Introduction to Statistical Machine Learning with Applications in Econometrics

Lecture 10: Ridge Regression, LASSO and Adaptive LASSO

Instructor: Ma, Jun

Renmin University of China

November 25, 2021

High-dimensional data

- In data sets with with more regressors than the number of observations, for the $n \times k$ matrix $\mathbf{X}$ collecting n observations on k variables, we have $k > n$ and therefore rank $(\mathbf{X}) \leq n < k$.
- The OLS ($\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$) cannot be computed since $\mathbf{X}^\top \mathbf{X}$ is non-singular.
- Perfect multicolinearity: there will be exact linear combinations among the regressors.
- In this lecture, still assume $n > k$ (classical environment).

Ridge regression

- Ridge regression estimator:

$$\widehat{\beta}_{\lambda}^{R} = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}b\|^2 + \lambda \|b\|^2,$$

  where $b = (b_1, ..., b_k)^\top$ and $\|b\| = \sqrt{\sum_{j=1}^{k} b_j^2}$.
- The first-order conditions:

$$-\mathbf{X}^\top \left(\mathbf{Y} - \mathbf{X}\widehat{\beta}_{\lambda}^{R}\right) + \lambda \widehat{\beta}_{\lambda}^{R} = \mathbf{0} \Longrightarrow \widehat{\beta}_{\lambda}^{R} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_k\right)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

- Ridge regression is biased:

$$\mathrm{E}\left[\widehat{\beta}_{\lambda}^{R} \mid \mathbf{X}\right] = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_k\right)^{-1} \mathbf{X}^\top \mathbf{X}\beta \neq \beta.$$

- While Ridge is biased, regularization reduces the variance of out-of-sample prediction.
- Larger values of λ would shrink the Ridge estimates more toward zero and prevent from overfitting. $\Longrightarrow$ More bias, less variance.
- By balancing between the variance and bias, it is possible to improve out-of-sample prediction over OLS.
- Every regressor is assigned a non-zero regression coefficient.

LASSO

- Suppose that the causal model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + U_i$$

  is sparse: only a few explanatory variables are relevant variables with non-zero coefficients.
- It is desirable to have an estimation procedure that can automatically in data-dependent manner shrink the coefficients on irrelevant regressors to zero.
- Neither OLS nor ridge can produce exactly zero regression coefficients.
- The LASSO estimator:

$$\widehat{\beta}_{\lambda}^{L} = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}b\|^2 + \lambda \|b\|_1 ,$$

  where $b = (b_1, ..., b_k)^\top$ and $\|b\|_1 = \sum_{j=1}^{k} |b_j|$.
- In general, the LASSO problem does not have an analytical solution, and therefore the LASSO estimates must be computed numerically.

- Ridge regression: for every λ, there exists $s > 0$ such that $\widehat{\beta}^R_\lambda$ solves

$$\begin{aligned} \widehat{\beta}^R_\lambda &= \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}b\|^2 \\ &\quad \text{subject to } \|b\|^2 \leq s. \end{aligned}$$

- LASSO: for every λ, there exists $s > 0$ such that $\widehat{\beta}^L_\lambda$ solves

$$\begin{aligned} \widehat{\beta}^L_\lambda &= \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \|\mathbf{Y} - \mathbf{X}b\|^2 \\ &\quad \text{subject to } \|b\|_1 \leq s. \end{aligned}$$

- An exceptional case with an analytical solution is when the regressors are orthogonal and normalized: $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_k$. We use it to illustrate the LASSO mechanism.
- Suppose that $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_k$ and let $\widehat{\beta}_\lambda = \left(\widehat{\beta}_{\lambda,1}, ..., \widehat{\beta}_{\lambda,k}\right)^\top$ denote the LASSO estimator

$$\widehat{\beta}_\lambda = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{Y} - \mathbf{X}b\|^2 + \lambda \|b\|_1 .$$

- Note that the sum of squares is scaled by $1/2$. This is without loss of generality since we can adjust λ accordingly.
- Assume that $k < n$. Let $\widetilde{\beta} = \left(\widetilde{\beta}_1, ..., \widetilde{\beta}_k\right)^\top$ denote the OLS.
- The LASSO estimator satisfies

$$\widehat{\beta}_{\lambda,j} = \operatorname{sgn}\left(\widetilde{\beta}_j\right)\left(\left|\widetilde{\beta}_j\right| - \lambda\right)_+ ,$$

  where $(x)_+ = \max\{x, 0\}$ and $\operatorname{sgn}(x)$ denotes the sign of x: $\operatorname{sgn}(x) = (-1)\,1(x < 0) + 1(x > 0)$.
- LASSO detects near-zero coefficients and shrink them to zero, which is equivalent to dropping such variables from the model.

- Similarly to the Ridge regression, LASSO estimates are biased due to shrinkage.
- Post-LASSO estimation: after LASSO, use OLS to regress the explained variable only on explanatory variables that survived LASSO selection.
- The motivation of Post-LASSO is to avoid the shrinkage bias. In such a case, LASSO is used only as a selection method.

Cross validation for LASSO

- $\lambda \downarrow 0$: $\widehat{\beta}_\lambda \to$ OLS; $\lambda \uparrow \infty$: $\widehat{\beta}_\lambda \to 0$.
- Randomly split the sample into a training set and a validation set:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_T \\ \mathbf{Y}_V \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} \mathbf{X}_T \\ \mathbf{X}_V \end{pmatrix},$$

  where $\mathbf{Y}_T$ is $n_T \times 1$, $\mathbf{Y}_V$ is $n_V \times 1$, $\mathbf{X}_T$ is $n_T \times k$, $\mathbf{X}_V$ is $n_V \times k$ and $n_V + n_T = n$.
- Fix any $\lambda > 0$, compute $\widehat{\beta}_{\lambda,T}$ using only observations in the training set:

$$\widehat{\beta}_{\lambda,T} = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \mathbf{Y}_T - \mathbf{X}_T b \right\|^2 + \lambda \left\| b \right\|_1 .$$

- The cross-validation (CV) estimate of the test MSE:

$$RSS_{CV}(\lambda) = \left\| \mathbf{Y}_V - \mathbf{X}_V \widehat{\beta}_{\lambda,T} \right\|^2 .$$

- Repeat the procedure K times with different training and validation sets. The resulting procedure is the K-fold cross validation.

- Randomly divide the data into K (approximately equal-sized) parts: $C_1, C_2, ..., C_K$ with $\cup_{j=1}^{K} C_j = \{1, 2, ..., n\}$, where C_j denote the indices of observations in part j. Let n_j denote the number of indices in C_j.
- For the $n \times k$ matrix $\mathbf{X}$, $\mathbf{X}_{C_j}$ denotes the $n_j \times k$ sub-matrix of $\mathbf{X}$ with observations in C_j. Similarly, $\mathbf{Y}_{C_j}$ denotes the $n_j \times 1$ sub-vector of $\mathbf{Y}$ with observations in C_j.
- Split

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_{C_1} \\ \vdots \\ \mathbf{Y}_{C_K} \end{pmatrix} \text{ and } \mathbf{X} = \begin{pmatrix} \mathbf{X}_{C_1} \\ \vdots \\ \mathbf{X}_{C_K} \end{pmatrix}.$$

- Denote $C_{-j} = \cup_{l \neq j} C_l$ and

$$\begin{aligned} \widehat{\beta}_{\lambda,-j} &= \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \left\| \mathbf{Y}_{C_{-j}} - \mathbf{X}_{C_{-j}} b \right\|^2 + \lambda \|b\|_1 \\ RSS_j(\lambda) &= \left\| \mathbf{Y}_{C_j} - \mathbf{X}_{C_j} \widehat{\beta}_{\lambda,-j} \right\|^2 . \end{aligned}$$

- Average to get the K-fold cross validation estimate of the test MSE:

$$RSS_{CV}(\lambda) = \frac{1}{K} \sum_{j=1}^{K} RSS_j(\lambda) .$$

- The standard error:

$$\begin{aligned} SE_{CV}(\lambda) &= \sqrt{\frac{\widehat{\mathrm{Var}}_{CV}(\lambda)}{K}}, \text{ where} \\ \widehat{\mathrm{Var}}_{CV}(\lambda) &= \frac{1}{K-1} \sum_{j=1}^{K} \left(RSS_j(\lambda) - RSS_{CV}(\lambda) \right)^2 . \end{aligned}$$

- Cross-validated choice of tuning parameter:

$$\widehat{\lambda}_{CV} = \underset{\lambda > 0}{\operatorname{argmin}} RSS_{CV}(\lambda) .$$

Inference-optimal choice of λ

- Choice of the tuning parameter λ that is optimal for inference should be such that all relevant regressors are included (non-zero coefficients) while all irrelevant regressors are excluded (coefficients shrunken to zero).
- Optimal prediction and inference can not be achieved simultaneously: the best λ from the prediction perspective is different from the best λ when the goal is to accurately find the relevant regressors.
- Suppose that the true model is

$$Y_i = \beta_1 X_{i,1} + U_i$$

with $\mathrm{E}\left[U_i \mid X_{i,1}, X_{i,2}\right] = 0$ and $\mathrm{E}\left[U_i^2 \mid X_{i,1}, X_{i,2}\right] = \sigma^2 > 0$. Assume that $\mathrm{E}\left[X_{i,1} X_{i,2}\right] = 0$.

- Let

$$\begin{pmatrix} \widehat{\beta}_{\lambda,1} \\ \widehat{\beta}_{\lambda,2} \end{pmatrix} = \underset{b_1, b_2}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Y_i - b_1 X_{i,1} - b_2 X_{i,2}\right)^2 + \lambda \left(|b_1| + |b_2|\right).$$

- Let $\left(X_{0,1}, X_{0,2}, Y_0\right)$ be an (independent) unseen observation. Denote $\widehat{Y}_0 = X_{0,1}\widehat{\beta}_{\lambda,1} + X_{0,2}\widehat{\beta}_{\lambda,2}$. Note that

$$Y_0 - \widehat{Y}_0 = U_0 - \left(\widehat{\beta}_{\lambda,1} - \beta_1\right) X_{0,1} - \widehat{\beta}_{\lambda,2} X_{0,2}.$$

- The test MSE:

$$\begin{aligned} \mathrm{E}\left[\left(Y_0 - \widehat{Y}_0\right)^2\right] &= \sigma^2 + \mathrm{E}\left[\left(\widehat{\beta}_{\lambda,1} - \beta_1\right)^2\right] \mathrm{E}\left[X_{0,1}^2\right] \\ &\quad + \mathrm{E}\left[\widehat{\beta}_{\lambda,2}^2\right] \mathrm{E}\left[X_{0,2}^2\right]. \end{aligned}$$

- Ideally, we would like $\mathrm{E}\left[\left(\widehat{\beta}_{\lambda,1} - \beta_1\right)^2\right]$ to be as small as possible and shrink $\widehat{\beta}_{\lambda,2} = 0$. The second condition requires heavy penalty (large λ). However, this results in large bias of $\widehat{\beta}_{\lambda,1}$.

Optimal rate for λ

- Suppose that

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{U} \\ \mathbf{U} \mid \mathbf{X} &\sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right) \\ \frac{1}{n}\mathbf{X}^\top \mathbf{X} &= \mathbf{I}_k . \end{aligned}$$

- The OLS:

$$\widetilde{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} = \beta + \frac{1}{n}\mathbf{X}^\top \mathbf{U} = \beta + \frac{1}{n}\sum_{i=1}^{n} X_i U_i .$$

- Define the LASSO:

$$\widehat{\beta}_\lambda = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{n} \|\mathbf{Y} - \mathbf{X}b\|^2 + \lambda \|b\|_1 .$$

Note that the RSS is scaled by $1/n$. Then,

$$\widehat{\beta}_{\lambda,j} = \operatorname{sgn}\left(\widetilde{\beta}_j\right)\left(\left|\widetilde{\beta}_j\right| - \lambda\right)_+ .$$

- To reduce the bias for the LASSO estimators of relevant regressors, we want to keep λ as small as possible.
- At the same time, we need λ to be large enough so that it can shrink to zero the coefficients on irrelevant regressors.
- If $\beta_j = 0$, then $\widetilde{\beta}_j \sim \mathrm{N}\left(0, \sigma^2/n\right)$ and $\sqrt{n}\widetilde{\beta}_j \sim \mathrm{N}\left(0, \sigma^2\right)$.
- $\lambda = \lambda_n$ should converge to zero at a rate that is slightly slower than $n^{-1/2}$: $\lambda_n = C \cdot \log(n)/\sqrt{n}$, where $C > 0$ is a constant. Then,

$$\begin{aligned}
\Pr\left[\left|\widetilde{\beta}_j\right| \leq \lambda_n\right] &= \Pr\left[\left|\sqrt{n}\widetilde{\beta}_j\right| \leq C \cdot \log(n)\right] \\
&= \Pr\left[-C \cdot \log(n) \leq \mathrm{N}\left(0, \sigma^2\right) \leq C \cdot \log(n)\right] \\
&\to 1
\end{aligned}$$

and therefore, $\Pr\left[\widehat{\beta}_{\lambda_n, j} = 0\right] \to 1$.

- If $\beta_l \neq 0$, then $\widetilde{\beta}_l \sim \mathrm{N}\left(\beta_l, \sigma^2/n\right)$ and

$$\begin{aligned}
&\Pr\left[\left|\widetilde{\beta}_l\right| \leq \lambda_n\right] = \Pr\left[\left|\sqrt{n}\left(\widetilde{\beta}_l - \beta_l\right) + \sqrt{n}\beta_l\right| \leq C \cdot \log(n)\right] \\
&= \Pr\left[-C \cdot \log(n) - \sqrt{n}\beta_l \leq \mathrm{N}\left(0, \sigma^2\right) \leq C \cdot \log(n) - \sqrt{n}\beta_l\right] \to 0
\end{aligned}$$

  and therefore, $\Pr\left[\widehat{\beta}_{\lambda_n, l} \neq 0\right] \to 1$.

- If $\|\mathbf{Y} - \mathbf{X}b\|^2$ is not standardized by $1/n$, the optimal rate for λ should be $\sqrt{n} \cdot \log(n)$.

Weighted LASSO

- LASSO solves:

$$\min_{b_1,\ldots,b_k} \sum_{i=1}^{n} \left(Y_i - b_1 X_{i,1} - \cdots - b_k X_{i,k}\right)^2 + \lambda \sum_{j=1}^{k} \left|b_j\right|.$$

- Weighted LASSO applies different weights to different coefficients. Let $w_1, ..., w_n$ be some (possibly data-dependent) non-negative weights. The weighted LASSO solves

$$\min_{b_1,\ldots,b_k} \sum_{i=1}^{n} \left(Y_i - b_1 X_{i,1} - \cdots - b_k X_{i,k}\right)^2 + \lambda \sum_{j=1}^{k} w_j \left|b_j\right|.$$

- The weighted LASSO allows for different amount of shrinkage and penalization for different coefficients.
- For example, by setting $w_1 = 0$, no shrinkage would be applied to the coefficient of the first regressor. This is useful for regressors that always should be included in the model.
- No penalty is typically applied to the intercept.

Adaptive LASSO (Zou, 2006)

- Let $\widetilde{\beta}_1, ..., \widetilde{\beta}_k$ denote the OLS estimates. The adaptive LASSO uses the weights

$$w_j = \frac{1}{\left|\widetilde{\beta}_j\right|}, \; j = 1, ..., k.$$

- Since $\widetilde{\beta}_j \to_p \beta_j$, $\widetilde{\beta}_j$ is a good initial guess for β_j.
- The shrinkage is tuned by these weights: when β_j is far away from zero, we expect that $\left|\widetilde{\beta}_j\right|$ is large, the weight is small and less shrinkage is imposed on the j-th coefficient; when β_j is indeed zero, we expect that $\left|\widetilde{\beta}_j\right|$ is small and heavy shrinkage is imposed.

Oracle procedure

- We call the smallest sub-model consisting of only explanatory variables with non-zero coefficients the "correct" model.
- A statistical procedure is an oracle procedure if with probability approaching one it selects the correct model and the estimator of the selected coefficients are asymptotically normal with no zero bias and an asymptotic variance you would get if you knew the correct model.
- For the linear model, denote $\mathcal{A} = \left\{ j : \beta_j \neq 0 \right\}$ (the indices of the relevant regressors), which characterizes the correct model.
- Suppose an estimation procedure produced a vector of estimates $\widehat{\beta} = \left(\widehat{\beta}_1, ..., \widehat{\beta}_k \right)^\top$. It is an oracle procedure if the following two conditions hold as $n \uparrow \infty$:
  - $\Pr\left[\widehat{\mathcal{A}} = \mathcal{A} \right] \to 1$, where $\widehat{\mathcal{A}} = \left\{ j : \widehat{\beta}_j \neq 0 \right\}$.
  - $\sqrt{n} \left(\widehat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}} \right) \to_d \mathrm{N}\left(0, V_{\mathcal{A}}\right)$, where $V_{\mathcal{A}}$ is the asymptotic variance matrix when the correct model $\mathcal{A}$ is known.

- The first condition requires that with the probability approaching one, the procedure selects the right regressors.
- The second condition states that the asymptotic distribution of the estimator for β's on the true regressors is the same as one would have obtain by regressing Y_i only on the regressors in $\mathcal{A}$.
- OLS is not an oracle procedure, as the probability that an OLS estimator is exactly equal to zero is zero. Similarly, the ridge regression is not an oracle procedure.
- Zou (2006) shows that LASSO in its original form is not an oracle procedure.

The oracle properties of the adaptive LASSO

- We assume the classical model with $k < n$ and i.i.d. observations (X_i, Y_i), $i = 1, 2, ..., n$ generated from the model:

$$\begin{aligned} Y &= X^\top \beta + U \\ \mathrm{E}\left[U \mid X\right] &= 0 \\ \mathrm{E}\left[U^2 \mid X\right] &= \sigma^2. \end{aligned}$$

- The homoskedasticity assumption $\mathrm{E}\left[U^2 \mid X\right] = \sigma^2$ can be dropped.
- Denote $\mathbf{V} = \mathrm{E}\left[XX^\top\right]$. The OLS estimator $\widetilde{\beta}$ is asymptotically normal:

$$\sqrt{n}\left(\widetilde{\beta} - \beta\right) \to_d \mathrm{N}\left(0, \sigma^2 \mathbf{V}^{-1}\right).$$

- The adaptive LASSO:

$$\widehat{\beta}_\lambda = \underset{b_1,\ldots,b_k}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - b_1 X_{i,1} - \cdots - b_k X_{i,k}\right)^2 + \lambda \sum_{j=1}^{k} \frac{|b_j|}{\left|\widetilde{\beta}_j\right|}.$$

- The second oracle property rules out asymptotic bias and requires a small λ, while the first oracle property requires a large λ.

- Assume that $\sqrt{n}\lambda_n \downarrow 0$ and $n\lambda_n \uparrow \infty$ as $n \uparrow \infty$. Let

$$\widehat{\mathcal{A}} = \left\{ j : \widehat{\beta}_{\lambda_n, j} \neq 0 \right\}.$$

- Then, the adaptive LASSO is an oracle procedure: as $n \uparrow \infty$,
  - $\Pr\left[\widehat{\mathcal{A}} = \mathcal{A}\right] \to 1$;
  - $\sqrt{n}\left(\widehat{\beta}_{\lambda_n, \mathcal{A}} - \beta_{\mathcal{A}}\right) \to_d \mathrm{N}\left(0, \sigma^2 \mathbf{V}_{\mathcal{A}}^{-1}\right)$, where $\mathbf{V}_{\mathcal{A}} = \mathrm{E}\left[X_{\mathcal{A}} X_{\mathcal{A}}^{\top}\right]$.
- As a model selection procedure, the adaptive LASSO does not need to assume homoskedasticity.
- The adaptive LASSO estimator for $\beta_{\mathcal{A}}$ is as good as the OLS with perfect knowledge of $\mathcal{A}$.

Choice of λ for adaptive LASSO

- The theorem requires a large penalty $n\lambda_n \uparrow \infty$ to detect the zero coefficients. To have smaller bias, we would like λ_n to be not too large.
- Typically, we choose

$$\lambda_n = C \cdot \frac{\log(n)}{n}.$$

- Here, $C > 0$ is a constant. We select C by cross validation.

Confidence intervals

- Run OLS with the regressors in $\widehat{\mathcal{A}}$ and get the standard errors.
- The confidence intervals for the non-zero coefficients can be centered around the post-LASSO OLS from the previous step or the adaptive LASSO estimator. Asymptotically, both are correct. Modified procedures are available and will be introduced later.