Some theoretical results on the grouped variables Lasso

Christophe Chesneau, Mohamed Hebiri

Some theoretical results on the grouped variables Lasso

Christophe Chesneau and Mohamed Hebiri

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

Abstract

We consider the linear regression problem with Gaussian error. We estimate the unknown parameters via an estimator constructed from a grouped variables penalty. It can be viewed as a slight modification of the Group Lasso estimator introduced by Yuan and Lin [15]. We establish several new theoretical results which prove that the considered estimator exploits more the sparsity in the model than the well-known Lasso estimator.

Key words: Lasso, Group Lasso, Variable selection, Sparsity.
1991 MSC: Primary: 62J07, Secondary: 62H20.

1 Introduction

We focus on the usual linear regression model:

$$y_i = x_i\beta^* + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where the design $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p$ is deterministic, $\beta^* = (\beta_1^*, ..., \beta_p^*)' \in \mathbb{R}^p$ is the unknown parameter vector of interest and $\varepsilon_1, ..., \varepsilon_n$, are i.i.d. centered Gaussian random variables with variance σ^2. We wish to estimate β^* in the sparse case i.e. when many of its unknown components are equal to zero. Thus, only a subset of the design variables $(x_{.,j})_j$ are truly of interest.

It is well known that, in such case, Ordinary Least Square and Ridge regression procedures lead to bad control of the variance in the estimation. Selection type procedures are then recommended. They are of the form:

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

where $X = (x_1, \ldots, x_n)'$, $Y = (y_1, \ldots, y_n)'$ and pen : $\mathbb{R}^p \to \mathbb{R}$, a positive convex function. For any vector $a = (a_1, \ldots, a_n)'$, we adopt the notation $\|a\|_n^2 = n^{-1} \sum_{i=1}^n |a_i|^2$. The Lasso procedure introduced by Tibshirani [13] seems to respond to our objective: it performs both regression parameters estimation and variable selection. In the literature, the theoretical and computational Lasso properties as well as its asymptotic results have been intensively studied. See, for instance, Efron et al. [6], Meinshausen and Bühlmann [12], Fan and Li [7], Knight and Fu [9], Zou [17] and Zhao and Yu [16], among others. Using the advantages of the Lasso l_1-penalty, many new penalized procedures have been proposed to solve the linear regression problem. We refer to Fan and Li [7], Meinshausen [11], Zou [17], Zou and Hastie [18] and Tibshirani et al. [14].

In this paper, we study Lasso-type procedures which take into account the group structure of the variables. We consider a slight modification of the Group Lasso procedure developed by Yuan and Lin [15]. This construction has the ability of selecting variables by groups and evaluating the estimation of the regression parameters inside the groups in a Ridge-type fashion. For the sake of clarity, we call our modified Group Lasso: Grouped Variables Lasso. Such procedures are really promising as they nicely combine Lasso and Ridge with a single control parameter. Here, we derive new theoretical results to the Grouped Variables Lasso estimator based on the sparsity of the model. We adopt the following criterion that measures the performance of a given estimator $\tilde{\beta}$ of β^*: find the best rate $\varphi_{n,p}$ (i.e. as small as possible) satisfying the following inequality:

$$\mathbb{P}\left(\|X\tilde{\beta} - X\beta^*\|_n^2 \leq \varphi_{n,p}\right) \geq 1 - u_{n,p},$$

where $u_{n,p}$ is a positive sequence of the form $n^{-\alpha}p^{-\gamma}$ with $\alpha, \gamma > 0$. We prove that, for the same $u_{n,p}$, the Group Lasso exploits the sparsity of the model more efficiently than the original Lasso. This can easily be seen through the form of $\varphi_{n,p}$ which depends on the sparsity. This study is in the same spirit as that of Bunea et al. [3] for the aggregation problem via the Lasso penalty.

The rest of this article is organized as follows. The Grouped Variables Lasso estimator is described in the next section. Section 3 presents the assumptions made on the model. The theoretical performances of the considered estimator are studied in Section 4. Proofs are given in Section 5.

2 The Grouped Variables Lasso (GVL) estimator

We define the estimator $\hat{\beta}$ by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\text{Argmin}} \left\{ \|Y - X\beta\|_n^2 + \text{pen}(\beta) \right\}, \tag{2.1}$$

where pen is the penalty function defined by

$$\text{pen}(\beta) = 2 \sum_{l=1}^{L} \sqrt{\sum_{j \in G_l} |w_{n,j}|^2 |\beta_j|^2}.$$

In this expression, L is a positive integer, $G = (G_l)_l$ is a sequence of sets satisfying $\cup_{l=1}^{L} G_l = \{1, ..., p\}$ and, for any $u \neq v$ with $u, v \in \{1, \ldots, L\}$, $G_u \cap G_v = \emptyset$. The sequence $w = (w_{n,j})_j$ is defined by

$$w_{n,j} = \lambda_{n,p}(\sqrt{n}\|x_{.,j}\|_n) = \lambda_{n,p} \sqrt{\sum_{i=1}^{n} |x_{i,j}|^2}, \tag{2.2}$$

with

$$\lambda_{n,p} = \kappa \sigma n^{-1} \sqrt{\log n + \log p}, \tag{2.3}$$

where κ is a real number greater than 1. Given the definition of $\hat{\beta}$, specific choices of the sequence G determine specific estimators.

- *The Lasso estimator:* The Lasso estimator $\hat{\beta}^L$ is defined by (2.1) with $L = p$ and, for any $l \in \{1, ..., L\}$, $G_l = \{l\}$.

- *The Grouped Variables Lasso (GVL) estimator:* Assume that $p/\lfloor \log p \rfloor$ is an integer where $\lfloor a \rfloor$ denotes the whole number part of a. The Grouped Variables Lasso estimator $\hat{\beta}^G$ is defined by (2.1) with $L = p/\lfloor \log p \rfloor$ and, for any $l \in \{1, ..., L\}$, G_l is the following set:

$$G_l = \left\{ k \in \{1, \ldots, p\} : (l-1)\lfloor \log p \rfloor + 1 \leq k \leq l \lfloor \log p \rfloor \right\}. \tag{2.4}$$

Comments on the GVL estimator. The GVL estimator is a slight modification of the Group Lasso estimator proposed by Yuan and Lin [15]. The only differences are the choice of the blocks G_l and the fact that, in our setting, we do not have $X'_{G_l} X_{G_l} = I_{Card(G_l)}$, where X_{G_l} is the restriction of X on the block G_l. Recent developments concerning the Group Lasso method can be found in Kim et al. [8] and Meier et al. [10].

If X is the identity matrix I_n, then the linear regression model (1.1) becomes the standard Gaussian sequence model. Moreover, each component of the GVL

estimator $\hat{\beta}^G$ in the block G_l can be expressed in the following explicit form

$$\hat{\beta}_i^G = \left(1 - \sqrt{n}\lambda_{n,n} / \sqrt{\sum_{j \in G_l} |y_j|^2}\right)_+ y_i,$$

where $\lambda_{n,n} = \kappa\sigma n^{-1}\sqrt{2\log n}$. The notation $(a)_+$ means $\max(a, 0)$. In this case, $\hat{\beta}^G$ can be viewed as a slight modification of the blockwise Stein estimator. This construction enjoys powerful theoretical properties in various statistical approaches (oracle inequalities, (near) minimax optimality,...). See, for instance, Cavalier and Tsybakov [4].

3 Assumptions

Here, we adopt the notations of the previous section.

Assumption (A1). Let $X = (x_{i,j})_{i,j}$ be a $n \times p$ matrix. Let us define the sequence $(v_{n,j})_j$ by

$$v_{n,j} = \sqrt{\sum_{i=1}^{n} |x_{i,j}|^2} \tag{3.1}$$

and the set $\mathcal{S}_2$ by $\mathcal{S}_2 = \{a = (a_j)_j \in \mathbb{Z}^*;\ \sum_{j \in G_l} |a_j|^2 \leq 1\}$. There exists a constant $C_p \geq 1$ such that

$$\sup_{l=1,...,L} \sup_{a \in \mathcal{S}_2} \left(\sum_{j \in G_l} \sum_{k \in G_l} a_j a_k |v_{n,j}|^{-1} |v_{n,k}|^{-1} \sum_{i=1}^{n} x_{i,j} x_{i,k} \right) \leq C_p.$$

Lemma 3.1 below proves that Assumption (A1) is satisfied by a large variety of matrix.

Lemma 3.1 *Let $X = (x_{i,j})_{i,j}$ be a $n \times p$ matrix. Suppose that, for any $j, k \in \{1, .., p\}$, each component of the $p \times p$ matrix $X'X$ satisfies*

$$\sum_{i=1}^{n} x_{i,j} x_{i,k} = z_j z_k b_{|j-k|}, \tag{3.2}$$

where $z = (z_u)_u$ denotes a positive sequence and $b = (b_u)_u$ denotes a sequence in $l_1(\mathbb{N})$ with $b_0 > 0$. Then X satisfies Assumption (A1) with $C_p = 1 + 2b_0^{-1}\|b\|_{l_1}$.

For any set $\mathcal{B} \subseteq \{1, ..., L\}$, we make the following assumption:

Assumption (A2)($\mathcal{B}$) . Let us consider the sequence $(v_{n,j})_j$ defined by (3.1). We have

$$\sup_{l\in\mathcal{B}}\sup_{u=1,...,L}\sqrt{\sum_{j\in G_l}\sum_{\substack{k\in G_u\\k\neq j}}|v_{n,j}|^{-2}|v_{n,k}|^{-2}|\sum_{i=1}^{n}x_{i,j}x_{i,k}|^2}\leq(32)^{-1}Card(\mathcal{B})^{-1}.$$

If, for each $l\in\{1,...,p/\lfloor\log p\rfloor\}$, the set G_l is defined by (2.4), then Assumption (A2)($\mathcal{B}$) can be viewed as a particular grouped version of the "local" mutual coherence condition considered by Bunea et al. [2] in the aggregation framework. This "local" mutual coherence condition has been introduced by Donoho et al. [5].

Remark 3.1 *For any two sets $\mathcal{B}_1$ and $\mathcal{B}_2$ such that $\mathcal{B}_1\subseteq\mathcal{B}_2\subseteq\{1,...,L\}$, Assumption (A2)($\mathcal{B}_2$) implies Assumption (A2)($\mathcal{B}_1$).*

Remark 3.2 *If $\mathcal{B}=\{1,...,L\}$ then Assumption (A2)($\mathcal{B}$) implies Assumption (A1) with $C_p=(32)^{-1}L^{-1}$. This is an immediate consequence of the Hölder inequality.*

Example. A simple example of $n\times p$ matrix $X=(x_{i,j})_{i,j}$ which satisfies Assumptions (A1) and (A2)($\mathcal{B}$) for any $\mathcal{B}\subseteq\{1,...,L\}$ is the one characterized by the equality $\sum_{i=1}^{n}x_{i,j}x_{i,k}=(32)^{-1}p^{-\alpha|j-k|}$, with $\alpha\geq 1$. A concise proof is given below.

Thanks to Lemma 3.1, Assumption (A1) is satisfied with $C_p=1+2p/(p-1)\leq 4$. Moreover, we have $(32)^{-1}\sup_{l=1,...,L}\sup_{u=1,...,L}\sqrt{\sum_{j\in G_l}\sum_{\substack{k\in G_u\\k\neq j}}p^{-2\alpha|j-k|}}\leq(32)^{-1}p^{-\alpha}\sup_{l=1,...,L}Card(G_l)\leq(32)^{-1}L^{-1}$ and Assumption (A2)($\mathcal{B}$) is satisfied.

4 Theoretical properties

In this section, we set theoretical results for the GVL estimator and the Lasso estimator. We show that the GVL estimator is better, in some sense, than the Lasso estimator. Let us mention that all results in the present section do not exclude the case $p\geq n$.

4.1 Main results

Theorem 4.1 below investigates the upper bound for the least square error of the GVL estimator and the Lasso estimator.

Theorem 4.1 *Let us consider the regression model (1.1). Let $\hat{\beta}$ be either the Lasso estimator $\hat{\beta}^L$ or the GVL estimator $\hat{\beta}^G$. Let $\Lambda_{n,p}$ be the random event defined by*

$$\Lambda_{n,p} = \left\{ \max_{l=1,\dots,L} \sqrt{\sum_{j \in G_l} |w_{n,j}|^{-2} |V_j|^2} \leq 2^{-1} \right\}, \tag{4.1}$$

where $V_j = n^{-1} \sum_{i=1}^n x_{i,j} \varepsilon_i$, $w_{n,j}$ is defined by (2.2),

- *$L = p$ and $G_l = \{l\}$ when $\hat{\beta} = \hat{\beta}^L$,*
- *$L = p/\lfloor \log p \rfloor$ and G_l is described by (2.4) when $\hat{\beta} = \hat{\beta}^G$.*

Let Θ be the sparsity set defined by:

$$\Theta = \left\{ l \in \{1, \dots, L\} : \text{ there exists an integer } j_0 \in G_l \text{ such that } \beta^*_{j_0} \neq 0 \right\}. \tag{4.2}$$

Suppose that X satisfies Assumption (A2)($\mathcal{B}$) for any set $\mathcal{B}$ such that $\Theta \subseteq \mathcal{B} \subseteq \{1, ..., L\}$. Then, on the event $\Lambda_{n,p}$, we have

$$\|X\hat{\beta} - X\beta^*\|_n^2 \leq s_{n,p} Card(\Theta), \tag{4.3}$$

where $s_{n,p} = 16 n \lambda_{n,p}^2$.

The proof of Theorem 4.1 is based on the 'argmin' definition of the estimators $\{\hat{\beta}^L, \hat{\beta}^G\}$ and some technical inequalities. The main contribution of Theorem 4.1 concerns the GVL estimator $\hat{\beta}^G$. The result obtained for the Lasso estimator is an adaptation of those in Bunea et al. [3]. We have formulated it in order to make easier the comparison with the GVL estimator. Let us just mention that the nature of Θ defined by (4.2) will play a crucial role in our comparative study. Further details are given in Subsection 4.2.

Remark 4.1 *The inequality (4.3) can be proved for any $\lambda_{n,p} \geq 0$ instead of the specific choice of $\lambda_{n,p}$ given by (2.3).*

Thanks to the definition of $\lambda_{n,p}$, Propositions 4.1 below proves that, under some assumptions on X, the inequality (4.3) of Theorem 4.1 is true with a high probability.

Proposition 4.1 *Let us consider the regression model (1.1). Let $\hat{\beta}$ be either the Lasso estimator $\hat{\beta}^L$ or the GVL estimator $\hat{\beta}^G$. Let Θ be the sparsity set defined by (4.2). Suppose that X satisfies Assumptions (A1) and (A2)($\mathcal{B}$) for any set $\mathcal{B}$ such that $\Theta \subseteq \mathcal{B} \subseteq \{1, ..., L\}$. Then we have*

$$\mathbb{P}\left(\|X\hat{\beta} - X\beta^*\|_n^2 \leq s_{n,p} Card(\Theta) \right) \geq 1 - u_{n,p}, \tag{4.4}$$

where $s_{n,p} = 16 \kappa^2 \sigma^2 n^{-1} (\log n + \log p)$ and $u_{n,p} = p(np)^{-(2^{-1}\kappa - 1)^2/(2C_p)}$.

The proof of Proposition 4.1 uses the result of Theorem 4.1 and a concentration inequality of the form $\mathbb{P}(\Lambda_{n,p}^c) \leq u_{n,p}$, where $\Lambda_{n,p}^c$ denotes the complementary of the set (4.1). The main interest of this proposition is developed in Subsection 4.2 below.

Corollary 4.1 below shows that, under some extra condition on p and with another choice of $\lambda_{n,p}$ in the definitions of $\{\hat{\beta}^L, \hat{\beta}^G\}$, the result of Proposition 4.1 holds for a smaller bound $s_{n,p}$.

Corollary 4.1 *Let us adopt the same mathematical framework of Proposition 4.1. If there exists a positive sequence* $v = (v_n)_n$ *such that* $\lim_{n\to\infty} v_n = \infty$ *and* $v_n \leq p$*, then the estimator* $\hat{\beta} \in \{\hat{\beta}^L, \hat{\beta}^G\}$ *defined with*

$$\lambda_{n,p} = \kappa\sigma n^{-1}\sqrt{\log p} \tag{4.5}$$

and $\kappa \geq 2(1+\sqrt{2C_p})$*, satisfies the inequality (4.4) with* $s_{n,p} = 16\kappa^2\sigma^2 n^{-1}\log p$ *and* $u_{n,p} = v_n^{1-(2^{-1}\kappa-1)^2/(2C_p)}$.

The proof of Corollary 4.1 is rigorously similar to the proof of Proposition (4.1). The restriction $v_n \leq p$ is only used to obtain the following inequality $\mathbb{P}(\Lambda_{n,p}^c) \leq p^{1-(2^{-1}\kappa-1)^2/(2C_p)} \leq v_n^{1-(2^{-1}\kappa-1)^2/(2C_p)}$. Note also that the main difference with Proposition 4.1 is that Corollary 4.1 excludes the case p constant whereas Proposition 4.1 does not.

4.2 Comparison with the Lasso

Starting from Proposition 4.1 (and Corollary 4.1), we can set a significant result concerning the superiority of the GVL estimator over the Lasso estimator. First of all, let us notice that the set Θ defined by (4.2) does the link between the considered estimators and the sparsity in the model. For the Lasso estimator, it can be reexpressed as

$$\Theta = \Theta_L = \{l \in \{1, ..., p\};\, \beta_l^* \neq 0\}.$$

For the GVL estimator, we have

$$\Theta = \Theta_G = \left\{l \in \{1, \ldots, L\} : \text{ there exists an integer } j_0 \in G_l \text{ such that } \beta_{j_0}^* \neq 0\right\},$$

where $L = p/\lfloor \log p \rfloor$ and G_l is described by (2.4). Therefore, for any β^*, the following inequality always holds:

$$Card(\Theta_G) \leq Card(\Theta_L).$$

It follows from Proposition 4.1 that, with a high probability, the GVL estimator has a smaller least squared error than the Lasso estimator. This is only due to the fact that the GVL estimator exploits more the sparsity in the model than the Lasso.

Moreover, in the sparse case, the quantity $Card(\Theta_G)$ can be asymptotically smaller than $Card(\Theta_L)$. For example, if $p = n$ and the unknown parameter vector $\beta^* = (\beta_1^*, ..., \beta_n^*)'$ is defined by

$$\beta^* = (\underbrace{1, \ldots, 1}_{\log n}, \underbrace{0, \ldots, 0}_{n - \log n}),$$

then $Card(\Theta_G) = 1$ and $Card(\Theta_L) = \log n$.

4.3 Discussion on Assumption (A2)($\mathcal{B}$)

According to Remark 3.1, Assumption (A2)(Θ) is clearly less restrictive than Assumption (A2)($\mathcal{B}$) with $\Theta \subseteq \mathcal{B} \subseteq \{1, ..., L\}$. However, it requires the knowledge of the set Θ by the statistician. Moreover, we understand that this assumption is related to the sparsity of the model. Indeed the more sparse β^* is, the easier Assumption (A2)(Θ) will be fulfilled. The best case being when the correlations between variables in groups belonging to Θ and those in the others groups are concentrated in a few correlation coefficients. The others are set to 0.

We now introduce a new assumption which can replace, in some cases, Assumption (A2)($\mathcal{B}$).

Assumption (A3). Here, we exclusively consider the case where $p \leq n$. Let us consider the $p \times p$ matrix Ψ defined by $\Psi = (\sum_{i=1}^n x_{i,j} x_{i,k})_{j,k}$. For any $p \geq 2$, there exists a constant $c_p > 0$ such that the matrix Z defined by

$$Z = \Psi - c_p \operatorname{diag}(\Psi),$$

is positive semi-definite.

Assumption (A3) is the same as in Bunea et al. [2, Assumption (A3)] which has been introduced in the aggregation framework. We then refer to Bunea et al. [2, Remarks 4-5] for more details. Assumption (A3) is, for instance, always fulfilled for positive matrices X'X. It is important to notice that this assumption can be helpful when the "group mutual coherence" is not small enough. In other words, Assumptions (A2)($\mathcal{B}$) and (A3) can recover different types of design matrices.

Moreover, we can rewrite Proposition 4.1 with Assumption (A3) instead of

Assumption (A2)($\mathcal{B}$) with $\Theta \subseteq \mathcal{B} \subseteq \{1, ..., L\}$. The only difference is the quantity $s_{n,p}$ which becomes $s_{n,p} = 16{c_p}^{-1}\kappa^2\sigma^2 n^{-1}(\log n + \log p)$, where c_p is the constant appearing in Assumption (A3).

5 Proofs of the main results

[Proof of Lemma 3.1.] For the sake of simplicity in exposition and without loss of generality, we work on the set $G_1 = \{1, ..., \lfloor \log p \rfloor\}$. Let us notice that, for any $u \in G_1$, we have $v_{n,u} = \sqrt{\sum_{i=1}^n |x_{i,u}|^2} = z_u\sqrt{b_0}$. Therefore,

$$\begin{aligned}
&\sum_{j\in G_1}\sum_{k\in G_1} a_j a_k v_{n,j}^{-1} v_{n,k}^{-1} \sum_{i=1}^n x_{i,j}x_{i,k} \\
&= b_0^{-1} \sum_{j=1}^{\lfloor \log p \rfloor}\sum_{k=1}^{\lfloor \log p \rfloor} a_j a_k b_{|j-k|} = \sum_{j=1}^{\lfloor \log p \rfloor} |a_j|^2 + 2b_0^{-1}\sum_{j=2}^{\lfloor \log p \rfloor}\sum_{k=1}^{j-1} a_j a_k b_{j-k} \\
&\leq \sum_{j=1}^{\lfloor \log p \rfloor} |a_j|^2 + b_0^{-1}\sum_{j=2}^{\lfloor \log p \rfloor}\sum_{u=1}^{j-1}(|a_j|^2 + |a_{j-u}|^2) b_u.
\end{aligned}$$

For any $a \in \mathcal{S}_2$, we have $\sum_{j=1}^{\lfloor \log p \rfloor} |a_j|^2 \leq 1$ and, a fortiori,

$$\sum_{j=2}^{\lfloor \log p \rfloor}\sum_{u=1}^{j-1} |a_j|^2 b_u = \sum_{j=2}^{\lfloor \log p \rfloor} |a_j|^2 \sum_{u=1}^{j-1} b_u \leq \|b\|_{l_1}$$

and

$$\sum_{j=2}^{\lfloor \log p \rfloor}\sum_{u=1}^{j-1} |a_{j-u}|^2 b_u = \sum_{u=1}^{\lfloor \log p \rfloor - 1} b_u \sum_{j=u+1}^{\lfloor \log p \rfloor} |a_{j-u}|^2 \leq \|b\|_{l_1}.$$

Therefore,

$$\sup_{a\in\mathcal{S}_2}\left(\sum_{j\in G_1}\sum_{k\in G_1} a_j a_k v_{n,j}^{-1} v_{n,k}^{-1} \sum_{i=1}^n x_{i,j}x_{i,k}\right) \leq (1 + 2b_0^{-1}\|b\|_{l_1}) = C_p.$$

This inequality can easily be extended to any set G_l. Thus, the matrix X satisfies Assumption (A1) with $C_p = 1 + 2b_0^{-1}\|b\|_{l_1}$.

[Proof of Theorem 4.1.] By definition of the penalized estimator (2.1), for any $\beta \in \mathbb{R}^p$, we have

$$\|X\hat{\beta} - X\beta^*\|_n^2 + 2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j|^2} - \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\hat{\beta}$$
$$\leq \|X\beta - X\beta^*\|_n^2 + 2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\beta_j|^2} - \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\beta.$$

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

$$\|X\hat{\beta} - X\beta^*\|_n^2 \leq 2\sum_{l=1}^{L}\left[\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\beta_j^*|^2} - \sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j|^2}\right] + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\left(\hat{\beta} - \beta^*\right). \tag{5.1}$$

Using the Hölder inequality, on the event $\Lambda_{n,p}$, we have

$$\begin{aligned}\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i x_i\left(\hat{\beta} - \beta^*\right) &= 2\sum_{l=1}^{L}\sum_{j\in G_l}V_j\left(\hat{\beta}_j - \beta_j^*\right)\\ &\leq 2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^{-2}|V_j|^2}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j - \beta_j^*|^2}\\ &\leq \sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j - \beta_j^*|^2}.\end{aligned} \tag{5.2}$$

It follows from (5.1), (5.2) and the definition of Θ that

$$\begin{aligned}&\|X\hat{\beta} - X\beta^*\|_n^2 + \sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j - \beta_j^*|^2}\\ &\leq 2\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j - \beta_j^*|^2} + 2\sum_{l=1}^{L}\left[\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\beta_j^*|^2} - \sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j|^2}\right]\\ &= 2\sum_{l\in\Theta}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j - \beta_j^*|^2} + 2\sum_{l\in\Theta}\left[\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\beta_j^*|^2} - \sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j|^2}\right].\end{aligned}$$

By the Minkowski inequality, for any $l \in \{1, ..., L\}$, we have

$$\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\beta_j^*|^2} - \sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j|^2} \leq \sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat{\beta}_j - \beta_j^*|^2}.$$

Therefore,

$$\|X\hat{\beta} - X\beta^*\|_n^2 + \sum_{l=1}^{L} \sqrt{\sum_{j\in G_l} |w_{n,j}|^2 |\hat{\beta}_j - \beta_j^*|^2} \leq 4 \sum_{l\in\Theta} \sqrt{\sum_{j\in G_l} |w_{n,j}|^2 |\hat{\beta}_j - \beta_j^*|^2}$$
$$\leq 4\sqrt{Card(\Theta)} \sqrt{\sum_{l\in\Theta} \sum_{j\in G_l} |w_{n,j}|^2 |\hat{\beta}_j - \beta_j^*|^2}. \tag{5.3}$$

Now, let us bound the term $\sum_{l\in\Theta} \sum_{j\in G_l} |w_{n,j}|^2 |\hat{\beta}_j - \beta_j^*|^2$. By a simple decomposition, we have

$$\|X\hat{\beta} - X\beta^*\|_n^2 =$$
$$\sum_{l\in\Theta} \sum_{j\in G_l} \left(n^{-1} \sum_{i=1}^{n} |x_{i,j}|^2 \right) |\hat{\beta}_j - \beta_j^*|^2 + n^{-1} \sum_{i=1}^{n} \left(\sum_{l\notin\Theta} \sum_{j\in G_l} x_{i,j} (\hat{\beta}_j - \beta_j^*) \right)^2$$
$$+ R(\Theta), \tag{5.4}$$

where

$$R(\Theta) = 2 \sum_{l\in\Theta} \sum_{l'\notin\Theta} \sum_{j\in G_l} \sum_{j'\in G_{l'}} \left(n^{-1} \sum_{i=1}^{n} x_{i,j} x_{i,j'} \right) (\hat{\beta}_j - \beta_j^*)(\hat{\beta}_{j'} - \beta_{j'}^*)$$
$$+ \sum_{l\in\Theta} \sum_{\substack{l'\in\Theta \\ l'\neq l}} \sum_{j\in G_l} \sum_{j'\in G_{l'}} \left(n^{-1} \sum_{i=1}^{n} x_{i,j} x_{i,j'} \right) (\hat{\beta}_j - \beta_j^*)(\hat{\beta}_{j'} - \beta_{j'}^*)$$
$$+ \sum_{l\in\Theta} \sum_{j\in G_l} \sum_{\substack{j'\in G_l \\ j'\neq j}} \left(n^{-1} \sum_{i=1}^{n} x_{i,j} x_{i,j'} \right) (\hat{\beta}_j - \beta_j^*)(\hat{\beta}_{j'} - \beta_{j'}^*).$$

By noticing that $n^{-1} \sum_{i=1}^{n} \left(\sum_{l\notin\Theta} \sum_{j\in G_l} x_{i,j} (\hat{\beta}_j - \beta_j^*) \right)^2 \geq 0$ and

$$|R(\Theta)| \leq 2 \sum_{l\in\Theta} \sum_{l'=1}^{L} \sum_{j\in G_l} \sum_{\substack{j'\in G_{l'} \\ j'\neq j}} |n^{-1} \sum_{i=1}^{n} x_{i,j} x_{i,j'}| |\hat{\beta}_j - \beta_j^*| |\hat{\beta}_{j'} - \beta_{j'}^*|,$$

the equality (5.4) implies that:

$$\sum_{l\in\Theta} \sum_{j\in G_l} |w_{n,j}|^2 |\hat{\beta}_j - \beta_j^*|^2$$
$$\leq \lambda_{n,p}^2 n \Big(\|X\hat{\beta} - X\beta^*\|_n^2 - R(\Theta) \Big) \leq \lambda_{n,p}^2 n \Big(\|X\hat{\beta} - X\beta^*\|_n^2 + |R(\Theta)| \Big)$$
$$\leq \lambda_{n,p}^2 n \Big(\|X\hat{\beta} - X\beta^*\|_n^2 +$$
$$2 \sum_{l\in\Theta} \sum_{l'=1}^{L} \sum_{j\in G_l} \sum_{\substack{j'\in G_{l'} \\ j'\neq j}} |n^{-1} \sum_{i=1}^{n} x_{i,j} x_{i,j'}| |\hat{\beta}_j - \beta_j^*| |\hat{\beta}_{j'} - \beta_{j'}^*| \Big). \tag{5.5}$$

Let us set $\Pi_{j,j'} = n^{-1}\sum_{i=1}^n |w_{n,j}|^{-1}|w_{n,j'}|^{-1}x_{i,j}x_{i,j'}$. The Cauchy-Schwarz inequality yields

$$
\begin{aligned}
&\sum_{l\in\Theta}\sum_{l'=1}^{L}\sum_{j\in G_l}\sum_{\substack{j'\in G_{l'}\\ j'\neq j}} |n^{-1}\sum_{i=1}^n x_{i,j}x_{i,j'}||\hat\beta_j-\beta_j^*||\hat\beta_{j'}-\beta_{j'}^*| \\
&=\sum_{l\in\Theta}\sum_{l'=1}^{L}\sum_{j\in G_l}\sum_{\substack{j'\in G_{l'}\\ j'\neq j}} |\Pi_{j,j'}||w_{n,j}||w_{n,j'}||\hat\beta_j-\beta_j||\hat\beta_{j'}-\beta_{j'}^*| \\
&\leq\sum_{l\in\Theta}\sum_{l'=1}^{L}\sqrt{\sum_{j\in G_l}\sum_{\substack{j'\in G_{l'}\\ j'\neq j}}|\Pi_{j,j'}|^2}\sqrt{\sum_{j\in G_l}\sum_{j'\in G_{l'}}|w_{n,j}|^2|w_{n,j'}|^2|\hat\beta_j-\beta_j|^2|\hat\beta_{j'}-\beta_{j'}^*|^2} \\
&\leq\sup_{l\in\Theta}\sup_{l'=1,...,L}\sqrt{\sum_{j\in G_l}\sum_{\substack{j'\in G_{l'}\\ j'\neq j}}|\Pi_{j,j'}|^2}\left(\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat\beta_j-\beta_j^*|^2}\right)^2 \\
&=B(\Theta).
\end{aligned}
$$

Combining (5.3), (5.5), the previous inequality and using an elementary inequality of convexity, we obtain

$$
\begin{aligned}
&\|X\hat\beta-X\beta^*\|_n^2+\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat\beta_j-\beta_j^*|^2} \\
&\leq 4n^{1/2}\lambda_{n,p}\sqrt{Card(\Theta)}\sqrt{\|X\hat\beta-X\beta^*\|_n^2+2B(\Theta)} \\
&\leq 4n^{1/2}\lambda_{n,p}\sqrt{Card(\Theta)}\sqrt{\|X\hat\beta-X\beta^*\|_n^2}+4\sqrt{2}n^{1/2}\lambda_{n,p}\sqrt{Card(\Theta)B(\Theta)}.
\end{aligned}
\tag{5.6}
$$

Assumption (A2)($\mathcal{B}$), with $\mathcal{B}$ such that $\Theta\subseteq\mathcal{B}\subseteq\{1,...,L\}$, yields

$$
4\sqrt{2}n^{1/2}\lambda_{n,p}\sqrt{Card(\Theta)B(\Theta)}\leq\sum_{l=1}^{L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^2|\hat\beta_j-\beta_j^*|^2}. \tag{5.7}
$$

It follows from (5.6) and (5.7) that

$$
\|X\hat\beta-X\beta^*\|_n^2\leq 4n^{1/2}\lambda_{n,p}\sqrt{Card(\Theta)}\|X\hat\beta-X\beta^*\|_n.
$$

Therefore,

$$
\|X\hat\beta-X\beta^*\|_n^2\leq 16n\lambda_{n,p}^2 Card(\Theta).
$$

This ends the proof of Theorem 4.1.

[Proof of Proposition 4.1.] We split this proof into two parts. The first part considers the case where $\hat{\beta} = \hat{\beta}^G$. The second part focuses on the case where $\hat{\beta} = \hat{\beta}^L$.

- *The case where* $\hat{\beta} = \hat{\beta}^G$. According to Theorem 4.1, it suffices to prove that

$$\mathbb{P}\left(\max_{l=1,\ldots,L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\right) \leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_p)}.$$

We have

$$\begin{aligned}\mathbb{P}\left(\max_{l=1,\ldots,L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\right) &\leq \sum_{l=1}^{L}\mathbb{P}\left(\sqrt{\sum_{j\in G_l}|w_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\right) \\ &\leq (p/\lfloor\log p\rfloor)\max_{l=1,\ldots,L}\mathbb{P}\left(\sqrt{\sum_{j\in G_l}|v_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\kappa\sigma n^{-1}\sqrt{\log n + \log p}\right).\end{aligned} \tag{5.8}$$

In order to bound this last term, we introduce the Borell inequality. For further details about this inequality, see, for instance, Adler [1].

Lemma 5.1 (The Borell inequality) *Let $\mathcal{D}$ be a subset of $\mathbb{R}$ and $(\eta_t)_{t\in\mathcal{D}}$ be a centered Gaussian process. Suppose that*

$$\mathbb{E}\left(\sup_{t\in\mathcal{D}}\eta_t\right) \leq N \quad \textit{and} \quad \sup_{t\in\mathcal{D}}Var(\eta_t) \leq Q.$$

Then, for any $x > 0$, we have

$$\mathbb{P}\left(\sup_{t\in\mathcal{D}}\eta_t \geq x + N\right) \leq \exp(-x^2/(2Q)). \tag{5.9}$$

Let us consider the set $\mathcal{S}_2$ defined by $\mathcal{S}_2 = \{a = (a_j) \in \mathbb{Z}^*;\ \sum_{j\in G_l}|a_j|^2 \leq 1\}$, and the centered Gaussian process $\mathcal{Z}(a)$ defined by

$$\mathcal{Z}(a) = \sum_{j\in G_l} a_j V_j v_{n,j}^{-1}.$$

By an argument of duality, we have

$$\sup_{a\in\mathcal{S}_2}\mathcal{Z}(a) = \sup_{a\in\mathcal{S}_2}\sum_{j\in G_l} a_j v_{n,j}^{-1}V_j = \sqrt{\sum_{j\in G_l}|v_{n,j}|^{-2}|V_j|^2}.$$

Let us investigate the upper bounds for $\mathbb{E}(\sup_{a\in\mathcal{S}_2}\mathcal{Z}(a))$ and $\sup_{a\in\mathcal{S}_2}Var(\mathcal{Z}(a))$, in turn.

The upper bound for $\mathbb{E}(\sup_{a\in\mathcal{S}_2}\mathcal{Z}(a))$. Since $V_j \sim \mathcal{N}(0, \sigma^2 n^{-2}\sum_{i=1}^n |x_{i,j}|^2)$, the Cauchy-Schwarz inequality yields

$$\begin{aligned}\mathbb{E}(\sup_{a\in\mathcal{S}_2}\mathcal{Z}(a)) &= \mathbb{E}\left(\sqrt{\sum_{j\in G_l}|v_{n,j}|^{-2}|V_j|^2}\right) \leq \sqrt{\sum_{j\in G_l}|v_{n,j}|^{-2}\mathbb{E}(|V_j|^2)} \\ &= \sigma n^{-1}\sqrt{\sum_{j\in G_l}\sum_{i=1}^n |v_{n,j}|^{-2}|x_{i,j}|^2} = \sigma n^{-1}\sqrt{\log p}.\end{aligned}$$

So, $N = \sigma n^{-1}\sqrt{\log p}$.

The upper bound for $\sup_{a\in\mathcal{S}_2} Var(\mathcal{Z}(a))$. We have

$$Var_f^n(\mathcal{Z}(a)) = \sum_{j\in G_l}\sum_{k\in G_l} a_j a_k v_{n,j}^{-1} v_{n,k}^{-1}\mathbb{E}(V_j V_k),$$

with $\mathbb{E}(V_j V_k) = n^{-2}\sum_{u=1}^n\sum_{v=1}^n x_{u,j}x_{v,k}\mathbb{E}(\epsilon_u\epsilon_v) = \sigma^2 n^{-2}\sum_{u=1}^n x_{u,j}x_{u,k}$. Using this with Assumption (A1), we obtain:

$$\sup_{a\in\mathcal{S}_2} Var(\mathcal{Z}(a)) = \sigma^2 n^{-2}\sup_{a\in\mathcal{S}_2}\left(\sum_{j\in G_l}\sum_{k\in G_l} a_j a_k v_{n,j}^{-1}v_{n,k}^{-1}\sum_{u=1}^n x_{u,j}x_{u,k}\right) \leq C_p\sigma^2 n^{-2}.$$

So, $Q = C_p\sigma^2 n^{-2}$.

Combining the obtained values of N and Q with Lemma 5.1, for any $l \in \{1, ..., L\}$, we have

$$\begin{aligned}&\mathbb{P}\left(\sqrt{\sum_{j\in G_l}|v_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\kappa\sigma n^{-1}\sqrt{\log n + \log p}\right) \\ &\leq \mathbb{P}\left(\sqrt{\sum_{j\in G_l}|v_{n,j}|^{-2}|V_j|^2} \geq (2^{-1}\kappa - 1)\sigma n^{-1}\sqrt{\log n + \log p} + \sigma n^{-1}\sqrt{\log p}\right) \\ &= \mathbb{P}\left(\sup_{t\in\mathcal{D}}\eta_t \geq (2^{-1}\kappa - 1)\sigma n^{-1}\sqrt{\log n + \log p} + N\right) \\ &\leq \exp\left(-(2^{-1}\kappa - 1)^2\sigma^2 n^{-2}\log(np)/(2Q)\right) = (np)^{-(2^{-1}\kappa-1)^2/(2C_p)}. \quad (5.10)\end{aligned}$$

Putting (5.8) and (5.10) together, we obtain

$$\mathbb{P}\left(\max_{l=1,...,L}\sqrt{\sum_{j\in G_l}|w_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\right) \leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_p)} = u_{n,p}.$$

This ends the proof of Proposition 4.1 when $\hat{\beta} = \hat{\beta}^G$.

- *The case where* $\hat{\beta} = \hat{\beta}^L$. According to Theorem 4.1, it suffices to prove that

$$\mathbb{P}\left(\max_{l=1,\dots,L} \sqrt{\sum_{j\in G_l} |w_{n,j}|^{-2}|V_j|^2} \geq 2^{-1}\right) \leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_p)},$$

i.e., due to the definition of L and G_l in the Lasso definition,

$$\mathbb{P}\left(\max_{l=1,\dots,p} |w_{n,l}|^{-1}|V_l| \geq 2^{-1}\right) \leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_p)}.$$

Since $V_j = n^{-1}\sum_{i=1}^n x_{i,j}\varepsilon_i \sim \mathcal{N}(0, \sigma^2 n^{-2}\sum_{i=1}^n |x_{i,j}|^2)$, an elementary Gaussian inequality gives

$$\begin{aligned}
\mathbb{P}\left(\max_{l=1,\dots,p} |w_{n,l}|^{-1}|V_l| \geq 2^{-1}\right) &\leq p \max_{l=1,\dots,p} \mathbb{P}\left(|w_{n,l}|^{-1}|V_l| \geq 2^{-1}\right) \\
&\leq p\exp\left(-n^2\kappa^2\lambda_{n,p}^2/(8\sigma^2)\right) \\
&= p(np)^{-\kappa^2/8} \\
&\leq p(np)^{-(2^{-1}\kappa-1)^2/(2C_p)} = u_{n,p}.
\end{aligned}$$

This ends the proof of Proposition 4.1 when $\hat{\beta} = \hat{\beta}^L$.

Acknowledgement. We would like to thank Professor Alexander Tsybakov and Professor Nicolas Vayatis for insightful comments.

References

[1] Adler, R. J. (1990). An introduction to continuity, extrema, and related topics for general gaussian processes. *Institute of Mathematical Statistics*, Hayward, CA.

[2] Bunea, F., Tsybakov, A., and Marten, H. (2007). Aggregation for gaussian regression. *Technical Report*.

[3] Bunea, F., Tsybakov, A., and Wegkamp, M. (2006). *Aggregation and Sparsity via l_1 Penalized Least Squares*, vol. pp. 379 - 391. Springer, New York, proceedings of the annual conference on learning theory, lecture notes in artificial intelligence (colt 2006) ed.

[4] Cavalier, L. and Tsybakov, A. (2001). Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Mathematical Methods of Statistics*, 10(3):247–282.

[5] Donoho, D. L., Elad, M., and Temlyakov, V. N. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 52(1):6–18.

[6] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499. With discussion, and a rejoinder by the authors.

[7] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

[8] Kim, Y., Kim, J., and Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica*, 16:375–390.

[9] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

[10] Meier, L., van de Geer, S., and Bühlmann, P. (2007). The group lasso for logistic regression. Technical report, Department of Statistics, Seminar For Statistics, ETH, Zurich.

[11] Meinshausen, N. (2005). Lasso with relaxation. *Technical Report*.

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

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

[14] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67(1):91–108.

[15] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 68(1):49–67.

[16] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. URL `citeseer.ist.psu.edu/zhao06model.html`.

[17] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

[18] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 67(2):301–320.