Tutorial 3: Shrinkage Effects of Ridge and Lasso

Teaching Team of STA314

Department of Statistical Sciences
University of Toronto

Monday September 23, 2024

Recall: Ridge Regression

$$\hat{\boldsymbol{\beta}}_\lambda^R = \underset{\boldsymbol{\beta}=(\beta_0,\ldots,\beta_p)\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{RSS} + \lambda \sum_{j=1}^{p}\beta_j^2 .$$

where $\lambda \geq 0$ is the tuning parameter and $\lambda \sum_{j=1}^{p} \beta_j^2$ is a shrinkage/regularization penalty.

Recall: Lasso Regression

The lasso coefficients, $\hat{\boldsymbol{\beta}}_\lambda^L$, minimize the quantity

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

In the case of the lasso, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly zero when the tuning parameter λ is sufficiently large.

Toy Example: the Shrinkage Effects of Ridge and Lasso

- Assume that $n = p$ and $\mathbf{X} = \mathbf{I}_n$. We force the intercept term $\beta_0 = 0$.
- In this way,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_p \end{bmatrix}.$$

- We assume

$$\mathbb{E}[\epsilon_j] = 0, \qquad \mathbb{E}[\epsilon_j^2] = \sigma^2, \qquad \forall j \in \{1, \dots, p\}.$$

Toy Example: OLS Estimator

- The OLS approach is to find $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p} (y_j - \beta_j)^2.$$

This gives the OLS estimator

$$\hat{\beta}_j = y_j, \qquad \forall j \in \{1, \ldots, p\}.$$

Toy Example: Ridge Estimator

- The ridge regression looks for $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

This leads to the ridge estimator

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda}, \qquad \forall j \in \{1, \ldots, p\}.$$

Since $\lambda \geq 0$, the magnitude of each estimated coefficient is proportionally shrinked towards 0.

Toy Example: Lasso Estimator

- Lasso looks for $\beta_1, \ldots, \beta_p$ that minimize

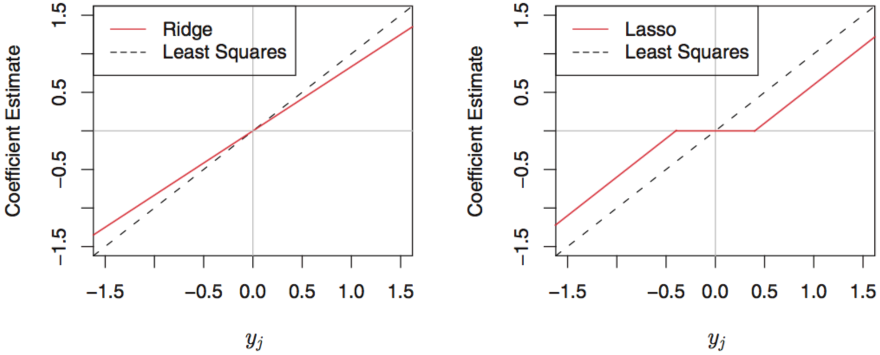
$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Which results ir

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases}$$

The estimated coefficients from Lasso are shrinked by a fixed amount, and equal to zero when the OLS estimate is in $[-\lambda/2, \lambda/2]$. The above shrinkage is known as **soft-thresholding**.

Toy Example: An Illustrative Figure



Toy Example: Bias and Variance of the OLS

Recall

$$y_j = \beta_j + \epsilon_j, \qquad \forall j \in \{1, \ldots, p\}.$$

For any $j \in \{1, \ldots, p\}$, the OLS estimator $\hat{\beta}_j = y_j$ satisfies

- **Bias**:

$$\mathbb{E}[\hat{\beta}_j] = \mathbb{E}[y_j] = \mathbb{E}[\beta_j + \epsilon_j] = \beta_j$$

$$\mathbb{E}[\hat{\beta}_j^R] - \beta_j = 0$$

- **Variance**:

$$\text{Var}(\hat{\beta}_j) = \text{Var}(\epsilon_j) = \sigma^2$$

Toy Example: MSE of the OLS

- **Mean squared error** of the jth coefficient:

$$\mathbb{E}\left[\left(\hat{\beta}_j - \beta_j\right)^2\right] = \left(\mathbb{E}[\hat{\beta}_j] - \beta_j\right)^2 + \mathrm{Var}(\hat{\beta}_j) = \sigma^2$$

- **Mean squared error** of all p coefficients:

$$\mathbb{E}\left[\sum_{j=1}^{p}\left(\hat{\beta}_j - \beta_j\right)^2\right] = p\sigma^2.$$

Toy Example: Bias and Variance of Ridge

Recall

$$y_j = \beta_j + \epsilon_j, \qquad \forall j \in \{1, \ldots, p\}.$$

For any $j \in \{1, \ldots, p\}$, the ridge estimator with tuning parameter λ,

$$\hat{\beta}_j^R = \frac{y_j}{1+\lambda},$$

satisfies

- **Bias**:

$$\mathbb{E}[\hat{\beta}_j^R] = \mathbb{E}\left[\frac{y_j}{1+\lambda}\right] = \mathbb{E}\left[\frac{\beta_j + \epsilon_j}{1+\lambda}\right] = \frac{\beta_j}{1+\lambda}$$

$$\mathbb{E}[\hat{\beta}_j^R] - \beta_j = \frac{-\lambda\beta_j}{1+\lambda}$$

- **Variance**:

$$\mathrm{Var}(\hat{\beta}_j^R) = \mathrm{Var}\left(\frac{\epsilon_j}{1+\lambda}\right) = \frac{\sigma^2}{(1+\lambda)^2}$$

Toy Example: MSE of the Ridge

- **Mean squared error** of the jth coefficient:

$$\mathbb{E}\left[\left(\hat{\beta}_j^R - \beta_j\right)^2\right] = \left(\mathbb{E}[\hat{\beta}_j^R] - \beta_j\right)^2 + \mathrm{Var}(\hat{\beta}_j^R)$$
$$= \left(\frac{\beta_j}{1+\lambda} - \beta_j\right)^2 + \frac{\sigma^2}{(1+\lambda)^2}$$
$$= \frac{\lambda^2\beta_j^2}{(1+\lambda)^2} + \frac{\sigma^2}{(1+\lambda)^2}.$$

  Recall that $\mathbb{E}[(\hat{\beta}_j - \beta_j)^2] = \sigma^2$.

- **Mean squared error** of all p coefficients:

$$\mathbb{E}\left[\sum_{j=1}^{p}\left(\hat{\beta}_j^R - \beta_j\right)^2\right] = \frac{\lambda^2 \sum_{j=1}^{p} \beta_j^2 + p\sigma^2}{(1+\lambda)^2}.$$

Quiz Time!