LASSO and Iterative Feature Selection: Oracle Inequalities and Numerical Performances

Pierre Alquier

LASSO AND ITERATIVE FEATURE SELECTION: ORACLE INEQUALITIES AND NUMERICAL PERFORMANCES

PIERRE ALQUIER

Abstract. We propose a general family of algorithms for regression estimation with quadratic loss. Our algorithms is able to select relevant functions into a large dictionary. We prove that some algorithms that have already been studied (LASSO, by Tibshirani [15], Iterative Feature Selection, in our paper [1], among others) belong to our family. We prove oracle-type inequalities in some particular cases, and compare numerical performances of LASSO and Iterative Feature Selection on a toy example.

Contents

1. Introduction

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

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

Date: October 24, 2007.
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.

and

$$\left((X_1, Y_1), ..., (X_n, Y_n)\right)$$

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:i

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

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 two 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 $\|.\|_n$).

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 $\|.\|_X$; $\Pi_{\mathcal{C}}(.)$ will denote the orthogonal projection on $\mathcal{C}$ with respect to the euclidian norm $\|.\|$.*

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 [15] and some variants or generalization like LARS by Efron, Hastie, Johnstone and Tibshirani [11], the Dantzig selector by Candes and Tao [6] and the Group LASSO by Yuan and Lin [17], or Iterative Feature Selection in our paper [1]. This paper proposes a general algorithm that contains LASSO and Iterative Feature Selection as a particular case.

A paper by Butucea, 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 Butucea and al. is writen is 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.

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 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 and explicit some particular cases (LASSO, Iterative Feature Selection and Group LASSO) 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 give some oracle inequalities under some hypothesis on the dictionary of functions.

Finally, section 4 is dedicated to simulations: we compare ordinary least square (OLS), LASSO and Iterative Feature Selection 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.

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 $\varepsilon \in [0,1]$ we have a bound $r(S_j, \varepsilon) \in \mathbb{R}$ such that*

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

Such confidence regions for $\overline{\alpha}_{S_j}$ can be obtained with standard techniques and various hypothesis on the probability P, we refer the reader to our previous work [1] for example.

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

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

We remark that the hypothesis implies that:

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

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 [13] 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 is [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:*

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

2.2. **General description of the algorithm.** Now let us choose $N \leq M$ and indices $(j_1, ..., j_N) \in \{1, ..., M\}^N$, the region:

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

is a closed, convex confidence region for $\overline{\alpha}$.

So we propose the following iterative algorithm.

- 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:

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

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

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

Proof. Let us 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.

2.3. **First particular case: 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:

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

Definition 2.4. *Let us put, for any $j \in \{1, ..., m\}$:*

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

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, 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 $\hat{\alpha}$ is given by:

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

and so:

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

We can write the programm in dual form:

$$\arg\min_{\alpha\in\mathbb{R}^m} \left\{ \|\alpha\|_X^2 - 2\sum_{j=1}^m \alpha_j\tilde{\alpha}_j + 2\sum_{j=1}^m \sqrt{r(\{j\},\varepsilon)}\,|\alpha_j| \right\}. \tag{2.1}$$

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

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

and so the programm is equivalent to:

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

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 [15]. An explicit algorithm to obtain the projection is given by Efron, Hastie, Johnstone and Tibshirani [11].

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 2.1 gives a form different of the usual LASSO programm for the cases where we do not use the empirical norm.

2.4. **Particular case: Iterative Feature Selection.** Here, 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:

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

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

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

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 an estimator:

$$\hat{\alpha} = \hat{\alpha}^{\hat{k}}$$

where

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

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

2.5. **Particular case: Generalization of 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 and so we obtain:

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

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

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

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

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

that is a generalization of the Group LASSO algorithm defined by Yuan and Lin [17] in the case of orthogonal basis functions and extended by Chesneau and Hebiri [8] to the general case.

3. Oracle inequalities in some particular cases

Some particular assumptions about the dictionnary of functions chosen by the statistician allow us to obtain oracle inequalities for some particular order of projection.

3.1. **Order on the dictionary of functions.** In this subsection, we assume that there is an order on the basis function: in some sense, the statistician knows that a function f_j with a small indice $j \in \{1, ..., m\}$ is more likely to be useful for his regression problem than another function $f_{j'}$ with a large indice $j' \in \{1, ..., m\}$.

A usual way to formalize this hypothesis is to make the following regularity assumption.

Definition 3.1. *We say that the ordered regularity assumption with order $\beta > 0$ and constant $C > 0$ is satisfied if, for any $j \in \{1, ..., m\}$, we have:*

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

Remark 3.1. This is a Sobolev-type regularity assumption, see Tsybakov [16] and the references therein for estimation in Sobolev spaces.

Note that this is equivalent to:

$$R\left(\overline{\alpha}_{\{1,...,j\}}\right) - R\left(\overline{\alpha}\right) \leq Cj^{-2\beta}.$$

Let us put $S_1 = \{1\}$, $S_2 = \{1, 2\}$, ..., $S_m = \{1, ..., m\}$ and so $M = m$ and follow the following iterative projection scheme:

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

Then we have the following result.

Theorem 3.1. *Let us assume that the CRA assumtion is satified. Then we have:*
(3.1)
$$P\left\{R\left(\hat{\alpha}\right) \leq R\left(\overline{\alpha}\right) + \inf_{1\leq j\leq m}\left[R\left(\overline{\alpha}_{\{1,...,j\}}\right) - R\left(\overline{\alpha}\right) + 4r(\{1,...,j\},\varepsilon)\right]\right\} \geq 1-\varepsilon.$$

If we assume moreover that there is a $k>0$ such that:

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

and that the ordered regularity assumption is satisfied with regularity β and constant C then we have:
(3.2)
$$P\left\{R\left(\hat{\alpha}\right) \leq R\left(\overline{\alpha}\right) + (2\beta+1)\,C^{\frac{1}{2\beta+1}}\left(\frac{2k\log\frac{m}{\varepsilon}}{\beta n}\right)^{\frac{2\beta}{2\beta+1}} + \left(\frac{4k\log\frac{m}{\varepsilon}}{n}\right)\right\} \geq 1-\varepsilon.$$

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:

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

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 3.1:

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

For Equation 3.2 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:

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

to conclude. □

Remark 3.2. Note that, as soon as $\beta > 1/2$, we reach the minimax rate of convergence up to a $\log m$ term. In general, as β is unknown to the statistician, we propose to take $m = n$ so we miss the optimal rate of convergence by a $\log n$ term. In this case, we propose the following modification (inspired by the block method, see Tsybakov [16] and the references therein for example). We take $S_h = \{1,...,2^h\}$ with $h\in\{1,...,m\}$ and

$$m = \left\lfloor\frac{\log n}{\log 2}\right\rfloor$$

as this choice ensures that $2^{m-1} < n \leq 2^m$ so we obtain:

$$R(\hat{\alpha}) \leq R(\overline{\alpha}) + \inf_{1 \leq h \leq m} \left[R\left(\overline{\alpha}_{\{1,\dots,2^h\}}\right) - R(\overline{\alpha}) + r(S_h, \varepsilon) \right]$$
$$\leq R(\overline{\alpha}) + \inf_{1 \leq h \leq m} \left[2^{-2\beta h} + \frac{2^h 4k \log \frac{\log\left\lfloor \frac{\log n}{\log 2} \right\rfloor}{\varepsilon}}{n} \right]$$

and so we can conclude:

$$R(\hat{\alpha}) \leq R(\overline{\alpha}) + \mathcal{O}\left[\left(\frac{\log \frac{\log n}{\varepsilon}}{n} \right)^{\frac{2\beta}{2\beta+1}} \right],$$

so in this case we reach the minimax rate of convergence up to a $\log \log n$ term.

3.2. **Nearly orthogonal dictionary of functions.**

Definition 3.2. *For a pair $(\delta, D) \in (\mathbb{R}_+)^2$, we say that the condition NO(δ, D) is satisfied if:*

$$\forall \alpha \in \mathbb{R}^m, \quad \delta \|\alpha\| \leq \|\alpha\|_X \leq D \|\alpha\|.$$

Remark 3.3. This condition was given by Kerkyacharian and Picard [12] to study the statistical properties of algorithms generalizing the idea of thresholding. The meaning of condition NO(δ, D) when $\delta \simeq 1$ and $D \simeq 1$ is clear: the norms $\|.\|$ and $\|.\|_X$ have the same behaviour, in other words the dictionary of functions $(f_j)_{j \in \{1,\dots,m\}}$ is "nearly orthogonal". Of course, as we are in a space of finite dimention m, we can always find a small δ (of the order $1/m$) and a large D (of the order m) such that condition NO(δ, D) is satisfied. However, Theorem 3.2 clearly shows that the results are interesting for values of δ and D that does not depend on m.

We will also use the following regularity hypothesis.

Definition 3.3. *We say that the general regularity assumption with order $\beta > 0$ and constant $C > 0$ if, for any $j \in \{1, ..., m\}$, we have:*

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

Remark 3.4. This is the type of regularity assumption used to define weak Besov spaces, see Cohen [9] and the references therein.

We still take $S_1 = \{1\}$, ..., $S_m = \{m\}$, and

$$\hat{\alpha} = \Pi^X_{\bigcap_{j=1}^m c\mathcal{R}(j,\varepsilon)}(0)$$

the LASSO estimator.

Theorem 3.2. *Let us assume that the CRA assumption is satisfied. For any $(\delta, D) \in (\mathbb{R}_+)^2$ such that condition NO(δ, D) is satisfied we have:*

$$P\left\{ R(\hat{\alpha}) \leq R(\overline{\alpha}) + \frac{1}{\delta^2} \inf_{S \subset \{1,\dots,m\}} \left[D^2 \left[R(\overline{\alpha}_S) - R(\overline{\alpha}) \right] + 4 \sum_{j \in S} r(\{j\}, \varepsilon) \right] \right\} \geq 1 - \varepsilon,$$

and so, if moreover 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:

$$P\left\{R\left(\hat{\alpha}\right)\leq R\left(\overline{\alpha}\right)+\frac{\left(2\beta+1\right)C^{\frac{1}{2\beta+1}}}{\delta^{2}}\left(\frac{2Dk\log\frac{m}{\varepsilon}}{\beta n}\right)^{\frac{2\beta}{2\beta+1}}+\frac{1}{\delta^{2}}\left(\frac{4k\log\frac{m}{\varepsilon}}{n}\right)\right\}\geq1-\varepsilon.$$

Proof. In this proof, we adopt the notation α' for the transposed vector of α, and let M denote the matrix:

$$M=\left(\left\langle e_{j},e_{k}\right\rangle_{X}\right)_{(j,k)\in\{1,...,m\}^{2}}=\left(P_{X}\left(f_{j}f_{k}\right)\right)_{(j,k)\in\{1,...,m\}^{2}}.$$

So note that for any $(\alpha,\beta)\in(\mathbb{R}^{m})^{2}$ we have:

$$\left\langle \alpha,\beta\right\rangle =\alpha'\beta$$

and

$$\left\langle \alpha,\beta\right\rangle_{X}=\alpha'M\beta.$$

We have:

$$\begin{aligned}R\left(\hat{\alpha}\right)-R\left(\overline{\alpha}\right)&=\left\|\hat{\alpha}-\overline{\alpha}\right\|_{X}^{2}=\left\|M^{\frac{1}{2}}\left(\hat{\alpha}-\overline{\alpha}\right)\right\|^{2}\\&\leq\frac{1}{\delta^{2}}\left\|M^{\frac{1}{2}}\left(\hat{\alpha}-\overline{\alpha}\right)\right\|_{X}^{2}=\frac{1}{\delta^{2}}\left\|M\left(\hat{\alpha}-\overline{\alpha}\right)\right\|^{2}=\frac{1}{\delta^{2}}\left(\hat{\alpha}-\overline{\alpha}\right)'MM\left(\hat{\alpha}-\overline{\alpha}\right)\\&=\frac{1}{\delta^{2}}Tr\left[\left(\hat{\alpha}-\overline{\alpha}\right)'MM\left(\hat{\alpha}-\overline{\alpha}\right)\right]=\frac{1}{\delta^{2}}Tr\left[M\left(\hat{\alpha}-\overline{\alpha}\right)\left(\hat{\alpha}-\overline{\alpha}\right)'M\right]\\&=\frac{1}{\delta^{2}}\sum_{j=1}^{m}e_{j'}M\left(\hat{\alpha}-\overline{\alpha}\right)\left(\hat{\alpha}-\overline{\alpha}\right)'Me_{j}=\frac{1}{\delta^{2}}\sum_{j=1}^{m}\left\langle \hat{\alpha}-\overline{\alpha},e_{j}\right\rangle_{X}^{2}.\end{aligned}$$

Now, note that the fact that for any $j\in\{1,...,m\}$, $\hat{\alpha}\in\mathcal{CR}(j,\varepsilon)$ implies that:

$$\forall j\in\{1,...,m\},\quad\left\langle \hat{\alpha}-\overline{\alpha},e_{j}\right\rangle_{X}^{2}\leq4r(\{j\},\varepsilon).$$

Moreover, the by the definition of $\hat{\alpha}$ we have:

$$\forall j\in\{1,...,m\},\quad\left\langle \hat{\alpha}-\overline{\alpha},e_{j}\right\rangle_{X}^{2}\leq\left\langle \overline{\alpha},e_{j}\right\rangle_{X}^{2}.$$

So, for any $S\subset\{1,...,m\}$ we have:

$$R\left(\hat{\alpha}\right)-R\left(\overline{\alpha}\right)\leq\frac{1}{\delta^{2}}\left[4\sum_{j\in S}r(\{j\},\varepsilon)+\sum_{j\notin S}\left\langle \overline{\alpha},e_{j}\right\rangle_{X}^{2}\right].$$

So, in order to proove Theorem 3.2 we just have to upper bound the second sum, using the same techniques:

$$\begin{aligned}\sum_{j\notin S}\left\langle \overline{\alpha},e_{j}\right\rangle_{X}^{2}&=\sum_{j=1}^{m}\left\langle \overline{\alpha}-\overline{\alpha}_{S},e_{j}\right\rangle_{X}^{2}=\sum_{j=1}^{m}e_{j'}M\left(\overline{\alpha}-\overline{\alpha}_{S}\right)\left(\overline{\alpha}-\overline{\alpha}_{S}\right)'Me_{j}\\&=Tr\left[M\left(\overline{\alpha}-\overline{\alpha}_{S}\right)\left(\overline{\alpha}-\overline{\alpha}_{S}\right)'M\right]=Tr\left[\left(\overline{\alpha}-\overline{\alpha}_{S}\right)'MM\left(\overline{\alpha}-\overline{\alpha}_{S}\right)\right]\\&=\left\|M\left(\overline{\alpha}-\overline{\alpha}_{S}\right)\right\|^{2}=\left\|M^{\frac{1}{2}}\left(\overline{\alpha}-\overline{\alpha}_{S}\right)\right\|_{X}^{2}\leq D^{2}\left\|M^{\frac{1}{2}}\left(\overline{\alpha}-\overline{\alpha}_{S}\right)\right\|^{2}\\&=D^{2}\left\|\overline{\alpha}-\overline{\alpha}_{S}\right\|_{X}^{2}=D^{2}\left[R\left(\overline{\alpha}_{S}\right)-R\left(\overline{\alpha}\right)\right].\end{aligned}$$

□

Remark 3.5. A particularly interesting case is the orthogonal case, where condition NO(1, 1) is satisfied. In this case:

$$\hat{\alpha}=\Pi_{\mathcal{CR}(m,\varepsilon)}^{X}\Pi_{\mathcal{CR}(m-1,\varepsilon)}^{X}...\Pi_{\mathcal{CR}(1,\varepsilon)}^{X}(0)=\Pi_{\bigcap_{j=1}^{m}\mathcal{CR}(j,\varepsilon)}^{X}(0)$$

so in this case, LASSO and Iterative Feature Selection (with $\kappa = 0$) are equivalent and actually we have:

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

that is a soft-thresholded estimator. See our previous work on Iterative Feature Selection [1] for a proof. Soft-thresholding is now a standard way to deal with selection of functions in an orthogonal family, see for example the seminal paper by Donoho and Johnstone [10] in the case of a wavelet basis. The bound just becomes:

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

4. Numerical simulations

4.1. **Motivation.** We compare here LASSO and Iterative Feature Selection on a toy example, introduced by Tibshirani [15]. 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, unsing the same confidence regions.

4.2. **Description of the experiments.** The model defined by Tibshirani [15] 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 device 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:

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

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

We will use the three particular values for β taken by Tibshirani [15]:

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

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 [14] for simulations. We implement Iterative Feature Selection as described in subsection 2.4 page 6, while using the standard OLS estimate and the LASSO estimator given by the LARS package described in [11]. 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.

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 and Iterative Feature Selection (IFS).

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

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

- both methods based on projection on random confidence regions clearly outperforms the OLS in the sparse cases, moreover they present the advantage of giving sparse estimates;

- 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 [15]);
- 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 remark 3.5 page 10;
- 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.

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.

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] Donoho, D., and Johnstone, I. Ideal spatial adaptation by wavelet shrinkage. *Biometrika 81* (1994), 425–455.

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

[12] Kerkyacharian, G., and Picard, D. Thresholding in learning theory. Preprint Laboratoire de Probabilits et Modles Alatoires, 2005.

[13] Panchenko, D. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability 31*, 4 (2003), 2068–2081.
[14] Team, R. D. C. R: A language and environment for statistical computing. Vienna, Austria. URL: http://www.R-project.org/, 2004.
[15] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 1 (1996), 267–288.
[16] Tsybakov, A. *Introduction l'Estimation Non-Paramtrique*. Springer, 2004.
[17] 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`