The Sparsity and Bias of The LASSO Selection In High-Dimensional Linear Regression

Cun-hui Zhang and Jian Huang

Presenter: Quefeng Li

Feb. 26, 2010
Zhao and Yu (2006) [presented by Jie Zhang] showed that under a strong irrepresentable condition, LASSO selects exactly the set of nonzero regression coefficients, provided that these coefficients are uniformly bounded away from zero at a certain rate. They also showed the sign consistency of LASSO.

Meinshausen and Buhlmann (2006) [presented by Jee Young Moon] showed that, for neighborhood selection in Gaussian graphical models, under a neighborhood stability condition, LASSO is consistent, even when \( p > n \).
Under a weaker sparse condition on coefficients and a sparse Riesz condition on the correlation of design variables, the authors showed the rate consistency as the following three aspects.

- LASSO selects a model of the correct order of dimensionality.
- LASSO controls the bias of the selected model.
- The $l_\alpha$-loss for the regression coefficients converge at the best possible rates under the given conditions.
Consider a linear regression model,

\[ y = \sum_{j=1}^{p} \beta_j x_j + \epsilon = X\beta + \epsilon. \]  

(1)

For a given penalty level \( \lambda \geq 0 \), the LASSO estimator of \( \beta \in \mathbb{R}^p \) is

\[ \hat{\beta} = \hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \| y - X\beta \|^2 / 2 + \lambda \| \beta \|_1 \right\}. \]  

(2)

In this paper,

\[ \hat{A} = \hat{A}(\lambda) = \{ j < p : \hat{\beta}_j \neq 0 \}, \]  

(3)

is considered as the model selected by the LASSO.
Sparsity Assumption on $\beta$ 

Assume there exists an index set $A_0 \subset \{1, \cdots, p\}$ such that

$$\#\{j \leq p : j \not\in A_0\} = q, \quad \sum_{j \in A_0} |\beta_j| \leq \eta_1. \quad (4)$$

Under this condition, there exists at most $q$ “large” coefficients and the $l_1$-norm of the “small” coefficients is no greater than $\eta_1$. Compared with the typical assumption,

$$|A_\beta| = q, \quad A_\beta \equiv \{j : \beta_j \neq 0\} \quad (5)$$

(4) is weaker.
Sparse Riesz Condition (SRC) on $X$

For $A \subset \{1, \cdots, p\}$, define $X_A \equiv (x_j, j \in A)$, $\Sigma_A \equiv X_A X'_A/n$.

The design matrix $X$ satisfies SRC with rank $q^*$ and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \leq \frac{\|X_Av\|^2}{n\|v\|^2} \leq c^*, \quad \forall A \text{ with } |A| = q^* \text{ and } v \in \mathbb{R}^{q^*} \quad (6)$$

or equivalently,

$$c_* \leq \|\Sigma_A\|_{(2,2)} \leq c^*, \quad \forall A \text{ with } |A| = q^* \text{ and } v \in \mathbb{R}^{q^*}$$

Note: (6) may not really be a sparsity condition.
A natural definition of the sparsity of the selected model is \( \hat{q} = O(q) \), where

\[
\hat{q} \equiv \hat{q}(\lambda) \equiv |\hat{A}| = \#\{j : \hat{\beta}_j \neq 0\} \tag{7}
\]

The selected model fits the mean \( \mathbf{X}\beta \) well if its bias

\[
\tilde{B} \equiv \tilde{B}(\lambda) \equiv \| (\mathbf{I} - \hat{\mathbf{P}}) \mathbf{X}\beta \|
\]

is small, where \( \hat{\mathbf{P}} \) is the projection from \( \mathbb{R}^n \) to the linear span of the selected \( \mathbf{x}_j \)'s and \( \mathbf{I} \equiv \mathbf{I}_{n \times n} \) is the identity matrix. Then, \( \tilde{B}^2 \) is the sum of squares of the part of the mean vector not explained by the selected model.
To measure the large coefficients missing in the selected model, we define

$$\zeta_\alpha \equiv \zeta_\alpha(\lambda) \equiv \left( \sum_{j \notin A_0} |\beta_j|^\alpha I\{\hat{\beta}_j = 0\} \right)^{1/\alpha}, \quad 0 \leq \alpha \leq \infty. \quad (9)$$

$\zeta_0$ is the number of $q$ largest $|\beta_j|$’s not selected, $\zeta_2$ is the Euclidean length of these missing large coefficients and $\zeta_\infty$ is their maximum.
A Simple Example

The example below indicates that, the following three quantities, are responsible benchmarks for $\tilde{B}^2$ and $n\zeta_2^2$,

$$\lambda\eta_1, \eta_2^2, \frac{q\lambda^2}{n},$$

(10)

where $\eta_2 \equiv \max_{A \subset A_0} \|\sum_{j \in A} \beta_j x_j\| \leq \max_{j \leq p} \|x_j\| \eta_1$.

Example

Suppose we have an orthonormal design with $X'X/n = I_p$ and i.i.d normal error $\epsilon \sim N(0, I_n)$. Then, (2) is the soft-threshold estimator with threshold level $\lambda/n$ for the individual coefficients: $\hat{\beta} = \text{sgn}(z_j)(|z_j - \lambda/n|)^+$, with $z_j \equiv x_j'y/n \sim N(\beta_j, 1/n)$ being the least-square estimator of $\beta_j$. If $|\beta_j| = \lambda/n$ for $j = 1, \cdots, q + \eta_1 n/\lambda$ and $\lambda/\sqrt{n} \to \infty$, then

$$P\{\hat{\beta}_j = 0 \approx 1/2\} \text{ so that } \tilde{B}^2 \approx 2^{-1}(q + \eta_1 n/\lambda)n(\lambda/n)^2 = 2^{-1}(q\lambda^2/n + \eta_1 \lambda).$$
The authors showed that LASSO is rate-consistent in model selection as, for a suitable $\alpha$ (e.g. $\alpha = 2$ or $\alpha = \infty$)

$$\hat{q} = O(q), \quad \tilde{B} = O_p(B), \quad \sqrt{n}\zeta_\alpha = O(B),$$ (11)

with possibility of $\tilde{B} = O(\eta_2)$ and $\zeta_\alpha = 0$ under stronger conditions, where $B \equiv \max(\sqrt{\eta_1 \lambda}, \eta_2, \sqrt{q \lambda^2 / n})$. 
Let
\[ M_1^* \equiv M_1^*(\lambda) \equiv 2 + 4r_1^2 + 4\sqrt{C}r_2 + 4C, \] (12)
\[ M_2^* \equiv M_2^*(\lambda) \equiv \frac{8}{3}\left\{ \frac{1}{4} + r_1^2 + r_2\sqrt{2C(1 + \sqrt{C})} + C\left(\frac{1}{2} + \frac{4}{3}C\right) \right\} \] (13)
and
\[ M_3^* \equiv M_3^*(\lambda) \equiv \frac{8}{3}\left\{ \frac{1}{4} + r_1^2 + r_2\sqrt{C(1 + 2\sqrt{1 + C})} \right. \\
\left. + \frac{3r_2^2}{4} + C\left(\frac{7}{6} + \frac{2}{3}C\right) \right\}, \] (14)
where
\[ r_1 \equiv r_1(\lambda) \equiv \left( \frac{c^*\eta_1 n}{q\lambda} \right)^{1/2}, \quad r_1 \equiv r_1(\lambda) \equiv \left( \frac{c^*\eta_2^2 n}{q\lambda^2} \right)^{1/2}, \quad C \equiv \frac{c^*}{c_*}, \] (15)
and \( \{q, \eta_1, \eta_2, c_*, c^*\} \) are as in (4), (10) and (6).
Since $r_j$ and $M_k^*$ in (12) - (15) are decreasing in $\lambda$. We define a lower bound for the penalty level as

$$\lambda_* \equiv \inf\{\lambda : M_1^*(\lambda)q + 1 \leq q^*\}, \quad \inf \emptyset \equiv \infty. \quad (16)$$

Let $\sigma \equiv (E\|\epsilon^2\|/n)^{1/2}$. With $\lambda_*$ in (16) and $c^*$ in (6), we consider the LASSO path for

$$\lambda \geq \max(\lambda_*, \lambda_{n,p}), \quad \lambda_{n,p} \equiv 2\sigma \sqrt{2(1 + c_0)c^*n \log(p \vee a_n)}, \quad (17)$$

with $c_0 \geq 0$ and $a_n \geq 0$ satisfying $p/(p \vee a_n)^{1+c_0} \approx 0$. For large $p$, the lower bound here is allowed to be of the order $\lambda_{n,p} \sim \sqrt{n \log p}$ with $a_n = 0$. 

Cun-hui Zhang and Jian Huang

Presenter: Quefeng Li

The Sparsity and Bias of The LASSO Selectic

Feb. 26, 2010 12 / 23
**Theorem 1** Suppose $\epsilon \sim N(0, \sigma^2 I)$, $q \geq 1$, and the sparsity (4) and sparse Riesz condition (6) hold. There then exists a set $\Omega_0$ in the sample space of $(X, \epsilon/\sigma)$, depending on $\{X\beta, c_0, a_n\}$ only, such that

$$P\{(X, \epsilon/\sigma) \in \Omega_0\} \geq 2 - \exp\left(\frac{2p}{(p \lor a_n)^{1+c_0}}\right) - \frac{2}{(p \lor a_n)^{1+c_0}} \approx 1 \quad (18)$$

and the following assertions hold in the event $(X, \epsilon/\sigma) \in \Omega_0$ for all $\lambda$ satisfying (17):

$$\hat{q}(\lambda) \leq \tilde{q}(\lambda) \equiv \#\{j : \hat{\beta}_j(\lambda) \neq 0 \text{ or } j \notin A_0\} \leq M_1^*(\lambda)q, \quad (19)$$

$$\tilde{B}^2(\lambda) = \|(I - \hat{P}(\lambda))X\beta\|^2 \leq M_2^*(\lambda)\frac{q\lambda^2}{c^*n}, \quad (20)$$

with $\hat{P}(\lambda)$ being the projection to the span of the selected design vectors $\{x_j, j \in \hat{A}(\lambda)\}$ and

$$\zeta_2^2(\lambda) = \sum_{j \notin A_0} |B_j|^2 I\{\hat{\beta}_j(\lambda) = 0\} \leq M_3^*(\lambda)\frac{q\lambda^2}{c^*c^*n^2}. \quad (21)$$
Remark 1 Conditions are impose on $X$ and $\beta$ jointly. We may first impose SRC on $X$. Given the configuration $\{q^*, c_*, c^*\}$ for the SRC and thus $C \equiv c^*/c_*$, (16) requires that $\{q, r_1, r_2\}$ satisfy

$$(2 + 4r_1^2 + 4\sqrt{Cr_2} + 4C)q + 1 \leq q^*.$$ 

Given $\{q, r_1, r_2\}$ and the penalty level $\lambda$, the condition on $\beta$ becomes

$$|A_0^c| \leq q, \quad \eta_1 \leq \frac{q\lambda r_1^2}{c^* n}, \quad \eta_2 \leq \frac{q\lambda^2 r_2^2}{c^* n}.$$ 

Remark 2 The condition $q \geq 1$ is not essential. Theorem 1 is still valid for $q = 0$ if we use $r_1 q = c^* \eta_1 n/\lambda$ and $r_2^2 q = c^* \eta_2^2 n/\lambda^2$ to recover (12), (13) and (14), resulting in

$$\hat{q}(\lambda) \leq 4c^* \frac{\eta_1 n}{\lambda}, \quad \tilde{B}^2(\lambda) \leq \frac{8}{3} \eta_1 \lambda, \quad \zeta_2^2 = 0.$$ 

The following result is an immediate consequence of Theorem 1.
Theorem 2 Suppose the conditions of Theorem 1 hold. Then, all variables with $\beta_j^2 > M^*_3(\lambda)q\lambda^2/(c^*c_*n^2)$ are selected with $j \in \hat{A}(\lambda)$, provided $(X, \epsilon/\sigma) \in \Omega_0$ and $\lambda$ is in the interval (17). Consequently, if $\beta_j^2 > M^*_3(\lambda)q\lambda^2/(c^*c_*n^2)$ for all $j \notin A_0$, then, for all $\alpha > 0$,

$$P\{A_0^c \subset \hat{A}, \tilde{B}(\lambda) \leq \eta_2 \text{ and } \zeta_\alpha(\lambda) = 0\} \geq 2 - \exp\left(\frac{2p}{(p \lor a_n)^{1+c_0}}\right) - \frac{2}{(p \lor a_n)^{1+c_0}} \approx 1$$ (22)
LASSO Estimation

Although it is not necessary to use LASSO for estimation once variable selection is done, the authors inspect implications of Theorem 1 for the estimation properties of LASSO. For simplicity, they confine this discussion to the special case where $c_*, c^*, r_1, r_2, c_0$ and $\sigma$ are fixed and

$$\lambda / \sqrt{n} \geq 2\sigma \sqrt{2(1 + c_0)c^* \log p} \to \infty.$$  

In this case, $M^*_K$ are fixed constants in (12), (13) and (14), and the required configurations for (4), (6) and (17) in Theorem 1 become

$$M^*_1q + 1 \leq q^*, \quad \eta_1 \leq \left( \frac{r_1^2}{c^*} \right) \frac{q\lambda}{n}, \quad \eta_2 \leq \left( \frac{r_2^2}{c^*} \right) \frac{q\lambda^2}{n}. \quad (23)$$

Of course, $p, q$ and $q^*$ are all allowed to depend on $n$, for example, $p \gg n > q^* > q \to \infty$. 
Theorem 3 Let $c_*, c^*, r_1, r_2, c_0$ and $\sigma$ be fixed and $1 \leq q \leq p \to \infty$. Let $\lambda/\sqrt{n} \geq 2\sigma \sqrt{2(1+c'_0)}c^*n \log p \to \infty$ with a fixed $c'_0 \geq c_0$ and $\Omega_0$ be as in Theorem 1. Suppose the conditions of Theorem 1 hold with configurations satisfying (23). There then exist constants $M_k^*$ depending only on $c_*, c^*, r_1, r_2, c'_0$ and a set $\tilde{\Omega}_1$ in the sample space of $(X, \epsilon/\sigma)$ depending only on $q$ such that

$$P\{(X, \epsilon/\sigma) \notin \Omega_0 \cap \tilde{\Omega}_q | X\}$$

$$\leq e^{2/p^{c_0}} - 1 + \frac{2}{p^{1+c_0}} + \left(\frac{1}{p^2} + \frac{\log p}{p^2/4}\right)^{(q+1)/2} \to 0$$

(24)

and the following assertions hold in the event $(X, \epsilon/\sigma) \in \Omega_0 \cap \tilde{\Omega}_q$:

$$\|X(\hat{\beta} - \beta)\| \leq M_4^* \sigma \sqrt{q \log p}$$

(25)

and, for all $\alpha \geq 1$,

$$\|\hat{\beta} - \beta\|_\alpha \equiv \left(\sum_{i=1}^{p} |\hat{\beta}_j - \beta_j|^\alpha\right)^{1/\alpha} \leq M_5^* \sigma q^{1/(\alpha^2)} \sqrt{\log p/n}.$$
In this section, the authors give some sufficient conditions for SRC for both deterministic and random design matrix. They consider the general form of sparse Riesz condition as

\[
c_{\star}(m) \equiv \min_{\|v\| = 1} \min_{|A| = m} \frac{\|X_A v\|^2}{n}, \quad c^\ast(m) \equiv \max_{\|v\| = 1} \max_{|A| = m} \frac{\|X_A v\|^2}{n},
\]

(27)
Proposition 1 Suppose that $\mathbf{X}$ is standardized with $\|\mathbf{x}_j\|^2/n = 1$. Let $\rho_{jk} = \mathbf{x}_j'\mathbf{x}_k/n$ be the correlation. If

$$
\max_{|A|=q^*} \inf_{\alpha \geq 1} \left\{ \sum_{j \in A} \left( \sum_{k \in A, k \neq j} |\rho_{kj}|^{\alpha/(\alpha-1)} \right)^{\alpha-1} \right\}^{1/\alpha} \leq \delta < 1,
$$

then the sparse Riesz condition (6) holds with rank $q^*$ and spectrum bounds $c_* = 1 - \delta$ and $c^* = 1 + \delta$. In particular, (6) holds with $c_* = 1 - \delta$ and $c^* = 1 + \delta$ if

$$
\max_{1 \leq j < k \leq p} |\rho_{jk}| \leq \frac{\delta}{q^* - 1}, \quad \delta < 1.
$$
Remark 3 If $\delta = 1/3$, then $C \equiv c^*/c_* = 2$ and Theorem 1 is applicable if $10q + 1 \leq q^*$ and $\eta_1 = 0$ in (4).
Random design matrices

Proposition 2 Suppose that the \(n\) rows of a random matrix \(X_{n \times p}\) are i.i.d. copies of a subvector \((\xi_{k1}, \cdots, \xi_{kp})\) of a zero-mean random sequence \(\{\xi_j, j = 1, 2, \cdots\}\) satisfying \(\rho^* \sum_{j=1}^{\infty} b_j^2 \leq E \left| \sum_{j=1}^{\infty} b_j \xi_j \right|^2 \leq \rho^* \sum_{j=1}^{\infty} b_j^2\).

Let \(c_*(m)\) and \(c^*(m)\) be as in (27).

(i) Suppose \(\{\xi_k, k \geq 1\}\) is a Gaussian sequence. Let \(\epsilon_k, k = 1, 2, 3, 4\) be positive constants in \((0, 1)\) satisfying \(m \leq \min(p, \epsilon_1^2, n), \epsilon_1 + \epsilon_2 < 1\) and \(\epsilon_3 + \epsilon_4 = \epsilon_2^2/2\). Then, for all \((m, n, p)\) satisfying \(\log(p/m) \leq \epsilon_3 n\),

\[
P\{\tau_* \rho_* \leq c_*(m) \leq c^*(m) \leq \tau^* \rho^*\} \geq 1 - 2e^{-n\epsilon_4}, \tag{30}\]

where \(\tau_* \equiv (1 - \epsilon_1 - \epsilon_2)^2\) and \(\tau^* \equiv (1 + \epsilon_1 + \epsilon_2)^2\).

(ii) Suppose \(\max_{j \leq p} \|\xi_j\|_{\infty} \leq K_n < \infty\). Then, for any \(\tau_* < 1 < \tau^*\), there exists a constant \(\epsilon_0 > 0\) depending only on \(\rho^*, \rho_*, \tau_*\) and \(\tau^*\) such that

\[
P\{\tau_* \rho_* \leq c_*(m) \leq c^*(m) \leq \tau^* \rho^*\} \to 1\]

for \(m \equiv m_n \leq \epsilon_0 K_n^{-1} \sqrt{n/\log p}\), provided \(\sqrt{n}/K_n \to \infty\).
Remark 4 By the Stirling formula, for $p/n \to \infty$, 

$$m \leq \epsilon_3 n / \log(p/n) \Rightarrow \log \left( \frac{p}{m} \right) \leq (\epsilon_3 + 0(1)) n.$$ 

Thus, Proposition 2(i) is applicable up to $p = e^{an}$ for some small $a > 0$. 
Thank You!