MVA720 Assignment

Discriminant Analysis

Ethan-Jöhl Smith (20452277)

Department of Statistics, University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

October 5, 2025

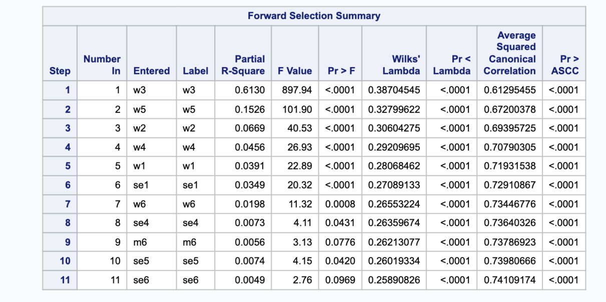
Question 1

Tumour dataset description

In this question, we analyse the "tumours" dataset, which contains data from 569 individuals whose breast masses were assessed to classify each tumour as benign (B) or malignant (M), our two population groups of interest. The dataset has no missing values and includes a sample of 357 (around 63%) benign and 212 (around 37%) malignant tumour cases. Each tumour was described by a set of 18 features derived from six real-valued characteristics – radius, texture, perimeter, area, smoothness, and compactness – each measured using three metrics: mean, standard error, and worst value.

Stepwise Discriminant Analysis

```
proc stepdisc data=tumours method=forward slentry=0.15 slstay=0.15 short;
class diagnosis;
var m1-m6 se1-se6 w1-w6;
run;
```



Forward Selection Summary

Step	Number In	Entered	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	w3	w3	0.6130	897.94	<.0001	0.38704545	<.0001	0.61295455	<.0001
2	2	w5	w5	0.1526	101.90	<.0001	0.32799622	<.0001	0.67200378	<.0001
3	3	w2	w2	0.0669	40.53	<.0001	0.30604275	<.0001	0.69395725	<.0001
4	4	w4	w4	0.0456	26.93	<.0001	0.29209695	<.0001	0.70790305	<.0001
5	5	w1	w1	0.0391	22.89	<.0001	0.28068462	<.0001	0.71931538	<.0001
6	6	se1	se1	0.0349	20.32	<.0001	0.27089133	<.0001	0.72910867	<.0001
7	7	w6	w6	0.0198	11.32	0.0008	0.26553224	<.0001	0.73446776	<.0001
8	8	se4	se4	0.0073	4.11	0.0431	0.26359674	<.0001	0.73640326	<.0001
9	9	m6	m6	0.0056	3.13	0.0776	0.26213077	<.0001	0.73786923	<.0001
10	10	se5	se5	0.0074	4.15	0.0420	0.26019334	<.0001	0.73980666	<.0001
11	11	se6	se6	0.0049	2.76	0.0969	0.25890826	<.0001	0.74109174	<.0001

Figure 1: Proc stepdisc results

Of the 18 features (predictor variables) in the dataset, the forward-selection method identified the following variables as grouping variables for use in ‘PROC DISCRIM‘: *m6, se1, se4, se5, se6, w1, w2, w3, w4, w5,* and *w6*. However, had we used the ’STEPWISE’ method instead, variable *w3* would have been eliminated at step 7, as its p-value (0.1747) would have exceeded the specified maximum significance level to stay (slstay = 0.15). The forward-selection method considers only the maximum significance level for adding variables (slentry = 0.15), while the ’STEPWISE’ method applies both slentry and slstay thresholds. For our analysis, we proceed with the grouping variables selected by the forward-selection method.

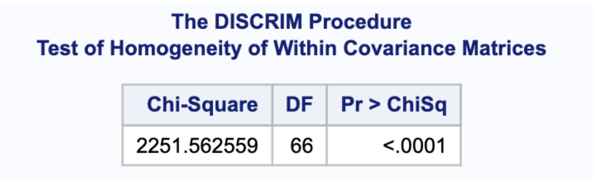
Create training and test/validation datasets

```
data train_data test_data;
set tumours;
if set = 'train' then output train_data;
else if set = 'validate' then output test_data;
run;
```

The tumours dataset was split into a training set of 353 observations and a test/validation set of 216 observations. To evaluate model performance, cross-validation will be conducted within 'PROC DISCRIM' on the training data. Proportional priors were selected, reflecting the assumption that individuals undergoing breast tissue analysis are likely motivated by health concerns, which suggests an imbalanced class/group distribution rather than an equal 50/50 split between malignant and benign cases. Proportional priors thus provide a data-informed approach to estimating group prior probabilities, assuming the sample is representative of the broader population.

```
proc discrim data = train_data POOL = TEST MANOVA METHOD= NORMAL crossvalidate
testdata = test_data;
class diagnosis;
priors proportional;
var &_stdvar;
```

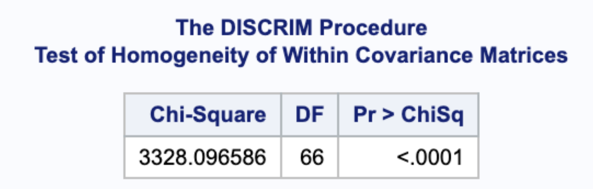
Barlett's test for equality of covariance matrix



The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2251.562559	66	<.0001

Figure 2: Bartlett's test results (Training dataset)



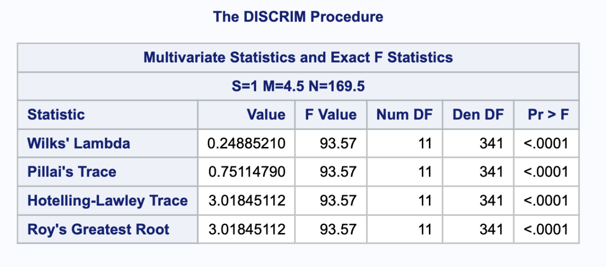
The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
3328.096586	66	<.0001

Figure 3: Bartlett's test results (Full dataset)

By specifying option 'POOL = TEST' in PROC DISCRIM, we test $H_0 : \mathbf{\Sigma_M} = \mathbf{\Sigma_B}$ against $H_A : \mathbf{\Sigma_M} \neq \mathbf{\Sigma_B}$. Since Bartlett's test yields a p-value < 0.0001, which is less than any reasonable significance level α, we reject H_0 in both the training and full datasets. This result indicates a statistically significant difference between the variance-covariance matrices of the benign and malignant tumour populations, suggesting non-homogeneity. Consequently, Quadratic Discriminant Analysis (QDA) is more appropriate than Linear Discriminant Analysis (LDA) for this data.

MANOVA with variables selected from the PROC STEPDISC



The DISCRIM Procedure

Multivariate Statistics and Exact F Statistics

S=1 M=4.5 N=169.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.24885210	93.57	11	341	<.0001
Pillai's Trace	0.75114790	93.57	11	341	<.0001
Hotelling-Lawley Trace	3.01845112	93.57	11	341	<.0001
Roy's Greatest Root	3.01845112	93.57	11	341	<.0001

Figure 4: MANOVA results

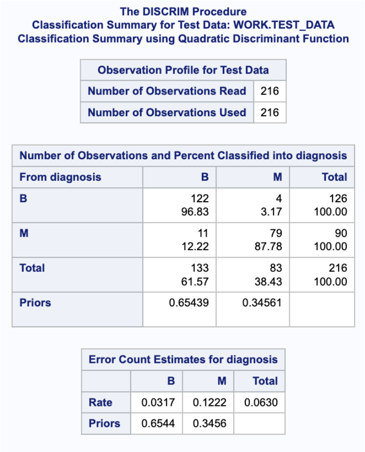
By using the MANOVA option, we test $H_0 : \boldsymbol{\mu_M} = \boldsymbol{\mu_B}$ against $H_A : \boldsymbol{\mu_M} \neq \boldsymbol{\mu_B}$.

All the p-values for the various types of test statistics are $< 0.0001 < \alpha$ (at any reasonable significance level), so we reject H_0 and conclude that there is a statistically significant difference in the population means between the benign and malignant tumour groups. This result suggests that performing discriminant analysis is likely to yield meaningful classification functions, as the set of 11 predictor variables (selected through the 'STEPDISC' procedure) provides sufficient separation between the groups.

For the following sections we consider the following code:

```
proc discrim data = train_data POOL = NO  METHOD= NORMAL simple outstat= info1
crossvalidate testdata = test_data;
class diagnosis;
priors proportional;
var &_stdvar;
```

Confusion matrix for the validation data and calculation of Performance Measures



The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST_DATA
Classification Summary using Quadratic Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	216
Number of Observations Used	216

Number of Observations and Percent Classified into diagnosis

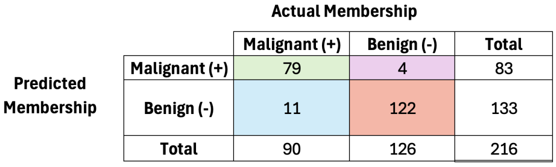
From diagnosis	B	M	Total
B	122 96.83	4 3.17	126 100.00
M	11 12.22	79 87.78	90 100.00
Total	133 61.57	83 38.43	216 100.00
Priors	0.65439	0.34561	

Error Count Estimates for diagnosis

	B	M	Total
Rate	0.0317	0.1222	0.0630
Priors	0.6544	0.3456	

Figure 5: Confusion matrix for test set

If we designate the Malignant (M) group as the "Positive" class and the Benign (B) group as the "Negative" class, we can reformulate the confusion matrix as follows:



		Actual Membership		
		Malignant (+)	Benign (-)	Total
Predicted Membership	Malignant (+)	79	4	83
	Benign (-)	11	122	133
	Total	90	126	216

Figure 6: Confusion matrix for test set (Reformulated)

In this confusion matrix, green indicates true positives (TP), purple indicates false positives (FP), blue indicates false negatives (FN), and red indicates true negatives (TN).

Specificity represents the proportion of tumour observations correctly classified as benign (predicted as benign (-) when they were actually benign (-)) out of all observations that were truly benign (-).

Sensitivity (Recall) represents the proportion of tumour observations correctly classified as malignant (predicted as malignant (+) when they were actually malignant (+)) out of all observations that were truly malignant (+).

Precision represents the proportion of tumour observations that were correctly classified as malignant (predicted as malignant (+) when they were actually malignant (+)) out of all observations predicted to be malignant (+).

Accuracy represents the proportion of tumour observations correctly classified as either malignant (predicted as malignant (+) when they were actually malignant (+)) or benign (predicted as benign (-) when they were actually benign (-)) out of all tumour observations.

In the test / validation set:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{79 + 122}{79 + 122 + 4 + 11} = \frac{201}{216} \approx 93.06\%$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{79}{79 + 4} = \frac{79}{83} \approx 95.18\%$$

$$\text{Sensitivity (Recall)} = \frac{TP}{TP + FN} = \frac{79}{90} \approx 87.78\%$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{122}{126} \approx 96.83\%$$

Fewer false positives (FPs, where benign tumours are incorrectly predicted as malignant) increase the precision metric, while fewer false negatives (FNs, where malignant tumours are incorrectly predicted as benign) improve the sensitivity metric. The false positive rate ($\frac{FP}{FP+TN}$) corresponds to a Type I error, while the false negative rate ($\frac{FN}{FN+TP}$) corresponds to a Type II error. In tumour diagnosis, minimising Type II errors is particularly important, as failing to identify a malignant tumour poses a greater risk to patient outcomes. The model demonstrates strong overall performance, with high accuracy, precision, and specificity; however, slightly lower sensitivity implies a higher risk of missed malignant cases (false negatives), which could adversely affect patient outcomes.

Graph of QDC scores

Figure 7 presents the QDC_{MB} classification scores for the test and validation tumour dataset, where magenta stars indicate benign classifications and light slate blue stars represent malignant classifications. Tumours with QDC_{MB} scores above 0 are classified as malignant, while those with scores below 0 are classified as benign. Despite the separation between the two groups suggested by centroids (mean QDC scores) of approximately -11.94 for benign and 234.79 for malignant cases – an overlap near zero suggests potential misclassifications or borderline cases. This overlap highlights the need for further examination of the classifier's sensitivity at the decision boundary to improve accuracy in ambiguous cases.

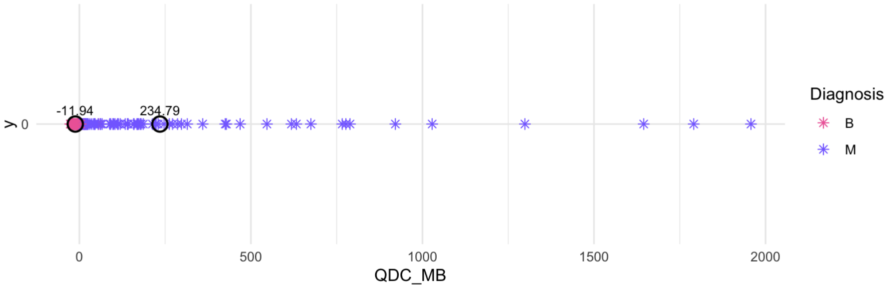


Figure 7: Graph of CDC scores and group centroids

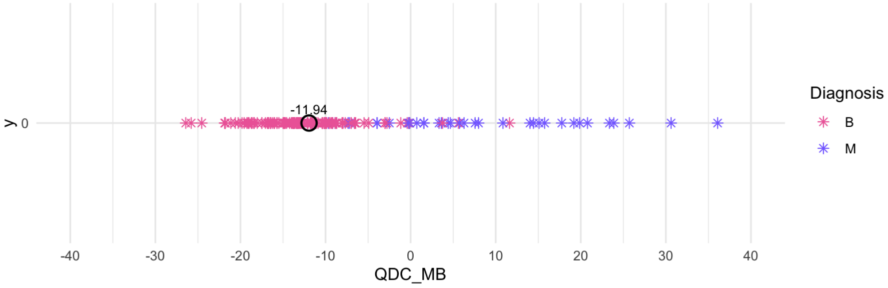


Figure 7 with the horizontal axis zoomed in around the scores for the Benign group

The QDC_{MB} values were calculated using the code provided below, and the resulting QDA output was exported as a CSV file. This CSV was then imported into R, where the plots were generated with centroids added for enhanced visualisation.

```
data quad_disc_func;
set info1;
if _type_ = "QUAD";

*priors;
data prior; set info1;
if _TYPE_ = "PRIOR";
keep m6 ; run;



*means;
data mean0; set info1;
if _TYPE_ = "MEAN" & diagnosis = 'B';
keep &_stdvar; run;

data mean1; set info1;
if _TYPE_ = "MEAN" & diagnosis = 'M';
keep &_stdvar; run;

*standard deviations;
```

```
data S0; set info1;
if _TYPE_ = "COV" & diagnosis = 'B';
keep &_stdvar; run;

data S1; set info1;
if _TYPE_ = "COV" & diagnosis = 'M';
keep &_stdvar; run;


*Calculate the Quadratic Score functions;

proc iml;
use prior; read all  into prior; print prior; p0=prior[1]; p1=prior[2];
use mean0; read all var {&_stdvar} into mean0;
use mean1; read all var {&_stdvar} into mean1;
use S0; read all var {&_stdvar} into S0;
detS0=det(S0);
use S1; read all var {&_stdvar} into S1;
detS1=det(S1);
use test_data;
read all var {&_stdvar} into test;
nn=nrow(test);

sQ0t1=-0.5*log(det(S0));
sQ1t1=-0.5*log(det(S1));
sQ0t3=log(p0);
sQ1t3=log(p1);
sQ0=J(nn,1,0);
sQ1=J(nn,1,0);

do i = 1 to nn;
sQ0t2=(test[i,]-mean0)*inv(S0)*(test[i,]-mean0)`;
sQ1t2=(test[i,]-mean1)*inv(S1)*(test[i,]-mean1)`;
sQ0[i,1] = sQ0t1+sQ0t3-0.5*sQ0t2;
sQ1[i,1] = sQ1t1+sQ1t3-0.5*sQ1t2;
end;

QDC10=sQ1-sQ0;
y= J(nn,1,0);
QDAmatrix= y  QDC10;
colNames= {"y","QDC_MB"};
```

```
create QDAdata from QDAmatrix[colname=colNames];
append from QDAmatrix;
close QDAdata;
quit;

data QDA;
merge test_data QDAdata;
run;


*Calculate group centroids (the average QDC value for each group);

proc means data=QDA; class diagnosis; var QDC_MB; run;

/* Plot QDC_MB values */
GOPTIONS RESET=GLOBAL;
SYMBOL1 VALUE=STAR CV=RED HEIGHT=1.2;
SYMBOL2 VALUE=STAR CV=BLACK HEIGHT=1.2;
proc gplot data=QDA;
plot y*QDC_MB=diagnosis;
run;
ODS RTF CLOSE;

GOPTIONS RESET=GLOBAL;
SYMBOL1 VALUE=STAR CV=RED HEIGHT=1.2;
SYMBOL2 VALUE=STAR CV=BLACK HEIGHT=1.2;

proc gplot data=QDA;
    plot y*QDC_MB=diagnosis /  haxis=(-40 to 40 by 10);
run;
quit;

ODS RTF CLOSE;

proc print data = quad_disc_func;
run;
```

Listing 1: Q1 SAS code

Question 2

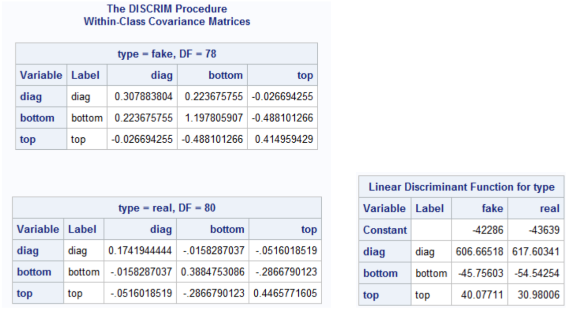
1 Swiss bank note dataset description

In this question, we consider the famous Swiss banknotes dataset. We have two population groups: *genuine/real* and *counterfeit/fake*. In the dataset, there is a sample of 100 fake notes and 100 real notes, each with observed measurements: *length* (length of the note), *left* (width of the left-hand side of the note), *right* (width of the right-hand side of the note), *bottom* (width of the bottom margin), *top* (width of the top margin), and *diag* (diagonal length of the printed area).

a) Linear Discriminant Analysis (LDA)

Consider the following code and output for performing a LDA on the independent variables *diag, bottom, and top*.

```
proc discrim data=swiss1 wcov pool=yes;
  class type;
  var diag bottom top;
  priors "real"=0.99 "fake"=0.01;
run;
```



The DISCRIM Procedure
Within-Class Covariance Matrices

type = fake, DF = 78

Variable	Label	diag	bottom	top
diag	diag	0.307883804	0.223675755	-0.026694255
bottom	bottom	0.223675755	1.197805907	-0.488101266
top	top	-0.026694255	-0.488101266	0.414959429

type = real, DF = 80

Variable	Label	diag	bottom	top
diag	diag	0.1741944444	-.0158287037	-.0516018519
bottom	bottom	-.0158287037	0.3884753086	-.2866790123
top	top	-.0516018519	-.2866790123	0.4465771605

Linear Discriminant Function for type

Variable	Label	fake	real
Constant		-42286	-43639
diag	diag	606.66518	617.60341
bottom	bottom	-45.75603	-54.54254
top	top	40.07711	30.98006

Figure 8: LDA code and relevant output

The test/validation set attained a 100% hit ratio.

The primary assumption of LDA is that the covariance matrices of the *real* (R) and *fake* (F) note populations are statistically indistinguishable (homogenous). Specifically, it is assumed that an observed sample unit $\boldsymbol{x} : p \times 1$ from population π_i follows a multivariate normal distribution, i.e., $(\boldsymbol{X} \mid \pi_i) \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, for $i = 1$ (F) and $i = 2$ (R).

Let the sample covariance matrix for the *fake* group/population be

$$\boldsymbol{S}_F = \boldsymbol{S}_1 = \begin{bmatrix} 0.307883804 & 0.223675755 & -0.026694255 \\ 0.223675755 & 1.197805907 & -0.488101266 \\ -0.026694255 & -0.488101266 & 0.414959429 \end{bmatrix}.$$

and the sample covariance matrix for the *real* group/population be

$$\boldsymbol{S}_R = \boldsymbol{S}_2 = \begin{bmatrix} 0.1741944444 & -.0158287037 & -0.0516018519 \\ -0.0158287037 & 0.3884753086 & -0.2866790123 \\ -0.0516018519 & -0.2866790123 & 0.4465771605 \end{bmatrix}.$$

.

Now, $n_1 = 100$ fake notes and $n_2 = 100$ real notes were observed.

Thus, the pooled covariance matrix is calculated as

$$\boldsymbol{S}_p = \frac{\sum_{i=1}^{2} (n_i - 1)\, \boldsymbol{S}_i}{\sum_{i=1}^{2} (n_i - 1)} \approx \begin{bmatrix} 0.2410 & 0.1039 & -0.0391 \\ 0.1039 & 0.7931 & -0.3874 \\ -0.0391 & -0.3874 & 0.4308 \end{bmatrix}.$$

We aim to classify a banknote of unknown origin, with measurements $\boldsymbol{x} = (x_1\ x_2,\ x_3) = (139.6,\ 8.0,\ 11.0)$ (diag, bottom, and top measurements, respectively), as either *fake* or *real* by assigning it to group i that has the highest posterior probability $\hat{p}\,(\pi_i \mid \boldsymbol{x})$. For this we use the linear discriminant score functions $s_R^L(\boldsymbol{x})$ and $s_F^L(\boldsymbol{x})$.

From the given output information, we get the estimated linear discriminant score functions

$$\widehat{s}_F^L(\boldsymbol{x}) = \left(\widehat{d}_{F0} + \ln \widehat{p}_F\right) + \sum_{j=1}^{3} \widehat{d}_{Fj} x_j = -42286 + 606.66518 \cdot (139.6) - 45.75603 \cdot (8.0) + 40.07711 \cdot (11.0) = 42479.26$$

and

$$\widehat{s}_R^L(\boldsymbol{x}) = \left(\widehat{d}_{R0} + \ln \widehat{p}_R\right) + \sum_{j=1}^{3} \widehat{d}_{Rj} x_j = -43639 + 617.60341 \cdot (139.6) - 54.54254 \cdot (8.0) + 30.98006 \cdot (11.0) = 42482.88$$

Thus, the estimated posterior probability that this note is *fake* is

$$\widehat{p}\,(\pi_F \mid \boldsymbol{x}) = \frac{e^{\widehat{s}_F^L(\boldsymbol{x})}}{e^{\widehat{s}_F^L(\boldsymbol{x})} + e^{\widehat{s}_R^L(\boldsymbol{x})}} = \frac{1}{1 + e^{[\widehat{s}_R^L(\boldsymbol{x}) - \widehat{s}_F^L(\boldsymbol{x})]}}$$

$$\approx \frac{e^{42479.26}}{e^{42479.26} + e^{42482.88}} \approx \frac{1}{1 + e^{3.62}} \approx 0.0262$$

Similarly, the estimated posterior probability that this note is *real* is $\widehat{p}\,(\pi_R \mid \boldsymbol{x}) \approx 0.9738$.

Note that, since $\widehat{s}_F^L(\boldsymbol{x})$ and $\widehat{s}_R^L(\boldsymbol{x})$ are so large, we employ the following factorisation:

$$\forall\, a \in \mathbb{R},\ b \in \mathbb{R}, \quad \frac{e^a}{e^a + e^b} = \frac{e^a}{e^a(1 + e^{b-a})} = \frac{1}{1 + e^{b-a}}.$$

Using Linear Discriminant Analysis (LDA), since the estimated posterior probability for the given note $\boldsymbol{x} = (x_1\ x_2,\ x_3) = (139.6,\ 8.0,\ 11.0)$ being *real* (97.38%) is significantly higher than the estimated probability for it being *fake* (2.62%), we assign this note $\boldsymbol{x}$ to the *real* group, deciding that it is a genuine Swiss banknote.

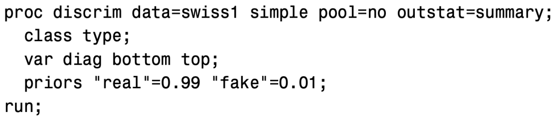
```
/* 1) Calculate the pooled cov matrix  */

proc iml;

*sample cov matrix for Fake group;

S1 = { 0.307883804 0.223675755 -0.026694255,
  0.223675755 1.197805907   -0.488101266,
  -0.026694255      -0.488101266       0.414959429};

n1 = 100; *100 fake notes;



*sample cov matrix for Real group;

S2 = { 0.1741944444      -.0158287037      -0.0516018519,
-0.0158287037      0.3884753086      -0.2866790123,
-0.0516018519      -0.2866790123      0.4465771605};

n2 = 100; *100 real notes;


print S1 , S2;

*sample pooled cov matrix;
Sp = (  (n1-1)*S1 + (n2-1)*S2) / ( (n1-1) + (n2-1) );

print Sp[format=8.4];

quit;



/* 2) Calculate the estimated posterior probability that the note x = ( diag =
    139.6, bottom = 8.0, top = 11.0) is fake */

proc iml;

d_F = { -42286, 606.66518,-45.75603  , 40.07711  };
x = { 1, 139.6, 8.0, 11.0};
```

```
sLF_x = d_F` * x;

print sLF_x[format= 8.4];

d_R = {-43639, 617.60341, -54.54254, 30.98006};

sLR_x = d_R` * x;

print sLR_x[format= 8.4];


*estimated posterior probability that the note is fake;


*phat_piF_x = exp(sLF_x) / ( exp(sLF_x) + exp(sLR_x) ) = exp(a) / [ exp(a) + exp
    (b) ] = 1 / (1 + exp(b-a)) where;

a = sLF_x;
b = sLR_x;


phat_piF_x = 1 / (1 + exp(b-a));


*estimated posterior probability that the note is real;

*phat_piR_x = exp(sLR_x) / ( exp(sLF_x) + exp(sLR_x) ) = exp(a) / [ exp(a) + exp
    (b) ] = 1 / (1 + exp(b-a)) where ;


a = sLR_x;
b = sLF_x;

phat_piR_x = 1 / (1 + exp(b-a));

print phat_piF_x[format= 8.4] phat_piR_x[format= 8.4] ;
```

Listing 2: Q2a LDA SAS code

b) Quadratic Discriminant Analysis (QDA)

Consider the following code for performing QDA on the independent variables *diag, bottom, and top*. The test/validation set attained a 100% hit ratio.



```
proc discrim data=swiss1 simple pool=no outstat=summary;
  class type;
  var diag bottom top;
  priors "real"=0.99 "fake"=0.01;
run;
```

Figure 9: QDA code

The primary assumption of QDA is that the covariance matrices of the *real* (R) and *fake* (F) note populations are statistically distinguishable (heterogeneous). Specifically, it is assumed that an observed sample unit $\boldsymbol{x} : p \times 1$ from population π_i follows a multivariate normal distribution, i.e., $(\boldsymbol{X} \mid \pi_i) \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, for $i = 1$ (F) and $i = 2$ (R).

A sample unit $\boldsymbol{x}$ is classified into population π_i (group i) that maximises the posterior probability:

$$p(\pi_i \mid \boldsymbol{x}) = \frac{e^{s_i^Q(\boldsymbol{x})}}{\sum_{j=1}^{2} e^{s_j^Q(\boldsymbol{x})}}, \quad s_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln p_i,$$

where $s_i^Q(\boldsymbol{x})$ is the quadratic score function for population π_i, with $i = 1$ (F) and $i = 2$ (R).

We aim to classify a banknote of unknown origin, with measurements $\boldsymbol{x} = (x_1\ x_2,\ x_3) = (139.6,\ 8.0,\ 11.0)$ (diag, bottom, and top measurements, respectively), as either *fake* or *real* by assigning it to group i that has the highest posterior probability $\hat{p}(\pi_i \mid \boldsymbol{x})$.

Given the following values (displayed in the report to four decimal places for brevity):

$$\mathbf{linear}_R \approx \begin{bmatrix} 934.5025 \\ 296.2182 \\ 320.8346 \end{bmatrix}, \quad \mathbf{quadratic}_R \approx \begin{bmatrix} -3.1947 & -0.7650 & -0.8602 \\ -0.7650 & -2.6289 & -1.7760 \\ -0.8602 & -1.7760 & -2.3591 \end{bmatrix}, \quad \mathbf{constant}_R \approx -68988.9704$$

and

$$\mathbf{linear}_F \approx \begin{bmatrix} 538.2175 \\ -128.0182 \\ -89.2253 \end{bmatrix}, \quad \mathbf{quadratic}_F \approx \begin{bmatrix} -2.0252 & 0.6244 & 0.6041 \\ 0.6244 & -0.9942 & -1.1293 \\ 0.6041 & -1.1293 & -2.4944 \end{bmatrix}, \quad \mathbf{constant}_F \approx -36358.8100$$

The estimated quadratic discriminant score functions for $i = R$ and $i = F$ are expressed as:

$$
\begin{aligned}
\widehat{s}_i^Q(\boldsymbol{x}) &= -\frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)'\,\boldsymbol{S}_i^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i) + \ln\widehat{p}_i \\
&= -\frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}\bar{\boldsymbol{x}}_i'\boldsymbol{S}_i^{-1}\bar{\boldsymbol{x}}_i - \frac{1}{2}\boldsymbol{x}'\boldsymbol{S}_i^{-1}\boldsymbol{x} + \bar{\boldsymbol{x}}_i'\boldsymbol{S}_i^{-1}\boldsymbol{x} + \ln\widehat{p}_i \\
&= (\bar{\boldsymbol{x}}_i'\boldsymbol{S}_i^{-1}\boldsymbol{x}) + (-\frac{1}{2}\boldsymbol{x}'\boldsymbol{S}_i^{-1}\boldsymbol{x}) + (-\frac{1}{2}\ln|\boldsymbol{S}_i| + \ln\widehat{p}_i - \frac{1}{2}\bar{\boldsymbol{x}}_i'\boldsymbol{S}_i^{-1}\bar{\boldsymbol{x}}_i) \\
&= \boldsymbol{x}'\cdot\textbf{linear}_i + \boldsymbol{x}'\cdot\textbf{quadratic}_i\cdot\boldsymbol{x} + \textbf{constant}_i.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\widehat{s}_R^Q(\boldsymbol{x}) &= (\bar{\boldsymbol{x}}_R'\boldsymbol{S}_R^{-1}\boldsymbol{x}) + (-\frac{1}{2}\boldsymbol{x}'\boldsymbol{S}_R^{-1}\boldsymbol{x}) + (-\frac{1}{2}\ln|\boldsymbol{S}_R| + \ln\widehat{p}_R - \frac{1}{2}\bar{\boldsymbol{x}}_R'\boldsymbol{S}_R^{-1}\bar{\boldsymbol{x}}_R) \\
&= (\boldsymbol{x}'\cdot\textbf{linear}_R) + (\boldsymbol{x}'\cdot\textbf{quadratic}_R\cdot\boldsymbol{x}) + (\textbf{constant}_R) \\
&\approx (136355.4781) + (-67375.5165) + (-68988.9704) \\
&\approx -9.0089
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{s}_F^Q(\boldsymbol{x}) &= (\bar{\boldsymbol{x}}_F'\boldsymbol{S}_F^{-1}\boldsymbol{x}) + (-\frac{1}{2}\boldsymbol{x}'\boldsymbol{S}_F^{-1}\boldsymbol{x}) + (-\frac{1}{2}\ln|\boldsymbol{S}_F| + \ln\widehat{p}_F - \frac{1}{2}\bar{\boldsymbol{x}}_F'\boldsymbol{S}_F^{-1}\bar{\boldsymbol{x}}_F) \\
&= (\boldsymbol{x}'\cdot\textbf{linear}_F) + (\boldsymbol{x}'\cdot\textbf{quadratic}_F\cdot\boldsymbol{x}) + (\textbf{constant}_F) \\
&\approx (73129.5461) + (-36781.7646) + (-36358.8100) \\
&\approx -11.0285
\end{aligned}
$$

Thus, the estimated posterior probability that this note is *fake* is

$$
\widehat{p}\,(\pi_F \mid \boldsymbol{x}) = \frac{e^{\widehat{s}_F^Q(\boldsymbol{x})}}{e^{\widehat{s}_F^Q(\boldsymbol{x})} + e^{\widehat{s}_R^Q(\boldsymbol{x})}} = \frac{1}{1 + e^{[\widehat{s}_R^Q(\boldsymbol{x}) - \widehat{s}_Q^L(\boldsymbol{x})]}}
$$

$$
\approx \frac{e^{-11.0285}}{e^{-11.0285} + e^{-9.0089}} \approx \frac{1}{1 + e^{2.0196}} \approx 0.1172
$$

Similarly, the estimated posterior probability that this note is *real* is $\widehat{p}\,(\pi_R \mid \boldsymbol{x}) \approx 0.8828$.

Using Quadratic Discriminant Analysis (QDA), since the estimated posterior probability for the given note $\boldsymbol{x} = (x_1, x_2, x_3) = (139.6, 8.0, 11.0)$ being *real* (88.28%) is significantly higher than the estimated probability of it being *fake* (11.72%), we assign this note $\boldsymbol{x}$ to the *real* group, concluding that it is a genuine Swiss banknote.

```
/* b) Quadratic Discriminant Analysis  */

proc iml;

x = { 139.6, 8.0, 11.0};
```

```
linear_R = {934.502525295218000, 296.218157161028000, 320.834571812728000};
linear_F = {538.217549887911000, -128.018223198110000, -89.225275429384800};



constant_R = -68988.970430357000000;
constant_F = -36358.810040127900000;

quadratic_R = {  -3.194692357743580 -0.764978796054153 -0.860221796945271,
-0.764978796054153 -2.628852119925690 -1.775977640616260,
 -0.860221796945271 -1.775977640616260 -2.359109795316430};

 quadratic_F = { -2.025208149262250 0.624367269724601 0.604138656279338,
 0.624367269724601 -0.994196289572104 -1.129270516008430,
 0.604138656279338 -1.129270516008430 -2.494391656570190};

 print linear_R [format = 12.4] linear_F [format = 12.4] ;
 print constant_R [format = 12.4] constant_F [format = 12.4] ;
 print quadratic_R [format = 12.4] quadratic_F [format = 12.4] ;

  linear_term = x‘ * linear_R;
  quadratic_term = x‘* quadratic_R * x;



  sQR_x = linear_term  + quadratic_term + constant_R ;

  print linear_term [format = 20.4] quadratic_term [format = 20.4]  sQR_x [
   format = 20.4];

  linear_term = x‘ * linear_F;
  quadratic_term = x‘* quadratic_F * x;

  sQF_x = linear_term  + quadratic_term + constant_F ;

   print linear_term [format = 20.4] quadratic_term [format = 20.4]  sQF_x [
   format = 20.4];



*estimated posterior probability that the note is fake;



*phat_piF_x = exp(sQF_x) / ( exp(sQF_x) + exp(sQR_x) ) = exp(a) / [ exp(a) + exp
```

```
    (b) ] = 1 / (1 + exp(b-a)) where;

a = sQF_x;
b = sQR_x;

diff1= b-a;
print diff1 [format= 8.4];

phat_piF_x = 1 / (1 + exp(b-a));

*estimated posterior probability that the note is real;

*phat_piR_x = exp(sQR_x) / ( exp(sQF_x) + exp(sQR_x) ) = exp(a) / [ exp(a) + exp
    (b) ] = 1 / (1 + exp(b-a)) where ;


a = sQR_x;
b = sQF_x;

phat_piR_x = 1 / (1 + exp(b-a));

diff2= b-a;
print diff2 [format= 8.4];

print phat_piF_x[format= 8.4] phat_piR_x[format= 8.4] ;

 quit;
```

Listing 3: Q2b QDA SAS code

In conclusion, Questions 2a) and 2b) highlight the important role of Bartlett's test in determining the homogeneity or heterogeneity of covariance matrices across population groups. This assessment directly informs the choice between Linear Discriminant Analysis (LDA), which assumes homogeneity, and Quadratic Discriminant Analysis (QDA), which accounts for heterogeneity. Bartlett's test ensures the selected method aligns with the data's structure. The observed discrepancy in the estimated posterior probabilities for the given note being fake (2.62% with LDA versus 11.72% with QDA) highlights the consequences of assuming covariance matrix homogeneity versus recognising heterogeneity. Despite identical priors (99% for the real group and 1% for the fake group), the results differ noticeably, illustrating the importance of aligning model assumptions with the data. While both methods classified the specific given note as *real*, this outcome might not hold for a different note, emphasising the need for robust model selection.

Bartlett’s test thus serves as a critical diagnostic tool, guiding the choice of a discriminant analysis method that ensures accurate, reliable, and data-appropriate classification outcomes.