LASSO, Iterative Feature Selection and the Correlation Selector

Oracle Inequalities and Numerical Performances

Pierre Alquier

LPMA, University Paris 7
& LS, CREST

July 24, 2008

Introduction: Confidence Region For β
- Setting of the problem
- Confidence region for β
- Estimation using confidence regions

Estimating $X\beta$
- General Remarks
- Iterative Feature Selection
- LASSO

Estimating β
- General Remarks
- The Dantzig Selector

Extensions
- General Remarks
- The Correlation Selector
- Estimation of $Z\beta$

Regression model

Regression model:

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim \mathcal{N}\left(X\beta, \sigma^2 I_n\right)$$

where X is a $n \times p$ matrix, $\beta \in \mathbb{R}^p$, and possibly:

$$p > n.$$

Quantity of interest

We may be interested in the estimation of several quantities:

- β itself (estimation and interpretation of the parameter);
- $X\beta$ (denoising problem);
- $Z\beta$ in general - for example transductive regression, we have:

  $$Y_i = X_i\beta + \varepsilon_i$$

  for $i \in \{1, ..., n\}$ where X_i is the i-th line of X. Moreover, a new set of X_i's is available: $X_{n+1}, ..., X_{n+m}$ and we want to predict the corresponding values. We define Z as the matrix which lines are the X_{n+i} and we want to predict: $Z\beta$.

Distribution of $X'(Y - X\beta)$

Note that $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ leads to

$$X'(Y - X\beta) \sim \mathcal{N}\left(0, \sigma^2(X'X)\right).$$

Usual normalization: $(X'X)_{i,i}/n = 1$.

So, for any $i \in \{1, ..., p\}$, $[X'(Y - X\beta)]_i \sim \mathcal{N}(0, n\sigma^2)$, and so

$$\mathbb{P}\left\{\left|[X'(Y - X\beta)]_i\right| \geq t\right\} \leq \exp\left(\frac{-t^2}{2n\sigma^2}\right).$$

Confidence regions for β

Union bound:

$$\mathbb{P}\left\{\exists i, \quad \left|[X'(Y - X\beta)]_i\right| \geq t\right\} \leq p \exp\left(\frac{-t^2}{2n\sigma^2}\right) = \varepsilon,$$

that leads to

$$\mathbb{P}\left\{\forall i \in \{1, ..., p\}, \beta \in CR_i(\varepsilon)\right\} \geq 1 - \varepsilon$$

where

$$CR_i(\varepsilon) = \left\{b \in \mathbb{R}^p, \left|[X'(Y - Xb)]_i\right| \leq \sigma\sqrt{2n \log \frac{p}{\varepsilon}}\right\}.$$

Intersection of the confidence regions

Remarks:

(1) Equivalently

$$\mathbb{P}\left\{\beta \in CR(\varepsilon)\right\} \geq 1 - \varepsilon$$

where

$$CR(\varepsilon) = \bigcap_{i=1}^{p} CR_i(\varepsilon),$$

$$CR(\varepsilon) = \left\{ b \in \mathbb{R}^p, \|X'(Y - Xb)\|_{\infty} \leq \sigma\sqrt{2n \log \frac{p}{\varepsilon}} \right\}.$$

(2) The same results could be obtained with non-gaussian noise (using Hoeffding's inequality for example)...

Estimation using $CR_i(\varepsilon)$

With large probability, the region $CR_i(\varepsilon)$ is closed, convex and contains β.
Let $d(.,.)$ be a distance on $\mathbb{R}^p$, and $\Pi_{CR_i(\varepsilon)}$ be the orthogonal projection onto $CR_i(\varepsilon)$ with respect to d. Then we have for any $b \in \mathbb{R}^p$,

$$d\left(\Pi_{CR_i(\varepsilon)} b, \beta\right) \leq d(b, \beta).$$

Consequence: for any estimator $\hat{\beta}$ of β, $\Pi_{CR_i(\varepsilon)}\hat{\beta}$ is a better estimator (at least for the distance d).

Estimation of $X\beta$: general remarks

We want to estimate $X\beta$.

Natural distance:

$$d(b, B) = \|Xb - XB\|_2.$$

Two methods: Iterative Feature Selection, and LASSO.

Iterative Feature Selection

Remarks on the confidence regions motivates the following algorithm:

- start from $\beta(0) = 0$;
- step n: choose $i(n)$ and $\beta(n+1) = \Pi_{CR_{i(n)}(\varepsilon)}\beta(n)$;
- stop when no (significative) improvement is possible.

Theorem: We have

$$\mathbb{P}\left\{\forall n \in \mathbb{N}, \|X[\beta(n+1) - \beta]\|_2^2 \leq \|X[\beta(n) - \beta]\|_2^2\right\} \geq 1 - \varepsilon.$$

LASSO (dual form)

Why not use a projection on $CR(\varepsilon) = \bigcap_i CR_i(\varepsilon)$?

$$\hat{\beta} \in \begin{cases} \arg\min_{b\in\mathbb{R}^p} \|Xb\|_2^2 \\ \\ s.t. \|X'(Y - Xb)\|_\infty \leq s = \sigma\sqrt{2n\log\frac{p}{\varepsilon}}. \end{cases}$$

Theorem: (Osborne, Presnell & Turlach, 2000) Let $\tilde{\beta}$ be a solution of the LASSO program (Tibshirani *et al.*, 1996)

$$\tilde{\beta} = \arg\min_{b\in\mathbb{R}^p} \left\{ \|Y - Xb\|_2^2 + 2s\|b\|_1 \right\},$$

then $X\tilde{\beta} = X\hat{\beta}$.

Oracle Inequality on the LASSO

Sparsity: a lot of β_i are equal to 0;

$$\|\beta\|_0 = \sum_{i=1}^{p} \mathbb{1}_{\beta_i \neq 0}.$$

Theorem: (Tsybakov *et al.*, 2007) If X satisfies

$$\max_{i,\beta_i \neq 0} \max_{j \neq i} \frac{(X'X)_{i,j}}{n} \leq \frac{1}{16\|\beta\|_0},$$

then

$$\mathbb{P}\left\{ \|X(\hat{\beta} - \beta)\|_2^2 \leq 16\sigma^2 \|\beta\|_0 \log \frac{p}{\varepsilon} \right\} \geq 1 - \varepsilon.$$

Estimation of β: general remarks

In this section we want to estimate β.

Natural distance:

$$d(b, B) = \|b - B\|_1.$$

Example: The Dantzig Selector.

The Dantzig Selector

Estimator introduced by Candes and Tao (2007), in order to estimate β.

$$\hat{\beta}_D \in \begin{cases} \arg\min_{b\in\mathbb{R}^p} \|b\|_1 \\ \\ s.t. \|X'(Y - Xb)\|_\infty \leq s. \end{cases}$$

Oracle inequalities: Candes and Tao (2007), Tsybakov *et al.* (2007), etc... With the same kind of hypothesis: sparsity for β, and parts of $X'X$ are nearly orthogonal.

General Remarks

The previous examples are a motivation to study the general family of estimators

$$\begin{cases} \arg\min_{b\in\mathbb{R}^p} d(b,0) \\ s.t. \|X'(Y-Xb)\|_\infty \leq s \end{cases}$$

with a general distance d.
In this last section, we exhibit another interesting estimator in this family (the Correlation Selector) and conclude by a discussion on the transductive regression case where we think one should take $d(b,B) = \|Z(b-B)\|_2$ to estimate $Z\beta$.

The Correlation Selector

We put

$$\hat{\beta}_{CS} \in \begin{cases} \arg\min_{b \in \mathbb{R}^p} \|(X'X)b\|_q^q \\ \\ s.t. \|X'(Y - Xb)\|_\infty \leq s \end{cases}$$

where we can prove that the solution does not depend on q.
We also prove:

$$(X'X)\hat{\beta}_{CS} = \begin{pmatrix} th[(X'Y)_1] \\ \vdots \\ th[(X'Y)_p] \end{pmatrix},$$

where $th(.)$ soft-thresholding function with threshold s.

Oracle Inequality for the Correlation Selector

Sparsity of the correlations: $(X'X)\beta = E(X'Y)$ is sparse.
Theorem: We have

$$\mathbb{P}\left\{\left\|\frac{(X'X)}{n}(\hat{\beta}_{CS} - \beta)\right\|_2^2 \leq 8\sigma^2 \|(X'X)\beta\|_0 \frac{\log \frac{p}{\varepsilon}}{n}\right\} \geq 1 - \varepsilon.$$

What about the estimation of $X\beta$?

Oracle Inequality for the Correlation Selector

Let us assume that for any b such that $[E(XY)]_i = 0 \Rightarrow b_i = 0$,

$$b'b \leq cb' \left(\frac{X'X}{n} \right) b.$$

Theorem: We have

$$\mathbb{P}\left\{ \left\| X(\hat{\beta}_{CS} - \beta) \right\|_2^2 \leq 8c\sigma^2 \|(X'X)\beta\|_0 \log \frac{p}{\varepsilon} \right\} \geq 1 - \varepsilon.$$

Experimental results (1/2)

Tibshirani's experiments, $n = 20$, $p = 8$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, columns of X gaussian with correlation $= 0.5$ and $\sigma \in \{1, 3\}$, theoretical value for s (not optimal in practice).

σ	OLS	LASSO	IFS	CS
3	**3.67** 1.84 8	**1.64** 1.25 4.64	**1.56** 1.20 4.62	**3.65** 1.96 8
1	**0.40** 0.22 8	**0.29** 0.19 5.42	**0.36** 0.23 5.70	**0.44** 0.23 8

Experimental results (2/2)

Same experiment with β such that $(X'X/20)\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$.

σ	OLS	LASSO	IFS	CS
3	**3.64** 1.99 8	**4.83** 2.53 5.98	**5.12** 2.64 6.05	**2.41** 1.92 8
1	**0.41** 0.21 8	**1.09** 1.72 7.11	**0.92** 0.48 7.40	**0.26** 0.19 8

Estimation of $Z\beta$

Natural to use the distance $d(b,B) = \|Z(b-B)\|_2$.
Estimator:

$$\begin{cases} \arg\min_{b\in\mathbb{R}^p} \|Zb\|_2^2 \\ \\ s.t. \|X'(Y-Xb)\|_\infty \leq s \end{cases}$$

estimates $Z\beta$ but with a very unnatural sparsity hypothesis for β. We have to replace the constraint $\|X'(Y-Xb)\|_\infty \leq s$ by another one of the type $\|P(Y-Xb)\|_\infty \leq s$ for some P linked with X and Z.
Work in progress with Mohamed Hebiri (Paris 7).