LASSO, Iterative Feature Selection and the Correlation Selector: Oracle Inequalities and Numerical Performances

Pierre Alquier

LASSO, ITERATIVE FEATURE SELECTION AND THE CORRELATION SELECTOR: ORACLE INEQUALITIES AND NUMERICAL PERFORMANCES

PIERRE ALQUIER

Abstract. We propose a general family of algorithms for regression estimation with quadratic loss. Our algorithms are able to select relevant functions into a large dictionary. We prove that a lot of algorithms that have already been studied for this task (LASSO and Group LASSO, Dantzig selector, Iterative Feature Selection, among others) belong to our family, and exhibit another particular member of this family that we call Correlation Selector in this paper. Using general properties of our family of algorithm we prove oracle inequalities for IFS, for the LASSO and for the Correlation Selector, and compare numerical performances of these estimators on a toy example.

Contents

Date: February 1, 2008.
2000 *Mathematics Subject Classification.* Primary 62G08; Secondary 62J07, 62G15, 68T05.
Key words and phrases. Regression estimation, statistical learning, confidence regions, shrinkage and thresholding methods, LASSO.
I Would like to thank Professors Olivier Catoni and Alexandre Tsybakov for useful remarks.

1. Introduction

1.1. **Setting of the problem.** Let $n \in \mathbb{N}\backslash\{0\}$. Let P be a probability distribution on

$$\left(\left(\mathcal{X} \times \mathbb{R}\right)^n, \left(\mathcal{B} \otimes \mathcal{B}_{\mathbb{R}}\right)^{\otimes n} \right)$$

and

$$\Big((X_1, Y_1), ..., (X_n, Y_n) \Big)$$

drawn from P.

For $i \in \{1, ..., n\}$, let p_i denote the marginal distribution of X_i under P, and let us put:

$$P_X = \frac{1}{n} \sum_{i=1}^{n} p_i.$$

We assume that P_X is known to the statistician.

Moreover, we put:

$$\overline{P} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}.$$

The statistician chooses a dictionary of functions: $(f_1, ..., f_m)$. For the sake of simplicity we assume that it is such that for any $j \in \{1, ..., m\}$ we have

$$P_X \left[f_j^2 \right] = 1.$$

Definition 1.1. *Let us put, for any $\alpha = (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m$ and $(x, y) \in \mathcal{X} \times \mathbb{R}$:*

$$l_\alpha(x, y) = \left(y - \sum_{j=1}^{m} \alpha_j f_j(x) \right).$$

We define:

$$r(\alpha) = \overline{P}\left(l_\alpha\right) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{m} \alpha_j f_j(X_i) \right]^2$$

and

$$R(\alpha) = P\left[r\left(\alpha\right)\right].$$

We put:

$$\overline{\alpha} \in \arg\min_{\alpha \in \mathbb{R}^m} R(\alpha).$$

For any $\alpha, \alpha' \in \mathbb{R}^m$ we put:

$$\left\langle \alpha, \alpha' \right\rangle_X = P_X \left[\sum_{j=1}^{m} \sum_{k=1}^{m} \alpha_j \alpha'_k f_j f_k \right],$$

and

$$\left\| \alpha \right\|_X = \sqrt{\left\langle \alpha, \alpha \right\rangle}.$$

Finally, we put $e_1 = (1, 0, ..., 0) \in \mathbb{R}^m$, ..., $e_m = (0, ..., 0, 1) \in \mathbb{R}^m$ the canonical basis of $\mathbb{R}^m$.

Let us remark that for any $\alpha \in \mathbb{R}^m$ we have

$$R(\alpha) - R\left(\overline{\alpha}\right) = \left\| \alpha - \overline{\alpha} \right\|_X^2.$$

Remark 1.1. We think of three cases of interest. If the pairs (X_i, Y_i) are i. i. d. we have $p_1 = ... = p_n = P_X$ and so P_X is the marginal distribution of X. It is assumed to be known to the statistician (restrictive hypothesis).

Another case of interest is when the values $X_1, ..., X_n$ are deterministic. In this case,

$$P_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

and so we obtain:

$$\langle \alpha, \alpha' \rangle_X = \frac{1}{n}\sum_{i=1}^{n} \left[\sum_{j,k} \alpha_j \alpha_k f_j(X_i) f_k(X_i) \right].$$

In this case $\|.\|_X$ is called the empirical norm (usually denoted $\|.\|_n$). This context is the one adopted in papers like Bunea, Tsybakov and Wegkamp [5].

The last setting we think of is probably the most important in the view of applications, it's the so-called transductive setting introduced by Vapnik [15]. We assume that there are deterministic pairs (x_i, y_i) for $i \in \{1, ..., (k+1)n\}$; we observe all the x_i but there is a cost for the observation of y_i, so we drawn n different values of i, uniformly on $\{1, ..., (k+1)n\}$, and let (X_i, Y_i) for $i \in \{1, ..., n\}$ denote the corresponding pairs. The idea is to guess the whole set of values $(y_i, i \in \{1, ..., (k+1)n\})$ from the restricted set $(Y_i, i \in \{1, ..., n\})$. Note that this matches with our context, with:

$$r(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{m} \alpha_j f_j(X_i) \right]^2,$$

$$R(\alpha) = \frac{1}{(k+1)n} \sum_{i=1}^{(k+1)n} \left[y_i - \sum_{j=1}^{m} \alpha_j f_j(x_i) \right]^2,$$

and

$$P_X = \frac{1}{(k+1)n} \sum_{i=1}^{(k+1)n} \delta_{x_i}.$$

See our previous paper [1] for more details.

Definition 1.2. *Let $\mathcal{C}$ be a closed, convex subset of $\mathbb{R}^d$. We let $\Pi_{\mathcal{C}}^X(.)$ denote the orthogonal projection on $\mathcal{C}$ with respect to the norm $\|.\|_X$.*

1.2. **Organization of the paper.** The aim of this paper is to propose a method to estimate the real regression function (say f) by selecting a few relevant functions among all the functions in the dictionary.

Recently, a lot of algorithms have been proposed for that purpose, let's cite among others the LASSO by Tibshirani [13] and some variants or generalization like LARS by Efron, Hastie, Johnstone and Tibshirani [10], the Dantzig selector by Candes and Tao [6] and the Group LASSO by Yuan and Lin [16], or Iterative Feature Selection in our paper [1]. This paper proposes a general algorithm that contains LASSO, Dantzig selector and Iterative Feature Selection as a particular case.

A paper by Bunea, Tsybakov and Wegkamp [5] gives sparsity oracle inequalities for the LASSO, that is inequalities that bounds the risk of the LASSO estimators in terms of the number of selected functions in the dictionary. This paper by Bunea and al. is writen in a different context than ours: random design with *unknown* distribution (in the case of a random design, our method require the knowledge of the distribution of the design). Another paper, by Bickel, Ritov and Tsybakov

[3] gives sparsity oracle inequalities for the LASSO and the Dantzig selector in the case of the deterministic design. However, in both papers the main results require the assumtion $\|f_j\|_\infty \leq L$ for some given L that is not necessary in our paper, and that prevents the use of popular basis of functions like wavelets. This is partly due to the use of Hoeffding's inequality.

Another particular case of our general algorithm leads to the definition of a new estimator, really easy to compute, that will be named Correlation Selector in this paper. Our paper uses a geometric point of view that allows to obtain simple sparsity oracle inequalities for the obtained estimator, in both deterministic design case and random design with known distribution. It also uses a (Bernstein's type) deviation inequality proved in a previous work [1] that is sharper than Hoeffding's inequality, and so get rid of the assumption of a (uniform) bound over the functions of the dictionary. Another improvement is that our method is valid for some types of data-dependant of dictionaries of function, for example the case where $m = n$ and

$$\{f_1(.), ..., f_m(.)\} = \{K(X_1, .), ..., K(X_n, .)\}$$

where K is a function $\mathcal{X}^2 \to \mathbb{R}$.

In Section 2, we give the general form for our algorithm under a particular assumption (CRA, Definition 2.2) that says that we are able to build some confidence region for the best value of α in some subspace of $\mathbb{R}^m$.

In Section 3, we show why Iterative Feature Selection, LASSO, Dantzig Selector among others are particular cases of our algorithm. We exhibit another particular case of interest (called the Correlation Selector in this paper). Moreover, when we can we try to prove some oracle inequalities for the obtained estimator.

Section 4 is dedicated to simulations: we compare ordinary least square (OLS), LASSO, Iterative Feature Selection and the Correlation Selector on a toy example. Simulations shows that both particular cases of our family of estimators (LASSO and Iterative Feature Selection) generally outperforms the OLS estimate. Moreover, LASSO performs generally better than Iterative Feature Selection, however, this is not always true: this fact leads to the conclusion that a data-driven choice of a particular algorithm in our general family could lead to optimal results.

After a conclusion, Section 6 is dedicated to some proofs.

2. General projection algorithms

2.1. **Additional notations and hypothesis.** We choose $M \in \mathbb{N}$ and $S_1 \subset \{1, ..., m\}$, ... , $S_M \subset \{1, ..., m\}$. We put, for every $S \subset \{1, ..., m\}$:

$$\mathcal{M}_S = \Big\{\alpha \in \mathbb{R}^m, \quad \ell \notin S \Rightarrow \alpha_\ell = 0\Big\}.$$

So every $\mathcal{M}_{S_j}$ is a submodel of the original model $\mathbb{R}^m$.

Definition 2.1. *We put, for every $S \subset \{1, ..., m\}$:*

$$\overline{\alpha}_S = \arg\min_{\alpha \in \mathcal{M}_S} R(\alpha).$$

Remark that for every $S \subset \{1, ..., m\}$,

$$\overline{\alpha}_S = \Pi^X_{\mathcal{M}_S}(\overline{\alpha}).$$

Moreover let us put:

$$\hat{\alpha}_S = \arg\min_{\alpha \in \mathcal{M}_S} r(\alpha).$$

Definition 2.2. *We say that the confidence region assumption (CRA) is satisfied if for $\varepsilon \in [0, 1]$ we have a bound $r(S_j, \varepsilon) \in \mathbb{R}$ such that*

$$P\Big[\forall j \in \{1, ..., M\}, \quad \left\|\overline{\alpha}_{S_j} - \hat{\alpha}_{S_j}\right\|^2_X \leq r(S_j, \varepsilon)\Big] \geq 1 - \varepsilon.$$

Definition 2.3. *We define, for any $\varepsilon > 0$ and $j \in \{1, ..., M\}$, the random set*

$$\mathcal{CR}(j, \varepsilon) = \left\{ \alpha \in \mathbb{R}^m, \quad \left\| \Pi^X_{\mathcal{M}_{S_j}}(\alpha) - \hat{\alpha}_{S_j} \right\|^2_X \leq r(S_j, \varepsilon) \right\}.$$

We remark that the hypothesis implies that

$$P\left[\forall j \in \{1, ..., M\}, \quad \overline{\alpha} \in \mathcal{CR}(j, \varepsilon) \right] \geq 1 - \varepsilon.$$

In our previous work [1] we examined different hypothesis on the probability P such that this hypothesis is satisfied. For example, using inequalities by Catoni [7] and Panchenko [11] we proved the following results (for models of dimension 1, that will be the most used in the sequel of this paper).

Lemma 2.1. *Let us assume that $P = \mathbb{P}_1 \otimes ... \otimes \mathbb{P}_n$. Let us assume that $Y_i = f(X_i) + \varepsilon_i$ with $\mathbb{P}(\varepsilon_i | X_i) = 0$,*

$$\sup_{i \in \{1, ..., n\}} \mathbb{P}_i\left(\varepsilon_i^2 | X_i\right) \leq \sigma^2$$

for some known σ and that $\|f\|_\infty \leq L$ for some known $L > 0$. If we take $S_j = \{j\}$ for any $j \in \{1, ..., m\}$, assumption CRA is satisfied, with

$$r(\{j\}, \varepsilon) = \frac{4\left(1 + \log \frac{2m}{\varepsilon}\right)}{n} \left[\frac{1}{n} \sum_{i=1}^n f_j^2(X_i) Y_i^2 + L^2 + \sigma^2 \right].$$

Remark 2.1. It is also shown in [1] that we are allowed to take

$$\{f_1, ..., f_m\} = \{K(X_1, .), ..., K(X_n, .)\}$$

for some function $\mathcal{X}^2 \to \mathbb{R}$, this being also true in the random design case, but we have to take

$$r(\{j\}, \varepsilon) = \frac{4\left(1 + \log \frac{4m}{\varepsilon}\right)}{n} \left[\frac{1}{n} \sum_{i=1}^n f_j^2(X_i) Y_i^2 + L^2 + \sigma^2 \right].$$

Lemma 2.2. *Let us assume that $P = \mathbb{P}_1 \otimes ... \otimes \mathbb{P}_n$ and that X_1, ..., X_n are deterministic. Let us assume that there is a $K > 0$ such that $\mathbb{P}_i(|Y_i| \leq K) = 1$ for any i. If we take $S_j = \{j\}$ for any $j \in \{1, ..., m\}$, assumption CRA is satisfied with*

$$r(\{j\}, \varepsilon) = \frac{8K^2\left(1 + \log \frac{2m}{\varepsilon}\right)}{n}.$$

A bound in the transductive case is also given in [1].

2.2. **General description of the algorithm.** We propose the following iterative algorithm. Let us choose a confidence level $\varepsilon > 0$ and a distance on $\mathcal{X}$, say $\delta(.,.)$.

- Step 0. Choose $\hat{\alpha}^0 = (0, ..., 0) \in \mathbb{R}^m$. Choose $\varepsilon \in [0, 1]$.
- General Step (k). Choose $N(k) \leq M$ and indices $(j_1^{(k)}, ..., j_N^{(k)}) \in \{1, ..., M\}^{N(k)}$ and put:

$$\hat{\alpha}^k \in \arg \min_{\alpha \in \bigcap_{\ell=1}^{N(k)} \mathcal{CR}(j_\ell^{(k)}, \varepsilon)} \delta\left(\alpha, \hat{\alpha}^{k-1}\right).$$

This algorithm is motivated by the following result.

Theorem 2.3. *When the CRA assumption is satisfied we have:*

$$P\left[\forall k \in \mathbb{N}, \quad \delta\left(\hat{\alpha}^k, \overline{\alpha}\right) \leq \delta\left(\hat{\alpha}^{k-1}, \overline{\alpha}\right) \leq ... \leq \delta\left(\hat{\alpha}^0, \overline{\alpha}\right) \right] \geq 1 - \varepsilon.$$

Moreover, if $\delta(x, x') = \|x - x'\|_X$ then

$$\hat{\alpha}^k = \Pi^X_{\bigcap_{\ell=1}^{N(k)} \mathcal{CR}(j_\ell^{(k)}, \varepsilon)}\left(\hat{\alpha}^{k-1}\right)$$

and we have the following:

$$P\left[\forall k \in \mathbb{N}, \quad R(\hat{\alpha}^k) \leq R(\hat{\alpha}^0) - \sum_{j=1}^{k} \left\|\hat{\alpha}^j - \hat{\alpha}^{j-1}\right\|_X^2\right] \geq 1-\varepsilon.$$

Proof. Let us assume that

$$\forall j \in \{1,...,M\}, \quad \left\|\overline{\alpha}_{S_j} - \hat{\alpha}_{S_j}\right\|_X \leq r(S_j,\varepsilon).$$

This is true with probability at least $1-\varepsilon$ according to assumption CRA. In this case we have seen that

$$\overline{\alpha} \in \bigcap_{\ell=1}^{N(k)} \mathcal{CR}(j_\ell^{(k)},\varepsilon)$$

that is a closed convex region, and so, by definition, $\delta\left(\hat{\alpha}^k,\overline{\alpha}\right) \leq \delta\left(\hat{\alpha}^{k-1},\overline{\alpha}\right)$ for any $k \in \mathbb{N}$. If δ is the distance associated with the norm $\|.\|_X$, let us choose $k \in \mathbb{N}$,

$$\begin{aligned} R(\hat{\alpha}^k) - R(\overline{\alpha}) = \left\|\hat{\alpha}^k - \overline{\alpha}\right\|_X^2 &= \left\|\Pi^X_{\bigcap_{\ell=1}^{N(k)} \mathcal{CR}(j_\ell^{(k)},\varepsilon)}\left(\hat{\alpha}^{k-1}\right) - \overline{\alpha}\right\|_X^2 \\ &\leq \left\|\hat{\alpha}^{k-1} - \overline{\alpha}\right\|_X^2 - \left\|\Pi^X_{\bigcap_{\ell=1}^{N(k)} \mathcal{CR}(j_\ell^{(k)},\varepsilon)}\left(\hat{\alpha}^{k-1}\right) - \hat{\alpha}^{k-1}\right\|_X^2 \\ &= R(\hat{\alpha}^{k-1}) - R(\overline{\alpha}) - \left\|\hat{\alpha}^k - \hat{\alpha}^{k-1}\right\|_X^2. \end{aligned}$$

A recurrence ends the proof. □

We choose as our estimator $\hat{\alpha} = \hat{\alpha}^k$ for some step $k \in \mathbb{N}$; the choice of the stopping step k will depend of the particular choices of the projections and is detailed in what follows.

3. Particular cases and oracle inequalities

We study some particular cases depending on the choice of the distance $\delta(.,.)$ and on the sets we are to project on.

Roughly, Iterative Feature Selection (at least as introduced in [1]) and LASSO corresponds to the choice $\delta(x,x') = \|x-x'\|_X$, and are studied first, together with their grouped variables generalizations.

Dantzig selector corresponds to the choice $\delta(x,x') = \|x-x'\|_1$ the ℓ_1 distance, is is studied in a second time, and can also be generalized to grouped variables selection.

Finally, the new Correlation Selector corresponds to a new choice for δ.

We just give an additionnal notation.

Definition 3.1. *Let us put, for any $j \in \{1,...,m\}$:*

$$\tilde{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i f_j(X_i).$$

3.1. The LASSO.

We first look at the case where $S_j = \{j\}$ for any $j \in \{1,...,m\}$ (and so $M=m$). In this case, we only use submodels of dimension 1.

Here, we use only one step where we project 0 onto the intersection of all the confidence regions and so we obtain:

$$\hat{\alpha}_{LASSO} = \hat{\alpha}^1 = \Pi^X_{\bigcap_{\ell=1}^{m} \mathcal{CR}(\ell,\varepsilon)}(0).$$

Note that we have:

$$\hat{\alpha}_{S_j} = \hat{\alpha}_{\{j\}} = (0,...,0,\tilde{\alpha}_j,0,...,0)$$

with the $\tilde{\alpha}_j$ in j-th position (see Definition 3.1), and that:

$$\mathcal{CR}(j,\varepsilon) = \left\{ \alpha = (\alpha_1, ..., \alpha_m) \in \mathbb{R}^m, \quad \tilde{\alpha}_j - r(\{j\},\varepsilon) \leq \langle \alpha, e_j \rangle_X \leq \tilde{\alpha}_j + r(\{j\},\varepsilon) \right\}.$$

The optimization program to obtain $\hat{\alpha}_{LASSO}$ is given by:

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,...,\alpha_m)\in\mathbb{R}^m} \|\alpha\|_X^2 \\ \\ \text{s. t.} \quad \alpha \in \bigcap_{\ell=1}^m \mathcal{CR}(\ell,\varepsilon) \end{cases}$$

and so:

$$\begin{cases} \arg\min_{\alpha\in\mathbb{R}^m} \|\alpha\|_X^2 \\ \\ \text{s. t.} \quad \forall j \in \{1, ..., m\}, \quad \left|\langle \alpha, e_j \rangle_X - \tilde{\alpha}_j\right| \leq \sqrt{r(\{j\},\varepsilon)} \end{cases} \tag{3.1}$$

Proposition 3.1. *Every solution of the program*

$$\arg\min_{\alpha\in\mathbb{R}^m} \left\{ \|\alpha\|_X^2 - 2\sum_{j=1}^m \alpha_j \tilde{\alpha}_j + 2\sum_{j=1}^m \sqrt{r(\{j\},\varepsilon)}\,|\alpha_j| \right\} \tag{3.2}$$

satisfies Program 3.1. In the case of a deterministic design, Program 3.2 is equivalent to:

$$\arg\min_{\alpha\in\mathbb{R}^m} \left\{ r(\alpha) + 2\sum_{j=1}^m \sqrt{r(\{j\},\varepsilon)}\,|\alpha_j| \right\}.$$

The proof is given in the end of the paper, in the section dedicated to proofs (more precisely Subsection 6.1 page 14).

Note that, if $r(\{j\},\varepsilon)$ does not depend on j, this is exactly the formulation of the original LASSO algorithm as introduced by Tibshirani [13]. An explicit algorithm to obtain the projection is given by Efron, Hastie, Johnstone and Tibshirani [10].

However, in the cases where $r(\{j\},\varepsilon)$ is not constant, the difference with the LASSO algorithm is the following: coordinates that are more difficult to estimate (because the confidence interval is larger) are more penalized.

Moreover, note that the program 3.2 gives a form different of the usual LASSO program for the cases where we do not use the empirical norm.

3.2. **Generalization: the Group LASSO.** Here we choose general subsets S_1, ..., $S_M \subset \{1, ..., N\}$.

As in the LASSO algorithm we only use one step where we project 0 onto the intersection of all the confidence regions,

$$\hat{\alpha}_{GLASSO} = \hat{\alpha}^1 = \Pi^X_{\bigcap_{\ell=1}^M \mathcal{CR}(\ell,\varepsilon)}(0).$$

The optimization program to obtain $\hat{\alpha}_{GLASSO}$ is given by

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,...,\alpha_m)\in\mathbb{R}^m} \|\alpha\|_X^2 \\ \\ \text{s. t.} \quad \forall j \in \{1, ..., M\}, \quad \left\| \Pi_{\mathcal{M}^X_{S_j}}(\alpha) - \hat{\alpha}_j \right\|_X \leq \sqrt{r(S_j,\varepsilon)}. \end{cases}$$

In the case of the empirical norm, this program is equivalent to the following:

$$\arg\min_{\alpha\in\mathbb{R}^m} \left\{ r(\alpha) + \sum_{j=1}^M \sqrt{r(S_j,\varepsilon)}\left\|\Pi_{\mathcal{M}_j}\alpha\right\|_X \right\},$$

that is a generalization of the Group LASSO algorithm defined by Yuan and Lin [16] in the case of orthogonal basis functions and extended by Chesneau and Hebiri [8] to the general case.

3.3. **Iterative Feature Selection.** As in the Group LASSO case, we choose general subsets S_1, ..., $S_m \subset \{1, ..., N\}$.

Moreover, instead of taking the intersection of every confidence region, we project on each of them iteratively. So the algorithm is the following:

$$\hat{\alpha}^0 = (0, ..., 0)$$

and at each step k we choose a $j(k) \in \{1, ..., m\}$ and

$$\hat{\alpha}^k = \Pi^X_{\mathcal{CR}(j(k),\varepsilon)} \left(\hat{\alpha}^{k-1}\right).$$

We choose a stoping step $\hat{k}$ and put

$$\hat{\alpha}_{IFS} = \hat{\alpha}^{\hat{k}}.$$

In the case where, as in the LASSO, we actually have $S_j = \{j\}$ for any j, this is exactly the Iterative Feature Selection algorithm that was introduced in Alquier [1], with the choice of $j(k)$:

$$j(k) = \arg\max_j \left\| \hat{\alpha}^{k-1} - \Pi^X_{\mathcal{CR}(j,\varepsilon)} \left(\hat{\alpha}^{k-1}\right) \right\|_X,$$

and the suggestion to take as a stoping step

$$\hat{k} = \inf \left\{ k \in \mathbb{N}^*, \quad \left\| \hat{\alpha}^k - \hat{\alpha}^{k-1} \right\|_X \leq \kappa \right\}$$

for some small $\kappa > 0$. In [1] is also given the explicit computation of every step of this algorithm.

3.4. **The Dantzig selector, and generalization to Group Dantzig selector.** The Dantzig selector is based on a change of distance δ. We choose

$$\delta(\alpha, \alpha') = \|\alpha - \alpha'\|_1 = \sum_{j=1}^m |\alpha_j - \alpha'_j|.$$

As is the LASSO case, we take $S_j = \{j\}$ and we make only one projection onto the intersection of every confidence region:

$$\hat{\alpha}_{DANTZIG} \in \arg \min_{\alpha \in \bigcap_{\ell=1}^m \mathcal{CR}(j,\varepsilon)} \|\alpha\|_1$$

and so $\hat{\alpha}_{DANTZIG}$ is the solution of the program:

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,...,\alpha_m)\in\mathbb{R}^m} \sum_{j=1}^m |\alpha_j| \\ \\ \text{s. t.} \quad \forall j \in \{1, ..., m\}, \quad \left| \langle \alpha, e_j \rangle_X - \tilde{\alpha}_j \right| \leq \sqrt{r(\{j\}, \varepsilon)}. \end{cases}$$

In the case where $r(\{j\}, \varepsilon)$ does not depend on j, this program is exactly the one proposed by Candes and Tao [6] to introduce the Dantzig selector.

Note that here again we can propose several changes in the algorithm: taking general S_j we obtain a Group Dantzig selector. Moreover, we can as in Iterative Feature Selection project successively onto the various confidence regions instead of projecting once onto their intersection.

3.5. **Oracle Inequalities for LASSO and Iterative Feature Selection.**

Theorem 3.2. *Let us assume that the CRA assumption is satified. Let us assume that we took $S_1 = \{1\}$, $S_2 = \{1, 2\}$, ..., $S_m = \{1, 2, ..., m\}$, that we use the Iterative Feature Selection estimator with stopping step m:*

$$\hat{\alpha}_{IFS} = \Pi^X_{\mathcal{CR}(m,\varepsilon)} ... \Pi^X_{\mathcal{CR}(1,\varepsilon)} 0.$$

Then we have:

(3.3)

$$P\left\{R\left(\hat{\alpha}_{IFS}\right) \leq R\left(\overline{\alpha}\right) + \inf_{1\leq j\leq m}\left[R\left(\overline{\alpha}_{\{1,...,j\}}\right) - R\left(\overline{\alpha}\right) + 4r(\{1,...,j\},\varepsilon)\right]\right\} \geq 1-\varepsilon.$$

The proof is given in Subsection 6.2 page 15.

Remark 3.1. This result is interesting in the case where the functions $f_1,...,f_m$ and P are such that there is a $\beta > 0$ and a constant $C > 0$ such that for any $j \in \{1,...,m\}$, we have:

$$\left\|\overline{\alpha}_{\{1,...,j\}} - \overline{\alpha}\right\|_X \leq Cj^{-\beta}.$$

This is a regularity assumption with an order on the family of functions, such an assumption is satisfied by functions in a Sobolev space, see Tsybakov [14] and the references therein for example. In this case, if we assume moreover that there is a $k > 0$ such that:

$$r(\{1,...,j\},\varepsilon) \leq \frac{jk\log\frac{m}{\varepsilon}}{n}$$

then we have:

(3.4)

$$P\left\{R\left(\hat{\alpha}\right) \leq R\left(\overline{\alpha}\right) + (2\beta+1)\,C^{\frac{1}{2\beta+1}}\left(\frac{2k\log\frac{m}{\varepsilon}}{\beta n}\right)^{\frac{2\beta}{2\beta+1}} + \left(\frac{4k\log\frac{m}{\varepsilon}}{n}\right)\right\} \geq 1-\varepsilon$$

(see also Subsection 6.2 for the proof).

Theorem 3.3. *Let us assume that the CRA assumption is satisfied. Let us assume that the functions $f_1,...,f_m$ are orthogonal with respect to $\langle .,. \rangle_X$. Let us assume that we choose $S_j = \{j\}$ for any $j \in \{1,...,m\}$, and*

$$\hat{\alpha}_{IFS} = \Pi^X_{\mathcal{CR}(m,\varepsilon)}...\Pi^X_{\mathcal{CR}(1,\varepsilon)}0.$$

Then

$$\hat{\alpha}_{LASSO} = \hat{\alpha}_{IFS} = \hat{\alpha}_{DANTZIG} = \sum_{j=1}^{m} sgn\left(\tilde{\alpha}_j\right)\left(\left|\tilde{\alpha}_j\right| - \sqrt{r(j,\varepsilon)}\right)_+ e_j$$

that is a soft-thresholded estimator, and

$$P\left\{R\left(\hat{\alpha}_{LASSO}\right) \leq R\left(\overline{\alpha}\right) + \inf_{S\subset\{1,...,m\}}\left[R\left(\overline{\alpha}_S\right) - R\left(\overline{\alpha}\right) + 4\sum_{j\in S} r(\{j\},\varepsilon)\right]\right\} \geq 1-\varepsilon.$$

For the proof, see Subsection 6.3 page 16.

Remark 3.2. We say that the general regularity assumption with order $\beta > 0$ and constant $C > 0$ if, for any $j \in \{1,...,m\}$, we have:

$$\inf_{\substack{S\subset\{1,...,m\}\\ |S|\leq j}} \left\|\overline{\alpha}_S - \overline{\alpha}\right\|_X \leq Cj^{-\beta}.$$

This is the kind of regularity satisfied by functions in weak Besov spaces, see Cohen [9] and the references therein, with f_j being wavelets. If the general regularity assumption is satisfied with regularity $\beta > 0$ and constant $C > 0$ and if there is a $k > 0$ such that

$$r(\{j\},\varepsilon) \leq \frac{k\log\frac{m}{\varepsilon}}{n},$$

then we have:

$$P\left\{R\left(\hat{\alpha}\right) \leq R\left(\overline{\alpha}\right) + (2\beta+1)\,C^{\frac{1}{2\beta+1}}\left(\frac{2k\log\frac{m}{\varepsilon}}{\beta n}\right)^{\frac{2\beta}{2\beta+1}} + \left(\frac{4k\log\frac{m}{\varepsilon}}{n}\right)\right\} \geq 1-\varepsilon.$$

3.6. **A new estimator: the Correlation Selector.** The idea of the Correlation Selector is to use the following norm:

$$\|\alpha\|_{csel} = \sum_{j=1}^{m} \langle e_j, \alpha\rangle_X^2 .$$

As is the LASSO case, we take $S_j = \{j\}$ and we make only one projection onto the intersection of every confidence region:

$$\hat{\alpha}_{csel} \in \arg \min_{\alpha \in \bigcap_{\ell=1}^{m} \mathcal{CR}(j,\varepsilon)} \|\alpha\|_{csel}$$

and so $\hat{\alpha}_{csel}$ is a solution of the program:

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,...,\alpha_m)\in\mathbb{R}^m} \sum_{j=1}^{m} \langle e_j, \alpha\rangle_X^2 \\ \\ \text{s. t.} \quad \forall j \in \{1,...,m\}, \quad \left|\langle \alpha, e_j\rangle_X - \tilde{\alpha}_j\right| \leq \sqrt{r(\{j\},\varepsilon)}. \end{cases}$$

This program can be solved for every $u_j = \langle e_j, \alpha\rangle_X$ individually: each of them is solution of

$$\begin{cases} \arg\min_u |u|^2 \\ \\ \text{s. t.} \quad \forall j \in \{1,...,m\}, \quad |u - \tilde{\alpha}_j| \leq \sqrt{r(\{j\},\varepsilon)}. \end{cases}$$

As a consequence,

$$u_j = \langle e_j, \hat{\alpha}_{csel}\rangle = sgn\left(\tilde{\alpha}_j\right)\left(|\tilde{\alpha}_j| - \sqrt{r(j,\varepsilon)}\right)_+$$

that does not depend on p. Note that u_j is a thresholded estimation of the correlation between Y and $f_j(X)$, this is what suggested the name "Correlation Selector". Let us put U the column vector that contains the u_j for $j \in \{1,...,m\}$ and M the matrix $(\langle e_i, e_j\rangle_X)_{i,j}$, then $\hat{\alpha}_{csel}$ is just a solution of $M\hat{\alpha}_{csel} = U$.

Remark 3.3. Note that the Correlation Selector has no reason to be sparse, however, the vector $M\hat{\alpha}_{csel}$ is sparse.

Finally we mention that here again, we can define some variants using grouped variables or iterative projections.

3.7. **Oracle inequality for the Correlation Selector.**

Theorem 3.4. *We have:*

$$P\left[\|\hat{\alpha}_{csel} - \overline{\alpha}\|_{csel}^2 \leq \inf_{S\subset\{1,...,m\}} \left(\sum_{j\notin S} \langle \overline{\alpha}, e_j\rangle_X^2 + 4\sum_{j\in S} r(\{j\},\varepsilon)\right)\right] \geq 1-\varepsilon.$$

Moreover, if we assume that there is a $D > 0$ such that for any $\alpha \in \mathcal{E}_m$, $\|\alpha\|_X \geq D\|\alpha\|$ where

$$\mathcal{E}_m = \left\{\alpha \in \mathbb{R}^m, \quad \langle \overline{\alpha}, e_j\rangle_X = 0 \Rightarrow \langle \alpha, e_j\rangle = 0\right\}$$

then we have:

$$P\left[R(\hat{\alpha}_{csel}) - R(\overline{\alpha}) \leq \frac{1}{D^2} \inf_{S\subset\{1,...,m\}} \left(\sum_{j\notin S} \langle \overline{\alpha}, e_j\rangle_X^2 + 4\sum_{j\in S} r(\{j\},\varepsilon)\right)\right] \geq 1-\varepsilon.$$

The proof can be found in Subection 6.4 page 17.

Remark 3.4. Note that if there is a $\overline{S}$ such that for any $j \notin \overline{S}$, $\langle \overline{\alpha}, e_j\rangle_X = 0$ and if $r(\{j\},\varepsilon) = k\log(m/\varepsilon)/n$ then we have:

$$P\left[\|\hat{\alpha}_{csel} - \overline{\alpha}\|_{csel,2}^2 \leq \frac{4k|\overline{S}|\log\frac{m}{\varepsilon}}{n}\right] \geq 1-\varepsilon,$$

and if moreover for any $\alpha \in \mathcal{E}_m$, $\|\alpha\|_X \geq D\|\alpha\|$ then

$$P\left[R(\hat{\alpha}_{csel}) - R(\overline{\alpha}) \leq \frac{4k|\overline{S}|\log\frac{m}{\varepsilon}}{D^2 n}\right] \geq 1-\varepsilon.$$

4. Numerical simulations

4.1. **Motivation.** We compare here LASSO, Iterative Feature Selection and Correlation Selector on a toy example, introduced by Tibshirani [13]. We also compare their performances to the ordinary least square (OLS) estimate as a benchmark. Note that we will not propose a very fine choice for the $r(\{j\},\varepsilon)$. The idea of these simulations is not to identify a good choice for the penalization in practice. The idea is to observe the similarity and differences between different order in projections in our general algorithm, using the same confidence regions.

4.2. **Description of the experiments.** The model defined by Tibshirani [13] is the following. We have:

$$\forall i \in \{1,...,20\}, \quad Y_i = \langle \beta, X_i\rangle + \varepsilon_i$$

with $X_i \in \mathcal{X} = \mathbb{R}^8$, $\beta \in \mathbb{R}^8$ and the ε_i are i. i. d. from a gaussian distribution with mean 0 and standard deviation σ.

The X_i's are i. i. d. too, and each X_i comes from a gaussian distribution with mean $(0,...,0)$ and with variance-covariance matrix:

$$\Sigma(\rho) = \left(\rho^{|i-j|}\right)_{\substack{i\in\{1,...,8\}\\ j\in\{1,...,8\}}}$$

for $\rho \in [0,1[$.

We will use the three particular values for β taken by Tibshirani [13]:

$$\beta^1 = (3, 1.5, 0, 0, 2, 0, 0, 0),$$
$$\beta^2 = (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5),$$
$$\beta^3 = (5, 0, 0, 0, 0, 0, 0, 0),$$

corresponding to a "sparse" situation (β^1), a "non-sparse" situation (β^2) and a "very sparse" situation (β^3).

We use two values for σ: 1 (the "low noise case") and 3 (the "noisy case").

Finally, we use two values for ρ: 0.1 ("weakly correlated variables") and 0.5 ("highly correlated variables").

We run each example (corresponding to a given value of β, σ and ρ) 250 times. We use the software R [12] for simulations. We implement Iterative Feature Selection as described in subsection 3.3 page 8, and the Correlation Selector, while using the standard OLS estimate and the LASSO estimator given by the LARS package described in [10]. The choice:

$$r(\{j\},\varepsilon) = \frac{\sigma}{3}\sqrt{\frac{\log m}{n}} = \frac{\sigma}{3}\sqrt{\frac{\log 8}{20}}$$

was not motivated by theoretical considerations but seems to perform well in practice.

4.3. **Results and comments.** The results are reported in Table 1.

The following remarks can easily be made in view of the results:

- both methods based on projection on random confidence regions using the norm $\|.\|_X$ clearly outperforms the OLS in the sparse cases, moreover they present the advantage of giving sparse estimates;

TABLE 1. Results of the Simulations. For each possible combination of β, σ and ρ, we report in a column the mean empirical loss over the 250 simulations, the standard deviation of this quantity over the simulations and finally the mean number of non-zero coefficients in the estimate, this for each estimate, ordinary least square (OLS), LASSO, Iterative Feature Selection (IFS) and Correlation Selector (C-SEL).

β	σ	ρ	OLS	LASSO	IFS	C-SEL
	3	0.5	**3.67**	**1.64**	**1.56**	**3.65**
β^1			1.84	1.25	1.20	1.96
(sparse)			8	4.64	4.62	8
	1	0.5	**0.40**	**0.29**	**0.36**	**0.44**
			0.22	0.19	0.23	0.23
			8	5.42	5.70	8
	3	0.1	**3.75**	**2.72**	**2.85**	**3.44**
			1.86	1.50	1.58	1.72
			8	5.70	5.66	8
	1	0.1	**0.40**	**0.30**	**0.31**	**0.43**
			0.19	0.19	0.19	0.20
			8	5.92	5.96	8
	3	0.5	**3.54**	**3.36**	**4.90**	**3.98**
β^2			1.82	1.64	1.58	1.85
(non sparse)			8	7.08	6.57	8
	1	0.5	**0.41**	**0.54**	**0.84**	**0.47**
			0.21	0.93	0.36	0.24
			8	7.94	7.89	8
	3	0.1	**3.78**	**3.82**	**4.50**	**4.01**
			1.78	1.51	1.59	1.86
			8	7.06	7.03	8
	1	0.1	**0.40**	**0.42**	**0.71**	**0.48**
			0.20	0.29	0.32	0.22
			8	7.98	7.98	8
	3	0.5	**3.55**	**1.65**	**1.59**	**3.42**
β^3			1.79	1.28	1.27	1.74
(very sparse)			8	4.48	4.49	8
	1	0.5	**0.40**	**0.18**	**0.17**	**0.46**
			0.21	0.14	0.14	0.25
			8	4.46	4.48	8
	3	0.1	**3.46**	**1.69**	**1.62**	**3.00**
			1.74	1.29	1.18	1.45
			8	4.92	4.92	8
	1	0.1	**0.40**	**0.20**	**0.19**	**0.44**
			0.20	0.14	0.14	0.24
			8	4.98	4.91	8

- in the non-sparse case, the OLS performs generally better than the other methods, but LASSO is very close, it is known that a better choice for the value $r(\{j\}, \varepsilon)$ would lead to a better result (see Tibshirani [13]);

- LASSO seems to be the best method on the whole set of experiments. In every case, it is never the worst method, and always performs almost as well as the best method;
- in the "sparse case" (β^1), note that IFS and LASSO are very close for the small value of ρ. This is coherent with the previous theory, see Theorem 3.3 page 9;
- IFS gives very bad results in the non-sparse case (β^2), but is the best method in the sparse case (β^3). This last point tends to indicate that different situations should lead to a different choice for the confidence regions we are to project on. However, theoretical results leading on that choice are missing;
- the Correlation Selector performs badly on the whole set of experiments. However, note that the good performances for LASSO and IFS occurs for sparse values of β, and the previous theory ensures good performances for C-SEL when $M\beta'$ is sparse where M is the covariance matrix of the X_i. In other words, two experiments where favorable to LASSO ans IFS, but there was no experiment favorable to C-SEL.

In order to illustrate this last point, we build a new experiment favorable to C-SEL. Note that we have

$$Y_i = \beta' X_i + \varepsilon_i = (M\beta)' M^{-1} X_i + \varepsilon_i \tag{4.1}$$

where M is the correlation matrix of the X_i. Let us put $\tilde{X}_i = M^{-1}X_i$ and $\tilde{\beta} = M\beta$, we have the following linear model:

$$Y_i = \tilde{\beta}' \tilde{X}_i + \varepsilon_i. \tag{4.2}$$

The sparsity of β gives advantage to the LASSO for estimating β in Model 4.1, it also gives an advantage to C-SEL for estimating $\tilde{\beta}$ in Model 4.2 (according to Remark 3.3 page 10).

We run again the experiments with $\beta = \beta^3$ and this time we try to estimate $\tilde{\beta}$ instead of β (so we act as if we had observed $\tilde{X}_i$ and not X_i).

Results are given in Table 2.

The correlation selector clearly outperforms the other methods in this case.

5. Conclusion

This paper provides a simple interpretation of well-known algorithms of statistical learning theory in terms of orthogonal projections on confidence regions. This very intuitive approach provides a very simple way to prove oracle inequalities.

Also note that this approach can be easily extended into general statistical problems with quadratic loss: in our paper [2], the Iterative Feature Selection method is generalized to the density estimation with quadratic loss problem, leading to a proposition of a LASSO-like programm for density estimation, that have also been proposed and studied by Bunea, Tsybakov and Wegkamp [4] under the name SPADES.

Simulations shows that methods based on confidence regions clearly outperforms the OLS estimate in most examples. However, theoretical results leading the statistician to a particular choice for the order of the successive projections are still missing. Moreover, more accurate values for $r(\{j\}, \varepsilon)$ would be needed in practice. More complete experimental studies are coming on this topic, including also various forms of group selectors, in joint works with Thomas Willer, Mohamed Hebiri and Christophe Chesneau.

TABLE 2. Results for the estimation of $\tilde{\beta}$. As previously, for each possible combination of σ and ρ, we report in a column the mean empirical loss over the 250 simulations, the standard deviation of this quantity over the simulations and finally the mean number of non-zero coefficients in the estimate, this for each estimate: OLS, LASSO, IFS and C-SEL.

β	σ	ρ	OLS	LASSO	IFS	C-SEL
β^1 (sparse)	3	0.5	**3.64** 1.99 8	**4.83** 2.53 5.98	**5.12** 2.64 6.05	**2.41** 1.92 8
	1	0.5	**0.41** 0.21 8	**1.09** 1.72 7.11	**0.92** 0.48 7.40	**0.26** 0.19 8
	3	0.1	**3.65** 1.71 8	**3.71** 1.96 6.25	**3.72** 1.99 6.28	**2.09** 1.40 8
	1	0.1	**0.40** 0.20 8	**0.47** 0.25 7.35	**0.55** 0.16 7.38	**0.23** 0.27 8

6. PROOFS

6.1. **Proof of Proposition 3.1.**

Proof. Let us remember program 3.1:

$$\begin{cases} \max_{\alpha\in\mathbb{R}^m} -\|\alpha\|_X^2 \\ \\ \text{s. t.} \quad \forall j\in\{1,...,m\}, \quad \left|\langle\alpha, e_j\rangle_X - \tilde{\alpha}_j\right| \leq \sqrt{r(\{j\},\varepsilon)}. \end{cases} \tag{6.1}$$

Let us write the lagrangian of this program:

$$\begin{aligned} \mathcal{L}(\alpha,\lambda,\mu) = &-\sum_i\sum_j \alpha_i\alpha_j \langle e_i, e_j\rangle_X \\ &+\sum_j \lambda_j \left[\sum_i \alpha_i \langle e_i, e_j\rangle_X - \tilde{\alpha}_j - \sqrt{r(\{j\},\varepsilon)}\right] \\ &+\sum_j \mu_j \left[-\sum_i \alpha_i \langle e_i, e_j\rangle_X + \tilde{\alpha}_j - \sqrt{r(\{j\},\varepsilon)}\right] \end{aligned}$$

with , for any j, $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j\mu_j = 0$. Any solution (α^*) of Program 3.1 must satisfy, for any j,

$$0 = \frac{\partial\mathcal{L}}{\partial\alpha_j}(\alpha^*,\lambda,\mu) = -2\sum_i \alpha_i^* \langle e_i, e_j\rangle_X + \sum_i (\lambda_i - \mu_i)\langle e_i, e_j\rangle_X ,$$

so for any j,

$$\sum_i \left\langle \frac{1}{2}(\lambda_i - \mu_i) e_i, e_j \right\rangle_X = \langle \alpha^*, e_j\rangle_X . \tag{6.2}$$

Note that this also implies that:

$$\begin{aligned}\|\alpha^*\|_X &= \left\langle \sum_i \alpha_i^* e_i, \sum_j \alpha_j^* e_j \right\rangle_X = \sum_i \alpha_i^* \left\langle e_i, \sum_j \alpha_j^* e_j \right\rangle_X \\ &= \sum_i \alpha_i^* \left\langle e_i, \sum_j \frac{1}{2}(\lambda_j - \mu_j) e_j \right\rangle_X = \sum_j \frac{1}{2}(\lambda_j - \mu_j) \left\langle \sum_i \alpha_i^* e_i, e_j \right\rangle_X \\ &= \sum_j \sum_i \frac{1}{2}(\lambda_j - \mu_j) \frac{1}{2}(\lambda_i - \mu_i) \langle e_i, e_j \rangle_X .\end{aligned}$$

Using these relations, the lagrangian may be written:

$$\begin{aligned}\mathcal{L}(\alpha^*, \lambda, \mu) = &- \sum_i \sum_j \frac{1}{2}(\lambda_i - \mu_i) \frac{1}{2}(\lambda_j - \mu_j) \langle e_i, e_j \rangle_X \\ &+ \sum_i \sum_j \frac{1}{2}(\lambda_i - \mu_i)(\lambda_j - \mu_j) \langle e_i, e_j \rangle_X \\ &- \sum_j (\lambda_j - \mu_j)\tilde{\alpha}_j + \sum_j (\lambda_j + \mu_j)\sqrt{r(\{j\}, \varepsilon)}.\end{aligned}$$

Note that the condition $\lambda_j \geq 0$, $\mu_j \geq 0$ and $\lambda_j \mu_j = 0$ means that there is a $\gamma_j \in \mathbb{R}$ such that $\gamma_j = 2(\lambda_j - \mu_j)$, $|\gamma_j| = 2(\lambda_j + \mu_j)$, and so $\mu_j = (\gamma_j/2)_-$ and $\lambda_j = (\gamma_j/2)_+$. Let also γ denote the vector which j-th component is exactly γ_j, we obtain:

$$\mathcal{L}(\alpha^*, \lambda, \mu) = \|\gamma\|_X^2 - 2\sum_j \gamma_j \tilde{\alpha}_j + 2\sum_j |\gamma_j| \sqrt{r(\{j\}, \varepsilon)}$$

that is maximal with respect to the λ_j and μ_j, so with respect to γ. So γ is the solution of Program 3.2.

Now, note that Equation 6.2 ensures that any solution α^* of Program 3.1 satisfies:

$$\left\langle \sum_i \gamma_i e_i, e_j \right\rangle_X = \langle \alpha^*, e_j \rangle_X .$$

We can easily see that $\alpha^* = \gamma$ is a possible solution.

If $\|.\|_X$ is the empirical norm we obtain:

$$\begin{aligned}\|\gamma\|_X^2 - 2\sum_{j=1}^m \gamma_j \tilde{\alpha}_j &= \frac{1}{n}\sum_{i=1}^n \left[\sum_{j=1}^m \gamma_j f_j(X_i)\right]^2 - 2\frac{1}{n}\sum_{i=1}^n Y_i \left[\sum_{j=1}^m \gamma_j f_j(X_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^n \left[Y_i - \sum_{j=1}^m \gamma_j f_j(X_i)\right]^2 - \frac{1}{n}\sum_{i=1}^n Y_i^2 \\ &= r(\gamma) - \frac{1}{n}\sum_{i=1}^n Y_i^2.\end{aligned}$$

□

6.2. **Proof of Theorem 3.2.**

Proof. Let us assume that the event:

$$\left\{ \forall j \in \{1, ..., M\}, \quad \left\| \overline{\alpha}_{S_j} - \hat{\alpha}_{S_j} \right\|_X \leq r(S_j, \varepsilon) \right\}$$

is satisfied (this is true with probability at least $1-\varepsilon$ thanks to assumption CRA). For any $j \in \{1,...,m\}$ we have:

$$\hat{\alpha}^j = \Pi^X_{\mathcal{CR}(j,\varepsilon)} ... \Pi^X_{\mathcal{CR}(2,\varepsilon)} \Pi^X_{\mathcal{CR}(1,\varepsilon)}(0),$$

and $\hat{\alpha} = \hat{\alpha}^m$. It is evident that $\hat{\alpha}^j \in \mathcal{CR}(j,\varepsilon)$. Moreover, we can prove that $\hat{\alpha}^j \in \mathcal{M}_{\{1,...,j\}}$ by recurrence. So we have:

$$\left\| \hat{\alpha}^j - \overline{\alpha}_{\{1,...,j\}} \right\|_X^2 \leq 4r(S_j, \varepsilon),$$

which means that:

$$R\left(\hat{\alpha}^j\right) \leq R\left(\overline{\alpha}_{\{1,...,j\}}\right) + 4r(S_j,\varepsilon).$$

Now, Theorem 2.3 ensures that:

$$R\left(\hat{\alpha}\right) = R\left(\hat{\alpha}^m\right) \leq R\left(\hat{\alpha}^j\right),$$

this proves Equation 3.3:

$$R\left(\hat{\alpha}\right) \leq R\left(\overline{\alpha}\right) + \inf_{1\leq j\leq m} \left[R\left(\overline{\alpha}_{\{1,...,j\}}\right) - R\left(\overline{\alpha}\right) + 4r(S_j,\varepsilon) \right].$$

□

For Equation 3.4 note that:

$$R\left(\hat{\alpha}\right) \leq R\left(\overline{\alpha}\right) + \inf_{1\leq j\leq m} \left[Cj^{-2\beta} + \frac{4jk\log\frac{m}{\varepsilon}}{n} \right],$$

and take:

$$j = \left\lfloor \left(\frac{\beta C n}{2k\log\frac{m}{\varepsilon}} \right)^{\frac{1}{2\beta+1}} \right\rfloor + 1$$

to conclude.

6.3. **Proof of Theorem 3.3.**

Proof. In the case of orthogonality, we have $\|.\|_X = \|.\|$ the euclidian norm. So $\hat{\alpha}_{LASSO}$ satisfies, according to its definition:

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,...,\alpha_m)\in\mathbb{R}^m} \sum_{j=1}^m \alpha_j^2 \\ \\ \text{s. t.} \quad \forall j \in \{1,...,m\}, \quad |\alpha_j - \tilde{\alpha}_j| \leq \sqrt{r(\{j\},\varepsilon)} \end{cases}$$

while $\hat{\alpha}_{DANTZIG}$ satisfies:

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,...,\alpha_m)\in\mathbb{R}^m} \sum_{j=1}^m |\alpha_j| \\ \\ \text{s. t.} \quad \forall j \in \{1,...,m\}, \quad |\alpha_j - \tilde{\alpha}_j| \leq \sqrt{r(\{j\},\varepsilon)}. \end{cases}$$

We can easily solve both problem by an individual optimization on each α_j and obtain the same solution

$$\alpha_j^* = sgn\left(\tilde{\alpha}_j\right)\left(|\tilde{\alpha}_j| - \sqrt{r(j,\varepsilon)} \right)_+.$$

For $\hat{\alpha}_{IFS}$ just note that in the case of orthogonality, sequential projections on each $\mathcal{CR}(j,\varepsilon)$ leads to the same result than the projection on their intersection, so $\hat{\alpha}_{IFS} = \hat{\alpha}_{LASSO}$. Then, let us choose $S \subset \{1,...,m\}$ and remark that

$$\begin{aligned} R\left(\hat{\alpha}_{LASSO}\right) - R\left(\overline{\alpha}\right) &= \left\|\hat{\alpha}_{LASSO} - \overline{\alpha}\right\|_X^2 = \left\|\hat{\alpha}_{LASSO} - \overline{\alpha}\right\|^2 \\ &= \sum_{j=1}^m \left\langle \hat{\alpha}_{LASSO} - \overline{\alpha}, e_j \right\rangle^2 \\ &= \sum_{j\in S} \left\langle \hat{\alpha}_{LASSO} - \overline{\alpha}, e_j \right\rangle^2 + \sum_{j\notin S} \left\langle \hat{\alpha}_{LASSO} - \overline{\alpha}, e_j \right\rangle^2. \end{aligned}$$

Now, with assumption CRA, with probability $1-\varepsilon$, for any j, $\overline{\alpha}$ satisfies the same constraint than the LASSO estimator so

$$|\langle \overline{\alpha}, e_j\rangle| \leq \sqrt{r(\{j\},\varepsilon)}$$

and so

$$|\langle \hat{\alpha}_{LASSO} - \overline{\alpha}, e_j\rangle| = \left|\alpha_j^* - \langle \overline{\alpha}, e_j\rangle\right| \leq \left|\alpha_j^* - \tilde{\alpha}_j\right| + |\langle \overline{\alpha}, e_j\rangle - \tilde{\alpha}_j| \leq 2\sqrt{r(\{j\},\varepsilon)}.$$

Moreover, let us remark that α_j^* is the number with the smallest absolute value satisfying this contraint, so

$$\left|\alpha_j^* - \langle \overline{\alpha}, e_j\rangle\right| \leq \max\left(\left|\alpha_j^*\right|, |\langle \overline{\alpha}, e_j\rangle|\right) \leq |\langle \overline{\alpha}, e_j\rangle|.$$

So we can conclude

$$\begin{aligned} R(\hat{\alpha}_{LASSO}) - R(\overline{\alpha}) \leq \sum_{j\in S} 4r(\{j\},\varepsilon) + \sum_{j\notin S} \langle \overline{\alpha}, e_j\rangle^2 &= 4\sum_{j\in S} r(\{j\},\varepsilon) + \|\overline{\alpha} - \overline{\alpha}_S\|^2 \\ &= 4\sum_{j\in S} r(\{j\},\varepsilon) + R(\overline{\alpha}_S) - R(\overline{\alpha}). \end{aligned}$$

□

6.4. **Proof of Theorem 3.4.**

Proof. Note that, for any S:

$$\begin{aligned} \|\hat{\alpha}_{csel} - \overline{\alpha}\|_{csel}^2 = \sum_{j=1}^{m} \langle \hat{\alpha}_{csel} - \overline{\alpha}, e_j\rangle_X^2 & \\ = \sum_{j\in S} \langle \hat{\alpha}_{csel} - \overline{\alpha}, e_j\rangle_X^2 + \sum_{j\notin S} \langle \hat{\alpha}_{csel} - \overline{\alpha}, e_j\rangle_X^2 & . \end{aligned}$$

By the constraint satisfied by $\hat{\alpha}_{csel}$ we have:

$$\langle \hat{\alpha}_{csel} - \overline{\alpha}, e_j\rangle_X^2 \leq 4r(\{j\},\varepsilon).$$

Moreover, we must remember that $u_j = \langle \hat{\alpha}_{csel}, e_j\rangle_X$ satisfies the program

$$\begin{cases} \arg\min_u |u| \\ \\ \text{s. t.} \quad \forall j \in \{1,...,m\}, \quad |u - \tilde{\alpha}_j| \leq \sqrt{r(\{j\},\varepsilon)}, \end{cases}$$

that is also satisfied by $\langle \overline{\alpha}, e_j\rangle_X$, so $|u_j| \leq |\langle \overline{\alpha}, e_j\rangle|$ and so

$$|u_j - \langle \overline{\alpha}, e_j\rangle| \leq \max(|u_j|, |\langle \overline{\alpha}, e_j\rangle|) = |\langle \overline{\alpha}, e_j\rangle|$$

and so we have the relation:

$$\langle \hat{\alpha}_{csel} - \overline{\alpha}, e_j\rangle_X^2 \leq \langle \overline{\alpha}, e_j\rangle_X^2.$$

So we obtain:

$$\|\hat{\alpha}_{csel} - \overline{\alpha}\|_{csel}^2 \leq \sum_{j\in S} 4r(\{j\},\varepsilon) + \sum_{j\notin S} \langle \overline{\alpha}, e_j\rangle_X^2$$

This proves the first inequality of the theorem. For the second one, we just have to prove that $M(\hat{\alpha}_{psel} - \overline{\alpha}) \in \mathcal{E}_m$. But this is trivial because of the relation:

$$\langle M(\hat{\alpha}_{csel} - \overline{\alpha}), e_j\rangle^2 = \langle \hat{\alpha}_{csel} - \overline{\alpha}, e_j\rangle_X^2 \leq \langle \overline{\alpha}, e_j\rangle_X^2.$$

□

References

[1] Alquier, P. Iterative feature selection in regression esimation. Preprint Laboratoire de Probabilits et Modles Alatoires (submitted to the Annales de l'IHP, accepted in 2007), 2005.

[2] Alquier, P. Density estimation with quadratic loss: A confidence intervals method. Preprint Laboratoire de Probabilits et Modles Alatoires (submitted to ESAIM PS, accepted in 2007), 2006.

[3] Bickel, P. J., Ritov, Y., and Tsybakov, A. Simultaneous analysis of lasso and dantzig selector. Preprint Submitted to the Annals of Statistics, 2007.

[4] Bunea, F., Tsybakov, A., and Wegkamp, M. Sparse density estimation with ℓ_1 penalties. In *Proceedings of 20th Annual Conference on Learning Theory (COLT 2007)* (2007), Springer-Verlag, pp. 530–543.

[5] Bunea, F., Tsybakov, A., and Wegkamp, M. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics 1* (2007), 169–194.

[6] Candes, E., and Tao, T. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics 35* (2007).

[7] Catoni, O. A pac-bayesian approach to adaptative classification. Preprint Laboratoire de Probabilits et Modles Alatoires, 2003.

[8] Chesneau, C., and Hebiri, M. Some theoretical results on the grouped variable lasso. Preprint Laboratoire de Probabilits et Modles Alatoires (submitted), 2007.

[9] Cohen, A. *Handbook of Numerical Analysis*, vol. 7. North-Holland, Amsterdam, 2000.

[10] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics 32*, 2 (2004), 407–499.

[11] Panchenko, D. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability 31*, 4 (2003), 2068–2081.

[12] R. A language and environment for statistical computing. By the R development core team, Vienna, Austria. URL: http://www.R-project.org/, 2004.

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

[14] Tsybakov, A. *Introduction l'Estimation Non-Paramtrique*. Springer, 2004.

[15] Vapnik, V. *The nature of statistical learning theory*. Springer, 1998.

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

CREST, and, Laboratoire de Probabilits et Modles Alatoires (Universit Paris 7), 175, rue du Chevaleret, 75252 Paris Cedex 05, France.

URL: `http://www.crest.fr/pageperso/alquier/alquier.htm`

E-mail address: `alquier@ensae.fr`