Generalization of l1 constraints for high dimensional regression problems

Pierre Alquier, Mohamed Hebiri

Generalization of ℓ_1 constraints for high dimensional regression problems

Pierre ALQUIER[(1)] and Mohamed HEBIRI[(2)]

(1, 2) LPMA, CNRS-UMR 7599,
Université Paris 7 - Diderot, UFR de Mathématiques,
175 rue de Chevaleret F-75013 Paris, France.

(1) CREST-LS,
3, avenue Pierre Larousse
92240 Malakoff, France.

Abstract

We consider the linear regression problem where the number p of covariates is possibly larger than the number n of observations. In the paper, we propose to approximate the unknown regression parameters under sparsity assumptions with a class of estimators that are motivated by geometrical considerations. Popular estimators based on the control of the ℓ_1 norm of the regression coefficients (such as the LASSO and the Dantzig selector for example) can be seen as special cases of our estimator for which we derive Sparsity Inequalities, i.e., bounds involving the sparsity of the parameter we try to estimate. In such a generalized setup, we show that it is possible to consider variations of the loss function to be minimized. In particular, under a suitable setting, we derive a new estimator that is a transductive version of the LASSO, and we analyze its performance with milder assumptions than in the well-known results about the "usual" LASSO.

Keywords: High-dimensional data, LASSO, Mutual coherence, Sparsity, Variable selection.

AMS 2000 subject classifications: Primary 62J05, 62J07; Secondary 62F25.

1 Introduction

In many modern applications, one has to deal with very large datasets. Regression problems may involve a large number of covariates, possibly larger than the sample size. In this situation, a major issue lies in dimension reduction which can be performed through the selection of a small amount of relevant covariates. For this purpose, numerous regression methods have been proposed in the literature, ranging from the classical information criteria such as C_p, AIC and BIC to the more recent regularization-based techniques such as the l_1 penalized least square estimator, known as the LASSO [Tib96], and the Dantzig selector [CT07] among many others. Regularized regression methods have recently witnessed several developments due to the attractive feature of computational feasibility, even for high dimensional data when the number of covariates p is large. In the present paper, we focus on the LASSO and the Dantzig selector with what they have in common: both obey to a geometric constraint which we now introduce. Consider the linear regression model $Y = X\beta^* + \varepsilon$, where Y is a vector in $\mathbb{R}^n$, $\beta^* \in \mathbb{R}^p$ is the parameter vector, X is an $n \times p$ real-valued matrix with possibly much fewer rows than columns, $n \ll p$, and ε is a random noise vector in $\mathbb{R}^n$. The analysis of regularized regression methods for high dimensional data usually involves a sparsity assumption on β^* through the *sparsity index* $\|\beta^*\|_0 = \sum_{j=1,\dots,p} \mathbb{I}(\beta^*_j \neq 0)$ where $\mathbb{I}(\cdot)$ is the indicator function. For any $q \geq 1$, $d \geq 0$ and $a \in \mathbb{R}^d$, denote by $\|a\|_q^q = \sum_{i=1}^d |a_i|^q$ and $\|a\|_\infty = \max_{1\leq i\leq d} |a_i|$, the ℓ_q and the ℓ_∞ norms respectively. When the design matrix X is normalized, the LASSO and the Dantzig selector minimize respectively $\|X\beta\|_2^2$ and $\|\beta\|_1$ under the constraint $\|X'(Y - X\beta)\|_\infty \leq s$ where s is a positive tuning parameter (e.g. [OPT00, Alq08] for the dual form of the LASSO). This geometric constraint is central in the approach developed in the present paper and we shall use it in a general perspective. In the sequel, we consider three specific problems in the high-dimensional setting (i.e., $p \geq n$):

Goal 1 - *Prediction:* The reconstruction of the signal $X\beta^*$ with the best possible accuracy is first considered. The quality of the reconstruction with an estimator $\hat{\beta}$ is often measured with the squared error $\|X\hat{\beta} - X\beta^*\|_2^2$. In the standard form, results are stated as follows: under assumptions on the matrix X and with high probability, the prediction error is bounded by $C \log(p) \|\beta^*\|_0$ where C is a positive constant. Such results for the prediction issue have been obtained in [BRT07, Bun07, BTW07a, BTW07b] for the LASSO and in [BRT07] for the Dantzig selector. We also refer to [Kol07, Kol08, MVdGB08, vdG08, DT07, CH08] for related works with different estimators (non-quadratic loss, penalties slightly different from l_1 and/or random design). The results obtained in the works above-mentioned are optimal up to a logarithmic factor as it has been proved in [BTW07a].

Goal 2 - *Estimation:* Another wishful thinking is that the estimator $\hat{\beta}$ is close to β^* in terms of the ℓ_q distance for $q \geq 1$. The estimation bound is of the form $C \|\beta^*\|_0 (\log(p)/n)^{q/2}$ where C is a positive constant. Such results are stated for the LASSO in [BTW07a, BTW07b] when $q = 1$, for the Dantzig selector in [CT07] when $q = 2$ and have been generalized in [BRT07] with $1 \leq q \leq 2$ for both the LASSO end the Dantzig selector.

Goal 3 - *Selection:* Since we consider variable selection methods, the identification of the true support $\{j : \beta_j^* \neq 0\}$ of the vector β^* is to be considered. One expects that the estimator $\hat{\beta}$ and the true vector β^* share the same support at least when n grows to infinity. This is known as the variable selection consistency problem and it has been considered for the LASSO estimator in several works [Bun07, MB06, MY09, Wai06, ZY06]. Recently, [Lou08] provided the variable selection consistency of the Dantzig selector. Other popular selection procedures, based on the LASSO estimator, such as the Adaptive LASSO [Zou06], the SCAD [FL01], the S-LASSO [Heb08] and the Group-LASSO [Bac08], have also been studied under this angle.

In the present paper, we address these three goals under a different sparsity assumption than the usual one. Namely, we relate the notion of sparsity to the sparsity index of the vector $P\beta^*$, for some matrix $P \in \mathbb{R}^{p \times p}$. Therefore, sparsity implies here that many components $(P\beta^*)_j$ are equal to 0. Naturally, when P equals I_p, the $p \times p$ identity matrix, we recover the standard assumption on the sparsity index of β^*. We consider a general family of estimators which are defined as solutions of different optimization functions but with the same set of constraint $\|X'(Y - X\beta)\|_\infty \leq s$ (when X is normalized). This family includes the LASSO and the Dantzig selector as special cases. We respond to the three goals described above but with some modifications. Concerning **Goal 1**, instead of the prediction of $X\beta^*$, we aim at recovering $Z\beta^*$ for some matrix $Z \in \mathbb{R}^{m \times p}$ with $m \in \mathbb{N}$. This matrix can be taken equal to X. However, different choices of matrices Z can be considered in such a way to cover other fields such as the transductive setting (Section 4.2). As far as estimation (**Goal 2**) and selection (**Goal 3**) are concerned, the whole study takes into consideration the sparse vector $P\beta^*$ instead of β^*. By exploiting the sparsity of $P\beta^*$, we provide similar results to those presented in the conventional case. However, we need assumptions which are less restrictive in some situations. We also show, in the high-dimensional case $p \gg n$, that it is possible to derive consistent results in situations where β^* is not sparse (in the usual sense).

The paper is organized as follows. In the next section, we specify the setting

and the estimators considered in this paper. A short description of the results stated in the sequel are also provided. In Section 3, we state our main results. More precisely, we present in Section 3.1, different assumptions used through the paper and compare them to the assumptions used in previous works. Using techniques from [BTW07b], we study the performance of the estimators in the different contexts (**Goal 1** to **Goal 3**). Applications of these results are then considered in Section 4: the transductive LASSO and the correlation selector [Alq08]. Finally Section 6 is dedicated to the proofs.

2 Model and estimator

In this section, we present the general setting. We first introduce the model and the estimators which are considered in the sequel. We also briefly present the results obtained in this paper with some technical arguments which should help the reader to better understand the progression of the paper.

Model. We focus on the usual linear regression model:

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

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 vector of interest and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. centered Gaussian random variables with known variance σ^2. Let X denote the matrix where the i-th line is x_i and the j-th column is X_j with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, p\}$. Then:

$$X = (x_1', \dots, x_n')' = (X_1, \dots, X_p).$$

For the sake of simplicity, it is often assumed that the observations are normalized in such a way that $X_j'X_j/n = 1$. In this paper, we will not make such an assumption, but we will discuss the consequences of such a normalization in the various results. For this purpose, let us introduce to following notation. For any $j \in \{1, \dots, p\}$,

$$\xi_j = \frac{X_j'X_j}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}^2 \quad \text{and} \qquad \Xi = \begin{pmatrix} \xi_1^{1/2} & & 0 \\ & \ddots & \\ 0 & & \xi_p^{1/2} \end{pmatrix}.$$

Let us also put $Y = (y_1, \dots, y_n)'$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$, so we have the following matricial form for Model (1): $Y = X\beta^* + \varepsilon$. As mentioned in Section 1, we assume that $P\beta^*$ is sparse and base our approach in exploiting this sparsity of the model through the *sparsity index* $\|P\beta^*\|_0 = \sum_{j=1\dots,p} \mathbb{I}((P\beta^*)_j \neq 0)$.

Estimator. Considering the case where $p \gg n$, dimension reduction is fundamental. It aims at producing consistent estimators while being easy to interpret. The estimators defined in this paper share the same constraint $\|\Xi^{-1}X'(Y - X\beta)\|_\infty \leq s$ where $s > 0$ is a tuning parameter to be specified later. Let us call this constraint the *Dantzig Constraint* ($DC(s)$), as it appears in the definition of the Dantzig selector [CT07] when $\xi_j = 1$. This constraint consists in a threshold on the correlation between a covariate X_j, $j \in \{1, \ldots, p\}$ and the residual $Y - X\beta$. Let us define the set:

$$DC(s) = \{\beta \in \mathbb{R}^p : \left\|\Xi^{-1}X'(Y - X\beta)\right\|_\infty \leq s\}, \tag{2}$$

where $s > 0$ is a tuning parameter depending on n and p to be specified later. This set is also interpreted as a confidence region for β^* [Alq08] and produce a geometrical motivation for the study of the following estimators:

$$\textbf{Program I:} \qquad \hat{\beta} = \underset{\beta \in DC(s)}{\operatorname{Argmin}} \|Z\beta\|_2^2, \tag{3}$$

$$\textbf{Program II:} \qquad \tilde{\beta} = \underset{\beta \in DC(s)}{\operatorname{Argmin}} \|\Xi P\beta\|_1, \tag{4}$$

where $Z \in \mathbb{R}^{m \times p}$ with $m \in \mathbb{N}$ and $DC(s)$ is the Dantzig Constraint given by (2). For unicity reasons, Let us assume from now on the following condition:

– *Kernel condition. The matrix Z is such that* $\ker Z = \ker X$.

The connection between the estimators $\hat{\beta}$ and $\tilde{\beta}$ defined respectively in (3) and (4) is made in the following way. The *Kernel condition* implies[1] that there exists an invertible matrix $P \in \mathbb{R}^{p \times p}$ such that

$$(X'X)P = (Z'Z). \tag{5}$$

When this matrix P coincides with the matrix P used in the definition of $\tilde{\beta}$ in (4), **Program I** and **Program II** will produce estimators that have, in theory and in practice, about the same performances. In this paper, both of the solutions $\hat{\beta}$ and $\tilde{\beta}$ are used to predict $Z\beta^*$ whenever $P\beta^*$ is sparse.

Sketch of main results and technical tools. Most of the results which are stated here rely on the exploitation the sparsity index $\|P\beta^*\|_0$, under assumptions on a symmetric matrix W, defined[2] such that

$$(Z'Z)W(X'X) = (X'X). \tag{6}$$

[1] See the section dedicated to proofs, more precisely Section 6.1 page 14 for a proof.
[2] Here again, the existence of such a W is given in Section 6.1.

In the sequel, in the high dimensional case (when $p \gg n$) and under the sparsity assumption $\|P\beta^*\|_0 \ll p$, we respond mainly to three objectives described in Section 1. Here is a sketch of the main results:
Goal 1 - *Prediction:* To ensure to reconstruction of $Z\beta^*$, we prove that, with high probability, $\|Z\beta - Z\beta^*\|_2^2 \leq C \log(p) \|P\beta^*\|_0$ where C is a positive constant and β is either $\hat{\beta}$ or $\tilde{\beta}$ defined in (3) and (4) respectively.
Goal 2 - *Estimation:* We hope that the solution β (where β is either $\hat{\beta}$ or $\tilde{\beta}$ defined in (3) and (4) respectively) is such that $P\beta$ is close to $P\beta^*$. We state that with high probability $\|P\beta - P\beta^*\|_1 \leq C\sqrt{\log(p)/n} \|P\beta^*\|_0$ where C is a positive constant.
Goal 3 - *Selection:* Variable selection consistency seems less interesting in our study as soon as $P \neq I_p$. However, we set that with high probability $\|P\beta - P\beta^*\|_\infty \leq C\sqrt{\log(p)/n}$ where C is a positive constant. One then can easily provide variable selection consistency results using such an inequality.
The results stated in the present paper can be interpret as *Sparsity Inequalities* (SIs), bounds which depend on the oracle vector β^* through the sparsity index $\|P\beta^*\|_0$. Furthermore, let us mention that one technical argument to provide our result is based on a dual form of **Program I**: given X, Z and β^*, the relation (5) permits us to introduce Γ, defined as

$$\Gamma = \{\gamma \in \mathbb{R}^p : \ X'X\gamma = Z'Z\beta^*\}.$$

This set consists of all vectors γ which belong to the space in which the transformation of β^* is sparse. We can define the sparsest vector in Γ by

$$\gamma^* \in \underset{\gamma \in \Gamma}{\text{Argmax}} \, \text{Card} \, \{j \in \{1, ..., p\}, \quad \gamma_j = 0\} = \underset{\gamma \in \Gamma}{\text{Argmax}} \, \|\gamma\|_0 \, . \tag{7}$$

As the matrix P is invertible, denote $\beta^{**} \in \mathbb{R}^p$ the vector such that $\beta^{**} = P^{-1}\gamma^*$ and consequently, we have $\|\gamma^*\|_0 = \|P\beta^{**}\|_0$, the sparsity index. Because of the Kernel condition, we have[3] $Z\beta^* = Z\beta^{**}$. Then estimating β^{**} instead of β^* does not affect the prediction objective **Goal 1**. From now on, for the sake of simplicity, let P_j denote the j-th row of P, so we can write, for any $j \in \{1, ..., p\}$: $\gamma_j^* = P_j\beta^{**}$. One of the main points in the proofs is to link the study of **Program I**, with the study of

$$\textbf{Program I-Dual} \qquad \hat{\gamma} = \underset{\gamma \in \mathbb{R}^p}{\text{Argmin}} \left\{ \|Y - X\gamma\|_2^2 + 2s \|\Xi\gamma\|_1 + \gamma' M \gamma \right\}, \tag{8}$$

for some matrix $M \in \mathbb{R}^{p \times p}$ related to Z. An explicit form of M will be provided in Section 3.2. Here again, note the form of the program when $\xi_j = 1$ for $j = 1, \ldots, p$:

$$\underset{\gamma \in \mathbb{R}^p}{\text{Argmin}} \left\{ \|Y - X\gamma\|_2^2 + 2s \|\gamma\|_1 + \gamma' M \gamma \right\}.$$

[3] See Section 6.1.

In this paper we present two applications based on the sparsity induced by the transformation $P\beta$. We consider in Section 4.1 the Correlation Selector introduced in [Alq08]. We also consider the Transductive LASSO. In such a case, one choice for Z may be the unlabeled dataset (More details are given in Section 4.2).

3 Main results

In this section we state all the theoretical results according to the solutions of **Program I** and **Program II**. We start with presenting different assumptions used through the paper in Section 3.1. This is the occasion to compare our hypotheses with the ones already used in the previous works mentioned above. Then, we study the performance of the estimators $\hat{\beta}$ and $\tilde{\beta}$ defined by **Programs I** and **II** respectively. We also relate the solutions of **Programs I** to those of **Program I-Dual** thanks to a link between Z and M (Section 3.2).

3.1 Assumptions

We present here the assumptions we need to state the Sparsity Inequalities provided in Sections 3.3 and 3.4. Note that they essentially involve the matrix Ω defined as follows:

$$\Omega = \frac{1}{n}(X'X)W(X'X). \tag{9}$$

We denote by $\Omega_{j,k}$, the (j,k) coefficient of Ω. We just remind that the definition of Ω involves the matrix W, given by (6). Using this notation we introduce the assumptions with more precision. The first assumption is used to respond to the prediction **Goal 1** and estimation **Goal 2** objectives.

– *Assumption (A1). There is a constant $c > 0$ such that, for any $\alpha \in \mathbb{R}^p$ such that*

$$\sum_{j:\gamma_j^*=0} \xi_j^{\frac{1}{2}} |\alpha_j| \leq 3 \sum_{j:\gamma_j^*\neq 0} \xi_j^{\frac{1}{2}} |\alpha_j| ,$$

where γ^* is given by (7), we have

$$\sum_{j:\gamma_j^*\neq 0} \alpha_j^2 \leq c\alpha'\Omega\alpha. \tag{10}$$

When we deal with variable selection **Goal 3**, we replace Assumption (A1) by the following:

– *Assumption (A2). Let us assume that , for any* $j \in \{1, ..., p\}$, $\xi_j = 1$ *and that*

$$\rho = \sup_{j\in\{1,...,p\}} \sup_{k\neq j} |\Omega_{j,k}| \leq \frac{\inf_{\gamma_j^* \neq 0} \Omega_{j,j}}{14\|\gamma^*\|_0}. \tag{11}$$

We now give some comments about the assumptions. First, note that both of these assumptions are modifications of the well-known *mutual coherence* condition introduced in [DET06] - but the mutual coherence condition is about the Gram matrix $n^{-1}X'X$ while the assumptions presented here involve the matrix Ω. Assumption (A2) is closer to the mutual coherence condition than Assumption (A1). It is also more restrictive. For a selection purpose **Goal 3**, the mutual coherence assumption is used in [Lou08]. Moreover Assumption (A1) can be seen as a modification of more general assumptions that can be found in [BRT07, Bun07, BTW07a, BTW07b]. For example, using a slight modification of a proof given in [BRT07], we can prove the following result.

Lemma 1. *Assumption (A2)* $\Rightarrow$ *Assumption (A1) with constant* $c = \frac{2}{\inf_{\gamma_j^* \neq 0} \Omega_{j,j}}$.

For the sake of completeness, the proof is given in Section 6 in its full length. It is known that in the high dimensional case ($p \gg n$), such assumptions are hard to relax when the considered estimators are solution of a convex minimization problem as in (3) and (4); see [BRT07] and [BTW07b, Remarks 4 and 5] for other comments on that topic. In the case where $n \geq p$, if $Z'Z$ and $X'X$ are invertible matrices, then we can find a constant c such that, for any $\alpha \in \mathbb{R}^p$, $\alpha'\alpha \leq c\alpha'\Omega\alpha$. Of course, this implies that Assumption (A1) is satisfied with this specific choice of the constant c.

3.2 Dual form of Program I

Let us put (remember that W is given by (6)):

$$M = (X'X)W(X'X) - (X'X). \tag{12}$$

Theorem 1. *All the solutions* $\hat{\beta}$ *of* **Program I** *are given by* $(Z'Z)\hat{\beta} = (X'X)\hat{\gamma}$ *where* $\hat{\gamma}$ *is any solution of* **Program I-Dual**, *with* M *given by* (12). *Moreover, when* $\hat{\gamma}$ *is unique, all the solutions of* **Program I** *give the same value to* $X\hat{\beta}$, *and also to* $Z\hat{\beta}$.

The proof is given in Section 6. Note that, taking $Z = X$ gives $M = 0$ and allows the choice $P = I_p$. So, $\hat{\gamma}$ is a solution the the LASSO program and we can take $\hat{\beta} = \hat{\gamma}$. Theorem 1 can be seen as a generalization of the dual form of the LASSO given in [Alq08, OPT00].

3.3 Sparse inequalities and sup-norm bound for Program I

In this section, we provide sparse inequalities (SIs) and a sup-norm bound for **Program I**. First Theorem 2 provides bounds on the squared error (corresponding to **Goal 1**) and to the distance between the estimated and true parameters (corresponding to **Goal 2**). The main key is to use the sparsity index $\|\gamma^*\|_0 = \|P\beta^{**}\|_0$ where γ^* is given by (7).

Theorem 2. *Let us consider the linear regression model* (1). *Let $\hat{\gamma}$ be any solution of the the quadratic* **Program I-Dual**. *Let $\hat{\beta} = P^{-1}\hat{\gamma}$, so by Theorem 1, $\hat{\beta}$ is a solution of* **Program I**. *Let us choose $\kappa > 2\sqrt{2}$ and $s = \kappa\sigma\sqrt{n\log(p)}$. Under Assumption (A1), with probability larger than $1 - p^{1-\frac{\kappa^2}{8}}$, we have*

$$\left\|Z(\hat{\beta} - \beta^{**})\right\|_2^2 = \left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 \leq 16c\kappa^2\sigma^2\log(p)\sum_{P_j\beta^{**}\neq 0}\xi_j, \tag{13}$$

and

$$\|\Xi\left(\hat{\gamma} - \gamma^*\right)\|_1 = \left\|\Xi P\left(\hat{\beta} - \beta^*\right)\right\|_1 \leq 16c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\sum_{P_j\beta^{**}\neq 0}\xi_j. \tag{14}$$

The proof of this result can be found in Section 6.

Corollary 3.1. *Under the conditions of Theorem 2, if we moreover assume that the matrix X is normalized in order to have $\xi_j = 1$ for any j, then we have, with probability larger than $1 - p^{1-\frac{\kappa^2}{8}}$,*

$$\left\|Z(\hat{\beta} - \beta^*)\right\|_2^2 \leq 16c\kappa^2\sigma^2\log(p)\left\|P\beta^{**}\right\|_0, \tag{15}$$

and

$$\left\|P\left(\hat{\beta} - \beta^{**}\right)\right\|_1 \leq 16c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\left\|P\beta^{**}\right\|_0. \tag{16}$$

Theorem 2 and its corollary state that with high probability, we can consistently perform prediction **Goal 1** and estimation **Goal 2**, exploiting the sparsity of the projected β^*, that is $\gamma^* = P\beta^{**}$. Note that the obtained rates are near optimal up to a logarithmic factor. Indeed, in our setting, it is proved in [BTW07a, Theorem 5.1] that the optimal rate for the l_2 risk (15) is $\log\left(\frac{p}{\|P\beta^{**}\|_0} + 1\right)\|P\beta^{**}\|_0$.

We provide now a bound on the sup-norm $\|\gamma^* - \hat{\gamma}\|_\infty$. As described in Remark 1, such a result would help us to easily get an estimator of γ^* which is consistent in variable selection **Goal 3**. That is, an estimator which succeed to recover the true support of γ^*, the sparse projection of β^*.

Theorem 3. *Let us consider the linear regression model* (1)*. Let* $\hat{\gamma}$ *be any solution of the quadratic* **Program I-Dual**. *Let us choose* $\kappa > 2\sqrt{2}$ *and* $s = \kappa\sigma\sqrt{n\log(p)}$. *Under Assumption (A2), with probability larger than* $1 - p^{1-\frac{\kappa^2}{8}}$, *we have simultaneously Inequalities* (15), (16) *and*

$$\left\|\hat{\gamma} - \gamma^*\right\|_\infty \leq \frac{3\kappa\sigma}{\inf_{1\leq j\leq p}\Omega_{j,j}}\sqrt{\frac{\log(p)}{n}}. \tag{17}$$

The proof of this result can be found in Section 6. Note also that these results generalize the results obtained in [Bun07, BTW07b, Lou08] as the LASSO can be seen as special cases of our estimator.

3.4 Sparsity Inequalities and sup-norm bound for Program II

For readability, let us recall **Program II**:

$$\begin{cases} \text{Argmin}_{\beta\in\mathbb{R}^p} \|\Xi P\beta\|_1 \\ \\ \text{s.t. } \left\|\Xi^{-1}X'\left(Y - X\beta\right)\right\|_\infty \leq s. \end{cases}$$

Here again we want to estimate $Z\beta^* = Z\beta^{**}$ when $P\beta^{**}$ is assumed to be sparse (as in Theorem 2). In such a context, analog results to those obtained in Section 3.3 can be obtained. First we state:

Theorem 4. *Let us assume that Assumption (A1) is satisfied. Let* $\tilde{\beta}$ *be a solution of* **Program II**. *Let us choose* $\kappa > 2\sqrt{2}$ *and* $s = \kappa\sigma\sqrt{n\log(p)}$. *Then with probability larger than* $1 - p^{1-\frac{\kappa^2}{8}}$, *we have*

$$\left\|Z(\tilde{\beta} - \beta^*)\right\|_2^2 \leq 9c\kappa^2\sigma^2\log(p)\sum_{P_j\beta^{**}\neq 0}\xi_j,$$

and

$$\left\|\Xi P\left(\tilde{\beta} - \beta^{**}\right)\right\|_1 \leq 6c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\sum_{P_j\beta^{**}\neq 0}\xi_j.$$

The proof is given in Section 6. In the same way as for the solution of **Program I**, we can provide an analog to Theorem 3.

Theorem 5. *With the notations of the previous theorem, under Assumption (A2), if we moreover assume that the matrix* X *is normalized in order to have* $\xi_j = 1$ *for any* j, *with probability greater than* $1 - p^{1-\frac{\kappa^2}{8}}$, *we have simultaneously*

$$\left\|Z(\tilde{\beta} - \beta^*)\right\|_2^2 \leq 9c\kappa^2\sigma^2\log(p)\left\|P\beta^{**}\right\|_0,$$

$$\left\| P\left(\tilde{\beta} - \beta^{**}\right)\right\|_1 \leq 6c\kappa\sigma\sqrt{\frac{\log(p)}{n}}\left\|P\beta^{**}\right\|_0,$$

and

$$\left\| P\left(\tilde{\beta} - \beta^{**}\right)\right\|_\infty \leq \frac{2\kappa\sigma}{\inf_{1\leq j\leq p}\Omega_{j,j}}\sqrt{\frac{\log(p)}{n}}.$$

Note that these results generalize the results obtained in [BRT07, Lou08] as the Dantzig selector can be seen as special cases of our estimator.

Remark 1. *Thanks to Theorems 3 and 5, we can easily construct a sign-consistent estimator (an estimator $\overline{\gamma}$ of the vector γ^* given by (7) such that it shares asymptotically and in probability, not only the same support (sparsity set) but also the same sign of its components with γ^*). This estimator $\overline{\gamma}$ is defined as a thresholded version of $\hat{\gamma}$ where $\hat{\gamma}$ is either solution of* **Program I** *or is equal to $P\tilde{\beta}$ with $\tilde{\beta}$ solution of* **Program II**. *The threshold used is respectively equal to the bound in the sup-norm result appearing in Theorem 3 and 5. Some more technical tools to establish the result are needed and we refer to [Bun07, Lou08] for more details.*

4 Applications

We present now two applications of the estimators considered in the previous sections.

4.1 The Correlation Selector

In [Alq08], an estimator is introduced for the case where most of the X_j's have a null correlation with Y while we think that all together, these covariates can provide a good prevision for Y: namely, we assume that the $(X'X)\beta^*$ is sparse.

Here (in this subsection only), let us assume that $(X'X)$ is invertible - this implies that $p \leq n$. Let us also assume that X is normalized, so $\xi_j = 1$ for any j. Then, if we take $Z = (X'X)$ then we can take $P = (X'X)$ too and $\Omega = I_p/n$, this means that Assumptions (A1) and (A2) are satisfied in any case. So, **Program I** involves the minimization of $\|(X'X)\beta\|_2^2$ while **Program II** involves the minimization of $\|(X'X)\beta\|_1$. Actually, it is proved in [Alq08] that the estimator defined by

$$\hat{\beta}_{CS} = \underset{\beta\in\mathbb{R}^p}{\text{Argmin}}\,\|(X'X)\beta\|_q^q \qquad \text{s.t. } \left\|X'\left(Y - X\beta\right)\right\|_\infty \leq s$$

for any $q \geq 1$ does not depend on q. Theorem 5 gives, with probability larger than $1 - n^{1-\frac{\kappa^2}{8}}$,

$$\left\|\frac{X'X}{n}(\hat{\beta}_{CS} - \beta^*)\right\|_2^2 \leq 9\kappa^2\sigma^2\frac{\log(p)}{n}\left\|(X'X)\beta^*\right\|_0, \tag{18}$$

$$\left\| \frac{X'X}{n} \left(\hat{\beta}_{CS} - \beta^* \right) \right\|_1 \leq 6\kappa\sigma \sqrt{\frac{\log(p)}{n}} \left\| (X'X)\beta^* \right\|_0 ,$$

and

$$\left\| \frac{X'X}{n} \left(\hat{\beta}_{CS} - \beta^* \right) \right\|_\infty \leq 2\kappa\sigma \sqrt{\frac{\log(p)}{n}}.$$

Note that Inequality (18) was already proved in [Alq08], but that the proof in [Alq08] could not be extended to the cases of the ℓ_1-norm and ℓ_∞-norm. However, the proof in [Alq08] allows to extend Inequality (18) to the case where $p \geq n$ without hypotheses; here, $\Omega = I/n$ would not be possible in this case and so we would have additional hypotheses. Finally, notice that Inequality (18) does not involve a natural norm. However, adding one more hypotheses, we obtain the following result.

Corollary 4.1. *Let us assume that there is a $\zeta > 0$ such that $\zeta(X'X)/n - I_p$ is definite positive. Then we have, with probability larger than $1 - n^{1-\frac{\kappa^2}{8}}$,*

$$\left\| X(\hat{\beta}_{CS} - \beta^*) \right\|_2^2 \leq 9\zeta\kappa^2\sigma^2 \log(p) \left\| (X'X)\beta^* \right\|_0 .$$

4.2 The transductive LASSO

By an application of Theorem 1, the equivalence between **Program I**, applied with and $Z = X$, and the LASSO, is clear. However, Theorem 2 allows to extend the LASSO to the so-called transductive setting introduced in [Vap98].

In this setting, we have $Y = \overline{X}\overline{\beta}^* + \varepsilon$ where $\overline{X}$ is a matrix containing the observation vectors $\overline{x}_i$ and $\overline{\beta}^*$ is "sparse". However, we are not interested in the estimation of $\overline{\beta}^*$ or $\overline{X}\overline{\beta}^*$: we have another set of m points - say $\overline{x}_{n+1}, \ldots, \overline{x}_{n+m}$. We choose

$$\overline{Z} = (\overline{x}'_{n+1}, \ldots, \overline{x}'_{n+m})',$$

satisfying $\ker \overline{Z} = \ker \overline{X}$ and so the objective is the estimation of $\overline{Z}\overline{\beta}^*$, that is the prediction of the value of the regression function on a particular set of points. We have also $\overline{P}$ such that $\overline{X}'\overline{X}\overline{P} = \overline{Z}'\overline{Z}$ and $\overline{W}$ such that $(\overline{Z}'\overline{Z})\overline{W}(\overline{X}'\overline{X}) = (\overline{X}'\overline{X})$.

It is argued in [Vap98] that in this setting, a bound on the general performances of an estimator of $\overline{\beta}^*$ is (often) useless and that the statistician should focus on a particular method to estimate $\overline{Z}\overline{\beta}^*$, that can be easier.

Note that a direct application of Theorem 2 (for example) would lead to an unsatisfying result as it would not assume that $\overline{\beta}^*$ is sparse, but $\overline{P}\overline{\beta}^*$. The problem can be solved in the following way. Note that

$$Y = \overline{X}\overline{\beta}^* + \varepsilon = (\overline{XP})(\overline{P}^{-1}\overline{\beta}^*) + \varepsilon = X\beta^* + \varepsilon$$

where we put $X = \overline{XP}$, $\beta^* = \overline{P}^{-1}\overline{\beta}^*$, and

$$\overline{Z\beta}^* = (\overline{ZP})(\overline{P}^{-1}\overline{\beta}^*) = Z\beta^*$$

where $Z = \overline{ZP}$. Note that the choice $W = \overline{P}^{-1}\overline{W}(\overline{P}^{-1})'$ satisfies $(Z'Z)W(X'X) = (X'X)$.

Note that ξ_j will still denote the j-th diagonal element of $X'X/n$. When we apply Theorem 2 to this setting, we have to make hypotheses about

$$\Omega = \frac{1}{n}(X'X)W(X'X) = \frac{1}{n}\overline{P}'(\overline{X}'\overline{X})\overline{P}\overline{P}^{-1}\overline{W}(\overline{P}^{-1})'\overline{P}'(\overline{X}'\overline{X})\overline{P}$$
$$= \frac{1}{n}\overline{P}'(\overline{X}'\overline{X})\overline{W}(\overline{X}'\overline{X})\overline{P} = \frac{1}{n}(\overline{Z}'\overline{Z}).$$

So we will not need any assumption about $\overline{X}$!

Corollary 4.2. *Let us assume that there is a constant $c > 0$ such that, for any $\alpha \in \mathbb{R}^p$ satisfying $\sum_{j:\overline{\beta}^*_j=0} \xi_j^{\frac{1}{2}} |\alpha_j| \leq 2\sum_{j:\overline{\beta}^*_j\neq 0} \xi_j^{\frac{1}{2}} |\alpha_j|$, we have $\alpha'\alpha \leq (c/n)\alpha'(\overline{Z}'\overline{Z})\alpha$. Let $\hat{\beta}$ be any solution of the the quadratic program*

$$\underset{\beta\in\mathbb{R}^p}{\text{Argmin}} \|Z\beta\|_2^2 \qquad s.t. \ \left\|\Xi^{-1}X'\left(Y - X\beta\right)\right\|_\infty \leq s$$

and let [4] $\tilde{\beta} = \overline{P}\hat{\beta}$. *Let us choose $\kappa > 2\sqrt{2}$ and $s = \kappa\sigma\sqrt{n\log(p)}$. Then with probability greater than $1 - p^{1-\frac{\kappa^2}{8}}$, we have the total error on the prevision of $x_{n+i}\beta^*$ for $1 \leq i \leq n$ that is given by*

$$\left\|\overline{Z}(\tilde{\beta} - \overline{\beta}^*)\right\|_2^2 \leq 16c\kappa^2\sigma^2\log(p)\sum_{\overline{\beta}^*\neq 0}\xi_j.$$

Note that in the theorems about LASSO, there are always hypotheses about the matrix X that is *given* to the statistician. Here, the only hypotheses is about $\overline{Z}$ that is *chosen* by the statistician. For example, if $\overline{Z}'\overline{Z}$ is not enough well conditioned, it is possible to "add vectors" in $\overline{Z}$, namely, to choose a larger m. *This is a real improvement.* However, there is a *price to pay* for this improvement, as explained now.

Remark that the matrix $\overline{Z}$ is not normalized. If one is to impose such a normalization, the more natural thing to do is to impose that the diagonal elements

[4] Equivalently, we could define

$$\tilde{\beta} = \arg\min_\beta \left\{-2Y'X\beta + \beta'\overline{Z}'\overline{Z}\beta + 2s\left\|\Xi\beta\right\|_1\right\}.$$

of $\overline{Z}'\overline{Z}$ are constant. But in this case, one can check that (in general) *it is no longer possible to normalize X in such a way that $\xi_j = 1$* for any j. So, without any assumption about a link between $\overline{Z}$ and $\overline{X}$ that would allow to have more information on X, it is not possible to control $\sum_{\overline{\beta}^* \neq 0} \xi_j$ by $\|\overline{\beta}^*\|_0$.

5 Conclusion

Based on a geometrical remark in [Alq08], we studied the family of estimators defined by

$$\underset{\beta \in DC(s)}{\text{Argmin}} \|\mathcal{M}\beta\|_q^q$$

that includes the LASSO, the Dantzig Selector, the Correlation Selector and the transductive LASSO in some particular cases: $q = 1$ and $q = 2$ for a quite general matrix $\mathcal{M}$, and $q \geq 1$ for the particular case $\mathcal{M} = (X'X)$.

Future works could include the theoretical study of our estimator for a general matrix $\mathcal{M}$ and $q \notin \{1, 2\}$, as well as an extension of the LARS algorithm to compute efficiently the solutions of Program (3) and (4).

6 Proofs

6.1 Basic algebra results

In this subsection, we prove the basic algebra results claimed in the introduction.

Proof that the kernel condition implies the existence of P. As $(Z'Z)$ is symmetric, we can diagonalize it in an orthogonal basis, given by a matrix Q: there is a $q \in \{0, ..., p\}$ and $\lambda_1, ..., \lambda_q > 0$ with

$$(Z'Z) = Q' \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right) Q, \text{ with } D = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_q \end{pmatrix}.$$

Remark also that this implies that $Ker(Z'Z) \perp Im(Z'Z)$. Now, remark that the kernel condition implies that $Ker(X'X) = Ker(Z'Z)$ and, because $(X'X)$ is symmetric, we have $Ker(X'X) \perp Im(X'X)$. This implies that $Im(Z'Z) = Im(X'X)$. So, $(X'X)$ can be "partially diagonalized" in the basis Q, in the sense that

$$(X'X) = Q' \left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array}\right) Q$$

where B is some invertible $q \times q$ matrix. Now, let us put

$$P = Q' \left(\begin{array}{c|c} B^{-1}D & 0 \\ \hline 0 & I_{p-q} \end{array} \right) Q.$$

We can easily check that P is invertible and that

$$(X'X)P = Q' \left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array} \right) QQ' \left(\begin{array}{c|c} B^{-1}D & 0 \\ \hline 0 & I_{p-q} \end{array} \right) Q = Q' \left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array} \right) Q = (Z'Z).$$

□ □

Proof of the existence of W. This proof uses the same arguments, we just put

$$W = Q' \left(\begin{array}{c|c} D^{-1} & 0 \\ \hline 0 & I_{p-q} \end{array} \right) Q$$

and we check that $(Z'Z)W(X'X) = (X'X)$ and that W is symmetric. □ □

Proof of $Z\beta^* = Z\beta^{**}$. We have $(Z'Z)\beta^* = (X'X)\gamma^* = (Z'Z)\beta^{**}$ by definition. So $(Z'Z)(\beta^* - \beta^{**}) = 0$ and so $(\beta^* - \beta^{**})'(Z'Z)(\beta^* - \beta^{**}) = 0$ and $Z(\beta^* - \beta^{**}) = 0$. □ □

6.2 Proof of Theorem 1

Proof. Let us remark that **Program I** can be written

$$\min_{\beta \in \mathbb{R}^p} \beta(Z'Z)\beta \qquad \text{s. t.} \quad \left\| \Xi^{-1}X'(Y - X\beta) \right\|_\infty \leq s \tag{19}$$

Let us write the Lagrangian of this program:

$$\mathcal{L}(\beta, \lambda, \mu) = \beta(Z'Z)\beta + \lambda'\Xi^{-1}\left[X'(X\beta - Y) - sE\right] + \mu'\Xi^{-1}\left[X'(Y - X\beta) - sE\right]$$

with $E = (1, ..., 1)'$, and for any j, $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j\mu_j = 0$. Any solution $\underline{\beta} = \underline{\beta}(\lambda, \mu)$ of **Program I** must satisfy

$$0 = \frac{\partial \mathcal{L}}{\partial \beta}(\underline{\beta}, \lambda, \mu) = 2\underline{\beta}'(Z'Z) + (\lambda - \mu)'\Xi^{-1}(X'X),$$

so

$$(Z'Z)\underline{\beta} = (X'X)\frac{1}{2}\Xi^{-1}(\mu - \lambda).$$

Note that the conditions $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j\mu_j = 0$ means that there is a $\gamma_j \in \mathbb{R}$ such that $\gamma_j = \xi_j^{\frac{1}{2}}(\mu_j - \lambda_j)/2$, $|\gamma_j| = \xi_j^{\frac{1}{2}}(\lambda_j + \mu_j)/2$, and so $\lambda_j = 2(\gamma_j/\xi_j^{\frac{1}{2}})_-$

and $\mu_j = 2(\gamma_j/\xi_j^{\frac{1}{2}})_+$, where $(a)_+ = max(a;0)$ and $(a)_- = max(-a;0)$. Let also γ denote the vector which j-th component is exactly γ_j, we obtain:

$$(Z'Z)\underline{\beta} = (X'X)\gamma. \tag{20}$$

Note that this also implies that:

$$\underline{\beta}'(Z'Z)\underline{\beta} = \underline{\beta}'(X'X)\gamma = \underline{\beta}'(Z'Z)W(X'X)\gamma = \gamma'(X'X)W(X'X)\gamma.$$

Using these relations, the Lagrangian may be written:

$$\begin{aligned}\mathcal{L}(\underline{\beta},\lambda,\mu) &= \gamma'(X'X)W(X'X)\gamma + 2\gamma'X'Y - 2\gamma'(X'X)\underline{\beta} - 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}|\gamma_j| \\ &= 2\gamma'X'Y - \gamma'(X'X)W(X'X)\gamma - 2s\left\|\Xi\gamma\right\|_1\end{aligned}$$

Note that λ and β, and so γ, should maximize this value. Hence, γ is to minimize

$$-2\gamma'X'Y + \gamma'(X'X)W(X'X)\gamma + 2s\|\Xi\gamma\|_1 + Y'Y$$

Now, note that

$$Y'Y - 2\gamma'X'Y = \|Y - X\gamma\|_2^2 - \gamma'(X'X)\gamma$$

and then γ is also to minimize

$$\|Y - X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma'\left[(X'X) - (X'X)W(X'X)\right]\gamma,$$

what is claimed in the theorem. Let $\underline{\gamma}$ denote a solution of this program. Equation (20) implies that if $\underline{\beta}$ is a solution of **Program I** then $(Z'Z)\underline{\beta} = (X'X)\underline{\gamma}$. Let $\overline{\beta}$ we another solution of **Program I**, note that we also have $(Z'Z)\overline{\beta} = (X'X)\underline{\gamma}$. Then $(Z'Z)(\underline{\beta} - \overline{\beta}) = 0$ so $(\underline{\beta} - \overline{\beta})$ belongs to $\ker Z$ and so to $\ker X$. This ends the proof. □ □

6.3 Proof of Lemma 1

Proof. Remember that Assumption (A2) implies among others that $\xi_j = 1$ for any j. Let $\alpha \in \mathbb{R}^p$ satisfy: $\sum_{j:\gamma_j^*=0}|\alpha_j| \leq 3\sum_{j:\gamma_j^*\neq 0}|\alpha_j|$. We have:

$$\begin{aligned}\alpha'\Omega\alpha &= \sum_{j:\gamma_j^*\neq 0}\Omega_{j,j}\alpha_j^2 + \sum_{j:\gamma_j^*=0}\sum_{k:\gamma_k^*=0}\Omega_{j,k}\alpha_j\alpha_k \\ &\quad +2\sum_{j:\gamma_j^*\neq 0}\sum_{k:\gamma_k^*=0}\Omega_{j,k}\alpha_j\alpha_k + \sum_{j:\gamma_j^*\neq 0}\sum_{\substack{k:\gamma_k^*\neq 0\\ k\neq j}}\Omega_{j,k}\alpha_k\alpha_j\end{aligned}$$

$$\geq \sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 + 2\sum_{j:\gamma_j^*\neq 0}\sum_{k:\gamma_k^*=0} \Omega_{j,k}\alpha_j\alpha_k + \sum_{j:\gamma_j^*\neq 0}\sum_{\substack{k:\,\gamma_k^*\neq 0\\ k\neq j}} \Omega_{j,k}\alpha_j\alpha_k.$$

So we have

$$\begin{aligned}\sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 &\leq \alpha'\Omega\alpha - 2\sum_{j:\gamma_j^*\neq 0}\sum_{k:\gamma_k^*=0} \Omega_{j,k}\alpha_j\alpha_k - \sum_{j:\gamma_j^*\neq 0}\sum_{\substack{k:\,\gamma_k^*\neq 0\\ k\neq j}} \Omega_{j,k}\alpha_j\alpha_k \\ &\leq \alpha'\Omega\alpha + \left(\sup_{\gamma_j^*\neq 0}\sup_{k\neq j}|\Omega_{j,k}|\right)\left[2\left(\sum_{\gamma_j^*\neq 0}|\alpha_j|\right)\left(\sum_{\gamma_k^*=0}|\alpha_k|\right) + \left(\sum_{\gamma_j^*\neq 0}|\alpha_j|\right)^2\right] \\ &\leq \alpha'\Omega\alpha + 7\left(\sup_{\gamma_j^*\neq 0}\sup_{k\neq j}|\Omega_{j,k}|\right)\left(\sum_{\gamma_j^*\neq 0}|\alpha_j|\right)^2 = \alpha'\Omega\alpha + 7\rho\left(\sum_{\gamma_j^*\neq 0}|\alpha_j|\right)^2. \quad (21)\end{aligned}$$

On the other hand, using the Cauchy-Schwarz inequality, we have

$$\left(\sum_{\gamma_j^*\neq 0}|\alpha_j|\right)^2 \leq \|\gamma^*\|_0 \sum_{\gamma_j^*\neq 0}\alpha_j^2 \leq \frac{\|\gamma^*\|_0}{\inf_{\gamma_j^*\neq 0}\Omega_{j,j}}\sum_{\gamma_j^*\neq 0}\Omega_{j,j}\alpha_j^2. \quad (22)$$

Combining (21) and (22), we obtain

$$\sum_{j:\gamma_j^*\neq 0} \Omega_{j,j}\alpha_j^2 \leq \frac{1}{1-7\frac{\|\gamma^*\|_0}{\inf_{\gamma_j^*\neq 0}\Omega_{j,j}}\rho}\alpha'\Omega\alpha.$$

Now, remember that we assumed that $\rho \leq \frac{\inf_{\gamma_j^*\neq 0}\Omega_{j,j}}{14\|\gamma^*\|_0}$ by hypotheses and conclude by

$$\sum_{\gamma_j^*\neq 0}\alpha_j^2 \leq \frac{1}{\inf_{\gamma_j^*\neq 0}\Omega_{j,j}}\sum_{j:\gamma_j^*\neq 0}\Omega_{j,j}\alpha_j^2 \leq \frac{2\alpha'\Omega\alpha}{\inf_{\gamma_j^*\neq 0}\Omega_{j,j}}.$$

□ □

6.4 A useful Lemma

Lemma 2. *Let $\Lambda_{n,p}$ be the random event defined by*

$$\Lambda_{n,p} = \left\{\forall j \in \{1,...,p\}, \quad 2|V_j| \leq s\xi_j^{\frac{1}{2}}\right\}, \quad (23)$$

where $V_j = \sum_{i=1}^n x_{i,j}\varepsilon_i$. Let us choose a $\kappa > 2\sqrt{2}$ and $s = \kappa\sigma\sqrt{n\log(p)}$. Then

$$\mathbb{P}(\Lambda_{n,p}) \geq 1 - p^{1-\frac{\kappa^2}{8}}.$$

Proof. Remember that $\xi_j = \frac{1}{n}\sum_{i=1}^n x_{i,j}^2$. Since $V_j = \sum_{i=1}^n x_{i,j}\varepsilon_i \sim \mathcal{N}(0, n\xi_j\sigma^2)$, an elementary Gaussian inequality gives

$$\begin{aligned}\mathbb{P}\left(\max_{l=1,\dots,p} s^{-1}\xi_j^{-\frac{1}{2}}|V_l| \geq 2^{-1}\right) &\leq p \max_{l=1,\dots,p} \mathbb{P}\left(s^{-1}\xi_j^{-\frac{1}{2}}|V_l| \geq 2^{-1}\right)\\ &\leq p\exp\left(-\kappa^2\log(p)/8\right) = p^{1-\kappa^2/8}.\end{aligned}$$

□ □

6.5 Proof of Theorem 2

The proof follows the technique used in [BTW07b]. We begin by a preliminary lemma.

Lemma 3. *Let us consider the regression model* (1)*. Let $\hat{\gamma}$ be a solution of Program* (8)*. Let us assume that $\Lambda_{n,p}$, the event defined in Lemma 2, is satisfied. Then*

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2 + s\left\|\Xi(\hat{\gamma}-\gamma^*)\right\|_1 \leq 4s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|. \tag{24}$$

Proof of Lemma 3. Let us remember the criterion (8):

$$\operatorname*{Argmin}_{\gamma\in\mathbb{R}^p}\left\{\|Y-X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma' M\gamma\right\}$$

with $Y = X\beta^* + \varepsilon$ and M is given by (12). First, let us prove that $X\beta^* = XW(X'X)\gamma^*$, we start from the relation $(Z'Z)\beta^* = (X'X)\gamma^* = (Z'Z)W(X'X)\gamma^*$ so $\beta^* - W(X'X)\gamma^* \in \ker(Z'Z) = \ker Z = \ker X$ and then $X\beta^* = XW(X'X)\gamma^*$. Therefore we have

$$\begin{aligned}\|Y-X\gamma\|_2^2 &= \|XW(X'X)\gamma^* - X\gamma + \varepsilon\|_2^2 = \|X[W(X'X)-I_p]\gamma^* + X(\gamma^*-\gamma)+\varepsilon\|_2^2\\ &= (\gamma^*)'[(X'X)W - I_p](X'X)[W(X'X)-Ip]\gamma^* + \|X\gamma^* - X\gamma\|_2^2\\ &\quad + \|\varepsilon\|_2^2 + 2\Big\{(\gamma^*)'[(X'X)W-I_p](X'X)(\gamma^*-\gamma)\\ &\quad + \varepsilon' X[W(X'X)-I_p)\gamma^* + \varepsilon' X(\gamma^*-\gamma)\Big\}.\end{aligned}$$

Then, since $M = (X'X)W(X'X) - (X'X)$, we have

$$\begin{aligned}&\operatorname*{Argmin}_{\gamma\in\mathbb{R}^p}\left\{\|Y-X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma' M\gamma\right\}\\ &= \operatorname*{Argmin}_{\gamma\in\mathbb{R}^p}\left\{\|X\gamma^*-X\gamma\|_2^2 + 2s\left\|\Xi\gamma\right\|_1 + \gamma' M\gamma - 2\varepsilon' X\gamma + 2(\gamma^*)'M(\gamma^*-\gamma)\right\}.\end{aligned}$$

Using now the definition of $\hat{\gamma}$ (it minimizes the above quantity) we obtain

$$\begin{aligned}\|X(\hat{\gamma}-\gamma^*)\|_2^2 &\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(\left|\gamma_j^*\right|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*)\\ &\qquad+\left[(\gamma^*)'M\gamma^*-\hat{\gamma}'M\hat{\gamma}-2(\gamma^*)'M(\gamma^*-\hat{\gamma})\right]\\ &\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(\left|\gamma_j^*\right|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*)-(\gamma^*-\hat{\gamma})'M(\gamma^*-\hat{\gamma}).\end{aligned}$$

As a consequence, replacing M by its definition we obtain

$$(\gamma^*-\hat{\gamma})'(X'X)W(X'X)(\gamma^*-\hat{\gamma})\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*).$$

Note that

$$\begin{aligned}(\gamma^*-\hat{\gamma})'(X'X)W(X'X)(\gamma^*-\hat{\gamma}) &= (\beta^*-\hat{\beta})'(Z'Z)W(X'X)(\gamma^*-\hat{\gamma})\\ &= (\beta^*-\hat{\beta})'(X'X)(\gamma^*-\hat{\gamma})=(\beta^*-\hat{\beta})'(Z'Z)(\beta^*-\hat{\beta}),\end{aligned}\quad(25)$$

then our bound so far is

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)+2\sum_{i=1}^{n}\varepsilon_i x_i(\hat{\gamma}-\gamma^*).\qquad(26)$$

Moreover, on the event $\Lambda_{n,p}$, we have

$$2\sum_{i=1}^{n}\varepsilon_i x_i\left(\hat{\gamma}-\gamma^*\right)=2\sum_{j=1}^{p}V_j\left(\hat{\gamma}_j-\gamma_j^*\right)\leq\sum_{j=1}^{p}s\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|.\qquad(27)$$

It follows from (26) and (27) that

$$\begin{aligned}\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2+s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right| &\leq 2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|+2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}|\gamma_j^*|-2s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}|\hat{\gamma}_j|\\ &\leq 2s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|+2s\sum_{\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left(|\gamma_j^*|-|\hat{\gamma}_j|\right)\\ &\leq 4s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|.\end{aligned}$$

This is the result claimed in the lemma. □ □

We are now ready to give the

Proof of Theorem 2. We apply Lemmas 2 and 3 and state that Inequality (24)

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2 + s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right| \leq 4s\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|.$$

holds with probability at least $1-p^{1-\frac{\kappa^2}{8}}$. This equation implies in particular that

$$\sum_{j:\gamma_j^*=0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right| \leq 3\sum_{j:\gamma_j^*\neq 0}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right|,$$

then taking $\alpha=\hat{\gamma}-\gamma^*$ in Assumption (A1), we obtain

$$\begin{aligned}\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2 + s\sum_{j=1}^{p}\xi_j^{\frac{1}{2}}\left|\hat{\gamma}_j-\gamma_j^*\right| &\leq 4s\sqrt{\left(\sum_{\gamma_j^*\neq 0}\xi_j\right)\left(\sum_{\gamma_j^*\neq 0}\left(\hat{\gamma}_j-\gamma_j^*\right)^2\right)} \\ &\leq 4s\sqrt{\left(\sum_{\gamma_j^*\neq 0}\xi_j\right)\frac{c}{n}(\hat{\gamma}-\gamma^*)'\Omega(\hat{\gamma}-\gamma^*)} = 4s\sqrt{\left(\sum_{\gamma_j^*\neq 0}\xi_j\right)\frac{c}{n}\|Z(\hat{\beta}-\beta^*)\|_2^2},\end{aligned}$$

where we used (25) in the last equality. As a consequence,

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2 \leq 4s\sqrt{\frac{c}{n}\sum_{\gamma_j^*\neq 0}\xi_j}$$

and so

$$\left\|Z(\hat{\beta}-\beta^*)\right\|_2^2 \leq \frac{16s^2c}{n}\sum_{\gamma_j^*\neq 0}\xi_j = 16c\kappa^2\sigma^2\log(p)\sum_{\gamma_j^*\neq 0}\xi_j$$

while

$$s\left\|\Xi(\hat{\gamma}-\gamma^*)\right\|_1 \leq 4s\sqrt{\frac{c}{n}\|Z(\hat{\beta}-\beta^*)\|_2^2\sum_{\gamma_j^*\neq 0}\xi_j}$$

which implies that

$$\left\|\Xi(\hat{\gamma}-\gamma^*)\right\|_1 \leq 4\sqrt{\frac{16c^2\kappa^2\sigma^2\log(p)}{n}}\sum_{\gamma_j^*\neq 0}\xi_j.$$

This ends the proof. □ □

6.6 Proof of Theorem 4

First, we give the following lemma.

Lemma 4. *We have, on the event* $\Lambda_{n,p}$,

$$\|Z(\tilde{\beta}-\beta^*)\|_2^2 \leq \frac{3s}{2}\left\|\Xi P\left(\tilde{\beta}-\beta^{**}\right)\right\|_1 \leq 3s\sum_{P_j\beta^{**}\neq 0}\xi_j^{\frac{1}{2}}\left|P_j\left(\tilde{\beta}-\beta^{**}\right)\right|. \quad (28)$$

Proof of Lemma 4. We have

$$\begin{aligned}\|Z(\tilde{\beta}-\beta^*)\|_2^2 &= \|Z(\tilde{\beta}-\beta^{**})\|_2^2 = (\tilde{\beta}-\beta^{**})'(Z'Z)(\tilde{\beta}-\beta^{**})\\ &= [P(\tilde{\beta}-\beta^{**})]'(X'X)(\tilde{\beta}-\beta^{**}) = [\Xi P(\tilde{\beta}-\beta^{**})]'\Xi^{-1}(X'X)(\tilde{\beta}-\beta^{**})\\ &\leq \|\Xi P(\tilde{\beta}-\beta^{**})\|_1\|\Xi^{-1}(X'X)(\tilde{\beta}-\beta^{**})\|_\infty \leq \frac{3s}{2}\|\Xi P(\tilde{\beta}-\beta^{**})\|_1, \quad (29)\end{aligned}$$

where we use the Dantzig constraint in the last inequality. Note that, by definition of $\tilde{\beta}$,

$$\begin{aligned}0 \leq \|\Xi P\beta^{**}\|_1 - \|\Xi P\tilde{\beta}\|_1 &= \sum_{P_j\beta^{**}\neq 0}\xi_j^{\frac{1}{2}}|P_j\beta^{**}| - \sum_{P_j\beta^{**}\neq 0}\xi_j^{\frac{1}{2}}\left|P_j\tilde{\beta}\right| - \sum_{P_j\beta^{**}=0}\xi_j^{\frac{1}{2}}\left|P_j\tilde{\beta}\right|\\ &\leq \sum_{P_j\beta^{**}\neq 0}\xi_j^{\frac{1}{2}}\left|P_j\beta^{**}-P_j\tilde{\beta}\right| - \sum_{P_j\beta^{**}=0}\xi_j^{\frac{1}{2}}\left|P_j\beta^{**}-P_j\tilde{\beta}\right|,\end{aligned}$$

that leads to Inequality (28). □ □

Proof of Theorem 4. We apply here Lemmas 2 and 4 and we obtain that with probability at least $1-p^{1-\frac{\kappa^2}{8}}$, we have Inequality (28). Now, let us remark that

$$\begin{aligned}\|Z(\beta^*-\tilde{\beta})\|_2^2 &\leq \frac{3s}{2}\|\Xi P\left(\beta^{**}-\tilde{\beta}\right)\|_1 \leq 3s\sum_{P_j\beta^{**}\neq 0}\xi_j^{\frac{1}{2}}\left|P_j\beta^{**}-P_j\tilde{\beta}\right|\\ &\leq 3s\sqrt{\left(\sum_{P_j\beta^{**}\neq 0}\xi_j\right)\left(\sum_{P_j\beta^{**}\neq 0}\left|P_j\beta^{**}-P_j\tilde{\beta}\right|^2\right)}\\ &\leq 3s\left(\sum_{P_j\beta^{**}\neq 0}\xi_j\right)^{\frac{1}{2}}\sqrt{\frac{c}{n}\|Z(\beta^*-\tilde{\beta})\|_2^2}.\end{aligned}$$

So we have,

$$\|Z(\tilde{\beta}-\beta^*)\|_2^2 \leq 9s^2\frac{c}{n}\sum_{P_j\beta^{**}\neq 0}\xi_j,$$

and as a consequence

$$\frac{3s}{2}\left\|\Xi P\left(\beta^{**}-\tilde{\beta}\right)\right\|_1 \leq 3s\left(\sum_{P_j\beta^{**}\neq 0}\xi_j\right)^{\frac{1}{2}}\sqrt{\frac{c}{n}\|Z(\beta^*-\tilde{\beta})\|_2^2} \leq 9s^2\frac{c}{n}\sum_{P_j\beta^{**}\neq 0}\xi_j,$$

this ends the proof. □ □

6.7 Proof of Theorems 3 and 5

Let us remind that Assumption (A2), involved in both theorems, implies among others that $\xi_j = 1$ for any j, so Ξ is the identity matrix.

Proof of Theorem 3. We can rewrite the fact that $\hat{\beta} = P\hat{\gamma}$ satisfies the Dantzig constraint:

$$\left\|\Omega(\hat{\gamma}-\gamma^*)-\frac{X'\varepsilon}{n}\right\|_\infty \leq \frac{s}{n}. \tag{30}$$

Recall that $\Lambda_{n,p} = \{\max_{j=1,\ldots,p} 2|V_j| \leq s\}$ with $V_j = X_j'\varepsilon$, then applying (30), we have on $\Lambda_{n,p}$ and for any $j \in \{1,\ldots,p\}$,

$$\begin{aligned}
\left|\Omega_{j,j}(\hat{\gamma}_j-\gamma_j^*)\right| &= \left|\{\Omega(\hat{\gamma}-\gamma^*)\}_j - \sum_{\substack{k=1\\k\neq j}}^{p}\Omega_{j,k}(\hat{\gamma}_k-\gamma_k^*)\right| \\
&\leq \frac{s}{n} + \left|\frac{X_j'\varepsilon}{n}\right| + \sum_{\substack{k=1\\k\neq j}}^{p}|\Omega_{j,k}(\hat{\gamma}_k-\gamma_k^*)| \\
&\leq \frac{3s}{2n} + \sum_{\substack{k=1\\k\neq j}}^{p}|\Omega_{j,k}(\hat{\gamma}_k-\gamma_k^*)| \leq \frac{3s}{2n} + \|\hat{\gamma}-\gamma^*\|_1\left(\sup_{k\neq j}|\Omega_{j,k}|\right)
\end{aligned}$$

which implies that

$$\|\hat{\gamma}-\gamma^*\|_\infty \leq \frac{1}{\inf_j \Omega_{j,j}}\left[\frac{3s}{2n} + \|\hat{\gamma}-\gamma^*\|_1\left(\sup_j\sup_{k\neq j}|\Omega_{j,k}|\right)\right]. \tag{31}$$

Now, remind that $\sup_j \sup_{k\neq j}|\Omega_{j,k}| = \rho$ is upper bounded by Assumption (A2). Moreover, Assumption (A2) implies, by Lemma 1, that Assumption (A1) is satisfied with $c = 2/(\inf_j \Omega_{j,j})$, so we can apply Theorem 2 to upper bound $\|\hat{\gamma}-\gamma^*\|_1$. This leads to

$$\|\hat{\gamma}-\gamma^*\|_\infty \leq \frac{1}{\inf_j \Omega_{j,j}}\left[\frac{3s}{2n} + 16c\kappa\sigma\|\gamma^*\|_0\sqrt{\frac{\log(p)}{n}} \times \frac{\inf_j \Omega_{j,j}}{14\|\gamma^*\|_0}\right]$$

writing $c = 2/(\inf_j \Omega_{j,j})$ and $s = \kappa\sigma\sqrt{n\log(p)}$ we obtain:

$$\|\hat{\gamma} - \gamma^*\|_\infty \leq \frac{3\kappa\sigma}{\inf_j \Omega_{j,j}}\sqrt{\frac{\log(p)}{n}},$$

that is the inequality stated in Theorem 3. □ □

Proof of Theorem 5. The preceding proof is also valid for Theorem 5. The only difference is that, at Inequality (31), upper bound the l_1 norm using Theorem 4 instead of Theorem 2. This replaces the constant 16 by a 6. □ □

[Alq08] P. Alquier. Lasso, iterative feature selection and the correlation selector: Oracle inequalities and numerical performances. *Electron. J. Stat.*, pages 1129–1152, 2008.

[Bac08] F. Bach. Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.*, 9:1179–1225, 2008.

[BRT07] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. Submitted to the Ann. Statist., 2007.

[BTW07a] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[BTW07b] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electron. J. Stat.*, 1:169–194, 2007.

[Bun07] F. Bunea. Consistent selection via the lasso for high dimensional approximating regression models. IMS Lecture Notes-Monograph Series, to appear, 2007.

[CH08] C. Chesneau and M. Hebiri. Some theoretical results on the grouped variables lasso. *Mathematical Methods of Statistics*, 17(4):317–326, 2008.

[CT07] E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35, 2007.

[DET06] D. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

[DT07] A. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. *COLT 2007 Proceedings. Lecture Notes in Computer Science 4539 Springer*, pages 97–111, 2007.

[FL01] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[Heb08] M. Hebiri. Regularization with the smooth-lasso procedure. Preprint LPMA, 2008.

[Kol07] V. Koltchinskii. Dantzig selector and sparsity oracle inequalities. Manuscript, 2007.

[Kol08] V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. Manuscript, 2008.

[Lou08] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[MB06] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

[MVdGB08] L. Meier, S. Van de Geer, and P. Bühlmann. High-dimensional additive modeling. Manuscript, 2008.

[MY09] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270, 2009.

[OPT00] M. Osborne, B. Presnell, and B. Turlach. On the LASSO and its dual. *J. Comput. Graph. Statist.*, 9(2):319–337, 2000.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[Vap98] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1998.

[vdG08] S. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[Wai06] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming. Technical report n. 709, Department of Statistics, UC Berkeley, 2006.

[Zou06] H. Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.

[ZY06] P. Zhao and B. Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.