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

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
Assume the following linear regression model:

\[ Y_n = X_n\beta^n + \epsilon_n \]

\( Y_n \) is an \( n \times 1 \) response;
\( X_n = (X^n_1, \ldots, X^n_p) = ((x_1)^T, \ldots, (x_n)^T)^T \) is the \( n \times p \) design matrix;
\( \beta^n \) 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

\[ \lambda \geq 0 \]
\[ \beta^n = (\beta^n_1, \ldots, \beta^n_q, \beta^n_{q+1}, \ldots, \beta^n_p)^T \]

Assume: \( \beta^n_j \neq 0 \) for \( j = 1, \ldots, q \) and \( \beta^n_j = 0 \) for \( j = q + 1, \ldots, p \)

\[ \beta^n_{(1)} = (\beta^n_1, \ldots, \beta^n_q), \quad \beta^n_{(2)} = (\beta^n_{q+1}, \ldots, \beta^n_p) \]

\[ X_n(1) = (X^n_1, \ldots, X^n_q), \quad X_n(2) = (X^n_{q+1}, \ldots, X^n_p) \]

\[ C^n = \frac{1}{n} X_n' X_n = \begin{pmatrix} C^n_{11} & C^n_{12} \\ C^n_{21} & C^n_{22} \end{pmatrix} \]

where

\[ C^n_{11} = \frac{1}{n} X_n(1)' X_n(1), \quad C^n_{12} = \frac{1}{n} X_n(1)' X_n(2), \quad C^n_{21} = \frac{1}{n} X_n(2)' X_n(1), \quad C^n_{22} = \frac{1}{n} X_n(2)' X_n(2) \]
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

\[ \hat{\beta}^n =_s \beta^n \iff sign(\hat{\beta}^n) = sign(\beta^n) \]
Consistency – History

- Knight and Fu (2000) have shown estimation consistency for Lasso for fixed $p$ and fixed $\beta^n$;

- 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 shown model selection consistency for both fixed $p$ and large $p$ problems.
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
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:** \( \exists \eta > 0, \text{ s.t. } \)

\[
| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}) | \leq 1 - \eta
\]

**Weak Irrepresentable Condition:**

\[
| C_{21}^n (C_{11}^n)^{-1} \text{sign}(\beta_{(1)}) | < 1
\]

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

$$A_n = \{ |(C^n_{11})^{-1} W^n(1)| < \sqrt{n}|\beta^n_{(1)}| - \frac{\lambda_n}{2n} |(C^n_{11})^{-1} \text{sign}(\beta^n_{(1)})| \}$$

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

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 $\beta^n$ are all fixed as $n \to \infty$.

Assume the following regularity conditions:

1. $C^n \to C > 0$, as $n \to \infty$; (1)

2. $\frac{1}{n} \max_{1 \leq i \leq n} (x_i^n)^T x_i^n) \to 0$, as $n \to \infty$. (2)
Theorem 1
For fixed $q, p$ and $\beta^n = \beta$, under regularity condition (1) and (2), \textit{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

$$P(\hat{\beta}^n(\lambda_n) =_{s} \beta^n) = 1 - o(e^{-nc})$$

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

The dimension of the designs $C^n$ and parameters $\beta^n$ 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 $M_1, M_2, M_3, M_4 > 0$,

$$\frac{1}{n} (X^n_i)'X^n_i \leq M_1, \text{for } \forall i,$$

$$\alpha' C^n_{11} \alpha \geq M_2, \text{for } \forall \||\alpha||^2 = 1,$$

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

$$n^{\frac{1-c_2}{2}} \min_{i=1,...,q} |\beta^n_i| \geq M_3.$$
Theorem 3

Assume $\epsilon_i^n$’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

$$P(\hat{\beta}^n(\lambda_n) = s \beta^n) \geq 1 - O\left(\frac{p_n n^k}{\lambda^{2k}}\right) \to 1 \text{ as } n \to \infty.$$
Theorem 4

Assume $\epsilon_i^n$'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 $\lambda_n \propto n^{1+c_4/2}$ with $c_3 < c_4 < c_2 - c_1$,

$$P(\hat{\beta}_n(\lambda_n) = s \beta^n) \geq 1 - o(e^{-n c_3}) \rightarrow 1 \text{ as } n \rightarrow \infty.$$
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 $i, j = 1, \ldots, p$, $C^n_{ij} = (\rho_n)^{|i-j|}$, for $|\rho_n| \leq c < 1$, then S.I.R. holds.
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
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 \( \text{sign}(\beta_{(1)}^n) = \text{sign}(\beta_{(1)}^n) \), and \( \beta_{(2)}^n = 0 \);

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

\[ |\hat{u}^n(1)| < |\beta_{(1)}^n|, \text{ and } \hat{u}^n(2) = 0 \]  

\((\star)\)

Another thing to notice is, since

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

then

\[ \hat{u}^n = \arg \min_{\tilde{u}^n} [||Y_n - X_n(u^n + \beta^n)||^2 + \lambda_n ||u^n + \beta^n||_1] \equiv \arg \min_{\tilde{u}^n} V_n(u^n) \]
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

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

\[
\left. \left| \frac{d}{d\beta_j} \left\| Y_n - X_n\beta \right\|_2^2 \right|_{\beta_j = \hat{\beta}_j^n} \right| \leq \lambda \quad \text{for } j \text{ s.t. } \hat{\beta}_j^n = 0
\] (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 (⋆).
Proposition 1 (cont.)

Proof (cont.)

So, assume:

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

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

\[
C_{11}^n(\sqrt{n}\hat{v}^n(1)) - W^n(1) = -\frac{\lambda_n}{2\sqrt{n}} \text{sign}(\beta^n_{(1)}) \tag{28}
\]

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 \( |\hat{v}^n(1)| < |\beta^n_{(1)}| \), then we finish the proof.
Proposition 1 (cont.)

Proof (cont.) - everything we know.

Aₙ implies:

\[ |(C^{n}_{11})^{-1} W^n(1)| < \sqrt{n}|\beta_{(1)}^n| - \frac{\lambda_n}{2n} |(C^{n}_{11})^{-1} \text{sign}(\beta_{(1)}^n)| \]  (31)

Bₙ and \( |C^{n}_{21}(C^{n}_{11})^{-1} \text{sign}(\beta_{(1)}^n)| \leq 1 - \eta \) (S.I.R.) implies:

\[ |C^{n}_{21}(C^{n}_{11})^{-1} W^n(1) - W^n(2)| \leq \frac{\lambda_n}{2\sqrt{n}} (1 - |C^{n}_{21}(C^{n}_{11})^{-1} \text{sign}(\beta_{(1)}^n)|) \]  (32)

as well as \( \hat{\nu}^n(2) = 0 \) and (28).
Then, (28) and (31) implies:

\[ |\hat{v}^n(1)| < |\beta^*_{(1)}| \tag{29} \]

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:

$$- \frac{1}{2\sqrt{n}} \frac{d\|Y_n - X_n(u^n + \beta^n)\|^2_2}{d(u^n_j + \beta^n_j)}|_{u^n_j=\hat{v}^n_j} = -\frac{\lambda_n}{2\sqrt{n}} \text{sign}(\hat{v}^n_j + \beta^n_j)$$

$$= -\frac{\lambda_n}{2\sqrt{n}} \text{sign}(\beta^n_j)$$

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

$$\frac{1}{2\sqrt{n}} \frac{d\|Y_n - X_n(u^n + \beta^n)\|^2_2}{d(u^n_j + \beta^n_j)}|_{\beta^n_j=\hat{v}^n_j} \leq \frac{\lambda_n}{2\sqrt{n}}$$

for $\hat{v}^n_j$ in $\hat{v}^n(2) = 0$
Theorem 3

\[ P(\hat{\beta}^n(\lambda_n) = s \beta^n) \geq 1 - O\left(\frac{p_n n^k}{\lambda^{2k}}\right) \rightarrow 1 \text{ as } n \rightarrow \infty. \]

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

Proof.

By proposition 1,

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

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

where

\[ z^n = (z_1^n, \ldots, z_q^n)' = (C_{11}^n)^{-1} W^n(1) \]
\[ \zeta^n = (\zeta_1^n, \ldots, \zeta_{p-q}^n)' = C_{21}^n(C_{11}^n)^{-1} W^n(1) - W^n(2) \]
\[ b = (b_1^n, \ldots, b_q^n) = (C_{11}^n)^{-1} \text{sign}(\beta_1^n). \]
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 \( \frac{\lambda_n}{\sqrt{n}} = o\left(n^{-\frac{c_2-c_1}{2}}\right) \)

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

\[
\sum_{i=1}^{p-q} P(|\zeta_i^n| \geq \frac{\lambda_i^n}{2\sqrt{n}} \eta_i) \leq \sum_{i=1}^{p-q} \frac{E|\zeta_i^n|^{2k}}{\left(\frac{\lambda_n}{\sqrt{n}} 2\eta_i\right)^{2k}} = (p - q)O\left(\frac{n^k}{\lambda_n^{2k}}\right) = O\left(\frac{pn^k}{\lambda_n^{2k}}\right)
\]
Theorem 3 (cont.)

Proof (cont.)

So,

\[ 1 - P(\hat{\beta}^n(\lambda_n) = \beta^n) \leq P(A^c_n) + P(B^c_n) \leq O\left(\frac{pn^k}{\lambda_{n}^{2k}}\right) \to 0 \text{ as } n \to \infty \]

for \( \frac{1}{p_n} (\frac{\lambda_n}{\sqrt{n}})^{2k} \to \infty \).
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 $\epsilon_i^n$, $z_i$’s and $\zeta_i$’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!