

# The Smooth-Lasso and other $\ell_1 + \ell_2$ -penalized methods

Mohamed Hebiri and Sara van de Geer

## Abstract

We consider the linear regression problem in the high dimensional setting, i.e., the number  $p$  of covariates can be much larger than the sample size  $n$ . In such a situation one often assumes sparsity of the regression vector, i.e., that it contains many zero components. We propose a Lasso-type estimator  $\hat{\beta}^{Quad}$  (where ‘*Quad*’ stands for quadratic), which is based on two penalty terms. The first one is the  $\ell_1$  norm of the regression coefficients used to exploit the sparsity of the regression as done by the Lasso estimator, whereas the second is a quadratic penalty term introduced to capture some additional information on the setting of the problem. We detail two special cases: the Elastic-Net  $\hat{\beta}^{EN}$ , introduced in [39], deals with sparse problems where correlations between variables may exist; and the S-Lasso<sup>1</sup>  $\hat{\beta}^{SL}$ , which responds to sparse problems where successive regression coefficients are known to vary slowly (in some situations, this can also be interpreted in terms of correlations between successive coefficients). From a theoretical point of view, we establish variable selection consistency results and show that  $\hat{\beta}^{Quad}$  achieves a Sparsity Inequality, i.e., a bound in terms of the number of non-zero components of the ‘true’ regression vector. These results are provided under a weaker assumption on the Gram matrix than the one used by the Lasso. In some (bad) situations this guarantees a significant improvement over the Lasso. Furthermore, a simulation study is conducted and shows that when we consider the estimation accuracy, the S-Lasso  $\hat{\beta}^{SL}$  performs better than known methods as the Lasso, the Elastic-Net  $\hat{\beta}^{EN}$ , and the Fused-Lasso (introduced in [28]), specifically when the regression vector is ‘smooth’, i.e., when the variations between successive coefficients of the unknown parameter of the regression are small. The study also reveals that the theoretical calibration of the tuning parameters imply a S-Lasso solution with close performance to the S-Lasso when the tuning parameters are chosen by 10 fold cross validation.

**Keywords:** Lasso, Elastic-Net, LARS, Sparsity, Variable selection, Restricted eigenvalues, High-dimensional data.

**AMS 2000 subject classifications:** Primary 62J05, 62J07; Secondary 62H20, 62F12.

## 1 Introduction

We focus on the usual linear regression model:

$$y_i = x_i \beta^* + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where the design  $x_i = (x_{i,1}, \dots, x_{i,p}) \in \mathbb{R}^p$  is deterministic,  $\beta^* = (\beta_1^*, \dots, \beta_p^*)' \in \mathbb{R}^p$  is the unknown parameter and  $\varepsilon_1, \dots, \varepsilon_n$  are independent identically distributed (i.i.d.) centered Gaussian random variables with known variance  $\sigma^2$ . We wish to estimate  $\beta^*$  in the sparse case, that is when many of its unknown components equal zero. Thus only a subset of the design covariates  $(X_j)_j$  is truly of interest where  $X_j = (x_{1,j}, \dots, x_{n,j})'$ ,  $j = 1, \dots, p$ . Moreover we are

---

<sup>1</sup>The S-Lasso estimator has initially been introduced in the paper titled *Regularization with the Smooth-Lasso procedure*, in [14]. Results can be found there for the this method which are not provided here, such as the theoretical performance when  $p \leq n$  and a simulation study from a variable selection point of view.

interested in the high dimensional problem where  $p \gg n$ , so that we 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:

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

where  $X = (x'_1, \dots, x'_n)'$ ,  $Y = (y_1, \dots, y_n)'$  and  $\operatorname{pen} : \mathbb{R}^p \rightarrow \mathbb{R}$  is a positive convex function called the penalty. For any vector  $a = (a_1, \dots, a_n)'$ , we have adopted the notation  $\|a\|_n^2 = n^{-1} \sum_{i=1}^n |a_i|^2$  (we denote by  $\langle \cdot, \cdot \rangle_n$  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 (see for instance [9, 13, 30]) are often computationally hard to optimize. Lasso-type procedures (modifications of the  $\ell_1$  penalized least square (Lasso) estimator introduced by [27]) have been extensively studied during the last few years. Between many others, see [3, 4, 7, 37] 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 this paper, we propose a novel estimator, denoted by  $\hat{\beta}^{Quad}$ , which is modification of the Lasso. 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 quadratic penalty  $\beta' \mathbf{J} \mathbf{J} \beta$  for some  $p \times m$  matrix  $\mathbf{J}$  ( $m \in \mathbb{N}^*$ ).

The matrix  $\mathbf{J}$  typically reflects some underlying geometry or structure in the true signal. More generally, the matrix  $\mathbf{J}$  can be chosen so that sparsity of  $\beta^*$  translates to some other desired behavior; this will depend, of course, on the context. There is a wide variety of interesting applications, and what we present below is not meant to be an exhaustive list, but rather a small set of illustrative examples that motivated our work on this problem in the first place.

We add this second term to the Lasso procedure for two major issues. First we exploit this second penalty in order to take into account some prior information on the data or the regression vector that the Lasso may not (as correlation between variables or a specified structure on the regression vector). Second the quadratic penalty is introduced to overcome (or to reduce) theoretical problems observed by the Lasso estimator. Indeed, in several works ([3, 4, 18, 21, 32, 34, 37, 38] among others) conditions to guarantee good performance in prediction, estimation or variable selection for the Lasso procedure are given. See also [31] for an overview of the conditions used to establish the theoretical results according to the Lasso. It was shown that the Lasso does not always ensure good performance when high correlations exist between the covariates. We establish theoretical results for  $\hat{\beta}^{Quad}$  that states that this estimator guarantees good performance under a weaker assumption than the Lasso estimator. The improvement is specifically observed when the Lasso achieves poor results.

Two particular cases of the estimator  $\hat{\beta}^{Quad}$  are mainly considered: the Elastic-Net, introduced in [39] to deal with problems where correlations between variables exist. It is defined with the quadratic penalty term  $\sum_{j=1}^p \beta_j^2$ . The second and novel procedure is called the *Smooth-Lasso* (*S-Lasso*) estimator. It is defined with the  $\ell_2$ -fusion penalty, i.e.,  $\sum_{j=2}^p (\beta_j - \beta_{j-1})^2$ . The  $\ell_2$ -fusion penalty was first introduced in [17]. This term helps to tackle situations where the regression vector is structured such that its coefficients vary slowly. Let us say in this case that the regression vector is ‘smooth’. Note however that our theoretical study takes into account a large amount of procedures such as the closely related procedure ‘Weighted Fusion’ introduced in [10], as detailed in Remark 1.

The main contribution of the paper is the introduction of the Smooth-Lasso estimator

which significantly improves (both in theory and in practice) the performance of the Lasso and the Elastic-Net in some situations. However the method appears as a special case of the estimator  $\hat{\beta}^{Quad}$ . This type of estimators aims to:

- capture the sparsity and some other structure (smoothness in the case of the S-Lasso);
- reduce the assumptions on the Gram matrix and provide theoretical guarantees in situations that are not suitable for the Lasso (correlations between successive covariates in the case of the S-Lasso).

From a practical point of view, some problems are also encountered when we solve the Lasso criterion (for instance with the LARS algorithm [12]). Indeed this algorithm fails to select a complete group of correlated covariates. Two major lacks follow. First the Lasso is not consistent neither in variable selection nor in estimation (bad reconstitution of  $\beta^*$ ). In this paper we focus on the estimation issue. We consider the case where the regression vector  $\beta^*$  is structured. We invoke the *S-Lasso* estimator to respond to such problems where the covariates are ranked so that the regression vector is ‘smooth’ (i.e., the vector  $\beta^*$  consists in small variations in its successive components). We will see through simulations that such situations support the use of the *S-Lasso* estimator. This estimator is inspired by the *Fused-Lasso* [28]. Both S-Lasso and Fused-Lasso combine a  $\ell_1$ -penalty with a fusion term [17]. The fusion term is suggested to make successive coefficients as close as possible to each other. The main difference between the two procedures is that we use the  $\ell_2$  distance between the successive coefficients (i.e., the  $\ell_2$ -fusion penalty) whereas the Fused-Lasso uses the  $\ell_1$  distance (i.e., the  $\ell_1$ -fusion penalty:  $\sum_{j=2}^p |\beta_j - \beta_{j-1}|$ ). Hence, compared to the Fused-Lasso, we sacrifice sparsity in changes between successive coefficients in the estimation of  $\beta^*$  in favor of an easier optimization due to the strict convexity of the  $\ell_2$  distance. This implies a large reduction of computational cost. However, sparsity is yet ensured by the Lasso penalty. The  $\ell_2$ -fusion penalty helps to provide ‘smooth’ solutions. Consequently, even if there is no perfect match between successive coefficients our results are still interpretable. From a theoretical point of view, the  $\ell_2$  distance also helps us to provide theoretical properties for the S-Lasso which in some situations appears to outperform the Lasso and the Elastic-Net [39]. Let us mention that variable selection consistency of the Fused-Lasso and the corresponding Fused adaptive Lasso has also been studied in [25] but in a different context from the one in the present paper. The results obtained in [25] 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 by [38], which is similar to the Lasso but with adaptive weights used to penalize each regression coefficient separately. This procedure reaches under certain (strong) conditions, ‘Oracles Properties’ (i.e., consistency in variable selection and asymptotic normality. See [38]). Another approach in the Relaxed Lasso [20], which 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 [33] with the Group-Lasso procedure which selects groups of correlated covariates instead of single covariates at each step. A first step to the variable selection consistency study has been proposed in [1] and Sparsity Inequalities were given in [8, 19]. Another choice of penalty has been proposed with the Elastic-Net [39] which has been studied for instance in [5, 15, 40]. It is in a unified fashion that we shall treat the S-Lasso and the Elastic-Net from a theoretical point of view.

The rest of the paper is organized as follows. In the next section, we introduce the estimator  $\hat{\beta}^{Quad}$  defined with the Lasso penalty on one hand and a quadratic penalty on the other hand. In particular, we define the S-Lasso estimator and the notion of smoothness. We also provide a way to solve the  $\hat{\beta}^{Quad}$  problem with the attractive property of piecewise linearity of its regularization path. Consistency in estimation and variable selection in the high dimensional case are considered in Section 3. We moreover provide some examples in favor of the Elastic-Net and the S-Lasso in Sections 3.1.1- 3.1.2, and technical issues in Section 3.3. We finally give experimental results in Section 4 which display the S-Lasso performance against some popular methods. All proofs are postponed to the Appendix section.

## 2 The S-Lasso procedure

In several applications and settings including macroeconomics, financial time series analysis and biological and medical sciences one often deals with data with given complex attributes. This is the case in trend filtering where the solution is assumed to be ‘smooth’. See [16] for a nice survey. Before going further, let us provide a definition of a ‘smooth’ vector:

**Definition 2.1** (Smoothness). *A vector  $\beta \in \mathbb{R}^p$  is  $\alpha_n$ -smooth (or simply smooth when we do not care of the rate) if*

$$\sum_{j=2}^p (\beta_j - \beta_{j-1})^2 \leq \alpha_n,$$

for some positive sequence  $\alpha_n$  which goes to zero when  $n$  goes to infinity.

In the above applications the regression vector  $\beta^*$  is smooth. Hence it is important to consider estimation methods which can reflect this aspect of the problem. It is often useful to assume that the regression vector is also sparse to be able to treat data such as spectrometry or some genomic ones, where both smoothness and sparsity appear simultaneously. For all these reasons it is worth introducing and analyzing a method which can reconstitute sparse and smooth regression vectors. Hence we define the S-Lasso estimator  $\hat{\beta}^{SL}$  as the solution of the optimization problem (2) when the penalty function is:

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

where  $\lambda$  and  $\mu$  are two positive parameters that control on one hand the sparsity of our estimator and its smoothness on the other hand. For any vector  $a = (a_1, \dots, a_p)'$  and integer  $q$ , we have used the notation  $|a|_q^q = \sum_{j=1}^p |a_j|^q$ . Note that when  $\mu = 0$ , the solution is the Lasso estimator so that it appears as a special case of the S-Lasso estimator. In a more general point of view we consider the following penalty

$$\text{pen}(\beta) = \lambda|\beta|_1 + \mu\beta' \mathbf{J} \mathbf{J} \beta, \quad (4)$$

where  $\mathbf{J}$  is any  $p \times p$  matrix. This penalty is a combination of the Lasso penalty and a quadratic penalty. The matrix  $\mathbf{J}$  typically reflects some underlying geometry or structure in the true signal (we refer to [29] for analogous ideas). Let us call  $\hat{\beta}^{Quad}$  the solution of the minimization problem (2)-(4). Note that the S-Lasso penalty can be seen as a particular case

of the penalty (4) when  $\mathbf{J}$  is given by

$$\mathbf{J} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & -1 & \ddots & \ddots & \vdots \\ 0 & 1 & -1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{pmatrix}, \quad (5)$$

and that the Elastic-Net corresponds to the case where  $\mathbf{J}$  is the identity matrix.

**Remark 1.** For any  $j, k \in \{1, \dots, p\}$ , denote by  $s_{j,k} = \text{sign}\left(\frac{X'_j X_k}{n}\right)$  the sign of the sample correlation between predictor variables  $j$  and  $k$ . Denote also by  $w_{j,k} \geq 0$  some predictor correlation driven weights. Given this notation, the Weighted Fusion introduced in [10] corresponds to the case where the  $k$ -th diagonal terms of  $\mathbf{J}$  equals  $w_{j,k}$  and  $(\mathbf{J})_{k,j} = (\mathbf{J})_{j,k} = -s_{j,k}w_{j,k}$  for  $j \neq k$ .

Now we deal with the solution  $\hat{\beta}^{Quad}$  of (2)-(4) and its computational cost. The following lemma shows that  $\hat{\beta}^{Quad}$  can be expressed as a Lasso solution by augmenting the data artificially.

**Lemma 1.** Given the dataset  $(X, Y)$  and the tuning parameters  $(\lambda, \mu)$ . Define the extended dataset  $(\tilde{X}, \tilde{Y})$  and  $\tilde{\varepsilon}$  by

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{n\mu}\mathbf{J} \end{pmatrix}, \quad \text{and} \quad \tilde{Y} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \tilde{\varepsilon} = \begin{pmatrix} \varepsilon \\ -\sqrt{n\mu}\mathbf{J}\beta^* \end{pmatrix},$$

where  $\mathbf{0}$  is a vector of size  $p$  containing only zeros,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$  is the noise vector and  $\mathbf{J}$  is the  $p \times p$  matrix given by the penalty (4) ( $\mathbf{J}$  is given by (5) in the case of the S-Lasso estimator). Then we have  $\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon}$ , and the estimator  $\hat{\beta}^{Quad}$ , solution of the minimization problem (2) with the penalty given by (4) (in the case of the S-Lasso, the penalty is given by (3)), is also the minimizer of the following Lasso-criterion

$$\frac{1}{n} \left| \tilde{Y} - \tilde{X}\beta \right|_2^2 + \lambda |\beta|_1. \quad (6)$$

This result is a consequence of simple algebra. It motivates the following comments on the estimator  $\hat{\beta}^{Quad}$ .

**Remark 2** (Regularization paths). LARS is an iterative algorithm introduced in [12]. A modification of LARS can be used to construct  $\hat{\beta}^{Quad}$ . For a fixed  $\mu$  (appearing in (3)), it constructs at each step an estimator based on the correlation between covariates and the current residual. Each step corresponds to a value of  $\lambda$ . Then for a fixed  $\mu$ , we get the evolution of the coefficients values of  $\hat{\beta}^{Quad}$  when  $\lambda$  varies. This evolution describes the regularization paths of  $\hat{\beta}^{Quad}$  which are piecewise linear ([26]). This property implies that (again for fixed  $\mu$ ) the S-Lasso problem can be solved with the same computational cost as the ordinary least square (OLS) estimate using the LARS algorithm.

### 3 Theoretical results in the high dimensional setting

In this section, we study the performance of the estimator  $\hat{\beta}^{Quad}$  in the high dimensional case. In particular, we provide a non-asymptotic bound on the squared risk. We also provide a

bound on the  $\ell_2$  estimation error of  $\hat{\beta}^{Quad}$ . This last result implies in particular the variable selection consistency of  $\hat{\beta}^{Quad}$ . Let

$$\tilde{J} = \mathbf{J}'\mathbf{J},$$

be the  $p \times p$  matrix where  $\mathbf{J}$  is the matrix appearing in the quadratic penalty (4). Since our main interest is devoted to the study of the S-Lasso estimator, we first focus on the case where the matrix  $\tilde{J}$  is sparse. We refer the reader to Section 3.3, where we address several technical points, among those is the study of the case where the matrix  $\tilde{J}$  is general.

All the results of this section are proved in Section 6. These theoretical contributions rely partly on Lemma 1. Let us finally mention that the tuning parameters  $\lambda$  and  $\mu$  will actually be chosen depending on the sample size  $n$ . We emphasize this dependency by adding a subscript  $n$  to these parameters.

### 3.1 Sparsity Inequality when $\tilde{J}$ is sparse

Now we establish a Sparsity Inequality (SI) achieved by the estimator  $\hat{\beta}^{Quad}$ , that is a bound on the squared risk that takes into account the sparsity of the regression vector  $\beta^*$ . More precisely, we prove that the rate of convergence of  $\hat{\beta}^{Quad}$  is  $\max(|\mathcal{A}^*| \log(n)/n; \mu_n^2 |\tilde{J}\beta^*|^2)$ , where  $\mathcal{A}^*$  is the sparsity set  $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$ . Then this rate depends not only on the sparsity index but also on  $|\tilde{J}\beta^*|$ . In the case of the S-Lasso, this last quantity is related to the smoothness of the vector  $\beta^*$ . Let us first establish the assumptions needed, and the setup of this contribution. Let  $\eta \in (0, 1)$  be a real number. We define the tuning parameter  $\lambda_n$  as follows

$$\lambda_n = 4\sqrt{2}\sigma \sqrt{\frac{\log(p/\eta)}{n}}, \quad (7)$$

and leave the calibration of  $\mu_n$  free. We discuss later (see Corollary 1 for instance) the choice for this last parameter. Our assumption on the Gram matrix  $\Psi^n$  involves the symmetric  $p \times p$  matrix  $K_n$  defined by

$$K_n = \Psi^n + \mu_n \tilde{J}. \quad (8)$$

Given the augmented dataset defined in Lemma 1, we note that  $K_n = n^{-1} \tilde{X}' \tilde{X}$ , which can be seen as an augmented Gram matrix. Let  $\Theta \subset \{1, \dots, p\}$  a set of indices. Using this notation, we formulate the following assumption:

**Assumption  $B(\Theta)$ :** Let  $K_n$  be the matrix given by (8) and let  $\varrho_n = 4\sqrt{|\mathcal{A}^*|} + \frac{4\mu_n}{\lambda_n} |\tilde{J}\beta^*|_2$ . There is a constant  $\phi > 0$  such that, for any  $\Delta \in \mathbb{R}^p$  that satisfies  $\sum_{j \notin \Theta} |\Delta_j| \leq \varrho_n \sqrt{\sum_{j \in \Theta} \Delta_j^2}$ , we have

$$\Delta' K_n \Delta \geq \phi \sum_{j \in \Theta} \Delta_j^2. \quad (9)$$

Here are some comments about this assumption:

- first of all, Assumption  $B(\Theta)$  is inspired by the Restricted Eigenvalue (RE) Assumption introduced in [3]. This RE Assumption is widely used in the literature. We then refer the reader to [3, 31] for instance for more details on this assumption. The main difference with the assumption we use is that in [3], the authors consider the case where  $K_n = \Psi^n$ , which matches with the Lasso estimator (that is  $\mu_n = 0$  in our setting);
- another minor difference is that the set on which the assumption should hold is larger in Assumption  $B(\Theta)$  than in the RE Assumption. Indeed, in Assumption  $B(\Theta)$ , the considered vectors  $\Delta$  should be such that  $\sum_{j \notin \Theta} |\Delta_j| \leq \varrho_n \sqrt{\sum_{j \in \Theta} \Delta_j^2}$ , whereas in [3],

the authors only need to consider vectors  $\Delta$  such that  $\sum_{j \notin \Theta} |\Delta_j| \leq cst \cdot \sum_{j \in \Theta} |\Delta_j|$  (see also [31]). We make this set larger to allow large values of the tuning parameter  $\mu_n$ . We will explain later why this is desirable;

- only small subsets of indices  $\Theta$  are considered in Assumption  $B(\Theta)$ . More precisely, let  $\mathcal{B} \subset \{1, \dots, p\}$  be a set of indices such that it includes  $\mathcal{A}^*$ , the true sparsity set. This set depends on  $\tilde{J}$  and on  $\mathcal{A}^*$ , and the sparser  $\tilde{J}$ , the smaller  $\mathcal{B}$ . For instance, in the case of the Elastic-Net,  $\mathcal{B} = \mathcal{A}^*$ , and in the case of the S-Lasso (that we will detail later), the set  $\mathcal{B}$  is such that  $|\mathcal{B}| \leq 3|\mathcal{A}^*|$ . Of course, the definition of  $\mathcal{B}$  depends on  $\mathcal{A}^*$ , but here we are only interested in the magnitude of  $|\mathcal{B}|$ . Thanks to the sparsity of  $\tilde{J}$ , we can assume that there exists a constant  $c_{\tilde{J}} \geq 1$  such that  $|\mathcal{B}| \leq c_{\tilde{J}}|\mathcal{A}^*|$ .

Given this new notation we can establish the main result which deals with the case of sparse matrices  $\tilde{J}$ :

**Theorem 1** ( $\tilde{J}$  sparse). *Let  $\mathcal{A}^*$  be the sparsity set. Let the tuning parameters  $(\lambda_n, \mu_n)$  be defined as in (7). Suppose that Assumption  $B(\mathcal{B})$  is satisfied with a set  $\mathcal{B} \supset \mathcal{A}^*$  such that  $|\mathcal{B}| \leq c_{\tilde{J}}|\mathcal{A}^*|$  for a given constant  $c_{\tilde{J}} \geq 1$ . Then with probability greater than  $1 - \eta$ , we have*

$$\begin{aligned} \left\| X\beta^* - X\hat{\beta}^{Quad} \right\|_n^2 &\leq \phi^{-1} (2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{J}\beta^*|_2)^2, \\ (\beta^* - \hat{\beta}^{Quad})' \tilde{J} (\beta^* - \hat{\beta}^{Quad}) &\leq \phi^{-1} \frac{(2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{J}\beta^*|_2)^2}{\mu_n}, \end{aligned} \quad (10)$$

and

$$|\beta^* - \hat{\beta}^{Quad}|_1 \leq 2\phi^{-1} \frac{(2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{J}\beta^*|_2)^2}{\lambda_n}.$$

Theorem 1 states that  $\hat{\beta}^{Quad}$  achieves the same SI which also brings into play the quantity  $|\tilde{J}\beta^*|_2$ . In the case, where  $K_n$  is invertible, the condition (9) is always satisfied for any  $\Delta \in \mathbb{R}^p$  with  $\phi$  larger than the smallest eigenvalue of  $K_n$ . Denote by  $\phi_0$  the quantity  $\phi$  in (9), when  $\mu_n = 0$  (and then  $K_n = \Psi^n$ ). When we consider the Lasso estimator, which corresponds to the case  $K_n = \Psi^n$ , the quantity  $\phi_0$  may be very small. On the other hand there might exist values of  $\mu_n$  that make  $\phi$  larger than  $\phi_0$ . This may help to state better bounds for  $\hat{\beta}^{Quad}$  than for the Lasso. Hence, larger values for  $\mu_n$  are desired, in order to control suitably the eigenvalues of  $K_n$ .

Let us now consider good choices for  $\mu_n$ . We focus here on the  $\ell_1$  estimation error (the same reasoning is true for the other errors). The rate of convergence is

$$\begin{cases} \frac{\lambda_n}{\phi} |\mathcal{A}^*| & \text{if } \mu_n |\tilde{J}\beta^*|_2 = \mathcal{O}(\lambda_n \sqrt{|\mathcal{A}^*|}) \text{ or even smaller in order,} \\ \frac{\mu_n^2}{\phi \lambda_n} |\tilde{J}\beta^*|_2^2 & \text{otherwise.} \end{cases}$$

Then the rate of convergence is worst than the usual one when  $\mu_n |\tilde{J}\beta^*|_2 \gg \lambda_n \sqrt{|\mathcal{A}^*|}$ . Given the above comments on  $\phi$  and on the rate of convergence, we propose the following compromise for the calibration of  $\mu_n$ :

**Corollary 1.** *In the same setting as in Theorem 1. Let  $\lambda_n = 4\sqrt{2}\sigma \sqrt{\frac{\log(p/\eta)}{n}}$  with  $\eta \in (0, 1)$ , and  $\mu_n = \frac{\lambda_n \sqrt{|\mathcal{A}^*|}}{2|\tilde{J}\beta^*|_2}$ . Then  $\varrho_n = 6\sqrt{|\mathcal{A}^*|}$  in Assumption  $B(\mathcal{B})$  and with probability greater than  $1 - \eta$ , we have*

$$\left\| X\beta^* - X\hat{\beta}^{Quad} \right\|_n^2 \leq \frac{36\sqrt{2}\sigma^2 \log(p/\eta)}{\phi} \frac{|\mathcal{A}^*|}{n},$$

and

$$|\beta^* - \hat{\beta}^{Quad}|_1 \leq \frac{72\sqrt{2}\sigma}{\phi} \sqrt{\frac{\log(p/\eta)}{n}} |\mathcal{A}^*|.$$

In our simulation study, we focus on this particular choice of  $\mu_n$ . However, in real applications, since it depends on the unknown regression vector  $\beta^*$ , we tune the parameters  $\lambda_n$  and  $\mu_n$  through a 2D ten fold cross validation over a grid.

In the above bound, which is almost the same as the Lasso, the influence of the second penalty term appears directly in  $\phi$ . We refer to Section 3.3 for other choices of  $\mu_n$  which are more suitable when we deal with a general (non sparse) matrix  $\tilde{J}$ .

**Remark 3.** *Corollary 1 states that  $\hat{\beta}^{Quad}$  improves the performance of the Lasso thanks to the quantity  $\phi$  introduced in Assumption  $B(\mathcal{B})$ . Let us denote by  $\phi_0$  the  $\phi$  obtained when  $\mu_n = 0$  (corresponding to the Lasso case). Since this quantity appears in the upper bound, we observe that the improvement using  $\hat{\beta}^{Quad}$  is significant in particular when the Lasso behaves poorly, that is when  $\phi_0$  is very small (much smaller than  $\mu_n$ ). Indeed, let us consider the clearer case, where  $\tilde{J}$  is diagonal (for instance the identity matrix corresponding to the Elastic-Net estimator) and then  $\mathcal{B} = \mathcal{A}^*$ . Then, Assumption  $B(\mathcal{A}^*)$  guaranties that  $\phi$  is at least  $\mu_n = cst \cdot \lambda_n$  and then, for instance, the prediction error  $\|X\beta^* - X\hat{\beta}^{Quad}\|_n^2$  is bounded by  $cst \cdot \sqrt{\log(p)}|\mathcal{A}^*|/n$ . Although not optimal, this bound is much better than to one achieved by the Lasso.*

**Remark 4.** *From the proofs of Theorem 1, the constant in the definition of the tuning parameter  $\lambda_n$  can be taken equal  $2\sqrt{2}$  instead of  $4\sqrt{2}$ . Such tuning is possible if we consider only the prediction error  $\|X\beta^* - X\hat{\beta}^{Quad}\|_n^2$  (cf. Proposition 1).*

In the next two paragraphs we treat the special cases of the Elastic-Net and the S-Lasso estimators.

### 3.1.1 Elastic-Net

The Elastic-Net corresponds to the case where  $\tilde{J}$  equals the identity matrix. The theoretical performance of this estimator has already been considered in papers as [5, 15]. In [15], the authors considered a version of the Irrepresentable Condition to establish their consistency results. This necessary and (almost) sufficient assumption for the variable selection task is harder to interpret than ours. The result in the present paper about the Elastic-Net are quite close to those in [5].

For any vector  $b \in \mathbb{R}^p$  and subset  $\Theta \subset \{1, \dots, p\}$ , let  $b_\Theta$  be the vector in  $\mathbb{R}^p$  such that  $(b_\Theta)_j = b_j$  if  $j \in \Theta$  and zero otherwise. According to Corollary 1, the Elastic-Net satisfies a Sparsity Inequality with  $\mathcal{B} = \mathcal{A}^*$ . Let us briefly compare the above results to those obtained in [5]. On one hand, Inequality (9) can be written as

$$\Delta' \Psi^n \Delta \geq (\phi - \mu_n) |\Delta_{\mathcal{A}^*}|_2^2 - \mu_n |\Delta_{(\mathcal{A}^*)^c}|_2^2, \quad (11)$$

where  $\Delta \in \Gamma := \{\Delta \in \mathbb{R}^p : |\Delta_{(\mathcal{A}^*)^c}|_1 \leq 6\sqrt{|\mathcal{A}^*|} |\Delta_{\mathcal{A}^*}|_2\}$ . Then Assumption  $B(\mathcal{A}^*)$  is close to *Condition Stabil* in [5, page 4] with however a few differences:

- The set  $\Gamma$  we use is larger than the set  $V_{4,0} := \{\Delta \in \mathbb{R}^p : |\Delta_{(\mathcal{A}^*)^c}|_1 \leq 4|\Delta_{\mathcal{A}^*}|_1\}$  in [5] since we use quadratic set of inequalities in  $\Gamma$ . Even though this difference is small, let us mention that we will establish in Section 3.3 theoretical guarantees which also require the same set  $V_{4,0}$ .

– In *Condition Stabil*, Equation (11) is replaced by

$$\Delta' \Psi^n \Delta \geq (\phi^{CS} - \mu_n) |\Delta_{\mathcal{A}^*}|_2^2,$$

where  $\Delta \in V_{4,0}$  and then we also have  $\phi^{CS} > \mu_n$ , but  $\mu_n$  is a bit smaller in [5] than in our study. Indeed,  $\mu_n = \frac{\lambda_n}{2|\beta^*|_\infty}$  in [5], whereas we replaced the  $\ell_\infty$  norm  $|\beta^*|_\infty$  by the mean  $\frac{|\beta^*|_2}{\sqrt{|\mathcal{A}^*|}}$ . This difference is not huge in the case of the Elastic-Net but can be noticed in some cases. Actually, in our setting, we have a bit more. Indeed,  $\phi$  can be much larger than  $\phi^{CS}$  since we subtract the term  $\mu_n |\Delta_{(\mathcal{A}^*)^c}|_2^2$  in (11), which can be large thanks to  $\mu_n$  (we expect  $|\Delta_{(\mathcal{A}^*)^c}|_2^2$  to be small).

According to the bounds for the  $\ell_1$  estimation error for instance, apart from the difference that  $\phi^{CS} < \phi$ , the bounds can be seen to be equivalent.

Let us finally mention that thanks to the fact that  $\phi > \phi_0$ , we conclude that the Elastic-Net achieves better performance than the Lasso. This is true even if we do not show that the Elastic-Net is particularly useful when correlations between variables exist.

Finally, we observe that in this case, Equation (10) is nothing but a SI on the  $\ell_2$  estimation error  $|\beta^* - \hat{\beta}^{Quad}|_2^2$ . Note however that the rate  $\lambda_n \sqrt{|\mathcal{A}^*|}$  (when  $\mu_n$  is defined as in Corollary 1) is not optimal, but has the advantage of not requiring a more restrictive assumption than Assumption  $B(\mathcal{A}^*)$ . Imposing Assumption  $B(\mathcal{B})$  to be satisfied with a set  $\mathcal{B}$  larger  $\mathcal{A}^*$ , a better rate of convergence can be reached (cf. Proposition 1).

### 3.1.2 Smooth-Lasso

The S-Lasso corresponds to the case where  $\tilde{\mathcal{J}}$  is given by (5). This estimator responds to problems where the regression vector is expected to be  $\alpha_n$ -smooth according to Definition 2.1. That is  $|\mathbf{J}\beta^*|_2 = \sum_{j=2}^p (\beta_j^* - \beta_{j-1}^*)^2 \leq \alpha_n$ . As a consequence we have the following worst case relation:  $|\tilde{\mathcal{J}}\beta^*|_2 \leq 7|\mathbf{J}\beta^*|_2$  (the constant 7 comes from computation and is not very accurate). Note also that in this case Assumption  $B(\Theta)$  is satisfied with a set  $\Theta = \mathcal{B}$  whose size is less than three times larger than  $\mathcal{A}^*$ . This set can be expressed by

$$\mathcal{B} = \{j \in \{2, \dots, p-1\} : \beta_j^* \neq 0, \beta_{j-1}^* \neq 0 \text{ or } \beta_{j+1}^* \neq 0\},$$

and Theorem 1 holds with  $c_{\tilde{\mathcal{J}}} = 3$ . Moreover, Equation (10) can be seen as a control on the ‘smoothness’ error  $\sum_{j=2}^p (\delta_j - \delta_{j-1})^2$ , where  $\delta_j$  is the components difference  $\beta_j^* - \hat{\beta}_j^{Quad}$ .

The S-Lasso is devoted to provide a smooth and sparse solution. This is true whatever the correlations between variables. However, it is interesting to remark how the smoothness has quite close interaction with correlations between successive variables. Indeed, when we deal with the S-Lasso estimator, the matrix  $\tilde{\mathcal{J}}$  is tridiagonal with its off-diagonal terms equal to  $-1$ . If we do not consider the diagonal terms, we remark that  $\Psi^n$  and  $K_n$  differ only in the terms on the second diagonals (i.e.,  $(K_n)_{j-1,j} \neq (\Psi^n)_{j-1,j}$  for  $j = 2, \dots, p$  as soon as  $\mu_n \neq 0$ ). Terms in the second diagonals of  $\Psi^n$  correspond to correlations between successive covariates.

When high correlations exist between successive covariates, a suitable choice of  $\mu_n$  makes Assumption  $B(\mathcal{B})$  satisfied. Hence, this assumption fits well with the setup where correlations between successive variables interfere. In many situations, we expect that the variables are ranked, such that not only the regression vector is ‘smooth’, but also successive covariates are correlated. In this case the S-Lasso estimator is particularly useful. We also observe how the ‘smoothness’ of the regression vector influences the control of the correlation on one hand (see Assumption  $B(\mathcal{B})$ ), and the prediction and the estimation errors on the other hand (as  $\phi$  depends on  $|\mathbf{J}\beta^*|_2$ ).

**Example.** Assume that  $n/4$  is an integer. First of all, let us define a smooth regression vector  $\beta^*$  with  $n/2$  non-zero components such that

$$\beta_j^* = 1 \quad \text{for } j = 1, \dots, n/4 - 1, \quad \text{and} \quad \beta_j^* = 1 - \frac{4}{n} \left( j - \frac{n}{4} \right) \quad \text{for } j = n/4, \dots, n/2.$$

This regression vector is piecewise linear (particular case of smoothness) to make the idea clear and for simplicity of computations. The vector is such that

$$|\beta^*|_2 = \sqrt{\frac{n}{3} - \frac{1}{2} + \frac{2}{3n}} = \mathcal{O}(\sqrt{n}), \quad \text{and} \quad |\mathbf{J}\beta^*|_2 = \sqrt{\frac{4}{n} - \frac{16}{n^2}} = \mathcal{O}(1/\sqrt{n}).$$

Then we can set the smoothness parameter  $\alpha_n = 4/\sqrt{n}$  in Definition 2.1.

Let us now consider the design matrix  $\Psi^n$ . Let  $\epsilon > 0$  be a real number. Let  $\Psi^n$  be a tridiagonal Gram matrix with diagonal elements equal 1 (normalized) and such that  $\Psi_{j,j-1}^n = \Psi_{k,k+1}^n = \epsilon$  for  $j = 2, \dots, p$  and  $k = 1, \dots, p-1$ . In such a case the spectrum of the Gram matrix lies in  $[1 - 2\epsilon, 1 + 2\epsilon]$ . Then  $\phi_0 \geq 1 - 2\epsilon$  (the  $\phi$  corresponding to the Lasso estimate ; that is when  $\mu_n = 0$ ). However, we do not know how far  $\phi_0$  is from  $1 - 2\epsilon$  and then we can only say the the prediction error of the Lasso  $\hat{\beta}^L$  is such that with high probability

$$\left\| X\beta^* - X\hat{\beta}^L \right\|_n^2 \leq \frac{16\sqrt{2}\sigma^2 \log(p/\eta)}{1 - 2\epsilon} \frac{|\mathcal{A}^*|}{n} = \mathcal{O}(\sigma^2 |\mathcal{A}^*|),$$

with the choice  $\epsilon = \frac{1}{2} - \frac{\log(p/\eta)}{2n}$ . Actually the above bound does not provide any control on the prediction error of the Lasso estimator.

Let us now focus on the Elastic-Net estimate  $\hat{\beta}^{EN}$ . According to Assumption  $B(\mathcal{A}^*)$ , we have to consider the spectrum of the matrix  $K_n^{EN} = \Psi^n + \mu_n I_p$ , where  $I_p$  is the identity matrix in  $\mathbb{R}^p$ . This spectrum lies in  $[1 - 2\epsilon + \mu_n, 1 + 2\epsilon + \mu_n]$ . Given the value of  $\epsilon$  and the  $|\beta^*|_2$ , we get the control

$$\left\| X\beta^* - X\hat{\beta}^{EN} \right\|_n^2 \leq \frac{1}{1 - 2\epsilon + \mu_n} (2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\beta^*|_2)^2 = \mathcal{O}(\sigma \sqrt{\log(p) |\mathcal{A}^*|}),$$

where we used the definition of  $\mu_n$  provided in Corollary 1. Let us mention that different value for  $\mu_n$  does not improve the bound. Hence in this case the Elastic-Net estimator does not control neither the prediction error.

On the other hand, in the case of the S-Lasso  $\hat{\beta}^{SL}$ , the eigenvalues of the matrix  $K_n^{SL} = \Psi^n + \mu_n \tilde{\mathcal{J}}$  lies in  $[1 + \mu_n - 2|\epsilon - \mu_n|, 1 + 2\mu_n + 2|\epsilon - \mu_n|]$ . We refer to [35] for more details on the eigenvalues of tridiagonal matrices. This interval is up to constants in the same order than the one of the Elastic-Net. By the sequel, we have the following control for the S-Lasso estimator (when  $\epsilon > \mu_n$ , otherwise the control is even better)

$$\left\| X\beta^* - X\hat{\beta}^{SL} \right\|_n^2 \leq \frac{1}{1 - 2\epsilon + 3\mu_n} (2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{\mathcal{J}}\beta^*|_2)^2 = \mathcal{O} \left( \sigma \frac{\sqrt{\log(p) |\mathcal{A}^*|}}{n} \right),$$

where here again, we considered the value of  $\mu_n$  given in Corollary 1. In this smooth context, the S-Lasso, is obviously the best method (in comparison to the Lasso and the Elastic-Net). Note that the last rate is better than the minimax rate under the sparsity assumption, that is  $\frac{\log(p/|\mathcal{A}^*|+1)|\mathcal{A}^*|}{n}$ . This is due to the fact that we also imposed a smoothness assumption, which has been nicely exploited by the S-Lasso estimator. Thus, the above minimax lower bounds cannot be applied anymore.

Let us conclude by the following remarks: in the above situation, we assume that the regression vector is smooth, but also that the successive covariates are correlated. This is the

best context for the Smooth-Lasso.

In the case where the regression vector is smooth, but we do not have a particular structure in the Gram matrix (say the variables are independent and  $\phi_0$  is a fixed positive constant), then the Lasso and the Elastic-Net (for instance with the value of  $\mu_n$  given in Corollary 1) reach the rate  $\sigma^2 \frac{\log(p)|\mathcal{A}^*|}{n}$ . Contrarily to the Elastic-Net, the value of  $\mu_n$  in the case of the S-Lasso which dependent on  $\alpha_n$  makes the bound better. Here again, if we consider the same regression vector  $\beta^*$  as in the above example, the rate is in order  $\mathcal{O}\left(\sigma \frac{\sqrt{\log(p)|\mathcal{A}^*|}}{n}\right)$ . We then have here again better performance than the Elastic-Net and the Lasso.

On the other hand, when the regression vector is not smooth (say,  $|\beta^*|_2$  and  $|\mathbf{J}\beta^*|_2$  are constants) and the design matrix is for instance as in the above example, the Lasso is not suitable. On the other hand, both of the Elastic-Net and the S-Lasso have comparable performance and their bound is in order  $\mathcal{O}(\sqrt{\log(p)|\mathcal{A}^*|/n})$ , which is much better than the Lasso (even not optimal).

### 3.2 Variable selection

Now we deal with variable selection. Let us first mention that the estimator  $\hat{\beta}^{Quad}$  and the Smooth-Lasso as particular case have not been introduced to tackle such an objective. Indeed they are more devoted to respond to the estimation criterion or more precisely to structural purposes. However, it is important to consider the variable selection ability of this method since we deal with sparse regression.

A large amount of work has been done on the topic of variable selection for Lasso-type methods. One important observation is that one has to make a compromise between not identifying a low signal level (that is, small in absolute value of the coefficients  $\beta_j^*$ ,  $j \in \mathcal{A}^*$ ) and imposing a strong restriction on the Gram matrix  $\Psi^n$ , which sometimes seems to be not realistic. Moreover, the question of the identifiability of  $\beta^*$  has also to be considered. Our approach consists in the choice of the middle road, that is, involving the less restrictive assumption on the Gram matrix that permit us to recover reasonably a low signal level. For this purpose we first provide a bound on the sup-norm  $|\beta^* - \hat{\beta}^{Quad}|_\infty$ , based on a control on the  $\ell_2$  estimation error.

To this end, we use Assumption  $B(\Theta)$  on the Gram matrix. Nevertheless the set  $\Theta$  should be larger than the one required in Theorem 1. To define it, let us denote by  $\mathcal{C}$  the index-set of the  $m$  largest components in absolute value of  $\beta^* - \hat{\beta}^{Quad}$  outside  $\mathcal{B}$ . Here  $\mathcal{B}$  is the set introduced in Theorem 1. In this setting  $m$  is an integer such that  $m + |\mathcal{B}| < p$ .

**Assumption  $B'(\mathcal{B} \cup \mathcal{C})$ :** Let  $K_n$  be the matrix given by (8) and let  $\varrho_n = 4\sqrt{|\mathcal{A}^*|} + \frac{4\mu_n}{\lambda_n} |\tilde{\mathcal{J}}\beta^*|_2$ . There is a constant  $\phi > 0$  such that, for any  $\Delta \in \mathbb{R}^p$  that satisfies  $\sum_{j \notin \mathcal{B}} |\Delta_j| \leq \varrho_n \sqrt{\sum_{j \in \mathcal{B}} \Delta_j^2}$ , we have

$$\Delta' K_n \Delta \geq \phi \sum_{j \in \mathcal{B} \cup \mathcal{C}} \Delta_j^2. \quad (12)$$

The above assumption differs from Assumption  $B(\Theta)$  only on the fact that we restrict  $\mathbb{R}^p$  in a different set than that is used in the condition (12). Obviously, Assumption  $B'(\mathcal{B} \cup \mathcal{C})$  implies Assumption  $B(\mathcal{B})$ .

**Proposition 1.** Let us consider the same setting as in Theorem 1 with the only difference that  $\lambda_n = 2\sqrt{2}\sigma \sqrt{\frac{\log(p/\eta)}{n}}$  with  $0 < \eta < 1$ . Under Assumption  $B'(\mathcal{B} \cup \mathcal{C})$  and with probability  $1 - \eta$

$$|\hat{\beta}^{Quad} - \beta^*|_\infty \leq |\hat{\beta}^{Quad} - \beta^*|_2 \leq \tilde{c}(\lambda_n \sqrt{|\mathcal{A}^*|} + \mu_n |\tilde{\mathcal{J}}\beta^*|_2),$$

where  $\tilde{c} = 2\phi^{-1}(1 + \frac{\rho_n}{\sqrt{m}})$ .

One can exploit the control provided in Proposition 1 to construct an thresholded version of  $\hat{\beta}^{Quad}$  which is consistent in variable selection. Such a construction has already been considered in several papers for the Lasso estimate. The closest methodology to ours is the one developed in [23].

Let us now provide a consistent version of the estimator  $\hat{\beta}^{Quad}$  for the selection purpose. Consider  $\hat{\beta}^{Th-Quad} = (\hat{\beta}_1^{Th-Quad}, \dots, \hat{\beta}_p^{Th-Quad})'$ , the thresholded  $\hat{\beta}^{Quad}$  estimator defined by

$$\hat{\beta}_j^{Th-Quad} = \hat{\beta}_j^{Quad} \quad \text{if } |\hat{\beta}_j^{Quad}| \geq \tilde{c}(\lambda_n \sqrt{|\mathcal{A}^*|} + \mu_n |\tilde{\mathcal{J}}\beta^*|_2),$$

and zero otherwise, where  $\tilde{c}$  is given in Proposition 1. This estimator consists of  $\hat{\beta}^{Quad}$  with its small coefficients reduced to zero. We then enforce the selection property of  $\hat{\beta}^{Quad}$ . Variable selection consistency of this estimator is established under one more restriction on the regression vector given now.

**Assumption C:** *The true regression vector  $\beta^*$  is such that*

$$\min_{j \in \mathcal{A}^*} |\beta_j^*| > 2\tilde{c}(\lambda_n \sqrt{|\mathcal{A}^*|} + \mu_n |\tilde{\mathcal{J}}\beta^*|_2),$$

where  $\tilde{c} = 2\phi^{-1}(1 + \frac{\rho_n}{\sqrt{m}})$  is from Proposition 1, and  $\phi$  is the term appearing in Assumption B'( $\mathcal{B} \cup \mathcal{C}$ ).

Here again, we observe how important the quantity  $\phi$  is. We want it to be as large as possible.

This assumption bounds from below the smallest regression coefficient in  $\beta^*$ . This is a common assumption to provide sign consistency in the high dimensional case. Indeed, this condition appears in [4, 18, 23, 32, 36, 37]. We refer to [18] for a longer discussion on how these works are related in terms of restriction according the threshold or the assumption on the Gram matrix. Now we can state the following sign consistency result.

**Theorem 2.** *Let us consider the thresholded estimator  $\hat{\beta}^{Th-Quad}$  as described above. In the same setting as in Proposition 1 and under Assumption B'( $\mathcal{B} \cup \mathcal{C}$ ) and also Assumption C*

$$\mathbb{P} \left( \text{Sgn}(\hat{\beta}^{Th-Quad}) = \text{Sgn}(\beta^*) \right) \geq 1 - \eta.$$

Note that all the remarks established in Sections 3.1.1 and 3.1.2 remain valid also for this variable selection result.

### 3.3 Technical advances

We devote this paragraph to several technical considerations. First we consider the case of a general matrix  $\tilde{\mathcal{J}}$ . Then we establish the variable selection consistency of the thresholded version of  $\hat{\beta}^{Quad}$ . Finally we provide a relaxation of the assumption on the noise. The reader who is not interested in these studies can skip them without consequences for the readability of the paper.

#### 3.3.1 General matrices $\tilde{\mathcal{J}}$

Theorem 1 deals with the case where  $\tilde{\mathcal{J}} = \mathbf{J}'\mathbf{J}$  is sparse. In that statement, Assumption B( $\mathcal{B}$ ) was needed with  $\mathcal{B} \supset \mathcal{A}^*$  which also depends on  $\tilde{\mathcal{J}}$ . More precisely  $\mathcal{B}$  contains the indices of components which interfere in the sparse product  $\beta^{*'}\tilde{\mathcal{J}}u$  for a given  $u \in \mathbb{R}^p$  (see the proof for

more details). This set is not too large compared to  $\mathcal{A}^*$  when we consider the case where  $\tilde{J}$  is sparse. This way to solve the problem allows us to choose  $\mu_n \sim \lambda_n \frac{\sqrt{|\mathcal{A}^*|}}{|\tilde{J}\beta^*|_2}$  (cf. Corollary 1). In what follows, we consider general  $p \times p$  matrices  $\tilde{J}$  (including naturally the sparse case) for which we only need a (adapted) RE Assumption. Here  $\mu_n$  is not a free parameter anymore, and is smaller than the one given in Corollary 1.

Let us first establish the assumptions needed, and the setup of this contribution. Let  $\eta \in (0, 1)$ . We define the regularization parameters  $\lambda_n$  and  $\mu_n$  in the following way:

$$\lambda_n = 8\sqrt{2}\sigma \sqrt{\frac{\log(p/\eta)}{n}}, \quad \text{and} \quad \mu_n = \lambda_n \frac{1}{8|\tilde{J}\beta^*|_\infty}. \quad (13)$$

We now state the adapted RE Assumption which differs from the usual one introduced in [3] only by the matrix to which we apply the assumption ( $K_n$  instead of  $\Psi^n$ ):

**Assumption RE:** *There is a constant  $\phi > 0$  such that, for any  $\Delta \in \mathbb{R}^p$  that satisfies  $\sum_{j \notin \mathcal{A}^*} |\Delta_j| \leq 4 \sum_{j \in \mathcal{A}^*} |\Delta_j|$ , we have*

$$\Delta' K_n \Delta \geq \phi \sum_{j \in \mathcal{A}^*} \Delta_j^2.$$

This assumption involves a set of linear inequalities. Then we clearly have  $\phi \geq \phi_0$  (the  $\phi$  corresponding to the Lasso; that is, when  $\mu_n = 0$ ). With this setting, we obtain the following result for a general matrix  $\tilde{J}$ .

**Theorem 3** (general  $\tilde{J}$ ). *Let  $\mathcal{A}^*$  be the sparsity set and let the tuning parameters  $(\lambda_n, \mu_n)$  be defined as in (13). If Assumption RE holds, then with probability greater than  $1 - \eta$ , we have*

$$\begin{aligned} \left\| X\beta^* - X\hat{\beta}^{Quad} \right\|_n^2 &\leq 4\phi^{-1} \lambda_n^2 |\mathcal{A}^*|, \\ (\beta^* - \hat{\beta}^{Quad})' \tilde{J} (\beta^* - \hat{\beta}^{Quad}) &\leq 4 \frac{|\tilde{J}\beta^*|_\infty}{\phi} \lambda_n |\mathcal{A}^*|, \end{aligned}$$

and

$$|\beta^* - \hat{\beta}^{Quad}|_1 \leq 8\phi^{-1} \lambda_n |\mathcal{A}^*|.$$

Similar bounds were provided for the Lasso estimator by [3]. Let us mention that the constants 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  $\ell_2$  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 ([6, Theorem 5.1]). Moreover, Theorem 3 states a control on an error which is linked to the expected prior information which suggested the use of the estimator  $\hat{\beta}^{Quad}$ .

The results provided in Theorem 1 and more precisely Corollary 1 differ from those established in Theorem 3 in a few points. First, the value of  $\mu_n$  is larger in the sparse case. Indeed,  $\mu_n$  equals  $\lambda_n \frac{\sqrt{|\mathcal{A}^*|}}{2|\tilde{J}\beta^*|_2}$  and  $\lambda_n \frac{1}{4|\tilde{J}\beta^*|_\infty}$  respectively in Corollary 1 and Theorem 3. The former value can be much larger for some regression vector  $\beta^*$ . Second, these values of  $\mu_n$  have an issue on the error bounds through  $\phi$ . As a consequence, the bounds in Corollary 1 are better than those in Theorem 3. Finally apart from the considerations on the quantity  $\phi$ , we observe a changing in the bound of  $(\beta^* - \hat{\beta}^{Quad})' \tilde{J} (\beta^* - \hat{\beta}^{Quad})$ . Indeed, the bound in Theorem 1 involves the term  $|\tilde{J}\beta^*|_2 \sqrt{|\mathcal{A}^*|}$  whereas in Theorem 3 appears  $|\tilde{J}\beta^*|_\infty |\mathcal{A}^*|$  which is obviously larger. We then have a better control on this error using the sparsity of the matrix  $\tilde{J}$ . Finally, we remark that the constant  $\kappa$  in Corollary 1 is smaller than the corresponding  $\kappa$  in Theorem 3. Nevertheless, one should mention that Assumption RE is less restrictive than Assumption  $B(\mathcal{B})$ .

### 3.3.2 Non-thresholded variable selector

In Section 3.2, we established variable selection consistency for a thresholded version of  $\hat{\beta}^{Quad}$  when  $\tilde{J}$  is sparse. One can also provide a variable selection consistency result directly for  $\hat{\beta}^{Quad}$  thanks to a different calibration of the tuning parameters. This result can be applied to general matrices  $\tilde{J}$ . The approach to prove the result is also different. We first provide a bound on the sup-norm  $|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_\infty$ . This can be easily done using theorems stated in Section 3.3.1. That is we use the  $\ell_1$  estimation error  $|\beta^* - \hat{\beta}^{Quad}|_1$ . Nevertheless, this implies that only ‘high’ levels of signal can be reconstituted, i.e., coefficients  $\beta_j^*$ ,  $j \in \mathcal{A}^*$  such that  $|\beta_j^*| \geq cst \cdot \lambda_n |\mathcal{A}^*|$ . Hence we favor to exploit here again a control on the  $\ell_2$  estimation error  $|\beta^* - \hat{\beta}^{Quad}|_2$  which by the sequel enables us to recover signals such  $|\beta_j^*| \geq cst \cdot \lambda_n \sqrt{|\mathcal{A}^*|}$  with the same assumption on the matrix  $K_n$ . Let us mention that  $\lambda_n \sqrt{|\mathcal{A}^*|}$  is not the best level which can be recovered. One can also get rid of the term  $\sqrt{|\mathcal{A}^*|}$  through a quite restrictive assumption on the correlations between variables such as the Mutual Coherence assumption:  $\max_{j \in \mathcal{A}^*} \max_{\substack{k \in \{1, \dots, p\} \\ k \neq j}} |(K_n)_{j,k}| \leq \frac{t}{|\mathcal{A}^*|}$ , where  $t$  is a small constant.

**Proposition 2.** *Let us consider the same setting as in Theorem 3 with the only difference that  $\lambda_n = 4\sqrt{2}\sigma\sqrt{\log(p/\eta)/n}$  and  $\mu_n = \lambda_n/(4|\tilde{J}\beta^*|_\infty)$  with  $0 < \eta < 1$ . Under Assumption RE, and with probability larger than  $1 - \eta$ , we have*

$$|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_\infty \leq |\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_2 \leq 2\phi^{-1}\lambda_n\sqrt{|\mathcal{A}^*|},$$

where  $\phi$  is the constant appearing in Assumption RE. Moreover, if  $\min_{j \in \mathcal{A}^*} |\beta_j^*| > 2\phi^{-1}\lambda_n\sqrt{|\mathcal{A}^*|}$ , we have

$$\mathbb{P}\left(\text{Sgn}(\hat{\beta}_{\mathcal{A}^*}^{Quad}) = \text{Sgn}(\beta_{\mathcal{A}^*}^*)\right) \geq 1 - \eta.$$

Proposition 2 is a trivial consequence of Theorem 3. A small proof is given in the Appendix section. This proposition directly underlines that under the Restrictive Eigenvalue Assumption, all non-zero components of  $\beta^*$  are detected by  $\hat{\beta}^{Quad}$ . Actually, in the setting of Proposition 2,  $\hat{\beta}^{Quad}$  contains too many non-zero components. More restrictions are needed in order to ensure the variable selection consistency of  $\hat{\beta}^{Quad}$ . Here is an additional assumption on the Gram matrix which controls the correlations between the truly relevant variables and those which are not.

**Assumption D:** *We assume that*

$$\max_{j \in \mathcal{A}^*} \max_{k \notin \mathcal{A}^*} |(K_n)_{j,k}| \leq \frac{t}{|\mathcal{A}^*|},$$

where  $t$  is a positive term smaller than  $\frac{\phi}{64}$ .

This assumption is quite close to the Mutual Coherence assumption which involves the Gram matrix  $\Psi^n$  instead of  $K_n$ . In addition, the Mutual Coherence assumption makes a restriction on correlations between all covariates.

**Theorem 4.** *Let consider the linear regression model (1). Let  $\lambda_n = 16\sigma\sqrt{\frac{\log(p/\sqrt{\eta p/(1+p)})}{n}}$  and  $\mu_n = \lambda_n/(4|\tilde{J}\beta^*|_\infty)$ . Under Assumptions RE-C and also Assumption D, we have*

$$\mathbb{P}(\hat{\mathcal{A}} \not\subseteq \mathcal{A}^*) \leq \eta,$$

and then

$$\mathbb{P}\left(\text{Sgn}(\hat{\beta}^{Quad}) = \text{Sgn}(\beta^*)\right) \geq 1 - \eta.$$

To prove the first claim, we use some arguments from [5]. The second point is a consequence of the first and of Proposition 2. There are essentially two differences between the setting of Theorem 4 and Proposition 2. First, we need for this last result a more restrictive assumption on the correlations between variables. However, this restriction is only between relevant variables and irrelevant covariates. This is a ‘quite’ reasonable assumption to identify the relevant variables, that is, the non-zero components of the vector  $\beta^*$ . Second, the minimal value of  $\lambda_n$  is larger in this last theorem. This suggests that we need a larger value of this tuning parameter to set to zero the irrelevant components. Note that we established the variable selection consistency of  $\hat{\beta}^{Quad}$  but with a value of the tuning parameter  $\mu_n$  smaller than with the thresholded version.

**Remark 5.** *The results of Theorem 4 can also be obtained under the more restrictive Mutual Coherence assumption:  $\max_{j \in \mathcal{A}^*} \max_{\substack{k \in \{1, \dots, p\} \\ k \neq j}} |(K_n)_{j,k}| \leq \frac{\tilde{t}}{|\mathcal{A}^*|}$ . Here, even the correlations between relevant variables are restricted but this restriction makes possible to recover even smaller signal. That is, we can detect coefficients of  $\beta^*$  such that  $|\beta_j^*| \geq cst \cdot \sqrt{\log(p)/n}$ .*

### 3.3.3 Non Gaussian noise with finite variance

Most of the results established for Lasso-type methods assume Gaussian or sub-Gaussian type noise [3, 5, 15, 32, 36]. Noise with exponential moment is studied in [4, 23]. Only a few references consider other type of noise. Noise with moment of order  $2k$ , where  $k \geq 1$  is an integer, is considered in [37], whereas in the paper [18], the author presents the case where the noise admits zero mean and finite variance. It is in the same spirit as that in this last reference that we establish this relaxation on the noise. According to the Elastic-Net, noise with moment of order  $2k + \delta$ , where  $k \geq 1$  is an integer and  $\delta > 0$  is a real, is considered in [40], but the authors considered only the case where  $p = \mathcal{O}(n)$ .

We assume that the noise random variables  $\varepsilon_1, \dots, \varepsilon_n$  are independent and admit zero mean and finite variance. That is  $\mathbb{E}\varepsilon_i = 0$  and  $\mathbb{E}\varepsilon_i^2 \leq \sigma^2$  for  $i = 1, \dots, n$  with  $\sigma^2 < \infty$ . In this generalization we also use a revisited version of Nemirovski’s Inequality established in [11]. One more restriction is needed on the sample points.

**Assumption E:** *There exists a positive constant  $L < \infty$  such that*

$$n^{-1} \sum_{i=1}^n \max_{j=1, \dots, p} x_{i,j}^2 \leq L.$$

Theorem 5 below extends the results in Corollary 1 of Section 3 to the non-Gaussian noise case. However, one is able to generalize all the results of that section in the same way.

**Theorem 5.** *Let consider the linear regression model (1) where the  $\varepsilon_i$ ’s are independent random variables with zero mean such that  $\mathbb{E}\varepsilon_i^2 \leq \sigma^2$  for  $i = 1, \dots, n$  with  $\sigma^2 < \infty$ . Denote by  $K_{Nem}$  the quantity  $K_{Nem} = \inf_{q \in [2, \infty] \cap \mathbb{R}} (q - 1)p^{2/q}$ , and let  $\lambda_n = 4\sigma \sqrt{\frac{K_{Nem}L}{n\eta}}$  with  $0 < \eta < 1$ . Let  $\mu_n = \frac{\lambda_n \sqrt{|\mathcal{A}^*|}}{2^{|\mathcal{J}\beta^*|_2}}$ . Assume also that Assumption B( $\mathcal{B}$ ) (where  $\varrho_n = 6\sqrt{|\mathcal{A}^*|}$ ) and Assumption E hold. Then with probability greater than  $1 - \eta$ , we have*

$$|\beta^* - \hat{\beta}^{Quad}|_1 \leq \frac{72\sigma}{\phi} \sqrt{\frac{K_{Nem}L}{n\eta}} |\mathcal{A}^*|.$$

Let us mention that  $2e \log(p) - 3e < K_{Nem} < 2e \log(p) - e$ . The rate of convergence in Theorems 5 is, up to constants, the same as in Corollary 1. However, the constants seem to be worse in the non-Gaussian case since it brings into play the constant  $L$  which can be large. This is the price to pay in the non-Gaussian noise.

**Remark 6.** *In the above theorem,  $\eta$  is fixed. However, one can set  $\eta$  depending on  $p$  (or on  $n$ ) in such way that it decreases to zero as  $p \rightarrow \infty$  (or  $n \rightarrow \infty$ ). It is interesting to note that in this case, we loose a small power  $\log(p)$  (or  $\log(n)$ ) in the rate of convergence when we consider non-Gaussian noise compared to the Gaussian case.*

Using similar reasoning as in Theorem 5 (cf. proof of Theorem 5), there is no major difficulty to extend the variable selection results established in Section 3.2 with Gaussian noise to the case where the noise is defined as above. This can be done using Lemma 3 instead of Lemma 2 in all the proofs.

## 4 Experimental Results

In this section we present the experimental performance of the estimator  $\hat{\beta}^{Quad}$ . In particular, we focus on two special cases: the Elastic-Net and the S-Lasso defined respectively with the penalties  $\text{pen}^{EN}(\beta) = \lambda|\beta|_1 + \mu|\beta|_2^2$  and  $\text{pen}^{SL}(\beta) = \lambda|\beta|_1 + \mu \sum_{j=2}^p (\beta_j - \beta_{j-1})^2$ . The Elastic-Net is useful when high correlations between variables appears, whereas the S-Lasso is devoted to problems where the regression vector  $\beta^*$  is ‘smooth’ (small variations in the values of the successive components of  $\beta^*$ ). We essentially are interested in the performance of these estimators w.r.t. their estimation accuracy, i.e., in terms of the estimation error  $|\hat{\beta} - \beta^*|_2$ , when  $\beta^*$  is known (simulated data). Indeed, the introduction of  $\hat{\beta}^{Quad}$  is motivated by a priori knowledge on the structure of the parameter  $\beta^*$ , or on the correlation between variables, and the purpose here is to see how this information can be taken into account to improve the reconstruction of the vector  $\beta^*$ . As benchmarks, we use the Lasso and the Fused-Lasso estimators, since the first is the reference method and the second is close in spirit to the S-Lasso estimator. Indeed, it is designed to produce solutions with equal values of the successive components of  $\beta^*$  (‘blocky’) [28]. Note also that in the pioneer paper of the Elastic-Net, a ‘corrected’ version of this estimator is proposed [39]. There is as yet no theoretical support for this method. Moreover, it outperforms the ‘non-corrected’ Elastic-Net (this ‘non-corrected’ Elastic-Net is denoted by naive in [39]) in only a very few of the situations we consider in this paper. We omitted the results for these ‘corrected’ versions to avoid digressions.

Except for the Fused-Lasso solution, all of the Lasso, the S-Lasso and the Elastic-Net solutions can be computed thanks to the LARS algorithm (cf. Lemma 1). However, we will not use this LARS algorithm in this study. Indeed, in order to be fair with all the methods, we used the same algorithm for all of them. We exploit an algorithm provided by J. Mairal<sup>2</sup> which is an implementation of a general convex program given by [24].

In all our experiments, the tuning parameters are chosen based on the 10 fold cross validation criterion (for the Fused-Lasso, the Elastic-Net and the S-Lasso, the cross validation is performed in a  $2d$  Grid), but we also display the results obtained based on the theoretical values. Note that for the Fused-Lasso, we considered the same theoretical values of the tuning parameters as for the S-Lasso as they are motivated by similar applications (this choice seems arbitrary but to our knowledge, no precise study has been made for the Fused-Lasso in the context we are considering). On the other hand, both Elastic-Net and the S-Lasso involve a sparse matrix  $\tilde{J}$  in the definition of the estimator  $\hat{\beta}^{Quad}$ . Then the theoretical values of

---

<sup>2</sup><http://www.di.ens.fr/~mairal/index.php>

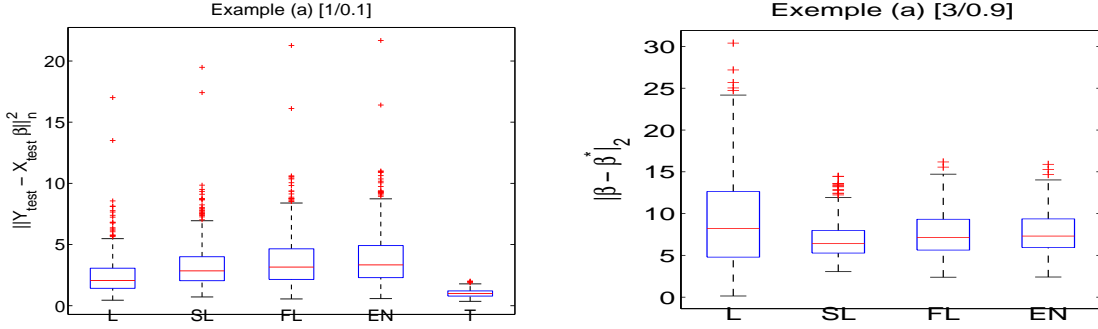


Figure 1: Performance of the Lasso (L), the S-Lasso (SL), the Fused-Lasso (FL) and the Elastic-Net (EN) applied to *Example (a)* and based on 500 replications. The tuning parameters are chosen based on the theoretical study. *Left*: Evaluation of the prediction error  $\|Y_{test} - X_{test}\hat{\beta}\|_n^2$ , in comparison with the performance of the truth (T), i.e.,  $\|Y_{test} - X_{test}\beta^*\|_n^2$ . *Right*: Evaluation of the  $\ell_2$  estimation error  $|\hat{\beta} - \beta^*|_2$ .

the tuning parameters are  $\lambda = 2\sqrt{2}\sigma\sqrt{\log(p)/n}$  and  $\mu = \lambda\sqrt{\mathcal{A}^*}/2|\tilde{\mathcal{J}}\beta^*|_2$ , in accordance with Corollary 1 and Proposition 1. These quantities depend on unknown parameters. They can be used only in the simulation study, and otherwise one needs to estimate  $|\tilde{\mathcal{J}}\beta^*|_2$ .

The different methods are applied to several simulation examples. They also have been applied to a *pseudo-real dataset* generated from the riboflavin dataset.

#### 4.1 Synthetic data

There are several parameters: the dimension  $p$ , the sample size  $n$  and the noise level  $\sigma$ . They will be specified in the setting of the experiments (that is in the different tabulars and figures captions). The first one is classical and has been introduced in the original paper of the Lasso [27]. The second one comes from the paper by [39]. Here we are interested to observe the performance of the procedures when groups of variables appear. The last two studies aim to determine the behavior of the methods when the regression vector is ‘smooth’.

*Example (a)  $[\sigma/\rho]$ : No particularities.* We fix  $p = 8$  and  $n = 20$ . Here only  $\beta_1$ ,  $\beta_2$  and  $\beta_5$  are nonzero and equal respectively 3, 1.5 and 2. Moreover, for  $j, k \in \{1, \dots, 8\}$ , the design correlation matrix  $\Psi$  is defined by  $\Psi_{j,k} = \rho^{-|j-k|}$  where  $\rho \in ]0, 1[$ .

*Example (b)  $[p/n/\sigma]$ : Groups.* We have  $\beta_j = 3$  for  $j \in \{1, \dots, 15\}$  and zero otherwise. We construct three groups of correlated variables:  $\Psi_{j,j} = 1$  for every  $j \in \{1, \dots, p\}$ ; for  $j \neq k$ ,  $\Psi_{j,k} \approx 1$  (actually  $\Psi_{j,k} = \frac{1}{1+0.01}$ , due to an extra noise variable) when  $(j, k)$  belongs to  $\{1, \dots, 5\}^2$ ,  $\{6, \dots, 10\}^2$  and  $\{11, \dots, 15\}^2$  and zero otherwise.

*Example (c)  $[p/n/\sigma]$ : Smooth regression vector.* The regression vector is given by  $\beta_j = (3 - 0.2j)^2$  for  $j = 1, \dots, 15$  and zero otherwise. Moreover, the correlations are described by  $\Psi_{j,k} = \exp(-|j - k|)$  for  $(j, k) \in \{1, \dots, p\}^2$ .

*Example (d)  $[p/n/\sigma]$ : High sparsity index and smooth regression vector.* The regression vector is such that  $\beta_j = (4 + 0.1j)^2$  for  $j \in \{1, \dots, 40\}$  and zero otherwise, and the correlations are the same as in *Example (c)*.

Except when  $p = 500$  where we run only 100 replications, we based all the experiments on 500 replications.

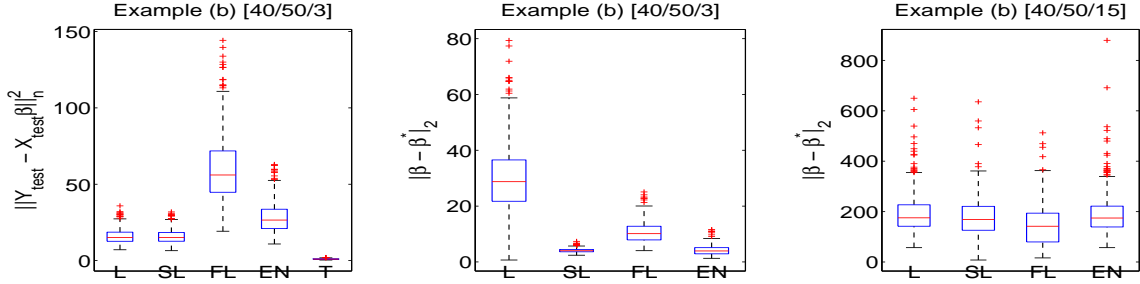


Figure 2: Performance of the Lasso (L), the S-Lasso (SL), the Fused-Lasso (FL) and the Elastic-Net (EN) applied to *Example (b)* and based on 500 replications. The tuning parameters are chosen based on the theoretical study in the first two plots, and by 10 fold cross validation in the third. *Left*: Evaluation of the prediction error  $\|Y_{test} - X_{test}\hat{\beta}\|_n^2$ , in comparison with the performance of the truth (T), i.e.,  $\|Y_{test} - X_{test}\beta^*\|_n^2$ . *Center-Right*: Evaluation of the  $\ell_2$  estimation error  $\|\hat{\beta} - \beta^*\|_2$ .

**Results.** The performance of the estimator  $\hat{\beta}$  (which can be the Lasso, the S-Lasso, the Elastic-Net or the Fused-Lasso) in terms of the prediction error  $\|Y_{test} - X_{test}\hat{\beta}\|_n^2$  (on a test set  $(Y_{test}, X_{test})$  of size  $n$ ) and the  $\ell_2$  estimation error  $\|\hat{\beta} - \beta^*\|_2$  are illustrated by boxplots in Figure 1 to Figure 4. For some of these experiments, the corresponding computational cost (in seconds) of each method is reported in Table 1. In what follows, we first compare the methods to each other in terms of their accuracy. Then we compare them in terms of their computational costs. Finally we provide some numerical justifications to the theoretical calibration of the tuning parameters of the S-Lasso procedure.

Methods comparison in terms of performance: Let us consider the different examples separately.

– *Example (a)*: when we consider the procedures induced by the cross validation criterion (for the choice of the tuning parameter), we notice that none of them outperforms the others even when  $\rho = 0.9$  (quite large correlation between successive variables). This is observed for both prediction and estimation errors. It is essentially due to the good behavior of the Lasso in such a situation where the regression vector is sparse but without any particular structure. Actually, this conclusion holds in almost all the cases even when the tuning parameters are chosen based on the theoretical study. However, two observations can be made. First, when both of  $\rho$  and  $\sigma$  are small, the Lasso estimator performs slightly better than the other methods. Moreover when  $\rho$  is large, a small improvement can be observed using the Fused-Lasso, the Elastic-Net and the S-Lasso methods when we care about the estimation error. This is illustrated in Figure 1 (left and right respectively) where we display the performance of the methods in terms of the prediction error in *Example (a)* [1/0.1] (left) and in terms of the estimation error in *Example (a)* [3/0.9] (right). For this example, the Lasso seems to be the best method since it involves only one tuning parameter. It moreover has a lower (mean) computational cost equal to 0.18 seconds for the Lasso (based on the cross validation criterion) as displayed in Table 1. The S-Lasso, the Elastic-Net and the Fused-Lasso computational costs are respectively 3.7, 3.6 and 4.2 seconds.

– *Example (b)*: with *Example (a)*, this example is the least favorable for the S-Lasso. Indeed, here the fifteen first coefficients equal 3. Then the value of the coefficients drops down directly to 0. There is a breaking point in the ‘smoothness’ in the true regression vector. Figure 5 displays the best reconstitution of the regression vector  $\beta^*$  using the S-Lasso solution (which minimizes the  $\ell_2$  estimation error since  $\beta^*$  is known). We observe the edge effects (breaking point in the ‘smoothness’) that the S-Lasso can meet due to the  $\ell_2$  fusion penalty term. However, even in this case, it seems that all the procedures perform in a similar way when the

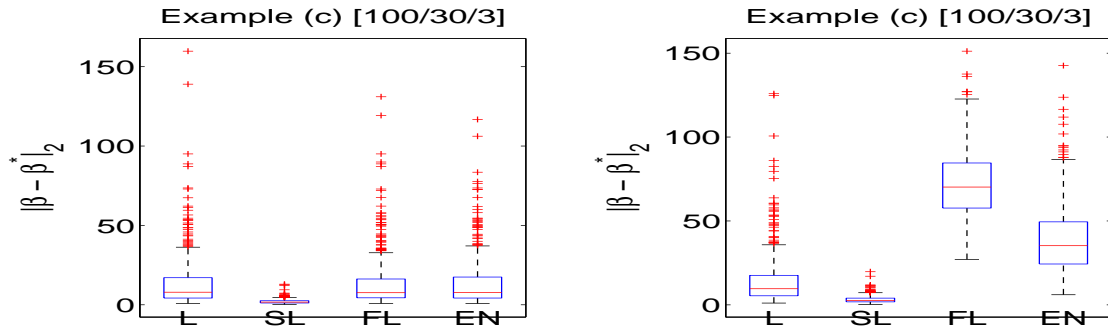


Figure 3: Evaluation of the  $\ell_2$  estimation error  $|\hat{\beta} - \beta^*|_2$  of the Lasso (L), the S-Lasso (SL), the Fused-Lasso (FL) and the Elastic-Net (EN) applied to *Example (c)* and based on 500 replications. *Left*: The tuning parameters are chosen by 10 fold cross validation. *Right*: The tuning parameters are chosen based on the theoretical study.

tuning parameters are chosen by cross validation. When the noise level is large ( $\sigma = 15$ ), let us nevertheless mention a (very) small improvement using the corrected versions of the S-Lasso and the Elastic-Net. Figure 2 (right) illustrates the performance of the methods in terms of the estimation error when they are applied to *Example (b)* [40/50/15]. The Fused-Lasso outperforms a little the other methods in this example (with this noise level) when we deal with the estimation performance.

On the other hand, when the methods are based on the theoretical calibration of the tuning parameters, two observations can be made regardless the noise level ( $1 \leq \sigma \leq 15$ ): the S-Lasso and the Elastic-Net provide good results whereas the Lasso has poor performance in terms of estimation error. This is illustrated in Figure 2 (left and center respectively) when the methods are applied to *Example (b)* [40/50/3]. Note moreover that a similar illustration is also obtained when  $p = 100$  and  $n = 40$ . Then the behavior of the different methods seems to be stable with the parameters  $p$ ,  $n$  and  $\sigma$ . This example is quite interesting since it points out that a good method for the prediction objective can be less efficient for the estimation objective (see the performance of the Lasso and the Elastic-Net).

– *Example (c)*: we consider several values of the sample size  $n$  and the dimension  $p$ . It turns out that here again, when  $p < n$  all the methods behave in the same way when the tuning parameters are chosen by cross validation (the S-Lasso induces just a small improvement). However when  $p > n$  the S-Lasso is by far better than the other methods. This is illustrated by Figure 3 (left) where  $\ell_2$  estimation error of each method applied to *Example (c)* [100/30/3] is displayed. The same plot is obtained for the prediction error.

Moreover when the tuning parameters are calibrated according to the theoretical study, the S-Lasso performs the best and the Fused-Lasso the worst. This appears to be true whatever the values of the parameters  $p$ ,  $n$  and  $\sigma$ . See for instance Figure 3 (right) where the different methods are applied to *Example (c)* [100/30/3] and for the estimation task (the same is obtained for the prediction objective).

Note that in this example, the Fused-Lasso and the Elastic-Net appear to be useless.

– *Example (d)*: this is with *Example (c)* the most favorable situation for the S-Lasso estimator where the regression vector is ‘smooth’ with a large amount of non-zero components. The S-Lasso estimator seems to dominate its opponents in all the cases, and regardless of the sample size  $n$ , the dimension  $p$  or the noise level  $\sigma$ . This observation holds for the  $\ell_2$  estimation and the prediction errors. Note that when the tuning parameters are chosen by cross validation, the Lasso, the Fused-Lasso and the Elastic-Net have quite close performance. Figure 4 illus-

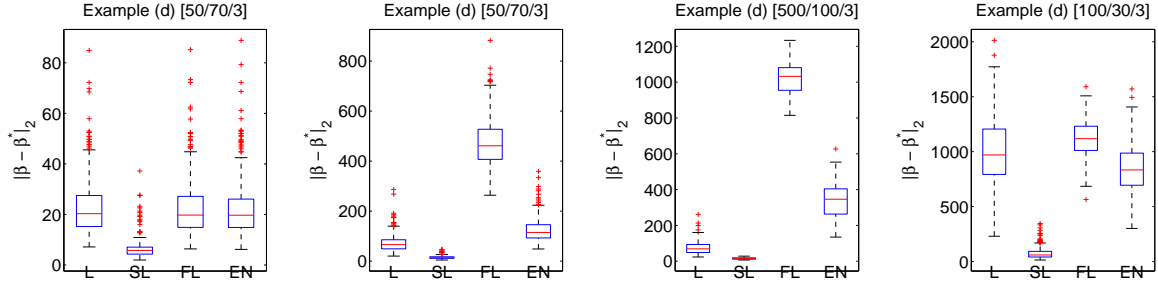


Figure 4: Evaluation of the  $\ell_2$  estimation error  $|\hat{\beta} - \beta^*|_2$  of the Lasso (L), the S-Lasso (SL), the Fused-Lasso (FL) and the Elastic-Net (EN) applied to *Example (d)* and based on 500 replications. *Left*: The tuning parameters are chosen by 10 fold cross validation. *Center-left*; *Center-right*; *Right*: The tuning parameters are chosen based on the theoretical study.

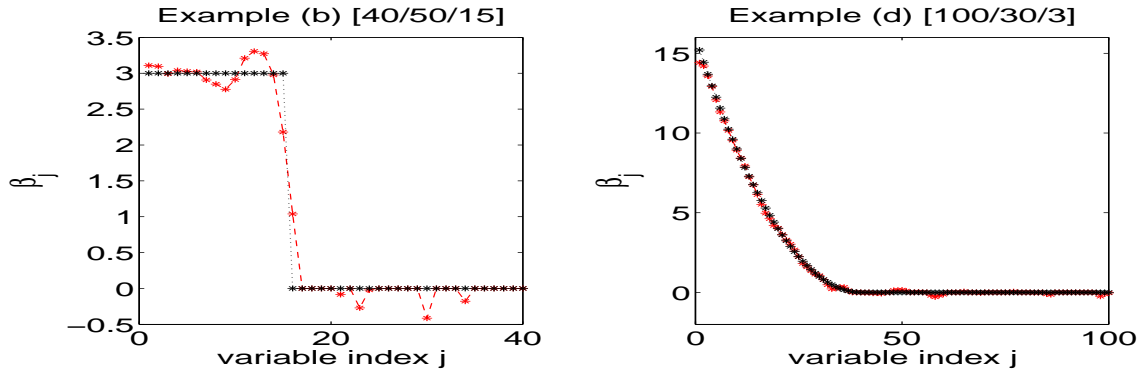


Figure 5: Best reconstitution of the regression vector  $\beta^*$  (black curve) by the SL-Lasso estimator (red curve). *Left*: Application to *Example (b)* [40/50/15]. *Right*: Application to *Example (d)* [100/30/3].

trates this fact when  $p < n$  for the estimation error (left: cross validation; center-left: theory). Moreover, Figure 4 (center-right and right) displays the performance of the methods when  $p > n$  when the tuning parameters are based on the theoretical study (note that ranking of the methods does not change from the case  $p < n$  when the tuning parameters are chosen by cross validation). Here an interesting observation follows from the experiments on *Example (d)* [100/30/3] (Figure 4-left). Indeed, here the sparsity index  $|\mathcal{A}^*| = 40$  and it is then larger than the sample size  $n = 30$ . In this case, the Lasso has poor performance. However, the S-Lasso is still good. Moreover, there even exists a pair  $(\lambda, \mu)$  (the pair minimizing the  $\ell_2$  estimation error since  $\beta^*$  is known) such that we have a good reconstitution on the regression vector  $\beta^*$  (see Figure 5-right).

*Methods comparison in terms of computational cost*: Table 1 displays the computational cost (in seconds) of each method on several examples. First note that the Fused-Lasso has the largest computational cost in all the simulations whereas the Lasso has the smallest. The Elastic-Net and the S-Lasso have intermediate computational costs but stay reasonable compared to the Fused-Lasso. More precisely, when the tuning parameters are chosen by cross validation, we remark that the computational costs of the S-Lasso and the Elastic-Net are about 30 times larger than the Lasso. This is partly explained by the number of values explored for the tuning parameter  $\mu$  (a grid with 20 elements). Actually, even for fixed  $\lambda$  and  $\mu$ , the computation cost of the Lasso is (a little) smaller than the computation costs of the S-Lasso and the Elastic-Net. This is observed for instance when we consider the solutions computed when the tuning parameters are chosen based on the theoretical study. The rea-

Table 1: Computational costs in seconds of the Lasso (L), the S-Lasso (SL), the Fused-Lasso (FL) and the Elastic-Net (EN) on several examples illustrated in the above figures. The parameter  $Tuning = Th$  or  $Tuning = Cv$  depending on whether we consider the methods with the tuning parameters based on the theoretical issue or on the 10 fold cross validation respectively.

METH.	TUNING	Ex.(A) [1/0.1]	Ex.(A) [3/0.9]	Ex.(B) [40/50/15]	Ex.(C) [30/50/3]	Ex.(D) [500/100/3]
L	$Th \cdot 10^{-4}$	$1.1 \pm 0.1$	$8 \pm 41$	$5 \pm 2$	$33 \pm 64$	$457 \pm 243$
	$Cv$	$0.18 \pm 0.01$	$0.5 \pm 0.2$	$0.5 \pm 0.1$	$1.1 \pm 0.3$	$12.3 \pm 4.9$
SL	$Th \cdot 10^{-4}$	$5.1 \pm 6.4$	$8 \pm 28$	$6 \pm 6$	$48 \pm 81$	$967 \pm 441$
	$Cv$	$3.7 \pm 0.1$	$11.1 \pm 1.3$	$10.2 \pm 2.0$	$36.2 \pm 9.1$	$648.3 \pm 219.2$
FL	$Th \cdot 10^{-4}$	$2.6 \pm 0.3$	$10.0 \pm 30.0$	$20 \pm 12$	$518 \pm 271$	$5996 \pm 2019$
	$Cv$	$4.2 \pm 0.2$	$14.1 \pm 1.6$	$38.3 \pm 5.8$	$245.6 \pm 64.3$	$\simeq 3 \cdot 10^5$
EN	$Th \cdot 10^{-4}$	$4.7 \pm 3.5$	$9 \pm 43$	$5 \pm 3$	$41 \pm 60$	$1022 \pm 432$
	$Cv$	$3.6 \pm 0.2$	$11.0 \pm 1.3$	$10.2 \pm 2.0$	$35.2 \pm 8.9$	$637.3 \pm 214.0$

son is that the S-Lasso and the Elastic-Net are solved thanks to a Lasso program applied to augmented data (cf. Lemma 1). Except on *Example* (a) where the increase of computational cost using the S-Lasso and the Elastic-Net is not justified (since the improvement using the Lasso-type methods is quite small), in most of the considered situations it is quite interesting to use the Elastic-Net and even more interesting to use the S-Lasso estimator. This is due to the ‘smoothness’ of the true regression vector.

On the other hand, the Fused-Lasso has a large computation cost due to the  $\ell_1$ -fusion penalty which is not strictly convex. Moreover, it does not improve enough the Lasso estimator in the situation we considered in this paper (as observed in the previous part).

In view of the computational costs related to *Example* (a) (the first two columns in Table 1), let us finally remark that these costs increase with  $\rho$ , the correlation level between variables, and  $\sigma$ , the noise level. We observe for instance that the mean computational cost of the Lasso estimator (when the tuning parameter is chosen by cross validation) is 1.1 seconds when  $\rho = 0.1$  and  $\sigma = 1$  and increases to 8 seconds when  $\rho = 0.9$  and  $\sigma = 3$ .

*S-Lasso; theory vs. cross validation:* Figure 6 resumes the comparison between the S-Lasso based on a theoretical choice of the tuning parameters (denoted by this part S-Lasso<sup>Th</sup>) and the S-Lasso where the tuning parameters are based on 10 fold cross validation (denoted here by S-Lasso<sup>Cv</sup>). First we can observe that the performance of both S-Lasso<sup>Th</sup> and S-Lasso<sup>Cv</sup> are close. Moreover given the results in the part ‘*Methods comparison in terms of performance*’, they both perform in a good way. However, it seems that S-Lasso<sup>Cv</sup> outperforms S-Lasso<sup>Th</sup> when we deal with the prediction task. This seems quite intuitive since by definition, the cross validation criterion attempts to provide good estimator for the prediction objective. According to the  $\ell_2$  estimation goal, we cannot conclude the superiority of one of the estimator on the other. Nevertheless, in the high dimensional setting *Example* (d) [500/100/ $\sigma$ ], it seems that S-Lasso<sup>Cv</sup> begins to become better.

Hence it turns out that the theoretical choice for  $\mu$  ( $\mu = \frac{\lambda\sqrt{A^*}}{2|\mathcal{J}\beta^*|_2}$ ) provides good performance both in terms of  $\ell_2$  estimation error and test error. Moreover, they are often close to the performance of the S-Lasso estimator based on the cross validation criterion. This is quite interesting since the computational cost of S-Lasso<sup>Th</sup> is much smaller than S-Lasso<sup>Cv</sup>. This study is actually more a verification of our theoretical choices of the tuning parameters than a rule to apply in practice. Indeed, since the theoretical choice of  $\mu$  depends on  $\beta^*$ , the corresponding estimator S-Lasso<sup>Th</sup> is unusable in real data problems.

Table 2: Median values of the tuning parameters  $(\lambda, \mu)$  of the S-Lasso for different ways of calibration: ‘*Cv*’ for cross validation; ‘*Th*’ for theoretical issue; ‘*Est*’ for  $\ell_2$  estimation error minimizers. The tuning parameters displayed here correspond to the experiments illustrated in Figure 6

TUNING	<i>Ex.</i> (A) [3/0.9]	<i>Ex.</i> (B) [100/40/3]	<i>Ex.</i> (C) [100/30/3]	<i>Ex.</i> (D) [500/50/3]
$\lambda^{Cv}$	0.4	0.5	0.3	1.0
$\mu^{Cv}$	0.0005	0.0003	0.2	0.1
$\lambda^{Th}$	2.7	2.8	1.1	2.1
$\mu^{Th}$	0.5110	1.3100	0.4	1.2
$\lambda^{Est}$	0.7	1.0	0.3	1.0
$\mu^{Est}$	0.2500	1.2500	0.3	2.0

*Tuning parameters:* Table 2 displays the values of the tuning parameters  $(\lambda, \mu)$  of the S-Lasso, when there are chosen by cross validation  $(\lambda^{Cv}, \mu^{Cv})$  and based on the theoretical issue  $(\lambda^{Th}, \mu^{Th})$ . We compare them to the values of these parameters  $(\lambda^{Est}, \mu^{Est})$  that minimize the  $\ell_2$  estimation error.

A first remark is that the values of the tuning parameters calibrated based on the theoretical study are always larger than those chosen by cross validation. This is not surprising since the theoretical calibration of the tuning parameters uses to provide such a behavior. Then it turns out that the theoretical considerations leads to ‘smoother’ solutions than the cross validation. Note however that this does not imply that the solution based on the theoretical issue is sparser since a larger  $\mu$  usually implies that the solution is less sparse.

In comparison to the best solution (where the tuning parameters minimize the  $\ell_2$  estimation error), there are two case. When the true regression vector is not smooth, it seems that these ‘best’ tuning parameters are closer those chosen by cross validation. When the true regression vector is smooth, they are closer to the tuning parameters calibrated based on the theoretical perspectives. To sum up, on can say that the best  $\lambda$  is close to one chosen by cross validation, whereas the best  $\mu$  is closer to the one based on the theory.

**Conclusion of the experimental results.** The S-Lasso has good performance when the regression vector is ‘smooth’ (*Examples* (c)-(d)). Nevertheless, even in situations made in favor of the Elastic-Net and the Fused-Lasso (*Examples* (b)), the S-Lasso performs similarly to the other methods when the tuning parameters are chosen based on the cross validation criterion. The S-Lasso is even better in these examples when the methods are constructed based on the theoretical considerations.

Moreover all the results according to the procedures for which the tuning parameters are chosen based on the theoretical study is a little unfair in disfavor of the Fused-Lasso. Indeed the rates of the tuning parameters have been calibrated based on a study made for the estimator  $\hat{\beta}^{Quad}$  (the Elastic-Net and the S-Lasso are two particular cases of this estimator). For the Lasso estimator, we also used the usual rate for  $\lambda$ . Even if the Fused-Lasso seems to be close to the S-Lasso, it turns out that similar choices for the tuning parameters lead to the worst results for the Fused-Lasso.

Based on results on *Examples* (c)-(d) it seems that the Fused-Lasso and the Elastic-Net imply a large bias for large values of  $\mu$  when the regression vector is smooth (also observed in [10]). They do not improve sufficiently the performance of the Lasso estimator in such situations. Even the ‘corrected’ Elastic-Net does not provide better results since the artificial correction seems to work for a small number of pairs  $(\lambda, \mu)$  that have to be chosen very carefully.

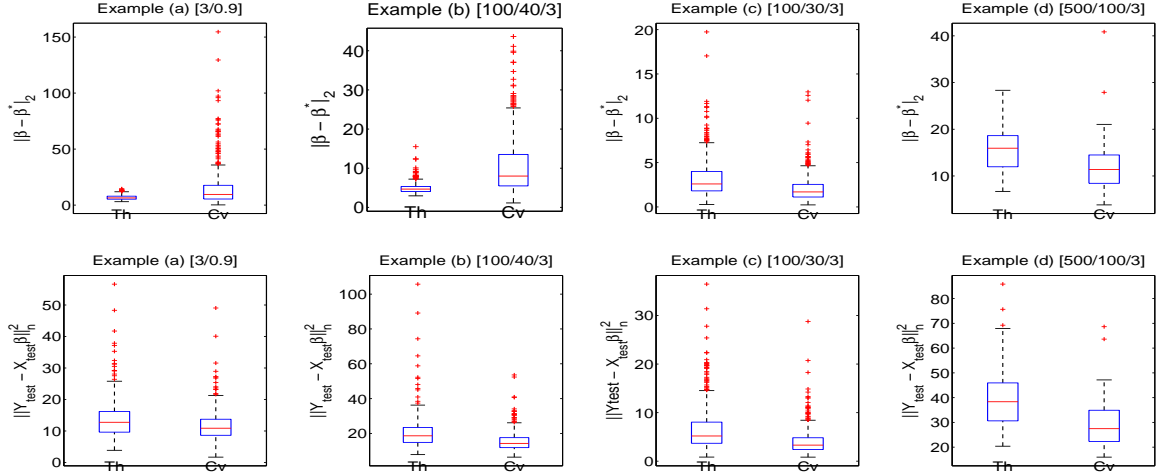


Figure 6: Evaluation of the  $\ell_2$  estimation error  $\|\hat{\beta} - \beta^*\|_2$  (top) and the prediction error  $\|Y_{test} - X_{test}\hat{\beta}\|_n^2$  (bottom) of the S-Lasso based on 500 replications. For each subplot: *Left*: The tuning parameters are chosen by 10 fold cross validation. *Right*: The tuning parameters are chosen based on the theoretical study. We refer to Table 2 for an evaluation of these tuning parameters

One can think of 2-stage methods to obtain better performance for the Fused-Lasso and the Elastic-Net (and also for the S-Lasso and the Lasso), where for instance an ordinary least squares is fitted based on the estimated support. This technique reduces of course the bias of the procedures and we refer to [2] for a nice theoretical study of such procedures. However we attempted here to compare the performance for the (1-stage) methods and observe how well the S-Lasso approaches the true regression vector.

## 4.2 Pseudo-real dataset

We apply all the methods we previously studied on artificially dataset generated from the riboflavin data. These data are about riboflavin (vitamin B2) production by *B. subtilis*. They kindly have been provided to us by DSM Nutritional Products (Switzerland). In the original data, the real-valued response variable is the logarithm of the riboflavin production rate, and there are  $p = 4088$  covariates measuring the logarithm of the expression level of 4088 genes that cover essentially the whole genome of *Bacillus subtilis*. The sample size is  $n = 71$ .

Here we are not interested in the riboflavin production, but only in a covariates matrix  $X$  coming from a real application. We use this design matrix to generate an artificial response vector thanks to a ‘smooth’ regression vector as in Equation (1). Let us mention that this trick to generate pseudo-real datasets has already been used in [22]. In what follows, we consider two different applications based on the real covariates matrix provided by the riboflavin dataset. In the first application, says *Application 1*, let us define  $X$  as the 1023 first covariates of the riboflavin dataset. Moreover let us define the regression vector  $\beta^*$ , such that  $\beta_j^* = 10 \cdot \exp -\frac{1}{1 - ((j-125)/125.1)^2}$  for  $j = 1, \dots, 250$  (cf. Figure 8), and the noise level  $\sigma = 3$ . Hence  $n = 71$  and  $p = 1023$  and then this is a high-dimensional setting with  $p \gg n$ , where the number of non-zero components (the sparsity index  $\mathcal{A}^*$ ) is larger than the sample size  $n$ . According to the second application, says *Application 2*, we restrict  $X$  to the 300 first covariates of the riboflavin dataset. The regression vector  $\beta^*$  is such that  $\beta_j^* = 10 \cdot \exp -\frac{1}{1 - ((j-25)/25.1)^2}$  for  $j = 1, \dots, 50$  (cf. Figure 8), and the noise level  $\sigma = 3$ . This is a more usual high-dimensional case where the sparsity index  $\mathcal{A}^*$  is smaller than the sample size  $n$ .

Let us now detail the obtained results for different experiences. First we mention that,

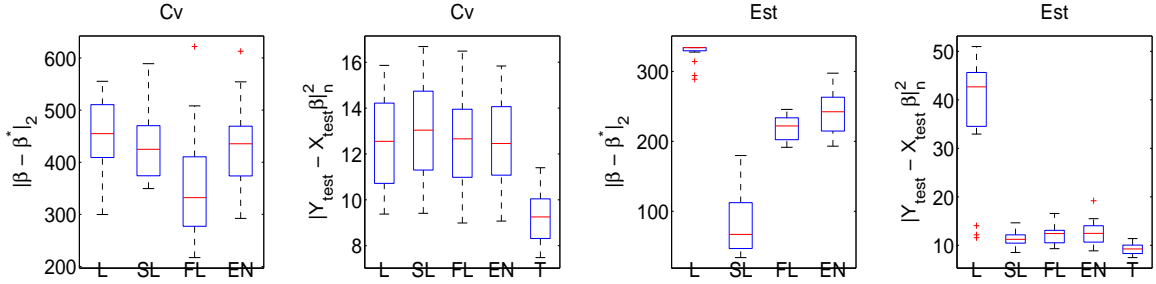


Figure 7: Evaluation of the  $\ell_2$  estimation error  $|\hat{\beta} - \beta^*|_2$  and the prediction error  $\|Y_{test} - X_{test}\hat{\beta}\|_n^2$  of the Lasso (L), the S-Lasso (SL), the Fused-Lasso (FL) and the Elastic-Net (EN) applied to the pseudo-real data, and based on 20 replications of *Application 2*. *Left; Center-left*: The tuning parameters are chosen by 10 fold cross validation. *Center-right; Right*: The tuning parameters minimize the estimation error.

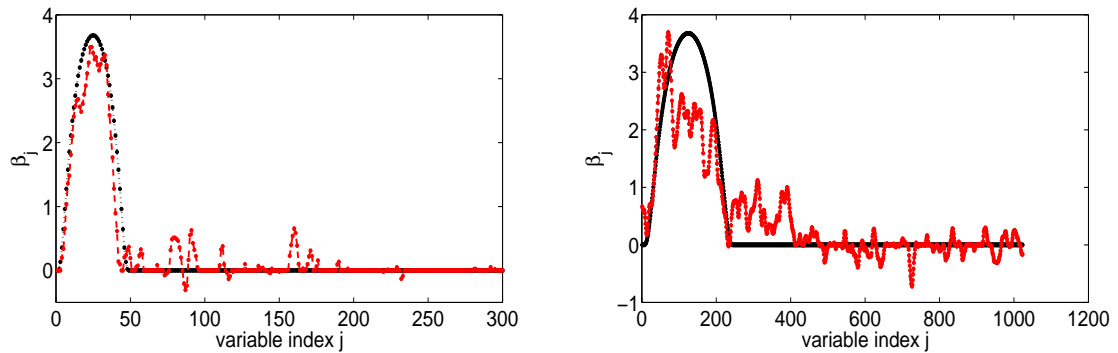


Figure 8: Best reconstitution of the regression vector  $\beta^*$  (black curve) by the SL-Lasso estimator (red curve). *Left*: On *Application 2*. *Right*: On *Application 1*.

with the exception of the S-Lasso, all the methods provide an estimation of the regression vector which is characterized by large variations in the values of the successive components when  $\mu$  is small (for the Elastic-Net and the Fused-Lasso), and by large bias when  $\mu$  is large. Hence we focus here on the S-Lasso estimator. Nevertheless, we display the comparison of all the methods in terms of accuracy in Figure 7 when the methods are applied to *Application 2*. Even though the S-Lasso estimator is outperformed when the tuning parameter is chosen by cross validation (by the Fused-Lasso for the estimation error and by all the methods for the prediction; cf. Figures 7 (left and center-left)), it turns out that we can disclose a S-Lasso solution which performs better than the other methods as displayed in Figures 7 (center-right and right). One of the best solution of the S-Lasso estimator in this *Application 2* can also be seen in Figure 8 (left). We observe how the S-Lasso succeed to reconstruct the ‘smooth’ regression vector  $\beta^*$ .

Finally, let us consider *Application 1*, and let’s recall that the sparsity index is here larger than the sample size. Figure 8 (right) displays the best reconstitution of the regression vector on this very difficult problem. We observe that the S-Lasso succeeds only partly to reconstruct the true regression vector. On the simulation study, we met a similar close situation in *Example (d)* [100/30/3] (cf. Figure 5), where the S-Lasso perfectly estimated  $\beta^*$ . However, the situation here is even more difficult since the sparsity index is much larger than the sample size and since many high and negative correlations appear between the covariates in the riboflavin dataset.

## 5 Conclusion

In this paper, we introduced the Lasso-type estimator  $\hat{\beta}^{Quad}$  which consists in two penalty terms: the  $\ell_1$  penalty which ensures sparsity; and a quadratic penalty which captures some structure in the regression vector. We showed that this estimator satisfies good theoretical performance specifically when the Lasso estimator might fail. As particular cases we considered the Elastic-Net and the S-Lasso. We pointed the interest to use such methods respectively when correlations between variables exist and when the regression vector is ‘smooth’.

In practice, we considered the performance of the S-Lasso estimator compared to the Lasso, the Elastic-Net and the Fused-Lasso in terms of prediction and estimation accuracy. We illustrated the superiority of the S-Lasso in several simulation experiments where the regression vector has a particular structure. We also observed that the theoretical calibration of the tuning parameters provides close performance as when they are chosen by 10 fold cross validation. The methods have also been applied to *pseudo real* examples based on the riboflavin dataset.

According to some simulation studies (as in *Example (d)* [100/30/ $\sigma$ ]), an interesting point would be whether the S-Lasso satisfies Sparsity Inequalities which can take into account more the ‘smoothness’ of the regression vector  $\beta^*$ . This is the topic of future works.

## 6 Proofs

We first provide two concentration results: the first deals with Gaussian noise and the second concerns noise admitting finite variance.

**Lemma 2.** *Let  $\eta \in (0, 1)$ . Let  $0 < \tau \leq 1$ , be a real number. Let  $\Lambda_{n,p}$  be the random event defined by  $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq \tau\lambda_n\}$  where  $V_j = n^{-1} \sum_{i=1}^n x_{i,j}\varepsilon_i$ . Let us define  $\lambda_n = \frac{2\sqrt{2}}{\tau}\sigma\sqrt{n^{-1}\log(p/\eta)}$ . Then*

$$\mathbb{P}\left(\max_{j=1,\dots,p} 2|V_j| \leq \tau\lambda_n\right) \geq 1 - \eta.$$

*Proof.* Since  $V_j \sim \mathcal{N}(0, n^{-1}\sigma^2)$  for any  $j \in \{1, \dots, p\}$ , an elementary Gaussian inequality gives

$$\begin{aligned} \mathbb{P}\left(\max_{j=1,\dots,p} |V_j| \geq \tau\lambda_n/2\right) &\leq p \max_{j=1,\dots,p} \mathbb{P}(|V_j| \geq \tau\lambda_n/2) \\ &\leq p \exp\left(-\frac{n}{2\sigma^2} \left(\frac{\tau\lambda_n}{2}\right)^2\right) = \eta. \end{aligned}$$

This ends the proof. □

**Lemma 3.** *Let  $\eta \in (0, 1)$ . Let  $0 < \tau \leq 1$ , be a real number. Denote also by  $L$  the constant such that  $n^{-1} \sum_{i=1}^n \max_{j=1,\dots,p} x_{i,j}^2 \leq L$ . Let  $\Lambda_{n,p}$  be the random event defined by  $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq \tau\lambda_n\}$  where  $V_j = n^{-1} \sum_{i=1}^n x_{i,j}\varepsilon_i$  is such that for any  $i = 1, \dots, n$ ,  $x_{i,j}^2 \leq L$  and the  $\varepsilon_i$ 's are independent random variables with zero mean and finite variance  $\mathbb{E}\varepsilon_i^2 \leq \sigma^2$ . Denote by  $K_{Nem}$  the quantity  $K_{Nem} = \inf_{q \in [2, \infty] \cap \mathbb{R}} (q-1)p^{2/q}$ . Then for  $\lambda_n = \frac{2\sigma}{\tau} \sqrt{\frac{K_{Nem}L}{n\eta}}$ , we have*

$$\mathbb{P}\left(\max_{j=1,\dots,p} 2|V_j| \leq \tau\lambda_n\right) \geq 1 - \eta.$$

*Proof.* This inequality use an inequality on the expectation of supremum of square of sum of independent random variables that can be found in [11, Theorem 2.2]. Let us mention that  $2e \log(p) - 3e < K_{Nem} < 2e \log(p) - e$ . Markov Inequality and Theorem 2.2 in [11] (with  $r = \infty$ ) imply

$$\begin{aligned} \mathbb{P} \left( \max_{j=1, \dots, p} |V_j| \geq \tau \lambda_n / 2 \right) &\leq \frac{4}{\tau^2 \lambda_n^2} \mathbb{E} \left( \max_{j=1, \dots, p} V_j^2 \right) \\ &\leq \frac{4K_{Nem}}{\tau^2 \lambda_n^2} \sum_{i=1}^n \mathbb{E} \left( \max_{j=1, \dots, p} n^{-2} x_{i,j}^2 \varepsilon_i^2 \right) \\ &\leq \frac{4\sigma^2 K_{Nem}}{\tau^2 n \lambda_n^2} n^{-1} \sum_{i=1}^n \max_{j=1, \dots, p} x_{i,j}^2 \leq \eta, \end{aligned} \quad (14)$$

where we used the definition of  $\lambda_n = \frac{2\sigma}{\tau} \sqrt{\frac{K_{Nem}L}{n\eta}}$  in the last inequality. Theorem 2.2 in [11] is used to obtain (14).  $\square$

*Proof of Theorem 1.* We provide a first result which may help the legibility of the paper. It states that the squared risk and the  $\ell_1$ -estimation error are controlled by the restricted  $\ell_2$ -estimation error  $|\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2$ .

**Proposition 3.** *Let  $\hat{\beta}^{Quad}$  be the estimator defined by (2)-(4) with tuning parameters  $\lambda_n$  and  $\mu_n$ . Let  $0 < \tau \leq 1$  be a real number. On the event  $\Lambda_{n,p} = \{\max_{j=1, \dots, p} 2|V_j| \leq \tau \lambda_n\}$  with  $V_j = n^{-1} \sum_{i=1}^n x_{i,j} \varepsilon_i$ , if  $\tau = 1/2$  we have*

$$\frac{1}{n} \left| \tilde{X} \beta^* - \tilde{X} \hat{\beta}^{Quad} \right|_2^2 + \frac{\lambda_n}{2} |\beta^* - \hat{\beta}^{Quad}|_1 \leq r_n |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2, \quad (15)$$

where  $r_n = 2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{J} \beta^*|_2$ , and  $\mathcal{B}$  is a set including  $\mathcal{A}^*$ .

*Proof.* Let first  $\tilde{X}$ ,  $\tilde{Y}$  and  $\tilde{\varepsilon}$  be the augmented dataset defined by

$$\tilde{X} = \begin{pmatrix} X \\ \sqrt{n\mu_n} \mathbf{J} \end{pmatrix}, \quad \text{and} \quad \tilde{Y} = \begin{pmatrix} Y \\ \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \tilde{\varepsilon} = \begin{pmatrix} \varepsilon \\ -\sqrt{n\mu_n} \mathbf{J} \beta^* \end{pmatrix},$$

where  $\mathbf{0}$  is a vector of size  $p$  containing only zeros and  $\mathbf{J}$  is the  $p \times p$  matrix given by (5). Then we have  $\tilde{Y} = \tilde{X} \beta^* + \tilde{\varepsilon}$ , and the estimator  $\hat{\beta}^{Quad}$ , solution of the minimization problem (2) with the penalty given by (4), is also the minimizer of

$$\frac{1}{n} \left| \tilde{Y} - \tilde{X} \beta \right|_2^2 + \lambda_n |\beta|_1.$$

Hence, by definition of the estimator  $\hat{\beta}^{Quad}$  we can write

$$\begin{aligned} &\frac{1}{n} \left| \tilde{Y} - \tilde{X} \hat{\beta}^{Quad} \right|_2^2 + \lambda_n |\hat{\beta}^{Quad}|_1 \leq \frac{1}{n} \left| \tilde{Y} - \tilde{X} \beta^* \right|_2^2 + \lambda_n |\beta^*|_1 \\ \iff &\frac{1}{n} \left| \tilde{X} \beta^* - \tilde{X} \hat{\beta}^{Quad} + \tilde{\varepsilon} \right|_2^2 - \frac{1}{n} |\tilde{\varepsilon}|_2^2 \leq \lambda_n |\beta^*|_1 - \lambda_n |\hat{\beta}^{Quad}|_1 \\ \iff &\frac{1}{n} \left| \tilde{X} \beta^* - \tilde{X} \hat{\beta}^{Quad} \right|_2^2 \leq \lambda_n \left[ |\beta^*|_1 - |\hat{\beta}^{Quad}|_1 \right] + \frac{2}{n} \tilde{\varepsilon}' \tilde{X} (\beta^* - \hat{\beta}^{Quad}). \end{aligned}$$

Let us now consider the term  $\frac{2}{n} \tilde{\varepsilon}' \tilde{X} (\beta^* - \hat{\beta}^{Quad})$ . By the definition of  $\tilde{X}$  and  $\tilde{\varepsilon}$ , we have the decomposition  $\frac{1}{n} \tilde{\varepsilon}' \tilde{X} (\beta^* - \hat{\beta}^{Quad}) = \frac{1}{n} \varepsilon' X (\beta^* - \hat{\beta}^{Quad}) - \mu_n \beta^{*'} \mathbf{J}' \mathbf{J} (\beta^* - \hat{\beta}^{Quad})$ . The first term

in this decomposition is quite common in the literature and we treat it using arguments which can be found for instance in [7]. We then need to adapt those arguments in order to deal with the second term of the decomposition  $\mu_n \beta^{*'} \mathbf{J}' \mathbf{J} (\beta^* - \hat{\beta}^{Quad})$  in the same time. Recall that  $\mathcal{A}^* = \{j : \beta_j^* \neq 0\}$  and that  $\mathbf{J}' \mathbf{J} = \tilde{\mathcal{J}}$ . Let  $0 < \tau \leq 1$  be a real number. Then, on the event  $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq \tau \lambda_n\}$  with  $V_j = n^{-1} \sum_{i=1}^n x_{i,j} \varepsilon_i$ , we have

$$\begin{aligned} \frac{1}{n} \left| \tilde{\mathcal{X}} \beta^* - \tilde{\mathcal{X}} \hat{\beta}^{Quad} \right|_2^2 &\leq \lambda_n \left[ |\beta^*|_1 - |\hat{\beta}^{Quad}|_1 \right] + \tau \lambda_n |\beta^* - \hat{\beta}^{Quad}|_1 \\ &\quad - 2\mu_n \beta^{*'} \tilde{\mathcal{J}} (\beta^* - \hat{\beta}^{Quad}). \end{aligned} \quad (16)$$

The remainder of this proof is linked to the way we choose to treat the term  $\mu_n \beta^{*'} \tilde{\mathcal{J}} (\beta^* - \hat{\beta}^{Quad})$  and in particular in the way we choose to link the RHS of Inequality (16) to the quantity  $|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_2$ . We obviously can write

$$-\mu_n \beta^{*'} \tilde{\mathcal{J}} (\beta^* - \hat{\beta}^{Quad}) = -\mu_n \beta_{\mathcal{B}}^{*'} \tilde{\mathcal{J}} (\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}) \leq \mu_n |\tilde{\mathcal{J}} \beta^*|_2 |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2,$$

where  $\mathcal{B}$  is the smallest set of indices such that the first equality holds. Note that the set  $\mathcal{B}$  includes  $\mathcal{A}^*$ , the true sparsity set, and is not much larger due to the sparsity of  $\tilde{\mathcal{J}}$ .

Now let  $\tau = 1/2$  in (16), add  $2^{-1} \lambda_n |\beta^* - \hat{\beta}^{Quad}|_1$  to both sides of this inequality. We then get

$$\begin{aligned} \frac{1}{n} \left| \tilde{\mathcal{X}} \beta^* - \tilde{\mathcal{X}} \hat{\beta}^{Quad} \right|_2^2 + \frac{\lambda_n}{2} |\beta^* - \hat{\beta}^{Quad}|_1 &\leq \lambda_n \left[ |\beta^*|_1 - |\hat{\beta}^{Quad}|_1 + |\beta^* - \hat{\beta}^{Quad}|_1 \right] \\ &\quad + 2\mu_n |\tilde{\mathcal{J}} \beta^*|_2 |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2 \\ &\leq 2\lambda_n \sum_{j \in \mathcal{A}} \left| \beta_j^* - \hat{\beta}_j^{Quad} \right| + 2\mu_n |\tilde{\mathcal{J}} \beta^*|_2 |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2 \\ &\leq r_n |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2, \end{aligned} \quad (17)$$

where  $r_n = 2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{\mathcal{J}} \beta^*|_2$ , since  $|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_1 \leq \sqrt{|\mathcal{A}^*|} |\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_2 \leq \sqrt{|\mathcal{A}^*|} |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2$ . In the second above inequality, we used the fact that  $|\beta_j^* - \hat{\beta}_j^{Quad}| + |\beta_j^*| - |\hat{\beta}_j^{Quad}| = 0$  for any  $j \notin \mathcal{A}$  and to the triangular inequality. This is the claim of Proposition 3 when  $\tilde{\mathcal{J}}$  is sparse.  $\square$

Let us now proof the main theorem. Thanks to Inequality (15) in Proposition 3, we easily obtain that

$$|\beta^* - \hat{\beta}^{Quad}|_1 \leq \varrho_n |\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2, \quad (18)$$

where  $\varrho_n := 2r_n/\lambda_n = 4\sqrt{|\mathcal{A}^*|} + \frac{2\mu_n}{\lambda_n} |\tilde{\mathcal{J}} \beta^*|_2$ . Then the vector  $\beta^* - \hat{\beta}^{Quad}$  is an admissible vector  $\Delta$  in Assumption  $B(\mathcal{B})$ . As a consequence, using this assumption in Equation (15), we get on one hand

$$\frac{1}{n} \left| \tilde{\mathcal{X}} \beta^* - \tilde{\mathcal{X}} \hat{\beta}^{Quad} \right|_2^2 \leq \frac{r_n}{\sqrt{\phi}} \sqrt{\frac{1}{n}} \left| \tilde{\mathcal{X}} \beta^* - \tilde{\mathcal{X}} \hat{\beta}^{Quad} \right|_2,$$

and a simple simplification leads to the first part of the result

$$\frac{1}{n} \left| \tilde{\mathcal{X}} \beta^* - \tilde{\mathcal{X}} \hat{\beta}^{Quad} \right|_2^2 \leq \phi^{-1} (2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{\mathcal{J}} \beta^*|_2)^2. \quad (19)$$

On the other hand, Inequality (18), combined to Assumption  $B(\mathcal{B})$  and Inequality (19), implies

$$|\beta^* - \hat{\beta}^{Quad}|_1 \leq 2\phi^{-1} \frac{(2\lambda_n \sqrt{|\mathcal{A}^*|} + 2\mu_n |\tilde{\mathcal{J}} \beta^*|_2)^2}{\lambda_n},$$

which is the desired bound on the  $\ell_1$  estimation error given in Theorem 1. The proof is completed when we use Lemma 2 with  $\tau = 1/2$  to control the probability of the event  $\Lambda_{n,p}$ .  $\square$

*Proof of Proposition 1.* We first provide a bound on  $|\beta_{\Theta}^* - \hat{\beta}_{\Theta}^{Quad}|_2$  for  $\Theta = \mathcal{B} \cup \mathcal{C}$ . Theorem 1 states a bounds on the prediction error and on the  $\ell_1$  estimation error under Assumption  $B(\mathcal{B})$ . Here we do not care about the  $\ell_1$  estimation error. Then one can observe that in the intermediate step between (16) and (17) in the previous proof, one can avoid the addition of the term  $\lambda_n/2|\beta^{Quad} - \beta^*|_1$ . As a consequence, we obtain (19) but with  $\tau = 1$  instead of  $1/2$  in (16). Apart from this value of  $\tau$  everything remains the same.

More particularly, thanks to (18) we can use Assumption  $B'(\mathcal{B} \cup \mathcal{C})$ , which directly implies that the following inequality holds  $|\beta_{\Theta}^* - \hat{\beta}_{\Theta}^{Quad}|_2 \leq \sqrt{\phi^{-1}} \sqrt{\frac{1}{n}} \left| \tilde{X}\beta^* - \tilde{X}\hat{\beta}^{Quad} \right|_2$ , with  $\Theta = \mathcal{B} \cup \mathcal{C}$ . Combining this inequality with (19), we easily get

$$|\beta_{\Theta}^* - \hat{\beta}_{\Theta}^{Quad}|_2 \leq 2\phi^{-1}(\lambda_n \sqrt{|\mathcal{A}^*|} + \mu_n |\tilde{J}\beta^*|_2), \quad (20)$$

with  $\Theta = \mathcal{B} \cup \mathcal{C}$ . Now, we consider the term  $|\beta_{\Theta^c}^* - \hat{\beta}_{\Theta^c}^{Quad}|_2$ . Denote by  $\delta$  the vector  $\delta = \beta^* - \hat{\beta}^{Quad}$  for shorten. For any  $p$ -dimensional vector  $a$ , let  $a_{(1)} \leq a_{(2)} \leq \dots a_{(p)}$  be the corresponding ranked sequence. Given this new notation, note that for any  $j \in [1, \dots, p]$ , the inequality  $|\delta_{\mathcal{B}^c}|_{(j)} \leq |\delta_{\mathcal{B}^c}|_1 \times j^{-1}$  holds. As a consequence

$$|\delta_{\Theta^c}|_2^2 \leq |\delta_{\mathcal{B}^c}|_1^2 \sum_{j \geq m+1} j^{-2} \leq m^{-1} |\delta_{\mathcal{B}^c}|_1^2,$$

where we recall that  $\Theta = \mathcal{B} \cup \mathcal{C}$ , with  $|\mathcal{B}| = m$ . Then using the last display with (18) yields to

$$|\delta_{\Theta^c}|_2 \leq \frac{\varrho_n}{\sqrt{m}} |\delta_{\mathcal{B}}|_2 \leq \frac{\varrho_n}{\sqrt{m}} |\delta_{\Theta}|_2,$$

where  $\varrho_n = 4\sqrt{|\mathcal{A}|} + \frac{4\mu_n}{\lambda_n} |\tilde{J}\beta^*|_2$ . Combine this last inequality with (20) implies

$$|\delta|_2 \leq \left(1 + \frac{\varrho_n}{\sqrt{m}}\right) |\delta_{\Theta}|_2 \leq 2\phi^{-1} \left(1 + \frac{\varrho_n}{\sqrt{m}}\right) (\lambda_n \sqrt{|\mathcal{A}^*|} + \mu_n |\tilde{J}\beta^*|_2).$$

Since  $|\delta|_{\infty} \leq |\delta|_2$ , we obtained the desired control on the sup-norm of  $\beta^* - \hat{\beta}^{Quad}$ .  $\square$

*Proof of Theorem 2.* This result is quite natural since it is a direct consequence of Proposition 1. We refer the reader to the proof of Theorem 2 in [18] for instance.  $\square$

*Proof of Theorem 3.* We consider now the case of general matrices  $\tilde{J}$ . Most of the proof is similar to the sparse case (Proof of Theorem 1 above). The same reasoning leads to (16) and the only different occurs when we deal with the term  $-\mu_n \beta^{*'} \tilde{J}(\beta^* - \hat{\beta}^{Quad})$ . We have here

$$-\mu_n \beta^{*'} \tilde{J}(\beta^* - \hat{\beta}^{Quad}) \leq \mu_n |\tilde{J}\beta^*|_{\infty} |\beta^* - \hat{\beta}^{Quad}|_1.$$

Then, if we set  $\tau = \frac{1}{4}$  and the tuning parameter  $\mu_n = \frac{\lambda_n}{8|\tilde{J}\beta^*|_{\infty}}$ , Inequality (16) becomes

$$\frac{1}{n} \left| \tilde{X}\beta^* - \tilde{X}\hat{\beta}^{Quad} \right|_2^2 \leq \lambda_n \left[ |\beta^*|_1 - |\hat{\beta}^{Quad}|_1 \right] + \frac{\lambda_n}{2} |\beta^* - \hat{\beta}^{Quad}|_1.$$

Add  $2^{-1} \lambda_n |\beta^* - \hat{\beta}^{Quad}|_1$  to both sides of the previous inequality and then thanks to the fact that  $|\beta_j^* - \hat{\beta}_j^{Quad}| + |\beta_j^*| - |\hat{\beta}_j^{Quad}| = 0$  for any  $j \notin \mathcal{A}^*$  and to the triangular inequality, the above inequality implies that (we refer to the proof of Proposition 3 for similar arguments).

$$\frac{1}{n} \left| \tilde{X}\beta^* - \tilde{X}\hat{\beta}^{Quad} \right|_2^2 + \frac{\lambda_n}{2} |\beta^* - \hat{\beta}^{Quad}|_1 \leq 2\lambda_n \sqrt{|\mathcal{A}^*|} |\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_2.$$

This above intermediate result is the analogous of Proposition 3 in the case where  $\tilde{J}$  is general. That is, we get a similar bound but depending on  $|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_2$  instead of  $|\beta_{\mathcal{B}}^* - \hat{\beta}_{\mathcal{B}}^{Quad}|_2$  and with  $r_n = 2\lambda_n\sqrt{|\mathcal{A}^*|}$ . Note also that (18) is replaced by the following linear inequality  $|\beta^* - \hat{\beta}^{Quad}|_1 \leq 4|\beta_{\mathcal{A}}^* - \hat{\beta}_{\mathcal{A}}^{Quad}|_1$ . Taking into account this changing, we use can use Assumption RE instead of Assumption  $B(\mathcal{B})$  and then a similar reasoning as in the proof of Theorem 1 leads to the desired results.  $\square$

*Proof of Proposition 2.* Using exactly the same reasoning as in the proof of Proposition 1 but based on Theorem 3 instead of Theorem 1 we obtain with probability at least  $1 - \eta$

$$|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_2 \leq 2\phi^{-1}\lambda_n\sqrt{|\mathcal{A}^*|}, \quad (21)$$

since here  $\tau$  becomes equal to  $1/2$  in Lemma 2. This completes the proof of the first part of the Proposition. We now show that  $\mathcal{A}^* \subset \hat{\mathcal{A}}$  with high probability. Thanks to (21), we have with high probability  $|\beta_{\mathcal{A}^*}^* - \hat{\beta}_{\mathcal{A}^*}^{Quad}|_\infty \leq U$  where we used  $U = 2\phi^{-1}\lambda_n\sqrt{|\mathcal{A}^*|}$  for short. But

$$|\hat{\beta}_{\mathcal{A}^*}^{Quad} - \beta_{\mathcal{A}^*}^*|_\infty \leq U \quad \Leftrightarrow \quad \beta_j^* - U \leq \hat{\beta}_j^{Quad} \leq \beta_j^* + U \quad \forall j \in \mathcal{A}^*.$$

Note that by assumption, we have  $|\beta_j^*| > U, \forall j \in \mathcal{A}^*$ . Then if we distinguish the case  $\beta_j^* > 0$  and the case  $\beta_j^* < 0$ , we easily conclude that  $\beta_j^* > 0$  implies  $\hat{\beta}_j^{Quad} > 0$  and  $\beta_j^* < 0$  implies  $\hat{\beta}_j^{Quad} < 0$ . This ables us to write

$$\mathbb{P}(\text{Sgn}(\hat{\beta}_{\mathcal{A}^*}^{Quad}) = \text{Sgn}(\beta_{\mathcal{A}^*}^*)) \geq \mathbb{P}(|\hat{\beta}_{\mathcal{A}^*}^{Quad} - \beta_{\mathcal{A}^*}^*|_\infty \leq U) \geq 1 - \eta,$$

and this naturally implies the that  $\mathcal{A}^* \subset \hat{\mathcal{A}}$  with high probability.  $\square$

*Proof of Theorem 4.* We now show that  $\hat{\mathcal{A}} \subset \mathcal{A}^*$  with high probability. This proof is quite inspired by the one by Bunea [5]. First of all, note that we can write the KKT conditions of the minimization problem (6) as

$$|K_n(\hat{\beta}^{Quad} - \beta^*) - \frac{X'_j\varepsilon}{n} + \mu_n\tilde{J}\beta^*|_\infty \leq \frac{\lambda_n}{2}. \quad (22)$$

Then all the solutions of the criterion (6) share the same active set

$$\hat{\mathcal{A}} = \left\{ j \in \{1, \dots, p\} : |(K_n(\hat{\beta}^{Quad} - \beta^*))_j - \frac{X'_j\varepsilon}{n} + \mu_n(\tilde{J}\beta^*)_j| = \frac{\lambda_n}{2} \right\}.$$

That is, all these solutions have non-zero components at the same positions. We now use this property to show that the estimator  $\hat{\beta}^{Quad}$  has non-zero components at the same positions as a well-controlled (but uncomputable) estimator on an event which occurs with high probability. For this purpose, let us consider the criterion

$$F(b) = \|Y - \sum_{j \in \mathcal{A}^*} X_j b_j\|_n^2 + \lambda_n \sum_{j \in \mathcal{A}^*} |b_j| + \mu_n b'_{\mathcal{A}^*} \mathbf{J}'_{\mathcal{A}^*} \mathbf{J}_{\mathcal{A}^*} b_{\mathcal{A}^*},$$

where recall that for any  $p$ -dimensional vector  $a$  and any set  $\Theta \subset \{1, \dots, p\}$ , the notation  $a_\Theta$  means that  $(a_\Theta)_j = a_j, \forall j \in \Theta$  and 0 otherwise. Moreover,  $\mathbf{J}_{\mathcal{A}^*}$  is such that  $(\mathbf{J}_{\mathcal{A}^*})_{j,k} = \mathbf{J}_{j,k}$  if  $j, k \in \mathcal{A}^*$  and 0 otherwise. Define the estimator

$$\hat{b} = \underset{b \in \mathbb{R}^p: b_{(\mathcal{A}^*)^c} = \mathbf{0}_p}{\text{Argmin}} F(b),$$

where  $\mathbf{0}_p$  is the zero in  $\mathbb{R}^p$ . Since we restricted  $\hat{b}$  to be zero when  $\beta^*$  is zero and that this is an information we do not have access to, we mention that the vector is not computable. Let us denote by  $\Omega$  the following event

$$\Omega = \bigcap_{k \notin \mathcal{A}^*} \left\{ \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) - \frac{X'_k \varepsilon}{n} + \mu_n \sum_{j \in \mathcal{A}^*} \tilde{J}_{j,k} \beta_j^* \right| < \frac{\lambda_n}{2} \right\}.$$

Observe how the event  $\Omega$  is inspired by the KKT conditions (22). Actually, on the event  $\Omega$ , the components  $\hat{b}_k$  with  $k \notin \mathcal{A}^*$  equals zero as they do not saturate KKT conditions. This makes the minimization of  $F(b)$  over  $b \in \mathbb{R}^p : b_{(\mathcal{A}^*)^c} = \mathbf{0}_p$  coincide with the minimization of the criterion (6) on  $\Omega$ . That is, the estimator  $\hat{b}$  turns out to be also solution of the original criterion (6) on  $\Omega$ . But  $\hat{\beta}^{Quad}$  is also solution of (6) and then, as we already pointed, this implies that on  $\Omega$ , both of  $\hat{\beta}^{Quad}$  and  $\hat{b}$  have non-zero components at the same positions and then,  $\hat{b}$  has non-zero components at components  $j \in \hat{\mathcal{A}}$ . Add the fact that by construction  $\hat{b}_{(\mathcal{A}^*)^c} = \mathbf{0}_p$ , then  $\hat{\mathcal{A}} \subset \mathcal{A}^*$  on the event  $\Omega$ . It then remains to prove that the event  $\Omega$  occurs with high probability. We have

$$\begin{aligned} \mathbb{P}(\hat{\mathcal{A}} \not\subset \mathcal{A}^*) &\leq \mathbb{P}(\Omega^c) \\ &\leq \sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) - \frac{X'_k \varepsilon}{n} + \mu_n \sum_{j \in \mathcal{A}^*} \tilde{J}_{j,k} \beta_j^* \right| \geq \frac{\lambda_n}{2} \right) \\ &\leq \sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) - \frac{X'_k \varepsilon}{n} \right| \geq \frac{\lambda_n}{2} - \mu_n |\tilde{J} \beta^*|_\infty \right) \\ &\leq \sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) - \frac{X'_k \varepsilon}{n} \right| \geq \frac{\lambda_n}{4} \right) \\ &\leq \sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) \right| \geq \frac{\lambda_n}{8} \right) + \sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \frac{X'_k \varepsilon}{n} \right| \geq \frac{\lambda_n}{8} \right) \quad (23) \end{aligned}$$

where we used the fact that for real number  $a$  and  $b$ , we have  $|a| + |b| \geq |a + b|$  in the third inequality and the fact that  $\mu_n = \frac{\lambda_n}{4|\tilde{J} \beta^*|_\infty}$  in the fourth one. Let us consider the last two terms

in the last display separately. i) First, since  $\lambda_n = 16\sigma \sqrt{\frac{\log(p/\sqrt{\eta p/(1+p)})}{n}}$ , and using close arguments to those employed in Lemma 2, we obtain  $\sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \frac{X'_k \varepsilon}{n} \right| \geq \frac{\lambda_n}{8} \right) \leq \eta \frac{1}{1+p}$ ; ii) according to  $\sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) \right| \geq \frac{\lambda_n}{8} \right)$ , we need to control  $\left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) \right|$  for every  $k \notin \mathcal{A}^*$ . On one hand, Assumption D implies that

$$\forall k \notin \mathcal{A}^* \quad \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) \right| \leq \sum_{j \in \mathcal{A}^*} |\hat{b}_j - \beta_j^*| t / |\mathcal{A}^*|. \quad (24)$$

By definition of  $\hat{b}$ , we just have to repeat the proof of Theorem 3 but with  $\hat{b}$  instead of  $\hat{\beta}^{Quad}$  and only on the true sparsity set  $\mathcal{A}^*$ . We get that on the event  $\Lambda_{n, \mathcal{A}^*} = \left\{ \max_{j \in \mathcal{A}^*} |X'_j \varepsilon| \leq \lambda_n / 8 \right\}$ , which is the same that  $\Lambda_{n,p}$  but using  $\mathcal{A}^*$  instead of  $\{1, \dots, p\}$ ,

$$\sum_{j \in \mathcal{A}^*} |\hat{b}_j - \beta_j^*| \leq 8\phi^{-1} \lambda_n |\mathcal{A}^*|.$$

Moreover, similar reasoning as in Lemma 2 leads to  $\mathbb{P}(\Lambda_{n,\mathcal{A}^*}^c) \leq \eta \frac{1}{1+p}$ . Combine this result with (24) and get

$$\begin{aligned} \sum_{k \notin \mathcal{A}^*} \mathbb{P} \left( \left| \sum_{j \in \mathcal{A}^*} (K_n)_{j,k} (\hat{b}_j - \beta_j^*) \right| \geq \frac{\lambda_n}{8} \right) &\leq p \mathbb{P} \left( \sum_{j \in \mathcal{A}^*} |\hat{b}_j - \beta_j^*| \geq \frac{|\mathcal{A}^*| \lambda_n}{8t} \right) \\ &\leq p \mathbb{P} \left( \sum_{j \in \mathcal{A}^*} |\hat{b}_j - \beta_j^*| \geq 8\phi^{-1} \lambda_n |\mathcal{A}^*| \right) \\ &\leq p \mathbb{P}(\Lambda_{n,\mathcal{A}^*}^c) \leq \eta \frac{p}{1+p}, \end{aligned}$$

provided that  $t \leq \frac{\phi}{64}$ . We finally conclude by this last inequality and (23) that  $\mathbb{P}(\hat{\mathcal{A}} \not\subseteq \mathcal{A}^*) \leq \eta \left( \frac{1}{1+p} + \frac{p}{1+p} \right) \leq \eta$ . Then we get the desired result.  $\square$

*Proof of Theorem 5.* This proof is almost the same as the one of Theorem 1. The only difference is the way to control the event  $\Lambda_{n,p} = \{\max_{j=1,\dots,p} 2|V_j| \leq \tau \lambda_n\}$  where  $V_j = n^{-1} \sum_{i=1}^n x_{i,j} \varepsilon_i$  when the noise admits only zero mean and finite variance. Then we do not use the concentration inequality provided in Lemma 2 for the Gaussian noise but an analog concentration inequality more adapted to this type of noise. This concentration inequality is given by Lemma 3 and we get

$$\mathbb{P} \left( \max_{j=1,\dots,p} 2|V_j| \leq \tau \lambda_n \right) \geq 1 - \eta,$$

for a value of  $\lambda_n = \frac{2\sigma}{\tau} \sqrt{\frac{KNemL}{n\eta}}$ . Then we set  $\tau = 1/2$  and we plug this new value of the tuning parameter  $\lambda_n$  instead to the one used to establish the previous results into Theorem 1. We just finish the proof by using the fact that  $\mu_n = \frac{\lambda_n \sqrt{|\mathcal{A}^*|}}{2|J\beta^*|_2}$  and we obtain the analog of Corollary 1.  $\square$

## References

- [1] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.
- [2] A. Belloni and V. Chernozhukov. Post- $\ell_1$ -Penalized estimation in high dimensional sparse linear regression models. Submitted, 2010.
- [3] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
- [4] F. Bunea. Consistent selection via the lasso for high dimensional approximating regression models. 2008. IMS Collections, B. Clarke and S. Ghosal Editors.
- [5] F. Bunea. Honest variable selection in linear and logistic regression models via  $\ell_1$  and  $\ell_1 + \ell_2$  penalization. *Electron. J. Stat.*, 2:1153–1194, 2008.
- [6] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

- [7] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.
- [8] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables lasso. *Math. Methods Statist.*, 17(4):317–326, 2008.
- [9] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.
- [10] Z. John Daye and X. Jessie Jeng. Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics & Data Analysis*, 53(4):1284–1298, 2009.
- [11] Lutz Dümbgen, Sara A. van de Geer, Mark C. Veraar, and Jon A. Wellner. Nemirovski’s inequalities revisited. *Amer. Math. Monthly*, 117(2):138–160, 2010.
- [12] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004. With discussion, and a rejoinder by the authors.
- [13] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.
- [14] M. Hebiri. Regularization with the smooth-lasso procedure. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2008.
- [15] J. Jia and B. Yu. On model selection consistency of elastic net when  $p \gg n$ . Tech. Report 756, Statistics, UC Berkeley, 2008.
- [16] S. Kim, K. Koh, S. Boyd, and D. Gorinevsky.  $l_1$  trend filtering. *SIAM Rev.*, 51(2):339–360, 2009.
- [17] S. R. Land and J. H. Friedman. Variable fusion: a new method of adaptive signal regression. *Manuscript*, 1996.
- [18] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.
- [19] L. Meier, S. van de Geer, and P. Bühlmann. The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(1):53–71, 2008.
- [20] N. Meinshausen. Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1):374–393, 2007.
- [21] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.
- [22] N. Meinshausen, L. Meier, and P. Bühlmann. p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104:1671–1681, 2009.
- [23] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.
- [24] Yu. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Sep 2007.

- [25] A. Rinaldo. Properties and refinements of the fused lasso. *Ann. Statist.*, 37(5B):2922–2952, 2009.
- [26] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Ann. Statist.*, 35(3):1012–1030, 2007.
- [27] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.
- [28] 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.
- [29] R.J. Tibshirani and J. Taylor. Regularization paths for least squares problems with generalized  $\ell_1$  penalties. Submitted, 2010.
- [30] 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.
- [31] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Elect. Journ. Statist.*, 3:1360–1392, 2009.
- [32] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -constrained quadratic programming. Manuscript, 2006.
- [33] 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.
- [34] M. Yuan and Y. Lin. On the non-negative garrote estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(2):143–161, 2007.
- [35] W-C. Yueh. Eigenvalues of several tridiagonal matrices. *Appl. Math. E-Notes*, 5:66–74 (electronic), 2005.
- [36] C-H. Zhang and J. Huang. The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.
- [37] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.
- [38] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.
- [39] 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.
- [40] H. Zou and H. Zhang. On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.*, 37(4):1733–1751, 2009.