

The
University
Of
Sheffield.

**PLEASE LEAVE THIS EXAM PAPER ON YOUR DESK.
DO NOT REMOVE IT FROM THE HALL.**

**Data Provided:
Neaves Tables
Graph Paper**

SCHOOL OF MATHEMATICS AND STATISTICS

MAS465

Autumn Semester 2011-2012

2 Hours

Multivariate Data Analysis

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. All answers will be marked but credit will be given for only the best **THREE** answers. All questions carry equal marks. Total marks 75.*

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

--	--	--	--	--	--	--	--	--

(This page is left blank)

- 1 An archaeologist has presented the following data: the mean lengths (in cm) of tibiae and femurs (i.e. shin and thigh bones) of 27 Romano-British and 20 Saxon male skeletons were as follows:

	Tibia	Femur
Romano-British	40.25	49.83
Saxon	38.95	48.63

The pooled within groups sample covariance matrix (on 45 d.f.) was

$$\begin{pmatrix} 4.031 & 3.027 \\ 3.027 & 3.021 \end{pmatrix}.$$

- (a) Do the data provide evidence that Romano-British and Saxon skeletons differ in their mean tibia lengths?
(4 marks)
- (b) Do the data provide evidence that Romano-British and Saxon skeletons differ in their mean femur lengths?
(4 marks)
- (c) Do the data provide evidence that Romano-British and Saxon skeletons differ in their overall size of leg measurements (i.e. shin and thigh together)?
(9 marks)
- (d) Summarize carefully and coherently your overall conclusions from the three answers given to parts (a), (b) and (c).
(4 marks)
- (e) What linear combination of tibia and femur lengths provides the greatest evidence of a difference in overall leg size between the two types of skeleton? What is the statistical significance of the difference between the two types measured in this direction?
(4 marks)

- 2 A study was made of samples of Romano-British glass from two known production centres at Chester and Doncaster. Particular attention was given to the quantities of potassium (K) and lithium (Li) since this is known to depend upon the local materials and so can be distinctive of the production centre. The 31 samples from Chester gave mean quantities for K and Li of 8.4 and 19.5 respectively with sample variances 1.84 and 2.23 respectively and covariance 0.24 (measurements are in parts per thousand). The measurements of the 41 samples from Doncaster gave mean results of 9.6 and 18.3 with sample variances 1.64 and 2.00 and sample covariance 0.18.
- (a) Calculate Fisher's linear discriminant function for classifying a piece of glass with quantities (x_1, x_2) of K and Li as coming from Chester. **(8 marks)**
- (b) In order to simplify the calculations it is decided to use a rule of classifying a piece of glass as coming from Chester if the difference in content of lithium and potassium is greater than 8 parts per thousand. Stating any distributional assumptions you make, estimate the probability of misclassifying a randomly selected Chester sample as being produced in Doncaster when using this rule. **(5 marks)**
- (c) With the aid of a rough diagram and without necessarily doing any detailed calculations, is the probability calculated in part (b) more or less than that using Fisher's linear discriminant function? **(3 marks)**
- (d) Estimate the probability of misclassifying a randomly selected Doncaster sample as being produced in Chester when using the simplified rule in part (b). **(4 marks)**
- (e) A piece of glass found in Buxton has potassium and lithium contents of 9.0 and 18.9 parts per thousand respectively. How would you classify the piece using each of the two rules specified in parts (a) and (b) above? **(5 marks)**

3 As part of an investigation into determining possible locations of early hominids (i.e. ancient ape-like ancestors of humans) outside Africa, data were collated giving the numbers of fossil species of non-hominid fauna in various categories found at 139 different sites in Africa and Asia. It is thought that the profile of non-hominid types of fauna is related to whether or not hominids are attracted to the site. These sites included 21 sites (which are all African) where early hominid fossil remains have been found; no hominids have been found in the other 118 sites, some of which are in Asia and the others in Africa. The five categories recorded were Herbivores, Herbivore/Root-Eaters, Carnivores, Carnivore/Scavengers and Omnivores. Given below is a record (edited in places) of various preliminary analyses of these data using R. (NB: Herbivores eat vegetation only, Carnivores eat primarily meat and Omnivores eat both.)

(a) The principal component analysis has been performed using the correlation matrix. Would you recommend instead using the covariance matrix? Justify your recommendation.

(2 marks)

(b) Basing your judgement on some informal graphical technique (which should be given), how many principal components would you recommend retaining for further exploratory analyses?

(3 marks)

(c) What features of the sites do the most important principal components reflect?

(7 marks)

(d) What characteristics of the sites (in terms of the categories of fossil species found at them) seem to be typical of the majority of the hominid sites?

(7 marks)

(e) Two further sites in Asia are under consideration for further intensive excavation in the hope of identifying early hominid fossils but resources are only sufficient for a single expedition to one of the sites. The numbers (respectively) of Herbivores, Herbivore/Root-Eaters, Carnivores, Carnivore/Scavengers and Omnivores recorded at Site A are 20, 2, 0, 1 and 0. At Site B they were 9, 0, 4, 1 and 1. Upon which site would you recommend concentrating the available resources? Justify your answer.

(6 marks)

```
> hominids[1:4,]
  Herbivores Herb.Root Carnivores Carn.Scavangers Omnivores Hominid
1          43         10          4              4          0          0
2          21          2          4              2          1          0
3           5          0          0              0          0          0
4           6          1          1              1          1          0

> hominids[117:121,]
  Herbivores Herb.Root Carnivores Carn.Scavangers Omnivores Hominid
117          8          0          2              1          0          0
118         12          0          4              1          0          0
119          0          0          1              0          0          1
120          8          1          0              0          0          1
121         25          3          1              0          0          1
```

Question 3 continued on next page

Question 3 continued

*** Summary Statistics for data in: hominids ***

Hominid:1

	herbivores	herb.root	carnivores	carn.scavengers	omnivores
Mean:	4.43	0.364	1.25	0.364	0.195
Total N:	118.00	118.000	118.00	118.000	118.000
Std Dev.:	8.11	1.152	2.30	0.724	0.398

Hominid:2

	herbivores	herb.root	carnivores	carn.scavengers	omnivores
Mean:	18.8	3.00	2.71	0.952	0.0952
Total N:	21.0	21.00	21.00	21.000	21.0000
Std Dev.:	13.5	2.35	3.23	1.284	0.3008

```
> attach(hominids)
```

```
> hom.pc<-prcomp(hominids[-6],scale=T)
```

```
> options(digits=2)
```

```
> summary(hom.pc)
```

Importance of components:

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.78	1.074	0.5647	0.4977	0.361
Proportion of Variance	0.63	0.231	0.0638	0.0495	0.026
Cumulative Proportion	0.63	0.861	0.9245	0.9740	1.000

```
> options(digits=1)
```

```
> hom.pc$rotation
```

	PC1	PC2	PC3	PC4	PC5
Herbivores	-0.5	0.23	-0.36	-0.003	-0.74
Herb.Root	-0.5	0.42	-0.38	-0.225	0.64
Carnivores	-0.5	-0.31	0.08	0.794	0.19
Carn.Scavengers	-0.5	0.03	0.79	-0.364	-0.03
Omnivores	-0.2	-0.82	-0.31	-0.432	0.05

```
> par(mfrow=c(2,2))
```

```
> screeplot(hominids,cor=T)
```

```
> plot(hom.pc$x[,1:2],type='n')
```

```
> points(hom.pc$x[1:118,1:2],pch=1)
```

```
> points(hom.pc$x[119:139,1:2],pch=15)
```

```
> plot(hom.pc$x[,2:3],type='n')
```

```
> points(hom.pc$x[1:118,2:3],pch=1)
```

```
> points(hom.pc$x[119:139,2:3],pch=15)
```

```
> plot(hom.pc$x[,3:4],type='n')
```

```
> points(hom.pc$x[1:118,3:4],pch=1)
```

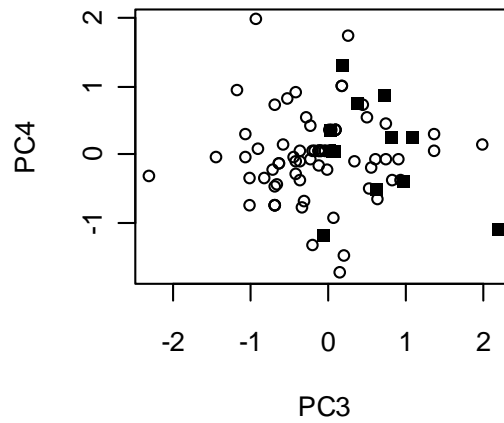
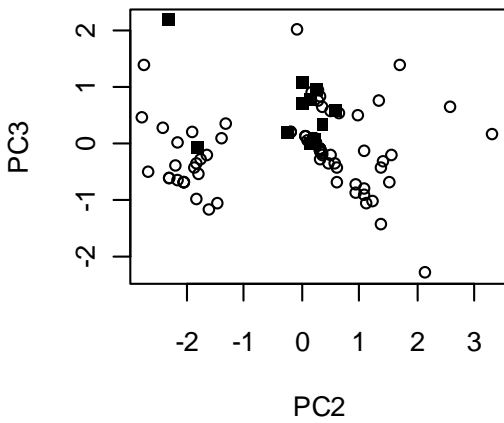
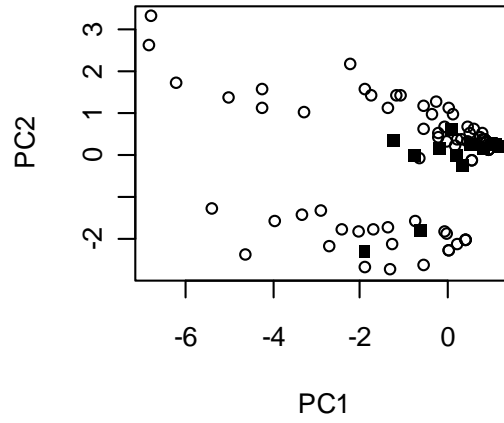
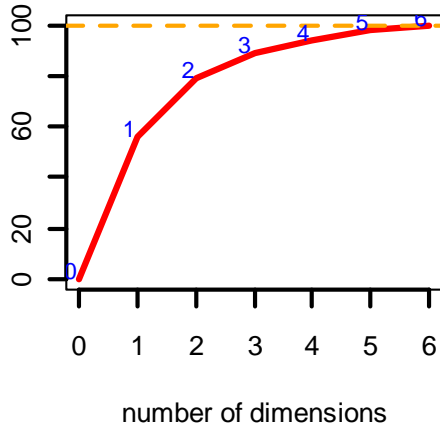
```
> points(hom.pc$x[119:139,3:4],pch=15)
```

Question 3 continued on next page

Question 3 continued

cumulative percentage of total variance

Scree plot of variances



v

- 4 Johnson & Wichern (2002) provide data giving the speeds in metres per second attained by men and women achieving the national record times (as they stood at the 1984 Los Angeles Olympics) in races over distances from 100 metres to 5000 metres in each of 55 different countries. These distances were, in metres, 100, 200, 400, 800, 1500 and either 3000 or 5000 (for women and men) and the marathon. A record of some multivariate analyses of these data is given below, where these various races are referred to as and m150m for the 1500 metres. The 3000m and 5000m are both coded as m5000m. Seven of the 55 countries are former 'Eastern block countries: China, Czechoslovakia, East Germany, Hungary, North Korea, Poland and Romania. Men and women are coded in the variable **gender** as 1 and 4 in the former Eastern block countries; 0 and 3 in the in the other 48 countries. Given below is a record (edited in places) of various preliminary analyses of these data using **R**.

- (a) What evidence is provided by the various analyses of data that the record times differ systematically between the four groups?
(6 marks)
- (b) Justifying your answer, what linear combination of the four measurements shows the greatest differences between the four groups?
(4 marks)
- (c) What is the statistical significance of the difference measured by this combination of variables?
(3 marks)
- (d) Overall, what is the major difference in record running speeds between men and women?
(3 marks)
- (e) In terms of running speeds what distinguishes between the record running speeds of women from the Eastern block and other countries?
(3 marks)
- (f) How strong is the evidence that there is a difference between record running speeds of women from the Eastern block and other countries?
(3 marks)
- (g) How strong is the evidence that there is a difference between record running speeds of men from the Eastern block and other countries?
(3 marks)

Question 4 continued on next page

Question 4 continued

```

> library(MASS)
> attach(arrowmin)

> olymprecs[53:58,]
  gender m100m m200m m400m m800m m1500m m5000m mmarathon country
53      0  10.1  10.1   9.1   7.7   7.1   6.3       5.5      usa
54      1   9.9  10.0   9.0   7.6   7.0   6.3       5.4      ussr
55      0   9.2   9.1   8.2   6.6   5.9   5.1       4.3    wsamoa
56      3   8.6   8.7   7.3   6.2   5.6   5.1       3.9  argentin
57      3   8.9   8.9   7.8   6.7   6.1   5.5       4.6  australi
58      3   8.7   8.7   7.9   6.7   5.9   5.4       4.4   austria

> ly.pc<-prcomp(olymprecs[,-c(1,9)],scale=T)
> options(digits=2)
> summary(oly.pc)
Importance of components:
                PC1      PC2      PC3      PC4      PC5      PC6      PC7
Standard deviation  2.553 0.5281 0.2982 0.23982 0.16059 0.1322 0.11370
Proportion of Variance 0.931 0.0398 0.0127 0.00822 0.00368 0.0025 0.00185
Cumulative Proportion 0.931 0.9710 0.9838 0.99197 0.99566 0.9981 1.00000

> oly.pc$rotation
      PC1      PC2      PC3      PC4      PC5      PC6      PC7
m100m  -0.37 -0.458 -0.382  0.267  0.254 -0.594 -0.123
m200m  -0.38 -0.417 -0.326  0.041 -0.071  0.698  0.287
m400m  -0.38 -0.272  0.351 -0.277 -0.651 -0.098 -0.384
m800m  -0.38 -0.043  0.566 -0.205  0.305 -0.175  0.604
m1500m -0.38  0.209  0.282  0.206  0.495  0.325 -0.579
m5000m -0.37  0.520 -0.049  0.603 -0.406 -0.085  0.231
mmarathon -0.37  0.476 -0.472 -0.634  0.063 -0.081 -0.029

> oly.lda<-lda(gender~as.matrix(olymprecs[,-
c(1,9)],prior=c(.25,.25,.25,.25)))
> oly.lda
Call:
lda(gender ~ as.matrix(olymprecs[,-c(1, 9)], prior = c(0.25,
0.25, 0.25, 0.25)))

Prior probabilities of groups:
  0      1      3      4
0.436 0.064 0.436 0.064
Group means
      m100m      m200m      m400m      m800m      m1500m      m5000m      marathon
0          9.5          9.5          8.6          7.4          6.8          6.0          5.1
1          9.7          9.7          8.8          7.5          6.9          6.2          5.3
3          8.6          8.4          7.4          6.4          5.8          5.3          4.1
4          8.8          8.8          8.0          6.9          6.2          5.7          4.4

```

Question 4 continued on next page

Question 4 continued

Coefficients of linear discriminants:

	LD1	LD2	LD3
m100m	0.902	2.08	-4.16
m200m	-0.294	2.47	2.88
m400m	-2.533	-5.64	-1.90
m800m	-3.248	0.18	4.49
m1500m	-2.827	2.18	1.91
m5000m	4.628	-4.80	-4.29
mmarathon	0.003	2.99	-0.39

Proportion of trace:

LD1	LD2	LD3
0.9563	0.0399	0.0038

># find predicted classifications**# without cross-validation/jackknifing**

```

> olydata<-data.frame(olymprecs[,-c(1,9)])
>
> oly.pred<-predict(oly.lda,olydata)
>
gender  0  1  3  4
      0 48  0  0  0
      1  7  0  0  0
      3  0  0 46  2
      4  0  0  1  6

```

find predicted classifications**# with cross-validation/jackknifing**

```

>
> oly.predCV<-predict(oly.lda,olydata,CV=T)
>
> table(gender,oly.predCV$class)

gender  0  1  3  4
      0 48  0  0  0
      1  7  0  0  0
      3  0  0 46  2
      4  0  0  1  6
>

```

Question 4 continued on next page

Question 4 continued

MANOVA : East vs west

```
> oly.manova<-manova(as.matrix(olymprecs[, -c(1,9)])~gender)
```

```
> summary(oly.manova)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
gender	1	0.758	45.7	7	102	<2e-16 ***

```
>
```

MANOVA on men: East vs west

```
> oly.men.manova<-manova(as.matrix(
+ olymprecs[1:55, -c(1,9)])~gender[1:55])
```

```
>
```

```
> summary(oly.men.manova)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
gender[1:55]	1	0.0721	0.522	7	47	0.81
Residuals	53					

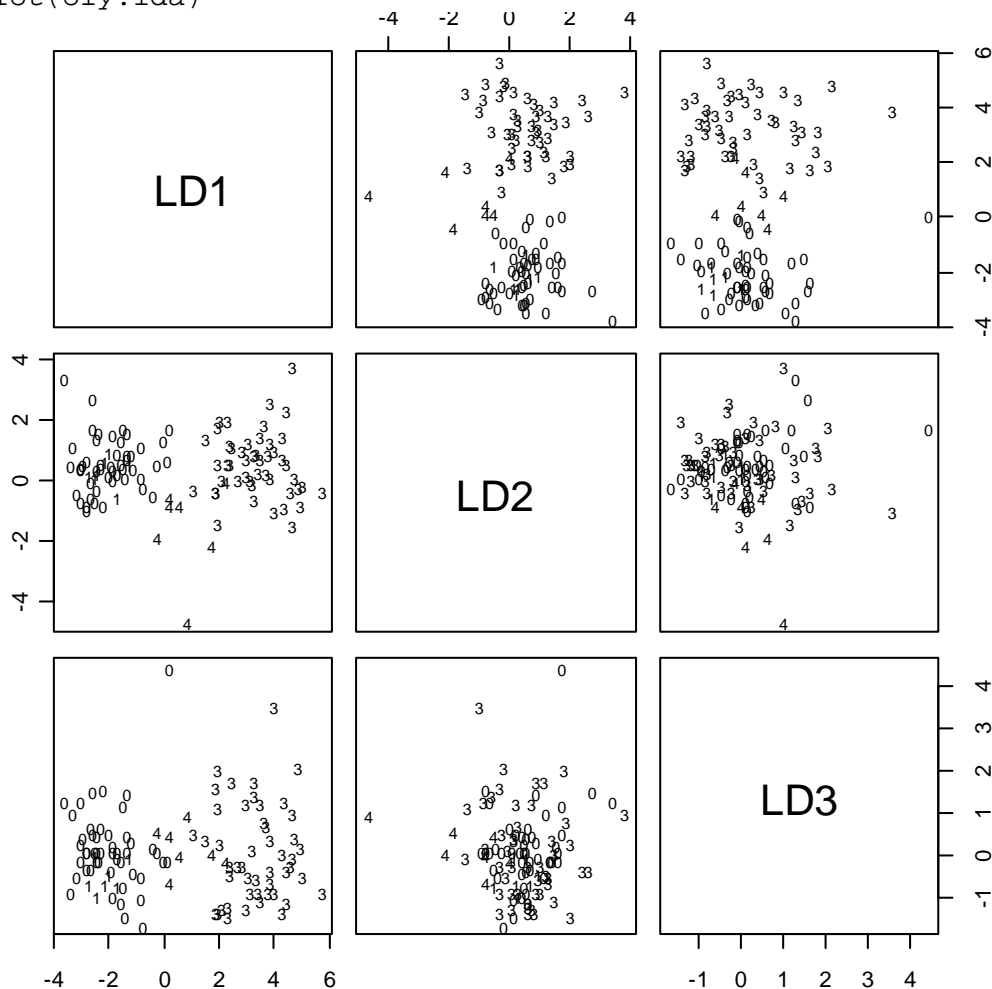
MANOVA on women: East vs west

```
> oly.women.manova<-manova(as.matrix
+ (olymprecs[56:110, -c(1,9)])~gender[56:110])
```

```
> summary(oly.women.manova)
```

	Df	Pillai	approx F	num Df	den Df	Pr(>F)
gender[56:110]	1	0.537	7.8	7	47	3e-06 ***

```
> plot(oly.lda)
```



End of Question Paper