Data provided: Graph Paper

MAS465

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2016–2017**

Multivariate Data Analysis **2 hours**

Marks will be awarded for your best ***three*** *answers.*

RESTRICTED OPEN BOOK EXAMINATION

Candidates may bring to the examination lecture notes and associated lecture material (but no textbooks) plus a calculator that conforms to University regulations.

There are 75 marks available on the paper.

Please leave this exam paper on your desk
Do not remove it from the hall

Registration number from U-Card (9 digits)
to be completed by student

Blank

1 (i) In a report from 1973, Weber reports on protein consumption in 25 European countries for nine food groups: Red meat, White meat, Eggs, Milk, Fish, Cereals, Starchy foods, Pulses and nuts, Fruits and vegetables. The R analysis of this data set follows on the next two pages.

(a) What is the correlation between red meat consumption and white meat consumption? ***(2 marks)***

(b) Why is the option `cor=TRUE` included included in the `princomp` command? What tends to happen if it is omitted? ***(2 marks)***

(c) The session has been edited so that only 6 components are listed. But how many would R have computed? ***(1 mark)***

(d) With the aid of an informal graphical technique, how many principal components would you retain in your study? Justify your answer. ***(3 marks)***

(e) Suggest characteristics of countries with high scores on the first principal component. ***(2 marks)***

(f) Albania has a very low consumption of fish. How might one expect to see that reflected in the PCA plots? Which country would you expect to have the highest consumption of fish? ***(3 marks)***

(g) Give an interpretation of the third and fourth principal components, and comment on the protein consumption of Finland with particular reference to these components. ***(4 marks)***

(ii) Find the principal components and the proportion of total variance explained by each when the covariance matrix is

$$S = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}.$$

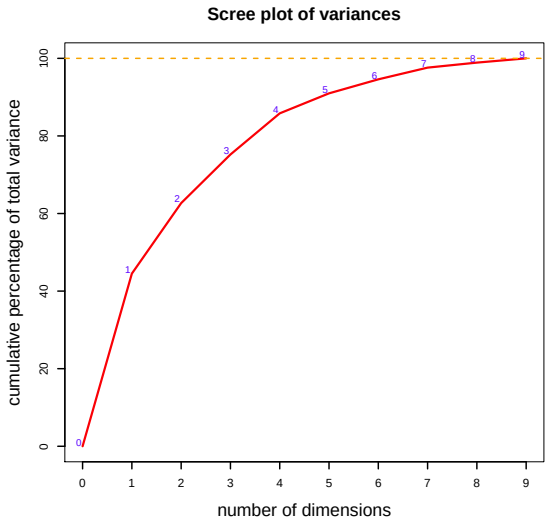
What is the possible range of ρ for S to be a variance matrix? ***(6 marks)***

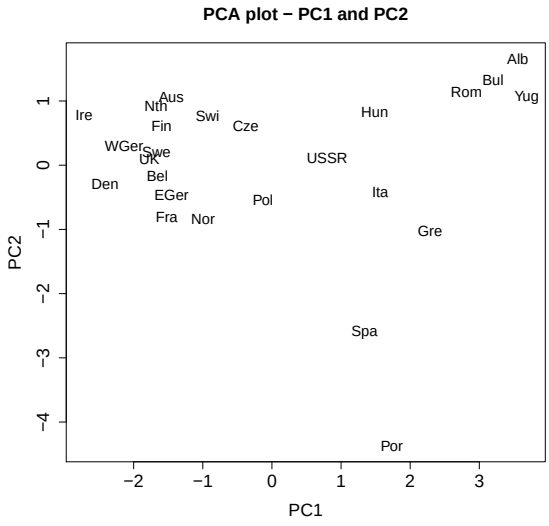
(iii) Explain briefly why multidimensional scaling might be regarded as a generalisation of principal components analysis. ***(2 marks)***

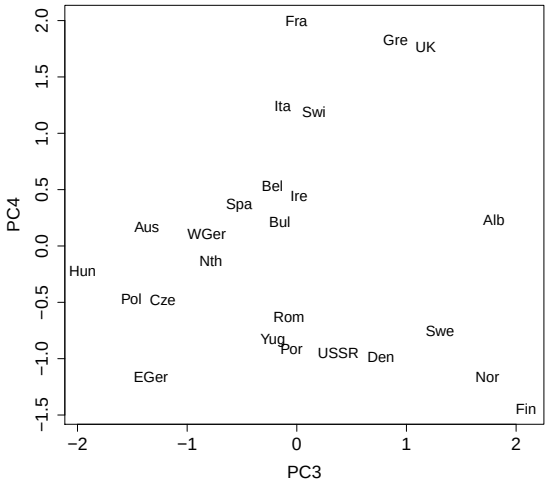
1 (continued)

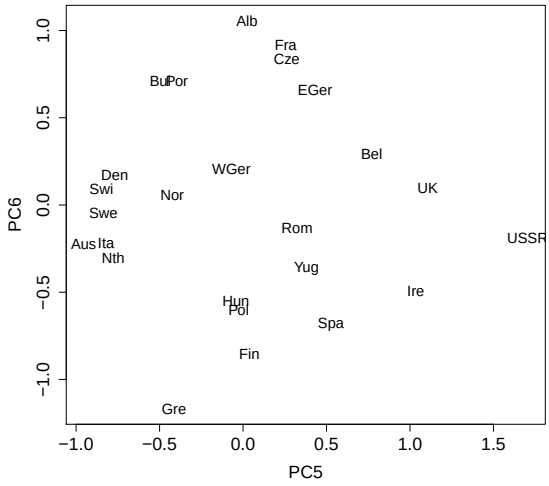
```
> attach(protein)
> protein[1:2,]
    RedMeat WhiteMeat Eggs Milk Fish Cereals Starch Nuts Fr.Veg
Alb    10.1       1.4  0.5  8.9  0.2      42    0.6  5.5    1.7
Aus     8.9      14.0  4.3 19.9  2.1      28    3.6  1.3    4.3
> apply(protein,2,mean)
  RedMeat WhiteMeat  Eggs  Milk  Fish  Cereals  Starch  Nuts  Fr.Veg
      9.8       7.9   2.9  17.1   4.3     32.2     4.3   3.1     4.1
> var(protein)
          RedMeat WhiteMeat   Eggs  Milk   Fish Cereals Starch  Nuts Fr.Veg
RedMeat     11.20      1.89  2.191  12.0   0.69  -18.36   0.74 -2.32 -0.448
WhiteMeat    1.89     13.65  2.561   7.4  -2.94  -16.78   1.89 -4.66 -0.409
Eggs         2.19      2.56  1.249   4.6   0.25   -8.74   0.83 -1.24 -0.092
Milk        11.96      7.39  4.570  50.5   3.33  -46.22   2.58 -8.76 -5.234
Fish         0.69     -2.94  0.249   3.3  11.58  -19.58   2.25 -0.99  1.634
Cereals    -18.36    -16.78 -8.738 -46.2 -19.58  120.45  -9.56 14.19  0.922
Starch       0.74      1.89  0.826   2.6   2.25   -9.56   2.67 -1.54  0.249
Nuts        -2.32     -4.66 -1.242  -8.8  -0.99   14.19  -1.54  3.94  1.343
Fr.Veg      -0.45     -0.41 -0.092  -5.2   1.63    0.92   0.25  1.34  3.254

> pr.pca<-princomp(protein,cor=TRUE)
> loadings(pr.pca)
Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
RedMeat   -0.303         0.298  0.646  0.322  0.460
WhiteMeat -0.311  0.237 -0.624        -0.300  0.121
Eggs      -0.427        -0.182  0.313        -0.361
Milk      -0.378  0.185  0.386        -0.200 -0.618
Fish      -0.136 -0.647  0.321 -0.216 -0.290  0.13D
Cereals    0.438  0.233                0.238
Starch    -0.297 -0.353 -0.243 -0.337  0.736 -0.148
Nuts       0.420 -0.143         0.330  0.151 -0.447
Fr.Veg     0.110 -0.536 -0.408  0.462 -0.234 -0.119

> summary(pr.pca)
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Standard deviation       2.00   1.28   1.06   0.98  0.681  0.570
Proportion of Variance   0.45   0.18   0.13   0.11  0.052  0.036
Cumulative Proportion    0.45   0.63   0.75   0.86  0.910  0.946
```

1 (continued)


Scree plot of variances
cumulative percentage of total variance
number of dimensions


PCA plot − PC1 and PC2
PC2
PC1
Alb
Bul
Rom
Yug
Aus
Nth
Ire
Swi
Cze
Fin
Hun
WGer
Swe
UK
USSR
Bel
Den
EGer
Pol
Ita
Fra
Nor
Gre
Spa
Por


PCA plot − PC3 and PC4
PC4
PC3
Fra
Gre
UK
Ita
Swi
Bel
Ire
Spa
Aus
Bul
Alb
WGer
Nth
Hun
Pol
Cze
Rom
Swe
Yug
Por
USSR
Den
EGer
Nor
Fin


PCA plot − PC5 and PC6
PC6
PC5
Alb
Fra
Cze
Bul
Por
EGer
Bel
WGer
Den
Swi
Nor
UK
Swe
Rom
USSR
Aus
Ita
Nth
Yug
Ire
Hun
Pol
Spa
Fin
Gre

2 Wolford and Hollingsworth (*Memory and Cognition*, 1974) were interested in the extent to which different consonants would be confused when a subject views them for a few milliseconds. The original aim of the study was to understand whether confusions occurred primarily because of auditory or visual short-term memory issues. We will consider a subset of 9 consonants from their list: D, F, H, M, N, Q, T, V and W, chosen because some pairs of these letters look similar (e.g., D and Q), and other pairs sounds similar (e.g., D and T).

The corresponding confusion matrix was formed, a submatrix of which is given below:

	D	F	H	M
D	–	5	6	4
F	7	–	6	2
H	4	4	–	11
M	3	2	19	–

An R analysis was performed, and some of the code and output is given on the following two pages.

(i) Give at least three reasons why the matrix given above is not suitable for use in the `cmdscale` command. What would you do to transform it into an appropriate form? ***(4 marks)***

(ii) After choosing one such method, techniques of multidimensional scaling and clustering were used to analyse the results. Give a reason why one should prefer non-metric to metric scaling in this example. ***(2 marks)***

(iii) Based on the list of eigenvalues, discuss how well a plot on the first two principal coordinates is likely to represent the data. ***(3 marks)***

(iv) Explain how you can see from the 2-dimensional plot on principal coordinates with the spanning tree that there is some distortion in the 2-dimensional plot. ***(3 marks)***

(v) Can we represent the data exactly in more dimensions? ***(2 marks)***

(vi) If we regard the pairs (D,T), (F,V), (H,W) and (M,N) as "sounding similar", can you find any (informal) evidence in the plots of principal coordinates that letters that *sound* similar are being confused? Is there any evidence in the cluster dendrogram? ***(4 marks)***

(vii) A k-means clustering was performed with various numbers of clusters. It was found that 5 clusters were needed so that the between clusters sum of squares accounted for 90.5% of the total sum of squares. Can you guess what the clusters were? ***(2 marks)***

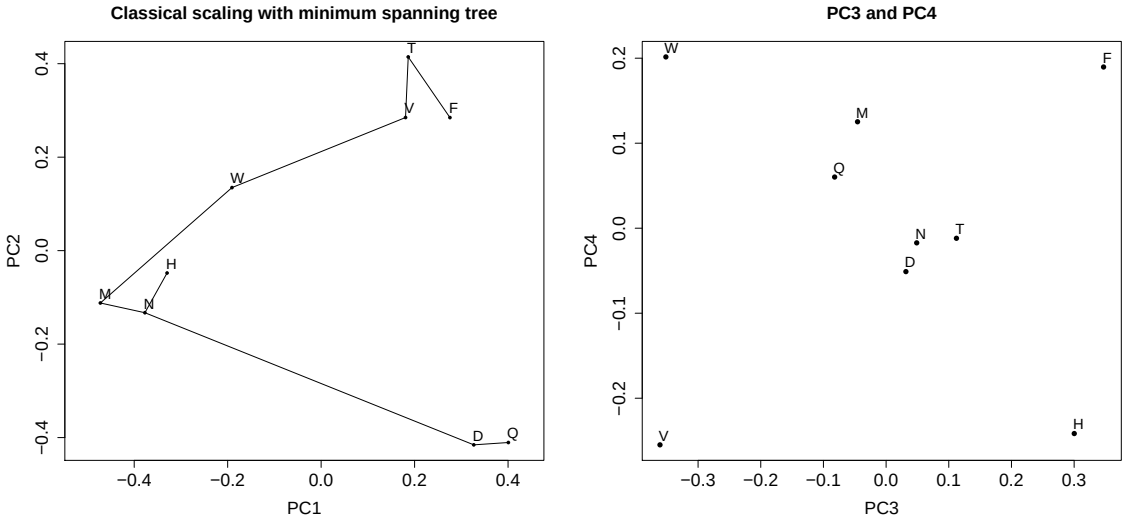
(viii) Suppose we have some method for converting the given confusion matrix into a dissimilarity matrix that starts:

```
    D   F   H  M
F  28
H  30  30
M  33  36  10
```

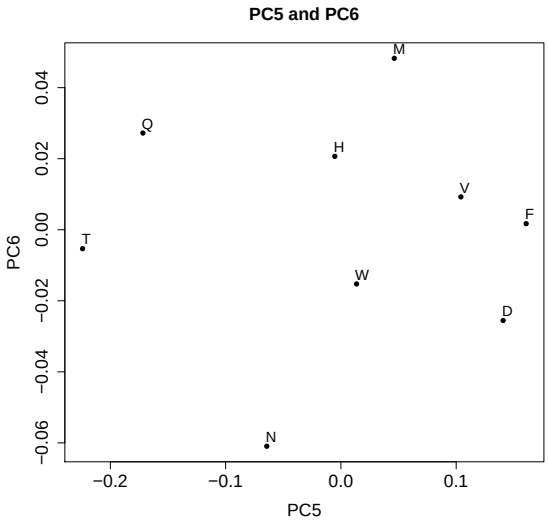
Work out the single linkage dendrogram for this subset of four letters. ***(5 marks)***

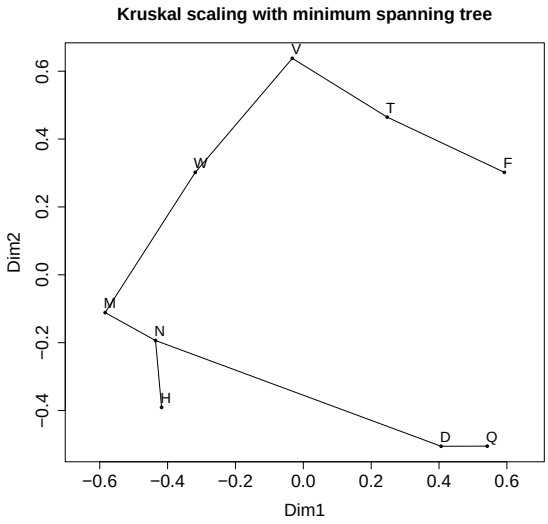
2 (continued)

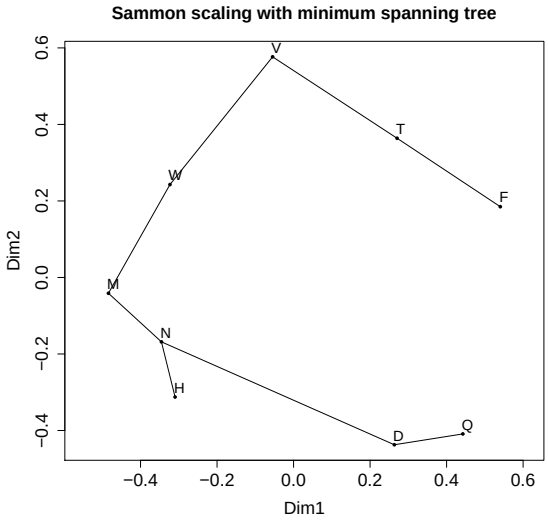
```
> wh.cmds<-cmdscale(wh,eig=TRUE)
> wh.cmds$eig
[1]  0.921  0.726  0.489  0.222  0.143  0.008  0.000  -0.009  -0.143
>
> wh.tr<-spantree(wh)
>
> plot(wh.tr,wh.cmds$points,pch=16,main="Classical scaling with minimum spanning tree")
> text(wh.cmds$points,labels=lets,adj=c(0.1,-0.4))
>
> plot(wh.cmds[,3],wh.cmds[,4],pch=16,xlab="PC3",ylab="PC4",main="PC3 and PC4")
> text(wh.cmds[,3],wh.cmds[,4],labels=lets,adj=c(0.1,-0.4))
>
> plot(wh.cmds[,5],wh.cmds[,6],pch=16,xlab="PC5",ylab="PC6",main="PC5 and PC6")
> text(wh.cmds[,5],wh.cmds[,6],labels=lets,adj=c(0.1,-0.4))
>
> wh.iso<-isoMDS(wh)
> plot(wh.tr,wh.iso$points,pch=16,main="Kruskal scaling with minimum spanning tree")
> text(wh.iso$points,labels=lets,adj=c(0.1,-0.4))
>
> wh.sam<-sammon(wh)
> wh.sam$stress
0.03133141
> plot(wh.tr,wh.sam$points,pch=16,main="Sammon scaling with minimum spanning tree")
> text(wh.sam$points,labels=lets,adj=c(0.1,-0.4))
>
> wh.cl<-hclust(as.dist(wh),method="average")
> plot(wh.cl)
```

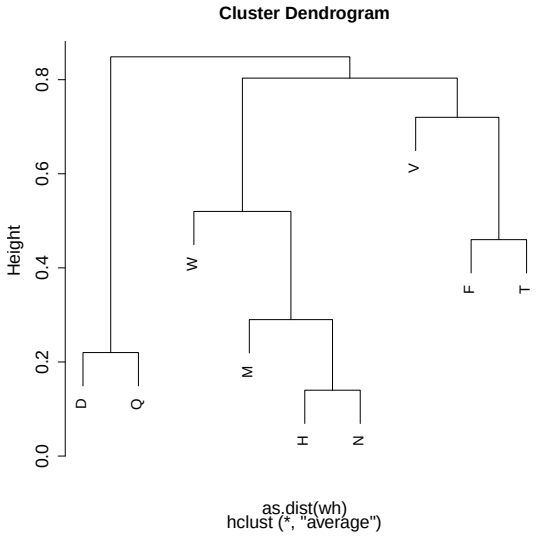


2 (continued)


PC5 and PC6
PC6
PC5
M
Q
H
V
F
T
W
D
N
−0.06
−0.04
−0.02
0.00
0.02
0.04
−0.2
−0.1
0.0
0.1


Kruskal scaling with minimum spanning tree
Dim2
Dim1
V
T
W
F
M
N
H
D
Q
−0.4
−0.2
0.0
0.2
0.4
0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Sammon scaling with minimum spanning tree
Dim2
Dim1
V
T
W
F
M
N
H
D
Q
−0.4
−0.2
0.0
0.2
0.4
0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Cluster Dendrogram
Height
as.dist(wh)
hclust (*, "average")
V
W
F
T
M
H
N
D
Q
0.0
0.2
0.4
0.6
0.8

3 Johnson and Wichern (2007) report on a study of 45 female hook-billed kites. The tail length and wing length are measured in centimetres for each bird. The mean tail length is 19.36cm, while the mean wing length is 27.98cm. The variance matrix is given by $S = \begin{pmatrix} 1.207 & 1.223 \\ 1.223 & 2.085 \end{pmatrix}$ (with the variable tail length appearing first, so that its variance is 1.207), with inverse $S^{-1} = \begin{pmatrix} 2.044 & -1.199 \\ -1.199 & 1.183 \end{pmatrix}$.

You may use the R output `qf(0.5,2,43)=3.214` and `qt(0.975,44)=2.015`.

(i) A more extended study of male kites has shown that the mean tail length for male birds is 19.75cm, and the mean wing length is 28.45cm.

(a) Compute a 95% confidence interval for the mean tail length of female kites. Can we reject the hypothesis that the mean tail length of female kites is 19.75cm at the 5% level of significance? ***(2 marks)***

(b) Similarly, can we reject the hypothesis that the mean wing length for female kites is 28.45cm at the 5% level of significance? ***(2 marks)***

(c) Test the multivariate hypothesis that the mean of the pairs of measurement $\overline{x} = (19.36, 27.98)$ for female kites is equal to $\mu_0 = (19.75, 28.45)$ at the 95% level. ***(6 marks)***

(d) Discuss briefly the results of parts (a)–(c). Illustrate your answer with a sketch of the multivariate confidence region. ***(4 marks)***

(ii) Suppose that $x_1, \ldots, x_n$ are independent observations of a bivariate normal distribution. We work with the principal components of the generated data, so that we suppose that the sample variance matrix is of the form $S = \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}$, with uncorrelated variables. Using these variables, we write $\overline{x}$ for the sample mean, and μ and Σ for the parameters of the distribution, so that our data comes from an $N_2(\mu, \Sigma)$ distribution.

Recall that the log-likelihood is given by

$$\ell(\mu, \Sigma) = -\tfrac{1}{2}(n-1)\mathrm{tr}(\Sigma^{-1}S) - \tfrac{1}{2}n\mathrm{tr}(\Sigma^{-1}(\overline{x}-\mu)(\overline{x}-\mu)') - n\log(2\pi) - \tfrac{1}{2}n\log|\Sigma|,$$

as $p = 2$, and you may assume that the unrestricted MLEs are given by $\hat{\mu} = \overline{x}$ and $\hat{\Sigma} = \frac{n-1}{n}S$.

(a) We wish to test the hypothesis that $\det\Sigma = \delta$, some fixed positive number. Under $H_0 : \det\Sigma = \delta$, we have $\hat{\mu} = \overline{x}$. Assuming also that $\hat{\Sigma}$ is diagonal, show that

$$\hat{\Sigma} = \sqrt{\frac{\delta}{ab}} \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}. \qquad \textit{(4 marks)}$$

(b) Hence give the likelihood ratio test for the null hypothesis that $\det\Sigma = \delta$, using Wilks's Theorem. ***(7 marks)***

4 Campbell and Mahon (1974) took various measurements of certain species of rock crab of genus *Leptograpsus*. Observations 1–50 were male, and observations 51–100 were female. Of interest in this question is whether we can predict the sex of the crab from the two dimension measurements, given in millimetres, as:

`FL` frontal lobe
`RW` rear width

Each line of the data set consists of 3 fields: firstly, the sex of the crab, and then the remaining 2 dimension measurements in the order above.

It seems that the dimensions of the two groups appear to be distributed as bivariate normal distributions with a similar variance matrix, which we take to be the same as the variance matrix in the R output which follows on the next page.

(i) Compute Fisher's linear discriminant function for this data set.

(5 marks)

(ii) Using the output from the previous part, classify the sex of a crab whose measurements are `FL` $= 13$ and `RW` $= 10$. ***(2 marks)***

(iii) By working out the probability of misclassification under the assumption that the two groups are distributed as bivariate normal distributions with the variance matrix in the R output, does Fisher's linear discriminant function classify the data set better or worse than expected? You are given that `pnorm(-1)=0.159`. ***(8 marks)***

(iv) Suppose that two bivariate normal distributions have means $\mu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ and $\mu_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and that both have variance matrix $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. Find the decision boundary (i.e., the set of points for which the probability densities agree) as a function of ρ.

(6 marks)

(v) Suppose that two bivariate normal distributions have means $\mu_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ and $\mu_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, and variances $\Sigma_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ respectively. Show that the decision boundary is given by a quadratic equation, which you should find.

(4 marks)

4 (continued)

```
> crab[1,]
  sex  FL  RW
1   M 8.1 6.7
> print(S<-var(crab[,-1]))
     FL   RW
FL 9.12 6.18
RW 6.18 5.20
> print(Sinv<-solve(S))
       FL     RW
FL  0.566 -0.673
RW -0.673  0.993
> print(crab.lda<-lda(sex~FL+RW))
Group means:
    FL   RW
F 13.3 12.1
M 14.8 11.7

Coefficients of linear discriminants:
     LD1
FL  1.21
RW -1.52
> crab.pred<-predict(crab.lda,data.frame(cbind(FL,RW)))
> crab.pred$class
  [1] M M M M M M F M M F F F M M M M M M M M M M M M
 [26] M M M M M M M M M M M M M M M M M M F M M M M M
 [51] F F F F M F F F F F F F F F F F F F F F F F F F
 [76] F F F F F F F F F F F F F F F F F F F F F F F F
Levels: F M
> table(sex,crab.pred$class)
sex  F  M
  F 49  1
  M  5 45
```

End of Question Paper