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 2010-2011 **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 Measurements (in mmHg) of systolic and diastolic blood pressure were made on 22 subjects, 11 of whom were treated with a placebo and 11 with a beta-blocker drug. The means of the systolic and diastolic blood pressures in the placebo readings were 134.1 and 88.7, respectively, with variances 46.0 and 55.8. The covariance between systolic and diastolic readings in this group was 45.4. The corresponding five summary results in the beta-blocker treated group were 126.9, 84.1, 49.5, 60.6 and 46.2.

(a) Do these data provide evidence of an overall difference in mean blood pressure between the placebo and beta-blocker groups?

(9 marks)

(b) What linear combination of systolic and diastolic pressures exhibits the greatest difference between the groups?

(5 marks)

(c) After the analysis had been performed the clinician involved in the study congratulated the statistician on using the multivariate test performed in part (a), saying that often studies only looked at systolic and diastolic measures separately, but said that ideally she would like to see the conventional mean summary measure of (systolic + $2 \times$ diastolic) / 3 included as a third variable in the multivariate test. What response should you give to this request and what justification would you give?

(4 marks)

(d) The clinician also mentioned that there were in fact only 11 subjects in all, not 22. All 11 subjects had first been treated with the placebo and then later with the beta-blocker.

i) How would you wish to modify your analysis to take advantage of this extra information?

(4 marks)

ii) What further sample covariances would be required to enable you to perform the calculations for this further analysis?

(3 marks)

2 Collett (2003) describes the measurements of fibrinogen and gamma globulin in 32 subjects whose ESR (Erythrocytes Sedimentation Rate) had also been determined. Subjects with ESR below 20 are considered healthy. The mean values of fibrinogen and gamma globulin of the 26 subjects with ESR < 20 were 2.65 and 35.12 respectively. The corresponding values of those not classified as healthy by their ESR measurement were 3.39 and 38.00. The pooled within group variances of the two measurements were 0.330 and 20.360 with covariance –0.102. A key aim of the study was to see if it was possible to predict whether subjects could be classified as healthy without determining their ESR.

(a) Estimate Fisher's linear discriminant function for classifying a subject as healthy or not on the basis of measurements of fibrinogen and gamma globulin.

(9 marks)

(b) Assuming that these measurements are adequately modelled by bivariate Normal distributions with a common variance and that the classification of subjects uses Fisher's discriminant function, estimate the probability of misclassifying a randomly selected healthy subject as unhealthy.

(6 marks)

(c) Suppose, for a particular subject, the gamma globulin measurement is not available. What value of fibrinogen should be used as a lower limit to ensure that the probability of missing an unhealthy subject is the same as that using the rule determined in part (a)?

(6 marks)

(d) What proportion of healthy subjects will be falsely diagnosed as unhealthy by the rule in part (c)?

(4 marks)

3 Measurements were made on the dimensions of the wings of two sub-species of the *spiravolarens* butterfly, *atalanta* and *britannicus*. The first is an infrequent summer visitor to the south coast and the second is a native species. A scientific museum has provided measurements of the lengths and widths of left and right wings of a sample of 26 specimens of *s.atalanta* from its reference collection and an amateur entomologist has provided similar measurements of 21 specimens of *s.britannicus* from his personal collection. The length of the wing is taken as the maximum diameter parallel to the body and the width as the maximum diameter orthogonal to it. The initial aims of the analysis are to gain insight into the nature of the variation in sizes of the wings, investigate the differences between the wings of the two sub-species and ultimately the possibility of using wing measurements as an aid to identification. Given below is a record of an **R** session (edited in places) performing various preliminary analyses both separately on the measurements and on the two sets of measurements grouped together.

(a) Principal component analyses have been performed on the correlation rather than the covariance matrices. Why in all three cases is this choice likely to be more informative than the alternative?

(3 marks)

(b) What features of the wings in the three analyses do the four principal components derived from the correlation matrices reflect?

(9 marks)

(c) Scatterplots have been produced from the scores on the four principal components calculated from the correlation matrix of the combined data. What recommendation would you give regarding the advisability of investigating further the use of these wing measurements for distinguishing *s.atalanta* from *s.britannicus*?

(4 marks)

(d) The amateur entomologist (who is also an amateur statistician) suggests extending the Principal Component Analysis by feeding the scores on the four principal components referred to in part (c) back into the PCA routine of the statistical package in the optimistic hope that 'better results' will be produced. What advice should you give to the entomologist regarding this idea?

(3 marks)

(e) What will the eigenvalues be if the PCA in (d) is performed on

i) The correlation matrix of the principal component scores?

(3 marks)

ii) The covariance matrix of the principal component scores?

(3 marks)

Question 3 continued on next page

Question 3 continued

Analysis of measurements of *spiravolarens* butterflies

```
***  Summary Statistics for data in:  butterflies ***
     $$$"Factor Summaries":
        species
    atalanta:26
 britannicus:21

     $$$"Numeric Summaries":
          left.length left.width right.length right.width
    Mean:        7.02       9.04         7.31       10.67
 Total N:       47.00      47.00        47.00       47.00
Std Dev.:        1.29       3.78         1.49        3.48
------------------------------------------------

species:atalanta
     $$$"Factor Summaries":
        species
    atalanta:26
 britannicus: 0
     species: 0

 $$$"Numeric Summaries":
          left.length left.width right.length right.width
    Mean:        6.99      10.67         7.16       10.79
 Total N:       26.00      26.00        26.00       26.00
Std Dev.:        1.45       3.08         1.38        3.71

------------------------------------------------
species:britannicus
     $$$"Factor Summaries":
        species
    atalanta: 0
 britannicus:21
     species: 0

     $$$"Numeric Summaries":
          left.length left.width right.length right.width
    Mean:        7.07       7.01         7.51       10.51
 Total N:       21.00      21.00        21.00       21.00
Std Dev.:        1.11       3.65         1.64        3.26
```

Question 3 continued on next page

Question 3 continued

Principal Component Analyses for *s.atlanta*:

```
> pca.atlantacor<- princomp(butterflies[1:26,-1],cor=T)
> summary(pca.atlantacor)
Importance of components:
                        Comp.1 Comp.2 Comp.3  Comp.4
    Standard deviation   1.48  1.263 0.4333 0.12881
Proportion of Variance   0.55  0.399 0.0469 0.00415
 Cumulative Proportion   0.55  0.949 0.9959 1.00000
> print(loadings(pca.atlantacor),cutoff=0.1)
             Comp.1 Comp.2 Comp.3 Comp.4
 left.length  0.530 -0.471 -0.342  0.617
  left.width  0.483  0.510 -0.611 -0.365
right.length  0.536 -0.456  0.396 -0.590
 right.width  0.445  0.558  0.594  0.372
```

Principal Component Analyses for *s.Britannicus*:

```
>
> pca.britancor<- princomp(butterflies[27:47,-1],cor=T)
> summary(pca.britancor)
Importance of components:
                        Comp.1 Comp.2 Comp.3 Comp.4
    Standard deviation  1.251  1.069  0.887  0.710
Proportion of Variance  0.391  0.286  0.197  0.126
 Cumulative Proportion  0.391  0.677  0.874  1.000
> print(loadings(pca.britancor),cutoff=0.1)
             Comp.1 Comp.2 Comp.3 Comp.4
 left.length -0.417 -0.683  0.214  0.560
  left.width  0.636  0.251  0.306  0.663
right.length -0.439  0.429  0.782 -0.102
 right.width  0.478 -0.535  0.499 -0.486
>
```

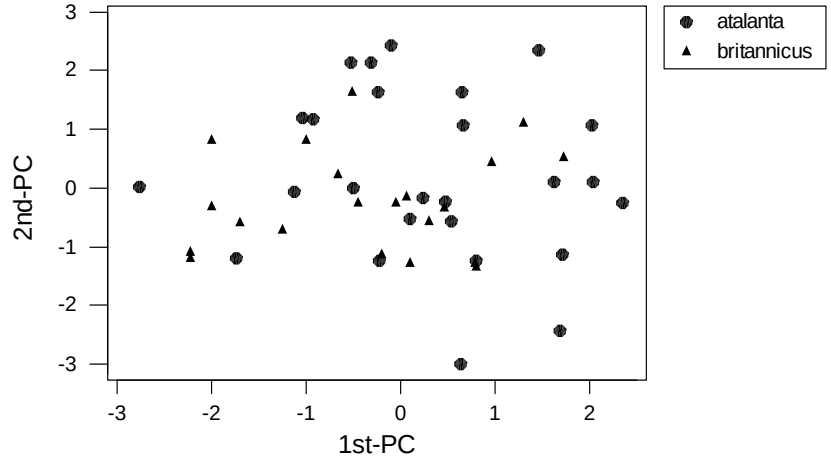
Principal Component Analyses for combined group:

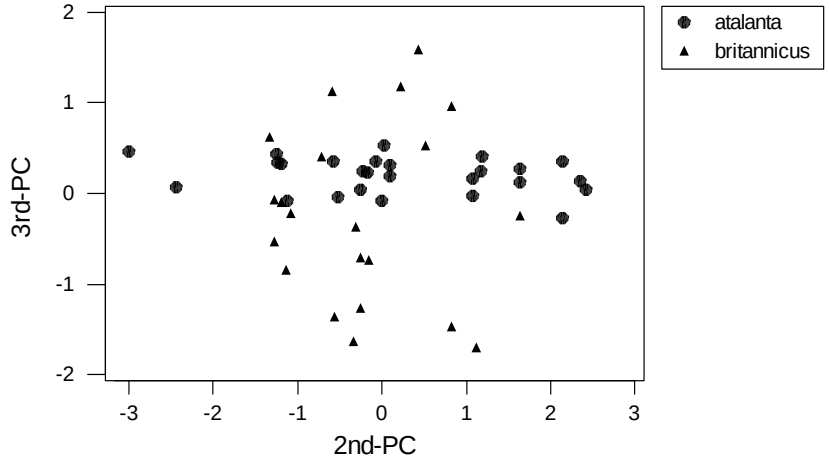
```
> pca.bothcor<- princomp(butterflies[,-1],cor=T)
> summary(pca.bothcor)
Importance of components:
                        Comp.1 Comp.2 Comp.3 Comp.4
    Standard deviation  1.261  1.226  0.700  0.646
Proportion of Variance  0.398  0.375  0.123  0.104
 Cumulative Proportion  0.398  0.773  0.896  1.000
> print(loadings(pca.bothcor),cutoff=0.1)
             Comp.1 Comp.2 Comp.3 Comp.4
 left.length -0.530 -0.472 -0.396 -0.583
  left.width  0.472 -0.523  0.584 -0.404
right.length -0.574 -0.407  0.469  0.534
 right.width  0.408 -0.582 -0.532  0.461
>
```

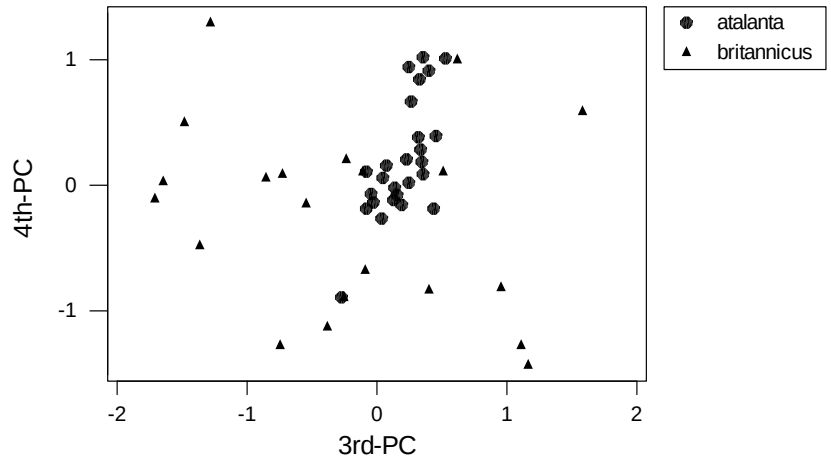
Question 3 continued on next page

Question 3 continued

Principal Component Score Plots for combined group:







4 Measurements (in tenths of a millimetre) were taken on large numbers of whelks (a type of shellfish) taken from two prehistoric middens (i.e. rubbish tips) and a modern beach adjacent to the middens in the island of Oronsay in the Hebrides in Scotland. Complete shells were selected so that all measurements could be taken accurately. The five measurements made were:

(1) *height* (i.e. overall length);
(2) *lengap1* (i.e. outer length of opening);
(3) *lengap2* (i.e. inner length of opening);
(4) *gapwidth* (i.e. width of opening);
(5) *lipthick* (i.e. thickness of outer lip).

Two new variables (*shape1* and *shape2*) were calculated to reflect the shape of the whelks. Other variables recorded were *site* and *period* (i.e. whether Prehistoric or Modern). The overall aim of the analysis was to investigate how the size and shape of the whelks varied with the different environments of the various beaches and middens. Various preliminary analyses using **R** were performed and the edited results are given in the next three pages**.**

(a) Linear discriminant analysis between the three sites has been performed using firstly the five measurements of the linear dimensions of the whelks, secondly using these with the addition of the two shape variables and thirdly the logarithms of the five dimension variables. A fourth analysis used just one of these sets to investigate the difference between modern and prehistoric. The analyses give summaries of the classifications using just the discriminant functions from the complete data. Those using cross-validation produced identical classifications in all cases. Noting that it is of interest to distinguish not only modern from prehistoric but also between the individual sites, which of these sets of predictors would you recommend, with justifications, for future use? In particular, is it better to use a rule based just on the modern/prehistoric classification if interest is in just this aspect?

(9 marks)

(b) Would it be worth investigating the use of the five logged variables together with the logged values of the two shape variables? (Justify your answer).

(6 marks)

(c) Noting the wide variation in the standard deviations of the five raw dimensions and the two derived shape variables, it is suggested that all the variables should be standardized first by subtracting their overall means and dividing by their standard deviations. What improvement, if any, will this provide for classifying whelk shells from this area (providing reasons for your answer)?

(5 marks)

(d) Summaries of the linear discriminant analyses indicate that the default choices of prior probabilities of class membership were taken. What effect on the analysis does this have and hence what you recommendation would you give regarding this choice in the context of the current study?

(5 marks)

Question 4 continued on next page

Question 4 continued

```
> library(MASS)
> whelks<-read.table("whelks.txt", header=T)
> whelks[1:5,]
     site  period height lengap1 lengap2 gapwidth lipthick
1 midden1 ancient   29.5    18.4    12.0      9.6      2.7
2 midden1 ancient   31.0    19.4    13.5      9.1      1.6
3 midden1 ancient   30.0    19.1    13.0     10.5      3.0
4 midden1 ancient   30.4    18.4    13.2     10.0      3.3
5 midden1 ancient   28.4    18.9    12.9     10.0      3.0
> attach(whelks)
> summary(whelks)
>
       site         period
   beach:78   prehistoric:129
 midden1:76         modern: 78
 midden2:53
        height lengap1 lengap2 gapwidth lipthick shape1 shape2
    Mean:  26.28 17.30 11.93  9.00   2.75 1.52  2.94
Std Dev.:   3.59  2.05  1.43  1.26   0.63 0.17  0.33

> whelksdat<-whelks[,-(1:2)]
> logwhelksdat<- log(whelksdat)
> shape1<-height/lengap1
> shape2<-height/gapwidth
> whelksmix<-as.data.frame(cbind
+(height,lengap1,lengap2,gapwidth,lipthick,shape1,shape2))
>
> whelk.lda<-lda(site~.,whelksdat)
> whelkmix.lda<-lda(site~.,whelksmix)
> logwhelk.lda<-lda(site~.,logwhelksdat)
>
> whelk.pred<-predict.lda(whelk.lda,whelksdat)
> whelkmix.pred<-predict.lda(whelkmix.lda,whelksmix)
> logwhelk.pred<-predict.lda(logwhelk.lda,logwhelksdat)
>
> whelk.predCV<-predict.lda(whelk.lda,whelksdat,CV=T)
> whelkmix.predCV<-predict.lda(whelkmix.lda,whelksmix,CV=T)
> logwhelk.predCV<-predict.lda(logwhelk.lda,logwhelksdat,CV=T)
>
```

Question 4 continued on next page

Question 4 continued

```
> table(site,whelk.pred$class)
        beach midden1 midden2
  beach    68       4       6
midden1     5      64       7
midden2    16       8      29
> table(site,whelkmix.pred$class)
        beach midden1 midden2
  beach    68       4       6
midden1     3      68       5
midden2    11       6      36
> table(site,logwhelk.pred$class)
        beach midden1 midden2
  beach    66       6       6
midden1     4      65       7
midden2    14       8      31
> whelkmixper.lda<-lda(period~.,whelksmix)
> whelkmixper.pred<-predict.lda(whelkmixper.lda,whelksmix)
> table(period,whelkmixper.pred$class)
        prehistoric modern
prehistoric     113     16
     modern      16     62
>> table(site,whelkmixper.pred$class)
        prehistoric modern
  beach      16     62
midden1      74      2
midden2      39     14

> whelk.lda
Prior probabilities of groups:
 beach midden1 midden2
 0.377   0.367   0.256
Coefficients of linear discriminants:
            LD1    LD2
  height  0.453 -0.193
 lengap1 -0.189  0.296
 lengap2 -0.142 -1.104
gapwidth  0.424  1.376
lipthick -1.693  0.202
Proportion of trace:
   LD1   LD2
 0.818 0.182
```

Question 4 continued on next page

Question 4 continued

```
> whelkmix.lda
Prior probabilities of groups:
 beach midden1 midden2
 0.377   0.367   0.256

Coefficients of linear discriminants:
            LD1      LD2
  height  1.083  -1.2878
 lengap1 -0.266   1.6287
 lengap2 -0.162  -1.1893
gapwidth -1.218   2.1667
lipthick -1.685   0.0688
  shape1 -0.672  12.7396
  shape2 -5.350   2.7516

Proportion of trace:
   LD1   LD2
 0.809 0.191

> logwhelk.lda
Prior probabilities of groups:
 beach midden1 midden2
 0.377   0.367   0.256

Coefficients of linear discriminants:
            LD1     LD2
  height  10.59  -5.301
 lengap1  -2.47   4.386
 lengap2  -3.73 -12.946
gapwidth   5.35  12.240
lipthick  -4.08   0.951

Proportion of trace:
   LD1   LD2
 0.816 0.184

> whelkmixper.lda
Prior probabilities of groups:
 prehistoric modern
   0.623  0.377

Coefficients of linear discriminants:
            LD1
  height -0.160
 lengap1 -0.662
 lengap2  0.763
gapwidth -0.205
lipthick  1.282
  shape1 -6.283
  shape2  2.717
>
```

End of Question Paper