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

Patrick J C Tardivel, Malgorzata Bogdan

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

Basis Pursuit (BP), Basis Pursuit DeNoising (BPDN) and LASSO are popular procedures for identifying important predictors in the high-dimensional linear regression model $Y = X\beta + \varepsilon$. When $\varepsilon = 0$, BP can recover the sign of β when this vector is identifiable with respect to the L_1 norm, while LASSO requires a much stronger irrepresentability condition. It is known that model selection properties of LASSO can be improved by hard-thresholding its estimates. In this article we support these findings by proving that thresholded LASSO, thresholded BPDN and thresholded BP can recover the sign of β if and only if β is identifiable with respect to the L1 norm. In particular when X has iid Gaussian entries and the numbers of predictors grow linearly with the sample size then these thresholded estimators can recover the sign of β if its limiting signal sparsity is below the Donoho-Tanner transition curve. This is in contrast with vanilla LASSO which, asymptotically, can recover the sign of β only if the signal sparsity tends to 0. Numerical experiments illustrate that contrary to irrepresentability condition, the identifiability condition seems to be not affected by the structure of correlations in the X matrix.

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

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

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, $\text{supp}(\beta) := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$.

1.1 BP, BPDN and LASSO

First appeared in compressed sensing [7], BP estimator is defined as the minimizer of the following optimization problem

$$\text{argmin } \|b\|_1 \text{ subject to } Y = Xb.$$

In the noiseless case, when $\varepsilon = 0$ and $Y = X\beta$, BP optimization problem allows to recover β if and only if it satisfies the following *identifiability* condition.

Definition 1 (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*

$$X\gamma = Xb \text{ and } \gamma \neq b \Rightarrow \|\gamma\|_1 > \|b\|_1. \tag{2}$$

Note that a geometrical characterization of the identifiablity condition is given in [23]. Under the identifiability assumption, which only depends on the sign of β (Proposition 2), β is sparse. Indeed, Lemma 3 in Tardivel et al. [26] 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 guarantee that β is identifiable with respect to the L_1 norm. For example the identifiability condition holds when $\|X_1\|_2 = \cdots = \|X_p\|_2 = 1$ an the number of nonzero elements of β satisfies the mutual coherence condition [12, 15, 17]:

$$k = \text{card}\{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\} \leq \frac{1}{2}\left(1 + \frac{1}{M}\right), \text{ where } M := \max_{i \neq j} |\langle X_i, X_j \rangle| \ , \tag{3}$$

In the particular case, where the elements of X are iid normal variables, the transition curve $\rho : (0, 1) \mapsto (0, 1)$ [11], described in Figure 1, characterizes the *identifiability* condition in terms of signal sparsity $\xi = k/n$, where k is the number of nonzero elements in β. Specifically, when $n/p \to \delta \in (0, 1)$ and $k/n \to \xi \in (0, 1)$, it is known that the *identifiability* condition is satisfied with probability converging to 1 when $\xi < \rho(\delta)$ or to 0 when $\xi > \rho(\delta)$.

To take into account the noise, Chen and Donoho [7] extended BP by proposing the following estimation algorithm:

$$\widehat{\beta}^{\text{L}} := \underset{b \in \mathbb{R}^p}{\text{argmin}} \ \frac{1}{2}\|Y - Xb\|_2^2 + \lambda \|b\|_1 \text{ with } \lambda = \sigma\sqrt{2 \log p}. \tag{4}$$

In the seminal paper of Tibshirani [27], the algorithm was extended for the estimation of general multiple regression models, where different values of λ are often more appropriate. The method gained a large popularity in a statistical community under the name of LASSO, while in the signal processing community it is often called

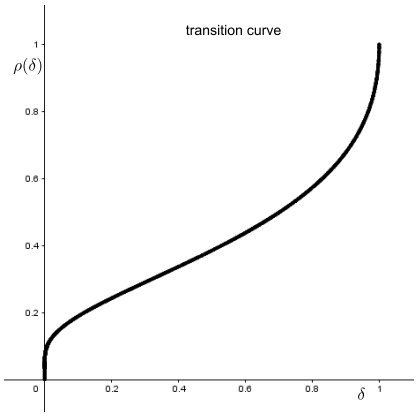

Figure 1: This figure illustrates the transition curve. Let ϕ and Φ be respectively the density and the cumulative distribution function of a $\mathcal{N}(0,1)$ distribution then, this curve is described parametrically as follows $\{(\delta, \rho(\delta)), \delta \in (0,1)\} = \left\{ \left(\frac{2\phi(t)}{t+2(\phi(t)-t\Phi(-t))}, \frac{\phi(t)-t\Phi(-t)}{\phi(t)} \right), t \in (0,+\infty) \right\}$ [2].

the Basis Pursuit Denoising (BPDN). In this article we will use the term BPDN for a slightly different form of this estimator, where the connection with BP is even more clear

$$\widehat{\beta}^{\mathrm{BPDN}} := \underset{b \in \mathbb{R}^p}{\operatorname{argmin}} \|b\|_1 \text{ subject to } \|Y - Xb\|_2^2 \leq R, \text{ where } R \geq 0. \tag{5}$$

Given a particular vector $Y \in \mathbb{R}^n$, there is a one-to-one correspondence between the tuning parameter $\lambda > 0$ and the regularization parameter $R > 0$, under which optimization problems given in (4) and (5) have the same minimizer (see *e.g* page 64 of [15]). However, the relationship between λ and R depends on the specific realization of Y and, given a fixed $\lambda > 0$ for LASSO, we cannot pick a fixed $R > 0$ for BPDN under which these both estimators are equal. To illustrate this fact, one may notice that $\widehat{\beta}^{\mathrm{BPDN}} = 0$ if and only if $\|Y\|_2^2 \leq R$ whereas for LASSO, $\widehat{\beta}^{\mathrm{LASSO}} = 0$ if and only if $\|X'Y\|_\infty \leq \lambda$.

1.2 Sign recovery by LASSO

Properties of the LASSO sign estimator $S(\widehat{\beta}^{\mathrm{L}}(\lambda)) := \left(S(\widehat{\beta}_1^{\mathrm{L}}(\lambda)), \ldots, S(\widehat{\beta}_p^{\mathrm{L}}(\lambda)) \right)$ (or properties of the active set estimator $\mathrm{supp}(\widehat{\beta}^{\mathrm{L}}(\lambda)) := \{i \in \{1,\ldots,p\} \mid \widehat{\beta}_i^{\mathrm{L}}(\lambda) \neq 0\}$) have been intensively studied [14, 19, 30, 33, 34]. Specifically, Zhao and Yu [33] and Zou [34] 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 2 (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 Theorem 2 of Wainwright [30], 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.

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 $\beta \in \mathbb{R}^p$, $I := \mathrm{supp}(\beta)$ 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.*

In case when the elements of X are iid random variables from the normal distributions, the irrepresentability condition is satisfied with a large probability if and only if $k \leq \frac{n}{2 \log p}$ (see [30]). This implies that LASSO cannot recover $S(\beta)$ in the linear sparsity of [11] if the limiting signal sparsity satisfies $\xi > 0$, even if $\xi < \rho(\delta)$. This observation is confirmed in [24], where the best achievable tradeoff diagram between the asymptotic True Positive Proportion and the False Discovery Proportion by LASSO is provided and illustrated with extensive computer simulations. The tradeoff diagram of [24] holds even in the noisless case ($\epsilon = 0$). Thus, in this case BP can recover $S(\beta)$ under a much wider range of sparsities than LASSO. This clearly illustrates that in case of gaussian matrices with independent entries the irrepresentability condition is a much stricter condition than the identifiability condition.

1.3 Sign recovery by thresholded LASSO

It is well known that LASSO can consistently estimate β under much weaker assumptions than the irrepresentable condition (see *e.g.* [20] or [29]). This implies that appropriately thresholded LASSO can recover $S(\beta)$ under weaker assumptions than the irrepresentable condition [21]. Concerning sign recovery by thresholded BP, first theoretical results were given by Saligrama and Zhao [22]. More recently, Descloux and Sardy [9] show that the stable nullspace property is a sufficient condition to recover the sign of β by thresholded BP.

To our knowledge, so far the stable nullspace property was the weakest sufficient condition known to guarantee recovery of $S(\beta)$ by thresholded BP and by thresholded LASSO. In Theorem 1 of this paper we show that the stable nullspace property can be relaxed by the weaker "identifiability" condition which is sufficient to recover $S(\beta)$ by thresholded BP, thresholded BPDN or thresholded LASSO, when the magnitudes of nonzero

elements of β are large enough. We also show that "identifiability" is necessary for recovering $S(\beta)$ by these procedures, independently of the signal magnitude.

Theorem 1 implies that in the linear sparsity regime of [11] for gaussian matrices, thresholded BP, thresholded BPDN and thresholded LASSO can recover $S(\beta)$ if the asymptotic sparsity of β is below the transition curve and the signal magnitude tends to infinity. This is in a stark contrast with the regular LASSO, which can not properly identify $S(\beta)$ under this regime, independently of the signal magnitude (see [24]).

1.4 Graphical illustration of the main result

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, 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 2, we provide the irrepresentability and the identifiability curves, representing the proportion of the sign vectors with k nonzero elements which satisfy the identifiability condition or the irrepresentability condition for two specific design matrices of dimensions 100×300.

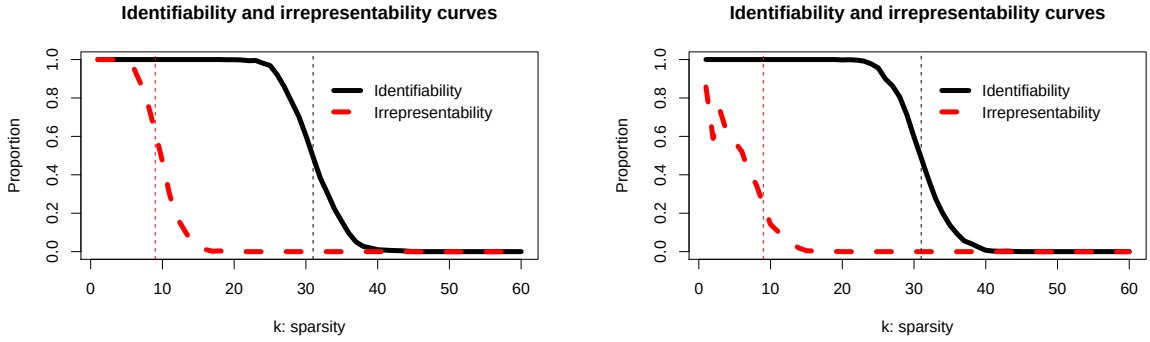

Figure 2: Identifiability and irrepresentability curves for two specific design matrices of dimensions 100×300. In both cases the rows of X were generated as iid random vectors from the multivariate normal distribution $N(0, \Sigma)$, where in the left panel $\Sigma = I$ and in the second panel Σ is of a compound symmetry form, with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.9$ for $i \neq j$. The x-axis represents the sparsity k and the y-axis represents the proportion of sign vectors satisfying the identifiability condition (resp. irrepresentability condition). In the left panel the vertical lines represent values $k = \frac{100}{2\log(200)}$ and $k = 100 \times \rho(100/300)$, with $\rho(\cdot)$ being the transition curve of Donoho and Tanner [11]. These values correspond to the asymptotic upper limits for k so the irrepresentability and identifiability conditions hold for Gaussian design matrices with independent entries (see [30] and [13]).

Figure 2 illustrates a great potential improvement for sign recovery by thresholded LASSO with respect to LASSO. When the columns of the design matrix are independent gaussian vectors than irrepresentability condition is satisfied with large probably only when $k \leq 10$, which is also a limiting sparsity for recovery of $S(\beta)$ by LASSO. Instead, the identifiability condition, which represents the limiting sparsity for recovery of $S(\beta)$ by thresholded LASSO holds when $k \leq 30$. The simulations are in a strong agreement with the asymptotic results given in [30] and [13], which predict the upper limits for k so the irrepresentability and identifiability conditions

hold for Gaussian design matrices with independent entries. Specifically, the right vertical line represents the sparsity predicted by the transition curve of Donoho and Tanner [11].

The difference between identifiability and irrepresentability curves becomes even more pronounced in the presence of strong correlations between different columns in the design matrix. As expected, irrepresentability curve is shrinked towards zero. Instead, and rather unexpectedly, the simulated identifiability curve remains intact.

1.5 Organization of the article

In Section 2, Theorem 1 shows that identifiability is a necessary and sufficient 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. Corollary 1 shows that under the asymptotic linear sparsity regime for Gaussian matrices thresholded LASSO and thresholded BPDN can recover the sign of β when $\xi < \rho(\delta)$. In Section 3, 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 4 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. Proofs are provided in appendix. In this section, we also formulate and discuss Proposition 3, which provides a tight upper bound for the probability to recover $S(\beta)$ by LASSO.

1.6 Notations and assumptions

In this article we always assume that design matrix X has columns in general position (see *e.g* [28] or the supplementary material for this manuscript). This assumption guarantees that the minimizer of (4) (resp. minimizer of (5)) 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 $(X_{11}, X_{12}, \ldots, X_{np})$ have a density on $\mathbb{R}^{np}$ then, almost surely, X is in general position [28].

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

- Let I be the subset of $\{1, \ldots, p\}$. We denote by $\overline{I}$ the complement of I, namely $\overline{I} := \{1, \ldots, p\} \setminus I$.
- 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*

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

Let us notice that identifiability and irrepresentability conditions just depends from s^0 (and do not depend on the magnitude of non-null components of $\beta^{(r)}$).

2 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 1 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 1 *Let X be a $n \times p$ matrix in general position and such that* $\mathrm{rank}(X) = n$ *and let $\widehat{\beta}$ be the LASSO or BPDN estimator with an arbitrary fixed value of the tuning parameter $\lambda > 0$ or with an arbitrary fixed regularization parameter $R \geq 0$.*

Necessary condition for sign recovery: *If $S(\beta)$ is not identifiable with respect to the L_1 norm then the sign estimator derived from thresholded LASSO or thresholded BPDN cannot recover $S(\beta)$. Indeed, for any fixed $\varepsilon \in \mathbb{R}^n$, the sign of at least one non-null component of β is not correctly estimate by $\widehat{\beta}(\varepsilon)$:*

$$\exists i \in \mathrm{supp}(\beta) \text{ such that } \widehat{\beta}_i(\varepsilon)\beta_i \leq 0.$$

Sufficient condition for sign recovery: *This condition is asymptotic. 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 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)

$$\forall i \in \mathrm{supp}(\beta^{(r)}),\ \widehat{\beta}_i(\varepsilon, r)\beta_i^{(r)} > 0.$$

ii)

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

Let us notice that the assumptions on X are very weak and generically hold when $n \leq p$. The assumption that $\mathrm{rank}(X) = n$ 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 1 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 1 suggests that $S(\beta)$ can be recovered by deriving sign estimators from the thresholded LASSO or thresholded BPDN. In Section 4 we show how the appropriate thresholds can be obtained with help from control variables constructed according to the knockoff methodology (see e.g. [1, 6]).

In the asymptotic linear sparsity regime for Gaussian matrices, the transition curve $\rho(\cdot)$ described in [11] allows to characterize the identifiability condition, and thus Theorem 1 provides the Corollary 1.

Corollary 1 *Let X be a $n \times p_n$ standard Gaussian matrix and let $\beta^{(n)} \in \mathbb{R}^{p_n}$ have k_n nonzero components.* **Necessary condition for sign recovery:** *If $n/p_n \to \delta \in (0,1)$ and $k_n/n \to \xi \in (0,1)$ and $\xi > \rho(\delta)$ then asymptotically the sign of at least one non-null component of $\beta^{(n)}$ is not correctly estimated by $\widehat{\beta}$:*

$$\lim_{n \to +\infty} \mathbb{P}_{X,\varepsilon}\left(\exists i \in \mathrm{supp}(\beta^{(n)}) \text{ such that } \widehat{\beta}_i\beta_i^{(n)} \leq 0\right) = 1.$$

Sufficient condition for sign recovery: *This condition is asymptotic on the signal strength. Given n, let $(\beta^{(n,r)})_{r \in \mathbb{N}}$ be a sequence satisfying Assumption 1 (where q does not depend on n). If $n/p_n \to \delta \in (0,1)$ and $k_n/n \to \xi \in (0,1)$ and $\xi < \rho(\delta)$ then asymptotically the estimator $\widehat{\beta}$ separates negative components of $\beta^{(n,r)}$, null components of $\beta^{(n,r)}$ and positive components of $\beta^{(n,r)}$:*

i)

$$\lim_{n \to +\infty} \lim_{r \to +\infty} \mathbb{P}_{X,\varepsilon}\left(\forall i \in \mathrm{supp}(\beta^{(n,r)}),\ \widehat{\beta}_i\beta_i^{(n,r)} > 0\right) = 1.$$

ii)

$$\lim_{n\to+\infty}\lim_{r\to+\infty}\mathbb{P}_{X,\varepsilon}\left(\max_{i\in\mathrm{supp}^{-}(\beta^{(n,r)})}\left\{\widehat{\beta}_i\right\}<\min_{i\notin\mathrm{supp}(\beta^{(n,r)})}\left\{\widehat{\beta}_i\right\}\leq\max_{i\notin\mathrm{supp}(\beta^{(n,r)})}\left\{\widehat{\beta}_i\right\}<\min_{i\in\mathrm{supp}^{+}(\beta^{(n,r)})}\left\{\widehat{\beta}_i\right\}\right)=1.$$

3 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:

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

The irrepresentability indicator indicates if the LASSO sign estimator can recover $S(\beta)$. Indeed, if $\phi^X_{\mathrm{IC}}(S(\beta))=0$ 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:

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

Such an identifiability indicator indicates if the sign estimators obtained by thresholded LASSO and thresholded BPDN can recover $S(\beta)$. Indeed, if $\phi^X_{\mathrm{Idtf}}(S(\beta))=0$ 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 $(X_i)_{i\in\mathrm{supp}(\beta)}$ are not linearly independent then β does not satisfy the identifiability condition. Consequently, when $\mathrm{card}(\mathrm{supp}(\beta))>n$ then $\phi^X_{\mathrm{IC}}(S(\beta))=\phi^X_{\mathrm{Idtf}}(S(\beta))=0$. 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 $s \in \{-1, 0, 1\}^p$, $\Phi^X_{\mathrm{IC}}(s) \leq \Phi^X_{\mathrm{Idtf}}(s)$.

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.

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

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

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 $\{u \in \{-1, 0, 1\}^p \mid \mathrm{card}(\mathrm{supp}(u)) = k\}$.

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 $\{u \in \{0, 1\}^p \mid \mathrm{card}(\mathrm{supp}(u)) = k\}$.

The results for first two settings are presented in Figure 2 in the Introduction. They illustrate that the irrepresentability condition is a much stronger condition than the identifiability condition. When comparing the graphs for Setting 1 and Setting 2 we can observe that the irrepresentability condition becomes substantially more stringent when the columns in design matrix are strongly correlated while the identifiability curve remains intact.

Figure 3 illustrates an interesting behavior of irrepresentability and identifiability curves in Setting 2 with positive components of the vector β. Here we can observe that the irrepresentability condition becomes even 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 3. 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 $k = 2$.

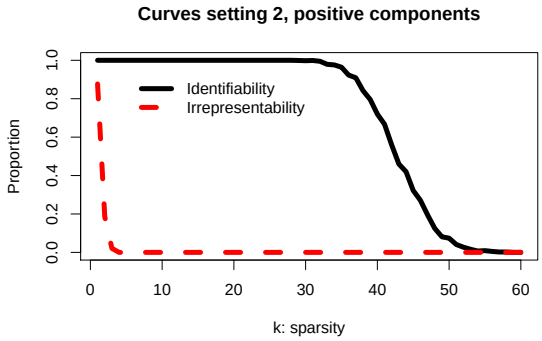

Figure 3: Graphs of the functions $k \mapsto p^X_{\mathrm{Idtf}+}(k)$ and $k \mapsto p^X_{\mathrm{IC}+}(k)$ in Setting 2 with positive components.

4 Numerical comparisons of sign estimators

Theorem 1 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 [34]. Indeed, as claimed in [34] or [18], 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.

4.1 Selection of the tuning parameter

As explained in [4, 31], 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.

4.1.1 Tuning parameter for LASSO sign estimator

When the sign $S(\beta)$ satisfies the irrepresentability condition, then by Proposition 3 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 3.1, 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 Proposition 3 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.

4.1.2 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 [3, 4, 24]. 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, 31]). As discussed in [4, 31], 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 $\lambda_{AMP}(k, t)$ 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. The values $\lambda_{AMP}(k, t)$ were then used for simulations in the independent case. In the "correlated" case these values turned out to be suboptimal, therefore in this case we additionally report results for the tuning parameters equal $0.5\lambda_{AMP}(k, t)$, which provide a substantially better empirical performance.

4.2 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 $\max\left\{|\widehat{\beta}_i|, i \notin \mathrm{supp}(\beta)\right\}$ 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 $\beta = \mathbf{0}$ [16]. For the BP estimator, Descloux and Sardy [9] 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. [24] for additional explanations).

The recently developed knockoff methodology [1, 6] 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 [6] 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 [6] (see Weinstein et al. [32] 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 [6, 32] 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 3.1, 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.

4.2.1 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*.

4.3 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 [9].

- 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.

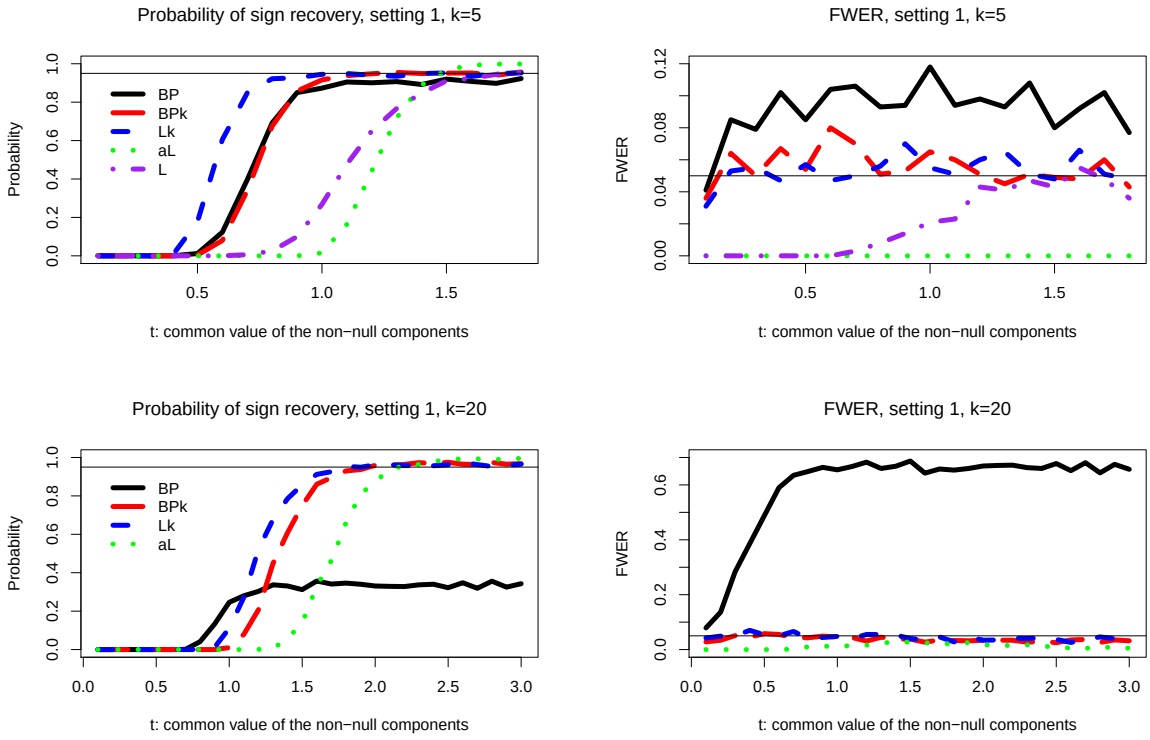

Figure 4: This figure provides the FWER and the probability to recover $S(\beta)$ for each sign estimators and when X is the design matrix given in setting 1. Graphics on the left provide the probability to recover $S(\beta)$ (on the y-axis) as a function of t, where t measures how large the non-null 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.

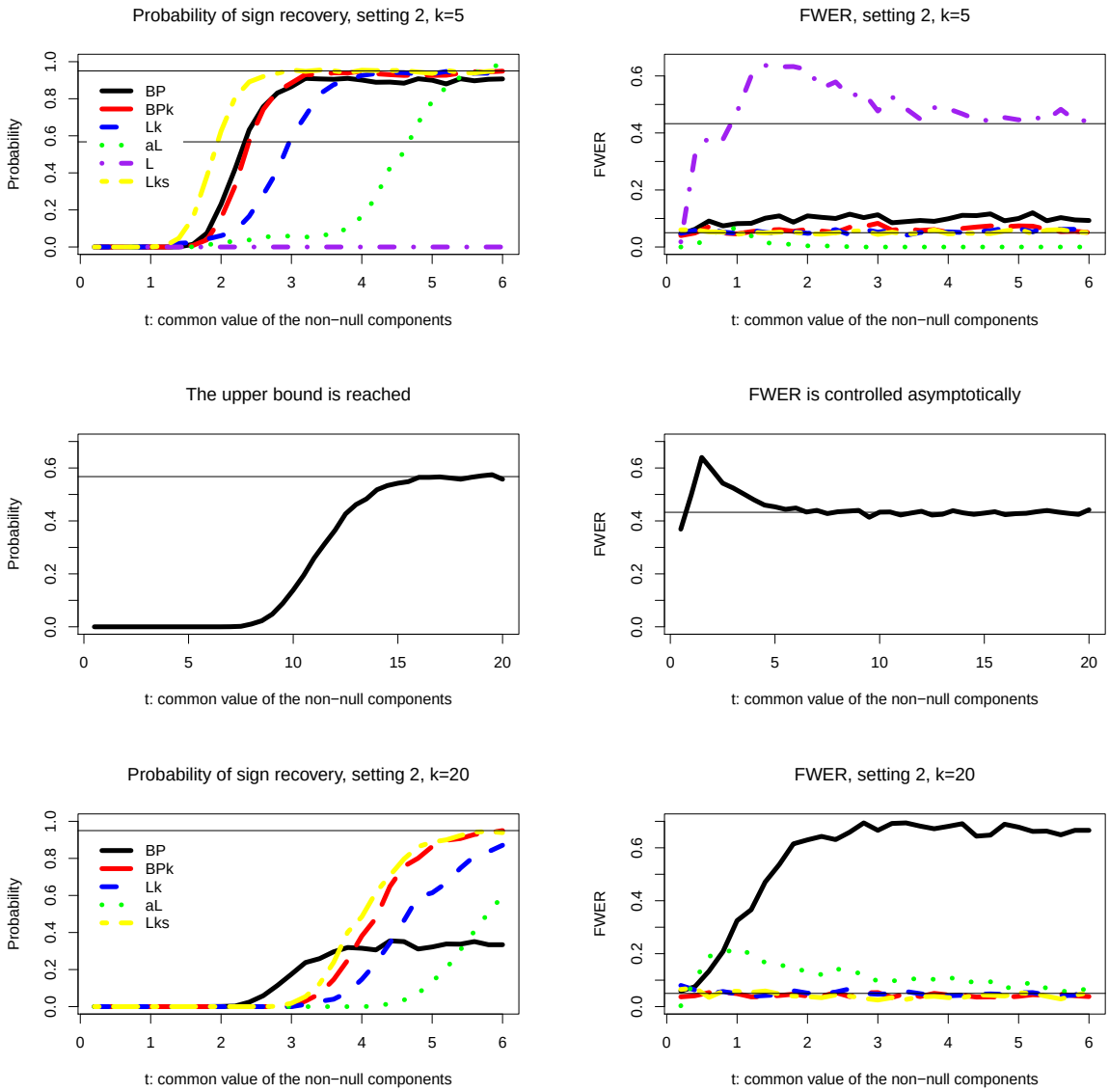

Figure 5: This figure provides the FWER and the probability to recover $S(\beta)$ for each sign estimators and when X is the design matrix given in setting 2. Graphics on the left provide the probability to recover $S(\beta)$ (on the y-axis) as a function of t (on the x-axis), where t measures how large 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 Proposition 3). One may notice that the upper-bound 0.55 is approximately reached and the FWER is approximately 0.45 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.

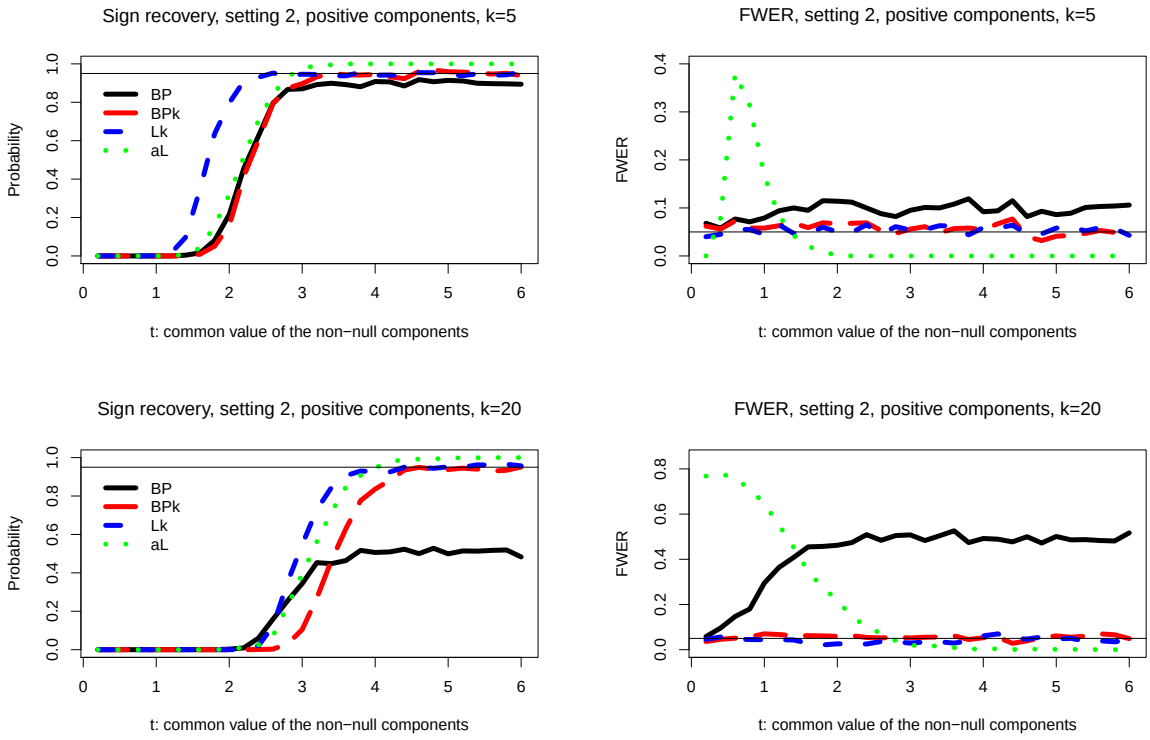

Figure 6: This figure provides the FWER and the probability to recover $S(\beta)$ for each sign estimator when X is the design matrix given in setting 2 and non-null components of β are positive. Graphics on the left provide the probability to recover $S(\beta)$ (on the y-axis) as a function of t (on 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 4-6 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 2 and 3, 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 [9] 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.

5 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.

6 appendix

6.1 Sign recovery with LASSO sign estimator

The upper bound given in Proposition 3 is given in Lemma 3 in Wainwright [30]. In this later article, this lemma is a technical result to establish the irrepresentable condition (Theorem 2 in Wainwright [30]). According to Proposition 3, when β is identifiable with respect to the L_1 norm, this upper bound is reached asymptotically once $\min\{|\beta_i|, i \in \text{supp}(\beta)\}$ tends to $+\infty$.

Proposition 3 *Let $I := \text{supp}(\beta)$ 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 $\zeta_{X,\lambda,S(\beta)} := X'_{\overline{I}} X_I (X'_I X_I)^{-1} S(\beta_I) + \frac{1}{\lambda} X'_{\overline{I}} \left(Id - X_I (X'_I X_I)^{-1} X'_I\right) \varepsilon$. Hereafter, let us remind the upper bound given by Wainwright [30]:*

Upper bound (Lemma 3 in Wainwright): *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 $1-\gamma$.

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

Remark 1 *Results given in Proposition 3 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 $\mathbb{P}(\forall i \notin \mathrm{supp}(\beta), \hat{\beta}_i^{\mathrm{L}}(\lambda) = 0)$.*

When ε has a covariance matrix $\sigma^2 Id_n$ one can obtain asymptotic results by taking β fixed and by letting σ tends to 0. However, contrarily to our asymptotic setting described in Assumption 1, such asymptotic results are very poor and do not give any information about the FWER. Indeed, when σ tends to 0 the upper bound tends to 0 or 1 depending on whether or not the irrepresentable condition holds on β.

Let us remind that the FWER is equal to $\mathbb{P}(\exists i \notin \mathrm{supp}(\beta), \widehat{\beta}_i^L \neq 0)$. According to Proposition 3, when non-null components of β are infinitely large, the FWER is controlled at level $1 - \mathbb{P}\left(\left\|\zeta_{X,\lambda,\mathrm{sign}(\beta)}\right\|_\infty < 1\right)$. To our knowledge, it is first theoretical result providing a formula for the FWER. Hereafter, let us provide some comments about the FWER control.

- To provide a specific value for λ allowing to control the FWER one needs to know the distribution of ε. For example, when the distribution of ε is known and $\beta = \mathbf{0}$ one controls the FWER at level α by taking λ as the $1-\alpha$ quantile of $\|X'\varepsilon\|_\infty$. Let us point out that when $\varepsilon_1, \ldots, \varepsilon_p$ are i.i.d and the variance σ^2 of these components is unknown then σ^2 can be consistently estimate as explained in [10].

- It is easier to control the FWER when X is a random matrix whose distribution is symmetric and invariant by columns permutation than when X is a fixed design matrix. Indeed, when X is random, the distribution of $\zeta_{X,\lambda,S(\beta)}$ just depends from the sparsity of β and not on $S(\beta)$.

- Let $k \in \{1, \ldots, p\}$ and let U_k be the set $\{u \in \{-1,0,1\}^p \mid \mathrm{card}(\mathrm{supp}(u)) = k\}$ and let us assume that $\beta \in U_k$. When X is a random matrix whose distribution is symmetric and invariant by columns permutation by taking λ_α such that $\mathbb{P}(\|\zeta_{X,\lambda_\alpha,u_0}\| < 1) = 1-\alpha$ one controls asymptotically the FWER at level α (where $u_0 := (1, \ldots, 1, 0, \ldots, 0) \in U_k$). Such a tuning parameter λ_α is easy to infer by Monte Carlo simulations. When X is fixed, by taking λ_α as follows

$$\frac{1}{\mathrm{card}(U_k)} \sum_{u\in U_k} \mathbb{P}(\|\zeta_{X,\lambda_\alpha,u}\| < 1) = 1-\alpha, \tag{8}$$

then the average value of the FWER with respect to $\beta \in U_k$ is $1-\alpha$ (where non-null components of β are infinitely large). Again λ_α is easy to infer. In the numerical experiment, the tuning parameter was selected by solving equation (8).

6.2 Proof of the Proposition 3

First, let us provide lemmas which are useful to prove both Proposition 3 and Theorem 1. Lemma 2 partially proves Proposition 3. 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 (4) 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

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

In addition, Cauchy-Schwarz inequality gives the following implications

$$\begin{aligned} & \frac{1}{2}\|\varepsilon + X\beta^{(r)} - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2^2 + \lambda\|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_1 \leq \frac{1}{2}\|\varepsilon\|_2^2 + \lambda\|\beta^{(r)}\|_1, \\ \Rightarrow \quad & -\|\varepsilon\|_2\|X\beta^{(r)} - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2 + \frac{1}{2}\|X\beta^{(r)} - X\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_2^2 + \lambda\|\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)\|_1 \leq \lambda\|\beta^{(r)}\|_1, \\ \Rightarrow \quad & -\frac{\|\varepsilon\|_2}{u_r}\left\|X\left(\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r) - \beta^{(r)}}{u_r}\right)\right\|_2 + \frac{1}{2}\left\|X\left(\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r) - \beta^{(r)}}{u_r}\right)\right\|_2^2 + \frac{\lambda}{u_r}\left\|\frac{\widehat{\beta}^{\mathrm{L}}(\varepsilon, r)}{u_r}\right\|_1 \leq \frac{\lambda}{u_r}. \end{aligned} \tag{10}$$

Because u_r tends to $+\infty$ then, according to (9), 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 (9) and (10) 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 Proposition 3:

Let us remind that the vector $\widehat{\beta}^{\mathrm{L}}(\lambda)$ is the LASSO estimator if and only if the following two inequalities occur simultaneously.

$$\begin{aligned} X_A'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) &= \lambda S(\widehat{\beta}_A^{\mathrm{L}}(\lambda)), \text{ where } A = \mathrm{supp}(\widehat{\beta}^{\mathrm{L}}(\lambda)), \quad (11) \\ \|X_{\overline{A}}'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda))\|_\infty &\leq \lambda. \quad (12) \end{aligned}$$

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

$$X_I'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) = \lambda s_I^0 \text{ and } \underbrace{\left\|X_{\overline{I}}' X_I (X_I' X_I)^{-1} \lambda s_I^0 + X_{\overline{I}}'\left(Id - X_I(X_I'X_I)^{-1}X_I'\right)\varepsilon\right\|_\infty}_{=\left\|\zeta_{X,\lambda,s^0}\right\|_\infty < 1} < \lambda. \quad (13)$$

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 (13) occurs then the following inequalities holds

$$\begin{aligned} \left\|X_{\overline{I}}' H(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) + X_{\overline{I}}'(Id - H)\varepsilon\right\|_\infty &< \lambda, \\ \left\|X_{\overline{I}}'\left(H(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) + (Id - H)\varepsilon\right)\right\|_\infty &< \lambda, \\ \left\|X_{\overline{I}}'\left(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda) + X_{\overline{I}}\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda) - HX_{\overline{I}}\widehat{\beta}_{\overline{I}}^{\mathrm{L}}(\lambda)\right)\right\|_\infty &< \lambda. \quad (14) \end{aligned}$$

Inequality (14) 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 (14)

implies that $\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) = \mathbf{0}$. Let us assume that $\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) \neq \mathbf{0}$ then, on the one hand, the following inequality occurs

$$\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)'v \leq \|\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)\|_1 \|v\|_\infty < \lambda \|\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)\|_1. \tag{15}$$

According to (11) the identity $\widehat{\beta}^{\mathrm{L}}_i(\lambda) X_i'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda)) = \lambda |\widehat{\beta}^{\mathrm{L}}_i(\lambda)|$ occurs. Consequently, on the other hand, the following inequalities hold

$$\begin{aligned}
\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)'v &= \widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)' X_{\overline{I}}' \left(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda) + X_{\overline{I}}\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) - H X_{\overline{I}}\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) \right), \\
&= \lambda \|\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)\|_1 + \widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda)' X_{\overline{I}}' (Id - H) X_{\overline{I}} \widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda), \\
&\geq \lambda \|\widehat{\beta}^{L}_{\overline{I}}(\lambda)\|_1.
\end{aligned} \tag{16}$$

The last inequality occurs because the projection matrix $Id - H$ is positive semi-definite. Inequalities (15) and (16) provide a contradiction which implies that $\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda) = \mathbf{0}$.
According to (11), the following implication holds

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

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

$$\lim_{r \to +\infty} \mathbb{P}(S(\widehat{\beta}^{\mathrm{L}}_I(\lambda, r)) = s^0_I) = \lim_{r \to +\infty} \mathbb{P}(X_I'(Y - X\widehat{\beta}^{\mathrm{L}}(\lambda, r)) = \lambda s^0_I) = 1. \tag{17}$$

Using this asymptotic result and since when (13) occurs then $\widehat{\beta}^{\mathrm{L}}_{\overline{I}}(\lambda, r) = \mathbf{0}$, one may deduce the following inequalities

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

Asymptotic full power and asymptotic control of the FWER) According to (17), 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.

$$\begin{aligned}\bar{\gamma} &\geq \limsup_{r\to+\infty} \mathbb{P}(S(\widehat{\beta}^L(\lambda,r)) = s^0),\\ &\geq \limsup_{r\to+\infty} \mathbb{P}\left(\forall i \in I, S(\widehat{\beta}_i^L(\lambda,r)) = s_i^0 \text{ and } \forall i \notin I, \widehat{\beta}_i^L(\lambda,r) = 0\right),\\ &\geq \limsup_{r\to+\infty} \mathbb{P}(\forall i \notin I, \widehat{\beta}_i^L(\lambda,r) = 0). \end{aligned} \tag{18}$$

The last inequality comes from (17). Similarly, we have

$$\gamma \leq \liminf_{r\to+\infty} \mathbb{P}(\forall i \notin I, \widehat{\beta}_i^L(\lambda,r) = 0). \tag{19}$$

Consequently, by taking the complement to 1 of the inequalities given in (18) and (19), 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 1

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

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

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

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 (5), one deduces the following inequality

$$\left\|\frac{1}{u_r}X\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon,r) - \frac{1}{u_r}X\beta^{(r)}\right\|_2 \leq \left\|\frac{1}{u_r}X\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon,r) - \frac{1}{u_r}Y\right\|_2 + \left\|\frac{1}{u_r}Y - \frac{1}{u_r}X\beta^{(r)}\right\|_2 \leq \frac{\sqrt{R}}{u_r} + \frac{\|Xu(\varepsilon)\|_2}{u_r}. \tag{20}$$

Because $\beta^{(r)} + u(\varepsilon)$ is an admissible point of problem (5) and because $\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r)$ is the minimizer of (5), one may deduce that the following inequalities hold

$$\frac{1}{u_r}\|\widehat{\beta}^{\mathrm{BPDN}}(\varepsilon, r)\|_1 \leq \frac{1}{u_r}\|\beta^{(r)} + u(\varepsilon)\|_1 \leq 1 + \frac{\|u(\varepsilon)\|_1}{u_r}. \tag{21}$$

Because u_r tends to $+\infty$ then, according to (21), 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 (20) and (21) 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 1 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 (4). 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 (5), which provides a contradiction. $\square$

For the proofs of Theorem 1 and the proof of Proposition 2 we need to introduce the following inequality which characterizes the identifiability condition [8]. 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

$$\forall h \in \ker(X) \setminus \{\mathbf{0}\}, \left| \sum_{i \in \mathrm{supp}(b)} S(b) h_i \right| < \sum_{i \notin \mathrm{supp}(b)} |h_i|. \tag{22}$$

Proof of Theorem 1:

Necessary condition: Let us assume that $S(\beta)$ is not identifiable with respect to the L_1 norm. Let us show that when the following events hold

$$\mathrm{supp}^-(\beta) \subset \mathrm{supp}^-(\widehat{\beta}(\varepsilon)) \text{ and } \mathrm{supp}^+(\beta) \subset \mathrm{supp}^+(\widehat{\beta}(\varepsilon)), \tag{23}$$

then inequality (22) occurs which contradicts that $S(\beta)$ is not identifiable with respect to the L_1 norm. Let $h \in \ker(X) \setminus \{\mathbf{0}\}$. On the one hand, when (23) occurs, we have

$$\begin{aligned}
\left| \sum_{i \in \mathrm{supp}(\beta)} S(\beta_i) h_i \right| &= \left| - \sum_{\mathrm{supp}^-(\beta)} h_i + \sum_{\mathrm{supp}^+(\beta)} h_i \right|, \\
&= \left| - \sum_{i \in \mathrm{supp}^-(\widehat{\beta}(\varepsilon))} h_i + \sum_{i \in \mathrm{supp}^-(\widehat{\beta}(\varepsilon)) \setminus \mathrm{supp}^-(\beta)} h_i + \sum_{i \in \mathrm{supp}^+(\widehat{\beta}(\varepsilon))} h_i - \sum_{i \in \mathrm{supp}^+(\widehat{\beta}(\varepsilon)) \setminus \mathrm{supp}^+(\beta)} h_i \right|, \\
&\leq \left| - \sum_{i \in \mathrm{supp}^-(\widehat{\beta}(\varepsilon))} h_i + \sum_{i \in \mathrm{supp}^+(\widehat{\beta}(\varepsilon))} h_i \right| + \sum_{i \in \mathrm{supp}(\widehat{\beta}(\varepsilon)) \setminus \mathrm{supp}(\beta)} |h_i|.
\end{aligned}$$

On the other hand, according to Lemma 4, $\widehat{\beta}(\varepsilon)$ is identifiable with respect to the L_1 norm then (22) occurs implying the following inequality

$$\begin{aligned}
&\left| - \sum_{i \in \mathrm{supp}^-(\widehat{\beta}(\varepsilon))} h_i + \sum_{i \in \mathrm{supp}^+(\widehat{\beta}(\varepsilon))} h_i \right| + \sum_{i \in \mathrm{supp}(\widehat{\beta}(\varepsilon)) \setminus \mathrm{supp}(\beta)} |h_i| \\
&\qquad < \sum_{i \notin \mathrm{supp}(\widehat{\beta}(\varepsilon))} |h_i| + \sum_{i \in \mathrm{supp}(\widehat{\beta}(\varepsilon)) \setminus \mathrm{supp}(\beta)} |h_i| = \sum_{i \notin \mathrm{supp}(\beta)} |h_i|.
\end{aligned}$$

Consequently the following inequality holds

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

which, according to (22), contradicts that $S(\beta)$ is not identifiable with respect to the L_1 norm.

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

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

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 $i \notin \text{supp}(s^0)$ (thus when $\beta_i^{(r)} = 0$) the following inequalities hold

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

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

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

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

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

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 proof of the sufficient condition. □

Proof of propositions

The proof of Proposition 1, provided below, is the one reported in the PhD manuscript of Tardivel [25].

Proof of Proposition 1: From Daubechies et al. [8], 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

$$\forall h \in \ker(X), \left| \sum_{i \in I} S(\beta_i) h_i \right| \leq \sum_{i \notin I} |h_i|. \tag{24}$$

We are going to show that when the irrepresentable condition holds for β then the inequality (22) 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

$$\begin{aligned} |h_I' S(\beta_I)| &= \left| h_{\overline{I}}' X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I) \right|, \\ &\leq \|h_{\overline{I}}\|_1 \|X_{\overline{I}}' X_I (X_I' X_I)^{-1} S(\beta_I)\|_\infty. \end{aligned} \tag{25}$$

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 (25) gives $|h_I' S(\beta_I)| \leq \|h_{\overline{I}}\|_1$. Thus, by the equivalence given in (24), 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 (22), 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 (4) (resp. problem (5)) is unique. Concerning LASSO, a sketch of proof given in Tibshirani [28] 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 (5) 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 $p - n$. Finally, a proof that the stable nullspace property implies the identifiability condition is given

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, Marc Lelarge, Andrea Montanari, et al. Universality in polytope phase transitions and message passing algorithms. The Annals of Applied Probability, 25(2):753–822, 2015.

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

[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] 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.

[7] Shaobing Chen and David Donoho. Basis pursuit. In Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, volume 1, pages 41–44. IEEE, 1994.

[8] 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.

[9] 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.

[10] Lee H Dicker. Variance estimation in high-dimensional linear models. Biometrika, 101(2):269–284, 2014.

[11] 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.

[12] 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.

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

[14] 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.

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

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

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

[18] Jian Huang, Shuangge Ma, and Cun-Hui Zhang. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica, 18(4):1603, 2008.

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

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

[21] Piotr Pokarowski, Wojciech Rejchel, Agnieszka Soltys, Michal Frej, and Jan Mielniczuk. Improving lasso for model selection and prediction. arXiv preprint arXiv:1907.03025, 2019.

[22] Venkatesh Saligrama and Manqi Zhao. Thresholded basis pursuit: Lp algorithm for order-wise optimal support recovery for sparse and approximately sparse signals from noisy random measurements. IEEE Transactions on Information Theory, 57(3):1567–1586, 2011.

[23] Ulrike Schneider and Patrick Tardivel. The geometry of uniqueness and model selection of penalized estimators including slope, lasso, and basis pursuit. arXiv preprint arXiv:2004.09106, 2020.

[24] 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.

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

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

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

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

[29] 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.

[30] 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.

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

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

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

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