Regularization with the Smooth-Lasso procedure

Mohamed Hebiri

▶ **To cite this version:**

Mohamed Hebiri. Regularization with the Smooth-Lasso procedure. 2008. hal-00260816v1

HAL Id: hal-00260816

https://hal.science/hal-00260816v1

Preprint submitted on 5 Mar 2008 (v1), last revised 15 Oct 2008 (v2)

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

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

hal-00260816, version 1 - 5 Mar 2008

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 variables (or predictors). The study covers the case when $n \ll p$, i.e. the number of observations is much smaller than the number of variables. Moreover, for fixed p, we establish asymptotic normality and consistency in variable selection results for our procedure. 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 variables exist. This procedure also appears to be a good challenger to the Elastic-Net [26].

Keywords: Lasso, LARS, Sparsity, Variable selection, Regularization paths

1 Introduction

We focus on the usual linear regression model:

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

*hebiri@math.jussieu.fr

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 variables $(x_{.,j})_j$ is truly of interest where $x_{.,j} = (x_{1,j}, \ldots, x_{n,j})'$, $j = 1, \ldots, p$. 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 $\|a\|_n^2 = n^{-1} \sum_{i=1}^n |a_i|^2$. The choice of the penalty appears to be crucial. Although well-suited for variable-selection purpose, Concave-type penalties ([8], [19] and [3]) are often computationally hard to optimize. Lasso-type procedures (modifications of the l_1 penalized least square (Lasso) estimator introduced by Tibshirani [17]) have been extensively studied during the last few years. They 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 [10]. We add it to the Lasso procedure in order to overcome the variable selection problems of the Lasso estimator. Indeed the Lasso estimator has good selection properties but fails in some situations. More precisely, in several works ([24], [12], [25] and [23]) 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 relevant (in the true model) and irrelevant (not in the true model) variables. We will give similar consistency conditions for the S-Lasso procedure. Problems are also encountered when we solve the Lasso criterion with the Lasso modification of the LARS algorithm [6]. Indeed this algorithm tends to select only one representer of each group of correlated variables. We attempt to respond to this problem in the case where the variables are ranked so that high correlations

can exist between successive variables. We will see through the simulations that such situations support the use of the *S-lasso* estimator. This estimator is inspired by the *Fused-Lasso* [18]. Both S-Lasso and Fused-Lasso combine a l_1-penalty with a fusion term [10]. 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: $\sum_{j=2}^{p} |\beta_j - \beta_{j-1}|$). The use of the l_2 distance can be relevant as: i) it forces successive coefficients to be close without a perfect match (sparsity between coefficients differences) which is the case for the l_1 distance proposed in the Fused-Lasso; ii) the l_2 distance is strictly convex so that we can more easily optimize the S-Lasso criterion than the Fused-Lasso. Actually, the sparsity is yet ensured by the Lasso penalty, we suggest the additional l_2 penalty mainly to catch correlations between variables. More relevant variables can then be selected due to correlations between them.

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 [25], 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 [11] 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 variables, group variable selection has been proposed by Yuan and Lin [22] with the Group-Lasso procedure which selects groups of correlated variables instead of single variables at each step. A first step to the consistency study has been proposed in [1] and Sparse Oracle Inequalities were given in [2]. Another choice of penalty has been proposed with the Elastic-Net [26]. It is in the same spirit that we shall treat the S-Lasso from a 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. We also establish a local bound for the regression coefficients. We give an estimate of the effective degree of freedom of the S-Lasso estimator in Section 4. We provide a way to control the variance of the estimator by scaling in Section 5 where a connection with soft-thresholding is also established. We finally give experimental results in Section 6 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

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

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 variables 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 [13]. 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 variables 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 variables. Then we are no longer limited by the sample size as for the Lasso [6].*

3 Theoretical properties of the S-Lasso estimator

In this section we introduce the theoretical results about the S-Lasso. We first establish a link between regression coefficients and correlation between variables. We then 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. Finally study for variable selection consistency. More precisely, we give conditions under which the S-Lasso estimator succeeds in finding the set of the non-zeros regression coefficients.

3.1 Local proximity

Here we show that there exists a link between the regression coefficients β_j and β_k of two variables $x_{.,j}$ and $x_{.,k}$ and these variables correlation. Remember that Y is centered and X standardized. Let us note $\rho = n^{-1}X'X$ the correlation matrix and $\rho(j,k) = n^{-1}x'_{.,j}x_{.,k}$ the sample correlation between variables j and k. Define $\Delta\beta_j = \beta_j - \frac{(\beta_{j+1}+\beta_{j-1})}{2}$ for $j \in 1, \ldots, p$ with the convention: $\beta_0 = \beta_1$ and $\beta_{p+1} = \beta_p$;

Theorem 1. *Given the data set (X, Y) and the parameters (λ, μ). Let $\hat{\beta}^{SL} = \hat{\beta}^{SL}(\lambda, \mu)$ be the S-Lasso estimator. Assume that $\hat{\beta}_j^{SL}\hat{\beta}_k^{SL} > 0$. Then for every $(j,k) \in \{1,\dots,p\}^2$, we have*

$$\left|\Delta\hat{\beta}_j^{SL} - \Delta\hat{\beta}_k^{SL}\right| \leq \frac{\|Y\|_n}{4\lambda_2}\sqrt{2\left(1-\rho(j,k)\right)}.$$

Remark 3. *Note that we obtained nearly the same bound as for the Elastic-Net procedure [26]. This is described as the "grouping effect". The main difference is that they bound $|\beta_j - \beta_k|$ whereas we bound $|\Delta\beta_j - \Delta\beta_k|$ which is a local version of the former.*

Theorem 1 states that the more correlated the variables $x_{.,j}$ and $x_{.,k}$ are, the smaller the difference between the local approximation $\Delta\hat{\beta}_j^{SL}$ and $\Delta\hat{\beta}_k^{SL}$ of their regression coefficients is. Then, in the high correlation case ($\rho(j,k) \simeq 1$), we have $\Delta\hat{\beta}_j^{SL} \simeq \Delta\hat{\beta}_k^{SL}$. This relation can be interpreted in two ways.

Either the quantities $\Delta\hat{\beta}_j^{SL}$ and $\Delta\hat{\beta}_k^{SL}$ are both close to 0; thus $\hat{\beta}_j^{SL}$ and $\hat{\beta}_k^{SL}$ are well approximated by $\frac{\hat{\beta}_{j-1}^{SL}+\hat{\beta}_{j+1}^{SL}}{2}$ and $\frac{\hat{\beta}_{k-1}^{SL}+\hat{\beta}_{k+1}^{SL}}{2}$ respectively. This is the more expected conclusion as the S-Lasso estimator (2)-(3) is mainly used when variables are ranked. Indeed, in such a problem it is obvious that the regression coefficient β_j^* of the variable $x_{.,j}$ depends on the coefficients β_{j-1}^* and β_{j+1}^*.

Or the quantities $\Delta\hat{\beta}_j^{SL}$ and $\Delta\hat{\beta}_k^{SL}$ are not close to 0 but are approximatively of the same order. This implies that correlated variables (even if their indexes distance $|j-k|$ is large) are sensitive in the same way to their neighboring values.

In both cases we conclude that there exists a link between $\hat{\beta}_j^{SL}$ and $\frac{\hat{\beta}_{j-1}^{SL}+\hat{\beta}_{j+1}^{SL}}{2}$. Moreover one can predict the behaviour of one variable in its neighborhood when analyzing the behavior of a correlated variable in its respective neighborhood.

3.2 Asymptotic Normality

In this section, we allow the regularization 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 variables 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 [9] gave the asymptotic distribution of the Lasso estimator. We provide here the asymptotic distribution to the S-Lasso.

Theorem 2. *Given the data set (X, Y), assume the correlation matrix verifies*

$$n^{-1}X'X \to \mathbf{C}, \quad \textit{when } n \to \infty,$$

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

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

where

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

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

Remark 4. *When $\kappa \neq 0$ is a finite constant: in this case v_n^{-1} is $\mathcal{O}_p(\sqrt{n})$ 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 $\mu_n / v_n \to \mu$.

3.3 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 true 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 $\mathcal{B} \in \{1, \dots, p\}$, 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 $j \in \{1, \dots, p\}$, the quantity Ω_j by

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

Now consider the following conditions: *for every* $j \in (\mathcal{A}^*)^c$

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

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

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 ([25], [24] and [23]). Now we state the consistency results

Theorem 3. *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 4. *If there exist sequences* (λ_n, μ_n) *such that the S-Lasso estimator is consistent in variable selection, 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 3 (more precisely $\lambda_n n^{1/2} \to \infty$), root-$n$ consistency and variables 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 5. *Here are special cases of condition* (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 [23] 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*

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

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

- *if* $(\mathbf{C}_{j,\mathcal{A}^*}\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*}^{-1}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*})\beta^*_{\mathcal{A}^*} > 0$*, then* γ *must be chosen between* $-\dfrac{1 + 2^{-1}\mathbf{C}_{j,\mathcal{A}^*}\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*}^{-1}\operatorname{Sgn}(\beta^*_{\mathcal{A}^*})}{(\mathbf{C}_{j,\mathcal{A}^*}\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*}^{-1}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*})\beta^*_{\mathcal{A}^*}}$ *and* $\dfrac{1 - 2^{-1}\mathbf{C}_{j,\mathcal{A}^*}\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*}^{-1}\operatorname{Sgn}(\beta^*_{\mathcal{A}^*})}{(\mathbf{C}_{j,\mathcal{A}^*}\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*}^{-1}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*})\beta^*_{\mathcal{A}^*}}$.
- *if* $(\mathbf{C}_{j,\mathcal{A}^*}\mathbf{C}_{\mathcal{A}^*,\mathcal{A}^*}^{-1}\widetilde{J}_{\mathcal{A}^*,\mathcal{A}^*} - \widetilde{J}_{j,\mathcal{A}^*})\beta^*_{\mathcal{A}^*} < 0$*, 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 more easily be satisfied.*

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 regularization parameters (λ_n, μ_n) to get consistency in variable selection. Note that Zou [25] showed that the Lasso estimator cannot be consistent in variable selection even if 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 both asymptotic normality and variable selection consistency.

4 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

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

where the expectation is taken over the random variable Y. The last term in this equation was called *optimism* [5]. Moreover, Tibshirani [17] 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

$$\mathbb{E}\left\{\|m-\hat{y}^{SL}\|_n^2\right\} = \mathbb{E}\left\{\|Y-\hat{y}^{SL}\|_n^2 - \sigma^2 + 2n^{-1}\,\mathrm{df}(\hat{y}^{SL})\sigma^2\right\}. \tag{8}$$

This final expression involves the degree of freedom which is unknown. Various methods exist to estimate the degree of freedom as bootstrap [7] or data perturbation methods [16]. We give an explicit form to the degree of freedom in order to reduce the computational cost as in [6] and [27].

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 variables 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 5. *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

$$\widetilde{J} = \begin{pmatrix} 1 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \ldots & 0 & -1 & 1 \end{pmatrix}. \tag{9}$$

As the estimation given in Theorem 5 has an important computational cost, we propose the following estimator of the degree of freedom of the S-Lasso estimator:

$$\widehat{\mathrm{df}}(\hat{y}_\zeta^{SL}) = \frac{|\mathcal{A}_\zeta| - 2}{1 + 2\mu} + \frac{2}{1 + \mu}, \tag{10}$$

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 6 (*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.*

5 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. the l_2-fusion) in order to find the constant we have to scale with.

Define

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

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 (9). Moreover as the design matrix X is standardized, the symmetric matrix $\mathbf{L}$

can be written

$$\mathbf{L} = \begin{pmatrix} 1+\mu & \rho(1,2)-\mu & \rho(1,3) & \dots & \rho(1,p) \\ & 1+2\mu & \rho(2,3)-\mu & \dots & \vdots \\ & & \ddots & \ddots & \rho(p-2,p-1) \\ & & & 1+2\mu & \rho(p-1,p)-\mu \\ & & & & 1+\mu \end{pmatrix},$$

where $\rho(j,k)$ is the correlation between the variables $x_{.,j}$ and $x_{.,k}$.

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 $\widetilde{\beta}$ 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

$$\left\|\widetilde{Y} - \widetilde{X}b\right\|_n^2 + r|b|_1. \tag{11}$$

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

$$\hat{\beta}^{SSL} = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}} \left\{ \frac{\nu_2}{\nu_1}\beta' \left(\frac{\frac{X'X}{n} + \mu\widetilde{J}}{c} \right) \beta - 2\frac{Y'X}{n}\beta + \lambda\sum_{j=1}^{p}|\beta_j| \right\}. \tag{12}$$

Equation (12) is only a rearrangement of the Lasso criterion (11). The SS-Lasso expression (12) emphasizes the importance of the scaling constant c. In a way, the SS-Lasso estimator stabilizes the Lasso estimator $\hat{\beta}^L$ (criterion (11) 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 7 (*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* (12)*, we have*

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

As a consequence, $(\hat{\beta}^{NSL}_{\infty})_j = \left(|Y'x_{.,j}| - \frac{\lambda}{2} \right)_+ \text{Sgn}(Y'x_{.,j})$ *which is the Univariate Soft Thresholding [4]. Hence, when* $\mu \to \infty$*, the NS-Lasso works as if all the variables were independent. The Lasso, which corresponds to the NS-Lasso when* $\mu = 0$*, often fails to select variables when high correlations exist between relevant and irrelevant variables. It seems that the NS-Lasso is able to avoid such problem by increasing* μ *and working as if all the variables 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.*

6 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 [26] even when large correlations between variables 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 variables and irrelevant selected variables.

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

$$y = x\beta^* + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1),\ x = (x_1, \ldots, x_p).$$

The first and the second examples were introduced in the original Lasso paper [17]. The third simulation creates a grouped variables situation. It was introduced in [26] and aims to point the efficiency of the Elastic-Net compared to the Lasso. The last simulation introduces large correlation between successive variables.

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

(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 variables are relevant.

(c) In the third example, we simulate 50 data with 40 variables. 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 variables generated as follows:

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

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

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

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

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

Validation. The selection of the regularization parameters λ and μ is based on the minimization of a BIC-type criterion [14]. 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 (10) if we consider the S-Lasso and denotes its analogous quantities if we consider the Lasso or the Elastic-Net. Such criterion provides an accurate estimator which enjoys good variable selection properties ([15] and [21]). 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 variable 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 variables and the number of noise variables that the procedures selected. Let us call SNR this ratio. Then we can express this ratio as

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

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.

Method	Example (a)	Example (b)	Example (c)	Example (d)
Lasso	3.8 [±1.2]	6.5 [±1.2]	6 [±2.0]	18.4 [±3.0]
E-Net	4.9 [±1.2]	6.9 [±0.6]	15.9 [±2.0]	20.5 [±3.4]
NS-Lasso	3.9 [±1.3]	6.5 [±0.9]	15.3 [±2.3]	18.9 [±2.5]
HS-Lasso	3.5 [±1.2]	5.9 [±1.2]	15 [±2.2]	18.1 [±3.1]

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

Method	Example (a)	Example (c)	Example (d)
Lasso	2.3 [±1.1]	2.9 [±1.7]	4.7 [±3.2]
E-Net	1.7 [±1.0]	13.1 [±4.4]	3.4 [±2.9]
NS-Lasso	2.5 [±1.0]	13.5 [±4.3]	6.8 [±4.8]
HS-Lasso	1.79 [±1.0]	11.4 [±4.6]	6.4 [±4.1]

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

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

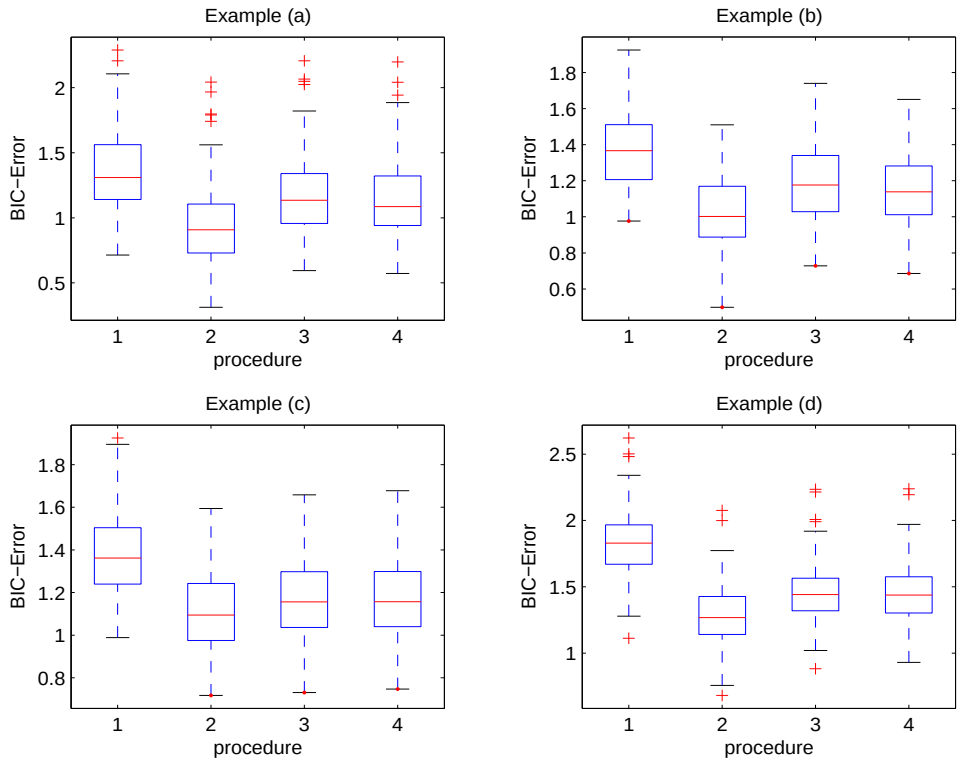

Figure 1: BIC error in each example. For each plot, we construct the boxplot for the procedure 1 = Lasso; 2 = Elastic-Net; 3 = NS-Lasso; 4 = HS-Lasso

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 variables are ranked. They both select more variables in the middle (that is variables 2 to 7) than the ones in the borders (variables 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 variables and less relevant ones tha n 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 variables than the S-Lasso procedures but it also selects more noise variables 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

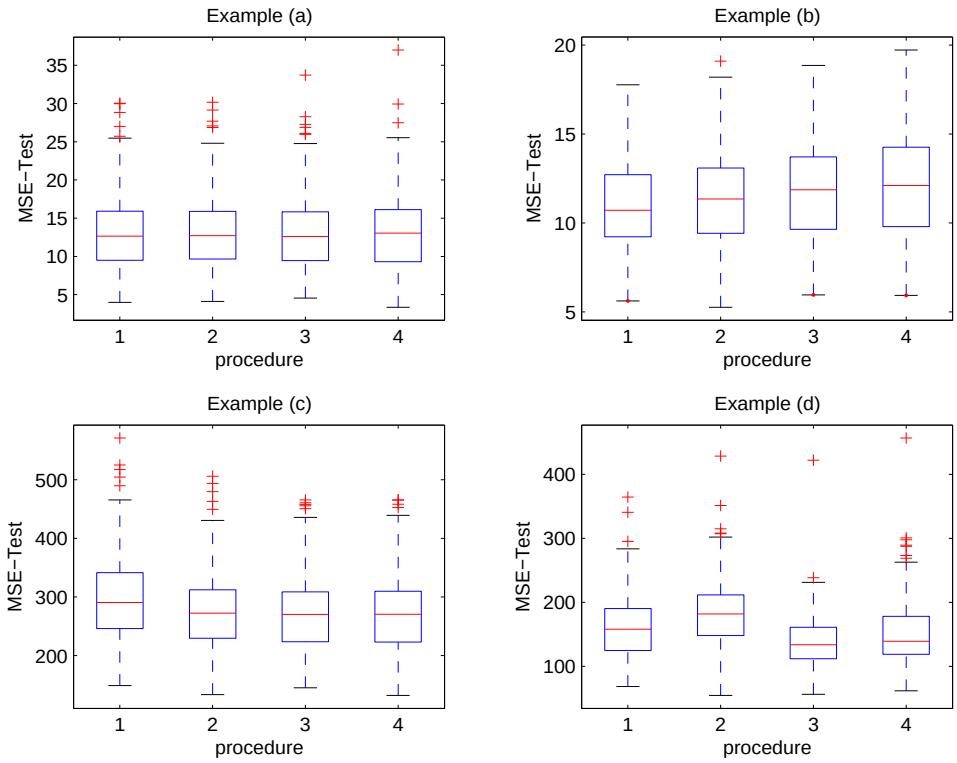

Figure 2: Test Error in each example. For each plot, we construct the boxplot for the procedure 1 = Lasso; 2 = Elastic-Net; 3 = NS-Lasso; 4 = HS-Lasso

in this example. However, it selects again less often relevant variables at the border than the Elastic-Net.

Example (d): we decompose the study into two parts. First, the independent part which considers variables $x_1, \ldots, x_{10}$ and $x_{26}, \ldots, x_{30}$. The second part considers the other variables which are dependent. Regarding the independent variables, 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 variables). For the dependent and relevant variables, the Lasso performs worst than the other procedures. It selects clearly less often these relevant variables. As in example (c), the reason is that the Lasso modification of the LARS algorithm tends to select only one representer of a group of highly correlated variables. 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 variables. In this example the Elastic-Net

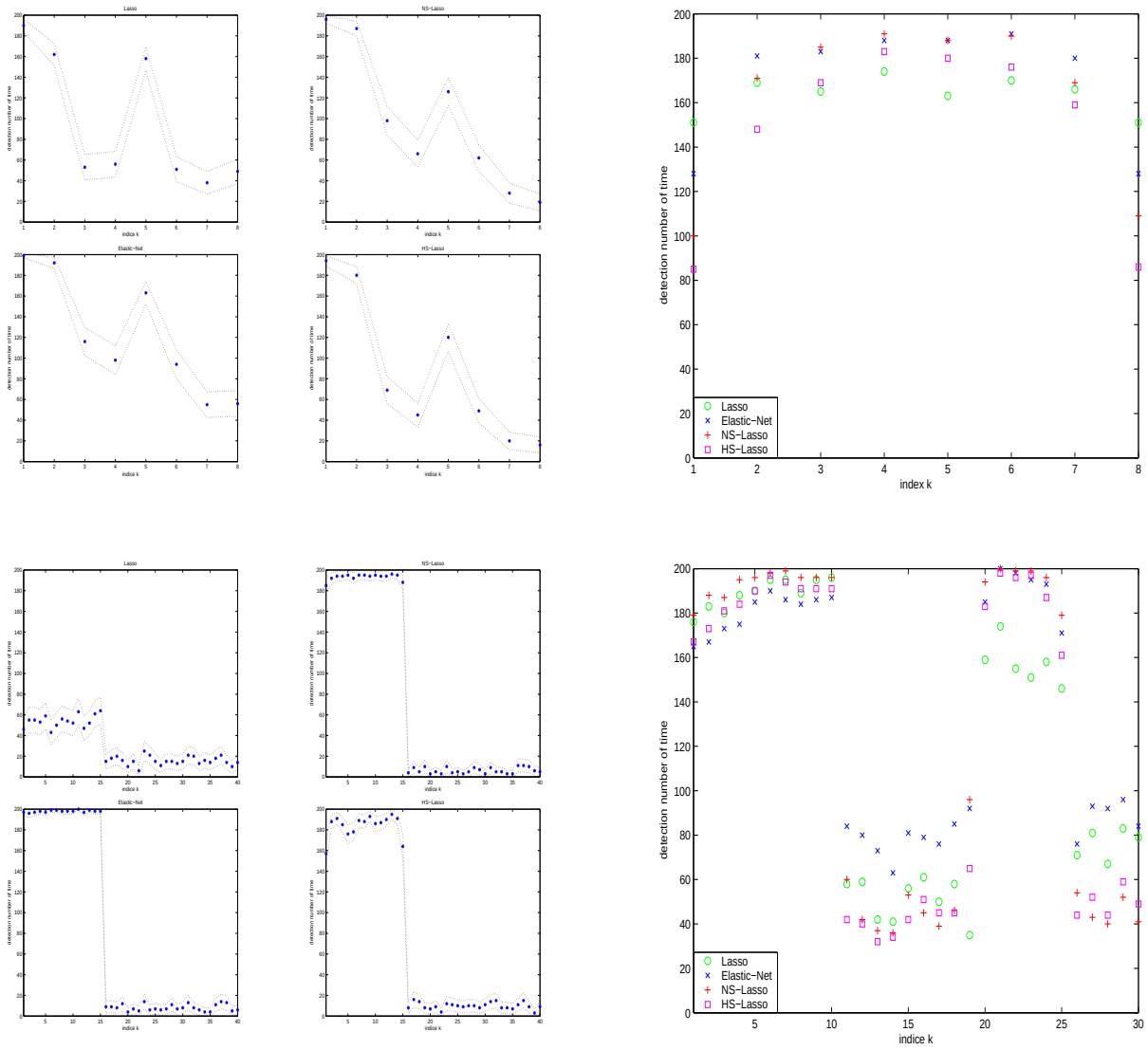

Figure 3: Number of variables detections for each procedure in all the examples (Top-Left: Example (a); Top-Right: Example (b); Bottom-Left: Example (a); Bottom-Right: Example (b))

correctly selects relevant variables but it is also the procedure which selects the more noise variables 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 variables x_{20}, x_{21}, x_{25} and x_{26} (that 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 variables and not only one representant as when we use the Lasso procedure. It also appears that the S-Lasso procedure has very good selection properties according to both relevant and noise variables. However it has slightly worst performance in the borders than in the centers of groups of variables (due to attractions of irrelevant variables). It almost always has a better SNR than the Elastic-Net, so we can take it as a good challenger for this procedure.

7 Conclusion

In this paper, we introduced a new procedure called the Smooth-Lasso which takes into account correlation between successive variables. We established its asymptotic distribution, provided consistency in variable selection results and concluded that the Smooth-Lasso can be both consistent in variable selection and asymptotically normal with a lower rate than $\sqrt{n}$. We also found that regression coefficients of two correlated variables highly depend on the correlation between these variables. Moreover, simulation studies showed that normalized versions of the Smooth-Lasso have nice properties of variable selection which is emphasized when high correlations exist between successive variables. It appears that the Smooth-Lasso almost always outperforms the Lasso and is a good challenger of the Elastic-Net.

Appendix A.

In this appendix we prove the main results:

Proof of Theorem 1. Using the definition of the S-Lasso criterion (2)-(3) we have:

$$-\frac{2}{n}x'_{.,j}(Y - X\hat{\beta}^{SL}) + \lambda\,\mathrm{Sgn}(\hat{\beta}_j^{SL}) + 2\mu\left[2\hat{\beta}_j^{SL} - (\hat{\beta}_{j+1}^{SL} + \hat{\beta}_{j-1}^{SL})\right] = 0 \quad \forall\, j \in \{1, \ldots, p\}$$

where $\hat{\beta}_0^{SL} = \hat{\beta}_1^{SL}$, $\hat{\beta}_{p+1}^{SL} = \hat{\beta}_p^{SL}$ and for any $x \in \mathbb{R}$, $\mathrm{Sgn}(x)$ equals 1, -1 or 0 respectively when x is bigger, smaller or equals 0. Now as in our choices $\hat{\beta}_j^{SL}$ and $\hat{\beta}_k^{SL}$ are non-zero and have the same sign, we then easily obtain

$$n^{-1}\left(x'_{.,j} - x'_{.,k}\right)(Y - X\hat{\beta}^{SL}) + 4\mu\left[\Delta\hat{\beta}_j^{SL} - \Delta\hat{\beta}_k^{SL}\right] = 0.$$

Or equivalently

$$\left[\Delta\hat{\beta}_j^{SL} - \Delta\hat{\beta}_k^{SL}\right] = \frac{1}{4n\mu}\left(x'_{.,j} - x'_{.,k}\right)(Y - X\hat{\beta}^{SL}). \tag{13}$$

On the other hand, by definition, the S-Lasso reaches its minimum at $\hat{\beta}^{SL}$. Hence

$$\|Y - X\hat{\beta}^{SL}\|_n^2 + \lambda|\hat{\beta}^{SL}|_1 + \mu\sum_{j=2}^{p}\left(\hat{\beta}_j^{SL} - \hat{\beta}_{j-1}^{SL}\right)^2 \leq \|Y\|_n^2,$$

and consequently, we obtain

$$\|Y - X\hat{\beta}^{SL}\|_n^2 \leq \|Y\|_n^2, \tag{14}$$

Moreover, since X is standardized,

$$\|x_{.,j} - x_{.,k}\|_n^2 = 2\left(1 - \rho(j,k)\right). \tag{15}$$

Combining (13), (14) and (15), we finally obtain

$$\begin{aligned}\frac{1}{\|Y\|_n}|\Delta\hat{\beta}_j^{SL} - \Delta\hat{\beta}_k^{SL}| &\leq \frac{1}{4n\mu}\sqrt{n}\,\|x_{.,j} - x_{.,k}\|_n\,\frac{\sqrt{n}\|Y - X\hat{\beta}^{SL}\|_n}{\|Y\|_n} \\ &\leq \frac{1}{2\mu}\sqrt{2\left(1 - \rho(j,k)\right)}.\end{aligned}$$

□

Proof of Theorem 2. Let Ψ_n be

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

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 $\hat{u} = \text{Argmin}_u \, \Psi_n(u) = \text{Argmin}_u \, V_n(u)$, 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 $\beta_j^* \neq 0$,

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

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

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

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 $V_n(u) \to V(u)$ 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 ([20]) lead to: $\hat{u}_n \to \text{Argmin}_u V(u)$. □

Proof of Theorem 3. 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*

$$\left. \frac{\|Y - X\beta\|_n^2 + \mu_n \beta' \widetilde{J} \beta}{d\beta_j} \right|_{\beta_j = \hat{\beta}_j^{SL}} = -\lambda_n \, \text{Sgn}(\hat{\beta}_j^{SL}) \qquad \text{for } j : \hat{\beta}_j^{SL} \neq 0, \quad (16)$$

$$\left| \left. \frac{\|Y - X\beta\|_n^2 + \mu_n \beta' \widetilde{J} \beta}{d\beta_j} \right|_{\beta_j = \hat{\beta}_j^{SL}} \right| \leq \lambda_n \qquad \text{for } j : \hat{\beta}_j^{SL} = 0. \quad (17)$$

Recall that $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$, the second result states that if we restrict ourselves to the variables 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 2. We now prove Theorem 3. 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 (16)-(17) 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 (16) is satisfied. We now must check the optimality condition (17). 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

$$n^{-1}(X'Y - X'X_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) = \mathbf{C}_{.,\mathcal{A}}(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) + \mathcal{O}_p(n^{-1/2}). \tag{18}$$

Moreover, the optimality condition (16) for the estimator $\tilde{\beta}$ can be written as

$$n^{-1}(X'_{.,\mathcal{A}}Y - X'_{.,\mathcal{A}}X_{.,\mathcal{A}}\tilde{\beta}_{\mathcal{A}}) = \frac{\lambda_n}{2}\operatorname{Sgn}(\tilde{\beta}_{\mathcal{A}}) - \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}}(\beta^*_{\mathcal{A}} - \tilde{\beta}_{\mathcal{A}}) + \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}}\beta^*_{\mathcal{A}}. \tag{19}$$

Combining (18) and (19), 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 size in the optimality condition (17)

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

converges in probability to

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

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 4. 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 (16) implies

$$\hat{\beta}^{SL}_{\mathcal{A}} = \left(\frac{X'_{.,\mathcal{A}} X_{.,\mathcal{A}}}{n} + \mu_n \widetilde{J}_{\mathcal{A},\mathcal{A}} \right)^{-1} \left(\frac{X'_{.,\mathcal{A}} Y}{n} - \frac{\lambda_n}{2} \operatorname{Sgn}(\hat{\beta}^{SL}_{\mathcal{A}}) \right). \tag{20}$$

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

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

Therefore,

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

with

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

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

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

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

For the other term, as $Y = X\beta^* + \varepsilon$ we have

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

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

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

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 (17) is not satisfied. Then $\hat{\beta}^{SL}$ is not optimal. We get a contradiction, which concludes the proof. □

Proof of Theorem 5. The proof of this theorem is essentially an adaptation of the one concerning the Lasso in [27]. We do not give the whole proof but only mention the important steps and let the reader refer to [27] 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 (20). This proof is the same as the one in [27] so that we omit it here. □

References

[1] Francis Bach. Consistency of the group lasso and multiple kernel learning. *Technical Report*, 2007.

[2] Christophe Chesneau and Mohamed Hebiri. Some theoretical results on the grouped variables lasso. *Technical Report*, 2007.

[3] Arnak Dalalyan and Alexandre 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.

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

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

[6] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression - with discussion. *Ann. Statist.*, 32(2):407–499, 2004.

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

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

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

[10] Stephanie R. Land and Jerome H. Friedman. Variable fusion: a new method of adaptive signal regression. *Technical Report*, 1996.

[11] Nicolai Meinshausen. Lasso with relaxation. *Technical Report*, 2005.

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

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

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

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

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

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

[18] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(1):91–108, 2005.

[19] 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.

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

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

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

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

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

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

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

[27] Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192, 2007.