Lecture 36: LASSO and Thresholding

LASSO estimator

Consider linear model $X = Z\beta + \varepsilon$, $\beta \in \mathbb{R}^p$ and $\text{Var}(\varepsilon) = \sigma^2 I_n$. The ridge regression estimator of $\beta$ is obtained from

$$\min_{\beta \in \mathbb{R}^p} (\|X - Z\beta\|^2 + \lambda \|\beta\|^2)$$

If we change the $L_2$ penalty $\|\beta\|^2$ to the $L_1$ penalty $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$, where $\beta_j$ is the $j$th component of $\beta$, then the LASSO estimator is from

$$\min_{\beta \in \mathbb{R}^p} (\|X - Z\beta\|^2 + \lambda \|\beta\|_1)$$

Difference between LASSO and ridge regression:

- LASSO estimator does not have an explicit form.
- When a component of $\beta$ is 0, its LASSO estimator may be 0, but its ridge regression estimator is never 0.
- The minimization for LASSO is still for a convex objective function, but the objective function is not always differentiable.
- If $p < n$, $Z$ can be deterministic; if $p \geq n$, $Z$ must be random.
Notation

\( \mathcal{A} \) = the set of indices of non-zero coefficients of \( \beta \)

\[ \beta = (\beta_\mathcal{A}, \beta_{\mathcal{A}^c}), \dim(\beta_\mathcal{A}) = q, \dim(\beta_{\mathcal{A}^c}) = p - q; \ X = (X_\mathcal{A}, X_{\mathcal{A}^c}) \]

\[ C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} X^\tau X_\mathcal{A} & X^\tau X_{\mathcal{A}^c} \\ X^\tau c X_\mathcal{A} & X^\tau c X_{\mathcal{A}^c} \end{pmatrix} = \frac{1}{n} X^\tau X \]

Consistency

The LASSO estimator \( \hat{\beta} \) of \( \beta \) is strongly sign consistent if there exists \( \lambda = \lambda_n \) not depending on \( Y \) or \( X \) such that

\[ \lim_{n \to \infty} P \left( \text{sign}(\hat{\beta}) = \text{sign}(\beta) \right) = 1 \]

which implies variable selection consistent (since \( \text{sign}(a) = 0 \) if \( a = 0 \)),

\[ \lim_{n \to \infty} P \left( \hat{\mathcal{A}} = \mathcal{A} \right) = 1 \]

where \( \hat{\mathcal{A}} \) is the index set of nonzero components of \( \hat{\beta} \).

Strong Irrepresentable Condition (SIC)

There exists a vector \( \eta \) whose components are positive such that

\[ |C_{21} C_{11}^{-1} \text{sign}(\beta_\mathcal{A})| \leq 1 - \eta \text{ component-wise, where } |a| = (|a_1|, |a_2|, ...) \]

for \( a = (a_1, a_2, ...) \) and 1 is the vector of ones.
Critical Lemma

Under the SIC,

\[ P \left( \text{sign} (\hat{\beta}) = \text{sign}(\beta) \right) \geq P(A_n \cap B_n), \]

where

\[ A_n = \left\{ |C_{11}^{-1} W_\mathcal{A}| < \sqrt{n} |\beta_\mathcal{A}| - \frac{\lambda_n}{2\sqrt{n}} |C_{11}^{-1} \text{sign}(\beta_\mathcal{A})| \right\} \]

\[ B_n = \left\{ |C_{21} C_{11}^{-1} W_\mathcal{A} - W_\mathcal{A}^c| \leq \frac{\lambda_n}{2\sqrt{n}} \eta \right\} \]

\[ W_\mathcal{A} = \frac{1}{\sqrt{n}} X_\mathcal{A}^\tau \varepsilon \quad W_\mathcal{A}^c = \frac{1}{\sqrt{n}} X_\mathcal{A}^c \varepsilon \]

Karush-Kuhn-Tucker (KKT) condition

\( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p) \) is the LASSO estimator if and only if

\[ \frac{\partial \| Y - X\beta \|^2}{\partial \beta_j} \bigg|_{\beta_j=\hat{\beta}_j} = \begin{cases} \lambda \text{sign}(\hat{\beta}_j) & \hat{\beta}_j \neq 0 \\ \text{bounded by } \lambda \text{ in absolute value} & \hat{\beta}_j = 0 \end{cases} \]
Proof of the Lemma

Let $\hat{u} = \hat{\beta} - \beta$ and $V_n(u) = \sum_{i=1}^{n} [(\epsilon_i - X_i u)^2 - \epsilon_i]^2 + \lambda_n \|u + \beta\|_1$

Then $\hat{u} = \text{argmin} \ V_n(u)$

It can be verified that the KKT condition is equivalent to

$$C_{11}(\sqrt{n}\hat{u}_A) - W_A = \frac{\lambda_n}{2\sqrt{n}} \text{sign}(\beta_A),$$  \hspace{1cm} (1)

$$-\frac{\lambda_n}{2\sqrt{n}} \mathbf{1} \leq C_{21}(\sqrt{n}\hat{u}_A) - W_A c \leq \frac{\lambda_n}{2\sqrt{n}} \mathbf{1},$$  \hspace{1cm} (2)

$$|\hat{u}_A| < |\beta_A|$$  \hspace{1cm} (3)

We now show that on $A_n \cap B_n$, a solution $\hat{u}$ satisfying (1) and $\hat{u}_A c = 0$ must satisfy (2) and (3), and hence $\hat{\beta} = \hat{u} + \beta$ is a LASSO estimator.

In fact, LASSO estimator is unique.

First, (1) and $A_n$ holds imply (3).

Second, (1) and $B_n$ holds and the SIC imply (2).

Finally, a sufficient condition for $\text{sign}(\hat{\beta}) = \text{sign}(\beta)$ is $|\hat{u}_A| < |\beta_A|$ and $\hat{u}_A c = 0$.

This proves that if $A_n \cap B_n$ holds, $\text{sign}(\hat{\beta}) = \text{sign}(\beta)$. 
Theorem (strong sign consistency of LASSO)

(i) Assume that $\varepsilon_i$’s are iid with $E(\varepsilon_i^{2k}) < \infty$ for an integer $k > 0$, and there are positive constants $c_1 < c_2 \leq 1$, $M_1$, $M_2$, $M_3$, such that

C1: $n^{-1}\|Z_j\|^2 \leq M_1$ for any $j = 1,...,p$, $Z_j$ is the $j$th column of $Z$;

C2: The smallest eigenvalue of $C_{11} \geq M_2$;

C3: $q = O(n^{c_1})$;

C4: $n^{(1-c_2)/2}\min_{j \in \mathcal{A}} |\beta_j| \geq M_3$;

C5: $p = o(n^{(c_2-c_1)k})$.

Under SIC, if $\lambda$ is chosen with $\lambda = o(n^{1+c_2-c_1}/2)$ and $pn^k/\lambda^{2k} = o(1)$, then

$$P\left(\text{sign}(\hat{\beta}) = \text{sign}(\beta)\right) \geq 1 - O(pn^k/\lambda^{2k})$$

(ii) Assume that $\varepsilon_i$’s are iid normal and C1-C4 hold, and

C5a: $p = O(e^{nc_3})$ with a constant $c_3$, $0 \leq c_3 < c_2 - c_1$.

Under SIC, if $\lambda$ is chosen with $\lambda \propto n^{(1+c_4)/2}$, $c_4$ is a constant, $c_3 < c_4 < c_2 - c_1$, then

$$P\left(\text{sign}(\hat{\beta}) = \text{sign}(\beta)\right) \geq 1 - O(e^{nc_3})$$
Proof.

\( z_j = \) the \( j \)th component of \( C_{11}^{-1} W_\mathcal{A} \), \( j = 1, \ldots, q \)

\( \zeta_j = \) the \( j \)th component of \( C_{21} C_{11}^{-1} W_\mathcal{A} - W_\mathcal{A}^c \), \( j = 1, \ldots, p - q \)

\( b_j = \) the \( j \)th component of \( C_{11}^{-1} \text{sign}(\beta_\mathcal{A}) \), \( j = 1, \ldots, q \)

The condition \( E(\varepsilon_i^{2k}) < \infty \) implies that \( E(z_j^{2k}) < \infty \) and \( E(\zeta_j^{2k}) < \infty \)

By the lemma,

\[
P \left( \text{sign}(\hat{\beta}) \neq \text{sign}(\beta) \right) \leq 1 - P(A_n \cap B_n)
\]

\[
\leq \sum_{j \in \mathcal{A}} P \left( |z_j| \geq \sqrt{n|\beta_j| - \lambda b_j / 2\sqrt{n}} \right)
\]

\[
+ \sum_{j \in \mathcal{A}^c} P \left( |\zeta_j| \geq \lambda \eta_j / 2\sqrt{n} \right)
\]

\[
\leq \sum_{j \in \mathcal{A}} \frac{E|z_j|^{2k}}{n^k \beta_j^{2k}} + \sum_{j \in \mathcal{A}^c} \frac{E|\zeta_j|^{2k}}{(2\lambda \eta_j)^{2k} / n^k}
\]

\[
= qO(n^{-kc_2}) + (p - q)O(n^k / \lambda^{2k})
\]

\[
= o(pn^k / \lambda^{2k}) + O(pn^k / \lambda^{2k}) = O(pn^k / \lambda^{2k})
\]

This proves (i).
For (ii), the normality of $\varepsilon_j$ implies that $z_j$ and $\zeta_j$ are normal.
Instead of using Markov inequality, using $1 - \Phi(t) \leq t^{-1} e^{-t^2/2}$ leads to the result (ii).

**Advantage and disadvantage of using LASSO**

- Variable selection and parameter estimation at the same time
- It is very good in estimation and prediction, but it is often too conservative in variable selection.
- Need SIC.
- Population version of SIC.

\[ |\Sigma_{21} \Sigma_{11}^{-1} \text{sign}(\beta_{\mathcal{A}})| \leq 1 - \eta, \quad \Sigma_{kj} \text{ are submatrices of } \Sigma = \text{Var}(z_j), \text{ if } z_j's \text{ are iid, } z_j \text{ is the } j\text{th row of } Z. \]

**Improvements**

- Adaptive LASSO
- Group LASSO
- Elastic net (other penalties)
- LASSO plus thresholding (ridge regression plus thresholding)
Variable selection by thresholding

Can we do variable selection using $p$-values?
Or, can we simply select variables by using the values $\hat{\beta}_j, j = 1, \ldots, p$?
Here $\hat{\beta}_j$ is the $j$th component of $\hat{\beta}$, the least squares estimator of $\beta$.
For simplicity, assume that $X|Z \sim N(Z\beta, \sigma^2 I)$.
Then
\[
\hat{\beta}_j - \beta_j = \sum_{i=1}^{n} l_{ij} \varepsilon_i \bigg| Z \sim N \left( 0, \sigma^2 \sum_{i=1}^{n} l_{ij}^2 \right)
\]
where $\varepsilon_i$ and $l_{ij}$ are the $i$th components of $\varepsilon = X - Z\beta$ and $(Z^\tau Z)^{-1} z_i$.
$z_j$ is the $j$th row of $Z$.
Because
\[
1 - \Phi(t) \leq \frac{\sqrt{2\pi}}{t} e^{-t^2/2}, \quad t > 0
\]
where $\Phi$ is the standard normal cdf,
\[
P \left( |\hat{\beta}_j - \beta_j| > t \sqrt{\text{var}(\hat{\beta}|Z)} \bigg| Z \right) \leq \frac{2\sqrt{2\pi}}{t} e^{-t^2/2}, \quad t > 0
\]
Let $J_j$ be the $p$-vector whose $j$th component is 1 and other components are 0:
\[
l_{ij}^2 = [J_j^c (Z^\tau Z)^{-1} z_i]^2 \leq J_j^c (Z^\tau Z)^{-1} J_j z_i^c (Z^\tau Z)^{-1} z_i
\]
\[ \sum_{i=1}^{n} l_{ij}^2 \leq c_j \sum_{i=1}^{n} z_i^\tau (Z^\tau Z)^{-1} z_i = pc_j \leq p/\eta_n \]

where \( c_j \) is the \( j \)th diagonal element of \((Z^\tau Z)^{-1}\) and \( \eta_n \) is the smallest eigenvalue of \( Z^\tau Z \).

Thus, for any \( j \),

\[ P \left( |\hat{\beta}_j - \beta_j| > t \sigma \sqrt{p/\eta_n} \mid Z \right) \leq \frac{2\sqrt{2\pi}}{t} e^{-t^2/2}, \quad t > 0 \]

and (letting \( t = a_n/(\sigma \sqrt{p/\eta_n}) \))

\[ P \left( |\hat{\beta}_j - \beta_j| > a_n \mid Z \right) \leq Ce^{-a_n^2 \eta_n/(2\sigma^2 p)} \]

for some constant \( C > 0 \),

\[ P \left( \max_{j=1,\ldots,p} |\hat{\beta}_j - \beta_j| > a_n \mid Z \right) \leq pCe^{-a_n^2 \eta_n/(2\sigma^2 p)} \]

Suppose that \( p/n \to 0 \) and \( p/(\eta_n \log n) \to 0 \) (typically, \( \eta_n = O(n) \)). Then, we can choose \( a_n \) such that \( a_n \to 0 \) and \( a_n^2 (\eta_n \log n/p) \to \infty \) such that
\[ P \left( \max_{j=1,\ldots,p} |\hat{\beta}_j - \beta_j| > ca_n \mid Z \right) = O(n^{-s}) \]

for any \( c > 0 \) and some \( s \geq 1 \); e.g.,

\[ a_n = M \left( \frac{p}{\eta_n \log n} \right)^{\alpha} \]

for some constants \( M > 0 \) and \( \alpha \in (0, \frac{1}{2}) \).

What can we conclude from this?

Let

\[ \mathcal{A} = \{ j : \beta_j \neq 0 \} \quad \text{and} \quad \hat{\mathcal{A}} = \{ j : |\hat{\beta}_j| > a_n \} \]

That is, \( \hat{\mathcal{A}} \) contains the indices of variables we select by thresholding \( |\hat{\beta}_j| \) at \( a_n \).

Selection consistency:

\[ P \left( \hat{\mathcal{A}} \neq \mathcal{A} \mid Z \right) \leq P \left( |\hat{\beta}_j| > a_n, j \notin \mathcal{A} \mid Z \right) + P \left( |\hat{\beta}_j| \leq a_n, j \in \mathcal{A} \mid Z \right) \]

The first term on the right hand side is bounded by

\[ P \left( \max_{j=1,\ldots,p} |\hat{\beta}_j - \beta_j| > a_n \mid Z \right) = O(n^{-s}) \]
On the other hand, if we assume that

$$\min_{j \in \mathcal{A}} |\beta_j| \geq c_0 a_n$$

for some $c_0 > 1$, then

$$P \left( |\hat{\beta}_j| \leq a_n, j \in \mathcal{A} \mid Z \right) \leq P \left( |\beta_j| - |\hat{\beta}_j - \beta_j| \leq a_n, j \in \mathcal{A} \mid Z \right)$$

$$\leq P \left( c_0 a_n - |\hat{\beta}_j - \beta_j| \leq a_n, j \in \mathcal{A} \mid Z \right)$$

$$\leq P \left( \max_{j=1,\ldots,p} |\hat{\beta}_j - \beta_j| \geq (c_0 - 1)a_n \mid Z \right)$$

$$= O(n^{-s})$$

Hence, we have consistency; in fact, the convergence rate is $O(n^{-s})$.

We can also obtain similar results by thresholding $|\hat{\beta}_j| / \sqrt{\sum_{i=1}^{n} l_{ij}^2}$.

This approach may not work if $p/n \not\to 0$.

If $p > n$, then $Z^\top Z$ is not of full rank.

There exist several other approaches for the case where $p > n$; e.g., we replace $(Z^\top Z)^{-1}$ by some matrix, or use ridge regression instead of LSE.