Effectively Solving Linear Ridge Regression

Student ID：518021911150 Class：F1803405 Name：周睿文

Problem:

- Numerical computation of matrix inversion of $Z^T Z$ is expensive
- Instead we could use singular value decomposition (SVD) to lower the computation cost:

  $$Z = UDV^T$$

  where:

  - $U = (u_1, u_2, \cdots, u_p)$ is an $n \times p$ orthogonal matrix
  - $D = \text{diag}(d_1, d_2, \cdots, d_p)$ is a $p \times p$ diagonal matrix
    consisting of the singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$
  - $V^T = (v_1^T, v_2^T, \cdots, v_p^T)$ is a $p \times p$ orthogonal matrix

- Proof:

$$\begin{aligned}\hat{\beta}_\lambda^{\text{ridge}} &= \left(Z^T Z + \lambda I_p\right)^{-1} Z^T y \\ &= V \operatorname*{diag}_j \left(\frac{d_j}{d_j^2 + \lambda}\right) U^T y\end{aligned}$$

Proof:

From the objective of linear ridge regression, we have:

$$J(\beta) = (y - Z\beta)^T (y - Z\beta) + \lambda \beta^T \beta$$

To minimize the cost function $J(\beta)$, we need:

$$\frac{\partial J}{\partial \beta} = -2Z^T (y - Z\beta) + 2\lambda\beta = 0$$

Hence we get:

$$\begin{aligned}Z^T y &= Z^T Z\beta + \lambda\beta \\ \beta &= \left(Z^T Z + \lambda I_p\right)^{-1} Z^T y\end{aligned}$$

As will be shown later, the $p \times p$ matrix $\left(Z^T Z + \lambda I_p\right)$ is invertible as long as $\lambda > 0$.

According to the definition of singular value decomposition, we let $Z = UDV^T$, where:

- $U = (u_1, u_2, \cdots, u_p)$ is an $n \times p$ orthogonal matrix
- $D = \mathrm{diag}(d_1, d_2, \cdots, d_p)$ is a $p \times p$ diagonal matrix
  consisting of the singular values $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$
- $V^T = (v_1^T, v_2^T, \cdots, v_p^T)$ is a $p \times p$ orthogonal matrix

Therefore, we obtain:

$$\begin{aligned} Z^T Z &= \left(UDV^T\right)^T \left(UDV^T\right) \\ &= VD^T U^T UDV^T \\ &= VD^T DV^T \\ &= VD^2 V^T \end{aligned}$$

As V is an orthogonal matrix, we have $V^T V = I_p$, so we can get:

$$\begin{aligned} Z^T Z + \lambda I_p &= VD^2 V^T + \lambda I_p \\ &= V\left(D^2 + \lambda I_p\right) V^T \\ &= V \operatorname*{diag}_j \left(d_j^2 + \lambda\right) V^T \end{aligned}$$

As long as $\lambda > 0$, we have $d_j^2 + \lambda > 0, \forall j \in \{0, 1, \cdots, p\}$, so we can get $\left(Z^T Z + \lambda I_p\right)^{-1}$ as following:

$$\begin{aligned} V \operatorname*{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) V^T \left(Z^T Z + \lambda I_p\right) &= V \operatorname*{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) V^T V \operatorname*{diag}_j \left(d_j^2 + \lambda\right) V^T \\ &= V \operatorname*{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) \operatorname*{diag}_j \left(d_j^2 + \lambda\right) V^T \\ &= VV^T = I_p \\ \Longrightarrow \left(Z^T Z + \lambda I_p\right)^{-1} &= V \operatorname*{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) V^T \end{aligned}$$

Therefore, we reach the final result:

$$\begin{aligned} \beta &= \left(Z^T Z + \lambda I_p\right)^{-1} Z^T y \\ &= V \operatorname*{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) V^T \left(VD^T U^T\right) y \\ &= V \operatorname*{diag}_j \left(\frac{1}{d_j^2 + \lambda}\right) \operatorname*{diag}_j \left(d_j\right) U^T y \\ &= V \operatorname*{diag}_j \left(\frac{d_j}{d_j^2 + \lambda}\right) U^T y \end{aligned}$$

■