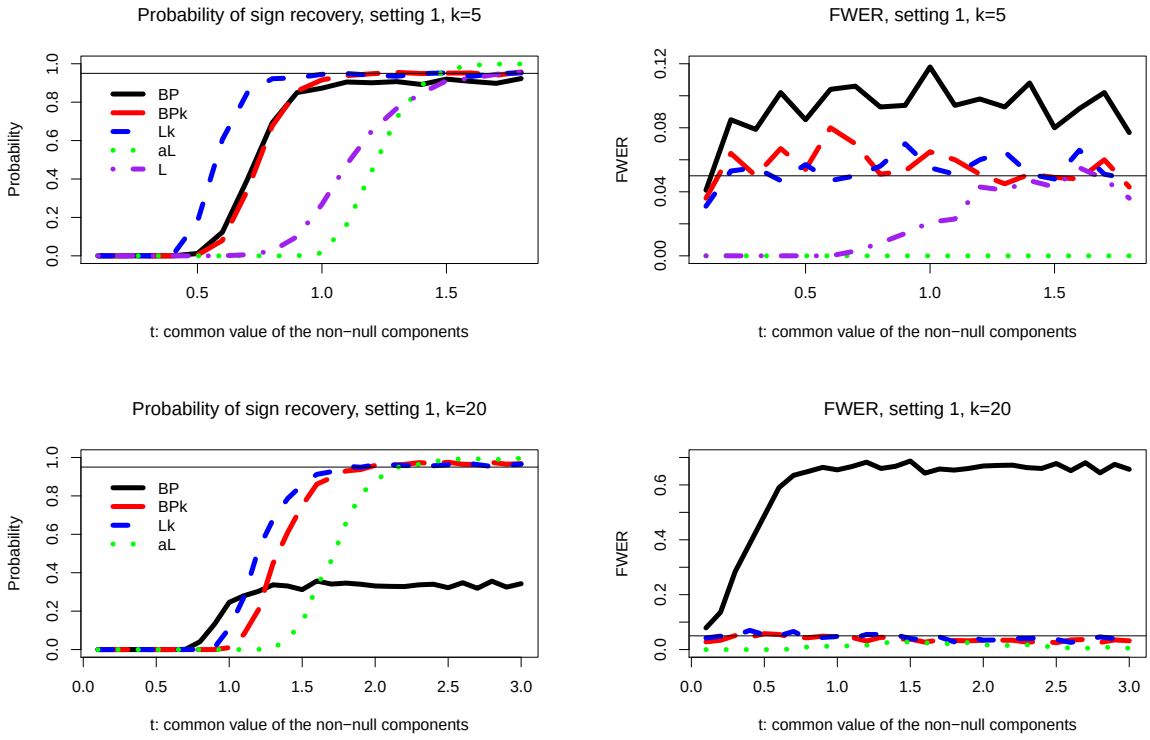

Figure 6: This figure provides the FWER and the probability to recover $S(\beta)$ for each sign estimators and when X is the design matrix given in setting 1. Graphics on the left provide the probability to recover $S(\beta)$ (on the y-axis) as a function of t, where t measures how large the non-null components of β are. Graphics on the right provide the FWER (on the y-axis) as a function of t (on the x-axis). Among these sign estimators, one may notice that the thresholded LASSO sign estimator is the one which recovers $S(\beta)$ with the largest probability. These sign estimators recover approximately $S(\beta)$ with a probability close to 0.95 when t is large.

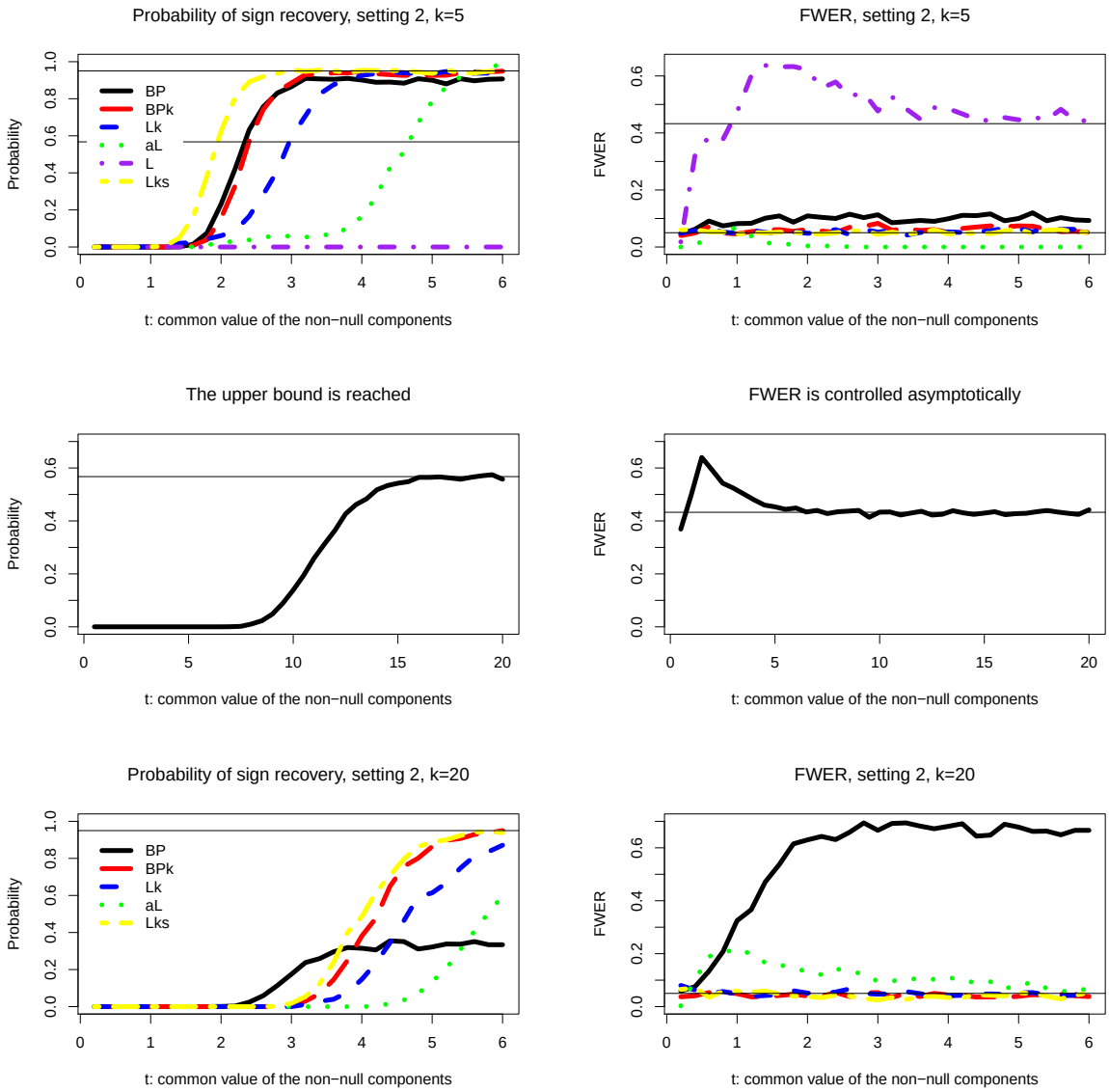

Figure 7: This figure provides the FWER and the probability to recover $S(\beta)$ for each sign estimators and when X is the design matrix given in setting 2. Graphics on the left provide the probability to recover $S(\beta)$ (on the y-axis) as a function of t (on the x-axis), where t measures how large the non-null components of β are. Graphics on the right provide the FWER (on the y-axis) as a function of t. The horizontal lines $y = 0.55$ and $y = 0.45$ represent respectively the average values of the upper bound for the probability of sign recovery and FWER associated with LASSO (see Theorem 1). One may notice that the upper-bound is approximately reached and the FWER is approximately controlled when t is very large as illustrated by graphics in the middle. Sign estimators (except LASSO sign estimator) recover approximately $S(\beta)$ with a probability close to 0.95 when t is large.

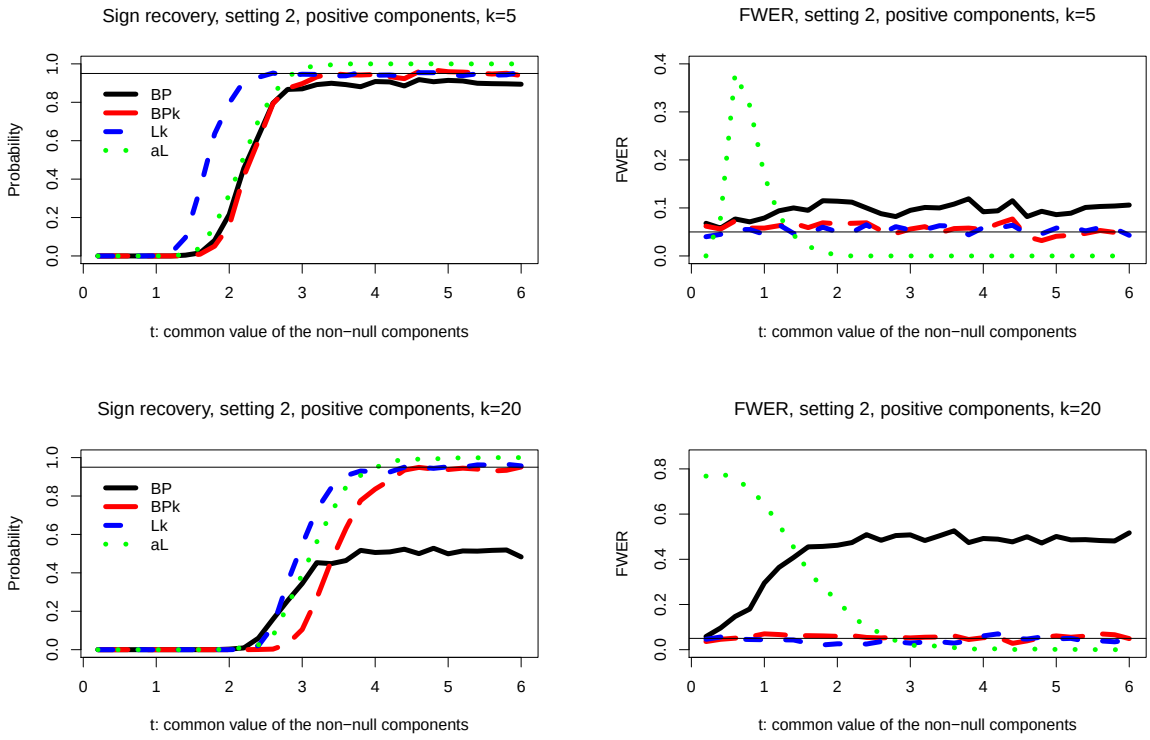

Figure 8: This figure provides the FWER and the probability to recover $S(\beta)$ for each sign estimator when X is the design matrix given in setting 2 and non-null components of β are positive. Graphics on the left provide the probability to recover $S(\beta)$ (on the y-axis) as a function of t (on the x-axis), where t measures how large the non-null components of β are. Graphics on the right provide the FWER (on the y-axis) as a function of t. These sign estimators recover approximately $S(\beta)$ with a probability close to 0.95 when t is large.

Figures 6-8 illustrate that the upper bound for the probability of sign recovery by LASSO is reached and the FWER is controlled when non-null component of β are large (*i.e* when t is large). On the other hand, thresholded LASSO and thresholded BP can appropriately identify $S(\beta)$ when the identifiability condition holds. Indeed, when $k \in \{5, 20\}$, as illustrated in Figures 4 and 5, the identifiability condition occurs and thus sign estimators derived from thresholded LASSO and thresholded BP recover $S(\beta)$ as soon as the threshold is well calibrated and the non-null components are large enough. In our simulated set-up, thresholded BP performs pretty well but is never optimal. Indeed using an appropriate tuning parameter λ, the probability to recover $S(\beta)$ is larger with thresholded LASSO than with thresholded BP. When entries of X are i.i.d $\mathcal{N}(0, 1)$, the optimal value of λ selected by AMP theory provides a thresholded LASSO for which the derived sign estimator is the best one to recover $S(\beta)$. One may notice that the threshold selection provided in Descloux and Sardy [11] does not allow to recover $S(\beta)$ with a large probability when β has lot of large components (intuitively when β is far from $\mathbf{0}$). Instead, our heuristic application of the knockoff methodology allows for almost perfect control of FWER at level 0.05. Consequently, when non-null components of β are large enough and when the threshold is given by knockoff methodology, sign estimator derived from thresholded LASSO (resp. thresholded BP) recovers $S(\beta)$ with a probability close to 0.95.

6 Conclusion

This article's main focus was on theoretical properties of sign estimators derived from LASSO, thresholded LASSO and thresholded BPDN. We provided an upper bound for LASSO sign recovery which is reached when non-null components of β are infinitely large and the identifiability condition holds. In addition, when the irrepresentable condition occurs (implying that the identifiability condition occurs), we have shown that λ can be selected appropriately in order to control asymptotically the FWER at an arbitrary level.

When $S(\beta)$ is identifiable with respect to the L_1 norm and when non-null components of β are infinitely large, we have shown that sign estimators derived from thresholded LASSO and thresholded BPDN recover $S(\beta)$. On the other hand, if $S(\beta)$ is not identifiable with respect to the L_1 norm, sign estimators derived from thresholded LASSO and thresholded BPDN cannot recover $S(\beta)$.

We have introduced identifiability curve (resp. irrepresentability curve) which is useful to know for which sparsity β is identifiable with respect to the L_1 norm (resp. for which sparsity β the irrepresentable condition holds).

The performances of sign estimators derived from LASSO, thresholded LASSO and thresholded BPDN depend obviously on the tuning parameter, the regularization parameter and the threshold. We have illustrated that AMP theory and knockoff methodology are useful to select these parameters. Our simulations show that thresholded LASSO and thresholded BPDN sign estimators outperform adaptive LASSO and LASSO sign estimators.

Acknowledgments

We would like to thank Emmanuel J. Candès and Wojciech Rejchel for helpful comments. The research of Małgorzata Bogdan was funded by the NCN grant 2016/23/B/ST1/00454. We gratefully acknowledge the grant of the Wrocław Center of Networking and Supercomputing (WCSS), where most of the computations were performed.

7 appendix

7.1 Proof of the Theorem 1

First, let us provide lemmas which are useful to prove both Theorems 1 and 2. Lemma 2 partially proves Theorem 1. Indeed, according to this Lemma, when $(\beta^{(r)})_{r\in\mathbb{N}}$ is a sequence of $\mathbb{R}^p$ satisfying assumptions 1 then the following asymptotic result holds

$$\lim_{r\to+\infty} \mathbb{P}\left(\forall i \in \mathrm{supp}(s^0), S(\widehat{\beta}_i^{\mathrm{L}}(\lambda, r)) = s_i^0\right) = 1.$$

Lemma 1 *Let $(\beta^{(r)})_{r\in\mathbb{N}}$ be a sequence of $\mathbb{R}^p$ satisfying the conditions* **1)** *and* **2)** *of Assumption 1, let us assume that s^0 is identifiable with respect to the L_1 norm and let us set $u_r = \|\beta^{(r)}\|_1$ then*

$$\lim_{r\to+\infty} \frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r) - \beta^{(r)}}{u_r} = 0.$$

Proof: Because $\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)$ is the LASSO estimator as defined in (2) then the following inequality occurs

$$\frac{1}{2}\|Y - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2^2 + \lambda\|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_1 \leq \frac{1}{2}\|Y - X\beta^{(r)}\|_2^2 + \lambda\|\beta^{(r)}\|_1.$$

Since $Y - X\beta^{(r)} = \varepsilon$ one may deduce the following inequalities

$$\begin{aligned} & \lambda\|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\|\beta^{(r)}\|_1, \\ \Rightarrow \quad & \|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)/u_r\|_1 \leq \frac{\|\varepsilon\|_2^2}{2\lambda u_r} + 1. \end{aligned} \tag{8}$$

In addition, Cauchy-Schwarz inequality gives the following implications

$$\begin{aligned} & \frac{1}{2}\|\varepsilon + X\beta^{(r)} - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2^2 + \lambda\|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\|\beta^{(r)}\|_1, \\ \Rightarrow \quad & -\|\varepsilon\|_2\|X\beta^{(r)} - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2 + \frac{1}{2}\|X\beta^{(r)} - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2^2 + \lambda\|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_1 \leq \lambda\|\beta^{(r)}\|_1, \\ \Rightarrow \quad & -\frac{\|\varepsilon\|_2}{u_r}\left\|X\left(\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r) - \beta^{(r)}}{u_r}\right)\right\|_2 + \frac{1}{2}\left\|X\left(\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r) - \beta^{(r)}}{u_r}\right)\right\|_2^2 + \frac{\lambda}{u_r}\left\|\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)}{u_r}\right\|_1 \leq \frac{\lambda}{u_r}. \end{aligned} \tag{9}$$

Because u_r tends to $+\infty$ then, according to (8), the sequence $((\widehat{\beta}^{\mathrm{L}}(\varepsilon,r)-\beta^{(r)})/u_r)_{r\in\mathbb{N}^*}$ is bounded since the following superior limit is finite

$$\limsup_{r\to+\infty}\left\|\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon,r)-\beta^{(r)}}{u_r}\right\|_1\leq 2.$$

Consequently, to prove that $\lim_{r\to+\infty}(\widehat{\beta}^{\mathrm{L}}(\varepsilon,r)-\beta^{(r)})/u_r=\mathbf{0}$ it is sufficient to show that $\mathbf{0}$ is the unique limit point of this sequence. Let $((\widehat{\beta}^{\mathrm{L}}(\varepsilon,\phi(r))-\beta^{(\phi(r))})/u_{\phi(r)})_{r\in\mathbb{N}^*}$ be a converging subsequence to l (with $\phi:\mathbb{N}^*\to\mathbb{N}^*$ strictly increasing) and without loss of generality, let us assume $\lim_{r\to+\infty}\widehat{\beta}^{\mathrm{L}}(\varepsilon,\phi(r))/u_{\phi(r)}=v$ and $\lim_{r\to+\infty}\beta^{(\phi(r))}/u_{\phi(r)}=v'$ so that $l=v-v'$. By (8) and (9) one may deduce that

$$Xv=Xv' \text{ and } \|v\|_1\leq 1.$$

Since, whatever $r\geq 0$, we have $S(\beta^{(\phi(r))}/u_{\phi(r)})=s^0$ where s^0 is identifiable with respect to the L_1 norm then, according to Proposition 2, one may deduce that $\beta^{(\phi(r))}/u_{\phi(r)}$ is an unitary vector satisfying the identifiability condition. Consequently, $\|v'\|_1=1$ and v' is identifiable with respect to the L_1 norm. Consequently, $v=v'$ and thus $l=\mathbf{0}$ is the unique limit point, which implies that

$$\lim_{r\to+\infty}\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon,r)-\beta^{(r)}}{u_r}=\mathbf{0}.$$

□

For the proof of Lemma 1, we have not used the third condition of Assumption 1. This condition, under which the smallest non-null component of $\beta^{(r)}$ is not asymptotically infinitely smaller than $\|\beta^{(r)}\|_\infty$, is useful to prove Lemma 2.

Lemma 2 *Let $(\beta^{(r)})_{r\in\mathbb{N}}$ be a sequence of $\mathbb{R}^p$ satisfying Assumption 1, then*

$$\lim_{r\to+\infty}\mathbb{P}(\forall i\in\mathrm{supp}(s^0),S(\widehat{\beta}_i^{\mathrm{L}}(\lambda,r))=s_i^0)=1.$$

Proof: Let ε be a fixed vector in $\mathbb{R}^p$. According to the third condition of Assumption 1 we have $\min\{|\beta_i^{(r)}|,i\in\mathrm{supp}(s^0)\}/\|\beta^{(r)}\|_\infty\geq q>0$, consequently the following inequalities occur

$$\forall i\in\mathrm{supp}(s^0),s_i^0\frac{\widehat{\beta}_i^{\mathrm{L}}(\varepsilon,\lambda,r)-\beta_i^{(r)}}{\|\beta^{(r)}\|_\infty}=\frac{s_i^0\widehat{\beta}_i^{\mathrm{L}}(\varepsilon,\lambda,r)}{\|\beta^{(r)}\|_\infty}-\frac{|\beta_i^{(r)}|}{\|\beta^{(r)}\|_\infty}\leq\frac{s_i^0\widehat{\beta}_i^{\mathrm{L}}(\varepsilon,\lambda,r)}{\|\beta^{(r)}\|_\infty}-q.$$

According to Lemma 1, the following inequality occurs

$$0=\liminf_{r\to+\infty}s_i^0\frac{\widehat{\beta}_i^{\mathrm{L}}(\varepsilon,\lambda,r)-\beta_i^{(r)}}{\|\beta^{(r)}\|_\infty}\leq\liminf_{r\to+\infty}\frac{s_i^0\widehat{\beta}_i^{\mathrm{L}}(\varepsilon,\lambda,r)}{\|\beta^{(r)}\|_\infty}-q.$$

Which implies that for r large enough $s_i^0 \widehat{\beta}_i^{\mathrm{L}}(\varepsilon, \lambda, r) > 0$ and thus $S(\widehat{\beta}_i^{\mathrm{L}}(\varepsilon, \lambda, r)) = s_i^0$. When ε is no longer fixed then, for $i \in \mathrm{supp}(s^0)$, almost surely $S(\widehat{\beta}_i^{\mathrm{L}}(r))$ converges to s_i^0 and consequently

$$\lim_{r \to +\infty} \mathbb{P}\left(\forall i \in \mathrm{supp}(s^0), S(\widehat{\beta}_i^{\mathrm{L}}(\lambda, r)) = s_i^0\right) = 1.$$

□

Proof of Theorem 1: Let A be the set $A := \mathrm{supp}(\widehat{\beta}^{\mathrm{L}}(\lambda))$.

Upper bound) Let us give two expressions met by the LASSO estimator as defined in (2). Vector $\widehat{\beta}^{\mathrm{L}}(\lambda)$ is the LASSO estimator if and only if the following two inequalities occur simultaneously.

$$X_A'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) = \lambda S(\widehat{\beta}_A^{\mathrm{L}}(\lambda)), \tag{10}$$

$$\|X_{\overline{A}}'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda))\|_\infty \leq \lambda. \tag{11}$$

These two expressions are given in Bühlmann and van de Geer [5] page 15 or in the proof of Theorem 1 of Zou [32]. Using equality (10) and inequality (11), we are going to show that if $S(\widehat{\beta}^{\mathrm{L}}(\lambda)) = S(\beta)$ then the following event holds

$$\left\| X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I) + \frac{1}{\lambda} X_{\overline{I}}' \left(Id - X_I (X_I' X_I)^{-1} X_I'\right) \varepsilon \right\|_\infty \leq 1.$$

Let us assume that $S(\widehat{\beta}^{\mathrm{L}}(\lambda)) = S(\beta)$ thus $A = I$ (where $I = \mathrm{supp}(\beta)$). Since $Y = X\beta + \varepsilon = X_I\beta_I + \varepsilon$ and $X\widehat{\beta}^{\mathrm{L}}(\lambda) = X_I \widehat{\beta}_I^{\mathrm{L}}(\lambda)$ then the equality (10) and the inequality (11) lead to the following expressions

$$X_I' \left(\varepsilon + X_I(\beta_I - \widehat{\beta}_I^{\mathrm{L}}(\lambda))\right) = \lambda S(\beta_I), \tag{12}$$

$$\left\| X_{\overline{I}}' \left(\varepsilon + X_I(\beta_I - \widehat{\beta}_I^{\mathrm{L}}(\lambda))\right) \right\|_\infty \leq \lambda. \tag{13}$$

Equality (12) assures that

$$\beta_I - \widehat{\beta}_I^{\mathrm{L}}(\lambda) = (X_I' X_I)^{-1} \left(\lambda S(\beta_I) - X_I' \varepsilon\right).$$

Let us notice that since $\ker(X_I) = \mathbf{0}$ then the Gram matrix $X_I' X_I$ is invertible. Using the previous expression in inequality (13) gives

$$\begin{aligned} \left\| X_{\overline{I}}' X_I (X_I' X_I)^{-1} (\lambda S(\beta_I) - X_I' \varepsilon) + X_{\overline{I}}' \varepsilon \right\|_\infty &\leq \lambda, \\ \left\| X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I) + \frac{1}{\lambda} X_{\overline{I}}' \left(Id - X_I (X_I' X_I)^{-1} X_I'\right) \varepsilon \right\|_\infty &\leq 1. \end{aligned}$$

Consequently, one may deduce the following inequality

$$\mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda)) = S(\beta)\right) \leq \underbrace{\mathbb{P}\left(\left\| X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I) + \frac{1}{\lambda} X_{\overline{I}}' \left(Id - X_I (X_I' X_I)^{-1} X_I'\right) \varepsilon \right\|_\infty \leq 1\right)}_{=\mathbb{P}\left(\left\|\zeta_{X,\lambda,S(\beta)}\right\|_\infty \leq 1\right)}.$$

Sharpness of the upper bound) Since the upper bound depends only on s^0 and not on how large the non-null components $\beta^{(r)}$ are then

$$\limsup_{r\to+\infty} \mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = s^0\right) \leq \mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty \leq 1\right).$$

Finally, it must be proven that $\liminf_{r\to+\infty} \mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = s^0\right) \geq \mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1\right)$. Let us remind that $I = \mathrm{supp}(s^0)$ and let us assume that the following events hold simultaneously

$$X_I'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) = \lambda s_I^0 \text{ and } \underbrace{\left\| X_{\overline{I}}' X_I (X_I' X_I)^{-1} \lambda s_I^0 + X_{\overline{I}}' \left(Id - X_I (X_I' X_I)^{-1} X_I'\right) \varepsilon \right\|_\infty < \lambda}_{=\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1}. \tag{14}$$

We aim to show that the inequalities given above imply that $\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) = \mathbf{0}$. For convenience, let us set H be the projection matrix $H := X_I (X_I' X_I)^{-1} X_I'$. When (14) occurs then the following inequalities holds

$$\begin{aligned}
\left\| X_{\overline{I}}' H (Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) + X_{\overline{I}}' (Id - H) \varepsilon \right\|_\infty &< \lambda, \\
\left\| X_{\overline{I}}' \left(H(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) + (Id - H)\varepsilon \right) \right\|_\infty &< \lambda, \\
\left\| X_{\overline{I}}' \left(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda) + X_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) - H X_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) \right) \right\|_\infty &< \lambda.
\end{aligned} \tag{15}$$

Inequality (15) comes from the following two identities

$$\begin{aligned}
HY &= H(X\beta^{(r)}) + H\varepsilon = H(X_I \beta_I^{(r)}) + H\varepsilon = X_I \beta_I^{(r)} + H\varepsilon = X(\beta^{(r)}) + H\varepsilon \text{ and,} \\
HX\widehat{\beta}^{\mathrm{L}}(\lambda) &= HX_I \widehat{\beta}_I^{\mathrm{L}}(\lambda) + HX_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) = X_I \widehat{\beta}_I^{\mathrm{L}}(\lambda) + HX_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) = X\widehat{\beta}^{\mathrm{L}}(\lambda) - X_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) + HX_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda).
\end{aligned}$$

Let v be the vector $v := X_{\overline{I}}' \left(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda) + X_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) - HX_{\overline{I}} \widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) \right)$. We are going to see that inequality (15) implies that $\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) = \mathbf{0}$. Let us assume that $\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) \neq \mathbf{0}$ then, on the one hand, the following inequality occurs

$$\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda)' v \leq \|\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda)\|_1 \|v\|_\infty < \lambda \|\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda)\|_1. \tag{16}$$

According to (10) the identity $\widehat{\beta}_i^{\mathrm{L}}(\lambda) X_i' (Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) = \lambda |\widehat{\beta}_i^{\mathrm{L}}(\lambda)|$ occurs. Consequently, on the other hand, the

following inequalities hold

$$\begin{aligned}\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)'v &= \widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)'X'_{\overline{I}}\left(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda) + X_{\overline{I}}\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) - HX_{\overline{I}}\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)\right),\\ &= \lambda\|\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)\|_1 + \widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)'X'_{\overline{I}}(Id-H)X_{\overline{I}}\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda),\\ &\geq \lambda\|\widehat{\beta}^{L}_{\overline{I}}(\lambda)\|_1. \end{aligned} \tag{17}$$

The last inequality occurs because the projection matrix $Id - H$ is positive semi-definite. Inequalities (16) and (17) provide a contradiction which implies that $\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) = \mathbf{0}$.
According to (10), the following implication holds

$$S(\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda, r)) = s^0_I \Rightarrow X'_I(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = \lambda s^0_I.$$

Because s^0 is identifiable with respect to the L_1 norm then, according to Lemma 2, the following convergence in probability occurs

$$\lim_{r\to+\infty}\mathbb{P}(S(\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda, r)) = s^0_I) = \lim_{r\to+\infty}\mathbb{P}(X'_I(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = \lambda s^0_I) = 1. \tag{18}$$

Using this asymptotic result and since when (14) occurs then $\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda, r) = \mathbf{0}$, one may deduce the following inequalities

$$\begin{aligned}\liminf_{r\to+\infty}\mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = s^0\right) &= \liminf_{r\to+\infty}\mathbb{P}\left(S(\widehat{\beta}^{\mathrm{L}}_I(\lambda, r)) = s^0_I \text{ and } \widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda, r) = \mathbf{0}\right),\\ &= \liminf_{r\to+\infty}\mathbb{P}(\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda, r) = \mathbf{0}),\\ &\geq \liminf_{r\to+\infty}\mathbb{P}\left(X'_I(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = s^0_I \text{ and } \left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1\right),\\ &\geq \liminf_{r\to+\infty}\mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1\right).\end{aligned}$$

Asymptotic full power and asymptotic control of the FWER) According to (18), asymptotically the power is equal to 1, namely $\lim_{r\to+\infty}\mathbb{P}(\forall i \in I, S(\widehat{\beta}^{\mathrm{L}}_i(\lambda, r)) = s^0_i) = 1$. Now let us prove that the FWER is controlled asymptotically. Let us remind that $\mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1\right) = \gamma$ and $\mathbb{P}\left(\left\|\zeta_{X,\lambda,s^0}\right\|_\infty \leq 1\right) = \bar{\gamma}$. Using asymptotic results given above one may deduce the following inequalities.

$$\begin{aligned}\bar{\gamma} &\geq \limsup_{r\to+\infty}\mathbb{P}(S(\widehat{\beta}^{L}(\lambda, r)) = s^0),\\ &\geq \limsup_{r\to+\infty}\mathbb{P}\left(\forall i \in I, S(\widehat{\beta}^{L}_i(\lambda, r)) = s^0_i \text{ and } \forall i \notin I, \widehat{\beta}^{L}_i(\lambda, r) = 0\right),\\ &\geq \limsup_{r\to+\infty}\mathbb{P}(\forall i \notin I, \widehat{\beta}^{L}_i(\lambda, r) = 0).\end{aligned} \tag{19}$$

The last inequality comes from (18). Similarly, we have

$$\gamma \leq \liminf_{r\to+\infty} \mathbb{P}(\forall i \notin I, \widehat{\beta}_i^L(\lambda, r) = 0). \tag{20}$$

Consequently, by taking the complement to 1 of the inequalities given in (19) and (20), one may deduce that

$$\liminf_{r\to+\infty} \mathbb{P}(\exists i \notin I, \widehat{\beta}_i^L(\lambda, r) \neq 0) \geq 1 - \bar{\gamma} \text{ and } \limsup_{r\to+\infty} \mathbb{P}(\exists i \notin I, \widehat{\beta}_i^L(\lambda, r) \neq 0) \leq 1 - \gamma.$$

□

Proof of Theorem 2

Lemma 3 provides the same result for BPDN as does Lemma 1 for LASSO. These both lemmas are the keystones to prove Theorem 2.

Lemma 3 *Let $(\beta^{(r)})_{r\in\mathbb{N}}$ be a sequence of $\mathbb{R}^p$ satisfying conditions* **1)** *and* **2)** *of Assumption 1, let us assume that s^0 is identifiable with respect to the L_1 norm and let set $u_r = \|\beta^{(r)}\|_1$ then*

$$\lim_{r\to+\infty} \frac{\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r) - \beta^{(r)}}{u_r} = 0.$$

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

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

Because $X(u(\varepsilon)) = \varepsilon$, we have $Y(\varepsilon) = X(\beta^{(r)} + u(\varepsilon))$ and because $\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r)$ is an admissible point of (3), one deduces the following inequality

$$\left\| \frac{1}{u_r} X\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r) - \frac{1}{u_r} X\beta^{(r)} \right\|_2 \leq \left\| \frac{1}{u_r} X\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r) - \frac{1}{u_r} Y \right\|_2 + \left\| \frac{1}{u_r} Y - \frac{1}{u_r} X\beta^{(r)} \right\|_2 \leq \frac{\sqrt{R}}{u_r} + \frac{\|Xu(\varepsilon)\|_2}{u_r}. \tag{21}$$

Because $\beta^{(r)} + u(\varepsilon)$ is an admissible point of problem (3) and because $\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r)$ is the minimizer of (3), one may deduce that the following inequalities hold

$$\frac{1}{u_r} \|\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r)\|_1 \leq \frac{1}{u_r} \|\beta^{(r)} + u(\varepsilon)\|_1 \leq 1 + \frac{\|u(\varepsilon)\|_1}{u_r}. \tag{22}$$

Because u_r tends to $+\infty$ then, according to (22), the sequence $((\widehat{\beta}^{\mathrm{L}}(\varepsilon, r) - \beta^{(r)})/u_r)_{r\in\mathbb{N}^*}$ is bounded since the following superior limit is finite

$$\limsup_{r\to+\infty} \left\| \frac{\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r) - \beta^{(r)}}{u_r} \right\|_1 \leq 2.$$

Consequently, to prove that $\lim_{r\to+\infty}(\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon,r)-\beta^{(r)})/u_r=\mathbf{0}$ it is sufficient to show that $\mathbf{0}$ is the unique limit point of this sequence. Let $((\widehat{\beta}^{\mathrm{L}}(\varepsilon,\phi(r))-\beta^{(\phi(r))})/u_{\phi(r)})_{r\in\mathbb{N}^*}$ be a converging subsequence to l (with $\phi:\mathbb{N}^*\to\mathbb{N}^*$ strictly increasing) and without loss of generality, let us assume $\lim_{r\to+\infty}\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon,\phi(r))/u_{\phi(r)}=v$ and $\lim_{r\to+\infty}b^{(\phi(r))}/u_{\phi(r)}=v'$ so that $l=v-v'$. By (21) and (22) one may deduce that

$$Xv=Xv' \text{ and } \|v\|_1\leq 1.$$

Since, whatever $r\geq 0$, we have $S(\beta^{(\phi(r))}/u_{\phi(r)})=s^0$ where s^0 is identifiable with respect to the L_1 norm then, according to Proposition 2, one may deduce that $\beta^{(\phi(r))}/u_{\phi(r)}$ is an unitary vector satisfying the identifiability condition. Consequently, $\|v'\|_1=1$ and v' is identifiable with respect to the L_1 norm. Consequently, $v=v'$ and thus $l=\mathbf{0}$ is the unique limit point, which implies that

$$\lim_{r\to+\infty}\frac{\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon,r)-\beta^{(r)}}{u_r}=\mathbf{0}.$$

□

Lemma 4 is useful to prove in Theorem 2 that when s^0 is not identifiable then sign estimator derived from thresholded LASSO cannot recover s^0.

Lemma 4 *Let X be a matrix in general position, then the random vector $\widehat{\beta}$ is identifiable with respect to X and the L_1 norm.*

Proof: Let us remind that when X is in general position then the minimizer $\widehat{\beta}$ is unique. Let us assume that $\widehat{\beta}$ is not identifiable with respect to X and the L_1 norm, then there exists $b\in\mathbb{R}^p$ such that $Xb=X\widehat{\beta}$ and $\|b\|_1\leq\|\widehat{\beta}\|_1$. Consequently, for LASSO, one may deduce that

$$\|Y-Xb\|^2+\lambda\|b\|_1\leq\|Y-X\widehat{\beta}^{\mathrm{L}}\|^2+\lambda\|\widehat{\beta}^{\mathrm{L}}\|_1.$$

This inequality contradicts $\widehat{\beta}^{\mathrm{L}}$ as the unique minimizer of (2). Similarly, when $\widehat{\beta}^{\mathrm{BPDN}}$ is not identifiable with respect to the L_1 norm then $\widehat{\beta}^{\mathrm{BPDN}}$ is not the unique minimizer of (3), which provides a contradiction. □

For the proofs of Theorem 2 and the proof of Proposition 2 we need to introduce the following inequality which characterizes the identifiability condition [10]. A vector $b\in\mathbb{R}^p$ is identifiable with respect to X and the L_1 norm if and only if the following inequality holds

$$\forall h\in\ker(X)\setminus\{\mathbf{0}\},\left|\sum_{i\in\mathrm{supp}(b)}S(b)h_i\right|<\sum_{i\notin\mathrm{supp}(b)}|h_i|. \tag{23}$$

Proof of Theorem 2: Let us remind that according to condition **3)** of Assumption 1 the following inequality holds

$$\forall r \in \mathbb{N}, \frac{\min\{|\beta_i^{(r)}|, i \in \mathrm{supp}(s^0)\}}{\|\beta^{(r)}\|_\infty} \geq q > 0.$$

According to Lemmas 1 and 3, when s^0 is identifiable with respect to the L_1 norm then

$$\lim_{r \to +\infty} \frac{\widehat{\beta}(\varepsilon, r) - \beta^{(r)}}{\|\beta^{(r)}\|_\infty} = 0.$$

Therefore, there exists $r_0(\varepsilon) \geq 0$ such that

$$\forall r \geq r_0(\varepsilon), \left\| \frac{\widehat{\beta}(\varepsilon, r) - \beta^{(r)}}{\|\beta^{(r)}\|_\infty} \right\|_\infty < q/2 \Leftrightarrow \forall i \in \{1, \ldots, p\}, \forall r \geq r_0(\varepsilon), \left| \frac{\widehat{\beta}_i(\varepsilon, r) - \beta_i^{(r)}}{\|\beta^{(r)}\|_\infty} \right| < q/2.$$

Consequently, when $r \geq r_0(\varepsilon)$, whatever $i \notin \mathrm{supp}(s^0)$ (thus when $\beta_i^{(r)} = 0$) the following inequalities hold

$$\begin{aligned} & \forall i \notin \mathrm{supp}(s^0), \left| \frac{\widehat{\beta}_i(\varepsilon, r)}{\|\beta^{(r)}\|_\infty} \right| < q/2, \\ \Rightarrow \quad & -\|\beta^{(r)}\|_\infty q/2 < \min_{i \notin \mathrm{supp}(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} \leq \max_{i \notin \mathrm{supp}(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} < \|\beta^{(r)}\|_\infty q/2. \end{aligned}$$

Whatever $i \in \mathrm{supp}^+(s^0)$ (thus when $\beta_i^{(r)} > 0$) the following inequalities hold

$$\begin{aligned} & \forall i \in \mathrm{supp}^+(s^0), \frac{\widehat{\beta}_i(\varepsilon, r)}{\|\beta^{(r)}\|_\infty} \geq -\left| \frac{\widehat{\beta}_i(\varepsilon, r) - \beta_i^{(r)}}{\|\beta^{(r)}\|_\infty} \right| + \frac{\beta_i^{(r)}}{\|\beta^{(r)}\|_\infty}, \\ \Rightarrow \quad & \min_{i \in \mathrm{supp}^+(s^0)} \left\{ \frac{\widehat{\beta}_i(\varepsilon, r)}{\|\beta^{(r)}\|_\infty} \right\} > -q/2 + q = q/2, \\ \Rightarrow \quad & \min_{i \in \mathrm{supp}^+(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} > \|\beta^{(r)}\|_\infty q/2. \end{aligned}$$

Whatever $i \in \mathrm{supp}^-(s^0)$ (thus when $\beta_i^{(r)} < 0$) the following inequalities hold

$$\begin{aligned} & \forall i \in \mathrm{supp}^+(s^0), \frac{\widehat{\beta}_i(\varepsilon, r)}{\|\beta^{(r)}\|_\infty} \leq \left| \frac{\widehat{\beta}_i(\varepsilon, r) - \beta_i^{(r)}}{\|\beta^{(r)}\|_\infty} \right| + \frac{\beta_i^{(r)}}{\|\beta^{(r)}\|_\infty}, \\ \Rightarrow \quad & \max_{i \in \mathrm{supp}^-(s^0)} \left\{ \frac{\widehat{\beta}_i(\varepsilon, r)}{\|\beta^{(r)}\|_\infty} \right\} < q/2 - q = -q/2, \\ \Rightarrow \quad & \max_{i \in \mathrm{supp}^-(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} < -\|\beta^{(r)}\|_\infty q/2. \end{aligned}$$

Finally, when $r \geq r_0(\varepsilon)$ we have

i)

$$\text{supp}^-(s^0) \subset \text{supp}^-(\widehat{\beta}_i(\varepsilon, r)) \text{ and } \text{supp}^+(s^0) \subset \text{supp}^+(\widehat{\beta}_i(\varepsilon, r)).$$

ii)

$$\max_{i \in \text{supp}^-(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} < \min_{i \notin \text{supp}(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} \leq \max_{i \notin \text{supp}(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\} < \min_{i \in \text{supp}^+(s^0)} \left\{ \widehat{\beta}_i(\varepsilon, r) \right\}.$$

These achieve the first part of the proof. Now, let us assume that s^0 is not identifiable with respect to the L_1 norm. Let us show that when the following events hold

$$\text{supp}^-(s^0) \subset \text{supp}^-(\widehat{\beta}) \text{ and } \text{supp}^+(s^0) \subset \text{supp}^+(\widehat{\beta}), \tag{24}$$

then inequality (23) occurs which contradicts that s^0 is not identifiable. Let $h \in \ker(X) \setminus \{\mathbf{0}\}$. On the one hand, when (24) occurs, we have

$$\left| \sum_{i \in \text{supp}(s^0)} s_i^0 h_i \right| = \left| - \sum_{\text{supp}^-(s^0)} h_i + \sum_{\text{supp}^+(s^0)} h_i \right| \leq \left| - \sum_{i \in \text{supp}^-(\widehat{\beta})} h_i + \sum_{i \in \text{supp}^+(\widehat{\beta})} h_i \right| + \sum_{i \in \text{supp}(\widehat{\beta}) \setminus \text{supp}(s^0)} |h_i|.$$

On the other hand, according to Lemma 4, $\widehat{\beta}$ is identifiable with respect to the L_1 norm then (23) occurs implying the following inequality

$$\left| - \sum_{i \in \text{supp}^-(\widehat{\beta})} h_i + \sum_{i \in \text{supp}^+(\widehat{\beta})} h_i \right| + \sum_{i \in \text{supp}(\widehat{\beta}) \setminus \text{supp}(s^0)} |h_i| < \sum_{i \notin \text{supp}(\widehat{\beta})} |h_i| + \sum_{i \in \text{supp}(\widehat{\beta}) \setminus \text{supp}(s^0)} |h_i| = \sum_{i \notin \text{supp}(s^0)} |h_i|.$$

Consequently the following inequality holds

$$\forall h \in \ker(X) \setminus \{\mathbf{0}\}, \left| \sum_{i \in \text{supp}(s^0)} s_i^0 h_i \right| < \sum_{i \notin \text{supp}(s^0)} |h_i|,$$

which, according to (23), contradicts that s^0 is not identifiable with respect to the L_1 norm. □

Proof of propositions

The proof of Proposition 1, provided below, is the one reported in the PhD manuscript of Tardivel [23].

Proof of Proposition 1: From Daubechies et al. [10], 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

$$\forall h \in \ker(X), \left| \sum_{i \in I} S(\beta_i) h_i \right| \leq \sum_{i \notin I} |h_i|. \tag{25}$$

We are going to show that when the irrepresentable condition holds for β then the inequality (23) holds.

Let $h \in \ker(X)$ and let us remind that h_I and $h_{\overline{I}}$ denote respectively vectors $(h_i)_{i\in I}$ and $(h_i)_{i\notin I}$. Then the following equality holds

$$\sum_{i\in I} S(\beta_i)h_i = h_I' S(\beta_I) = h_I' X_I' X_I (X_I' X_I)^{-1} S(\beta_I).$$

Because $\mathbf{0} = Xh = X_I h_I + X_{\overline{I}} h_{\overline{I}}$, one may deduce the following inequalities

$$\begin{aligned} |h_I' S(\beta_I)| &= \left|h_{\overline{I}}' X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\right|, \\ &\leq \|h_{\overline{I}}\|_1 \|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty. \end{aligned} \tag{26}$$

Consequently, when the irrepresentable condition holds for β, namely when $\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(b_I^*)\|_\infty \leq 1$, then the inequality (26) gives $|h_I' S(\beta_I)| \leq \|h_{\overline{I}}\|_1$. Thus, by the equivalence given in (25), 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.* 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. □

Let us notice that when the inequality in the irrepresentable condition is strict, Theorem 1 remains true without assuming that X is in general position.

Proof of Proposition 2: Because b is identifiable with respect to the L_1 norm and because $S(\tilde{b}) = S(b)$ implies $\text{supp}(\tilde{b}) = \text{supp}(b)$, then the following inequality holds

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

Consequently, according to (23), parameter $\tilde{b}$ is identifiable with respect to the L_1 norm. □

Supplementary material

We have already said that when X is in general position the minimizer of problem (2) (resp. problem (3)) is unique. Concerning LASSO, a sketch of proof given in Tibshirani [26] shows the uniqueness of the LASSO estimator when X is in general position. In order to provide a self-contained article, we show that when X

is in general position, the minimizer of problem (3) is unique when $R = 0$ as well as when $R > 0$. We have already stressed that when β is identifiable with respect to the L_1 norm then β is sparse. We show that when the identifiability holds for β then the family $(X_i)_{i \in \text{supp}(\beta)}$ is linearly independent and thus the number of components of β equal to 0 is larger than $p - n$.

References

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

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

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

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

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

[6] 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.

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] Charles Dossal, Marie-Line Chabanol, Gabriel Peyré, and Jalal Fadili. Sharp support recovery from noisy random measurements by 1-minimization. *Applied and Computational Harmonic Analysis*, 33(1):24–43, 2012.

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

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

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

[20] J. Huang, S. Ma, and C.-H. Zhang. Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

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

[22] 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.

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

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

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

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

[27] 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.

[28] S. Wang, H. Weng, and A. Maleki. Which bridge estimator is the best for variable selection ? *arxiv*, 2018.

[29] Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is optimal for variable selection? *arXiv preprint arXiv:1705.08617*, 2017.

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

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

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