Week 36 exercise 1

1a) Show that the optimal parameters

$$\hat{\beta}_{ridge} = (\underline{X}^T \underline{X} + \lambda \underline{I})^{-1} \underline{X}^T \underline{y}$$

with $I \in \mathbb{R}^{p \times p}$ being the identity matrix.

We have to solve the Ridge regression minimization problem

$$\min_{\underline{\beta} \in \mathbb{R}^p} \frac{1}{n} \| \underline{y} - \underline{X}\underline{\beta} \|_2^2 + \lambda \|\underline{\beta}\|_2^2 \quad (1)$$

The minimization problem is a question of finding the optimal vector $\underline{\beta}$!

Since this is a sum we split it into two parts. I start with the latter and utilize the identity $\frac{\partial}{\partial \underline{x}} \|\underline{x}\|_2 = \underline{x}$

$$\frac{\partial}{\partial \underline{\beta}} \left[\lambda \|\underline{\beta}\|_2^2 \right] = 2\lambda \underline{\beta} \quad (2)$$

Secondly we have the first term

$$\frac{1}{n}\frac{\partial}{\partial \beta}\left[\|y - X\beta\|_2^2\right]$$

$$= \frac{1}{n}\frac{\partial}{\partial u}\left[\|u\|_2^2\right]\frac{\partial}{\partial \beta}\left[y - X\beta\right]$$

$$= \frac{1}{n}\left(2(y - X\beta)\right)(-X)$$

$$= -\frac{2}{n}(y - X\beta)X$$

$$= \frac{2}{n}X^T(X\beta - y) \qquad (3)$$

To do the actual optimization we now find the extremes by (3)+(2)=0

$$\frac{2}{n}X^T(X\beta - y) + 2\lambda\beta = 0$$

$$\frac{2}{n}X^TX\beta - \frac{2}{n}X^Ty + 2\lambda\beta = 0$$

By multiplying by $n/2$ we get

$$X^TX\beta + n\lambda\beta = X^Ty$$

Extracting $\underline{\beta}$ from LH we have to multiply $\underline{I}$ into the equation for it to be valid. Further since n is a const we can absorb it into λ, $\lambda = \lambda' = n\lambda$

$$\left(\underline{X}^T\underline{X} + \lambda \underline{I}\right)\underline{\beta} = \underline{X}^T\underline{Y}$$

Assuming invertability

$$\Rightarrow \underline{\hat{\beta}}_{ridge} = \left(\underline{X}^T\underline{X} + \lambda\underline{I}\right)^{-1}\underline{X}^T\underline{Y}$$ ∎

1b) $\underline{\tilde{y}}_{OLS} = \underline{X}\,\underline{\beta}$ (1)

For $\underline{\beta}$ we have the sol. from 1a)

$$\underline{\beta} = (\underline{X}^T\underline{X})^{-1}\,\underline{X}^T\underline{y} \quad (2)$$

and $\underline{X}$ can be expressed by the SVD as

$$\underline{X} = \underline{U}\,\underline{\Sigma}\,\underline{V}^T \quad (3)$$

where:

$\underline{U} \in \mathbb{R}^{n\times n}$, $\underline{V} \in \mathbb{R}^{p\times p}$,

$\underline{\Sigma} \in \mathbb{R}^{n\times p}$ matrix on the form

$$\underline{\Sigma} = \begin{bmatrix} \underline{D} & 0 \\ 0 & 0 \end{bmatrix}, \quad \underline{D} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}$$

where σ's are the first $r = \text{Rank}(\underline{X})$ singular values > 0.

Combining (1) and (2):

$$\hat{\underline{y}}_{OLS} = \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}$$

To replace the $\underline{X}$ matrices we have

↳ $\underline{X}$: (3)

↳ $(\underline{X}^T\underline{X})^{-1}$: $(\underline{U}^T\underline{\Sigma}^T\underline{V}\underline{U}\underline{\Sigma}\underline{V}^T)^{-1}$

$$= (\underline{V}\underline{\Sigma}^T\underline{\Sigma}\underline{V}^T)^{-1}$$

utilysing the orthogonality of $\underline{V}$ we get

$$(\underline{X}^T\underline{X})^{-1} = \underline{V}^T(\underline{\Sigma}^T\underline{\Sigma})^{-1}\underline{V}$$

↳ $\underline{X}^T = \underline{U}^T\underline{\Sigma}^T\underline{V}$

This yields us

$$\tilde{y}_{ols} = \underline{U}\underline{\Sigma}\underline{V}^T\underline{V}^T(\underline{\Sigma}^T\underline{\Sigma})^{-1}\underline{V}\underline{U}^T\underline{\Sigma}^T\underline{V}$$

$$= \underline{U}\,\underline{\Sigma}\,\underline{V}^T\underline{V}\,(\underline{\Sigma}^T\underline{\Sigma})^{-1}\,\underline{V}^T\underline{V}\,\underline{\Sigma}^T\underline{U}^T$$

$$= \underline{U}\underline{\Sigma}\;\underbrace{I_p}_{p\times p}\;\underbrace{\frac{1}{\underline{\Sigma}^T\underline{\Sigma}}}_{\underline{\Sigma}^T\underline{\Sigma}}\;\underbrace{I_p}_{p\times p}\;\underline{\Sigma}^T\underline{U}^T \qquad (4)$$

$n\times n \cdot n\times p = n\times p$ $\quad$ $p\times n \cdot n\times p = p\times p$ $\quad$ $p\times n \cdot n\times n = p\times n$

From the dimentions we see that the products add up. Further we can observe what each product will behave as:

$\underline{U}\underline{\Sigma}$: The rows/colums $=0$ in $\underline{\Sigma}$ will go to 0 for the product thus we will have a $n\times p$ matrix containing the prod. of the s.v. and the eigenvectors of the design matrix e.g $\sigma_i \underline{u}_1$ where $\underline{U} = [\underline{u}_1, \underline{u}_2 \cdots$

$(\underline{\Sigma}^T\underline{\Sigma})^{-1}$: will yield a diagonal matrix with $\frac{1}{\sigma_i^2}$ in the entries.

$\Sigma^T U^T$: again the 0 of Σ^T makes the matrixproduct where the singular values in one col and the transposed U: vectors filling the remaining cols

However we can easally see that we can cancel matrices in (4):

$$\tilde{y}_{OLS} = U \Sigma \frac{1}{\Sigma^T \Sigma} \Sigma^T U^T y \quad (4)$$

$$= U U^T y$$

For $i \neq j$ the inner product $= 0$ therefor

$$\tilde{y}_{OLS} = \left(u_0 u_0^T + \ldots + u_p u_p^T \right) y$$

$$= \left[\sum_{j=1}^{D} u_j u_j^T \right] y$$

$$\tilde{y}_{ridge} = \underline{X}\,\beta_{ridge}$$

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

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

$$\left(\underline{U}\,\underline{\Sigma}\,\underline{V}^T\right)^T y$$

Using $\underline{U}^T\underline{U} = I$ and $\underline{V}^T\underline{V} = I$

$$\tilde{y}_{ridg} = \underline{U}\,\underline{\Sigma}\left(\underline{\Sigma}^2 + \lambda \underline{I}\right)^{-1}\underline{\Sigma}\,\underline{U}^T y$$

Since $\Sigma\Sigma^{-1} = I$ we get

$$= \underline{U}\left(\underline{\Sigma}^2 + \lambda \underline{I}\right)^{-1}\underline{U}^T y$$

Similary as for OLS $j \neq i$ yields an inner prod $= 0$ and we therefor only get one sumation

Further $(\underline{\Sigma}^2 + \lambda \underline{I})^{-1}$ will yield

$$\frac{\sigma_j^2}{\sigma_j^2 + \lambda}$$ giving us

$$\tilde{\underline{Y}}_{ridge} = \sum_{j=1}^{p} \underline{u}_j \frac{\sigma_j^2}{\sigma_j^2 + \lambda} \underline{u}_j^T \underline{Y}$$

The interpetation of this is that the $\underline{Y}$ is the weigthed sum of the $\underline{u}_j$'s with the weight being dictated by the regularization parametre.

A $\lambda = 0$ will yield $\hat{Y}_{OLS}$ since $\frac{\sigma_j^2}{\sigma_j^2 + 0} = 1$. However a large λ will decrease the value of the sum since

$$\lim_{\lambda \to \infty} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} = 0$$

This will stchalize the model.