Exercise 5

Group: Mara-Eliana Popescu, Andrei Prioteasa

1 Bias and variance of ridge regression

To show: $\mathbb{E}[\hat{\beta}_\tau] = S_\tau^{-1} S \beta^*$

$\mathrm{Cov}[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1} \sigma^2$

with $S := X^\top X$ and $S_\tau := X^\top X + \tau \mathbb{1}_D$

Proof: solution to ridge regression proven in the lecture

- $\mathbb{E}[\hat{\beta}_\tau] = \mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top y]$

$$= \mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top (X\beta^* + \varepsilon)]$$

$$= \mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^* + (X^\top X + \tau \mathbb{1}_D)^{-1} X^\top \varepsilon]$$

$$= \mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^*] + \mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top \varepsilon]$$

$$= (X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^* + \mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top \varepsilon]$$

only ε is a random variable, thus

$$\mathbb{E}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^*] = (X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^*$$

$$= (X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^* + (X^\top X + \tau \mathbb{1}_D)^{-1} X^\top \underbrace{\mathbb{E}[\varepsilon]}_{=0}$$

$$= (X^\top X + \tau \mathbb{1}_D)^{-1} X^\top X \beta^*$$

$$= S_\tau^{-1} S \beta^*$$

- $\mathrm{Cov}[\hat{\beta}_\tau] = \mathrm{Cov}[(X^\top X + \tau \mathbb{1}_D)^{-1} X^\top (X\beta^* + \varepsilon)]$

$= \mathrm{Cov}[\underbrace{S_\tau^{-1} S \beta^*}_{\text{constant}} + S_\tau^{-1} X^\top \varepsilon]$

$= \mathrm{Cov}[S_\tau^{-1} X^\top \varepsilon]$ $\quad$ ($\mathrm{Cov}[AY] = A\,\mathrm{Cov}[Y]$ for matrix A and a random variable Y)

$= S_\tau^{-1} X^\top \underbrace{\mathrm{Cov}[\varepsilon]}_{=\sigma^2 \mathbb{1}_N} (S_\tau^{-1} X^\top)^\top$

$= S_\tau^{-1} X^\top \sigma^2 \mathbb{1}_N X (S_\tau^{-1})^\top$

$= S_\tau^{-1} \underbrace{X^\top X}_{=S} (S_\tau^{-1})^\top \sigma^2$

$S_\tau^\top = (X^\top X + \tau \mathbb{1}_D)^\top = (X^\top X)^\top + \tau \mathbb{1}_D^\top$

$= X^\top X + \tau \mathbb{1}_D = S_\tau$

S_τ symmetric and nonsingular $\Rightarrow$ S_τ^{-1} is also symmetric, i.e. $(S_\tau^{-1})^\top = S_\tau^{-1}$

Therefore: $\mathrm{Cov}[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1} \sigma^2$ $\blacksquare$

2. LDA - Derivation from the LSE

$$\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i^* - x_i \cdot \beta)^2$$

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - x_i \beta)^2 = \sum_{i=1}^{N} \frac{\partial}{\partial \beta} (y_i^* - x_i \beta)^2 =$$

$$\sum_{i=1}^{N} 2 (y_i^* - x_i \beta)(-x_i^T) \overset{!}{=} 0$$

$$\Leftrightarrow \sum_{i=1}^{N} (y_i^* - x_i \beta) x_i^T = 0$$

$$\Leftrightarrow \sum_{i=1}^{N} y_i^* x_i^T - \sum_{i=1}^{N} x_i \beta x_i^T = 0$$

$$\Leftrightarrow \frac{1}{N} \sum_{i=1}^{N} y_i^* x_i^T - \frac{1}{N} \sum_{i=1}^{N} x_i \beta x_i^T = 0 \qquad (*)$$

$$\boxed{\frac{1}{N} \sum_{i=1}^{N} y_i^* x_i^T} = \frac{1}{N} \sum_{i: y_i^* = -1} -x_i^T + \frac{1}{N} \sum_{i: y_i^* = 1} x_i^T$$

$$= \frac{1}{2} (-\mu_{-1}^T + \mu_1^T)$$

$$= \boxed{\frac{1}{2} (\mu_1 - \mu_{-1})^T}$$

$$\boxed{\frac{1}{N} \sum_{i=1}^{N} x_i \beta x_i^T} \underset{\substack{\uparrow \\ x_i \beta \text{ is a scalar}}}{=} \frac{1}{N} \sum_{i=1}^{N} x_i^T x_i \beta$$

$$= \left[\frac{1}{N} \sum_{i: y_i^* = -1} x_i^T x_i + \frac{1}{N} \sum_{i: y_i^* = 1} x_i^T x_i \right] \beta$$

$$= \frac{1}{N} \Big\{ \sum_{i: y_i^* = -1} \left[(x_i - \mu_{-1})^T (x_i - \mu_{-1}) + x_i^T \mu_{-1} + \mu_{-1}^T \cdot x_i - \mu_{-1}^T \mu_{-1} \right]$$

$$+ \sum_{i: y_i^* = 1} \left[(x_i - \mu_1)^T (x_i - \mu_1) + x_i^T \mu_1 + \mu_1^T x_i - \mu_1^T \mu_1 \right] \Big] \beta$$

$$= \frac{1}{N}\Big[\sum_{i:y_i^*=-1}(x_i-\mu_{-1})^\top(x_i-\mu_{-1}) + \sum_{i:y_i^*=1}(x_i-\mu_1)^\top(x_i-\mu_1)\Big]\beta +$$

$$\Big[\frac{1}{N}\sum_{i:y_i^*=-1} x_i^\top \mu_{-1} + \frac{1}{N}\sum_{i:y_i^*=-1}\mu_{-1}^\top x_i - \frac{1}{N}\cdot\frac{N}{2}\mu_{-1}^\top\mu_{-1}\Big]\beta +$$

$$\Big[\frac{1}{N}\sum_{i:y_i^*=1} x_i^\top \mu_1 + \frac{1}{N}\sum_{i:y_i^*=1}\mu_1^\top x_i - \frac{1}{N}\cdot\frac{N}{2}\mu_1^\top\mu_1\Big]\beta$$

$$= \Sigma\beta + \Big[\frac{1}{2}\mu_{-1}^\top\mu_{-1} + \cancel{\frac{1}{2}\mu_{-1}^\top}\mu_{-1} - \frac{1}{2}\cancel{\mu_{-1}^\top}\mu_{-1}\Big]\beta +$$

$$\Big[\frac{1}{2}\mu_1^\top\mu_1 + \cancel{\frac{1}{2}\mu_1^\top}\mu_1 - \frac{1}{2}\mu_1^\top\cancel{\mu_1}\Big]\beta$$

$$= \Sigma\beta + \left(\frac{1}{2}\mu_{-1}^\top\mu_{-1} + \frac{1}{2}\mu_1^\top\mu_1\right)\beta$$

$$= \Sigma\beta + \frac{1}{4}\left(2\mu_{-1}^\top\mu_{-1} + 2\mu_1^\top\mu_1\right)\beta$$

$$= \Sigma\beta + \frac{1}{4}\left(\mu_{-1}^\top\mu_{-1} + \mu_{-1}^\top\mu_{-1} + \mu_1^\top\mu_1 + \mu_1^\top\mu_1\right)\beta$$

$$\overset{\mu_1=-\mu_{-1}}{=} \Sigma\beta + \frac{1}{4}\left(\mu_{-1}^\top\mu_{-1} - \mu_{-1}^\top\mu_1 - \mu_1^\top\mu_{-1} + \mu_1^\top\mu_1\right)\beta$$

$$= \boxed{\Sigma\beta + \frac{1}{4}(\mu_1-\mu_{-1})^\top(\mu_1-\mu_{-1})\beta}$$

Hence, $(*)$ can be rewritten as:

$$\Sigma\beta + \frac{1}{4}(\mu_1-\mu_{-1})^\top(\mu_1-\mu_{-1})\beta = \frac{1}{2}(\mu_1-\mu_{-1})^\top$$

$\blacksquare$