Chapter 3 Ridge Regression and Kernel Methods

AMA4680
Fu Guanxing
guanxing.fu@polyu.edu.hk

Outline

- Ridge regression.
- LASSO.
- Kernel method.

Ridge Regression

- **Motivating reason for ridge regression**
- Introduction of ridge regression

Singular value decomposition (SVD)

Theorem (SVD). Any m by n matrix A can be factored into

$$A = U\Sigma V' = (\text{orthogonal})(\text{diagonal})(\text{orthogonal}).$$

The columns of $U = U_{m\times m}$ are eigenvectors of AA', and the columns of $V = V_{n\times n}$ are eigenvectors of $A'A$. The r singular values on the diagonal of $\Sigma = \Sigma_{m\times n}$ are the square roots of the nonzero eigenvalues of both AA' and $A'A$.

- No need to know how to calculate U, Σ and V; python does it!
- If $n = m$,
- If $n > m$,
- If $n < m$,

OLS is sometimes sensitive (1/4)

- Consider the linear regression

$$Y = Z\beta + \varepsilon$$

- Assume $Z'Z$ invertible. Least square estimate:

$$\hat{\beta} = (Z'Z)^{-1}Z'y.$$

- Collinearity vs. perfect collinearity.
- One application of SVD: OLS is sensitive to small change of data, when collinearity exists.[1]

[1] This is one problem arising from collinearity, which does not violate Gauss-Markov assumptions.

OLS is sometimes sensitive (2/4)

- [2]Using SVD to Z:

$$Z = \underbrace{\begin{bmatrix} Z_{1,1} & \cdots & Z_{1,r+1} \\ \vdots & \ddots & \vdots \\ Z_{n,1} & \cdots & Z_{n,r+1} \end{bmatrix}}_{n\times(r+1)} = U\Sigma V'$$

$$= \underbrace{\begin{bmatrix} U_{1,1} & \cdots & U_{1,n} \\ \vdots & \ddots & \vdots \\ U_{n,1} & \cdots & U_{n,n} \end{bmatrix}}_{U,\ n\times n} \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_{r+1} \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}}_{\Sigma,\ n\times(r+1)} \underbrace{\begin{bmatrix} V_{1,1} & \cdots & V_{r+1,1} \\ \vdots & \ddots & \vdots \\ V_{1,r+1} & \cdots & V_{r+1,r+1} \end{bmatrix}}_{V',\ (r+1)\times(r+1)}.$$

- $\sigma_i \neq 0$, for $i = 1, \cdots, r+1$.
- Generally, $n > r+1$. The analysis also applies to the case $n \leq r+1$, when $\Sigma = ...$

[2]In linear regression, the first column of Z is $(1, 1, \cdots, 1)'$. The following analysis applies to general Z.

OLS is sometimes sensitive (3/4)

- Calculate $(Z'Z)^{-1}Z'$:

$$Z'Z = V\Sigma'U'U\Sigma V' = V\Sigma'\Sigma V'$$

$$= V \underbrace{\begin{bmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & \ddots & \vdots \\ & & \sigma_{r+1} & 0 & \cdots & 0 \end{bmatrix}}_{\Sigma',\ (r+1)\times n} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_{r+1} & \\ 0 & \cdots & 0 & \\ \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & \end{bmatrix}}_{\Sigma,\ n\times(r+1)} V'$$

$$= V\,\mathrm{diag}\{\sigma_1^2, \cdots, \sigma_{r+1}^2\}V'$$

$$\Rightarrow (Z'Z)^{-1} = V\,\mathrm{diag}\left\{\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \cdots, \frac{1}{\sigma_{r+1}^2}\right\}V' = V(\Sigma'\Sigma)^{-1}V'$$

$$\Rightarrow (Z'Z)^{-1}Z' = V\,\mathrm{diag}\left\{\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \cdots, \frac{1}{\sigma_{r+1}^2}\right\} \underbrace{\begin{bmatrix} \sigma_1 & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & \ddots & \vdots \\ & & \sigma_{r+1} & 0 & \cdots & 0 \end{bmatrix}}_{\Sigma',\ (r+1)\times n} U'$$

OLS is sometimes sensitive (4/4)

- OLS estimator follows

$$(Z'Z)^{-1}Z'y = V \underbrace{\begin{bmatrix} \sigma_1^{-1} & & & 0 & \cdots & 0 \\ & \ddots & & \vdots & \ddots & \vdots \\ & & \sigma_{r+1}^{-1} & 0 & \cdots & 0 \end{bmatrix}}_{(\Sigma'\Sigma)^{-1}\Sigma',\ (r+1)\times n} \underbrace{\begin{bmatrix} (U'y)_1 \\ \vdots \\ (U'y)_n \end{bmatrix}}_{U'y,\ n\times 1}$$

$$= V \underbrace{\begin{bmatrix} \sigma_1^{-1}(U'y)_1 \\ \vdots \\ \sigma_{r+1}^{-1}(U'y)_{r+1} \end{bmatrix}}_{(\Sigma'\Sigma)^{-1}\Sigma'U'y,\ n\times 1}$$

- Recall SVD: σ_i is the square root of nonzero eigenvalues of $Z'Z$. Collinearity implies that $\det(Z'Z)$ is small $\Leftrightarrow$ $\sigma_1^2 \cdots \sigma_{r+1}^2$ small $\Leftrightarrow$ some $|\sigma_i|$ must be small!
- OLS estimator is sensitive to the change of y.
- Solution: change $(Z'Z)^{-1}$ to

$$(Z'Z + \lambda I)^{-1} = V \operatorname{diag}\left(\frac{1}{\sigma_1^2 + \lambda}, \cdots, \frac{1}{\sigma_{r+1}^2 + \lambda}\right) V', \quad \lambda > 0.$$

$\curvearrowright$ **ridge regression!**

Ridge Regression

- Motivating reason for ridge regression
- **Introduction of ridge regression**

Ridge regression

- W.l.o.g., we consider the linear regression without intercept:

$$Y = \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_{r+1} Z_{r+1} + \varepsilon.$$

- **Ridge regression** is defined by[3]:

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \arg\min_{\beta \in \mathbb{R}^{r+1}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_1 z_{1i} - \cdots - \beta_{r+1} z_{r+1,i})^2 + \lambda \sum_{j=1}^{r+1} \beta_j^2 \\ &= \arg\min_{\beta \in \mathbb{R}^{r+1}} \frac{1}{n} \|y - Z\beta\|^2 + \lambda \|\beta\|^2.\end{aligned}$$

  Here λ is called the **tuning parameter**, which controls the strength of the penalty term. Note that:

  - When $\lambda = 0$, we go back to the OLS estimate.
  - When $\lambda \to \infty$, we have $\hat{\beta}^{\text{ridge}} = 0$.

[3] $\lambda \|\beta\|^2$ is a penalization: large β cannot be optimal $\Rightarrow$ address the issue of collinearity.

Solve ridge regression

- Let

$$\mathcal{E}(\beta) := \frac{1}{n}\|y - Z\beta\|^2 + \lambda\|\beta\|^2.$$

- Least square estimate in ridge regression is

$$\hat{\beta}^{\text{ridge}} = (Z'Z + n\lambda I)^{-1} Z'y.$$

- Recall that

$$Z'Z = V\Sigma' U'U\Sigma V' = V\Sigma'\Sigma V' = V\text{diag}\{\sigma_1^2, \cdots, \sigma_{r+1}^2\}V',$$

  which implies that

$$Z'Z + n\lambda I = V\text{diag}\{\sigma_1^2 + n\lambda, \cdots, \sigma_{r+1}^2 + n\lambda\}V'.$$

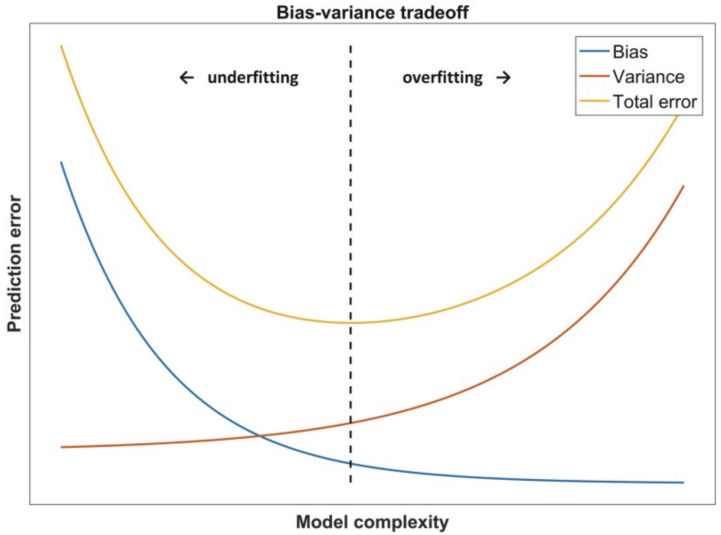
- Estimators with regularization are robust.

Bias-variance trade-off (1/2)

- Let β be a vector of (unknown) parameters, and $\hat{\beta}$ be a vector of estimators. Define MSE (mean squared error) and bias by

$$\text{MSE}(\hat{\beta}) = \mathbb{E}\left[\|\hat{\beta} - \beta\|^2\right], \quad \text{Bias}(\hat{\beta}) = \beta - \mathbb{E}[\hat{\beta}],$$

- $\text{MSE}(\hat{\beta}) = \left\|\text{Bias}(\hat{\beta})\right\|^2 + \text{Tr}(\text{Cov}(\hat{\beta}))$.



- • Less model parameters ⇒ underfitting ⇒ **large bias**.
  • More model parameters ⇒ learning focuses too much on some specific dataset ⇒ overfitting ⇒ does not fit new data well[4] ⇒ **large variance**.
- Tradeoff!
- The bias-variance tradeoff is the property of a model that the variance of the estimators across samples can be reduced by increasing the bias in the estimators.

[4] The *generalization ability* is weak.

Bias-Variance trade-off (2/2)

Theorem

$$\mathrm{MSE}(\hat{\beta}) = \left\|\mathrm{Bias}(\hat{\beta})\right\|^2 + \mathrm{Tr}(\mathrm{Cov}(\hat{\beta})).$$

- Definition of trace: for $C \in \mathbb{R}^{n\times n}$, $\mathsf{Tr}(C) := \mathsf{Trace}(C) := \sum_{i=1}^{n} C_{ii}$.
- For $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times m}$,

$$\begin{aligned}
\mathsf{Tr}(AB) = \sum_{i=1}^{m} (AB)_{ii} &= \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} B_{ji} \right) \\
&= \sum_{j=1}^{n} \sum_{i=1}^{m} B_{ji} A_{ij} \\
&= \sum_{j=1}^{n} (BA)_{jj} \\
&= \mathsf{Tr}(BA).
\end{aligned}$$

Bias of OLS and ridge regression (1/1)

- Bias of OLS: Recall $\hat{\beta}^{\text{ols}} = (Z'Z)^{-1}Z'(Z\beta + \varepsilon) = \beta + (Z'Z)^{-1}Z'\varepsilon$. It implies that $\mathbb{E}[\hat{\beta}^{\text{ols}}] = \beta$ and $\text{Bias}(\hat{\beta}^{\text{ols}}) = 0$.
- Bias of ridge: Recall

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= (Z'Z + n\lambda I)^{-1}Z'(Z\beta + \varepsilon) \\ &= (Z'Z + n\lambda I)^{-1}Z'Z\beta + (Z'Z + n\lambda I)^{-1}Z'\varepsilon,\end{aligned}$$

which implies that

$$\begin{aligned}\text{Bias}(\hat{\beta}^{\text{ridge}}) &= \beta - \mathbb{E}[\hat{\beta}^{\text{ridge}}] = \beta - (Z'Z + n\lambda I)^{-1}Z'Z\beta \\ &= \beta - (Z'Z + n\lambda I)^{-1}(Z'Z + n\lambda I - n\lambda I)\beta \\ &= \beta - (Z'Z + n\lambda I)^{-1}(Z'Z + n\lambda I)\beta + (Z'Z + n\lambda I)^{-1}n\lambda\beta \\ &= n\lambda(Z'Z + n\lambda I)^{-1}\beta.\end{aligned}$$

Remark. In ridge regression, even if singular value of Z is very small, $(Z'Z + n\lambda I)^{-1}$ can be controlled. However, the bias in not zero.

Variance of OLS and ridge regression (1/3)

- Variance of OLS:

$$\mathrm{Tr}(\mathrm{Cov}(\hat{\beta}^{\mathrm{ols}})) = \sigma^2 \sum_{i=1}^{r+1} \frac{1}{\sigma_i^2}.$$

- Variance of ridge regression:

$$\mathrm{Tr}(\mathrm{Cov}(\hat{\beta}^{\mathrm{ridge}})) = \sigma^2 \sum_{i=1}^{r+1} \frac{\sigma_i^2}{(\sigma_i^2 + n\lambda)^2}.$$

- When there is some $\sigma_i \to 0$, we see that $\mathrm{Tr}(\mathrm{Cov}(\hat{\beta}^{\mathrm{ols}})) \to \infty$ but $\mathrm{Tr}(\mathrm{Cov}(\hat{\beta}^{\mathrm{ridge}}))$ does not.
- We see that, when some $\sigma_i \to 0$, $\mathrm{MSE}(\hat{\beta}^{\mathrm{ols}}) \to \infty$ but $\mathrm{MSE}(\hat{\beta}^{\mathrm{ridge}})$ remains finite. By using the regularization parameter λ, some bias is introduced to $\hat{\beta}^{\mathrm{ridge}}$ yet the total MSE is possibly decreased. This phenomenon is referred to as the *bias-variance trade-off*.

Variance of OLS and ridge regression (2/3)

- Variance of OLS:

Recall $\hat{\beta}^{\text{ols}} = (Z'Z)^{-1}Z'(Z\beta + \varepsilon) = \beta + (Z'Z)^{-1}Z'\varepsilon$. Then

$$\begin{aligned}\text{Tr}(\text{Cov}(\hat{\beta}^{\text{ols}})) &= \text{Tr}\left((Z'Z)^{-1}Z'\text{Cov}(\varepsilon)Z(Z'Z)^{-1}\right) \\ &= \text{Tr}\left(\sigma^2(Z'Z)^{-1}Z'IZ(Z'Z)^{-1}\right) = \sigma^2\text{Tr}((Z'Z)^{-1}).\end{aligned}$$

Recall $Z = U\text{diag}\{\sigma_1, \ldots, \sigma_{r+1}\}V'$. Then

$$\begin{aligned}\text{Tr}((Z'Z)^{-1}) &= \text{Tr}\{(V\text{diag}\{\sigma_1, \cdots, \sigma_{r+1}\}U'U\text{diag}\{\sigma_1, \cdots, \sigma_{r+1}\}V')^{-1}\} \\ &= \text{Tr}\{(V\text{diag}\{\sigma_1^2, \cdots, \sigma_{r+1}^2\}V')^{-1}\} \\ &= \text{Tr}\{V\text{diag}\{\sigma_1^{-2}, \cdots, \sigma_{r+1}^{-2}\}V'\} \\ &= \text{Tr}\{\text{diag}\{\sigma_1^{-2}, \cdots, \sigma_{r+1}^{-2}\}V'V\} = \text{Tr}\{\text{diag}\{\sigma_1^{-2}, \cdots, \sigma_{r+1}^{-2}\}\} \\ &= \sum_{i=1}^{r+1}\frac{1}{\sigma_i^2}.\end{aligned}$$

Variance of OLS and ridge regression (3/3)

- Variance of ridge regression:

  Recall $\hat{\beta}^{\text{ridge}} - \mathbb{E}[\hat{\beta}^{\text{ridge}}] = (Z'Z + n\lambda I)^{-1} Z'\varepsilon$.

LASSO

LASSO: introduction

- The **l**east **a**bsolute **s**hrinkage and **s**election **o**perator (LASSO) is defined by

$$\begin{aligned}\hat{\beta}^{\text{lasso}} &= \arg\min_{\beta\in\mathbb{R}^p} \|y - Z\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i| \\ &= \arg\min_{\beta\in\mathbb{R}^p} \|y - Z\beta\|^2 + \lambda \|\beta\|_1 .\end{aligned}$$

- The only difference between the lasso problem and ridge regression is that the latter uses an ℓ_2 **regularization** $\|\beta\|_2^2$, while the former uses an ℓ_1 **regularization** $\|\beta\|_1$.
- The **tuning parameter** λ controls the strength of the penalty, and (like ridge regression) as $\lambda \searrow 0$ we get that $\hat{\beta}^{\text{lasso}}$ goes back to the least square estimate; as $\lambda \nearrow \infty$, $\hat{\beta}^{\text{lasso}} \to 0$.
- Both ridge regression and LASSO can address the issue of overfitting.
- Why LASSO? ℓ_1 regularization leads to *sparsity* (some weights in the regression are zero) $\Rightarrow$ *feature selection* (refer to AMA4602)

Proposition. $\|\hat{\beta}^{\text{ridge}}\|_2$ and $\|\hat{\beta}^{\text{lasso}}\|_1$ decrease as a function of its tuning parameter λ.