Lecture 34: Ridge regression and LASSO

Ridge regression

Consider linear model $X = Z\beta + \varepsilon$, $\beta \in \mathscr{R}^p$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$.

The LSE is obtained from the minimization problem

$$\min_{\beta \in \mathscr{R}^p} \|X - Z\beta\|^2 \qquad (1)$$

A type of shrinkage estimator is obtained though (1) by adding a penalty on $\|\beta\|^2$, i.e.,

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

where $\lambda \geq 0$ is a constant controlling the penalization.

$$\frac{\partial}{\partial \beta}(\|X - Z\beta\|^2 + \lambda\|\beta\|^2) = -2Z^\tau(X - Z\beta) + 2\lambda\beta$$

which gives the solution to (2) as

$$\widehat{\beta}_\lambda = (Z^\tau Z + \lambda I_p)^{-1} Z^\tau X$$

This estimator is better than the LSE when $Z^\tau Z$ is nearly singular.

This gives a class of estimators called ridge regression estimators; in particular, $\lambda = 0$ gives the LSE.

Bias and covariance matrix

$$E(\widehat{\beta}_\lambda) = (Z^\tau Z + \lambda I_p)^{-1} Z^\tau E(X) = (Z^\tau Z + \lambda I_p)^{-1} Z^\tau Z \beta$$

The bias of $\widehat{\beta}_\lambda$ is then

$$b(\beta) = (Z^\tau Z + \lambda I_p)^{-1} Z^\tau Z \beta - \beta = -\lambda (Z^\tau Z + \lambda I_p)^{-1} \beta$$

The bias is not 0, but converges to 0 as $\lambda \to 0$.

$$\begin{aligned}
\mathrm{Var}(\widehat{\beta}_\lambda) &= (Z^\tau Z + \lambda I_p)^{-1} Z^\tau \mathrm{Var}(X) Z (Z^\tau Z + \lambda I_p)^{-1} \\
&= \sigma^2 (Z^\tau Z + \lambda I_p)^{-1} Z^\tau Z (Z^\tau Z + \lambda I_p)^{-1} \\
&= \sigma^2 (Z^\tau Z + \lambda I_p)^{-1} - \sigma^2 \lambda (Z^\tau Z + \lambda I_p)^{-2}
\end{aligned}$$

It can be seen that the variance converges to 0 if $\lambda \to \infty$ and to $\sigma^2 (Z^\tau Z)^{-1}$ if $\lambda \to 0$.

Combining the bias and variance, we get

$$\begin{aligned}
E\|\widehat{\beta}_\lambda - \beta\|^2 &= \|b(\beta)\|^2 + E\|\widehat{\beta}_\lambda - E\widehat{(\beta}_\lambda)\|^2 \\
&= \lambda^2 \|(Z^\tau Z + \lambda I_p)^{-1} \beta\|^2 + \sigma^2 \mathrm{tr}[Z^\tau Z (Z^\tau Z + \lambda I_p)^{-2}]
\end{aligned}$$

Theorem (Comparison between ridge regression and LSE)

Let $\widehat{\beta} = \widehat{\beta}_0$ be the LSE.

(i) If $0 < \lambda < 2\sigma^2/\|\beta\|^2$, then $E\|\widehat{\beta}_\lambda - \beta\|^2 < E\|\widehat{\beta} - \beta\|^2$.

(ii) Assume that the smallest eigenvalue of $Z^\tau Z = O(n)$.
If $\lambda > 2\sigma^2/\|\beta\|^2$, then $E\|\widehat{\beta}_\lambda - \beta\|^2 > E\|\widehat{\beta} - \beta\|^2$ for sufficiently large n; if $\lambda = 2\sigma^2/\|\beta\|^2$, then $E\|\widehat{\beta}_\lambda - \beta\|^2 = E\|\widehat{\beta} - \beta\|^2 + O(n^{-3})$.

Proof.

Let

$$\begin{aligned} A = {} & \sigma^2(Z^\tau Z)^{-1} - \sigma^2(Z^\tau Z + \lambda I_p)^{-1} Z^\tau Z (Z^\tau Z + \lambda I_p)^{-1} \\ & - \lambda^2 (Z^\tau Z + \lambda I_p)^{-1} \beta\beta^\tau (Z^\tau Z + \lambda I_p)^{-1} \end{aligned}$$

Then

$$\begin{aligned} (Z^\tau Z + \lambda I_p) A (Z^\tau Z + \lambda I_p) & = \sigma^2 (Z^\tau Z + \lambda I_p)(Z^\tau Z)^{-1}(Z^\tau Z + \lambda I_p) \\ & \quad - \sigma^2 Z^\tau Z - \lambda^2 \beta\beta^\tau \\ & = 2\lambda\sigma^2 I_p + \lambda^2\sigma^2 (Z^\tau Z)^{-1} - \lambda^2 \beta\beta^\tau \end{aligned}$$

Hence

$$A = (Z^\tau Z + \lambda I_p)^{-1} [2\lambda\sigma^2 I_p - \lambda^2\beta\beta^\tau + \lambda^2\sigma^2 (Z^\tau Z)^{-1}] (Z^\tau Z + \lambda I_p)^{-1}$$

Assume $\lambda > 0$ and $\beta \neq 0$.
Then

$$A > \lambda^2\sigma^2(Z^\tau Z + \lambda I_p)^{-1}(Z^\tau Z)^{-1}(Z^\tau Z + \lambda I_p)^{-1}$$

if and only if

$$2\sigma^2\lambda^{-1}I_p - \beta\beta^\tau > 0 \qquad \text{equivalent to} \qquad \lambda < 2\sigma^2/\|\beta\|^2$$

This can be shown as follows. If $2\sigma^2\lambda^{-1}I_p - \beta\beta^\tau > 0$, then $0 < \beta^\tau(2\sigma^2\lambda^{-1}I_p - \beta\beta^\tau)\beta = 2\sigma^2\lambda^{-1}\|\beta\|^2 - \|\beta\|^4$, which means $\lambda < 2\sigma^2/\|\beta\|^2$. On the other hand, if $\lambda < 2\sigma^2/\|\beta\|^2$, then $(2\sigma^2\lambda^{-1}I_p - \beta\beta^\tau)/\|\beta\|^2 = (2\sigma^2\lambda^{-1}\|\beta\|^{-2} - 1)I_p + I_p - \beta\beta^\tau/\|\beta\|^2 > 0$, because $I_p - \beta\beta^\tau/\|\beta\|^2$ is a projection matrix whose eigenvalues are either 0 or 1.

Since $\mathrm{Var}(\widehat{\beta}) = \sigma^2(Z^\tau Z)^{-1}$, using the formula for $\mathrm{Var}(\widehat{\beta}_\lambda)$ we obtain

$$E\|\widehat{\beta} - \beta\|^2 - E\|\widehat{\beta}_\lambda - \beta\|^2 = \mathrm{tr}(A)$$

Thus, (i) follows, and (ii) and (iii) follow from

$$\lambda^2\sigma^2(Z^\tau Z + \lambda I_p)^{-1}(Z^\tau Z)^{-1}(Z^\tau Z + \lambda I_p)^{-1} \leq \lambda^2\sigma^2(Z^\tau Z)^{-3}$$

The ridge regression is better if the noise to signal ratio is large.

High dimension problems

The dimension of β in a linear model is p (Z is $n \times p$)
In traditional applications: $p << n$; p is fixed when $n \to \infty$.
In modern applications, p is large; $p = p_n$ increases as n increases.

- $p = O(n^k)$: polynomial-type divergence rate
- $p = O(e^{n^\nu})$: ultra-high dimension, where ν is a constant < 1.

Non-identifiability of β

- $r = r_n$: rank of Z.
- The dimension of $\mathscr{R}(Z)$ is $r \leq n$.
- If $p > n$, then β is not identifiable.
  This means that there are β and $\tilde{\beta}$, $\beta \neq \tilde{\beta}$ but $Z\beta = Z\tilde{\beta}$ so that the data generated under the models with β and $\tilde{\beta}$ are the same.
- It is not possible to estimate all components of β consistently; we are not able to estimate something out of the data range.
- We can estimate consistently some useful functions of β.
- We can estimate the projection of β onto $\mathscr{R}(Z)$.
- Estimation of the projection is sufficient for many problems

Projection

- Singular value decomposition: $Z = PDQ^\tau$
  P: $n \times r$ matrix with $P^\tau P = I_r$ (identity matrix)
  Q: $p \times r$ matrix with $Q^\tau Q = I_r$
  D: $r \times r$ diagonal matrix of full rank
- Projection of β onto $\mathscr{R}(Z)$:
  $\theta = Z^\tau (ZZ^\tau)^- Z\beta = QQ^\tau \beta \in \mathscr{R}(Z)$
- $Z\theta = PDQ^\tau (QQ^\tau \beta) = PDQ^\tau \beta = Z\beta$
- The model

$$Y = Z\beta + \varepsilon \qquad \text{is the same as} \qquad Y = Z\theta + \varepsilon$$

Ridge regression estimator of θ

$$\widehat{\theta} = (Z^\tau Z + h_n I_p)^{-1} Z^\tau X \qquad h_n > 0$$

We only need to invert an $n \times n$ matrix, because

$$(Z^\tau Z + h_n I_p)^{-1} Z^\tau = Z^\tau (ZZ^\tau + h_n I_n)^{-1}$$

$\widehat{\theta}$ is always in $\mathscr{R}(Z)$

Derivation of the bias of ridge regression estimator

Let $\Gamma = \begin{pmatrix} Q & Q_\perp \end{pmatrix}$, $Q^\tau Q_\perp = 0$, $\Gamma\Gamma^\tau = \Gamma^\tau\Gamma = I_p$.
Then

$$
\begin{aligned}
\mathrm{bias}(\widehat{\theta}) &= E(\widehat{\theta}) - \theta \\
&= (Z^\tau Z + h_n I_p)^{-1} Z^\tau Z\theta - \theta \\
&= -(h_n^{-1} Z^\tau Z + I_p)^{-1}\theta \\
&= -\Gamma(h_n^{-1}\Gamma^\tau Z^\tau Z\Gamma + I_p)^{-1}\Gamma^\tau QQ^\tau\theta \\
&= -\begin{pmatrix} Q & Q_\perp \end{pmatrix} \begin{pmatrix} (h_n^{-1}D^2 + I_r)^{-1} & 0 \\ 0 & I_{p-r} \end{pmatrix} \begin{pmatrix} Q^\tau \\ Q_\perp^\tau \end{pmatrix} QQ^\tau\theta \\
&= -\begin{pmatrix} Q(h_n^{-1}D^2 + I_r)^{-1} & Q_\perp \end{pmatrix} \begin{pmatrix} Q^\tau\theta \\ 0 \end{pmatrix} \\
&= -Q(h_n^{-1}D^2 + I_r)^{-1} Q^\tau\theta \\
&= -Q \begin{pmatrix} (1 + d_{1n}/h_n)^{-1} & & \\ & \ddots & \\ & & (1 + d_{rn}/h_n)^{-1} \end{pmatrix} Q^\tau\theta
\end{aligned}
$$

where $d_{jn} > 0$ is the jth diagonal element of D^2 (eigenvalue of $Z^\tau Z$).

Thus,

$$\|\mathrm{bias}(\widehat{\theta})\|^2 = \theta^\tau Q(h_n^{-1}D^2 + I_r)^{-2}Q^\tau\theta$$
$$\leq \max_{1\leq j\leq r}(1 + d_{jn}/h_n)^{-2}\theta^\tau QQ^\tau\theta$$
$$\leq h_n^2 d_{1n}^{-2}\|\theta\|^2$$

For the variance,

$$\mathrm{Var}(\widehat{\theta}) = \sigma^2(Z^\tau Z + h_n I_p)^{-1}Z^\tau Z(Z^\tau Z + h_n I_p)^{-1}$$
$$\leq \sigma^2 h_n^{-1} I_p$$

Theorem (Consistency of $\widehat{\theta}$)

Assume that

(C1) $d_{1n}^{-1} = O(n^{-\eta})$, $\eta \leq 1$ and η does not depend on n.

(C2) $\|\theta\| = O(n^\tau)$, $\tau < \eta$ and τ does not depend on n.

Then

(i) As $n \to \infty$, $E(\ell^\tau\widehat{\theta} - \ell^\tau\theta)^2 = O(h_n^{-1}) + O(h_n^2 n^{-2(\eta-\tau)})$ uniformly over p-dimensional deterministic vector ℓ with $\|\ell\| = 1$.

(ii) $n^{-1}E\|Z\widehat{\theta} - Z\theta\|^2 = O(r_n n^{-1}) + O(h_n^2 n^{-(1+\eta-2\tau)})$.

Remarks

- (C2) means that θ is sparse; without any condition, the order of $\|\theta\|^2$ could be p.
- $\|\theta\| \leq \|\beta\|$ so that (C2) holds if β is sparse.
- For any fixed $\ell'\theta$, $\ell'\widehat{\theta}$ is consistent if $h_n \to \infty$ and $h_n n^{-(\eta-\tau)} \to 0$.
- $\widehat{\theta}$ is not sparse even if θ is sparse.
- Typically $r_n/n \not\to 0$ so $\widehat{\theta}$ is not L_2-consistent.
- The reason (ii) is interesting is that

$$n^{-1}E\|Z\widehat{\theta} - Z\theta\|^2 = n^{-1}E\|X_* - Z\widehat{\theta}\|^2 - \sigma^2,$$

where X_* is an independent copy of X and $n^{-1}E\|X_* - Z\widehat{\theta}\|^2$ is the average prediction mean squared error.

Problem of the ridge regression estimator

When $p < n$, $\theta = \beta$ has many zero components, the ridge regression estimator does not have any zero components, although it has many small components.

LASSO estimator

Consider linear model $X = Z\beta + \varepsilon$, $\beta \in \mathscr{R}^p$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$.

The ridge regression estimator of β is obtained from

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

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

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

Difference between LASSO and ridge regression:

- LASSO estimator does not have an explicit form.
- When a component of β 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.
- Although LASSO is still defined when $p > n$, it is usually used in the case where $p < n$.
- If $p < n$, Z can be deterministic or random.

Notation

$\mathscr{A}$ = the set of indices of non-zero coefficients of β

$\beta = (\beta_{\mathscr{A}}, \beta_{\mathscr{A}^c})$, $\dim(\beta_{\mathscr{A}}) = q$, $\dim(\beta_{\mathscr{A}^c}) = p - q$; $X = (X_{\mathscr{A}}, X_{\mathscr{A}^c})$

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

Consistency

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

$$\lim_{n\to\infty} P\left(\mathrm{sign}(\widehat{\beta}) = \mathrm{sign}(\beta)\right) = 1$$

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

$$\lim_{n\to\infty} P\left(\widehat{\mathscr{A}} = \mathscr{A}\right) = 1$$

where $\widehat{\mathscr{A}}$ is the index set of nonzero components of $\widehat{\beta}$.

Strong Irrepresentable Condition (SIC)

There exists a vector η whose components are positive such that $|C_{21} C_{11}^{-1} \mathrm{sign}(\beta_{\mathscr{A}})| \le 1 - \eta$ component-wise, where $|a| = (|a_1|, |a_2|, \ldots)$ for $a = (a_1, a_2, \ldots)$ and 1 is the vector of ones.

Critical Lemma

Under the SIC,

$$P\left(\operatorname{sign}(\widehat{\beta}) = \operatorname{sign}(\beta)\right) \geq P(A_n \cap B_n),$$

where

$$A_n = \left\{ |C_{11}^{-1} W_{\mathscr{A}}| < \sqrt{n}|\beta_{\mathscr{A}}| - \frac{\lambda_n}{2\sqrt{n}} |C_{11}^{-1} \operatorname{sign}(\beta_{\mathscr{A}})| \right\}$$

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

$$W_{\mathscr{A}} = \frac{1}{\sqrt{n}} X_{\mathscr{A}}^{\tau} \varepsilon \qquad W_{\mathscr{A}^c} = \frac{1}{\sqrt{n}} X_{\mathscr{A}^c}^{\tau} \varepsilon$$

Karush-Kuhn-Tuker (KKT) condition

$\widehat{\beta} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)$ is the LASSO estimator if and only if

$$\left. \frac{\partial \|Y - X\beta\|^2}{\partial \beta_j} \right|_{\beta_j = \widehat{\beta}_j} = \begin{cases} \lambda \operatorname{sign}(\widehat{\beta}_j) & \widehat{\beta}_j \neq 0 \\ \text{bounded by } \lambda \text{ in absolute value} & \widehat{\beta}_j = 0 \end{cases}$$

Proof of the Lamma

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

Then $\widehat{u} = \operatorname{argmin} V_n(u)$

It can be verified that the KKT condition is equivalent to

$$C_{11}(\sqrt{n}\widehat{u}_{\mathscr{A}}) - W_{\mathscr{A}} = \frac{\lambda_n}{2\sqrt{n}} \operatorname{sign}(\beta_{\mathscr{A}}), \tag{3}$$

$$-\frac{\lambda_n}{2\sqrt{n}} 1 \le C_{21}(\sqrt{n}\widehat{u}_{\mathscr{A}}) - W_{\mathscr{A}^c} \le \frac{\lambda_n}{2\sqrt{n}} 1, \tag{4}$$

$$|\widehat{u}_{\mathscr{A}}| < |\beta_{\mathscr{A}}| \tag{5}$$

We now show that on $A_n \cap B_n$, a solution $\widehat{u}$ satisfying (3) and $\widehat{u}_{\mathscr{A}^c} = 0$ must satisfy (4) and (5), and hence $\widehat{\beta} = \widehat{u} + \beta$ is a LASSO estimator.

In fact, LASSO estimator is unique.

First, (3) and A_n holds imply (5).

Second, (3) and B_n holds and the SIC imply (4).

Finally, a sufficient condition for $\operatorname{sign}(\widehat{\beta}) = \operatorname{sign}(\beta)$ is $|\widehat{u}_{\mathscr{A}}| < |\beta_{\mathscr{A}}|$ and $\widehat{u}_{\mathscr{A}^c} = 0$.

This proves that if $A_n \cap B_n$ holds, $\operatorname{sign}(\widehat{\beta}) = \operatorname{sign}(\beta)$.

Theorem (strong sign consistency of LASSO)

(i) Assume that ε_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 jth column of Z;
C2: The smallest eignvalue of $C_{11} \geq M_2$;
C3: $q = O(n^{c_1})$;
C4: $n^{(1-c_2)/2} \min_{j\in\mathscr{A}} |\beta_j| \geq M_3$;
C5: $p = o(n^{(c_2-c_1)k})$.

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

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

(ii) Assume that ε_i's are iid normal and C1-C4 hold, and

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

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

$$P\left(\mathrm{sign}(\widehat{\beta}) = \mathrm{sign}(\beta)\right) \geq 1 - O(e^{n^{c_3}})$$

Proof.

$z_j =$ the jth component of $C_{11}^{-1} W_{\mathscr{A}}$, $j = 1, ..., q$
$\zeta_j =$ the jth component of $C_{21} C_{11}^{-1} W_{\mathscr{A}} - W_{\mathscr{A}^c}$, $j = 1, ..., p - q$
$b_j =$ the jth component of $C_{11}^{-1} \mathrm{sign}(\beta_{\mathscr{A}})$, $j = 1, ..., q$
The condition $E(\varepsilon_i^{2k}) < \infty$ implies that $E(z_j^{2k}) < \infty$ and $E(\zeta_j^{2k}) < \infty$
By the lemma,

$$\begin{aligned}
P\left(\mathrm{sign}(\widehat{\beta}) \neq \mathrm{sign}(\beta)\right) &\leq 1 - P(A_n \cap B_n) \\
&\leq \sum_{j \in \mathscr{A}} P\left(|z_j| \geq \sqrt{n}|\beta_j| - \lambda b_j / 2\sqrt{n}\right) \\
&\quad + \sum_{j \in \mathscr{A}^c} P\left(|\zeta_j| \geq \lambda \eta_j / 2\sqrt{n}\right) \\
&\leq \sum_{j \in \mathscr{A}} \frac{E|z_j|^{2k}}{n^k \beta_j^{2k}} + \sum_{j \in \mathscr{A}^c} \frac{E|\zeta_j|^{2k}}{(2\lambda \eta_j)^{2k} / n^k} \\
&= qO(n^{-kc_2}) + (p - q)O(n^k / \lambda^{2k}) \\
&= o(pn^k / \lambda^{2k}) + O(pn^k / \lambda^{2k}) = O(pn^k / \lambda^{2k})
\end{aligned}$$

This proves (i).

For (ii), the normality of ε_j implies that z_j and ζ_j are normal.
Instead of using Markov inequality, using $1-\Phi(t) \le t^{-1}e^{-t^2/2}$ 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.
  $|\Sigma_{21}\Sigma_{11}^{-1}\mathrm{sign}(\beta_{\mathscr{A}})| \le 1-\eta$, Σ_{kj} are submatrices of $\Sigma = \mathrm{Var}(z_j)$, if z_j's are iid, z_j is the jth row of Z.

Improvements

- Adaptive LASSO
- Group LASSO
- Elastic net (other penalties)
- LASSO plus thresholding (ridge regression plus threshodling)