Lecture September 11

Interpretation of OLS & Ridge

$$\tilde{y}^{OLS} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}\, y = Ay$$

$$A^2 = \left(X(X^TX)^{-1}X^T\right)^2 A = X(X^TX)^{-1}X^T$$

$$= A \in \mathbb{R}^{n\times n}$$

$\tilde{y}^{OLS}$ is an orthogonal projection of y onto the space spanned by the columns of X

with $\hat{\beta}$ we have the residuals

$$\hat{\varepsilon} = y - \tilde{y}^{OLS} = y - X\hat{\beta}$$

$$= \left(I - X(X^TX)^{-1}X^T\right) y$$

The residuals are the projections of y onto the orthogonal complement of the space spanned by the columns

$I = A + Q$

$Q = I - A$

of X

<u>SVD</u> $\quad X \in \mathbb{R}^{n \times p} \; (\mathbb{C}^{n \times p})$

$$X = U \Sigma V^T$$

$U \in \mathbb{R}^{n \times n} \qquad U^T U = U U^T = \mathbb{1}$

$\Sigma \in \mathbb{R}^{n \times p}$

$\sigma_i \geq \sigma_{i+1} \leftarrow \sigma_1 \geq \sigma_2 \geq \sigma_3 \ldots \geq \sigma_p \geq 0$

$V \in \mathbb{R}^{p \times p} \qquad V V^T = V^T V = \mathbb{1}$

Example

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

$$\Sigma \Sigma^T = \underset{3 \times 2}{\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}} \overset{2 \times 3}{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}$$

$$= U \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} U^T$$

$$\Sigma^T \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \left(= \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \; \sigma_p^2 \end{bmatrix} \right.$$

standard case $n > p$
$(n \gg p)$

$$\tilde{y}^{OLS} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

$$= U \Sigma V^T \big(V \Sigma^T \underbrace{U^T U}_{= I} \Sigma V^T \big)^{-1} \times V \Sigma^T U^T y$$

$$V \Sigma^2 V^T = \Sigma^2 \in \mathbb{R}^{p \times p}$$

$$= U \Sigma \underbrace{V^T \frac{1}{\Sigma^2} V}_{I} \Sigma^T U^T y$$

$$= U \Sigma \frac{1}{\Sigma^2} \Sigma^T U^T y$$

$\mathbb{R}^{p \times p}$

$U \in \mathbb{R}^{n \times n}$

$\Sigma \Sigma^T \in \mathbb{R}^{n \times n}$

$$U \frac{\Sigma \Sigma^T}{\Sigma^2} U^T = \boxed{U U^T \neq I}$$

$$= \sum_{j=0}^{p-1} u_j u_j^T$$

$\Rightarrow$

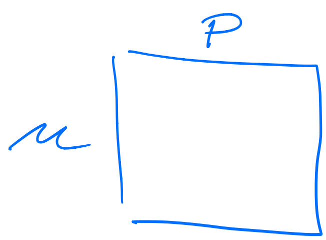
$$\tilde{y}^{OLS} = \sum^{p-1} u_j u_j^T y$$

$j=0$

$u^T y$ are the coordinates of y wrt to the orthonormal U

$$U = [u_0\ u_1\ \dots\ u_{n-1}]$$

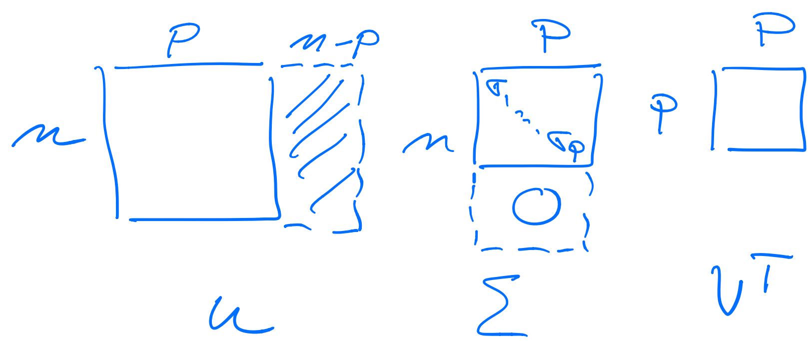
X



$$= \begin{bmatrix} u_{00} & u_{01} & \dots \\ u_{10} & u_{11} & \\ u_{20} & \vdots & \\ \vdots & \vdots & \\ u_{n-1\,0} & u_{n-1\,1} & \end{bmatrix}$$

$U \in \mathbb{R}^{n\times n}$

$UU^T \in \mathbb{R}^{n\times n}$



$$X^T X \propto \mathrm{cov}[X]$$

$$\mathrm{var}(\hat\beta) \propto (X^T X)^{-1}$$

$$X^T X = V\Sigma^T U^T U \Sigma V^T = V\Sigma^2 V^T$$

$(X^T X) V = V \Sigma^2 \Rightarrow$ eigenvectors of $X^T X$ are V with eigen values Σ^2

$$X X^T = U \Sigma V^T V \Sigma^T U^T$$

$$= U(\Sigma \Sigma^T) U^T \Rightarrow$$

$$(X X^T) U = U(\Sigma \Sigma^T)$$

eigenvectors of $X X^T$ are U with eigenvalues $\Sigma \Sigma^T$

$$U = \mathrm{evec}(X X^T)$$

$$V = \mathrm{evec}(X^T X)$$

$$\Sigma^2 = \mathrm{eval}(X X^T) = \mathrm{eval}(X^T X)$$

(non-zero)

Ridge

$$\hat{\beta}^{Ridge} = (\underbrace{X^T X + \lambda \mathbb{I}}_{\in \mathbb{R}^{p \times p}})^{-1} X^T y$$

$$\Rightarrow \tilde{y}^{Ridge} = X \hat{\beta}^{Ridge}$$

$$= U \Sigma V^T \left(V \Sigma^T \underbrace{U^T U}_{I} \Sigma V^T + \lambda I \right)^{-1}$$

$$\times \; V \Sigma^T U^T y$$

$$= \left(\sum_{j=0}^{p-1} u_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} u_j^T \right) y$$

$\lambda \geq 0$ then we have

$$\frac{\sigma_j^2}{\sigma_j^2 + \lambda} \leq 1$$

Ridge shrinks the coordinates of y by a factor $\frac{\sigma_j^2}{\sigma_j^2 + \lambda}$

A greater amount of shrinkage is applied to the coordinates of basis vector with smaller values of σ_j^2.

Example

$$X^T X = X X^T = \mathbb{1} \qquad p = n$$

$$\hat{\beta}^{Ridge} = (\mathbb{I} + \lambda \mathbb{I})^{-1} X^T y$$

$$\Rightarrow \quad \tilde{y}^{Ridge} = \frac{1}{1+\lambda} \tilde{y}^{OLS}$$

$$X^T X = V \Sigma^2 V^T = C[X] \cdot n$$

$$\boxed{(X^T X) V = V \Sigma^2}$$ eigvalues of the covariance matrix and the singular values

$$\sigma_j^2 / n \qquad \sigma_j \geq \sigma_{j+1}$$

The first component v_1 has the largest singular value σ_1^2

With Ridge we are effectively shrinking those components with small covariance/variance (small eigenvalues σ_{ii}^2)

v_p has eigenvalue σ_p^2

Ridge shrinks these values by $\dfrac{\sigma_j^2}{\sigma_j^2+\lambda}$

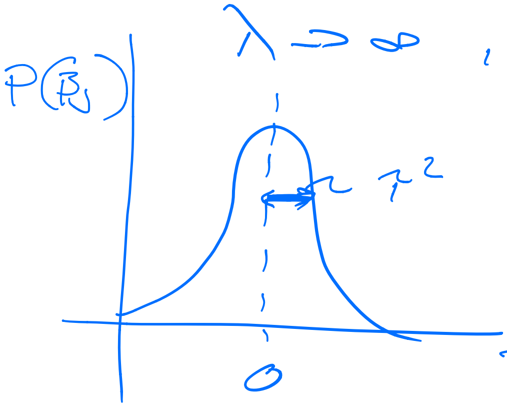
degrees of freedom

$$df(\lambda) = Tr\left[X(X^TX+\lambda I)^{-1}X^T\right]$$

$$= \sum_{j=0}^{p-1} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$$

$\lambda = 0$ then $df(\lambda) = p$

$\lambda \to \infty$, $df(\lambda) \to 0$



$$var(\hat{\beta}^{OLS})$$

$$= \sigma^2 (X^TX)^{-1}$$

$$\sigma^2 \left(X^TX\right)^{-1}_{jj} = \sigma_j^{-2}$$

$$(X^TX)\underline{v} = v\Sigma^2$$

$var(\beta_j) =$

$$var(\hat{\beta}^{Ridge}) = \sigma^2\left[X^TX+\lambda I\right]^{-1} \underline{X^TX}$$

$$\times \left\{ \left[X^T X + \lambda I\right]^{-1} \right\}^T$$

Ridge if $var(\hat{\beta}_j^{OLS})$ is large or small σ_j^2 from $X^T X$, with λ in Ridge we can shrink $var(\hat{\beta}_j^{Ridge})$ to smaller values.