Electronic Journal of Statistics
Vol. 2 (2008) 1129–1152
ISSN: 1935-7524
DOI: 10.1214/08-EJS288

LASSO, Iterative Feature Selection and the Correlation Selector: Oracle inequalities and numerical performances

Pierre Alquier*

Laboratoire de Probabilités et Modèles Aléatoires (Université Paris 7)
175, rue du Chevaleret
75252 Paris Cedex 05, France

CREST, LS
3, avenue Pierre Larousse
92240 Malakoff, France
e-mail: alquier@ensae.fr
url: http://alquier.ensae.net/

Abstract: We propose a general family of algorithms for regression estimation with quadratic loss, on the basis of geometrical considerations. These algorithms are able to select relevant functions into a large dictionary. We prove that a lot of methods that have already been studied for this task (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.

AMS 2000 subject classifications: Primary 62G08; secondary 62J07, 62G15, 68T05.
Keywords and phrases: Regression estimation, statistical learning, confidence regions, shrinkage and thresholding methods, LASSO.

Received August 2008.

Contents

*I Would like to thank Professors Olivier Catoni, Alexandre Tsybakov, Mohamed Hebiri and Joseph Salmon as well as the anonymous referees for useful remarks.

1. Introduction

1.1. The regression problem

In this paper, we study the linear regression problem: we observe n pairs (X_i, Y_i) with $Y_i = f(X_i) + \varepsilon_i$ for a noise $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ to be specified later.

The idea is that the statistician is given (or chooses) a dictionary of functions: $(f_1, \dots, f_m)$, with possibly $m > n$, and he wants to build a "good" estimation of f of the form $\alpha_1 f_1 + \cdots + \alpha_m f_m$.

Actually, we have to precise two things: what is the distribution of the pairs (X_i, Y_i), and what is the criterion for a "good" estimation. We are going to consider two cases.

1.2. Deterministic and random design

1.2.1. Deterministic design case

In this case the values $X_1, \dots, X_n$ are deterministic, and the ε_i are i. i. d. according to some distribution $\mathbb{P}$ with $\mathbb{E}_{\varepsilon\sim\mathbb{P}}(\varepsilon) = 0$ and $\mathbb{E}_{\varepsilon\sim\mathbb{P}}(\varepsilon^2) < \infty$. In this case, the distance between f and $\alpha_1 f_1 + \cdots + \alpha_m f_m$ will be measured in terms of the so-called empirical norm.

Definition 1.1. *For any* $\alpha = (\alpha_1, \dots, \alpha_m) \in \mathbb{R}^m$ *and* $\alpha' = (\alpha'_1, \dots, \alpha'_m) \in \mathbb{R}^m$ *we put*

$$\|\alpha - \alpha'\|_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{j=1}^{m} \alpha_j f_j(X_i) - \sum_{j=1}^{m} \alpha'_j f_j(X_i) \right]^2$$

and

$$\overline{\alpha}_n \in \arg\min_{\alpha \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \left[f(X_i) - \sum_{j=1}^{m} \alpha_j f_j(X_i) \right]^2.$$

1.2.2. Random design case

In this case, we assume that the pairs (X_i, Y_i) are i. i. d. according to some distribution $\mathbb{P}$, that the marginal distribution of every X_i is $\mathbb{P}_X$, and that we still have $\mathbb{E}_{(X,Y)\sim\mathbb{P}}(\varepsilon) = 0$ and $\mathbb{E}_{(X,Y)\sim\mathbb{P}}(\varepsilon^2) < \infty$. The distance will be measured by the $\mathfrak{L}^2$ distance with respect to $\mathbb{P}_X$.

Definition 1.2. *For any* $\alpha, \alpha' \in \mathbb{R}^m$ *we put*

$$\|\alpha - \alpha'\|_X^2 = \mathbb{E}_{X\sim\mathbb{P}_X} \left\{ \left[\sum_{j=1}^{m} \alpha_j f_j(X) - \sum_{j=1}^{m} \alpha'_j f_j(X) \right]^2 \right\}$$

and

$$\overline{\alpha}_X \in \arg\min_{\alpha \in \mathbb{R}^m} \mathbb{E}_{X\sim\mathbb{P}_X} \left\{ \left[f(X) - \sum_{j=1}^{m} \alpha_j f_j(X) \right]^2 \right\}.$$

Moreover, we make the following restrictive hypothesis: the statistician knows $\mathbb{P}_X$.

1.3. General notations

Now, we assume that we are in one of the two cases defined previously. However, as the results we want to state are the same in both settings, we introduce the following notation.

Definition 1.3. *We introduce the general norm*

$$\|\alpha - \alpha'\|_{GN}$$

that is simply $\|\alpha - \alpha'\|_n$ *if we are in the deterministic design case and and* $\|\alpha - \alpha'\|_X$ *if we are in the random design case. Moreover, we will let* $\overline{\alpha}$ *denote* $\overline{\alpha}_n$ *or* $\overline{\alpha}_X$ *according to the case.*

In any case, we let P *denote the distribution of the sample* $(X_i, Y_i)_{i=1,\dots,n}$.

In order so simplify the notations, we assume that the functions f_j of the dictionary are normalized, in the sense that $\frac{1}{n}\sum_{i=1}^{n} f_j^2(X_i) = 1$ if we are in the deterministic design case and that $\mathbb{E}_{X\sim\mathbb{P}_X}\left[f_j(X)\right]^2 = 1$ if we are in the random design case. Note that this could be simply written in terms of the general norm: if we put $e_1 = (1, 0, \ldots, 0), \ldots, e_m = (0, \ldots, 0, 1)$ the canonical basis of $\mathbb{R}^m$, we just have to assume that for any $j \in \{1, \ldots, m\}$, $\|e_j\|_{GN} = 1$.

Finally, let us mention that $\langle ., . \rangle_{GN}$ will denote the scalar product associated to the norm $\|.\|_{GN}$ while we will use the notation $\|.\|$ for the euclidian norm in $\mathbb{R}^m$ and $\langle ., . \rangle$ for the associated scalar product.

1.4. Previous works and organization of the paper

The aim of this paper is to propose a method to estimate the real regression function (say f) on the basis of the dictionary $(f_1, \ldots, f_m)$, that have good performances even if $m > n$.

Recently, a lot of algorithms have been proposed for that purpose, let's cite among others the bridge regression by Frank and Friedman [14], and a particular case of bridge regression called LASSO by Tibshirani [19], some variants or generalization like LARS by Efron, Hastie, Johnstone and Tibshirani [13], the Dantzig selector by Candes and Tao [9] and the Group LASSO by Bakin [3], Yuan and Lin [21] and Chesneau and Hebiri [11] or iterative algorithms like Iterative Feature Selection in our paper [2] or greedy algorithms in Barron, Cohen, Dahmen and DeVore [4]. This paper proposes a general method that contains LASSO, Dantzig selector and Iterative Feature Selection as a particular case.

Note that in the case where m/n is small, we can use the ordinary least square estimate. The risk of this estimator is roughly in m/n. But when $m/n > 1$, this estimator isn't even properly defined. The idea of all the mentioned works is the following: if there is a "small" vector space $F \subset \mathbb{R}^m$ such that $\overline{\alpha} \in F$, one could build a constrained estimator with a risk in $\dim(F)/n$. But can we obtain such a result if F is unknown? For example, a lot of papers study the *sparsity* of $\overline{\alpha}$, this means that F is the span of a few e_j, or, in other words, that $\overline{\alpha}$ have only a small number (say p) of non-zero coordinates: an estimator that selects automatically p relevant coordinates and achieving a risk close to p/n is said to satisfy a "sparsity oracle inequality". A paper, by Bickel, Ritov and Tsybakov [5] gives sparsity oracle inequalities for the LASSO and the Dantzig selector in the case of the deterministic design. Another paper by Bunea, Tsybakov and Wegkamp [8] gives sparsity oracle inequalities for the LASSO. This paper is written in a more general context than ours: random design with *unknown* distribution (in the case of a random design, remember that our method require the knowledge of the distribution of the design). However, the main results require the assumption $\|f_j\|_\infty \leq L$ for some given L, what is not necessary in our paper, and prevents the use of popular basis of functions like wavelets. This is due to the use of Hoeffding's inequality in the technical parts of the paper.

Our paper uses a geometric point of view. This allows to build a general method of estimation and to obtain simple sparsity oracle inequalities for the obtained estimator, in both deterministic design case and random design with known distribution. It uses a (Bernstein's type) deviation inequality proved in a previous work [2] that is sharper than Hoeffding's inequality, and so gets 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 functions, for example the case where $m = n$ and

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

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

In Section 2, we give the general form for our algorithm under a particular assumption, Assumption **(CRA)**, that says 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 (IFS), 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, we prove some oracle inequalities for the obtained estimators: roughly, LASSO, Dantzig Selector and IFS performs well when the vector $\overline{\alpha}$ is sparse (which means that a lot of its coordinates, $\langle \overline{\alpha}, e_j \rangle = \overline{\alpha}_j$ are equal to zero) or approximately sparse (a lot of coordinates are nearly equal to zero), while the Correlation Selector performs well when a lot of $\langle \overline{\alpha}, e_j \rangle_{GN}$ are almost equal to zero (in the deterministic design case, $\langle \overline{\alpha}, e_j \rangle_{GN} = \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} f_j(X_i)Y_i)$ while in the random design case, $\langle \overline{\alpha}, e_j \rangle_{GN} = \mathbb{E}(f_j(X)Y)$, so in any case, this quantity is a measure of the correlation between the variable Y and the j-th function in the dictionary). So, intuitively, the Correlation Selector gives good results when most of the functions in the dictionary have weak correlation with Y, but we expect that altogether these functions can bring a good prediction for Y.

In order to prove oracle inequalities, some types of orthogonality (or approximate orthogonality, in some sense) are required on the dictionary of functions. Our results are the following: under orthogonality on the dictionary of functions, and *using only general properties of our family of estimators*, we have a sparse oracle inequality. Under an approximate orthogonality condition taken from Bickel, Ritov and Tsybakov [5], the result can be extended for the LASSO and the Dantzig selector (with a proof taken from [5]). Some remarks by Huang, Cheang and Barron [15] show that these results can be extended to IFS with a slight modification of the estimator. Finally, the central result for the Correlation Selector does not require any hypothesis on the dictionary of functions but concerns a measure of the risk that is not natural, we obtain a result on the risk measured by $\|.\|_{GN}$ under an assumption very close to the one in [5] - here again, the proof uses only general properties of our family of estimators.

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 5), Section 6 is dedicated to some proofs.

2. General projection algorithms

2.1. Additional notations and hypothesis

Definition 2.1. *Let $\mathcal{C}$ be a closed, convex subset of $\mathbb{R}^d$. We let $\Pi_{\mathcal{C}}^{GN}(.)$ denote the orthogonal projection on $\mathcal{C}$ with respect to the norm $\|.\|_{GN}$:*

$$\Pi_{\mathcal{C}}^{GN}(\alpha) = \arg\min_{\beta\in\mathcal{C}} \|\alpha - \beta\|_{GN}.$$

For a generic distance δ, we will use the notation $\Pi_{\mathcal{C}}^{\delta}(.)$ for the orthogonal projection on $\mathcal{C}$ with respect to δ.

We put, for every $j \in \{1,\dots,m\}$:

$$\mathcal{M}_j = \left\{\alpha \in \mathbb{R}^m, \quad \ell \neq j \Rightarrow \alpha_\ell = 0\right\} = \left\{\alpha e_j, \alpha \in \mathbb{R}\right\}.$$

Definition 2.2. *We put, for every $j \in \{1,\dots,m\}$:*

$$\overline{\alpha}^j = \arg\min_{\alpha\in\mathcal{M}_j} \|\overline{\alpha} - \alpha e_j\|_{GN} = \Pi_{\mathcal{M}_j}^{GN}(\overline{\alpha}).$$

Moreover let us put:

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

Remark 2.1. In the deterministic design case ($\|.\|_{GN} = \|.\|_n$) we have

$$\overline{\alpha}^j = \left[\frac{1}{n}\sum_{i=1}^{n} f_j(X_i)f(X_i)\right]e_j$$

and in the random design case we have

$$\overline{\alpha}^j = \mathbb{E}_{X\sim\mathbb{P}_X}\left[f_j(X)f(X)\right]e_j$$

so in any case, $\hat{\alpha}^j$ is an estimator of $\overline{\alpha}^j$.

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

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

In our previous work [2] we examined different hypothesis on the probability P such that this hypothesis is satisfied. For example, using inequalities by Catoni [10] and Panchenko [17] we proved the following results.

Lemma 2.1. *Let us assume that* $\|f\|_\infty \leq L$ *for some known* L*. Let us assume that* $\mathbb{E}_{\mathbb{P}}(\varepsilon^2) \leq \sigma^2$ *for some known* $\sigma^2 < \infty$*. Then 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 2.2. It is also shown in [2] that we are allowed to take

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

for some function $K : \mathcal{X}^2 \to \mathbb{R}$ (this allows for $f(x)$ a kernel estimator of the form $\sum_{i=1}^n \alpha_i K(X_i,x)$), even in the random design case, but we have to take

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

in this case.

Lemma 2.2. *Let us assume that there is a* $K > 0$ *such that* $\mathbb{P}(|Y| \leq K) = 1$*. Then Assumption* **(CRA)** *is satisfied with*

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

Definition 2.3. *When* **(CRA)** *is satisfied, we define, for any* $\varepsilon > 0$ *and* $j \in \{1,\dots,m\}$*, the random set*

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

This can easily be interpreted: Assumption **(CRA)** says that there is a confidence region for $\overline{\alpha}^j$ in the small model $\mathcal{M}_j$; $\mathcal{CR}(j,\varepsilon)$ is the set of all vectors falling in this confidence region when they are orthogonally projected on $\mathcal{M}_j$.

We remark that the hypothesis implies that

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

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

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

This algorithm is motivated by the following result.

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

$$P\Big[\forall k \in \mathbb{N}, \quad \delta\big(\hat{\alpha}(k), \overline{\alpha}\big) \leq \delta\big(\hat{\alpha}(k-1), \overline{\alpha}\big) \leq \cdots \leq \delta\big(\hat{\alpha}(0), \overline{\alpha}\big)\Big] \geq 1 - \varepsilon. \quad (2.1)$$

So, our algorithm builds a sequence of $\hat{\alpha}(k)$ that gets closer to $\overline{\alpha}$ (according to δ) at every step. Moreover, if $\delta(x, x') = \|x - x'\|_{GN}$ then

$$\hat{\alpha}(k) = \Pi^{GN}_{\bigcap_{\ell=1}^{N(k)} \mathcal{CR}(j_\ell^{(k)}, \varepsilon)}\big(\hat{\alpha}(k-1)\big)$$

and we have the following:

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

Proof. Let us assume that

$$\forall j \in \{1, \ldots, M\}, \quad \left\|\overline{\alpha}^j - \hat{\alpha}^j\right\|^2_{GN} \leq r(S_j, \varepsilon).$$

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 $\|.\|_{GN}$, let us choose $k \in \mathbb{N}$,

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

A recurrence ends the proof. □

Remark 2.3. We choose our estimator $\hat{\alpha} = \hat{\alpha}(k)$ for some step $k \in \mathbb{N}$; the choice of the stopping step k will depend on the particular choices of the projections and is detailed in what follows. But remark that **there is no bias-variance balance involved in the choice of** k as Theorem 2.3 shows that overfitting is not possible for large values of k.

3. Particular cases and oracle inequalities

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

Roughly, LASSO and Iterative Feature Selection (at least as introduced in [2]) correspond to the choice $\delta(\alpha,\alpha') = \|\alpha - \alpha'\|_{GN}$, and are studied first.

Dantzig selector corresponds to the choice $\delta(\alpha,\alpha') = \|\alpha-\alpha'\|_1$ the ℓ_1 distance, it is studied in a second time.

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

3.1. The LASSO

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

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

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

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

and so:

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

Proposition 3.1. *Every solution of the program*

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

satisfies Program 3.1. Moreover, all the solutions α of Program 3.1 have the same risk value $\|\alpha - \overline{\alpha}\|^2_{GN}$. Finally, in the ***deterministic design case****, Program 3.2 is equivalent to:*

$$\arg\min_{\alpha\in\mathbb{R}^m} \left\{ \frac{1}{n}\sum_{i=1}^n \left[Y_i - \sum_{j=1}^m \alpha_j f_j(X_i)\right]^2 + 2\sum_{j=1}^m \sqrt{r(j,\varepsilon)}\,|\alpha_j| \right\}. \tag{3.3}$$

The proof is given ay the end of the paper (in Subsection 6.1 page 1147).

Note that, if $r(j,\varepsilon)$ does not depend on j, Program 3.3 is exactly one of the formulations of the LASSO estimator studied first by Tibshirani [19]. In the particular **deterministic design case**, this dual representation was already known and introduced by Osborne, Presnell and Turlach [16].

However, in the cases where $r(j,\varepsilon)$ is not constant, the difference with the LASSO algorithm is the following: the harder the coordinates are to be estimated, the more penalized they are.

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

3.2. Iterative Feature Selection (IFS)

As in the LASSO case we use the distance $\delta(\alpha,\beta) = \|\alpha - \beta\|_{GN}$.

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

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

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

$$\hat{\alpha}(k) = \Pi^{GN}_{\mathcal{CR}(j(k),\varepsilon)}\big(\hat{\alpha}(k-1)\big).$$

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

$$\hat{\alpha}^{IFS} = \hat{\alpha}(\hat{k}).$$

This is exactly the Iterative Feature Selection algorithm that was introduced in Alquier [2], with the choice of $j(k)$:

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

and the suggestion to take as a stopping step

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

for some small $\kappa > 0$.

Remark 3.1. In Alquier [2], it is proved that:

$$\hat{\alpha}(k) = \hat{\alpha}(k-1) + sgn\,(\beta_k)\left(|\beta_k| - \sqrt{r(j(k),\varepsilon)}\right)_+ e_{j(k)} \tag{3.4}$$

where

$$\beta_k = \frac{1}{n}\sum_{i=1}^{n} f_j(X_i)\left[Y_i - \sum_{\ell=1}^{m}\hat{\alpha}(k)_\ell f_\ell(X_i)\right].$$

So this algorithm looks quite similar to a greedy algorithm, as it is described by Barron, Cohen, Dahmen and DeVore [4]. Actually, it would be a greedy algorithm if we replace $r(j,\varepsilon)$ by 0 (such a choice is however not possible here): it is a soft-thresholded version of a greedy algorithm. Such greedy algorithms were studied in a recent paper by Huang, Cheang and Barron [15] under the name "penalized greedy algorithm", in the case $\|.\|_{GN} = \|.\|_n$.

Note that in Iterative Feature Selection, every selected feature actually improves the estimator: $\|\hat{\alpha}(k) - \overline{\alpha}\|^2_{GN} \leq \|\hat{\alpha}(k-1) - \overline{\alpha}\|^2_{GN}$ (Equation 2.1).

3.3. The 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 make only one projection onto the intersection of every confidence region:

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

and so $\hat{\alpha}^{DS}$ is the solution of the program:

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

In the case where $r(j,\varepsilon)$ does not depend on j, and where $\|.\|_{GN} = \|.\|_n$, this program is exactly the one proposed by Candes and Tao [9] when they introduced the Dantzig selector.

3.4. Oracle Inequalities for the LASSO, the Dantzig Selector and IFS

Definition 3.1. *For any $S \subset \{1,\dots,m\}$ let us put*

$$\mathcal{M}_S = \{\alpha \in \mathbb{R}^m, \quad j \notin S \Rightarrow \alpha_j = 0\}$$

and

$$\overline{\alpha}_S = \arg \min_{\alpha \in \mathcal{M}_S} \|\alpha - \overline{\alpha}\|_{GN}.$$

Every $\mathcal{M}_S$ is a submodel of $\mathbb{R}^m$ of dimension $|S|$ and $\overline{\alpha}_S$ is the best approximation of $\overline{\alpha}$ in this submodel.

Theorem 3.2. *Let us assume that assumption* **(CRA)** *is satisfied. Let us assume that the functions $f_1,\dots,f_m$ are orthogonal with respect to $\langle .,. \rangle_{GN}$. In this case the order of the projections in Iterative Feature Selection does not affect the obtained estimator, so we can set*

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

Then

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

is a soft-thresholded estimator, and

$$P\left\{ \left\|\hat{\alpha}^{L} - \overline{\alpha}\right\|^2_{GN} \leq \inf_{S \subset \{1,\dots,m\}} \left[\|\overline{\alpha}_S - \overline{\alpha}\|^2_{GN} + 4 \sum_{j \in S} r(j,\varepsilon) \right] \right\} \geq 1 - \varepsilon.$$

For the proof, see Subsection 6.2 page 1149.

Remark 3.2. We call "general regularity assumption with order $\beta > 0$ and constant $C > 0$":

$$\forall j \in \{1, \dots, m\}, \qquad \inf_{\substack{S \subset \{1, \dots, m\} \\ |S| \leq j}} \|\overline{\alpha}_S - \overline{\alpha}\|_{GN} \leq C j^{-\beta}.$$

This is the kind of regularity satisfied by functions in weak Besov spaces, see Cohen [12] 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:

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

Now, note that the orthogonality assumption is very restrictive. Usual results about LASSO or Dantzig Selector usually involve only approximate orthogonality, see for example Cands and Tao [9], Bunea [6], Bickel, Ritov and Tsybakov [5] and Bunea, Tsybakov and Wegkamp [8], and sparsity (the fact that a lot of the coordinates of $\overline{\alpha}$ are null), as for example the following result, which is a small variant of a result in [8], that is reminded here in order to provide comparison with the results coming later in the paper.

Theorem 3.3 (Variant of Bunea, Tsybakov and Wegkamp [8])**.** *Let us assume that we are in the* ***deterministic design case****, that assumption* **(CRA)** *is satisfied, and that $r(j, \varepsilon) = r(\varepsilon)$ does not depend on j (this is always possible by taking $r(\varepsilon) = \sup_{j \in \{1, \dots, m\}} r(j, \varepsilon)$). Moreover, we assume that there is a constant D such that, for any $\alpha \in \mathcal{F}_m$,*

$$\|\alpha\|_{GN} \geq D \|\alpha\|$$

where

$$\mathcal{F}_m = \left\{ \alpha \in \mathbb{R}^m, \quad \sum_{j : \overline{\alpha}_j = 0} |\alpha_j| \leq 3 \sum_{j : \overline{\alpha}_j \neq 0} |\alpha_j| \right\}.$$

Then

$$P\left\{ \left\|\hat{\alpha}^L - \overline{\alpha}\right\|_{GN}^2 \leq \frac{16}{D^2} \Big| \{j : \overline{\alpha}_j \neq 0\} \Big| r(\varepsilon) \right\} \geq 1 - \varepsilon.$$

The only difference with the original result in [8] is that $r(\varepsilon)$ is given in a general form here, so we are allowed to use different values for $r(\varepsilon)$ depending on the context, see the discussion of Hypothesis **(CRA)** in the beginning of the paper. Similar results are available for the Dantzig selector, see Candes and Tao [9], and Bickel, Ritov and Tsybakov [5].

Remark 3.3. We can wonder how IFS performs when the dictionary is not orthogonal. Actually, the study of penalized greedy algorithm in Huang, Cheang and Barron [15] leads to the following conclusion in the deterministic design case: there are cases where IFS can be really worse than LASSO. However, the authors proposes a modification of the algorithm, called "relaxed penalized greedy algorithm"; if we apply this modification here we obtain

$$\hat{\alpha}(k) = \gamma_k \hat{\alpha}(k-1) + sgn\left(\beta_k\right)\left(\left|\tilde{\beta}_k\right| - \sqrt{r(j(k),\varepsilon)}\right)_+ e_{j(k)}$$

instead of equation 3.4, where

$$\tilde{\beta}_k = \frac{1}{n}\sum_{i=1}^{n} f_j(X_i)\left[Y_i - \gamma_k \sum_{\ell=1}^{m} \hat{\alpha}(k)_\ell f_\ell(X_i)\right],$$

and at each step we have to minimize the empirical least square error with respect to $\gamma_k \in [0,1]$. Such a modification ensures that the estimators given by the k-th step of the algorithm become equivalent to the LASSO when k grows, for more details see [15] (note that the interpretation in terms of confidence regions and the property $\|\hat{\alpha}(k) - \overline{\alpha}\|_{GN}^2 \leq \|\hat{\alpha}(k-1) - \overline{\alpha}\|_{GN}^2$ are lost with this modification).

3.5. A new estimator: the Correlation Selector

The idea of the Correlation Selector is to use

$$\|\alpha\|_{CS} = \sum_{j=1}^{m} \langle e_j, \alpha\rangle_{GN}^2 .$$

We make only one projection onto the intersection of every confidence region:

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

and so $\hat{\alpha}^{CS}$ is a solution of the program:

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

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

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

As a consequence,

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

that does not depend on p. Note that u_j is a thresholded estimator 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 $j \in \{1,\dots,m\}$ and M the matrix $(\langle e_i, e_j\rangle_{GN})_{i,j}$, then $\hat{\alpha}^{CS}$ is just any solution of $\hat{\alpha}^{CS}M = U$.

Remark 3.4. Note that the Correlation Selector has no reason to be sparse, however, the vector $\hat{\alpha}^{CS}M$ is sparse. An interpretation of this fact is given in the next subsection.

3.6. Oracle inequality for the Correlation Selector

Theorem 3.4. *We have:*

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

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

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

then we have:

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

The proof can be found in Subsection 6.3 page 1150.

Remark 3.5. Note that the result on $\left\|\hat{\alpha}^{CS} - \overline{\alpha}\right\|_{CS}$ does not require any assumption on the dictionary of functions. However, this quantity does not have, in general, an interesting interpretation. The result about the quantity of interest, $\left\|\hat{\alpha}^{CS} - \overline{\alpha}\right\|_{GN}^2$, requires that a part of the dictionary is almost orthogonal, this condition is to be compared to the one in Theorem 3.3.

Remark 3.6. Note that if there is a $\overline{S}$ such that for any $j \notin \overline{S}$, $\langle\overline{\alpha}, e_j\rangle_{GN} = 0$ and if $r(j,\varepsilon) = k\log(m/\varepsilon)/n$ then we have:

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

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

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

The condition that for a lot of j, $\langle \overline{\alpha}, e_j \rangle_{GN} = 0$ means that most of the functions in the dictionary are not correlated with Y. In terms of sparsity, it means that the vector $\overline{\alpha}M$ is sparse. So, intuitively, the Correlation Selector will perform well when most of the functions in the dictionary have weak correlation with Y, but we expect that altogether these functions can bring a reasonable prediction for Y.

4. Numerical simulations

4.1. Motivation

We compare here LASSO, Iterative Feature Selection and Correlation Selector on a toy example, introduced by Tibshirani [19]. 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 [19] is the following. We have:

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

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 $(0, \dots, 0)$ and with variance-covariance matrix:

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

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

We will use the three particular values for β taken by Tibshirani [19]:

$$\begin{aligned}
\beta^1 &= (3, 1.5, 0, 0, 2, 0, 0, 0), \\
\beta^2 &= (1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5), \\
\beta^3 &= (5, 0, 0, 0, 0, 0, 0, 0),
\end{aligned}$$

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 [18] for simulations. We implement Iterative Feature Selection as described in subsection 3.2 page 1138, and the Correlation Selector, while using the standard OLS estimate and the LASSO estimator given by the LARS package described in [13]. Note that we use the estimators defined in the **deterministic design case**, this means that we consider $\|.\|_{GN} = \|.\|_n$ (the empirical norm) as our criterion here. The choice:

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

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 $\|.\|_{GN} = \|.\|_n$ clearly outperform 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 [19]);
- 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 3.2 page 1139;
- 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 βM is sparse where M is the covariance matrix of the X_i. In other words, two experiments where favorable to LASSO and 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

$$Y_i = \langle X_i, \beta \rangle + \varepsilon_i = \left\langle X_i M^{-1}, \beta M \right\rangle + \varepsilon_i \tag{4.1}$$

TABLE 1

Results of 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, for 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

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

$$Y_i = \left\langle \tilde{X}_i, \tilde{\beta} \right\rangle + \varepsilon_i. \tag{4.2}$$

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 3.4 page 1142).

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).

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

Results are given in Table 2.

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

5. Conclusion

5.1. Comments on the results of the paper

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 also provides tools to prove oracle inequalities.

Simulations shows that methods based on confidence regions clearly outperforms the OLS estimate in most examples. Actually, the theoretical results and the experiments lead to the following conclusion: in the case where we think that $\overline{\alpha}$ is sparse, that means, if we assume that only a few functions in the dictionary are relevant, we should use the LASSO or the Dantzig Selector (we know that these estimators are almost equivalent since [5]); IFS can be seen as a good algorithmic approximation of the LASSO in the orthogonal case. In the other cases, we should think of another method of approximation (LARS, relaxed greedy algorithm...). When $\overline{\alpha}M$ is sparse, i. e. almost all the functions in the dictionary are uncorrelated with Y, then we the Correlation Selector seems to be a reasonable choice. This is, in some way, the "desperate case", where for example for various reason a practitioner thinks that he has the good set of variables to explain Y, but he realizes that only a few of them are correlated with Y and that methods based on the selection of a small subset of variables (LASSO, ...) leads to unsatisfying results.

5.2. Extentions

First, note that all the results given here in the deterministic design case ($\|.\|_{GN} = \|.\|_n$) and in the random design case ($\|.\|_{GN} = \|.\|_X$) can be extended to another

kind of regression problem: the transductive case, introduced by Vapnik [20]. In this case, we assume that m more pairs $(X_{n+1}, Y_{n+1}),\dots, (X_{n+m}, Y_{n+m})$ are drawn (i. i. d. from $\mathbb{P}$), and that $X_{n+1},\dots, X_{n+m}$ are given to the statistician, whose task is now to predict the missing values $Y_{n+1},\dots, Y_{n+m}$. Here, we can introduce the following criterion

$$\|\alpha - \alpha'\|^2_{trans} = \frac{1}{m}\sum_{i=n+1}^{m}\left[\sum_{j=1}^{m}\alpha_j f_j(X_i) - \sum_{j=1}^{m}\alpha'_j f_j(X_i)\right]^2.$$

In [2], we argue that this case is of considerable interest in practice, and we show that Assumption **(CRA)** can be satisfied in this context. So, the reader can check that all the results in the paper can be extended to the case $\|.\|_{GN} = \|.\|_{trans}$.

Also note that this approach can easily be extended into general statistical problems with quadratic loss: in our paper [1], the Iterative Feature Selection method is generalized to the density estimation with quadratic loss problem, leading to a proposition of a LASSO-like program for density estimation, that have also been proposed and studied by Bunea, Tsybakov and Wegkamp [7] under the name SPADES.

5.3. Future works

Future works on this topic include a general study of the projection into the intersection of the confidence regions

$$\begin{cases} \arg\min_{\alpha=(\alpha_1,\dots,\alpha_m)\in\mathbb{R}^m} \delta(\alpha, 0) \\ \text{s. t.} \quad \forall j \in \{1,\dots,m\}, \quad \left|\langle \alpha, e_j\rangle_{GN} - \tilde{\alpha}_j\right| \leq \sqrt{r(j,\varepsilon)} \end{cases}$$

for a generic distance $\delta(.,.)$.

A generalization to confidence regions defined by grouped variables, that would include the Group LASSO studied by Bakin [3], Yuan and Lin [21] and Chesneau and Hebiri [11] as a particular case is also feasible.

A more complete experimental study, including comparison of various choices for $\delta(.,.)$ and for $r(j,\varepsilon)$ based *on theoretical results or on heuristics* would be of great interest.

6. Proofs

6.1. Proof of Proposition 3.1

Proof. Let us remember program 3.1:

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

Let us write the lagrangian of this program:

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

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,

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

so for any j,

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

Note that this also implies that:

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

Using these relations, the lagrangian may be written:

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

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:

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

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

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

$$\left\langle \sum_i \gamma_i e_i, e_j \right\rangle_{GN} = \langle \alpha^*, e_j \rangle_{GN} .$$

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

In the case where $\|.\|_{GN}$ is the empirical norm $\|.\|_n$ we obtain:

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

□

6.2. *Proof of Theorem 3.2*

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

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

while $\hat{\alpha}^{DS}$ satisfies:

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

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 $\hat{\alpha}^{IFS} = \hat{\alpha}^L$. Then, let us choose $S \subset \{1,\ldots,m\}$ and remark that

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

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

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

and so

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

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

$$\left\|\hat{\alpha}^L - \overline{\alpha}\right\|_{GN}^2 \leq \sum_{j\in S} 4r(j,\varepsilon) + \sum_{j\notin S} \langle \overline{\alpha}, e_j\rangle^2 = 4\sum_{j\in S} r(j,\varepsilon) + \left\|\overline{\alpha} - \overline{\alpha}_S\right\|^2.$$

□

6.3. *Proof of Theorem 3.4*

Proof. Note that, for any S:

$$\left\|\hat{\alpha}^{CS} - \overline{\alpha}\right\|_{CS}^2 = \sum_{j=1}^m \langle \hat{\alpha}_{CS} - \overline{\alpha}, e_j\rangle_{GN}^2$$
$$= \sum_{j\in S} \left\langle \hat{\alpha}^{CS} - \overline{\alpha}, e_j\right\rangle_{GN}^2 + \sum_{j\notin S} \left\langle \hat{\alpha}^{CS} - \overline{\alpha}, e_j\right\rangle_{GN}^2.$$

By the constraint satisfied by $\hat{\alpha}^{CS}$ we have:

$$\left\langle \hat{\alpha}^{CS} - \overline{\alpha}, e_j\right\rangle_{GN}^2 \leq 4r(j,\varepsilon).$$

Moreover, we must remember that $u_j = \left\langle \hat{\alpha}^{CS}, e_j\right\rangle_{GN}$ satisfies the program

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

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

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

and so we have the relation:

$$\left\langle \hat{\alpha}^{CS} - \overline{\alpha}, e_j\right\rangle_{GN}^2 \leq \langle \overline{\alpha}, e_j\rangle_{GN}^2.$$

So we obtain:

$$\left\|\hat{\alpha}^{CS}-\overline{\alpha}\right\|_{CS}^2 \leq \sum_{j\in S} 4r(j,\varepsilon) + \sum_{j\notin S} \langle \overline{\alpha}, e_j\rangle_{GN}^2$$

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

$$\left\langle (\hat{\alpha}^{CS}-\overline{\alpha})M, e_j\right\rangle^2 = \left\langle \hat{\alpha}^{CS}-\overline{\alpha}, e_j\right\rangle_{GN}^2 \leq \langle \overline{\alpha}, e_j\rangle_{GN}^2 .$$

□

References

[1] Alquier, P. Density estimation with quadratic loss: A confidence intervals method. *ESAIM P&S 12* (2008), 438–463. MR2437718

[2] Alquier, P. Iterative feature selection in regression estimation. *Annales de l'Institut Henri Poincaré, Probability and Statistics 44*, 1 (2008), 47–88.

[3] Bakin, S. *Adaptative Regression and Model Selection in Data Mining Problems*. PhD thesis, Australian National University, 1999.

[4] Barron, A., Cohen, A., Dahmen, W., and DeVore, R. Adaptative approximation and learning by greedy algorithms. *The annals of statistics 36*, 1 (2008), 64–94. MR2387964

[5] Bickel, P. J., Ritov, Y., and Tsybakov, A. Simultaneous analysis of lasso and dantzig selector. Annals of Statistics (to appear).

[6] Bunea, F. Honest variable slection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization. preprint available on arxiv (0808.4051), 2008.

[7] 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. MR2397610

[8] Bunea, F., Tsybakov, A., and Wegkamp, M. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics 1* (2007), 169–194. MR2312149

[9] Candes, E., and Tao, T. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics 35* (2007). MR2382644

[10] Catoni, O. A pac-bayesian approach to adaptative classification. Preprint Laboratoire de Probabilités et Modèles Aléatoires, 2003.

[11] Chesneau, C., and Hebiri, M. Some theoretical results on the grouped variable lasso. Preprint Laboratoire de Probabilités et Modèles Aléatoires (submitted), 2007.

[12] Cohen, A. *Handbook of Numerical Analysis*, vol. 7. North-Holland, Amsterdam, 2000. MR1804747

[13] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. Least angle regression. *The Annals of Statistics 32*, 2 (2004), 407–499. MR2060166

[14] Frank, L., and Friedman, J. A statistical view on some chemometrics regression tools. *Technometrics 16* (1993), 499–511.

[15] Huang, C., Cheang, G. L. H., and Barron, A. Risk of penalized least squares, greedy selection and l1 penalization for flexible function libraries. preprint, 2008.

[16] Osborne, M., Presnell, B., and Turlach, B. On the lasso and its dual. *Journal of Computational and Graphical Statsitics 9* (2000), 319–337. MR1822089

[17] Panchenko, D. Symmetrization approach to concentration inequalities for empirical processes. *The Annals of Probability 31*, 4 (2003), 2068–2081. MR2016612

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

[19] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B 58*, 1 (1996), 267–288. MR1379242

[20] Vapnik, V. *The nature of statistical learning theory.* Springer, 1998. MR1641250

[21] 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. MR2212574