Question 1:

1. $X^T X \beta = X^T y$, $X = S \Sigma_r Q^T$, $X^T X = Q \Sigma_r^2 Q^T$

$$X^T y = Q \Sigma_r S^T y$$

$$\hat{\beta}_{LS} = (X^T X)^{-1} X^T y = Q \Sigma_r^{-2} Q^T Q \Sigma_r S^T y = Q \Sigma_r^{-1} S^T y$$

$$= \sum_{i=1}^{p} \frac{s_i^T y}{\sigma_i} q_i = \sum_{i=1}^{p} \hat{b}_i$$

$$\hat{b}_i = \frac{s_i^T y}{\sigma_i} q_i$$

2. The ridge regression estimate is given by

$$\hat{\beta}_{RR} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

$$\hat{\beta}_{RR} = \frac{\partial}{\partial \beta} (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

$$= -2 X^T (y - X\beta) + 2\lambda \beta = -2 X^T y + 2 X^T X \beta + 2\lambda \beta = 0$$

$$(X^T X + \lambda I_p)\beta = X^T y \rightarrow \hat{\beta}_{RR} = (X^T X + \lambda I_p)^{-1} X^T y$$

$$\hat{\beta}_{RR} = (Q \Sigma_r^2 Q^T + \lambda Q Q^T)^{-1} Q \Sigma_r S^T y$$

$$= [Q(\Sigma_r^2 + \lambda I) Q^T]^{-1} Q \Sigma_r S^T y$$

$$= Q (\Sigma_r^2 + \lambda I)^{-1} \Sigma_r S^T y = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} s_i^T y \, q_i$$

$$= \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \frac{s_i^T y}{\sigma_i} q_i = \sum_{i=1}^{p} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \hat{b}_i$$

3: $f(\tilde{v}_i) = \begin{cases} 1 & \text{if } i \le k < r \\ 0 & \text{if } i > k \end{cases}$

$f(\tilde{v}_i) = \frac{\tilde{v}_i^2}{\tilde{v}_i^2 + \lambda}$

4:

$$E(\hat{\beta}_{RR}) = E[(X^TX + \lambda I_p)^{-1} X^T y] = (X^TX + \lambda I_p)^{-1} X^T E(y)$$

$$= (X^TX + \lambda I_p)^{-1} X^TX\beta = (X^TX + \lambda I_p)^{-1}(X^TX + \lambda I_p - \lambda I_p)\beta$$

$$= \beta - \lambda(X^TX + \lambda I_p)^{-1}\beta$$

$E[\hat{\beta}_{RR}] \neq \beta$ [illegible]

biased.

5:

$$Var[\hat{\beta}_{RR}] = Var[(X^TX + \lambda I_p)^{-1} X^T y]$$

$$= (X^TX + \lambda I_p)^{-1} X^T Var(y) X (X^TX + \lambda I_p)^{-T}$$

$$= \sigma^2 (X^TX + \lambda I_p)^{-1} X^TX (X^TX + \lambda I_p)^{-T} = \sigma^2 W_\lambda (X^TX)^{-1} W_\lambda^T$$

$W_\lambda = (X^TX + \lambda I_p)^{-1} X^TX$. $W_\lambda^{-1} = (X^TX)^{-1}(X^TX + \lambda I_p) = I_p + \lambda(X^TX)^{-1}$

$$Var(\hat{\beta}_{LS}) - Var(\hat{\beta}_{RR}) = \sigma^2[(X^TX)^{-1} - W_\lambda (X^TX)^{-1} W_\lambda^T]$$

$$\sigma^2 W_\lambda \{[I + \lambda(X^TX)^{-1}](X^TX)^{-1}[I + \lambda(X^TX)^{-1}]^T - (X^TX)^{-1}\} W_\lambda^T$$

$$\sigma^2 W_\lambda \{\lambda^2 (X^TX)^{-2} + \lambda^2 (X^TX)^{-3}\} W_\lambda^T$$

$$= \sigma^2 (X^T X + \lambda I_p)^{-1} [2\lambda I_p + \lambda^2 (X^T X)^{-1}] (X^T X + \lambda I_p)^{-T}$$

- The difference is non-negative definite as each component in the matrix product is non-negative definite. Then

$$Var(\hat{\beta}_{LS}) \geq Var(\hat{\beta}_{RR})$$

- The variance of the ML estimator is larger than that of the ridge estimator.

- $\lim_{\lambda \to 0} Var(\hat{\beta}_{RR}) = Var(\hat{\beta}_{LS})$

$\lim_{\lambda \to \infty} Var(\beta_{RR}) = \lim_{\lambda \to \infty} \sigma^2 W_\lambda (X^T X) W_\lambda = 0_p$

The variance of the ridge estimator decreases towards zero as the penalty parameter becomes large.

6: The ridge regression estimator is linear in y

$$\hat{\beta}_{RR} \sim N\left((X^T X + \lambda I_p)^{-1} X^T X \beta, \sigma^2 (X^T X + \lambda I_p)^{-1} X^T X [(X^T X + \lambda I_p)^{-1}]^T\right)$$

7: The MSE of the ridge regression estimator is

$MSE(\hat{\beta}_{RR}) = E[(W_\lambda \hat{\beta}_{LS} - \beta)^T (W_\lambda \hat{\beta}_{LS} - \beta)]$ we drop LS

$= E[\hat{\beta}^T W_\lambda^T W_\lambda \hat{\beta}] - E[\beta^T W_\lambda \hat{\beta}] - E[\hat{\beta}^T W_\lambda^T \beta] + E[\beta^T \beta]$

$= E[\hat{\beta}^T W_\lambda^T W_\lambda \hat{\beta}] - E[\beta^T W_\lambda^T W_\lambda \hat{\beta}] - E[\hat{\beta}^T W_\lambda^T W_\lambda \beta] + E[\beta^T W_\lambda^T W_\lambda \beta]$
$- E[\beta^T W_\lambda^T W_\lambda \beta] + E[\beta^T W_\lambda^T W_\lambda \hat{\beta}] + E[\beta^T W_\lambda^T W_\lambda \beta] - E[\beta^T W_\lambda \hat{\beta}^T]$

$$- E[\hat{B}^T W_\lambda^T B] + E[B^T B]$$

$$= E[(\hat{B}-B)^T W_\lambda^T W_\lambda (\hat{B}-B)] - B^T W_\lambda^T W_\lambda B + B^T W_\lambda^T W_\lambda B + B^T W_\lambda^T W_\lambda B$$
$$- B^T W_\lambda B - B^T W_\lambda^T B + B^T B$$

$$= E[(\hat{B}-B)^T W_\lambda^T W_\lambda (\hat{B}-B)] + B^T [W_\lambda - I_p]^T [W_\lambda - I_p] B$$

$$= \sigma^2 tr[W_\lambda (X^T X)^{-1} W_\lambda^T] + B^T [W_\lambda - I_p]^T [W_\lambda - I_p] B$$

where we used $E[\xi^T A \xi] = tr(A\Sigma_\xi) + \mu_\xi^T A \mu_\xi$ for $\xi \sim N(\mu_\xi, \Sigma_\xi)$

$\lim_{\lambda \to \infty} MSE(\hat{B}_{RR}) = B^T B$ which is the squared bias.

5: In the case of orthonormal design matrix: $X^T X = I_p$

- $\hat{B}_{RR} = (X^T X + \lambda I_p)^{-1} X^T y$

$$= (X^T X + \lambda I_p)^{-1} (X^T X) \hat{B}$$

$$= (I_p + \lambda I_p)^{-1} \hat{B} = (1+\lambda)^{-1} \hat{B}$$

- $E(\hat{B}_{RR}) = E\{(1+\lambda)^{-1} \hat{B}\} = (1+\lambda)^{-1} E(\hat{B}) = (1+\lambda)^{-1} B \neq B$

The estimator and its expectation vanishes as $\lambda \to \infty$.

- $Var(\hat{B}_{RR}) = \sigma^2 W_\lambda (X^T X)^{-1} W_\lambda = \sigma^2 (I_p + \lambda I_p)^{-1} I_p [(I_p + \lambda I_p)^{-1}]^T$

$$= \sigma^2 (1+\lambda)^{-2} I_p.$$

which vanishes as $\lambda \to \infty$.

- $MSE(\hat{B}_{RR}) = p\sigma^2 (1+\lambda)^{-2} + \lambda^2 (1+\lambda)^{-2} B^T B$

which is minimized for $\lambda = p\sigma^2 / B^T B$

9 Increasing norm with decreasing λ

$$\|\hat{\beta}_{rr}\|_2^2 = y^T S \Sigma_r (\Sigma_r^2 + \lambda I)^{-1} \Sigma^T \Sigma (\Sigma_r^2 + \lambda I)^{-1} \Sigma_r S^T y$$

$$= y^T S \Sigma_r (\Sigma_r^2 + \lambda I)^{-2} \Sigma_r S^T y$$

$$= (S^T y)^T \Sigma_r (\Sigma_r^2 + \lambda I)^{-2} \Sigma_r (S^T y)$$

$$= \sum_{i=1}^{r} \frac{\sigma_i^2}{(\sigma_i^2 + \lambda)^2} (s_i^T y)^2$$

As $\lambda \to 0$, $\frac{\sigma_i^2}{(\sigma_i^2+\lambda)^2}$ increases and also $\|\hat{\beta}_{rr}\|_2^2$.

Question 2:

1: The number of degrees of freedom is obtained

$$\sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i)$$

for least square: $\hat{y} = X(X^T X)^{-1} X^T y$

$$\text{cov}(\hat{y}, y) = \text{cov}(X(X^T X)^{-1} X^T y, y) = X(X^T X)^{-1} X^T \text{cov}(y, y)$$

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

- The values $\text{cov}(\hat{y}_i, y_i)$ are the diagonal values of the matrix given by $\text{cov}(\hat{y}, y)$

$$\sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i) = \text{tr}\left[X(X^T X)^{-1} X^T\right] \sigma^2 = \text{tr}\left[(X^T X)^{-1} X^T X\right] \sigma^2 = p\sigma^2$$

- In the orthonormal design matrix case: $X^T X = I_p$

$$\sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i) = p\sigma^2.$$

8: for ridge regression: $\hat{y} = X(X^TX + \lambda I)^{-1}X^Ty$

$\text{cov}(\hat{y}, y) = X(X^TX + \lambda I)^{-1}X^T\sigma^2$

$$\sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i) = \text{tr}\{X(X^TX + \lambda I)^{-1}X^T\}\sigma^2 = \text{tr}\{(X^TX + \lambda I)^{-1}X^TX\}\sigma^2$$

For orthonormal design matrix X

$$\sum_{i=1}^{n} \text{cov}(\hat{y}_i, y_i) = \frac{p\sigma^2}{1+\lambda}$$