Theoretical Questions

4.2 – a) The LDA decision rule is defined as

$$G(x) = \text{argmax}_k \, \delta_k(x)$$

$$\delta(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\Sigma}^{-1} \hat{\mu}_k + \log \pi_k \qquad (4.10 \text{ in ESL})$$

Therefore $G(x) = 2$ iff. $\text{argmax}_k (\delta_1, \delta_2) = 2$

$$\Leftrightarrow \delta_2 > \delta_1$$

$$\Leftrightarrow x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) - \frac{1}{2} (\hat{\mu}_2 - \hat{\mu}_1)^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) + \log \left(\frac{N_2/N}{N_1/N} \right) > 0$$

$$\Leftrightarrow x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2} (\mu_2 - \mu_1)^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) + \log \left(\frac{N_2}{N_1} \right)$$

b) $\hat{\beta} = (X^T X)^{-1} X^T Y$

$$X^T X = \begin{pmatrix} 1_{N_1}^T & 1_{N_2}^T \\ X_1^T & X_2^T \end{pmatrix} \begin{pmatrix} 1_{N_1} & X_1 \\ 1_{N_2} & X_2 \end{pmatrix} = \begin{pmatrix} N_1 + N_2 & N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2 \\ N_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2^T & X_1^T X_1 + X_2^T X_2 \end{pmatrix}$$

$$\sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T = N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T$$

c) $(\hat{\mu}_2 - \hat{\mu}_1)^T$ is a $1 \times p+1$ vector and β is a $(p+1) \times 1$ vector, therefore $(\hat{\mu}_2 - \hat{\mu}_1)^T \beta$ is a scalar. Therefore $\hat{\Sigma}_B \beta$ is proportional to $(\hat{\mu}_2 - \hat{\mu}_1)$.

d)

4.3 - LDA using $\hat{Y}$ is identical to LDA in the original space

iff. $G_k(x) = G_k(y) \quad \forall k \in K$

Let k be some class in the set K, and let

$$\delta_k(x) = x^T \Sigma^{-1} \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \Sigma^{-1} \hat{\mu}_k$$

be the discriminant function for LDA with X as a predictor.

Want to prove that

$$\delta_k(x) \text{ is maximal} \iff \delta_k(y) \text{ is maximal} \quad (*)$$

$\boxed{\Rightarrow}$

Let $\ell \in K$ s.t $\ell \neq k$. We have then

$$\delta_k(x) > \delta_\ell(x)$$

$$\iff x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k > x^T \Sigma^{-1} \mu_\ell - \frac{1}{2} \mu_\ell^T \Sigma^{-1} \mu_\ell + \log \pi_\ell$$

$$\iff \log\left(\frac{\pi_k}{\pi_\ell}\right) - \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) + x^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) > 0$$

Since $\hat{\mu}_i = 1_{N_i}^T X_i / N_i$, the sample means $\hat{\mu}_i$ can be transformed by B, just like any other point in $\mathbb{R}^p$.

$$\iff -\frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) + x^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) > 0$$

$$\iff x^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell) > \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \Sigma^{-1}(\hat{\mu}_k - \hat{\mu}_\ell)$$

~~since~~ multiplying both sides by $(\hat{\mu}_k - \hat{\mu}_\ell)^{-1} \Sigma$

$$\iff x^T > \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T$$

$$\iff x^T \hat{B} > \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_\ell)^T \hat{B}$$

$$\iff \hat{y} > \frac{1}{2}(\hat{\mu}_k' + \hat{\mu}_\ell')^T$$

where μ_i' is the average of the transformed $\hat{y}_i$: ~~vector~~ sample:

$$\mu_i' = \left[1_N^T X_i / N_i\right] \hat{B} = \frac{1}{N_i} 1_N^T Y$$

The other direction of (*) can be obtained by applying a similar procedure as for $\boxed{\Rightarrow}$.