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

Previous Results Review

- 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 $p > n$.

Main Objective of This Paper

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_α-loss for the regression coefficients converge at the best possible rates under the given conditions.

Consider a linear regression model,

$$\mathbf{y} = \sum_{j=1}^{p} \beta_j \mathbf{x}_j + \boldsymbol{\epsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1}$$

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

$$\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\lambda) \equiv \underset{\boldsymbol{\beta}}{\text{argmin}}\{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2/2 + \lambda\|\boldsymbol{\beta}\|_1\}. \tag{2}$$

In this paper,

$$\hat{A} \equiv \hat{A}(\lambda) \equiv \{j < p : \hat{\beta}_j \neq 0\}, \tag{3}$$

is considered as the model selected by the LASSO.

Sparsity Assumption on $\boldsymbol{\beta}$

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

$$\#\{j \leq p : j \notin A_0\} = q, \ \sum_{j \in A_0} |\beta_j| \leq \eta_1. \tag{4}$$

Under this condition, there exists at most q "large" coefficients and the l_1-norm of the "small" coefficients is no greater than η_1.
Compared with the typical assumption,

$$|A_{\boldsymbol{\beta}}| = q, \ A_{\boldsymbol{\beta}} \equiv \{j : \beta_j \neq 0\} \tag{5}$$

(4) is weaker.

Sparse Riesz Condition (SRC) on $\mathbf{X}$

For $A \subset \{1, \cdots, p\}$, define $\mathbf{X}_A \equiv (\mathbf{x}_j, j \in A)$, $\mathbf{\Sigma}_A \equiv \mathbf{X}_A \mathbf{X}_A'/n$.
The design matrix $\mathbf{X}$ satisfies SRC with rank q^* and spectrum bounds $0 < c_* < c^* < \infty$ if

$$c_* \leq \frac{\|\mathbf{X}_A \mathbf{v}\|^2}{n\|\mathbf{v}\|^2} \leq c^*, \quad \forall A \text{ with } |A| = q^* \text{ and } \mathbf{v} \in \mathbb{R}^{q^*} \tag{6}$$

or equivalently,

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

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

$$\tilde{B} \equiv \tilde{B}(\lambda) \equiv \|(\mathbf{I} - \hat{\mathbf{P}})\mathbf{X}\beta\| \tag{8}$$

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

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

ζ_0 is the number of q largest $|\beta_j|$'s not selected, ζ_2 is the Euclidean length of these missing large coefficients and ζ_∞ is their maximum.

A Simple Example

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

$$\lambda\eta_1, \eta_2^2, \frac{q\lambda^2}{n}, \tag{10}$$

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

Example

Suppose we have an orthonormal design with $\mathbf{X}'\mathbf{X}/n = \mathbf{I}_p$ and i.i.d normal error $\boldsymbol{\epsilon} \sim N(0, \mathbf{I}_n)$. Then, (2) is the soft-threshold estimator with threshold level λ/n for the individual coefficients: $\hat{\boldsymbol{\beta}} = \text{sgn}(z_j)(|z_j - \lambda/n|)^+$, with $z_j \equiv \mathbf{x}_j'\mathbf{y}/n \sim N(\beta_j, 1/n)$ being the least-square estimator of β_j. If $|\beta_j| = \lambda/n$ for $j = 1, \cdots, q + \eta_1 n/\lambda$ and $\lambda/\sqrt{n} \to \infty$, then $P\{\hat{\beta}_j = 0 \approx 1/2\}$ so that $\tilde{B}^2 \approx 2^{-1}(q + \eta_1 n/\lambda)n(\lambda/n)^2 = 2^{-1}(q\lambda^2/n + \eta_1\lambda)$.

Goal

The authors showed that LASSO is rate-consistent in model selection as, for a suitable α (e.g. $\alpha = 2$ or $\alpha = \infty$)

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

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

$$M_1^* \equiv M_1^*(\lambda) \equiv 2 + 4r_1^2 + 4\sqrt{C}r_2 + 4C, \tag{12}$$

$$M_2^* \equiv M_2^*(\lambda) \equiv \frac{8}{3}\{\frac{1}{4} + r_1^2 + r_2\sqrt{2C}(1 + \sqrt{C}) + C(\frac{1}{2} + \frac{4}{3}C)\} \tag{13}$$

and

$$\begin{aligned} M_3^* \equiv M_3^*(\lambda) \equiv \frac{8}{3}\{\frac{1}{4} + r_1^2 + r_2\sqrt{C}(1 + 2\sqrt{1 + C}) \\ + \frac{3r_2^2}{4} + C(\frac{7}{6} + \frac{2}{3}C)\}, \end{aligned} \tag{14}$$

where

$$r_1 \equiv r_1(\lambda) \equiv \left(\frac{c^*\eta_1 n}{q\lambda}\right)^{1/2}, \quad r_1 \equiv r_1(\lambda) \equiv \left(\frac{c^*\eta_2^2 n}{q\lambda^2}\right)^{1/2}, \quad C \equiv \frac{c^*}{c_*}, \tag{15}$$

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 λ. We define a lower bound for the penalty level as

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

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

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

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 $a_n = 0$.

Theorem 1 Suppose $\epsilon \sim N(0, \sigma^2 \mathbf{I}), q \geq 1$, and the sparsity (4) and sparse Riesz condition (6) hold. There then exists a set Ω_0 in the sample space of $(\mathbf{X}, \epsilon/\sigma)$, depending on $\{\mathbf{X}\beta, c_0, a_n\}$ only, such that

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

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

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

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

with $\hat{\mathbf{P}}(\lambda)$ being the projection to the span of the selected design vectors $\{\mathbf{x}_j, j \in \hat{A}(\lambda)\}$ 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 $\mathbf{X}$ and β jointly. We may first impose SRC on $\mathbf{X}$. Given the configuration $\{q^*, c_*, c^*\}$ for the SRC and thus $C \equiv c^*/c_*$, (16) requires that $\{q, r_1, r_2\}$ satisfy $(2 + 4r_1^2 + 4\sqrt{C}r_2 + 4C)q + 1 \leq q^*$. Given $\{q, r_1, r_2\}$ and the penalty level λ, the condition on β becomes

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

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 $(\mathbf{X}, \epsilon/\sigma) \in \Omega_0$ and λ 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 $\alpha > 0$,

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

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 σ 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

$$M_1^* q + 1 \leq q^*, \qquad \eta_1 \leq \left(\frac{r_1^2}{c^*}\right)\frac{q\lambda}{n}, \qquad \eta_2 \leq \left(\frac{r_2^2}{c^*}\right)\frac{q\lambda^2}{n}. \tag{23}$$

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 σ be fixed and $1 \le q \le p \to \infty$. Let $\lambda/\sqrt{n} \ge 2\sigma\sqrt{2(1+c_0')c^* n \log p} \to \infty$ with a fixed $c_0' \ge c_0$ and Ω_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 $(\mathbf{X}, \epsilon/\sigma)$ depending only on q such that

$$P\{(\mathbf{X}, \epsilon/\sigma) \notin \Omega_0 \cap \tilde{\Omega}_q | \mathbf{X}\} \le e^{2/p^{c_0}} - 1 + \frac{2}{p^{1+c_0}} + \left(\frac{1}{p^2} + \frac{\log p}{p^2/4}\right)^{(q+1)/2} \to 0 \tag{24}$$

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

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\| \le M_4^* \sigma \sqrt{q \log p} \tag{25}$$

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

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_\alpha \equiv \left(\sum_{i=1}^{p} |\hat{\beta}_j - \beta_j|^\alpha\right)^{1/\alpha} \le M_5^* \sigma q^{1/(\alpha\wedge 2)} \sqrt{\log p/n}. \tag{26}$$

Sufficient Condition for SRC

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

$$c_*(m) \equiv \min_{|A|=m} \min_{\|\mathbf{v}\|=1} \|\mathbf{X}_A \mathbf{v}\|^2/n, c^*(m) \equiv \max_{|A|=m} \max_{\|\mathbf{v}\|=1} \|\mathbf{X}_A \mathbf{v}\|^2/n, \tag{27}$$

Deterministic design matrices

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

$$\max_{|A|=q^*} \inf_{\alpha \geq 1} \left\{ \sum_{j \in A} \left(\sum_{k \in A, k \neq j} |\rho_{kj}|^{\alpha/(\alpha-1)} \right)^{\alpha-1} \right\}^{1/\alpha} \leq \delta < 1, \tag{28}$$

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

$$\max_{1 \leq j < k \leq p} |\rho_{jk}| \leq \frac{\delta}{q^* - 1}, \quad \delta < 1. \tag{29}$$

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 $\mathbf{X}_{n\times p}$ are i.i.d. copies of a subvector $(\xi_{k_1},\cdots,\xi_{k_p})$ of a zero-mean random sequence $\{\xi_j, j=1,2,\cdots\}$ satisfying $\rho_*\sum_{j=1}^{\infty} b_j^2 \le \mathrm{E}\left|\sum_{j=1}^{\infty} b_j\xi_j\right|^2 \le \rho^*\sum_{j=1}^{\infty} b_j^2$.
Let $c_*(m)$ and $c^*(m)$ be as in (27).
(i) Suppose $\{\xi_k, k\ge 1\}$ is a Gaussian sequence. Let $\epsilon_k, k=1,2,3,4$ be positive constants in $(0,1)$ satisfying $m\le\min(p,\epsilon_1^2,n)$, $\epsilon_1+\epsilon_2<1$ and $\epsilon_3+\epsilon_4=\epsilon_2^2/2$. Then, for all (m,n,p) satisfying $\log\binom{p}{m}\le\epsilon_3 n$,

$$P\{\tau_*\rho_*\le c_*(m)\le c^*(m)\le\tau^*\rho^*\}\ge 1-2e^{-n\epsilon_4}, \tag{30}$$

where $\tau_*\equiv(1-\epsilon_1-\epsilon_2)^2$ and $\tau^*\equiv(1+\epsilon_1+\epsilon_2)^2$.
(ii) Suppose $\max_{j\le p}\|\xi_{k_j}\|_\infty\le K_n<\infty$. Then, for any $\tau_*<1<\tau^*$, there exists a constant $\epsilon_0>0$ depending only on ρ^*,ρ_*,τ_* and τ^* such that

$$P\{\tau_*\rho_*\le c_*(m)\le c^*(m)\le\tau^*\rho^*\}\to 1$$

for $m\equiv m_n\le\epsilon_0 K_n^{-1}\sqrt{n/\log p}$, provided $\sqrt{n}/K_n\to\infty$.

Remark 4 By the Stirling formula, for $p/n \to \infty$,

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

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

Thank You!