Homework of Chapter 4

Chen Cheng, 1130339005

January 11, 2015

Ex. 4.2

Solution.
(a) Using LDA rule, we know:

$$Pr(G=1|\mathbf{X}=x) = \frac{f_1(x)\pi_1}{f_1(x)\pi_1 + f_2(x)\pi_2}$$

$$Pr(G=2|\mathbf{X}=x) = \frac{f_2(x)\pi_2}{f_1(x)\pi_1 + f_2(x)\pi_2}$$

From the problem description, we can get $\pi_1 = \frac{N_1}{N}$ and $\pi_1 = \frac{N_2}{N}$
Then,

$$\begin{aligned}
&log\frac{Pr(G=2|\mathbf{X}=x)}{Pr(G=1|\mathbf{X}=x)} \\
=&log\frac{\pi_2}{\pi_1} - \frac{1}{2}(\hat{\mu}_2+\hat{\mu}_1)^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2-\hat{\mu}_1) + x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2-\hat{\mu}_1) \\
=&log\frac{N_2}{N_1} - \frac{1}{2}(\hat{\mu}_2^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_2^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1 + \hat{\mu}_1^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_1^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1) + x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2-\hat{\mu}_1) \\
=&log\frac{N_2}{N} - log\frac{N_1}{N} - \frac{1}{2}(\hat{\mu}_2^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_1^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1) + x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2-\hat{\mu}_1)
\end{aligned}$$

Therefore,

$$\begin{aligned}
&log\frac{Pr(G=2|\mathbf{X}=x)}{Pr(G=1|\mathbf{X}=x)} > 0 \\
\Leftrightarrow & x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2-\hat{\mu}_1) > log\frac{N_1}{N} - log\frac{N_2}{N} + \frac{1}{2}(\hat{\mu}_2^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_1^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1)
\end{aligned}$$

(b) We reorder the x_i so that $x_i(1 \leq i \leq N_1)$ are in class 1 and $x_{N_1+i}(1 \leq i \leq N_2)$ are in class 2.
So,

$$RSS(\beta, \beta_0) = \sum_{i=1}^{N_1}(-\frac{N}{N_1} - \beta_0 - \beta^T x_i)^2 + \sum_{i=1}^{N_2}(\frac{N}{N_2} - \beta_0 - \beta^T x_{N_1+i})^2$$

Since we consider minimization of the least squares criterion, we can get

$$\frac{\partial}{\partial \beta_0} RSS(\beta, \beta_0) = -2\sum_{i=1}^{N_1}(-\frac{N}{N_1} - \beta_0 - \beta^T x_i) - 2\sum_{i=1}^{N_2}(\frac{N}{N_2} - \beta_0 - \beta^T x_{N_1+i}) = 0$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{N_1}(\frac{N}{N_1} + \beta_0 + \beta^T x_i) &= \sum_{i=1}^{N_2}(\frac{N}{N_2} - \beta_0 - \beta^T x_{N_1+i}) \\
N + N_1\beta_0 + \beta^T \sum_{i=1}^{N_1} x_i &= N - N_2\beta_0 - \beta^T \sum_{i=1}^{N_2} x_{N_1+i} \\
N\beta_0 &= -\beta^T \sum_{i=1}^{N_2} x_i - \beta^T \sum_{i=1}^{N_1} x_{N_1+i} \\
&= -N_2\beta^T \hat{\mu}_2 - N_1\beta^T \hat{\mu}_1 \\
\beta_0 &= -\frac{1}{N}\beta^T(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)
\end{aligned}$$

We can also get

$$\frac{\partial}{\partial \beta} RSS(\beta, \beta_0) = -2\sum_{i=1}^{N_1} x_i(-\frac{N}{N_1} - \beta_0 - \beta^T x_i) - 2\sum_{i=1}^{N_2} x_{N_1+i}(\frac{N}{N_2} - \beta_0 - \beta^T x_{N_1+i}) = 0$$

Therefore,

$$\sum_{i=1}^{N_1} x_i(\frac{N}{N_1} + \beta_0 + \beta^T x_i) = \sum_{i=1}^{N_2} x_{N_1+i}(\frac{N}{N_2} - \beta_0 - \beta^T x_{N_1+i})$$

$$N\hat{\mu}_1 + N_1\hat{\mu}_1\beta_0 + \sum_{i=1}^{N_1} x_i\beta^T x_i = N\hat{\mu}_2 - N_2\hat{\mu}_2\beta_0 - \sum_{i=1}^{N_2} x_{N_1+i}\beta^T x_{N_1+i}$$

$$N\hat{\mu}_1 + N_1\hat{\mu}_1\beta_0 + \sum_{i=1}^{N_1} x_i x_i^T \beta = N\hat{\mu}_2 - N_2\hat{\mu}_2\beta_0 - \sum_{i=1}^{N_2} x_{N_1+i} x_{N_1+i}^T \beta$$

$$N\hat{\mu}_1 + N_1\hat{\mu}_1\beta_0 + N_1 E_1(xx^T)\beta = N\hat{\mu}_2 - N_2\hat{\mu}_2\beta_0 - N_2 E_2(xx^T)\beta$$

$$N\hat{\mu}_1 + N_1\hat{\mu}_1\beta_0 + ((N_1-1)\hat{\mathbf{\Sigma}} + N_1\hat{\mu}_1\hat{\mu}_1^T)\beta = N\hat{\mu}_2 - N_2\hat{\mu}_2\beta_0 - ((N_2-1)\hat{\mathbf{\Sigma}} + N_2\hat{\mu}_2\hat{\mu}_2^T)\beta$$

$$N(\hat{\mu}_2 - \hat{\mu}_1) = (N_1\hat{\mu}_1 + N_2\hat{\mu}_2)\beta_0 + ((N-2)\hat{\mathbf{\Sigma}} + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T)\beta$$

We combine this equation with

$$\beta_0 = -\frac{1}{N}\beta^T(N_2\hat{\mu}_2 + N_1\hat{\mu}_1) = -\frac{1}{N}(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)^T\beta$$

We can get

$$\left[-\frac{1}{N}(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)^T + ((N-2)\hat{\mathbf{\Sigma}} + N_1\hat{\mu}_1\hat{\mu}_1^T + N_2\hat{\mu}_2\hat{\mu}_2^T)\right]\beta = N(\hat{\mu}_2 - \hat{\mu}_1)$$

$$\left[\frac{N_1N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T + (N-2)\hat{\mathbf{\Sigma}}\right]\beta = N(\hat{\mu}_2 - \hat{\mu}_1)$$

So, we get

$$\left[\frac{N_1N_2}{N}\hat{\mathbf{\Sigma}}_{\mathbf{B}} + (N-2)\hat{\mathbf{\Sigma}}\right]\beta = N(\hat{\mu}_2 - \hat{\mu}_1)$$

(c) Since $\hat{\mathbf{\Sigma}}_{\mathbf{B}} = (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$, and product $(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$ is a scalar. Then, $\hat{\mathbf{\Sigma}}_{\mathbf{B}}$ is in the direction $(\hat{\mu}_2 - \hat{\mu}_1)$.
Let $c = \frac{N_1N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)^T\beta$, then we have

$$c(\hat{\mu}_2 - \hat{\mu}_1) + (N-2)\hat{\mathbf{\Sigma}}\beta = N(\hat{\mu}_2 - \hat{\mu}_1)$$

So,

$$\beta = \frac{N-c}{N-2}\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \propto \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

(d) We suppose that the target code of class 1 and class 2 are y_1 and y_2. Similar to (b), we firstly let the partial derivative of β_0 and β to be 0.
Consider the partial derivative of β_0 be 0:

$$\frac{\partial}{\partial\beta_0}RSS(\beta,\beta_0) = -2\sum_{i=1}^{N_1}(y_1 - \beta_0 - \beta^T x_i) - 2\sum_{i=1}^{N_2}(y_2 - \beta_0 - \beta^T x_{N_1+i}) = 0$$

Therefore,

$$\beta_0 = \frac{1}{N}[N_1 y_1 + N_2 y_2 - \beta^T (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)]$$

Consider the partial derivative of β to be 0:

$$\frac{\partial}{\partial \beta} RSS(\beta, \beta_0) = -2 \sum_{i=1}^{N_1} x_i (y_1 - \beta_0 - \beta^T x_i) - 2 \sum_{i=1}^{N_2} x_{N_1+i} (y_2 - \beta_0 - \beta^T x_{N_1+i}) = 0$$

So,

$$\sum_{i=1}^{N_1} x_i (-y_1 + \beta_0 + \beta^T x_i) = \sum_{i=1}^{N_2} x_{N_1+i} (y_2 - \beta_0 - \beta^T x_{N_1+i})$$

$$-N_1 y_1 \hat{\mu}_1 + N_1 \hat{\mu}_1 \beta_0 + \sum_{i=1}^{N_1} x_i \beta^T x_i = N_2 y_2 \hat{\mu}_2 - N_2 \hat{\mu}_2 \beta_0 - \sum_{i=1}^{N_2} x_{N_1+i} \beta^T x_{N_2+i}$$

$$-N_1 y_1 \hat{\mu}_1 + N_1 \hat{\mu}_1 \beta_0 + \sum_{i=1}^{N_1} x_i x_i^T \beta = N_2 y_2 \hat{\mu}_2 - N_2 \hat{\mu}_2 \beta_0 - \sum_{i=1}^{N_2} x_{N_1+i} x_{N_2+i}^T \beta$$

$$-N_1 y_1 \hat{\mu}_1 + N_1 \hat{\mu}_1 \beta_0 + N_1 E_1 (x x^T) \beta = N_2 y_2 \hat{\mu}_2 - N_2 \hat{\mu}_2 \beta_0 - N_2 E_2 (x x^T) \beta$$

$$-N_1 y_1 \hat{\mu}_1 + N_1 \hat{\mu}_1 \beta_0 + ((N_1 - 1) \hat{\mathbf{\Sigma}} + N_1 \hat{\mu}_1 \hat{\mu}_1^T) \beta = N_2 y_2 \hat{\mu}_2 - N_2 \hat{\mu}_2 \beta_0 - ((N_2 - 1) \hat{\mathbf{\Sigma}} + N_2 \hat{\mu}_2 \hat{\mu}_2^T) \beta$$

$$N_1 y_1 \hat{\mu}_1 + N_2 y_2 \hat{\mu}_2 = (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \beta_0 + ((N-2) \hat{\mathbf{\Sigma}} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T) \beta$$

We combine this equation with

$$\beta_0 = \frac{1}{N}[N_1 y_1 + N_2 y_2 - \beta^T (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)] = \frac{1}{N}[N_1 y_1 + N_2 y_2 - (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2)^T \beta]$$

$$\left[-\frac{1}{N} (N_2 \hat{\mu}_2 + N_1 \hat{\mu}_1)(N_2 \hat{\mu}_2 + N_1 \hat{\mu}_1)^T + ((N-2) \hat{\mathbf{\Sigma}} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T) \right] \beta$$

$$= \frac{N_1 N_2}{N} (y_2 - y_1)(\hat{\mu}_2 - \hat{\mu}_1)$$

$$\left[\frac{N_1 N_2}{N} (\hat{\mu}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T + (N-2) \hat{\mathbf{\Sigma}} \right] \beta = \frac{N_1 N_2}{N} (y_2 - y_1)(\hat{\mu}_2 - \hat{\mu}_1)$$

Similar to (c), we let $N' = \frac{N_1 N_2}{N}(y_2 - y_1)$ and $c = \frac{N_1 N_2}{N}(\hat{\mu}_2 - \hat{\mu}_1)^T \beta$.
we have

$$\beta = \frac{N' - c}{N - 2} \hat{\mathbf{\Sigma}}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) \propto \hat{\mathbf{\Sigma}}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

(e) According to the result of (b) and (c), we know $\hat{\beta}_0 = -\frac{1}{N}(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)^T\beta$ and $\hat{\beta} = \frac{N-c}{N-2}\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$ Then, we have

$$
\begin{aligned}
f =& \hat{\beta}_0 + \hat{\beta}^T x \\
=& \hat{\beta}_0 + x^T\hat{\beta} \\
=& -\frac{N-c}{N(N-2)}(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) + \frac{N-c}{N-2}x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1)
\end{aligned}
$$

Therefore,(why N-c¿0?)

$$
\begin{aligned}
& f > 0 \\
\Longleftrightarrow & x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{N}(N_2\hat{\mu}_2 + N_1\hat{\mu}_1)^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) \\
\Longleftrightarrow & x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{N_2}{N}\hat{\mu}_2^T\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{N_2}{N}\hat{\mu}_2^T\hat{\Sigma}^{-1}\hat{\mu}_1 + \frac{N_1}{N}\hat{\mu}_1^T\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{N_1}{N}\hat{\mu}_1^T\hat{\Sigma}^{-1}\hat{\mu}_1
\end{aligned}
$$

Compare this to the LDA rule

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

We can find that the results of two rules are the same only when $N_1 = N_2$:

$$
x^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_2 - \hat{\mu}_1^T\hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_1)
$$

□

Ex. 4.3

Solution. According to the relationship between $\hat{\mathbf{Y}}$ and $\mathbf{X}$ described in the problem, we can get

$$\begin{aligned}
\pi'_k = \pi_k =& \frac{N_k}{N} \\
\hat{\mu}'_k =& \frac{1}{N_k}\sum_{c_i=k}\hat{y}_i = \frac{1}{N_k}\sum_{c_i=k}\hat{\mathbf{B}}^T x_i = \hat{\mathbf{B}}^T\hat{\mu}_k \\
\hat{\mathbf{\Sigma}}' =& \frac{1}{N-K}\sum_{k=1}^{K}\sum_{c_i=k}(\hat{y}_i - \hat{\mu}'_k)(\hat{y}_i - \hat{\mu}'_k)^T \\
=& \frac{1}{N-K}\sum_{k=1}^{K}\sum_{c_i=k}(\hat{\mathbf{B}}^T x_i - \hat{\mathbf{B}}^T\hat{\mu}_k)(\hat{\mathbf{B}}^T x_i - \hat{\mathbf{B}}^T\hat{\mu}_k)^T \\
=& \hat{\mathbf{B}}^T\left[\frac{1}{N-K}\sum_{k=1}^{K}\sum_{c_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T\right]\hat{\mathbf{B}} \\
=& \hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}}
\end{aligned}$$

where c_i is the class of the i-th example.
Consider the LDA using $\hat{\mathbf{Y}}$:

$$\begin{aligned}
& log\frac{Pr(G=k|\hat{\mathbf{X}}=\hat{y})}{Pr(G=l|\hat{\mathbf{Y}}=\hat{y})} \\
=& log\frac{\pi'_k}{\pi'_l} - \frac{1}{2}(\hat{\mu}'_k + \hat{\mu}'_l)^T\hat{\mathbf{\Sigma}}'^{-1}(\hat{\mu}'_k - \hat{\mu}'_l) + \hat{y}^T\hat{\mathbf{\Sigma}}'^{-1}(\hat{\mu}'_k - \hat{\mu}'_l) \\
=& log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\hat{\mathbf{B}}^T\hat{\mu}_k + \hat{\mathbf{B}}^T\hat{\mu}_l)^T(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}(\hat{\mathbf{B}}^T\hat{\mu}_k - \hat{\mathbf{B}}^T\hat{\mu}_l) + (\hat{\mathbf{B}}^T x)^T(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}(\hat{\mathbf{B}}^T\hat{\mu}_k - \hat{\mathbf{B}}^T\hat{\mu}_l) \\
=& log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\hat{\mu}_k + \hat{\mu}_l)^T\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T(\hat{\mu}_k - \hat{\mu}_l) + x^T\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T(\hat{\mu}_k - \hat{\mu}_l)
\end{aligned}$$

In order to show that LDA using $\hat{\mathbf{Y}}$ is identical to LDA in the original space, we only need to prove $\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T(\hat{\mu}_k - \hat{\mu}_l) = \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_k - \hat{\mu}_l)$.
Since $\mathbf{Y}$ is an indicator response matrix, let y_k be the kth-column of $\mathbf{Y}$, we have $N_k\hat{\mu}_k = \sum_{c_i=k} x_i = \mathbf{X}^T y_k$.

Then, we can get

$$\begin{aligned}
\hat{\mathbf{\Sigma}} =& \frac{1}{N-K} \sum_{k=1}^{K} \sum_{c_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T \\
=& \frac{1}{N-K} \left(\sum_{i=1}^{N} x_i x_i^T - \sum_{k=1}^{K} N_k \hat{\mu}_k \hat{\mu}_k^T \right) \\
=& \frac{1}{N-K} \left(\mathbf{X}^T\mathbf{X} - \sum_{k=1}^{K} \frac{\mathbf{X}^T y_k y_k^T \mathbf{X}}{N_k} \right) \\
=& \frac{1}{N-K} \left(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{Y}\mathbf{D}\mathbf{Y}^T\mathbf{X} \right)
\end{aligned}$$

where $\mathbf{D} = diag(\frac{1}{N_1}, \frac{1}{N_2}, .., \frac{1}{N_K})$

Therefore, we can compute $\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}}$ as follows:

$$\begin{aligned}
\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}} =& \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \left[\frac{1}{N-K} (\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{Y}\mathbf{D}\mathbf{Y}^T\mathbf{X}) \right] (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
=& \frac{1}{N-K} \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}[\mathbf{I} - \mathbf{D}\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\
=& \frac{1}{N-K} \mathbf{Q}(\mathbf{I} - \mathbf{D}\mathbf{Q})
\end{aligned}$$

where $\mathbf{Q} = \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

Then, we have

$$\begin{aligned}
& \hat{\mathbf{\Sigma}}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T\mathbf{X}^T\mathbf{Y} \\
=& \frac{1}{N-K} \left(\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\mathbf{Y}\mathbf{D}\mathbf{Y}^T\mathbf{X} \right) (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \left[\frac{1}{N-K} \mathbf{Q}(\mathbf{I} - \mathbf{D}\mathbf{Q}) \right]^{-1} \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
=& \mathbf{X}^T\mathbf{Y}[\mathbf{I} - \mathbf{D}\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}](\mathbf{I} - \mathbf{D}\mathbf{Q})^{-1}\mathbf{Q}^{-1}\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
=& \mathbf{X}^T\mathbf{Y}(\mathbf{I} - \mathbf{D}\mathbf{Q})(\mathbf{I} - \mathbf{D}\mathbf{Q})^{-1}\mathbf{Q}^{-1}\mathbf{Q} \\
=& \mathbf{X}^T\mathbf{Y}
\end{aligned}$$

Since

$$\mathbf{X}^T\mathbf{Y} = [\mathbf{X}^T y_1 \quad \mathbf{X}^T y_2 \quad \dots \quad \mathbf{X}^T y_K] = [N_1\hat{\mu}_1 \quad N_2\hat{\mu}_2 \quad \dots \quad N_K\hat{\mu}_K]$$

We can get,

$$
\begin{aligned}
&\hat{\mathbf{\Sigma}}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T[N_1\hat{\mu}_1 \quad N_2\hat{\mu}_2 \quad \dots \quad N_K\hat{\mu}_K] = [N_1\hat{\mu}_1 \quad N_2\hat{\mu}_2 \quad \dots \quad N_K\hat{\mu}_K] \\
\Longrightarrow &\hat{\mathbf{\Sigma}}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T N_k\hat{\mu}_k = N_k\hat{\mu}_k \quad (k = 1, 2, ..., K) \\
\Longrightarrow &\hat{\mathbf{\Sigma}}\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T\hat{\mu}_k = \hat{\mu}_k \\
\Longrightarrow &\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T\hat{\mu}_k = \hat{\mathbf{\Sigma}}^{-1}\hat{\mu}_k \\
\Longrightarrow &\hat{\mathbf{B}}(\hat{\mathbf{B}}^T\hat{\mathbf{\Sigma}}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^T(\hat{\mu}_k - \hat{\mu}_l) = \hat{\mathbf{\Sigma}}^{-1}(\hat{\mu}_k - \hat{\mu}_l)
\end{aligned}
$$

Hence, we achieve the conclusion of the problem.

□