Theoretical Understanding the Regularization Effect of Ridge

Xin Bing

Department of Statistical Sciences
University of Toronto

A theoretical understanding of the role of regularization

Consider the linear regression

$$Y = X^\top \boldsymbol{\beta} + \epsilon.$$

Suppose we have i.i.d. observations $(x_1, y_1), \ldots, (x_n, y_n)$. Further assume the design matrix $\mathbf{X}$ is deterministic and orthonormal, i.e.

$$\frac{1}{n}\mathbf{X}^\top \mathbf{X} = \mathbf{I}_p.$$

Consider the ridge estimator $\hat{\boldsymbol{\beta}}_\lambda^R$ of $\boldsymbol{\beta}$ for any given regularization parameter $\lambda \geq 0$. Let $\hat{\boldsymbol{\beta}}$ be the OLS estimator of $\boldsymbol{\beta}$.

We now contrast the behaviour of the ridge estimator with that of the OLS estimator side by side.

- Criteria:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{X}\beta||_2^2$$

$$\hat{\beta}_\lambda^R = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, ||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2.$$

- Closed-form solutions:

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \frac{1}{n}\mathbf{X}^\top\mathbf{y}$$

$$\hat{\beta}_\lambda^R = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y} = \frac{1}{n+\lambda}\mathbf{X}^\top\mathbf{y}.$$

- We examine their statistical properties of estimating β in terms of
  - bias
  - variance
  - mean squared error

Bias of the OLS and ridge estimators

- OLS: unbiased

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top \mathbf{y}\right] = \mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right]$$
$$= \frac{1}{n}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top \boldsymbol{\epsilon}\right] = \boldsymbol{\beta}.$$

- Ridge: biased

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda^R] = \mathbb{E}\left[\frac{1}{n+\lambda}\mathbf{X}^\top \mathbf{y}\right] = \mathbb{E}\left[\frac{1}{n+\lambda}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right]$$
$$= \frac{1}{n+\lambda}\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \mathbb{E}\left[\frac{1}{n+\lambda}\mathbf{X}^\top \boldsymbol{\epsilon}\right]$$
$$= \frac{n}{n+\lambda}\boldsymbol{\beta}$$
$$= \boldsymbol{\beta} - \frac{\lambda}{n+\lambda}\boldsymbol{\beta}.$$

Variance of the OLS and ridge estimators

- OLS:

$$\mathrm{Cov}(\hat{\beta}) = \frac{1}{n^2}\mathbf{X}^\top \mathrm{Cov}(\mathbf{y})\mathbf{X} = \frac{\sigma^2}{n^2}\mathbf{X}^\top\mathbf{X} = \frac{\sigma^2}{n}\mathbf{I}_p.$$

- Ridge:

$$\mathrm{Cov}(\hat{\beta}^R_\lambda) = \frac{1}{(n+\lambda)^2}\mathbf{X}^\top \mathrm{Cov}(\mathbf{y})\mathbf{X} = \frac{\sigma^2 n}{(n+\lambda)^2}\mathbf{I}_p.$$

Estimation error of the OLS and ridge estimators

- OLS:

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2] &= \mathrm{Cov}(\hat{\boldsymbol{\beta}}) + \left\|\mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}\right\|_2^2 \\ &= \underbrace{\frac{\sigma^2 p}{n}}_{\textbf{Variance}} + \underbrace{0}_{\textbf{Bias}}.\end{aligned}$$

- Ridge:

$$\begin{aligned}\mathbb{E}[\|\hat{\boldsymbol{\beta}}_\lambda^R - \boldsymbol{\beta}\|_2^2] &= \mathrm{Cov}(\hat{\boldsymbol{\beta}}_\lambda^R) + \left\|\mathbb{E}[\hat{\boldsymbol{\beta}}_\lambda^R] - \boldsymbol{\beta}\right\|_2^2 \\ &= \underbrace{\frac{\sigma^2 pn}{(n+\lambda)^2}}_{\textbf{Variance}} + \underbrace{\left(\frac{\lambda}{n+\lambda}\right)^2 \|\boldsymbol{\beta}\|_2^2}_{\textbf{Bias}}.\end{aligned}$$

Remark: Ridge estimator has smaller variance by paying extra bias as the price. **This is the essential idea of regularization!** The balance between variance and bias of ridge is controlled by the magnitude of λ.

Same phenomenon for prediction

Since we predict $X = x$ by

- OLS:

$$\hat{y} = x^\top \hat{\boldsymbol{\beta}}$$

- Ridge:

$$\hat{y}_\lambda^R = x^\top \hat{\boldsymbol{\beta}}_\lambda^R$$

Regularization controlled by λ has the same effects on prediction MSE.

Same phenomenon for the Lasso

The same idea holds for the Lasso. But the analysis of the MSE estimation error of the Lasso is less straightforward than that of Ridge.