Suggestions - Problem Set 3

4.2 (a) Show the discriminant condition (1) takes the form

\[ x^T \Sigma^{-1} (\mu_2 - \mu_1) > \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \ln \frac{N_1}{N} - \ln \frac{N_2}{N}, \]

as desired. We then replace the quantities \( \mu_i, \Sigma_i \) by their estimates to get the proper form for this discriminant.

(b) Here using the output notations \( y_1 = -\frac{N}{N_1}, y_2 = \frac{N}{N_2} \) for classes 1 and 2 respectively, you want to minimize

\[ \sum_{i=1}^{N} (y_i - \beta_0 - \beta^T x_i)^2 = (y - \hat{X} \hat{\beta})^2, \]

where \( \hat{\beta} = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} \), letting \( X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix} \) and \( \hat{X} = [1_N \ X] = \begin{bmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix} \).

In general vectors/matrices with a ~ on them can represent vectors augmented with 1’s (and in some cases 0’s).

Use the usual least squares so that

\[ \hat{\beta} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T y \]

approximates \( \hat{\beta} \).

Thus

\[ \hat{X}^T \hat{X} \hat{\beta} = \hat{X}^T y. \quad (1) \]

First consider the right hand side, \( \hat{X}^T y \). Without loss you can arrange the data so the first \( N_1 \) examples \((x_i, y_i)\) are in the first class and the last \( N_2 \) are in the second.

Thus show the right side of (1) becomes:

\[ \hat{X}^T y = \begin{bmatrix} 1_N^T \\ X^T \end{bmatrix} y = \begin{bmatrix} 1_N^T y \\ X^T y \end{bmatrix} = \begin{bmatrix} 0 \\ \hat{X}^T y \end{bmatrix}. \]

Meantime show

\[ X^T y = - \frac{N}{N_1} \sum_{j=1}^{N_1} x_i + \frac{N}{N_2} \sum_{j=N_1+1}^{N} x_i = N(\hat{\mu}_2 - \hat{\mu}_1). \]

So
To calculate the left side of (1), you can write

\[
\mathbf{X}^T \mathbf{y} = \begin{bmatrix} N(\mu_2 - \mu_1) \\
0 \end{bmatrix}.
\]  

(2)

Let

\[
\mathbf{X} = \begin{bmatrix}
1 & \mathbf{x}_1^T \\
1 & \mathbf{x}_2^T \\
1 & \mathbf{x}_N^T \\
1 & \mathbf{x}_{N+1}^T \\
\vdots & \vdots \\
1 & \mathbf{x}_N^T 
\end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\
\mathbf{X}_2 
\end{bmatrix},
\]

\[
\mathbf{M} = \begin{bmatrix}
\frac{1}{N_1} & \mathbf{1}_{N_1}^T & \mathbf{1}_{N_1} \\
\frac{1}{N_2} & \mathbf{1}_{N_2}^T & \mathbf{1}_{N_2}
\end{bmatrix} = [1_N \quad \mathbf{M}]
\]

i.e. \( \mathbf{M} \) is the matrix whose first \( N_1 \) rows are copies of \( \mu_1^T \), and whose last \( N_2 \) rows are copies of \( \mu_2^T \). Here \( 1_N \) is always a column vector of length \( N \) with all 1's.

Then show

\[
\mathbf{X}^T \mathbf{X} = \begin{bmatrix} \mathbf{1}_N^T \\
\mathbf{X}^T \end{bmatrix} \mathbf{1}_N \mathbf{X} = \begin{bmatrix} \mathbf{1}_N^T & \mathbf{1}_N^T \mathbf{X} \\
\mathbf{X}^T & \mathbf{X}^T \mathbf{X} \end{bmatrix} = \begin{bmatrix} N & \mathbf{N}_1 \mu_1^T + \mathbf{N}_2 \mu_2^T \\
\mathbf{X}^T & \mathbf{X}^T \mathbf{X} \end{bmatrix}
\]

Thus show

\[
\mathbf{X}^T \mathbf{X} \hat{\beta} = \begin{bmatrix} N \\
\mathbf{X}^T \mathbf{X} \end{bmatrix} \begin{bmatrix} \beta_0 \\
\hat{\beta} \end{bmatrix} = \begin{bmatrix} N \hat{\beta}_0 + (\mathbf{N}_1 \mu_1 + \mathbf{N}_2 \mu_2) \hat{\beta} \\
(\mathbf{N}_1 \mu_1 + \mathbf{N}_2 \mu_2) \hat{\beta}_0 + \mathbf{X}^T \mathbf{X} \hat{\beta} \end{bmatrix}.
\]

Now from the relationship

\[
\hat{\mathbf{y}} = \mathbf{X} \hat{\beta} = \begin{bmatrix} \mathbf{1}_N & \mathbf{X} \end{bmatrix} \begin{bmatrix} \beta_0 \\
\hat{\beta} \end{bmatrix} = \hat{\beta}_0 \mathbf{1}_N + \mathbf{X} \hat{\beta},
\]

if you average over the entries \( \hat{\mathbf{y}} \) of \( \hat{\mathbf{y}} \), show

\[
0 = \mathbf{1}_N^T \hat{\mathbf{y}} = N \hat{\beta}_0 + \mathbf{1}_N^T \mathbf{X} \hat{\beta} = N \hat{\beta}_0 + (\mathbf{N}_1 \mu_1 + \mathbf{N}_2 \mu_2)^T \hat{\beta},
\]

so

\[
0 = N \hat{\beta}_0 + (\mathbf{N}_1 \mu_1 + \mathbf{N}_2 \mu_2)^T \hat{\beta},
\]
\[ \hat{\beta}_0 = - \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T \hat{\beta}, \]

so now
\[
\hat{X}^T \hat{X} \hat{\beta} = \begin{bmatrix}
0 \\
- \left( N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2 \right) \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T \hat{\beta} + X^T X \hat{\beta}
\end{bmatrix}. \tag{3}
\]

You can write
\[
X^T X = (X - M)^T (X - M) + X^T M + M^T X - M^T M
\]

But show
\[
(X - M)^T (X - M) = (N - 2) \bar{\Sigma}
\]
\[
X^T M = \sum_{j=1}^{N_1} x_j \hat{\mu}_1^T + \sum_{j=N_1+1}^{N_2} x_j \hat{\mu}_2^T = N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T
\]
\[
M^T M = N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T.
\]

Thus
\[
X^T X = (N - 2) \bar{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T.
\]

So by (3) above, show
\[
\hat{X}^T \hat{X} \hat{\beta}
\]
\[
= \begin{bmatrix}
0 \\
- \left( N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2 \right) \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T + (N - 2) \bar{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T \end{bmatrix} \hat{\beta}. \tag{4}
\]

Now show the bottom term coefficient is
\[
- \left( N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2 \right) \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T + (N - 2) \bar{\Sigma} + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T
\]
\[
= (N - 2) \bar{\Sigma} + \left[ \frac{N_1 N_2}{N^2} \right] (\hat{\mu}_1 - \hat{\mu}_2)(\hat{\mu}_1 - \hat{\mu}_2)^T
\]
Now use (1), (2), (4) and (5).

\[(5) \quad (N - 2) \bar{\Sigma} + \left[ \frac{N_1 N_2}{N} \right] \Sigma_B \]

(e) It follows that

\[\hat{\beta} = (\hat{\beta}_2 - \hat{\mu}_1)(\hat{\mu}_2 - \hat{\mu}_1)^T \beta = [\hat{\mu}_2 - \hat{\mu}_1] \beta (\hat{\mu}_2 - \hat{\mu}_1),\]

which is clearly in the direction of \((\hat{\mu}_2 - \hat{\mu}_1)\), since \([\hat{\mu}_2 - \hat{\mu}_1] \beta\) is a scalar (why?)

Finally from (4.56),

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

\[= \text{(scalar) } \bar{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1).\]

(d) Changing the coding for the two \(y\) values transforms the pair of numbers \(- \frac{N}{N_1}\) and \(\frac{N}{N_2}\) respectively into another pair \(a\) and \(b\) of possible \(y\) values. Show that there is a linear scalar transformation \(y^* = cy + d = f(y)\) such that \(f\left(- \frac{N}{N_1}\right) = a\) and \(f\left(\frac{N}{N_2}\right) = b\).

What are \(c\) and \(d\)? Now show that if \(y\) has only entries \(- \frac{N}{N_1}\) and \(\frac{N}{N_2}\), then in their places the vector \(y^* = cy + d1_N\) will have \(a\) and \(b\) respectively. Further show that if we replace \(y\) by \(y^*\) in the dataset \(\{(x_i, y_i)\}_{i=1}^N\), then we will have a new

\[\hat{y}^* = X^T \hat{\beta}^* = X(X^T X)^{-1} X^T y^* = Hy = H(cy + d1_N) = c\hat{y} + d1_N.\]

(why is \(H1_N = 1_N\); recall \(H\) is a projection). Thus the transformation to \(\hat{y}\) is exactly the same as the transformation to \(y\) above. Show in fact that the transformation acts in exactly the same way on each component of \(\hat{y}\). Now show that the final selection of classes based on the new \(\hat{y}^* = \begin{bmatrix} \hat{y}_{1}^* \\ \hat{y}_{2}^* \\ \vdots \\ \hat{y}_{N}^* \end{bmatrix}\) will be based on each entry \(\hat{y}_i\), and whether it is closer to \(a\) (choose class 1) or \(b\) (choose class 2). Show that \(\hat{y}_i^*\) is closer to \(a\) iff \(\hat{y}_i\) is closer to \(- \frac{N}{N_1}\).

(e) Now you have \(\hat{\beta}\) and \(\beta_0\) and the regression function

\[\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_T x.\]
From part (c), \( \hat{\beta} = k \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) \) for some \( k \). Thus show from above that

\[
\hat{\beta}_0 = - \left( \frac{N_1}{N} \hat{\mu}_1 + \frac{N_2}{N} \hat{\mu}_2 \right)^T k \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1).
\]

Recall the group targets (y-values) on which we have trained the regression are:

\[
\text{class 1: } y = - \frac{N}{N_1}; \quad \text{class 2: } y = \frac{N}{N_2}.
\]

For an input test vector \( \mathbf{x} \), show that the corresponding \( y \) will be in class 1 if \( f(\mathbf{x}) \) is closer to \( - \frac{N}{N_1} \), than to \( \frac{N}{N_2} \), and otherwise class 2. Show \( y \) should be assigned to class 2 if

\[
f(\mathbf{x}) > \frac{1}{2} \left( \frac{-N}{N_1} + \frac{N}{N_2} \right).
\]

Show from above that the criterion for class 2 assignment is:

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

or

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

Is this the same as the LDA criterion in (a)? Now assume \( N_1 = N_2 = N/2 \) - what happens then?

### 4.3

Recall the LDA criterion for choosing the group \( l \) out of groups \( 1, ..., K \) given a test feature vector \( \mathbf{x} \) is

\[
l = \arg \max_k \delta_k(\mathbf{x}),
\]

i.e., finding the \( l = k \) which makes \( \delta_k(\mathbf{x}) \) the largest. Here as usual

\[
\delta_k(\mathbf{x}) = \mathbf{x}^T \hat{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \hat{\Sigma}^{-1} + \ln \pi_k.
\]

This problem is related to the discussion in section 4.2 involving the use of a regression approach to distinguish among the \( K \) predictions. This works by choosing targets (representatives of the \( K \) classes to be set equal to the response variable \( y \)) as follows.

For a vector \( \mathbf{x} \) whose class is \( k \), we choose the response variable to be \( \mathbf{y} = (0, ..., 0, 1, 0..., 0) \) (a row vector), with a 1 only in the \( k^{th} \) position. Then if we are given a training set \( \tau = \{ (\mathbf{x}_i, y_i) \}_{i=1}^N \), the responses are no longer \( y_i = 0 \) or 1, but vectors \( \mathbf{y}_i \) with a 1 in the \( k^{th} \) position if the class assigned to \( \mathbf{x}_i \) is group \( k \).
As shown in the text, the appropriate regression here works exactly as in the case the responses \( y_i \) are scalars, except that the usual vector

\[
\mathbf{y} = \begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_N
\end{bmatrix}
\]

with each row a scalar (0 or 1) representing the class of the \( x_i \) is replaced by a matrix

\[
\mathbf{Y} = \begin{bmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_N
\end{bmatrix},
\]

with each class indicator \( y_i \) indicating the class through the position of its only entry 1 (note again that each \( y_i \) is a row vector).

By adding the usual column of 1's we form

\[
\tilde{\mathbf{Y}} = \begin{bmatrix}
  1 & y_1 \\
  \vdots & \vdots \\
  1 & y_N
\end{bmatrix}.
\]

Otherwise the regression process is the same, with the vector \( \mathbf{y} \) replaced by the matrix \( \mathbf{Y} \). Now following the regression discussion in the text, the usual estimated value \( \hat{\mathbf{y}} \) of \( \mathbf{y} \) is replaced using the same formula to get an estimated value \( \hat{\mathbf{Y}} \) of \( \mathbf{Y} \):

\[
\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y},
\]

which has exactly the same form as standard regression, with \( \mathbf{y} \) replaced by \( \mathbf{Y} \). Notice

\[
\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}
\]

with

\[
\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.
\]

Note that as in our general regression discussion the matrix \( \mathbf{X} \) is assumed to already contain an initial column of 1's. Row-wise, defining \( \tilde{\mathbf{y}}_i \) to be the \( i^{th} \) row of

\[
\tilde{\mathbf{Y}} = \begin{bmatrix}
  \tilde{y}_1 \\
  \tilde{y}_2 \\
  \vdots \\
  \tilde{y}_N
\end{bmatrix},
\]

equation (8a) is equivalent to \( \hat{\mathbf{y}}_i = \tilde{x}_i^T\hat{\mathbf{B}} \); here and elsewhere the tilde \( \sim \) on a vector means we have added 1's in the initial position:

\[
\hat{x}_i = \begin{bmatrix}
  1 \\
  \mathbf{x}_i
\end{bmatrix}.
\]
Note that whereas previously $\hat{\beta}$ was given by the same formula (8c), now $\hat{\mathbf{B}}$ is a matrix instead of a vector. Whereas previously we had $\hat{y}_i = \hat{x}_i^T \hat{\beta}$ as the estimated value of $y_i$ within the dataset, we now have instead

$$\hat{y}_i = \hat{x}_i^T \hat{\mathbf{B}},$$

where $\hat{y}_i$ is a vector (the $i^{th}$ row of $\hat{\mathbf{Y}}$).

We are asking what would happen if we simply replace this training set $\tau = \{(x_i, y_i)\}$ with a new training set replacing the input vectors $x_i$ by the corresponding estimates $\hat{x}_i^T$, so that the training set now looks like $\tau' = \{ (\hat{y}_i^T, y_i) \}_{i=1}^N$. Note that we are using the transpose in $\hat{y}_i^T$ because we want it to be a column vector (why?), replacing the original column input vector $x_i$. Equivalently, we are replacing the training matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_N^T \end{bmatrix}$$

by the response vector set $\hat{\mathbf{Y}} = \begin{bmatrix} 1 & \hat{y}_1 \\ \vdots & \vdots \\ 1 & \hat{y}_N \end{bmatrix}$.

We wish to show that if we use this new dataset in both training and testing, then we will still get the same class predictions for a new test vector $\mathbf{x}$, but using LDA (not regression here).

Show that given that $l = \arg \max_k \delta_k(x)$, we just need to check how the computation of the $\delta_k(x)$ changes using the new data set. Note the training data now have the form

$$\tau' = \{ (\hat{y}_i^T, y_i) \}_{i=1}^N = \{ (\hat{\mathbf{B}}^T \hat{x}_i, y_i) \}_{i=1}^N.$$

The original discriminant function has the form

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

where $\hat{\pi}_k = \frac{N_k}{N}$. Show the new discriminant function $\tilde{\delta}_k(x)$ has the form

$$\tilde{\delta}_k(\hat{y}^T) = \hat{y}^T \Sigma^{-1} \mu_k^* - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k^* + \ln \hat{\pi}_k,$$

where

$$\mu_k^* = \frac{1}{N_k} \sum_{j=1, g(j) = k}^N \hat{y}_j = \frac{1}{N_k} \sum_{j=1, g(j) = k}^N \hat{\mathbf{B}}^T \hat{x}_j = \hat{\mathbf{B}}^T \hat{\mu}_k,$$

$$\Sigma^* = \frac{1}{N - K} \sum_{j=1}^N (\hat{y}_j - \mu_{g(j)}) (\hat{y}_j - \mu_{g(j)})^T.$$
vectors of interest based on their individual groups, but with $x_i$ replaced by $\hat{y}_i^T$. Here $g(j)$ represents the group (out of the $K$ total) of the $j^{th}$ sample.

You wish to show that the modified discriminant functions make the same decisions as the original ones, i.e., that whenever $y = x^T \hat{B}$,

$$\delta_k^*(\hat{y}^T) > \delta_l^*(\hat{y}^T) \iff \delta_k(x) > \delta_l(x).$$

But note that $\delta_k^*(\hat{y}^T)$ has the form (10), with

$$\mu_k^* = (\hat{\mu}_k^T \hat{B})^T = \hat{B}^T \hat{\mu}_k$$

and show

$$\Sigma^* = \frac{1}{N-K} \sum_{j=1}^N (\hat{y}_i^T - \mu_{g(i)})^T (\hat{y}_i^T - \mu_{g(i)})^T$$

$$= \frac{1}{N-K} \sum_{j=1}^N \hat{B}^T (\hat{x}_i - \hat{\mu}_{g(i)}) (\hat{x}_i - \hat{\mu}_{g(i)})^T \hat{B}$$

$$= \hat{B}^T \hat{\Sigma} \hat{B}$$

where, because our vectors are augmented to have a 1 in the first position (and thus are of length $p+1$) we must also augment the covariance matrix $\hat{\Sigma}$ in order to be $(p+1) \times (p+1)$, by adding a first row and first column of 0’s. That is, we define

$$\Sigma = \begin{bmatrix} 0 & \mathbf{0}_p \\ \mathbf{0}_p & \Sigma \end{bmatrix},$$

where $\mathbf{0}_p$ is a column vector of length $p$ with all zeroes, and the upper left corner is $1 \times 1$. Of course $\hat{\Sigma}$ is the estimator of $\Sigma$.

Thus show we can write

$$\delta_k^*(\hat{y}^T) = \hat{y} \Sigma^*-1 \mu_k^* - \frac{1}{2} \mu^* \Sigma^*-1 \mu_k^* + \ln \hat{\pi}_k$$

$$= \hat{x}^T \hat{B} (\hat{B}^T \hat{\Sigma} \hat{B})^{-1} \hat{B}^T \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{B} (\hat{B}^T \hat{\Sigma} \hat{B})^{-1} \hat{B}^T \hat{\mu}_k + \ln \hat{\pi}_k$$

(10a)

Now to define the square root of a matrix. For any $p \times p$ square symmetric invertible matrix $A$, assume that $\{\lambda_i, \mathbf{a}_i\}_{i=1}^p$ are its eigenvalues and corresponding eigenvectors. For a function $f(x)$ define $f(A)$ to be the matrix with the same eigenvectors $\mathbf{a}_i$, but eigenvalues $f(\lambda_i)$. Thus $A^{-1/2}$ would have $\{\lambda_i^{-1/2}, \mathbf{a}_i\}_{i=1}^p$ as it eigenvalue-eigenvector pairs.
Now we replace our dataset $x_i \rightarrow \Sigma^{-1/2} x_i$. This leads to the replacement

$$X^T \rightarrow X^T \Sigma^{-1/2}.$$

Clearly $\hat{Y}$ in (7) does not change under this transformation (why?). However, now with the transformed $X$ values, show we have

$$V(X) = V(\Sigma^{-1/2} X) = \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = I.$$

This transformed dataset thus leads to the same $\hat{Y}$, but gives a linear discriminant function which has changed to (now $\hat{\Sigma} = I$)

$$\delta_k(x) = x^T \hat{\mu}_k - \frac{1}{2} \hat{\mu}_k^T \hat{\mu}_k + \ln \hat{\pi}_k,$$  

(11)

Show this is actually identical to that before — remember that the new $x^T$ equals the old $x^T$ times $\Sigma^{-1/2}$ and we have also computed the $\mu_k$ from the new dataset; thus the classes obtained from using the discriminants in (11) (using the new transformed dataset) will be identical to what the predicted classes were before.

Also show the transformed discriminant function (now changing these $x^T$ into the unchanged $\hat{Y}$ and forming the resulting discriminant) must be exactly the same as when we replaced by $\hat{Y}$ before, since the dataset $\{\hat{y}_i, y_i\}_{i=1}^N$ is identical (see (9a)). Thus we need only show the result of this problem for the new (transformed) dataset $\{x_i, \hat{y}_i\}$, where the new $x_i$ are defined as above.

Show using the same argument we will again replace our dataset so that each current datapoint $x_i$ will be replaced by $x_i - \mu$, with $\mu$, i.e. the overall current mean of all $x_i$ (regardless of class). Show this does not change the covariance $\hat{\Sigma}$, and in terms of the new dataset the identical discriminant functions will now be

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

(13)

(again with $x$ obtained from the new mean-subtracted dataset). Show by translating all data by the same amount $\mu$ will not change the relative sizes of the discriminant functions, and so if we replace the old discriminant function (13) by

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

(15)

then clearly this will not affect whether $\delta_k(x) > \delta_l(x)$ or not. Furthermore, since $\hat{Y}$ is a standard regression estimator (just with multiple columns), show a translation of the dataset will not affect the predictions, so that with this new dataset the $\hat{Y}$ we obtain is identical to the previous one.
Thus at this point you have reduced the problem to having a dataset $\mathbf{X}$ with empirical mean 0 and standard deviation 1 for each coordinate, and we still have the same discriminant functions (15) and estimator $\hat{\mathbf{Y}}$, derived in the same way from the new data and the outcome matrix $\mathbf{Y}$.

This means we are using the discriminant function (11) above; show equation (10a) becomes

$$
\delta_k^*(\mathbf{y}^T) = \mathbf{x}^T \hat{\mathbf{B}} (\hat{\mathbf{B}}^T \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^T \mu_k - \frac{1}{2} \hat{\mu}_k^T (\hat{\mathbf{B}} (\hat{\mathbf{B}}^T \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^T \mu_k + \ln \hat{\mu}_k.
$$

But show that $\mathbf{H} \equiv \hat{\mathbf{B}} (\hat{\mathbf{B}}^T \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^T$ is just the projection onto the column space of $\hat{\mathbf{B}}$ (see the discussion on p. 46 of the hat function, which projects $\mathbf{y}$ onto the column space of $\mathbf{X}$, giving $\hat{\mathbf{y}}$).

Now show that $\hat{\mu}_k$ is in the column space of $\hat{\mathbf{B}}$. Note our assumptions of covariance $I$ and mean 0 for the $\mathbf{x}_i$, it follows that $\mathbf{X}^T \mathbf{X} = I$, so $\mathbf{B} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{Y}$. Thus the $k^{th}$ column $\mathbf{b}_k$ of $\mathbf{B}$ is just $\mathbf{b}_k = \mathbf{X}^T \mathbf{y}_k$, where $\mathbf{y}_k$ denotes the $k^{th}$ column (not row) of $\mathbf{Y}$. But from the definition of $\mathbf{X}^T = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_N]$ and $\mathbf{y}_k$ (whose $i^{th}$ entry is 0 unless $\mathbf{x}_i$ is in class $k$), the column $\mathbf{b}_k$ must be just $\mathbf{b}_k = \sum_{i \in G_k} \mathbf{x}_i$, i.e., a multiple of $\hat{\mu}_k$. Thus, clearly all $\hat{\mu}_k$ are in the column space of $\hat{\mathbf{B}}$, and hence $\hat{\mathbf{H}} \hat{\mu}_k = \hat{\mu}_k$ for all $k$.

Thus by (17) show

$$
\delta_k^*(\hat{\mathbf{y}}^T) = \mathbf{x}^T \mathbf{H} \mu_k - \frac{1}{2} \hat{\mu}_k^T \mathbf{H} \mu_k + \ln \hat{\pi}_k = \mathbf{x}^T \mu_k - \frac{1}{2} \mu_k^T \mu_k + \ln \hat{\pi}_k,
$$

that is, the discriminant from the $\hat{\mathbf{Y}}$-based discriminant function gives identical values to the discriminant (15), which we have shown gives identical choices to the discriminant based on the original dataset, as desired.