Regularization with the Smooth-Lasso procedure

Mohamed Hebiri

Regularization with the Smooth-Lasso procedure

Mohamed Hebiri*

Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris 7 - Diderot, UFR de Mathématiques,
175 rue de Chevaleret F-75013 Paris, France.

Abstract

We consider the linear regression problem. We propose the S-Lasso procedure to estimate the unknown regression parameters. This estimator enjoys sparsity of the representation while taking into account correlation between successive covariates (or predictors). The study covers the case when $p \gg n$, i.e. the number of covariates is much larger than the number of observations. In the theoretical point of view, for fixed p, we establish asymptotic normality and consistency in variable selection results for our procedure. When $p \geq n$, we provide variable selection consistency results and show that the S-Lasso achieved a Sparsity Inequality, i.e., a bound in term of the number of non-zero components of the oracle vector. It appears that the S-Lasso has nice variable selection properties compared to its challengers. Furthermore, we provide an estimator of the effective degree of freedom of the S-Lasso estimator. A simulation study shows that the S-Lasso performs better than the Lasso as far as variable selection is concerned especially when high correlations between successive covariates exist. This procedure also appears to be a good challenger to the Elastic-Net [36].

Keywords: Lasso, LARS, Sparsity, Variable selection, Regularization paths, Mutual coherence, High-dimensional data.
AMS 2000 subject classifications: Primary 62J05, 62J07; Secondary 62H20, 62F12.

*hebiri@math.jussieu.fr

1 Introduction

We focus on the usual linear regression model:

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where the design $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter and $\varepsilon_1, \ldots, \varepsilon_n$, are independent identically distributed (i.i.d.) centered Gaussian random variables with known variance σ^2. We wish to estimate β^* in the sparse case, that is when many of its unknown components equal zero. Thus only a subset of the design covariates $(\xi_j)_j$ is truly of interest where $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$. Moreover the case $p \gg n$ is not excluded so that we can consider p depending on n. In such a framework, two main issues arise: i) the interpretability of the resulting prediction; ii) the control of the variance in the estimation. Regularization is therefore needed. For this purpose we use selection type procedures of the following form:

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \left\{ \|Y - X\beta\|_n^2 + \text{pen}(\beta) \right\}, \tag{2}$$

where $X = (x_1', \ldots, x_n')'$, $Y = (y_1, \ldots, y_n)'$ and pen : $\mathbb{R}^p \to \mathbb{R}$ is a positive convex function called the penalty. For any vector $a = (a_1, \ldots, a_n)'$, we have adopted the notation $\|a\|_n^2 = n^{-1} \sum_{i=1}^n |a_i|^2$ (we denote by $< \cdot, \cdot >_n$ the corresponding inner product in $\mathbb{R}^n$). The choice of the penalty appears to be crucial. Although well-suited for variable selection purpose, Concave-type penalties ([12], [27] and [6]) are often computationally hard to optimize. Lasso-type procedures (modifications of the l_1 penalized least square (Lasso) estimator introduced by Tibshirani [25]) have been extensively studied during the last few years. Between many others, see [2, 4, 34] and references inside. Such procedures seem to respond to our objective as they perform both regression parameters estimation and variable selection with low computational cost. We will explore this type of procedures in our study.

In the paper, we propose a novel modification of the Lasso we call the *Smooth-lasso* (*S-lasso*) estimator. It is defined as the solution of the optimization problem (2) when the penalty function is a combination of the Lasso penalty (i.e., $\sum_{j=1}^p |\beta_j|$) and the l_2-fusion penalty (i.e., $\sum_{j=2}^p (\beta_j - \beta_{j-1})^2$). The l_2-fusion penalty was first introduced in [15]. We add it to the Lasso procedure in order to overcome the variable selection problems observed by the Lasso estimator. Indeed the Lasso estimator has good selection properties but fails in some situations. More precisely, in several works ([2, 16, 18, 29, 32, 34, 35] among others) conditions for the consistency in variable selection of the Lasso procedure are given. It was shown that the Lasso is

not consistent when high correlations exist between the covariates. We give similar consistency conditions for the S-Lasso procedure and show that it is consistent in variable selection in much more situations than the Lasso estimator. From a practical point of view, problems are also encountered when we solve the Lasso criterion with the Lasso modification of the LARS algorithm [10]. Indeed this algorithm tends to select only one representing covariates in each group of correlated covariates. We attempt to respond to this problem in the case where the covariates are ranked so that high correlations can exist between successive covariates. We will see through simulations that such situations support the use of the *S-lasso* estimator. This estimator is inspired by the *Fused-Lasso* [26]. Both S-Lasso and Fused-Lasso combine a l_1-penalty with a fusion term [15]. The fusion term is suggested to catch correlations between covariates. More relevant covariates can then be selected due to correlations between them. The main difference between the two procedures is that we use the l_2 distance between the successive coefficients (i.e., the l_2-fusion penalty) whereas the Fused-Lasso uses the l_1 distance (i.e., the l_1-fusion penalty: $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$). Hence, compared to the Fused-Lasso, we sacrifice sparsity between successive coefficients in the estimation of β^* in favor of an easier optimization due to the strict convexity of the l_2 distance. However, since sparsity is yet ensured by the Lasso penalty. The l_2-fusion penalty helps us to catch correlations between covariates. Consequently, even if there is no perfect match between successive coefficients our result are still interpretable. Moreover, when successive coefficients are significantly different, a perfect match seems to be not really adapted. In the theoretical point of view, The l_2 distance also helps us to provide theoretical properties for the S-Lasso which in some situations appears to outperforms the Lasso and the Elastic-Net [36], another Lasso-type procedure. Let us mention that variable selection consistency of the Fused-Lasso and the corresponding Fused adaptive Lasso has also been studied in [20] but in a different context from the one in the present paper. The result obtained in [20] are established not only under the sparsity assumption, but the model is also supposed to be *blocky*, that is the non-zero coefficients are represented in a block fashion with equal values inside each block.

Many techniques have been proposed to solve the weaknesses of the Lasso. The Fused-Lasso procedure is one of them and we give here some of the most popular methods; the Adaptive Lasso was introduced in [35], which is similar to the Lasso but with adaptive weights used to penalize each regression coefficient separately. This procedure reaches 'Oracles Properties' (i.e. consistency in variable selection and asymptotic normality). Another approach is used in the Relaxed Lasso [17] and aims to doubly-control the Lasso estimate: one parameter to control variable selection and the other to control shrinkage of the selected coefficients. To overcome

the problem due to the correlation between covariates, group variable selection has been proposed by Yuan and Lin [31] with the Group-Lasso procedure which selects groups of correlated covariates instead of single covariates at each step. A first step to the consistency study has been proposed in [1] and Sparsity Inequalities were given in [5]. Another choice of penalty has been proposed with the Elastic-Net [36]. It is in the same spirit that we shall treat the S-Lasso from a some theoretical point of view.

The paper is organized as follows. In the next section, we present one way to solve the S-Lasso problem with the attractive property of piecewise linearity of its regularization path. Section 3 gives theoretical performances of the considered estimator such as consistency in variable selection and asymptotic normality when $p \leq n$ whereas consistency in estimation and variable selection in the high dimensional case are considered in Section 4. We also give an estimate of the effective degree of freedom of the S-Lasso estimator in Section 5. Then, we provide a way to control the variance of the estimator by scaling in Section 6 where a connection with soft-thresholding is also established. A generalization and comparative study to the Elastic-Net is done in Section 7. We finally give experimental results in Section 8 showing the S-Lasso performances against some popular methods. All proofs are postponed to an Appendix section.

2 The S-Lasso procedure

As described above, we define the S-Lasso estimator $\hat{\beta}^{SL}$ as the solution of the optimization problem (2) when the penalty function is:

$$\text{pen}(\beta) = \lambda|\beta|_1 + \mu \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 , \tag{3}$$

where λ and μ are two positive parameters that control the smoothness of our estimator. For any vector $a = (a_1, \ldots, a_p)'$, we have used the notation $|a|_1 = \sum_{j=1}^{p} |a_j|$. Note that when $\mu = 0$, the solution is the Lasso estimator so that it appears as a special case of the S-Lasso estimator. Now we deal with the resolution of the S-Lasso problem (2)-(3) and its computational cost. From now on, we suppose w.l.o.g. that $X = (x_1, \ldots, x_n)'$ is standardized (that is $n^{-1} \sum_{i=1}^{n} x_{i,j}^2 = 1$ and $n^{-1} \sum_{i=1}^{n} x_{i,j} = 0$) and $Y = (y_1, \ldots, y_n)'$ is centered (that is $n^{-1} \sum_{i=1}^{n} y_i = 0$). The following lemma shows that the S-Lasso criterion can be expressed as a Lasso criterion by augmenting the data artificially.

Lemma 1. *Given the data set* (X, Y) *and* (λ, μ)*. Define the extended dataset* $(\widetilde{X}, \widetilde{Y})$ *by*

$$\widetilde{X} = \frac{1}{\sqrt{1+\mu}} \begin{pmatrix} X \\ \sqrt{n\mu}\mathbf{J} \end{pmatrix} \quad \textit{and} \quad \widetilde{Y} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ *is a vector of size* p *containing only zeros and* $\mathbf{J}$ *is the* $p \times p$ *matrix*

$$\mathbf{J} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & -1 & \ddots & \ddots & \vdots \\ 0 & 1 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix}. \tag{4}$$

Let $r = \lambda/\sqrt{1+\mu}$ *and* $b = \sqrt{1+\mu}\,\beta$*. Then the S-Lasso criterion can be written*

$$\left\| \widetilde{Y} - \widetilde{X} b \right\|_n^2 + r|b|_1.$$

Let $\hat{b}$ *be the minimizer of this Lasso-criterion, then*

$$\hat{\beta}^{SL} = \frac{1}{\sqrt{1+\mu}}\,\hat{b}.$$

This result is a consequence of simple algebra. Lemma 1 motivates the following comments on the S-Lasso procedure.

Remark 1 (*Regularization paths*)**.** *The S-Lasso modification of the LARS is an iterative algorithm. For a fixed* μ *(appearing* (3)*), it constructs at each step an estimator based on the correlation between covariates and the current residue. Each step corresponds to a value of* λ*. Then for a fixed* μ*, we get the evolution of the S-Lasso estimator coefficients values when* λ *varies. This evolution describes the regularization paths of the S-Lasso estimator which are piecewise linear [21]. This property implies that the S-Lasso problem can be solved with the same computational cost as the ordinary least square (OLS) estimate using the Lasso modification version of the LARS algorithm.*

Remark 2 (*Implementation*)**.** *The number of covariates that the LARS algorithm and its Lasso version can select is limited by the number of rows in the matrix* X*. Applied to the augmented data* $(\widetilde{X}, \widetilde{Y})$ *introduced in Lemma 1, the Lasso modification of the LARS algorithm is able to select all the* p *covariates. Then we are no longer limited by the sample size as for the Lasso [10].*

3 Theoretical properties of the S-Lasso estimator when $p \leq n$

In this section we introduce the theoretical results according to the S-Lasso with a moderate sample size ($p \leq n$). We first provide rates of convergence of the S-Lasso estimator and show how through a control on the regularization parameters we can establish root-n consistency and asymptotic normality. Then we look for variable selection consistency. More precisely, we give conditions under which the S-Lasso estimator succeeds in finding the set of the non-zero regression coefficients. We show that with a suitable choice of the tuning parameter (λ, μ), the S-Lasso is consistent in variable selection. All the results of this section are proved in Appendix A.

3.1 Asymptotic Normality

In this section, we allow the tuning parameters (λ, μ) to depend on the sample size n. We emphasize this dependence by adding a subscript n to these parameters. We also fix the number of covariates p. Let us note $\mathbb{I}(\cdot)$ the indicator function and define the sign function such that for any $x \in \mathbb{R}$, $\mathrm{Sgn}(x)$ equals 1, -1 or 0 respectively when x is bigger, smaller or equals 0. Knight and Fu [14] gave the asymptotic distribution of the Lasso estimator. We provide here the asymptotic distribution to the S-Lasso. Let $\mathbf{C}_n = n^{-1}X'X$, be Gram matrix, then

Theorem 1. *Given the data set (X, Y), assume the correlation matrix verifies*

$$\mathbf{C}_n \to \mathbf{C}, \quad \textit{when } n \to \infty,$$

in probability where $\mathbf{C}$ is a positive definite matrix. If there exists a sequence v_n such that $v_n \to 0$ and the regularization parameters verify $\lambda_n v_n^{-1} \to \lambda \geq 0$ and $\mu_n v_n^{-1} \to \mu \geq 0$. Then, if $(\sqrt{n}v_n)^{-1} \to \kappa \geq 0$, we have

$$v_n^{-1}(\hat{\beta}^{SL} - \beta^*) \xrightarrow{\mathcal{D}} \underset{u \in \mathbb{R}^p}{\mathrm{Argmin}}\, V(u), \quad \textit{when } n \to \infty,$$

where

$$\begin{aligned} V(u) = -2\kappa u^T W + u^T \mathbf{C} u \;&+\; \lambda \sum_{j=1}^{p} \left\{ u_j \,\mathrm{Sgn}(\beta_j^*)\mathbb{I}(\beta_j^* \neq 0) + |u_j|\,\mathbb{I}(\beta_j^* = 0) \right\} \\ &+\; 2\mu \sum_{j=2}^{p} \left\{ (u_j - u_{j-1})(\beta_j^* - \beta_{j-1}^*)\mathbb{I}(\beta_j^* \neq \beta_{j-1}^*) \right\}, \end{aligned}$$

with $W \sim \mathcal{N}(0, \sigma^2 \mathbf{C})$.

Remark 3. *When $\kappa \neq 0$ is a finite constant: in this case v_n^{-1} is $\mathcal{O}(\sqrt{n})$ so that the estimator $\hat{\beta}^{SL}$ is root-n consistent. Moreover when $\lambda = \mu = 0$, we obtain the following standard regressor asymptotic normality: $\sqrt{n}(\hat{\beta}^{SL} - \beta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2 \mathbf{C}^{-1})$. When $\kappa = 0$: in this case, the rate of convergence is slower than $\sqrt{n}$ so that we no longer have the optimal rate. Moreover the limit is not random anymore.*

Note first that the correlation penalty does not alter the asymptotic bias when successive regression coefficients are equal. We also remark that the sequence v_n must be chosen properly as it determines our convergence rate. We would like v_n to be as close as possible to $1/\sqrt{n}$. This sequence is calibrated by the user such that $\lambda_n/v_n \to \lambda$ and $\mu_n/v_n \to \mu$.

3.2 Consistency in variable selection

In this section, variable selection consistency of the S-Lasso estimator is considered. For this purpose, we introduce the following sparsity sets: $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$ and $\mathcal{A}_n = \{j : \hat{\beta}_j^{SL} \neq 0\}$. The set $\mathcal{A}^*$ consists of the non-zero coefficients in the vector of the oracle regression vector β^*. The set $\mathcal{A}_n$ consists of the non-zero coefficients in the S-Lasso estimator $\hat{\beta}_j^{SL}$ and is also called the active set of this estimator. Before stating our result, let us introduce some notations. For any vector $a \in \mathbb{R}^p$ and any set of indexes $\mathcal{B} \in \{1, \ldots, p\}$, denote by $a_{\mathcal{B}}$ the restriction of the vector a to the indexes in $\mathcal{B}$. In the same way, if we note $|\mathcal{B}|$ the cardinal of the set $\mathcal{B}$, then for any $s \times q$ matrix M, we use the following convention: i) $M_{\mathcal{B},\mathcal{B}}$ is the $|\mathcal{B}| \times |\mathcal{B}|$ matrix consisting of the lines and rows of M whose indexes are in $\mathcal{B}$; ii) $M_{.,\mathcal{B}}$ is the $s \times |\mathcal{B}|$ matrix consisting of the rows of M whose indexes are in $\mathcal{B}$; iii) $M_{\mathcal{B},.}$ is the $|\mathcal{B}| \times q$ matrix consisting of the lines of M whose indexes are in $\mathcal{B}$. Moreover, we define $\widetilde{J}$ the $p \times p$ matrix $\mathbf{J}'\mathbf{J}$ where $\mathbf{J}$ was defined in (4). Finally we define for $j \in \{1, \ldots, p\}$, the quantity $\Omega_j = \Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)$ by

$$\Omega_j = \mathbf{C}_{j,\mathcal{A}^*}(\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*} + \mu \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*})^{-1} \left(2^{-1} \operatorname{Sgn}(\beta_{\mathcal{A}^*}^*) + \frac{\mu}{\lambda} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta_{\mathcal{A}^*}^*\right) - \frac{\mu}{\lambda} \widetilde{J}_{j,\mathcal{A}^*} \beta_{\mathcal{A}^*}^*, \quad (5)$$

where $\mathbf{C}$ is defined as in Theorem 1. Now consider the following conditions: *for every* $j \in (\mathcal{A}^*)^c$

$$|\Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)| < 1, \quad (6)$$

$$|\Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)| \leq 1. \quad (7)$$

These conditions on the correlation matrix $\mathbf{C}$ and the regression vector $\beta_{\mathcal{A}^*}^*$ are the analogues respectively of the sufficient and necessary conditions derived for the Lasso ([35], [34] and [32]). Now we state the consistency results

Theorem 2. *If condition* (6) *holds, then for every couple of regularization parameters* (λ_n, μ_n) *such that* $\lambda_n \to 0$, $\lambda_n n^{1/2} \to \infty$ *and* $\mu_n \to 0$, *the S-Lasso estimator* $\hat{\beta}^{SL}$ *as defined in* (2)-(3) *is consistent in variable selection. That is*

$$\mathbb{P}(\mathcal{A}_n = \mathcal{A}^*) \to 1, \quad \text{when } n \to \infty.$$

Theorem 3. *If there exist sequences* (λ_n, μ_n) *such that* β^{SL} *converges to* β^* *and* $\mathcal{A}_n$ *converges to* $\mathcal{A}^*$ *in probability, then condition* (7) *is satisfied.*

We just have established necessary and sufficient conditions to the selection consistency of the S-Lasso estimator. Due to the assumptions needed in Theorem 2 (more precisely $\lambda_n n^{1/2} \to \infty$), root-$n$ consistency and variable selection consistency cannot be treated here simultaneously. We may want to know if the S-Lasso estimator can be consistent with a slower rate than $n^{1/2}$ and consistent in variable selection in the same time.

Remark 4. *Here are special cases of conditions* (6)- (7).
When $\mu = 0$ *and* $\mu/\lambda = 0$*: these conditions are exactly the sufficient and necessary conditions of the Lasso estimator. In this case Yuan and Lin [32] showed that the condition* (6) *becomes necessary and sufficient for the Lasso estimator consistency in variable selection.*
When $\mu = 0$ *and* $\mu/\lambda = \gamma \neq 0$*: in this case, condition* (6) *becomes*

$$\sup_{j \in (\mathcal{A}^*)^c} |\mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} (2^{-1} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*}) + \gamma \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta^*_{\mathcal{A}^*}) - \gamma \widetilde{J}_{j,\mathcal{A}^*} \beta^*_{\mathcal{A}^*}| < 1.$$

Here a good calibration of γ *leads to consistency in variable selection:*

- *if* $(\mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*} > 0$, *then* γ *must be chosen between* $-\dfrac{1 + 2^{-1} \mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*})}{(\mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*}}$ *and* $\dfrac{1 - 2^{-1} \mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \operatorname{Sgn}(\beta^*_{\mathcal{A}^*})}{(\mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*}}$.
- *if* $(\mathbf{C}_{j,\mathcal{A}^*} \mathbf{C}^{-1}_{\mathcal{A}^*,\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*}) \beta^*_{\mathcal{A}^*} < 0$, *then* γ *must be chosen between the same quantities but with inversion in their order.*

When $\mu \neq 0$ *and* $\mu/\lambda = \gamma \neq 0$*: this case is similar to the previous. In addition, it allows to have another control on the condition through a calibration with* μ*, so that condition* (6) *can be satisfied with a better control.*

We conclude that if we sacrifice the optimal rate of convergence (i.e. root-n consistency), we are able through a proper choice of the tuning parameters (λ_n, μ_n)

to get consistency in variable selection. Note that Zou [35] showed that the Lasso estimator cannot be consistent in variable selection even with a slower rate of convergence than $\sqrt{n}$. He then added weights to the Lasso (i.e. the adaptive Lasso estimator) in order to get Oracles Properties (that is both asymptotic normality and variable selection consistency). Note that we can easily adapt techniques used in the adaptive Lasso to provide a weighted S-Lasso estimator which achieved the Oracles Properties.

4 Theoretical results when dimension p is larger than sample size n

In this section, we propose to study the performance of the S-Lasso estimator in the high dimensional case. In particular, we provide a non-asymptotic bound on the squared risk. We also provide bound on the estimation risk under the sup-norm (i.e., the l_∞-norm: $\|\hat{\beta}^{SL} - \beta^*\|_\infty = \sup_j |\hat{\beta}_j^{SL} - \beta_j^*|$). This last result helps us to provide a variable selection consistent estimator obtained through thresholding the S-Lasso estimator. The results of this section are proved in Appendix B.

4.1 Sparsity Inequality

Now we establish a Sparsity Inequality (SI) achieved by the S-Lasso estimator, that is a bound on the squared risk that takes into account the sparsity of the oracle regression vector β^*. More precisely, we prove that the rate of convergence is $|\mathcal{A}^*| \log(n)/n$. For this purpose, we need some assumptions on the Gram matrix $\mathbf{C}_n$ which is normalized in our setting. Recall that $\xi_j = (x_{1,j}, \ldots, x_{n,j})'$. Then we define the regularization parameters λ_n and μ_n in the following forms:

$$\lambda_n = \kappa_1 \sigma \sqrt{\frac{\log(p)}{n}}, \quad \text{and} \quad \mu_n = \kappa_2 \sigma^2 \frac{\sqrt{\log(p)}}{n}, \tag{8}$$

where $\kappa_1 > 2\sqrt{2}$ and κ_2 is positive constants. Let us define the maximal correlation quantity $\rho_1 = \max_{j \in \mathcal{A}^*} \max_{\substack{k \in \{1,\ldots,p\} \\ k \neq j}} |(\mathbf{C}_n)_{j,k}|$. Using these notations, we formulate the following assumptions:

- *Assumption (A1). The true regression vector β^* is such that there exists a finite constant L_1 such that:*

$$\beta_{\mathcal{A}^*}^{*\prime} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta_{\mathcal{A}^*}^* \leq L_1 \log(p) |\mathcal{A}^*|, \tag{9}$$

where $\widetilde{J} = \mathbf{J}'\mathbf{J}$ *where* $\mathbf{J}$ *was defined in* (4).

- *Assumption (A2). We have:*

$$\rho_1 \leq \frac{1}{16|\mathcal{A}^*|}. \tag{10}$$

Note that Assumption (A1) is not restrictive. A sufficient condition is that the larger non-zero component of $\beta^*_{\mathcal{A}^*}$ is bounded by $L_1 \log(p)$ which can be very large. Assumption (A2) is the well-known coherence condition considered in [3], which has been introduced in [7]. Most of SIs provided in the literature use such a condition. We refer to [3] for more details.
Theorem 4 below provides an upper bound for the squared error of the estimator $\hat{\beta}^{SL}$ and for its l_1 estimation error which takes into account the sparsity index $|\mathcal{A}^*|$.

Theorem 4. *Let us consider the linear regression model* (1). *Let* $\hat{\beta}^{SL}$ *be S-Lasso estimator. Let* $\mathcal{A}^*$ *be the sparsity set. Suppose that* $p \geq n$ *(and even* $p \gg n$*). If Assumptions (A1)–(A2) hold, then with probability greater than* $1 - u_{n,p}$*, we have*

$$\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 \leq c_2 \frac{\log(p)|\mathcal{A}^*|}{n}, \tag{11}$$

and

$$|\hat{\beta}^{SL} - \beta^*|_1 \leq c_1 \sqrt{\frac{\log(p)}{n}} |\mathcal{A}^*|, \tag{12}$$

where $c_2 = (16\kappa_1^2 + L_1\kappa_2)\sigma^2$, $c_1 = (16\kappa_1 + L_1\kappa_1^{-1}\kappa_2)\sigma$ *and where* $u_{n,p} = p^{1-\kappa_1^2/8}$ *with* κ_1 *and* κ_2*, the constants appearing in* (8).

The proof of Theorem 4 is based on the 'argmin' definition of the estimator $\hat{\beta}^{SL}$ and some technical concentration inequalities. Similar bounds were provided for the Lasso estimator in [4]. Let us mention that the constants c_1 and c_2 are not optimal. We focused our attention on the dependency on n (and then on p and $|\mathcal{A}^*|$). It turns out that our results are near optimal. For instance, for the l_2 risk, the S-Lasso estimator reaches nearly the optimal rate $\frac{|\mathcal{A}^*|}{n} \log(\frac{p}{|\mathcal{A}^*|} + 1)$ up to a logarithmic factor [3, Theorem 5.1].

4.2 Sup-norm bound and variable selection

Now we provide a bound on the sup-norm $\|\beta^* - \hat{\beta}^{SL}\|_\infty$. Thanks to this result, one may be able to define a rule in order to get a variable selection consistent estimator

when $p \gg n$. That is, we can construct an estimator which succeeds to recover the support of β^* in high dimensional settings.
Small modifications are to be imposed to provide our selection results in this section. Let K_n be the symmetric $p \times p$ matrix defined by $K_n = \mathbf{C}_n + \mu_n \widetilde{J}$. Instead of Assumption (A2), we will consider the following

- *Assumption (A3). We assume that*

$$\max_{\substack{j,\, k \in \{1,\ldots,p\} \\ k \neq j}} |(K_n)_{j,k}| \leq \frac{1}{16|\mathcal{A}^*|}.$$

Remark 5. *Note that the matrix $\widetilde{J}$ is tridiagonal with its off-diagonal terms equal to -1. If we do not consider the diagonal terms, we remark that $\mathbf{C}_n$ and K_n differ only in the terms on the second diagonals (i.e., $(K_n)_{j-1,j} \neq (\mathbf{C}_n)_{j-1,j}$ for $j = 2, \ldots, p$ as soon as $\mu_n \neq 0$). Then, as we do not consider the diagonal terms in Assumptions (A2) and (A3), they differ only in the restriction they impose to terms on the second diagonals. Terms in the second diagonals of $\mathbf{C}_n$ correspond to correlations between successive covariates. Then when high correlations exist between successive covariates, a suitable choice of μ_n makes Assumption (A3) satisfied while Assumptions (A2) does not. Hence, Assumption (A3) fits better with setup considered in the paper.*

In the sequel, a convenient choice of the tuning parameter μ_n is $\mu_n = \kappa_3 \sigma / \sqrt{n \log(p)}$, where $\kappa_3 > 0$ is a constant. Moreover, from Assumption (A1), we have $\beta^{*'}_{\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} \beta^*_{\mathcal{A}^*} \leq L_1 \log(p) |\mathcal{A}^*|$. This inequality guarantees the existence of a constant $L_2 > 0$ such that $\|\widetilde{J}\beta^*\|_\infty \leq L_2 \log(p)$.

Theorem 5. *Let us consider the linear regression model* (1). *Let $\lambda_n = \kappa_1 \sigma \sqrt{\log(p)/n}$ and $\mu_n = \kappa_3 \sigma / \sqrt{n \log(p)}$ with $\kappa_1 > 2\sqrt{2}$ and $\kappa_3 > 0$. Suppose that $p \geq n$ (and even $p \gg n$). Under Assumptions (A1) and (A3) and with probability greater than $1 - p^{1 - \frac{\kappa_1^2}{8}}$, we have*

$$\|\hat{\beta}^{SL} - \beta^*\|_\infty \leq \tilde{c} \sqrt{\frac{\log(p)}{n}},$$

where $\tilde{c}$ equals to

$$\frac{1}{1 + \frac{B\sigma}{n}} \left(\frac{3}{4} + \frac{1}{\alpha - 1} + \frac{4L_1 B}{9\alpha^2 A^2} + \frac{2L_1 B}{3\alpha A^2} + \sqrt{\frac{2L_1 B}{3\alpha(\alpha-1)A^2} + \frac{8L_1\, L_2 B^2}{9\alpha(\alpha-1)A^4}\lambda_n} + (\frac{4L_2 B}{3A^2} + \frac{L_2 B}{A^2})\lambda_n \right).$$

Note that the leading term in $\tilde{c}$ is $\frac{3}{4}+\frac{1}{\alpha-1}+\frac{4L_1B}{9\alpha^2A^2}+\frac{2L_1B}{3\alpha A^2}+\sqrt{\frac{2L_1B}{3\alpha(\alpha-1)A^2}}$. One may find back the result obtained for the Lasso by setting L_1 to zero [16]. Secondly, the calibration of μ_n aims at making the convergence rate under the sup-norm equal to $\sqrt{\log(p)/n}$. On one hand, the proof of Theorem 5 allows us to choose this parameter with a faster convergence to zero without affecting the rate of convergence. On the other hand, a more restrictive Assumption (A1) on $\beta^{*'}_{\mathcal{A}^*}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta^*_{\mathcal{A}^*}$ and $\|\widetilde{J}\beta^*\|_\infty$ can be formulated in order to make μ_n converge slower to zero. If we let $\beta^{*'}_{\mathcal{A}^*}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*}\beta^*_{\mathcal{A}^*} \leq L_1|\mathcal{A}^*|$ in Assumption (A1), we can set μ_n as $\mathcal{O}(\sqrt{\log(p)/n})$, the slower convergence we can get for μ_n.
Let us now provide a consistent version of the S-Lasso estimator. Consider $\hat{\beta}^{ThSL}$, the thresholded S-Lasso estimator defined by $\hat{\beta}^{ThSL}=\hat{\beta}^{SL}\mathbb{I}(\hat{\beta}^{SL}\geq\tilde{c}\sqrt{\log(p)/n})$ where $\tilde{c}$ is given in Theorem 5. This estimator consists of the S-Lasso estimator with its small coefficients reduced to zero. We then enforce the selection property of the S-Lasso estimator. Variable selection consistency of this estimator is established under one more restriction:

- *Assumption (A4). The smallest non-zero coefficient of β^* is such that there exists a constant $c_l>0$ with*

$$\min_{j\in\mathcal{A}^*}|\beta^*_j| > c_l\sqrt{\frac{\log(p)}{n}}.$$

Assumption (A4) bounds from below the smallest regression coefficient in β^*. This is a common assumption to provide sign consistency in the high dimensional case. This condition appears in [19, 29, 33, 34] but with a larger (in term of sample size n dependence) and then more restrictive threshold. We refer to [16] for a longer discussion. An equivalent lower bound in the oracle regression coefficients can be found in [2, 16]. With this new assumption, we can state the following sign consistency result.

Theorem 6. *Let us consider the thresholded S-Lasso estimator $\hat{\beta}^{ThSL}$ as described above. Choose moreover $\lambda_n=\kappa_1\sigma\sqrt{\log(p)/n}$ and $\mu_n=\kappa_3\sigma/\sqrt{n\log(p)}$ with the positive constants $\kappa_1>2\sqrt{2}$ and κ_3. Under Assumptions (A1), (A3) and (A4), if $c_l>2\tilde{c}$ with $\tilde{c}$ is given by Theorem 5, with probability greater than $1-p^{1-\frac{\kappa_1^2}{8}}$, we have*

$$\mathrm{Sgn}(\hat{\beta}^{ThSL})=\mathrm{Sgn}(\beta^*), \tag{13}$$

and then as $n\to+\infty$

$$\mathbb{P}(\mathrm{Sgn}(\hat{\beta}^{ThSL})=\mathrm{Sgn}(\beta^*))\to 1. \tag{14}$$

Remark 6. *As observed in Remark 5, Assumption (A3) is more easily satisfied when correlation exists between successive covariates. Then in situations where the correlation matrix* $\mathbf{C}_n$ *is tridiagonal with its off-diagonal terms equal to* δ *with* $\delta \in [0,1]$*, the constant* κ_3 *appearing in the definition of* μ_n *can be adjusted in order to get Assumption (A3) satisfied.*

5 Model Selection

As already said [Remark 1 in Section 2], each step of the S-Lasso version of the LARS algorithm provides an estimator of β^*. In this section, we are interested in the choice of the best estimator according to its prediction accuracy. For a new $n \times p$ matrix x_{new} of instances (independent of X), denote $\hat{y}^{SL} = \hat{\beta}^{SL} x_{new}$ the estimator of its unknown response value y_{new} and $m = \mathbb{E}(y_{new}|x_{new})$. We aim to minimize the true risk $\mathbb{E}\left\{\|m - \hat{y}^{SL}\|_n^2\right\}$. First, we easily obtain

$$\mathbb{E}\left\{\|m - \hat{y}^{SL}\|_n^2\right\} = \mathbb{E}\{\|Y - \hat{y}^{SL}\|_n^2 - \sigma^2 + 2n^{-1}\sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i^{SL})\},$$

where the expectation is taken over the random variable Y. The last term in this equation was called *optimism* [9]. Moreover, Tibshirani [25] links this quantity to the *degree of freedom* $\mathrm{df}(\hat{y}^{SL})$ of the estimator $\hat{y}^{SL}$, so that the above equality becomes

$$\mathbb{E}\left\{\|m - \hat{y}^{SL}\|_n^2\right\} = \mathbb{E}\left\{\|Y - \hat{y}^{SL}\|_n^2 - \sigma^2 + 2n^{-1}\,\mathrm{df}(\hat{y}^{SL})\sigma^2\right\}. \tag{15}$$

This final expression involves the degree of freedom which is unknown. Various methods exist to estimate the degree of freedom as bootstrap [11] or data perturbation methods [24]. We give an explicit form to the degree of freedom in order to reduce the computational cost as in [10] and [37].

Degrees of freedom: the degree of freedom is a quantity of interest in model selection. Before stating our result, let us introduce some useful properties about the regularization paths of the S-Lasso estimator:
Given a response Y, and a regularization parameter $\mu \geq 0$, there is a finite sequence $0 = \lambda^{(K)} < \lambda^{(K-1)} < \ldots < \lambda^{(0)}$ such that $\hat{\beta}^{SL} = \mathbf{0}$ for every $\lambda \geq \lambda^{(0)}$. In this notation, superscripts correspond to the steps of the S-Lasso version of the LARS algorithm.
Given a response Y, and a regularization parameter $\mu \geq 0$, for $\lambda \in (\lambda^{(k+1)}, \lambda^{(k)})$, the same covariates are used to construct the estimator. Let us note $\mathcal{A}_\zeta$ the active set for a fixed couple $\zeta = (\lambda, \mu)$ and $X_{.,\mathcal{A}_\zeta}$ the corresponding design matrix.

In what follows, we will use the subscript ζ to emphasize the fact that the considered quantity depends on ζ.

Theorem 7. *For fixed $\mu \geq 0$ and $\lambda > 0$, an unbiased estimate of the effective degree of freedom of the S-Lasso estimate is given by*

$$\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) = \mathrm{Tr}\left[X_{.,\mathcal{A}_\zeta}\left(X'_{.,\mathcal{A}_\zeta}X_{.,\mathcal{A}_\zeta} + \mu\widetilde{J}_{\mathcal{A}_\zeta,\mathcal{A}_\zeta}\right)^{-1}X'_{.,\mathcal{A}_\zeta}\right],$$

where $\widetilde{J} = \mathbf{J}'\mathbf{J}$ is defined by

$$\widetilde{J} = \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}. \tag{16}$$

As the estimation given in Theorem 7 has an important computational cost, we propose the following estimator of the degree of freedom of the S-Lasso estimator:

$$\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) = \frac{|\mathcal{A}_\zeta| - 2}{1 + 2\mu} + \frac{2}{1+\mu}, \tag{17}$$

which is very easy to compute. Let $\mathbf{I}_s$ be the $s \times s$ identity matrix where s is an integer. We found the former approximation of the degree of freedom under the orthogonal covariance matrix assumption (that is $n^{-1}X'X = \mathbf{I}_p$). Moreover we approximate the matrix $(\mathbf{I}_{|\mathcal{A}_\lambda|} + \mu\widetilde{J}_{\mathcal{A}_\lambda,\mathcal{A}_\lambda})$ by the diagonal matrix with $1+\mu$ in the first and the last terms, and $1+2\mu$ in the others.

Remark 7 (*Comparison to the Lasso and the Elastic-Net)***.** *A similar work leads to an estimation of the degree of freedom of the Lasso: $\widehat{\mathrm{df}}(\hat{y}_\zeta^L) = |\mathcal{A}_\zeta|$ and to an estimation of the degree of freedom of the Elastic-Net estimator: $\widehat{\mathrm{df}}(\hat{y}_\zeta^{EN}) = |\mathcal{A}_\zeta|/(1+\mu)$. These approximations of the degrees of freedom provide the following comparison for a fixed ζ: $\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) \leq \widehat{\mathrm{df}}(\hat{y}_\zeta^{EN}) \leq \widehat{\mathrm{df}}(\hat{y}_\zeta^{L})$. A conclusion is that the S-Lasso estimator is the one which penalizes the smaller models, and the Lasso estimator the larger. As a consequence, the S-Lasso estimator should select larger models than the Lasso or the Elastic-Net estimator.*

6 The Normalized S-Lasso estimator

In this section, we look for a scaled S-Lasso estimator which would have better empirical performance than the original S-Lasso presented above. The idea behind this study is to better control shrinkage. Indeed, using the S-Lasso procedure (2)-(3) induces double shrinkage: one using the Lasso penalty and the other using the fusion penalty. We want to undo the shrinkage implied by the fusion penalty as shrinkage is already ensured by the Lasso penalty. We then suggest to study the S-Lasso criterion (2)-(3) without the Lasso penalty (i.e. with only the l_2-fusion penalty) in order to find the constant we have to scale with.

Define

$$\tilde{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \|Y - X\beta\|_n^2 + \mu \sum_{j=2}^{p} (\beta_j - \beta_{j-1})^2 .$$

We easily obtain $\tilde{\beta} = ((X'X)/n + \mu\widetilde{J})^{-1}(X'Y)/n := \mathbf{L}^{-1}(X'Y)/n$ where $\widetilde{J}$ is given by (16). Moreover as the design matrix X is standardized, the symmetric matrix $\mathbf{L}$ can be written

$$\mathbf{L} = \begin{pmatrix} 1+\mu & \frac{\xi_1'\xi_2}{n} - \mu & \frac{\xi_1'\xi_3}{n} & \dots & \frac{\xi_1'\xi_p}{n} \\ & 1+2\mu & n^{-1}\xi_2'\xi_3 - \mu & \dots & \vdots \\ & & \ddots & \ddots & \frac{\xi_{p-2}'\xi_p}{n} \\ & & & 1+2\mu & \frac{\xi_{p-1}'\xi_p}{n} - \mu \\ & & & & 1+\mu \end{pmatrix}.$$

In order to get rid of the shrinkage due to the fusion penalty, we force $\mathbf{L}$ to have ones (or close to a diagonal of ones) in its diagonal elements. Then we scale the estimator $\tilde{\beta}$ by a factor c. Here are two choice we will use in the following of the paper: i) the first is $c = 1 + \mu$ so that the first and the last diagonal elements of $\mathbf{L}^{-1}$ become equal to one; ii) the second is $c = 1 + 2\mu$ which offers the advantage that all the diagonal elements of $\mathbf{L}^{-1}$ become equal to one except the first and the last. This second choice seems to be more appropriate to undo this extra shrinkage and specially in high dimensional problem.

We first give a generalization of Lemma 1.

Lemma 2. *Given the dataset (X, Y) and (λ_1, μ). Define the augmented dataset $(\widetilde{X}, \widetilde{Y})$ by*

$$\widetilde{X} = \nu_1^{-1} \begin{pmatrix} X \\ \sqrt{n\mu}\mathbf{J} \end{pmatrix} \quad \text{and} \quad \widetilde{Y} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix},$$

where ν_1 is a constant which depends only on μ and $\mathbf{J}$ is given by (4)*. Let $r = \lambda/\nu_1$ and $b = (\nu_2/c)\beta$ where ν_2 is a constant which depends only on μ, and c is the scaling constant which appears in the previous study. Then the S-Lasso criterion can be written*

$$\left\|\widetilde{Y} - \widetilde{X}b\right\|_n^2 + r|b|_1. \tag{18}$$

Let $\hat{b}$ be the minimizer of this Lasso-criterion, then we define the Scaled Smooth Lasso (SS-Lasso) by

$$\hat{\beta}^{SSL} = \hat{\beta}^{SSL}(\nu_1, \nu_2, c) = (c/\nu_2)\,\hat{b}.$$

Moreover, let $\widetilde{J} = \mathbf{J}'\mathbf{J}$. Then we have

$$\hat{\beta}^{SSL} = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}} \left\{ \frac{\nu_2}{\nu_1}\beta' \left(\frac{\frac{X'X}{n} + \mu\widetilde{J}}{c} \right) \beta - 2\frac{Y'X}{n}\beta + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{19}$$

Equation (19) is only a rearrangement of the Lasso criterion (18). The SS-Lasso expression (19) emphasizes the importance of the scaling constant c. In a way, the SS-Lasso estimator stabilizes the Lasso estimator $\hat{\beta}^L$ (criterion (18) based in (X, Y) instead of $(\widetilde{X}, \widetilde{Y})$) as we have

$$\hat{\beta}^L = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}} \left\{ \beta' \left(\frac{X'X}{n} \right) \beta - 2\frac{Y'X}{n}\beta + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

The choice of ν_1 and ν_2 should be linked to this scaling constant c in order to get better empirical performances and to have less parameters to calibrate. Let us define some specific cases. i) *Case 1: When $\nu_1 = \nu_2 = \sqrt{1+\mu}$ and $c = 1$*: this is the "original" S-Lasso estimator as seen in Section 2. ii) *Case 2: When $\nu_1 = \nu_2 = \sqrt{1+\mu}$ and $c = 1 + \mu$*: we call this scaled S-Lasso estimator Normalized Smooth Lasso (NS-Lasso) and we note it $\hat{\beta}^{NSL}$. In this case, we have $\hat{\beta}^{NSL} = (1 + \mu\hat{\beta}^{SL})$. iii) *Case 3: When $\nu_1 = \nu_2 = \sqrt{1+2\mu}$ and $c = 1 + 2\mu$*: we call this scaled version Highly Normalized Smooth Lasso (HS-Lasso) and we note it $\hat{\beta}^{HSL}$.

Others choices are possible for ν_1 and ν_2 in order to better control shrinkage. For instance we can consider a compromise between the NS-Lasso and the HS-Lasso by defining $\nu_1 = 1 + \mu$ and $\nu_2 = 1 + 2\mu$.

Remark 8 (*Connection with Soft Thresholding*)**.** *Let us consider the limit case of the NS-Lasso estimator. Note $\hat{\beta}^{NSL}_{\infty} = \lim_{\mu\to\infty} \hat{\beta}^{NSL}$, then using* (19)*, we have*

$$\hat{\beta}^{NSL}_{\infty} = \underset{\beta}{\text{Argmin}}\{\beta'\beta - 2Y'X\beta + \lambda|\beta|_1\}.$$

As a consequence, $(\hat{\beta}_{\infty}^{NSL})_j = \left(|Y'\xi_j| - \frac{\lambda}{2}\right)_+ \mathrm{Sgn}(Y'\xi_j)$ *which is the Univariate Soft Thresholding [8]. Hence, when* $\mu \to \infty$*, the NS-Lasso works as if all the covariates were independent. The Lasso, which corresponds to the NS-Lasso when* $\mu = 0$*, often fails to select covariates when high correlations exist between relevant and irrelevant covariates. It seems that the NS-Lasso is able to avoid such problem by increasing* μ *and working as if all the covariates were independent. Then for a fixed* λ*, the control of the regularization parameter* μ *appears to be crucial. When we vary it, the NS-Lasso bridges the Lasso and the Soft Thresholding.*

7 Extension and comparison

All results obtained in the present paper can be generalized to all penalized least square estimators for which the penalty term can be written as:

$$\mathrm{pen}(\beta) = \lambda|\beta|_1 + \beta' M \beta, \tag{20}$$

where M is $p \times p$ matrix. In particular, our study can be extended for instance to the Elastic-Net estimator with the special choice $M = \mathbf{I}_p$. Such an observation underlines the superiority of the S-Lasso estimator on the Elastic-Net in some situations. Indeed, let us consider the variable selection consistency in the high dimensional setting (cf. Section 4.2). Regarding the Elastic-Net, Assumption (A3) becomes

- *Assumption (A3-EN). We assume that*

$$\max_{\substack{j,\, k \in \{1,\dots,p\} \\ k \neq j}} |(\mathbf{C}_n)_{j,k} + \mu_n \mathbf{I}_p| \leq \frac{1}{16|\mathcal{A}^*|}. \tag{21}$$

Since the identity matrix is diagonal and since the maximum in (21) is taken over indexes $k \neq j$, condition (21) reduces to $\max_{\substack{j,\, k \in \{1,\dots,p\} \\ k \neq j}} |(\mathbf{C}_n)_{j,k}| \leq \frac{1}{16|\mathcal{A}^*|}$. This makes Assumption (A3-EN) similar to the assumption needed to get the variable selection consistency of the Lasso estimator [2]. Hence, we get no gain to use the Elastic-Net in a variable selection consistency point of view in our framework. This ables us to think that the S-Lasso outperforms the Elastic-Net at least on examples as the one in Remark 6. Recently, Jia and Yu [13] studied the variable selection consistency of the Elastic-Net under an assumption called *Elastic Irrepresentable Condition*:

- *(EIC). There exists a positive constant* θ *such that for any* $j \in (\mathcal{A}^*)^c$

$$|\mathbf{C}_{j,\mathcal{A}^*}(\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*} + \mu \mathbf{I}_{\mathcal{A}^*})^{-1} \left(2^{-1}\,\mathrm{Sgn}(\beta^*_{\mathcal{A}^*}) + \frac{\mu}{\lambda}\beta^*_{\mathcal{A}^*}\right)| \leq 1 - \theta.$$

This condition can be seen as a generalization of the *Irrepresentable Condition* involved in the Lasso variable selection consistency.
Let us discuss how the two assumptions can be compared in the case $p \gg n$. First, note that Assumption (A3-EN), as well as EIC suggests low correlations between covariates. Moreover Assumption (A1), (A4) and (A3-EN) seem more restrictive than EIC as all the correlations are constrained in (21). However, EIC is harder to interpret in term of the coefficients of the regression vector β^*. It also depends on the sign of β^*. The main difference is that the consistency result in the present paper holds uniformly on the solutions of the Elastic Net criterion while the result from [13] hinges upon the existence of a consistent solution for variable selection. Obviously, this is more restrictive as we are certain to provide the sign-consistent solution under the EIC. Finally, we have also provided results on the sup-norm and sparsity inequalities on the squared risk of our estimators. Such results are new for estimators defined with the penalty (20), including the S-Lasso and the Elastic-Net.

8 Experimental results

In the present section we illustrate the good prediction and selection properties of the NS-Lasso and the HS-Lasso estimators. For this purpose, we compare it to the Lasso and the Elastic-Net. It appears that S-Lasso is a good challenger to the Elastic-Net [36] even when large correlations between covariates exist. We further show that in most cases, our procedure outperforms the Elastic-Net and the Lasso when we consider the ratio between the relevant selected covariates and irrelevant selected covariates.

Simulations:
Data. Four simulations are generated according to the linear regression model

$$y = x\beta^* + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1),\ x = (\xi_1, \ldots, \xi_p) \in \mathbb{R}^p.$$

The first and the second examples were introduced in the original Lasso paper [25]. The third simulation creates a grouped covariates situation. It was introduced in [36] and aims to point the efficiency of the Elastic-Net compared to the Lasso. The last simulation introduces large correlation between successive covariates.

(a) In this example, we simulate 20 observations with 8 covariates. The true regression vector is $\beta^* = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ so that only three covariates are truly relevant. Let $\sigma = 3$ and the correlation between ξ_j and ξ_k such that $\mathrm{Cov}(\xi_j, \xi_k) = 2^{-|j-k|}$.

(b) The second example is the same as the first one, except that we generate 50 observations and that $\beta_j^* = 0.85$ for every $j \in \{1, \ldots, 8\}$ so that all the covariates are relevant.

(c) In the third example, we simulate 50 data with 40 covariates. The true regression vector is such that $\beta_j^* = 3$ for $j = 1, \ldots, 15$ and $\beta_j^* = 0$ for $j = 16, \ldots, 40$. Let $\sigma = 15$ and the covariates generated as follows:

$$\begin{aligned}
\xi_j &= Z_1 + \varepsilon_j, \quad Z_1 \sim \mathcal{N}(0,1), \qquad j = 1, \ldots, 5, \\
\xi_j &= Z_2 + \varepsilon_j, \quad Z_2 \sim \mathcal{N}(0,1), \qquad j = 6, \ldots, 10, \\
\xi_j &= Z_3 + \varepsilon_j, \quad Z_3 \sim \mathcal{N}(0,1), \qquad j = 11, \ldots, 15,
\end{aligned}$$

where ε_j, $j = 1, \ldots, 15$, are i.i.d. $\mathcal{N}(0, 0.01)$ variables. Moreover for $j = 16, \ldots, 40$, the ξ_j's are i.i.d $\mathcal{N}(0,1)$ variables.

(d) In the last example, we generate 50 data with 30 covariates. The true regression vector is such that

$$\begin{aligned}
\beta_j &= \quad 3 - 0.1j \qquad && j = 1, \ldots, 10, \\
\beta_j &= -5 + 0.3j \qquad && j = 20, \ldots, 25, \\
\beta_j &= \quad 0 \qquad && \text{for the others } j.
\end{aligned}$$

The noise is such that $\sigma = 9$ and the correlations are such that $\mathrm{Cov}(\xi_j, \xi_k) = \exp\left(-\frac{|j-k|}{2}\right)$ for $(j,k) \in \{11, \ldots, 25\}^2$ and the others covariates are i.i.d. $\mathcal{N}(0,1)$, also independent from $\xi_{11}, \ldots, \xi_{25}$. In this model there are big correlation between relevant covariates and even between relevant and irrelevant covariates.

Validation. The selection of the tuning parameters λ and μ is based on the minimization of a BIC-type criterion [22]. For a given $\hat{\beta}$ the associated BIC error is defined as:

$$\mathrm{BIC}(\hat{\beta}) = \|Y - X\hat{\beta}\|_n^2 + \frac{\log(n)\sigma^2}{n}\widehat{\mathrm{df}}(\hat{\beta}),$$

where $\widehat{\mathrm{df}}(\hat{\beta})$ is given by (17) if we consider the S-Lasso and denotes its analogous quantities if we consider the Lasso or the Elastic-Net. Such a criterion provides an

Method	Example (a)	Example (b)	Example (c)	Example (d)
Lasso	3.8 [±0.1]	6.5 [±0.1]	6 [±0.1]	18.4 [±0.2]
E-Net	4.9 [±0.1]	6.9 [±0.1]	15.9 [±0.1]	20.5 [±0.2]
NS-Lasso	3.9 [±0.1]	6.5 [±0.1]	15.3 [±0.2]	18.9 [±0.2]
HS-Lasso	3.5 [±0.1]	5.9 [±0.1]	15 [±0.1]	18.1 [±0.2]

Table 1: Mean of the number of non-zero coefficients [and its standard error] selected respectively by the Lasso, the Elastic-Net (E-Net), the Normalized Smooth Lasso (NS-Lasso) and the Highly Smooth Lasso (HS-Lasso) procedures.

accurate estimator which enjoys good variable selection properties ([23] and [30]). In simulation studies, for each replication, we also provide the Mean Square Error (MSE) of the selected estimator on a new and independent dataset with the same size as training set (that is n). This gives an information on the robustness of the procedures.

Interpretations. All the results exposed here are based on 200 replications. Figure 1 and Figure 2 give respectively the BIC error and the test error of the considered procedures in each example. According to the selection part, Figure 3 shows the frequencies of selection of each covariate for all the procedures, and Table 1 shows the mean of the number of non-zeros coefficients that each procedure selected. Finally for each procedure, Table 2 gives the ratio between the number of relevant covariates and the number of noise covariates that the procedures selected. Let us call SNR this ratio. Then we can express this ratio as

$$\text{SNR} = \frac{\sum_{j \in \mathcal{A}_n} \mathbb{I}(j \in \mathcal{A}^*)}{\sum_{j \in \mathcal{A}_n} \mathbb{I}(j \notin \mathcal{A}^*)}.$$

This is a good indication of the selection power of the procedures.

As the Lasso is a special case of the S-Lasso and the Elastic-Net, the Lasso BIC error (Figure 1) is always larger than the BIC error for the other methods. These two seem to have equivalent BIC errors. When considering the test error (Figure 2), it seems again that all the procedures are similar in all of the examples. They manage to produce good prediction independently of the sparsity of the model.

The more attractive aspect concerns variable selection. For this purpose we treat each example separately.

Example (a): the Elastic-Net selects a model which is too large (Table 1). This is reflected by the worst SNR (Table 2). As a consequence, we can observe in Figure 3

Method	Example (a)	Example (c)	Example (d)
Lasso	2.3 [±0.1]	2.9 [±0.1]	4.7 [±0.2]
E-Net	1.7 [±0.1]	13.1 [±0.3]	3.4 [±0.2]
NS-Lasso	2.5 [±0.1]	13.5 [±0.3]	6.8 [±0.3]
HS-Lasso	1.79 [±0.1]	11.4 [±0.3]	6.4 [±0.3]

Table 2: Mean of the ratio between the number of relevant covariates and the number of noise covariates (SNR) [and its standard error] that each of the Lasso, the Elastic-Net, the NS-Lasso and the HS-Lasso procedures selected.

that it also includes the second covariate more often than the other procedures. This is due to the "grouping effect" as the first covariate is relevant. For similar reasons, the S-Lasso often selects the second covariate. However, this covariate is less selected than by the Elastic-Net as the S-Lasso seems to be a little bit disturbed by the third covariate which is irrelevant. This aspect of the S-Lasso procedure is also present in the selection of the covariate 5 as its neighbor covariates 4 and 6 are irrelevant. We can also observe that the S-Lasso procedure is the one which selects less often irrelevant covariates when these covariates are far away from relevant ones (in term of indices distance). Finally, even if the Lasso procedure selects less often the relevant covariates than the Elastic-Net and the S-Lasso procedures, it also has as good SNR. The Lasso presents good selection performances in this example.

Example (b): we can see in Figure 3 how the S-Lasso and Elastic-Net selection depends on how the covariates are ranked. They both select more covariates in the middle (that is covariates 2 to 7) than the ones in the borders (covariates 1 and 8) than the Lasso. We also remark that this aspect is more emphasized for the S-Lasso than for the Elastic-Net.

Example (c): the Lasso procedure performs poorly. It selects more noise covariates and less relevant ones than the other procedures (Figure 3). It also has the worst SNR (Table 2). In this example, Figure 3 also shows that the Elastic-Net selects more often relevant covariates than the S-Lasso procedures but it also selects more noise covariates than the NS-lasso procedure. Then even if the Elastic-Net has very good performance in variable selection, the NS-Lasso procedure has similar performances with a close SNR (Table 2). The NS-Lasso appears to have very good performance in this example. However, it selects again less often relevant covariates at the border than the Elastic-Net.

Example (d): we decompose the study into two parts. First, the independent part which considers covariates $\xi_1, \ldots, \xi_{10}$ and $\xi_{26}, \ldots, \xi_{30}$. The second part considers the

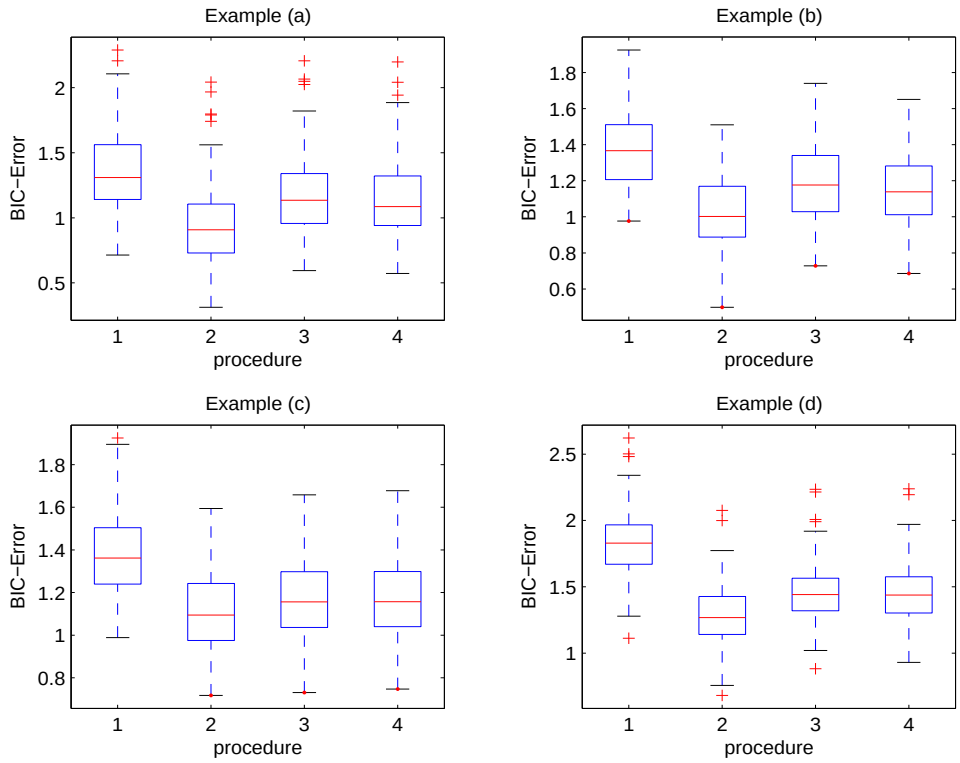

Figure 1: BIC error in each example. For each plot, we construct the boxplot for the procedure 1 = Lasso; 2 = Elastic-Net; 3 = NS-Lasso; 4 = HS-Lasso

other covariates which are dependent. Regarding the independent covariates, Figure 3 shows that all the procedures perform roughly in the same way, though the S-Lasso procedure enjoys a slightly better selection (in both relevant and noise group of covariates). For the dependent and relevant covariates, the Lasso performs worst than the other procedures. It selects clearly less often these relevant covariates. As in example (c), the reason is that the Lasso modification of the LARS algorithm tends to select only one representative of a group of highly correlated covariates. The high value of the SNR for the Lasso (when compared to the Elastic-Net) is explained by its good performance when it treat noise covariates. In this example the Elastic-Net correctly selects relevant covariates but it is also the procedure which selects the more noise covariates and has the worst SNR. We also note that both the NS-Lasso and HS-Lasso outperform the Lasso and Elastic-Net. This gain is emphasized especially in the center of the groups. Observe that for the covariates $\xi_{20}, \xi_{21}, \xi_{25}$ and ξ_{26} (that

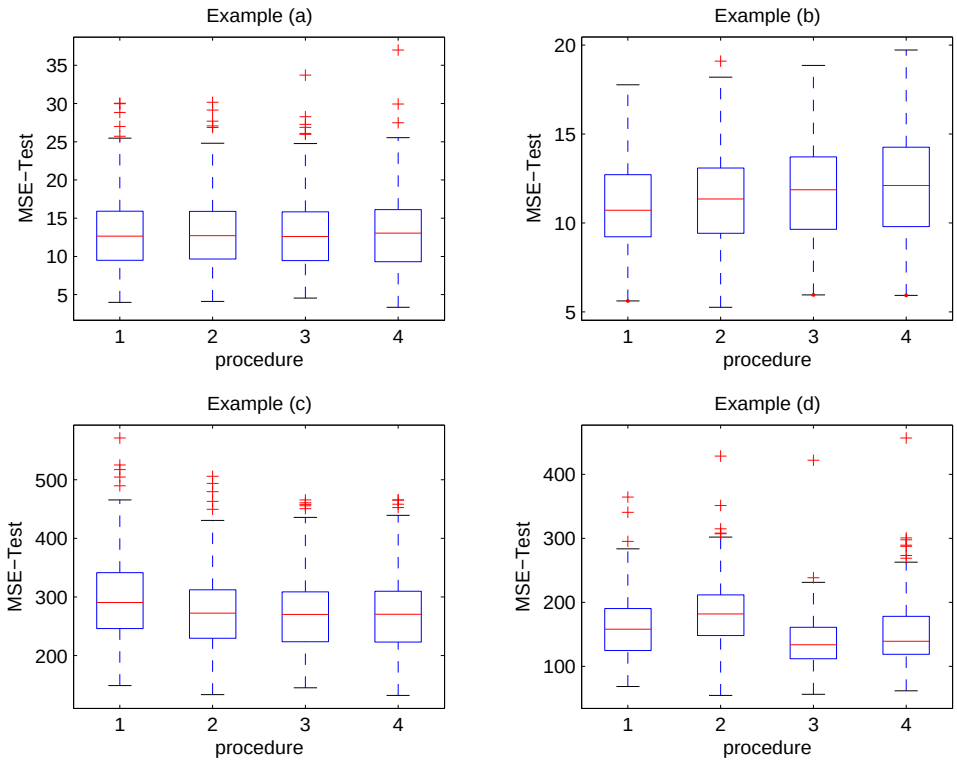

Figure 2: Test Error in each example. For each plot, we construct the boxplot for the procedure 1 = Lasso; 2 = Elastic-Net; 3 = NS-Lasso; 4 = HS-Lasso

is the borders), the NS-Lasso and HS-Lasso have slightly worst performance than in the center of the groups. This is again due to the attraction we imposed by the fusion penalty (3) in the S-Lasso criterion.

Conclusion of the experiments. The S-Lasso procedure seems to respond to our expectations. Indeed, when successive correlations exist, it tends to select the whole group of these relevant covariates and not only one representing the group as done by the Lasso procedure. It also appears that the S-Lasso procedure has very good selection properties according to both relevant and noise covariates. However it has slightly worst performance in the borders than in the centers of groups of covariates (due to attractions of irrelevant covariates). It almost always has a better SNR than the Elastic-Net, so we can take it as a good challenger for this procedure.

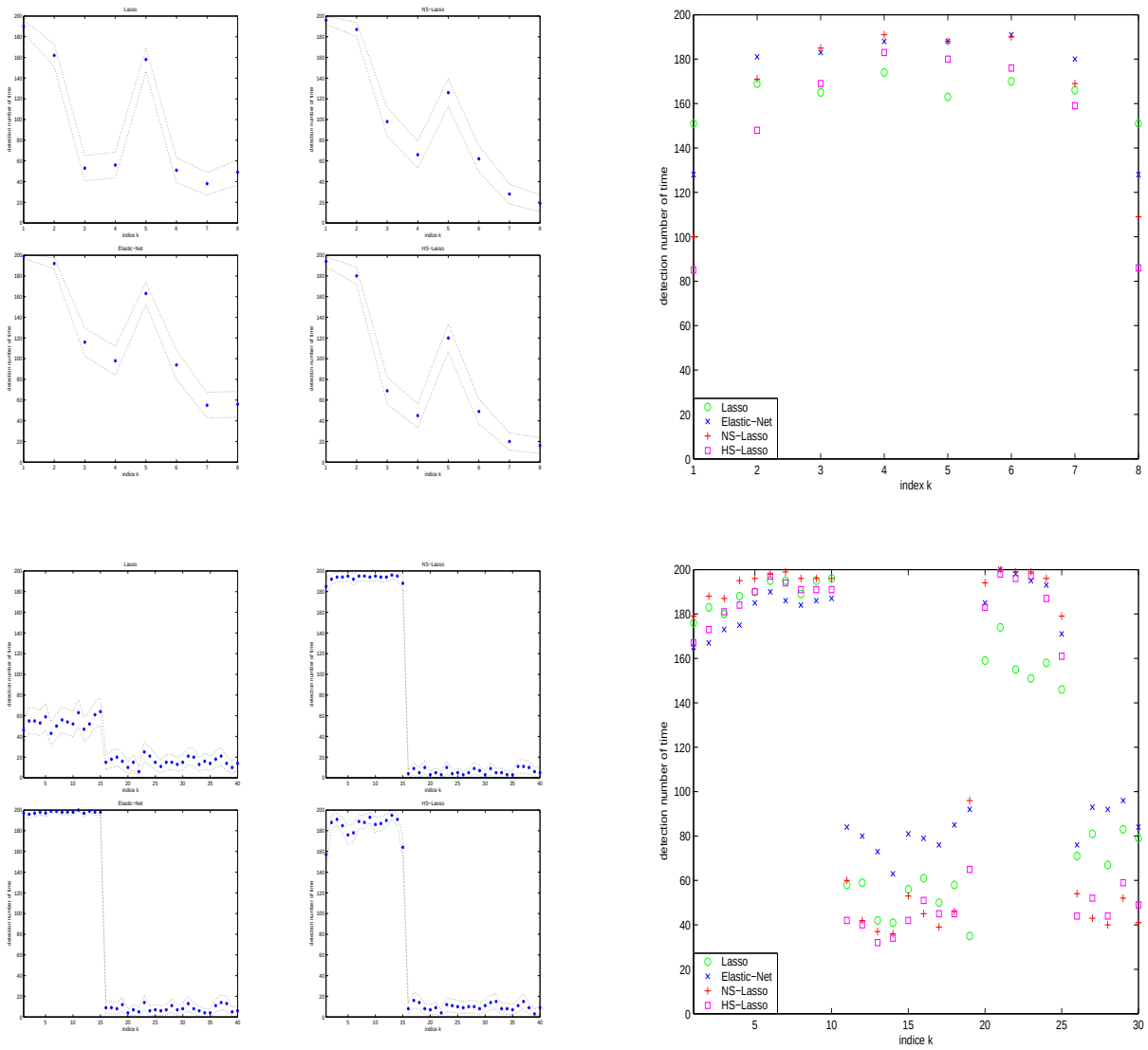

Figure 3: Number of covariates detections for each procedure in all the examples (Top-Left: Example (a); Top-Right: Example (b); Bottom-Left: Example (a); Bottom-Right: Example (b))

9 Conclusion

In this paper, we introduced a new procedure called the Smooth-Lasso which takes into account correlation between successive covariates. We established several theoretical results. The main conclusions are that when $p \leq n$, the S-Lasso is consistent in variable selection and asymptotically normal with a rate lower than $\sqrt{n}$. In the high dimensional setting, we provided a condition related to the coherence mutual condition, under which the thresholded version of the Smooth-Lasso is consistent in variable selection. This condition is fulfilled when correlations between successive covariates exist. Moreover, simulation studies showed that normalized versions of the Smooth-Lasso have nice properties of variable selection which are emphasized when high correlations exist between successive covariates. It appears that the Smooth-Lasso almost always outperforms the Lasso and is a good challenger of the Elastic-Net.

Appendix A.

Since the matrix $\mathcal{C}_n + \mu_n \tilde{J}$ plays a crucial role in the proves, we use to shorten the notation $K_n = \mathcal{C}_n + \mu_n \tilde{J}$ and when $p \leq n$ we define $K = \mathcal{C} + \mu \tilde{J}$, its limit.
In this appendix we prove the results when $p \leq n$.

Proof of Theorem 1. Let Ψ_n be

$$\begin{aligned} \Psi_n(u) = \|Y - X(\beta^* + v_n u)\|_n^2 \quad &+ \quad \lambda_n \sum_{j=1}^{p} |\beta_j^* + v_n u_j| \\ &+ \quad \mu_n \sum_{j=2}^{p} \left(\beta_j^* - \beta_{j-1}^* + v_n(u_j - u_{j-1})\right)^2, \end{aligned}$$

for $u = (u_1, \ldots, u_p)' \in \mathbb{R}^p$ and let $\hat{u} = \text{Argmin}_u \, \Psi_n(u)$. Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$, we then

have

$$\begin{aligned}
\Psi_n(u) - \Psi_n(0) &=: V_n(u) \\
&= v_n^2\, u' \left(\frac{X'X}{n} \right) u - 2 \frac{v_n}{\sqrt{n}} \frac{\varepsilon' X}{\sqrt{n}} u + v_n \lambda_n \sum_{j=1}^{p} v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*| \right) \\
&\quad + v_n \mu_n \sum_{j=2}^{p} v_n^{-1} \left\{ \left(\beta_j^* - \beta_{j-1}^* + v_n (u_j - u_{j-1}) \right)^2 - \left(\beta_j^* - \beta_{j-1}^* \right)^2 \right\} \\
&= v_n^2 \left[u' \left(\frac{X'X}{n} \right) u - \frac{2}{v_n \sqrt{n}} \frac{\varepsilon' X}{\sqrt{n}} u + \frac{\lambda_n}{v_n} \sum_{j=1}^{p} v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*| \right) \right. \\
&\quad \left. + \frac{\mu_n}{v_n} \sum_{j=2}^{p} v_n^{-1} \left\{ \left(\beta_j^* - \beta_{j-1}^* + v_n (u_j - u_{j-1}) \right)^2 - \left(\beta_j^* - \beta_{j-1}^* \right)^2 \right\} \right] \\
&= v_n^2 V_n(u).
\end{aligned}$$

Note that $\hat{u} = \operatorname{Argmin}_u \Psi_n(u) = \operatorname{Argmin}_u V_n(u)$, we then have to consider the limit distribution of $V_n(u)$. First, we have $\frac{X'X}{n} \to \mathbf{C}$. Moreover, as $1/(v_n\sqrt{n}) \to \kappa$ and as given X, the random variable $\frac{\varepsilon' X}{\sqrt{n}} \xrightarrow{\mathcal{D}} W$, with $W \sim \mathcal{N}(0, \sigma^2 \mathbf{C})$, the Slutsky theorem implies that

$$\frac{2}{v_n \sqrt{n}} \frac{\varepsilon' X}{\sqrt{n}} u \xrightarrow{\mathcal{D}} 2\kappa W' u.$$

Now we treat the last two terms. If $\beta_j^* \neq 0$,

$$v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*| \right) \to u_j \operatorname{Sgn}(\beta_j^*),$$

and is equal to $|u_j|$ otherwise. Then, as

$$\frac{\lambda_n}{v_n} \sum_{j=1}^{p} v_n^{-1} \left(|\beta_j^* + v_n u_j| - |\beta_j^*| \right) \to \lambda \sum_{j=1}^{p} \left\{ u_j \operatorname{Sgn}(\beta_j^*) \mathbb{I}(\beta_j^* \neq 0) + |u_j| \, \mathbb{I}(\beta_j^* = 0) \right\},$$

For the remaining term, we show that if $\beta_j \neq \beta_{j-1}$,

$$v_n^{-1} \left\{ \left(\beta_j^* - \beta_{j-1}^* + v_n (u_j - u_{j-1}) \right)^2 - \left(\beta_j^* - \beta_{j-1}^* \right)^2 \right\} \to 2(u_j - u_{j-1})(\beta_j^* - \beta_{j-1}^*),$$

and is equal to $\frac{(u_j - u_{j-1})^2}{n}$ otherwise. But μ_n converge to 0, implies that

$$\frac{\mu_n}{v_n} \sum_{j=2}^{p} v_n^{-1} \left\{ \left(\beta_j^* - \beta_{j-1}^* + v_n (u_j - u_{j-1}) \right)^2 - \left(\beta_j^* - \beta_{j-1}^* \right)^2 \right\} \to$$

$$2\mu \sum_{j=2}^{p} \left\{ (u_j - u_{j-1})(\beta_j^* - \beta_{j-1}^*) \mathbb{I}(\beta_j^* \neq \beta_{j-1}^*) \right\}.$$

Therefore we have $V_n(u) \to V(u)$ in probability, for every $u \in \mathbb{R}^p$. And since $\mathbf{C}$ is a positive defined matrix, $V(u)$ has a unique minimizer. Moreover as $V_n(u)$ is convex, standard M-estimation results [28] lead to: $\hat{u}_n \to \text{Argmin}_u V(u)$. □

Proof of Theorem 2. We begin by giving two results which we will use in our proof. The first one concerns the optimality conditions of the S-Lasso estimator. Recall that by definition

$$\hat{\beta}^{SL} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \, \|Y - X\beta\|_n^2 + \lambda_n |\beta|_1 + \mu_n \beta' \widetilde{J} \beta.$$

Note $f(a)|_{a=a_0}$ the evaluation of the function f at the point a_0. As the above problem is a non-differentiable convex problem, classical tools lead to the following optimality conditions for the S-Lasso estimator:

Lemma 3. *The vector $\hat{\beta}^{SL} = (\hat{\beta}_1^{SL}, \ldots, \hat{\beta}_p^{SL})'$ is the S-Lasso estimate as defined in* (2)-(3) *if and only if*

$$\left. \frac{\|Y - X\beta\|_n^2 + \mu_n \beta' \widetilde{J} \beta}{d\beta_j} \right|_{\beta_j = \hat{\beta}_j^{SL}} = -\lambda_n \, \text{Sgn}(\hat{\beta}_j^{SL}) \qquad \text{for } j : \hat{\beta}_j^{SL} \neq 0, \quad (22)$$

$$\left| \left. \frac{\|Y - X\beta\|_n^2 + \mu_n \beta' \widetilde{J} \beta}{d\beta_j} \right|_{\beta_j = \hat{\beta}_j^{SL}} \right| \leq \lambda_n \qquad \text{for } j : \hat{\beta}_j^{SL} = 0. \quad (23)$$

Recall that $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$, the second result states that if we restrict ourselves to the covariates which we are after (i.e. indexes in $\mathcal{A}^*$), we get a consistent estimate as soon as the regularization parameters λ_n and μ_n are properly chosen.

Lemma 4. *Let $\tilde{\beta}_{\mathcal{A}^*}$ a minimizer of*

$$\|Y - X_{\mathcal{A}^*} \beta_{\mathcal{A}^*}\|_n^2 + \lambda_n \sum_{j \in \mathcal{A}^*} |\beta_j| + \mu_n \beta'_{\mathcal{A}^*} \widetilde{J}_{\mathcal{A}^*, \mathcal{A}^*} \beta_{\mathcal{A}^*}.$$

If $\lambda_n \to 0$ and $\mu_n \to 0$, then $\tilde{\beta}_{\mathcal{A}^*}$ converges to $\beta_{\mathcal{A}^*}^*$ in probability.

This lemma can be see as a special and restricted case of Theorem 1. We now prove Theorem 2. Let $\tilde{\beta}_{\mathcal{A}^*}$ as in Lemma 4. We define an estimator $\tilde{\beta}$ by extending $\tilde{\beta}_{\mathcal{A}^*}$ by zeros on $(\mathcal{A}^*)^c$. Hence, consistency of $\tilde{\beta}$ is ensure as a simple consequence of Lemma 4. Now we need to prove that with probability tending to one, this estimator is optimal for the problem (2)-(3). That is the optimal conditions (22)-(23) are fulfilled with probability tending to one.

From now on, we denote $\mathcal{A}$ for $\mathcal{A}^*$. By definition of $\tilde{\beta}_{\mathcal{A}}$, the optimality condition (22) is satisfied. We now must check the optimality condition (23). Combining the fact that $Y = X\beta^* + \varepsilon$ and the convergence of the matrix $X'X/n$ and the vector $\varepsilon' X/\sqrt{n}$, we have

$$n^{-1}(X'Y - X'X_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) = \mathbf{C}_{\cdot,\mathcal{A}}(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) + \mathcal{O}_p(n^{-1/2}). \tag{24}$$

Moreover, the optimality condition (22) for the estimator $\tilde{\beta}$ can be written as

$$n^{-1}(X'_{\cdot,\mathcal{A}}Y - X'_{\cdot,\mathcal{A}}X_{\cdot,\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) = \frac{\lambda_n}{2}\operatorname{Sgn}(\tilde{\beta}_{\mathcal{A}}) - \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}}(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) + \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}. \tag{25}$$

Combining (24) and (25), we easily obtain

$$(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) = (\mathbf{C}_{\mathcal{A},\mathcal{A}} + \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}})^{-1}\left(\frac{\lambda_n}{2}\operatorname{Sgn}(\tilde{\beta}_{\mathcal{A}}) + \mu_n\widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}\right) + \mathcal{O}_p(n^{-1/2}).$$

Since $\tilde{\beta}$ is consistent and $\lambda_n n^{1/2} \to \infty$, for each $j \in \mathcal{A}^c$, the left hand side in the optimality condition (23)

$$\frac{1}{\lambda_n n}(\xi'_j Y - \xi'_j X_{\cdot,\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) - \frac{\mu_n}{\lambda_n}\widetilde{J}_{j,\mathcal{A}}\tilde{\beta}_{\mathcal{A}} =: L_j^{(n)},$$

converges in probability to

$$\mathbf{C}_{j,\mathcal{A}}(K_{\mathcal{A},\mathcal{A}})^{-1}\left(2^{-1}\operatorname{Sgn}(\beta^*_{\mathcal{A}}) + \frac{\mu}{\lambda}\widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}\right) - \frac{\mu}{\lambda}\widetilde{J}_{j,\mathcal{A}}\beta^*_{\mathcal{A}} =: L_j.$$

By condition (6), this quantity is strictly smaller than one. Then

$$\lim_{n\to\infty}\mathbb{P}\left(\forall j \in \mathcal{A}^c,\, |L_j^{(n)}| \leq 1\right) \geq \prod_{j\in\mathcal{A}^c}\mathbb{P}\left(|L_j| \leq 1\right) = 1,$$

which ends the proof. □

Proof of Theorem 3. We prove the theorem by contradiction by assuming that there exists a $j \in (\mathcal{A}^*)^c$ such that there exists a $i \in \mathcal{A}^*$ and

$$|\Omega_j(\lambda, \mu, \mathcal{A}^*, \beta^*)| > 1,$$

where the Ω_j are given by (5). Since $\mathcal{A}_n = \mathcal{A}^*$ with probability tending to one, optimality condition (22) implies

$$\hat{\beta}^{SL}_{\mathcal{A}} = ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \left(\frac{X'_{.,\mathcal{A}} Y}{n} - \frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) \right). \tag{26}$$

Using this expression of $\hat{\beta}^{SL}_{\mathcal{A}}$ and $Y = X_{.,\mathcal{A}} \beta^*_{\mathcal{A}} + \varepsilon$, then for every $j \in \mathcal{A}^c$,

$$\begin{aligned}
\frac{\xi'_j Y}{n} - \frac{\xi'_j X_{.,\mathcal{A}} \hat{\beta}^{SL}_{\mathcal{A}}}{n} &= \frac{\xi'_j Y}{n} - \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} Y}{n} \\
&\quad + \frac{\lambda_n}{2} \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) \\
&= \frac{\xi'_j Y}{n} - \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} \varepsilon}{n} - \frac{\xi'_j X_{.,\mathcal{A}}}{n} \beta^*_{\mathcal{A}} \\
&\quad + \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \left(\frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) + \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}} \beta^*_{\mathcal{A}} \right).
\end{aligned}$$

Therefore,

$$n^{-1}(\xi'_j Y - \xi'_j X_{.,\mathcal{A}} \hat{\beta}^{SL}_{\mathcal{A}}) - \mu_n \widetilde{J}_{j,\mathcal{A}} \beta^{SL}_{\mathcal{A}} = A_n + B_n,$$

with

$$\begin{cases}
A_n = \frac{\xi'_j Y}{n} - \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} \varepsilon}{n} - \frac{\xi'_j X_{.,\mathcal{A}}}{n} \beta^*_{\mathcal{A}} \\
B_n = \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \left(\frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) + \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}} \beta^*_{\mathcal{A}} \right) - \mu_n \widetilde{J}_{j,\mathcal{A}} \hat{\beta}^{SL}_{\mathcal{A}}.
\end{cases}$$

We treat this two terms separately. First as $\hat{\beta}^{SL}_{\mathcal{A}}$ converges in probability to $\beta^*_{\mathcal{A}}$ and empirical covariance matrices convergence, the sequence B_n/λ_n converges to

$$B = \mathbf{C}_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} (2^{-1} \lambda \operatorname{Sgn}(\beta^*_{\mathcal{A}}) + \mu \lambda^{-1} \widetilde{J}_{\mathcal{A},\mathcal{A}} \beta^*_{\mathcal{A}}) - \mu \lambda^{-1} \widetilde{J}_{j,\mathcal{A}} \beta^*_{\mathcal{A}}.$$

By assumption $|B| > 1$. This implies that $\mathbb{P}\left(B_n/\lambda_n \geq (1 + |B|)/2\right)$ converges to one.

With regard to the other term, since $Y = X\beta^* + \varepsilon$ we have

$$\begin{aligned}
A_n &= \frac{\xi'_j \varepsilon}{n} - \frac{\xi'_j X_{.,\mathcal{A}}}{n} ((K_n)_{\mathcal{A},\mathcal{A}})^{-1} \frac{X'_{.,\mathcal{A}} \varepsilon}{n} \\
&= n^{-1} \sum_{k=1}^{n} \varepsilon_k (x_{k,j} - \mathbf{C}_{j,\mathcal{A}} (K_{\mathcal{A},\mathcal{A}})^{-1} x'_{k,\mathcal{A}}) + o_p(n^{-1/2}) \\
&= n^{-1} \sum_{k=1}^{n} c_n + o_p(n^{-1/2}) = C_n + o_p(n^{-1/2}),
\end{aligned}$$

where c_n are i.i.d. random variables with mean 0 and variance:

$$\begin{aligned} s^2 = \mathbb{V}ar(c_k) &= \mathbb{E}(c_k^2) = \mathbb{E}[\mathbb{E}(c_k^2|X)] \\ &= \mathbb{E}\left[\mathbb{E}(\varepsilon_k^2|X)(x_{k,j} - \mathbf{C}_{j,\mathcal{A}}(K_{\mathcal{A},\mathcal{A}})^{-1}x'_{k,\mathcal{A}})^2\right] \\ &= \sigma^2\mathbb{E}\left[\mathbf{C}_{j,j} + \mathbf{C}_{j,\mathcal{A}}(K_{\mathcal{A},\mathcal{A}})^{-1}\mathbf{C}_{\mathcal{A},\mathcal{A}}(K_{\mathcal{A},\mathcal{A}})^{-1}\mathbf{C}_{\mathcal{A},j}\right. \\ &\quad \left. -2\mathbf{C}_{j,\mathcal{A}}(K_{\mathcal{A},\mathcal{A}})^{-1}\mathbf{C}_{\mathcal{A},j}\right]. \end{aligned}$$

Thus, by the central limit theorem, $n^{1/2}C_n$ is asymptotically normal with mean 0 and covariance matrix s^2/n, which is finite. Thus $\mathbb{P}(n^{1/2}A_n > 0)$ converges to $1/2$.

Finally, $\mathbb{P}((A_n + B_n)/\lambda_n > (1 + |B|)/2)$ is asymptotically bounded below by $1/2$. Thus $|(A_n + B_n)/\lambda_n|$ is asymptotically bigger than 1 with a positive probability, that is to say the optimality condition (23) is not satisfied. Then $\hat{\beta}^{SL}$ is not optimal. We get a contradiction, which concludes the proof. □

Appendix B.

In this appendix we mainly prove the results when $p \geq n$.

Proof of Theorem 4. Using the definition of the penalized estimator (2)–(3), for any $\beta \in \mathbb{R}^p$, we have

$$\begin{aligned} &\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 - \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i \hat{\beta}^{SL} + \lambda_n|\hat{\beta}^{SL}|_1 + \mu_n(\hat{\beta}^{SL})'\widetilde{J}\hat{\beta}^{SL} \\ \leq\ &\|X\beta - X\beta^*\|_n^2 - \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i \beta + \lambda_n|\beta|_1 + \mu_n\beta'\widetilde{J}\beta. \end{aligned}$$

Therefore, if we chose $\beta = \beta^*$, we obtain the following inequalities:

$$\begin{aligned} \|X\hat{\beta}^{SL} - X\beta^*\|_n^2 &\leq \lambda_n\sum_{j=1}^p \left(|\beta_j^*| - |\hat{\beta}_j^{SL}|\right) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i(\hat{\beta}^{SL} - \beta^*) \\ &\quad +\mu_n(\beta^{*'}\widetilde{J}\beta^* - (\hat{\beta}^{SL})'\widetilde{J}\hat{\beta}^{SL}) \\ &\leq \lambda_n\sum_{j=1}^p \left(|\beta_j^*| - |\hat{\beta}_j^{SL}|\right) + \frac{2}{n}\sum_{i=1}^n \varepsilon_i x_i(\hat{\beta}^{SL} - \beta^*) \\ &\quad +\mu_n\beta^{*'}\widetilde{J}\beta^*, \end{aligned} \tag{27}$$

as $\beta'\widetilde{J}\beta \geq 0$ for any $\beta \in \mathbb{R}^p$. In order to control (27), we use in a first time Assumption (A1) so that $\mu_n\beta^{*'}\widetilde{J}\beta^* \leq L_1\kappa_2\sigma^2\dfrac{\log(p)|\mathcal{A}^*|}{n}$. Second we bound the residual term

in the same way as in [4]. Then, we only present here the main lines. Recall that $\mathcal{A} = \mathcal{A}^* = \{j : \beta_j^* \neq 0\}$. Then, on the event $\Lambda_{n,p} = \{\max_{j=1,\ldots,p} 4|V_j| \leq \lambda_n\}$ with $V_j = n^{-1}\sum_{i=1}^n x_{i,j}\varepsilon_i$, we have

$$\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 + 2^{-1}\lambda_n \sum_{j=1}^p \left|\hat{\beta}_j^{SL} - \beta_j^*\right| \leq \lambda_n \sum_{j\in\mathcal{A}} \left|\hat{\beta}_j^{SL} - \beta_j^*\right| + L_1\kappa_2\sigma^2\frac{\log(p)|\mathcal{A}|}{n}. \quad (28)$$

This inequality is obtained thanks to the fact that $|\beta_j^* - \hat{\beta}_j^{SL}| + |\beta_j^*| - |\hat{\beta}_j^{SL}| = 0$ for any $j \notin \mathcal{A}$ and to the triangular inequality. The rest of the proof consists in bounding this term $\lambda_n \sum_{j\in\mathcal{A}} \left|\hat{\beta}_j^{SL} - \beta_j^*\right|$. Using similar arguments as in [4], we can write

$$\begin{aligned}\sum_{j\in\mathcal{A}}(\hat{\beta}_j^{SL} - \beta_j^*)^2 \;&\leq\; \|X\hat{\beta}^{SL} - X\beta^*\|_n^2 + 2\rho_1 \sum_{k\in\mathcal{A}} |\hat{\beta}_k^{SL} - \beta_k^*| \sum_{j=1}^p |\hat{\beta}_j^{SL} - \beta_j^*| \\ &\quad -\rho_1 \left(\sum_{j\in\mathcal{A}} |\hat{\beta}_j^{SL} - \beta_j^*|\right)^2. \end{aligned} \quad (29)$$

But $\left(\sum_{j\in\mathcal{A}} \left|\hat{\beta}_j^{SL} - \beta_j^*\right|\right)^2 \leq |\mathcal{A}| \sum_{j\in\mathcal{A}}(\hat{\beta}_j^{SL} - \beta_j^*)^2$, then

$$\begin{aligned}&\left(\sum_{j\in\mathcal{A}} \left|\hat{\beta}_j^{SL} - \beta_j^*\right|\right)^2 \\ &\leq |\mathcal{A}| \;\left\{\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 + 2\,\rho_1 \sum_{k\in\mathcal{A}} |\hat{\beta}_k^{SL} - \beta_k^*| \sum_{j=1}^p |\hat{\beta}_j^{SL} - \beta_j^*|\right. \\ &\quad \left. - \rho_1 \left(\sum_{j\in\mathcal{A}} |\hat{\beta}_j^{SL} - \beta_j^*|\right)^2\right\}. \end{aligned} \quad (30)$$

A simple optimization implies

$$\sum_{j\in\mathcal{A}} \left|\hat{\beta}_j^{SL} - \beta_j^*\right| \leq \frac{2\rho_1|\mathcal{A}| \sum_{j=1}^p |\hat{\beta}_j^{SL} - \beta_j^*|}{1 + \rho_1|\mathcal{A}|} + \frac{\sqrt{|\mathcal{A}|}\,\|X\hat{\beta}^{SL} - X\beta^*\|_n^2}{1 + \rho_1|\mathcal{A}|}. \quad (31)$$

Now, use Assumption (A2) to bound the left hand side of the inequality (31) and combine this to (28) to get

$$\|X\hat{\beta}^{SL} - X\beta^*\|_n^2 + \lambda_n \sum_{j=1}^p \left|\hat{\beta}_j^{SL} - \beta_j^*\right| \leq 16\lambda_n^2|\mathcal{A}| + L_1\kappa_2\sigma^2\frac{\log(p)|\mathcal{A}|}{n}. \quad (32)$$

This proves (11). Finally (12) follows directly by dividing by λ_n both sides of this last inequality. A concentration inequality to bound $\mathbb{P}\left(\max_{j=1,\dots,p} 4|V_j| \leq \lambda_n\right)$ allows us to conclude the proof. □

Lemma 5. *Let $\Lambda_{n,p}$ be the random event defined by $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 4|V_j| \leq \lambda_n\}$ where $V_j = n^{-1}\sum_{i=1}^n x_{i,j}\varepsilon_i$. Let us choose a $\kappa_1 > 2\sqrt{2}$ and $\lambda_n = \kappa_1\sigma\sqrt{n^{-1}\log(p)}$. Then*

$$\mathbb{P}\left(\max_{j=1,\dots,p} 4|V_j| \leq \lambda_n\right) \geq 1 - p^{1-\frac{\kappa_1^2}{8}}.$$

Proof. Since $V_j \sim \mathcal{N}(0, n^{-1}\sigma^2)$, an elementary Gaussian inequality gives

$$\begin{aligned}
\mathbb{P}\left(\max_{j=1,\dots,p} \lambda_n^{-1}|V_j| \geq 4^{-1}\right) &\leq p \max_{j=1,\dots,p} \mathbb{P}\left(\lambda_n^{-1}|V_j| \geq 4^{-1}\right) \\
&\leq p\exp\left(-\kappa_1^2\log(p)/8\right) \\
&= p^{1-\kappa_1^2/8}.
\end{aligned}$$

This ends the proof. □

Proof of Theorem 5. Through this proof, for any $a \in \mathbb{R}^p$, let us denote by $a_{\mathcal{A}}$, the p-dimensional vector such that $(a_{\mathcal{A}})_j = a_j$ if $j \in \mathcal{A}$ and zero otherwise. Moreover, we recall that $K_n = \mathbf{C}_n + \mu_n\widetilde{J}$. Now, note that we can write the KKT conditions (22)-(23) as

$$\|K_n(\hat{\beta}^{SL} - \beta^*) - \frac{X'\varepsilon}{n} + \mu_n\widetilde{J}\beta^*\|_\infty \leq \frac{\lambda_n}{2}. \tag{33}$$

Recall that $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq \lambda_n\}$ with $V_j = \frac{\xi_j'\varepsilon}{n}$, then applying (33) and Assumption (A4), we have on $\Lambda_{n,p}$ and for any $j \in \{1,\dots,p\}$

$$\begin{aligned}
|(K_n)_{j,j}(\hat{\beta}_j^{SL} - \beta_j^*)| &= |\{K_n(\hat{\beta}^{SL} - \beta^*)\}_j - \sum_{\substack{k=1\\k\neq j}}^p (K_n)_{j,k}(\hat{\beta}_k^{SL} - \beta_k^*) + \mu_n(\widetilde{J}\beta^*)_j| \\
&\leq \frac{\lambda_n}{2} + |\frac{\xi_j'\varepsilon}{n}| + \sum_{\substack{k=1\\k\neq j}}^p |(K_n)_{j,k}(\hat{\beta}_k^{SL} - \beta_k^*) + \mu_n(\widetilde{J}\beta^*)_j| \\
&\leq \frac{3\lambda_n}{4} + \frac{1}{3\alpha|\mathcal{A}|}|\hat{\beta}^{SL} - \beta^*|_1 + \mu_n|(\widetilde{J}\beta^*)_j|.
\end{aligned}$$

Then

$$\|K_n(\hat{\beta}^{SL} - \beta^*)\|_\infty \leq \frac{3\lambda_n}{4} + \frac{1}{3\alpha|\mathcal{A}|}|\hat{\beta}^{SL} - \beta^*|_1 + \mu_n\|\widetilde{J}\beta^*\|_\infty. \tag{34}$$

Let us now bound $|\hat{\beta}^{SL} - \beta^*|_1$. Thanks to (27), we can write

$$\lambda_n|\hat{\beta}^{SL}|_1 \leq \lambda_n|\beta^*|_1 + \frac{2}{n}\sum_{i=1}^{p}\varepsilon_i x_i(\hat{\beta}^{SL} - \beta^*) + \mu_n\beta^{*\prime}\widetilde{J}\beta^*$$
$$\overset{\text{on } \Lambda_{n,p}}{\Longleftrightarrow} \lambda_n|\hat{\beta}^{SL}|_1 \leq \lambda_n|\beta^*|_1 + \frac{\lambda_n}{2}|\hat{\beta}^{SL} - \beta^*|_1 + \mu_n\beta^{*\prime}\widetilde{J}\beta^*.$$

Dividing by λ_n, and adding $2^{-1}|\hat{\beta}^{SL} - \beta^*|_1 - |\hat{\beta}^{SL}|_1$, we get on the event $\Lambda_{n,p}$

$$\begin{aligned}
2^{-1}|\hat{\beta}^{SL} - \beta^*|_1 &\leq (|\hat{\beta}^{SL} - \beta^*|_1 + |\beta^*|_1 - |\hat{\beta}^{SL}|_1) + \frac{\mu_n}{\lambda_n}\beta^{*\prime}\widetilde{J}\beta^* \\
\iff \quad |\hat{\beta}^{SL} - \beta^*|_1 &\leq 2|\hat{\beta}^{SL}_{\mathcal{A}} - \beta^*_{\mathcal{A}}|_1 + 2\frac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}^*} \qquad (35) \\
\iff \quad |\hat{\beta}^{SL} - \beta^*|_1 &\leq 2\sqrt{|\mathcal{A}|}\,\|\hat{\beta}^{SL}_{\mathcal{A}} - \beta^*_{\mathcal{A}}\|_2 + 2\frac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}^*}, \qquad (36)
\end{aligned}$$

where we used the Cauchy Schwarz inequality in the last line. Combine (34) and (36), we easily get

$$\begin{aligned}
\|\hat{\beta}^{SL} - \beta^*\|_\infty \leq \frac{1}{1+\mu_n} \quad &\Big(\tfrac{3\lambda_n}{4} + \tfrac{2}{3\alpha|\mathcal{A}|}\sqrt{|\mathcal{A}|}\,\|\hat{\beta}^{SL}_{\mathcal{A}} - \beta^*_{\mathcal{A}}\|_2 \\
&+\mu_n\|\widetilde{J}\beta^*\|_\infty + \tfrac{2\mu_n}{3\alpha\lambda_n|\mathcal{A}|}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}\Big). \qquad (37)
\end{aligned}$$

The final step consists in bounding $\|\hat{\beta}^{SL}_{\mathcal{A}} - \beta^*_{\mathcal{A}}\|_2$. First, using the KKT condition (33), we remark that $\|K_n(\hat{\beta}^{SL} - \beta^*)\|_\infty \leq 3\lambda_n/4 + \mu_n\|\widetilde{J}\beta^*\|_\infty$ on $\Lambda_{n,p}$. This and equation (36) lied to

$$\begin{aligned}
(\hat{\beta}^{SL} - \beta^*)' K_n(\hat{\beta}^{SL} - \beta^*) &\leq \|K_n(\hat{\beta}^{SL} - \beta^*)\|_\infty\,|\hat{\beta}^{SL} - \beta^*|_1 \\
&\leq (\frac{3\lambda_n}{4} + \mu_n\|\widetilde{J}\beta^*\|_\infty)(2\sqrt{|\mathcal{A}|}\|\hat{\beta}^{SL}_{\mathcal{A}} - \beta^*_{\mathcal{A}}\|_2 + 2\frac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}). \qquad (38)
\end{aligned}$$

On the other hand, using Assumption (A4), and similar arguments as in [16],

$$\begin{aligned}
\frac{(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'K_n(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} &= \frac{(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'\,\mathrm{diag}(K_n)(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\
&\quad + \frac{(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'(K_n-\mathrm{diag}(K_n))(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\
&\geq 1-\frac{1}{3\alpha|\mathcal{A}|}\sum_{j,k=1}^{p}\frac{|(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})_j|\,|(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})_k|}{\|\hat\beta_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\
&\geq 1-\frac{1}{3\alpha|\mathcal{A}|}\frac{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_1^2}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2},
\end{aligned}$$

where we used in the second inequality the fact that $\mathrm{diag}(K_n)$ has larger diagonal elements than 1 since the diagonal elements in $\mathbf{C}_n$ and $\widetilde{J}$ are respectively equal to 1 and larger than 0. Now, twice using Assumption (A4), one deduces

$$\begin{aligned}
\frac{(\hat\beta^{SL}-\beta^*)'K_n(\hat\beta^{SL}-\beta^*)}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} &\geq \frac{(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})'K_n(\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}})}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}+\frac{(\hat\beta^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c})'K_n(\hat\beta^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c})}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\
&\geq 1-\frac{1}{3\alpha|\mathcal{A}|}\frac{|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1^2}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}-\frac{|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1|\hat\beta^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c}|_1}{3\alpha|\mathcal{A}|\,\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\
&\geq 1-\frac{|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1^2}{\alpha|\mathcal{A}|\,\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}-\frac{2\mu_n\,\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}{3\alpha\lambda_n|\mathcal{A}|}\frac{|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2} \\
&\geq (1-\frac{1}{\alpha})-\frac{2\mu_n\,\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}{3\alpha\lambda_n|\mathcal{A}|}\frac{|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1}{\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2}.
\end{aligned}$$

where we used the fact that (35) implies $|\hat\beta^{SL}_{\mathcal{A}^c}-\beta^*_{\mathcal{A}^c}|_1 \leq 2|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1+2\frac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}$ in the third line. The last inequalities can be summed-up by

$$(\hat\beta^{SL}-\beta^*)'K_n(\hat\beta^{SL}-\beta^*) \geq (1-\frac{1}{\alpha})\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2^2-\frac{2\mu_n\,\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}{3\alpha\lambda_n|\mathcal{A}|}|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}|_1. \quad (39)$$

Let us consider (38) and (39). An optimization work over $\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2$ provides us the following bound:

$$\begin{aligned}
\|\hat\beta^{SL}_{\mathcal{A}}-\beta^*_{\mathcal{A}}\|_2 \leq\ & (\tfrac{\alpha}{\alpha-1})\left[(\tfrac{3\lambda_n}{2}+2\mu_n\|\widetilde{J}\beta^*\|_\infty)\sqrt{|\mathcal{A}|}+\tfrac{2\mu_n}{3\alpha\lambda_n\sqrt{|\mathcal{A}|}}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}\right] \\
&+\sqrt{\tfrac{\alpha}{\alpha-1}\left(\tfrac{3\lambda_n}{2}+2\mu_n\|\widetilde{J}\beta^*\|_\infty\right)\tfrac{\mu_n}{\lambda_n}\beta^{*\prime}_{\mathcal{A}}\widetilde{J}\beta^*_{\mathcal{A}}}. \quad (40)
\end{aligned}$$

Thanks to Assumption (A1), $\beta^{*'}_{\mathcal{A}}\widetilde{J}\beta^{*}_{\mathcal{A}} \leq L_1 \log(p)|\mathcal{A}|$ and $\|\widetilde{J}\beta^*\|_\infty \leq L_2 \log(p)$. Moreover the tuning parameters λ_n and μ_n are chosen in the form $\lambda_n = \kappa_1\sigma\sqrt{\log(p)/n}$ and $\mu_n = \kappa_3\sigma/n$. Then we conclude from (37) and (40)

$$\|\hat{\beta}^{SL} - \beta^*\|_\infty \leq \frac{1}{1+\frac{\kappa_3\sigma}{n}} \quad \Big(\tfrac{3}{4} + \tfrac{1}{\alpha-1} + \tfrac{4L_1\kappa_3}{9\alpha^2\kappa_1^2} + \tfrac{2L_1\kappa_3}{3\alpha\kappa_1^2} + \sqrt{\tfrac{2L_1\kappa_3}{3\alpha(\alpha-1)\kappa_1^2} + \tfrac{8L_1\,L_2\kappa_3^2}{9\alpha(\alpha-1)\kappa_1^4}\lambda_n} + (\tfrac{4L_2\kappa_3}{3\kappa_1^2} + \tfrac{L_2\kappa_3}{\kappa_1^2})\lambda_n\Big)\,\lambda_n.$$

This ends the proof. □

Proof of Theorem 7. The proof of this theorem is essentially an adaptation of the one concerning the Lasso in [37]. We do not give the whole proof but only mention the important steps and let the reader refer to [37] for more details. The main points in the proof are Stein's lemma and these few facts:

- For every couple (λ, μ), the S-Lasso estimator is a continuous function of Y.
- For every couple $(\lambda, \mu) = \zeta$, the active set $\mathcal{A}_\zeta$ and the sign vector of $\hat{\beta}^{SL}_\zeta$ which we denote by Sgn_ζ are piecewise constant with respect to Y, out of a set with Lebesgue measure equal to 0.

The detailed proof uses these points and the explicit form of the estimator $\hat{\beta}^{SL}$ given by (26). This proof is the same as the one in [37] so that we omit it here. □

References

[1] F. R. Bach. Consistency of the group Lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[2] F. Bunea. Consistent selection via the lasso for high dimensional approximating regression models. IMS Collections, B. Clarke and S. Ghosal Editors, 2008.

[3] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[4] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.

[5] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables lasso. *Technical Report*, 2007.

[6] A. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. *20th Annual Conference on Learning Theory, COLT 2007 Proceedings. Lecture Notes in Computer Science 4539 Springer*, pages 97–111, 2007.

[7] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[8] D.L. Donoho and I.M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[9] B. Efron. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, 81(394):461–470, 1986.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression - with discussion. *Ann. Statist.*, 32(2):407–499, 2004.

[11] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[12] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[13] J.. Jia and B. Yu. On model selection consistency of the elastic net when p ¿¿ n. *Technical Report*, 2008.

[14] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 2000.

[15] S.R. Land and J.H. Friedman. Variable fusion: a new method of adaptive signal regression. *Technical Report*, 1996.

[16] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[17] N. Meinshausen. Lasso with relaxation. *Technical Report*, 2005.

[18] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[19] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist., to appear, 2006.

[20] A. Rinaldo. Properties and refinements of the fused lasso. *Technical Report*, 2008.

[21] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Ann. Statist.*, 35(3):1012–1030, 2007.

[22] G. Schwartz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[23] J. Shao. An asymptotic theory for linear model selection - with comments. *Statist. Sinica*, 7(2):221–264, 1997.

[24] X. Shen and J. Ye. Adaptive model selection. *J. Amer. Statist. Assoc.*, 97(457):210–221, 2002.

[25] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[26] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[27] A.B. Tsybakov and S.A. Van de Geer. Square root penalty: adaptation to the margin in classification and in edge estimation. *Ann. Statist.*, 33(3):1203–1224, 2005.

[28] A.W. Van Der Vaart. Asymptotic statistics. *Cambridge Univ. Press*, 1998.

[29] M.J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming. Technical report n°709, Department of Statistics, UC Berkeley, 2006.

[30] Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[31] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67, 2006.

[32] M. Yuan and Y. Lin. On the non-negative garrotte estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2):143–161, 2007.

[33] Cun-Hui Zhang and Jian Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

[34] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[35] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

[36] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

[37] H. Zou, T. Hastie, and R. Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007.