On Model Selection Consistency Of Lasso

Peng Zhao, Bin Yu
Department of Statistics
University of California, Berkeley

Jie Zhang

UW-Madison

February 12, 2010

Outline

1. Review
   - LASSO
   - Consistency
2. Important Definitions
   - Sign Consistency
   - Irrepresentable Conditions
3. Results
   - Proposition 1
   - In the setting of small p and q
   - In the setting of large p and q
   - Sufficient Conditions for S.I.R.
4. Proofs

Outline

LASSO – Definition

Assume the following linear regression model:

$$Y_n = X_n\beta^n + \epsilon_n$$

Y_n is an $n \times 1$ response;
$X_n = (X_1^n, \ldots, X_p^n) = ((x_1)^T, \ldots, (x_n)^T)^T$ is the $n \times p$ design matrix;
β^n is the $p \times 1$ vector of model coefficients.

Lasso estimator is:

$$\hat{\beta}^n(\lambda) = \arg\min_{\beta}[||Y_n - X_n\beta||_2^2 + \lambda||\beta||_1]$$

with

$$\lambda \geq 0$$

LASSO – Notation

$$\beta^n = (\beta_1^n, \ldots, \beta_q^n, \beta_{q+1}^n, \ldots, \beta_p^n)^T$$

Assume: $\beta_j^n \neq 0$ for $j = 1, \ldots, q$ and $\beta_j^n = 0$ for $j = q+1, \ldots, p$

$$\beta_{(1)}^n = (\beta_1^n, \ldots, \beta_q^n),\ \beta_{(2)}^n = (\beta_{q+1}^n, \ldots, \beta_p^n)$$

$$X_n(1) = (X_1^n, \ldots, X_q^n),\ X_n(2) = (X_{q+1}^n, \ldots, X_p^n)$$

$$C^n = \frac{1}{n} X_n' X_n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}$$

where

$C_{11}^n = \frac{1}{n} X_n(1)' X_n(1)$, $C_{12}^n = \frac{1}{n} X_n(1)' X_n(2)$, $C_{21}^n = \frac{1}{n} X_n(2)' X_n(1)$, $C_{22}^n = \frac{1}{n} X_n(2)' X_n(2)$

Consistency – Definition

- Estimation consistency:

$$\hat{\beta}^n - \beta^n \to_p 0, \text{ as } n \to \infty$$

- Model selection consistency:

$$P(\{i : \hat{\beta}_i^n \neq 0\} = \{i : \beta_i^n \neq 0\}) \to 1, \text{ as } n \to \infty$$

- Sign consistency:

$$P(\hat{\beta}^n =_s \beta^n) \to 1, \text{ as } n \to \infty$$

where

$$\hat{\beta}^n =_s \beta^n \Leftrightarrow sign(\hat{\beta}^n) = sign(\beta^n)$$

Consistency – History

- Knight and Fu(2000) have shown estimation consistency for Lasso for fixed p and fixed β^n;
- Meinshausen and Buhlmann(2006) have shown that Lasso is consistent in estimating the dependency between Gaussian variables even when p grows faster than n;
- Zhao and Yu(2006) have show model selection consistency for both fixed p and large p problems.

Outline

Important Definitions

Definition 1

Lasso is **Strongly Sign Consistent** if $\exists \lambda_n = f(n)$, s.t.

$$\lim_{n \to \infty} P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1$$

Lasso is **General Sign Consistent** if

$$\lim_{n \to \infty} P(\exists \lambda \geq 0, \hat{\beta}^n(\lambda) =_s \beta^n) = 1$$

Important Definition

Definition 2

Strong Irrepresentable Condition: $\exists \eta > 0$, s.t.

$$|C_{21}^n (C_{11}^n)^{-1} sign(\beta_{(1)}^n)| \leq \mathbf{1} - \eta$$

Weak Irrepresentable Condition:

$$|C_{21}^n (C_{11}^n)^{-1} sign(\beta_{(1)}^n)| < \mathbf{1}$$

Where **1** is a p-q by 1 vector of 1's, and the inequality holds element-wise.

Outline

The big result

Proposition 1

Assume Strong Irrepresentable Condition holds with a constant $\eta > 0$, then

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq P(A_n \cap B_n)$$

for

$$A_n = \{|(C_{11}^n)^{-1} W^n(1)| < \sqrt{n}|\beta_{(1)}^n| - \frac{\lambda_n}{2n}|(C_{11}^n)^{-1} sign(\beta_{(1)}^n)|\}$$

$$B_n = \{|C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)| \leq \frac{\lambda_n}{2\sqrt{n}}\eta\}$$

where

$$W^n(1) = \frac{1}{\sqrt{n}} X_n(1)' \epsilon_n \text{ and } W^n(2) = \frac{1}{\sqrt{n}} X_n(2)' \epsilon_n$$

Small p and q – Assumptions

Classical setting: q,p and β^n are all fixed as $n \to \infty$.

Assume the following regularity conditions:

$$C^n \to C > 0, \text{ as } n \to \infty; \tag{1}$$

$$\frac{1}{n} \max_{1 \leq i \leq n} ((x_i^n)^T x_i^n) \to 0, \text{ as } n \to \infty. \tag{2}$$

Small p and q – Theorems

Theorem 1

For fixed q,p and $\beta^n = \beta$, under regularity condition (1) and (2), Lasso is strongly sign consistent if Strong Irrepresentable Condition holds. That is, when Strong Irrepresentable Condition holds, for $\forall \lambda_n$ that satisfies $\lambda_n/n \to 0$ and $\lambda_n/n^{\frac{1+c}{2}} \to \infty$ with $0 \leq c < 1$, we have

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) = 1 - o(e^{-n^c})$$

Theorem 2

For fixed q,p and $\beta^n = \beta$, under regularity condition (1) and (2), Lasso is general sign consistent only if there exists N so that Weak Irrepresentable Condition holds for $n > N$.

Large p and q – Assumptions

The dimension of the designs C^n and parameters β^n grow as n grows, then, p_n and q_n are allowed to grow with n.

Assume the following conditions: $\exists 0 \leq c_1 < c_2 \leq 1$ and $M_1, M_2, M_3, M_4 > 0$,

$$\frac{1}{n}(X_i^n)' X_i^n \leq M_1, for \forall i, \tag{3}$$

$$\alpha' C_{11}^n \alpha \geq M_2, for \forall ||\alpha||_2^2 = 1, \tag{4}$$

$$q_n = O(n^{c_1}), \tag{5}$$

$$n^{\frac{1-c_2}{2}} \min_{i=1,\ldots,q} |\beta_i^n| \geq M_3. \tag{6}$$

Large p and q – Theorems

Theorem 3

Assume ϵ_i^n's are i.i.d. random variables with $E(\epsilon_i^n)^{2k} < \infty$ for an integer $k > 0$. Under conditions (3)(4)(5)(6), Strong Irrepresentable Condition implies that Lasso has strong sign consistency for $p_n = o(n^{(c_2 - c_1)k})$. In particular, for $\forall \lambda_n$ that satisfies $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2 - c_1}{2}})$ and $\frac{1}{p_n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$, we have

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - O(\frac{p_n n^k}{\lambda^{2k}}) \to 1 \text{ as } n \to \infty.$$

Large p and q – Theorems

Theorem 4

Assume ϵ_i^n's are i.i.d. Gaussian random variables. Under conditions (3)(4)(5)(6),if there exists $0 \leq c_3 < c_2 - c_1$ *for which* $p_n = O(e^{n^{c_3}})$*, then Strong Irrepresentable Condition implies that Lasso has strong sign consistency.*

In particular, for $\lambda_n \propto n^{\frac{1+c_4}{2}}$ *with* $c_3 < c_4 < c_2 - c_1$*,*

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - o(e^{-n^{c_3}}) \to 1 \text{ as } n \to \infty.$$

Sufficient Conditions for S.I.R

Corollary 1 (Constant Positive Correlation)

Suppose C^n *has 1's on the diagonal, and there exists* $c > 0$ *such that* $0 < C^n_{ij} = r_n \leq \frac{1}{1+cq}$*, then S.I.R. holds.*

Corollary 2 (Bounded Correlation)

Suppose C^n *has 1's on the diagonal and bounded correlation* $|C^n_{ij}| \leq \frac{c}{2q-1}$ *for a constant* $0 < c < 1$*, then S.I.R. holds*

Corollary 3 (Power Decay Correlation)

Suppose for any $i, j = 1, \dots, p$, $C^n_{ij} = (\rho_n)^{|i-j|}$*, for* $|\rho_n| \leq c < 1$*, then S.I.R. holds.*

Outline

Proposition 1

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq P(A_n \cap B_n)$$

Proof.

Need to show: $A_n \cap B_n$ implies $sign(\beta^{\hat{n}}_{(1)}) = sign(\beta^n_{(1)})$, and $\beta^{\hat{n}}_{(2)} = 0$;

If define $\hat{u}^n = \hat{\beta}^n - \beta^n$, a sufficient condition for $\hat{\beta}^n(\lambda_n) =_s \beta^n$ is:

$$|\hat{u}^n(1)| < |\beta^n_{(1)}|, \text{ and } \hat{u}^n(2) = 0 \qquad (\star)$$

Another thing to notice is, since

$$\hat{\beta}^n = \arg\min_{\beta} ||Y_n - X_n\beta||_2^2 + \lambda||\beta||_1$$

then

$$\begin{aligned} \hat{u}^n &= \arg\min_{u^n}[||Y_n - X_n(u^n + \beta^n)||_2^2 + \lambda_n||u^n + \beta^n||_1] \\ &\equiv \arg\min_{u^n} V_n(u^n) \end{aligned}$$

Proposition 1(cont.)

Lemma 2 (Karush-Kuhn-Tucker condition)

$\hat{\beta}^n = (\hat{\beta}_1^n, \ldots, \hat{\beta}_p^n)$ *are the Lasso estimates as defined above, if and only if*

$$\frac{d||Y_n - X_n\beta||_2^2}{d\beta_j}\Big|_{\beta_j=\hat{\beta}_j^n} = \lambda sign(\hat{\beta}_j^n) \quad \text{for } j \text{ s.t. } \hat{\beta}_j^n \neq 0 \tag{7}$$

$$\left|\frac{d||Y_n - X_n\beta||_2^2}{d\beta_j}\Big|_{\beta_j=\hat{\beta}_j^n}\right| \leq \lambda \quad \text{for } j \text{ s.t. } \hat{\beta}_j^n = 0 \tag{8}$$

Proof(cont.)

To take advantage of the KKT condition, it's natural to think that whether applying (7) and (8) can generate the desired result in ($\star$). □

Proposition 1(cont.)

Proof(cont.)

So, assume:

$$\hat{v}^n(2) = 0$$

and $\hat{v}^n(1)$ is the solution of:

$$C_{11}^n(\sqrt{n}\hat{v}^n(1)) - W^n(1) = -\frac{\lambda_n}{2\sqrt{n}} sign(\beta_{(1)}^n) \tag{28}$$

a If $\hat{v}^n$ satisfies KKT condition (7) and (8), then $\hat{v}^n$ is one Lasso estimator which minimize $V_n(u^n)$.
Then, by uniqueness of Lasso estimator, $\hat{u}^n = \hat{v}^n$.

b If $\hat{v}^n(1)$ satisfying (28) can imply $|\hat{v}^n(1)| < |\beta_{(1)}^n|$, then we finish the proof.

□

Proposition 1(cont.)

Proof(cont.) - everything we know.

A_n implies:

$$|(C_{11}^n)^{-1}W^n(1)| < \sqrt{n}|\beta_{(1)}^n| - \frac{\lambda_n}{2n}|(C_{11}^n)^{-1}sign(\beta_{(1)}^n)| \quad (31)$$

B_n and $|C_{21}^n(C_{11}^n)^{-1}sign(\beta_{(1)}^n)| \leq \mathbf{1} - \eta$ (S.I.R.) implies:

$$|C_{21}^n(C_{11}^n)^{-1}W^n(1) - W^n(2)| \leq \frac{\lambda_n}{2\sqrt{n}}(\mathbf{1} - |C_{21}^n(C_{11}^n)^{-1}sign(\beta_{(1)}^n)|) \quad (32)$$

as well as $\hat{v}^n(2) = 0$ and (28). $\square$

Proposition 1(cont.)

Proof(cont.)

Then, (28) and (31) implies:

$$|\hat{v}^n(1)| < |\beta^n_{(1)}| \tag{29}$$

And, (28) and (32) implies:

$$-\frac{\lambda_n}{2\sqrt{n}}\mathbf{1} \leq C^n_{21}(\sqrt{n}\hat{v}^n(1)) - W^n(2) \leq \frac{\lambda_n}{2\sqrt{n}}\mathbf{1} \tag{30}$$

□

Proposition 1(cont.)

Proof(cont.)

With $\hat{v}^n(2) = 0$ and (29), (28) and (30) are exactly the KKT condition. Because:

$$
\begin{aligned}
-\frac{1}{2\sqrt{n}} \frac{d||Y_n - X_n(u^n + \beta^n)||_2^2}{d(u_j^n + \beta_j^n)}\Big|_{u_j^n = \hat{v}_j^n} &= -\frac{\lambda_n}{2\sqrt{n}} sign(\hat{v}_j^n + \beta_j^n) \\
&= -\frac{\lambda_n}{2\sqrt{n}} sign(\beta_j^n)
\end{aligned}
$$

for $\hat{v}_j^n$ in $\hat{v}^n(1)$

$$
\frac{1}{2\sqrt{n}} \left| \frac{d||Y_n - X_n(u^n + \beta^n)||_2^2}{d(u_j^n + \beta_j^n)}\Big|_{v_j^n = \hat{v}_j^n} \right| \leq \frac{\lambda_n}{2\sqrt{n}}
$$

for $\hat{v}_j^n$ in $\hat{v}^n(2) = 0$

Theorem 3

$$P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \geq 1 - O(\frac{p_n n^k}{\lambda^{2k}}) \to 1 \text{ as } n \to \infty.$$

$$p_n = o(n^{(c_2 - c_1)k}), \text{ and } q_n = O(n^{c_1})$$

Proof.

By proposition 1,

$$1 - P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \leq 1 - P(A_n \cap B_n) \leq P(A_n^c) + P(B_n^c)$$

$$\leq \sum_{i=1}^{q} P(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n)) + \sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_i^n}{2\sqrt{n}} \eta_i)$$

where

$$z^n = (z_1^n, \ldots, z_q^n)' = (C_{11}^n)^{-1} W^n(1)$$

$$\zeta^n = (\zeta_1^n, \ldots, \zeta_{p-q}^n)' = C_{21}^n (C_{11}^n)^{-1} W^n(1) - W^n(2)$$

$$b = (b_1^n, \ldots, b_q^n) = (C_{11}^n)^{-1} sign(\beta_{(1)}^n).$$

Theorem 3(cont.)

Proof(cont.)

In order to apply Markov's Inequality, need to have $E(z_i^n)^{2k} < \infty$ and $E(\zeta_i^n)^{2k} < \infty$. By condition $E(\epsilon_i^n)^{2k} < \infty$ and condition (3) (4), and

$$E(\alpha'\epsilon^n) \leq (2k-1)!!||\alpha||_2^2 E(\epsilon_i^n)^{2k}$$

$E(z_i^n)^{2k} < \infty$ and $E(\zeta_i^n)^{2k} < \infty$ are guaranteed.

Then, by Markov's Inequality, for $\frac{\lambda_n}{\sqrt{n}} = o(n^{\frac{c_2-c_1}{2}})$

$$\sum_{i=1}^{q} P(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n)) \leq \sum_{i=1}^{q} \frac{E|z_i^n|^{2k}}{(\sqrt{n}\beta_i^n)^{2k}} = qO(n^{-kc_2}) = o(\frac{pn^k}{\lambda_n^{2k}})$$

$$\sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_i^n}{2\sqrt{n}}\eta_i) \leq \sum_{i=1}^{p-q} \frac{E|\zeta_i^n|^{2k}}{(\frac{\lambda_n}{\sqrt{n}} 2\eta_i)^{2k}} = (p-q)O(\frac{n^k}{\lambda_n^{2k}}) = O(\frac{pn^k}{\lambda_n^{2k}})$$

Theorem 3(cont.)

Proof(cont.)

So,

$$1 - P(\hat{\beta}^n(\lambda_n) =_s \beta^n) \leq P(A_n^c) + P(B_n^c) \leq O(\frac{pn^k}{\lambda_n^{2k}}) \to 0 \text{ as } n \to \infty$$

for $\frac{1}{p_n}(\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty$. $\square$

Theorem 4

- The inequility:

$$1 - P(\hat{\beta}^n(\lambda_n) =_s \beta^n)$$
$$\leq \sum_{i=1}^{q} P(|z_i^n| \geq \sqrt{n}(|\beta_i^n| - \frac{\lambda_n}{2n} b_i^n)) + \sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_i^n}{2\sqrt{n}} \eta_i)$$

  still holds.
- From the normal assumption of ϵ_i^n, z_i's and ζ_i's are also normal. Rewrite the probabilities above as $1 - \Phi(f(n))$ and $1 - \Phi(g(n))$.
- Use the inequality:

$$1 - \Phi(t) < t^{-1} e^{-\frac{1}{2}t^2}$$

  then, the summation of is bounded by $o(e^{-n_3^c})$.

Thank You!