SCHOOL OF MATHEMATICS AND STATISTICS  

Multivariate Data Analysis  

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
As part of an investigation into determining possible locations of diamond deposits in Australia, data were collated giving the numbers of geographical micro-deposits in various categories found at 90 different sites. These sites included 11 sites (numbered 80 to 90) where diamonds have been found; no diamonds have been found in the other 79 sites (numbered 1 to 79). The five categories recorded were Igneous (ign), Igneous/Calcific (ign.calc), Sedimentary (sed), Metamorphic Sedimentary (meta.sed) and Amorphous (amo).

Given below is an edited record of various preliminary analyses of these data using R.

(i) The principal component analysis has been performed using the correlation matrix. Would you recommend instead using the variance matrix? Justify your recommendation.  
   (2 marks)

(ii) With the aid of an informal graphical technique, how many principal components would you recommend retaining for further exploratory analyses?  
   (3 marks)

(iii) What features of the sites do the three most important principal components reflect?  
   (4 marks)

(iv) What characteristics of the sites (in terms of the categories of deposits found at them) seem to be typical of the majority of the diamond sites? Explain your answers.  
   (5 marks)

(v) Two additional sites are under consideration for further intensive excavation in the hope of identifying diamond deposits, but resources are only sufficient for a single expedition to one of the sites. The numbers (respectively) of Igneous, Igneous/Calcific, Sedimentary, Metamorphic Sedimentary and Amorphous recorded at Site A are 6, 0, 4, 1 and 1. At Site B, they were 7, 3, 0, 1 and 0. Upon which site would you recommend concentrating the available resources?  
   (5 marks)

(vi) A colleague notices that the analysis uses the function princomp, and believes that prcomp is meant to have certain advantages numerically. Looking up the help page, he spots that prcomp uses the formula $S = \frac{1}{n-1}(X - \overline{X})(X - \overline{X})'$ for the variance matrix, whereas princomp uses $S = \frac{1}{n}(X - \overline{X})(X - \overline{X})'$. What differences, if any, would this make to the R analysis below? And would it have any effect on your answer to part (ii)? Justify your answers.  
   (4 marks)

(vii) After projecting the data onto the principal components, suppose that each principal component is scaled to have standard deviation equal to 1. What is the variance matrix of the resulting set? Justify your answer.  
   (2 marks)

```r
> attach(diamonds)
> library(MASS)
> apply(diamonds[1:79,-6],2,mean)
   ign ign.calc sed meta.sed amo
5.443  0.4684  1.3544  0.40506  0.18987
> apply(diamonds[1:79,-6],2,sdev)
   ign ign.calc sed meta.sed amo
9.316  1.3759  2.4075  0.75987  0.39471
```
> apply(diamonds[80:90,-6],2,mean)
  ign ign.calc sed meta.sed amo
  20.182 3.7273 3.4545 1.1818 0.09091
> apply(diamonds[80:90,-6],2,sdev)
  ign ign.calc sed meta.sed amo
  14.586 2.6867 3.5879 1.4709 0.30151

> dia.pca<-princomp(diamonds[-6],cor=T)
> summary(dia.pca)

Importance of components:

            Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Standard deviation 1.79353 1.04910 0.536479 0.521141 0.351077
Proportion of Variance 0.64335 0.22012 0.057562 0.054318 0.024651
Cumulative Proportion 0.64335 0.86347 0.921031 0.975349 1.000000

> loadings(dia.pca)

             Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
ign           0.522 -0.123  0.329  0.776
ign.calc      0.471 -0.387  0.520  0.598
sed           0.473  0.293 -0.357  0.734 -0.155
meta.sed      0.490 -0.137 -0.633 -0.581
amo           0.205  0.855  0.304 -0.349 -0.114

> par(mfrow=c(1,2))
> plot(dia.pca[,1:2],type='n')
> points(dia.pca[1:79,1:2],pch=1)
> points(dia.pca[80:90,1:2],pch=15)
> plot(dia.pca[,2:3],type='n')
> points(dia.pca[1:79,2:3],pch=1)
> points(dia.pca[80:90,2:3],pch=15)
In a study of English dialects, 8 villages in the East Midlands (a region to the south of Sheffield) were compared to see whether they used the same word for 60 everyday items. Two villages were selected from each of Lincolnshire (Lin1 and Lin2), Nottinghamshire (Not1 and Not2), Leicestershire (Lei1 and Lei2) and Northamptonshire (Nth1 and Nth2). Additionally, one village from each of Cambridgeshire (Cam) and Bedfordshire (Bed) were added to the group. The measure of similarity between two villages is the percentage of items for which the same word is used. An R analysis using the similarity matrix was performed with a view to producing a graphical representation of the 10 dialects. Some of the results are given below, followed by a map, and plots of some of the principal coordinates against each other. The final pair of plots involve superimposing the minimum spanning tree onto the first plot, and then onto the result of using non-metric scaling.

(i) With the aid of an informal graphical technique, how many dimensions would you recommend to provide an adequate representation of the data? (6 marks)

(ii) What interpretations can you give of the plots? (9 marks)

(iii) All of the eigenvalues in this analysis are positive. If one or more of them had been negative, what modifications would you make to the analysis and interpretation of scatterplots? (3 marks)

(iv) Comment on the differences between the results in the final pair of plots. Which two villages should be furthest apart in the Kruskal scaling? Which two villages should be next furthest apart? Comment on whether this happens in the final plot, and if not, suggest reasons. (7 marks)

> dialects
  Lin1 Lin2 Not1 Not2 Lei1 Lei2 Nth1 Nth2 Bed Cam
Lin1 100 71 63 63 41 25 22 22 29 16
Lin2 71 100 64 66 36 25 24 20 32 26
Not1 63 64 100 71 42 32 32 27 31 30
Not2 63 66 71 100 50 39 36 36 44 33
Lei1 41 36 42 50 100 64 38 45 45 47
Lei2 25 25 32 39 64 100 51 54 53 49
Nth1 22 24 32 36 38 51 100 63 60 42
Nth2 22 20 27 36 45 54 63 100 61 44
Bed 29 32 31 44 45 53 60 61 100 54
Cam 16 26 30 33 47 49 42 44 54 100
> dia<-as.dist(100-dialects)
> diasmall.cmds<-cmdscale(dia)
> dia.cmds<-cmdscale(dia,eig=TRUE,k=9)
> dia.cmds$eig
  [1] 8.577e+03 2.396e+03 1.886e+03 1.084e+03 7.104e+02 5.515e+02 3.638e+02
 [8] 2.576e+02 3.997e+00 2.969e-13
> diasmall.cmds
  [,1]  [,2]
Lin1 -42.007 4.038
Lin2 -39.647 2.940
Not1 -32.038 1.331
Not2 -24.504 2.618
Lei1 5.091 -27.338
Lei2 24.870 -16.659
Nth1 27.531 24.830
Nth2 31.733 15.117
Bed 22.904 11.874
Cam 26.067 -18.751

5 Question 2 continued on next page
Classical scaling – minimum spanning tree

Non-metric (Kruskal) scaling – minimum spanning tree
Measurements were taken on a sample of children in a European town on their second birthdays. The overall sizes of the children were assessed by two measurements, the height and the chest circumference. In total, the sample consisted of 31 boys and 25 girls. The mean lengths obtained are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Height (cm)</th>
<th>Chest (cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>83.26</td>
<td>59.55</td>
</tr>
<tr>
<td>Girls</td>
<td>80.79</td>
<td>58.28</td>
</tr>
</tbody>
</table>

The variance matrix for the group of 31 boys is \( S_B = \begin{pmatrix} 25.72 & 12.09 \\ 12.09 & 8.36 \end{pmatrix} \) (so the variance of the height is 25.72 and the variance for the chest is 8.36), while the variance matrix for the group of 25 girls is \( S_G = \begin{pmatrix} 22.32 & 11.91 \\ 11.91 & 8.65 \end{pmatrix} \).

(i) Calculate the pooled within groups sample variance matrix (on 54 d.f.).

(ii) Do the data provide evidence that the boys are taller than the girls?

(iii) Do the data provide evidence that the boys have larger chest measurements than the girls?

(iv) Test the hypothesis that the height and chest measurements of the group of boys is the same as that of the girls. Compare your answers with parts (ii) and (iii), and summarise your conclusions.

(v) The experiment was partly conducted to compare the results with an earlier large study on a group of Australian boys on their second birthdays. The variance in the Australian study was found to be \( \begin{pmatrix} 25.89 & 13.01 \\ 13.01 & 10.21 \end{pmatrix} \). Use a likelihood-ratio test to test the hypothesis that the variance of the European boys is the same as that found in the Australian study. You may assume any standard results on MLEs, and may also assume that the sample size is sufficiently large that Wilks’s Theorem applies.
Johnson and Wichern (2002) report on a study into potential haemophilia A carriers, consisting of a group of 30 subjects without the haemophilia gene (the non-carrier group), and a group of 22 subjects who were known haemophilia carriers (the carrier group). Measurements were made of two variables; $X_1$ is related to antihaemophiliac factor activity, and $X_2$ to antihaemophilic-like antigens. (Since the quantities involved were recorded on a logarithmic scale, some of the entries are negative.)

The investigators provided information

$$\bar{x}_N = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix}, \quad \bar{x}_C = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix},$$

and

$$S^{-1} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix},$$

where $\bar{x}_N$ and $\bar{x}_C$ denote the sample mean for readings of $X_1$ and $X_2$ for the noncarrier group and carrier group respectively, and $S$ is the pooled sample variance matrix.

(i) Estimate Fisher’s linear discriminant function for classifying a subject as in the carrier group or not on the basis of the measurements of $X_1$ and $X_2$. (8 marks)

(ii) Informal investigations suggest that the data for each group is reasonably well approximated by a bivariate normal distribution, and, further, that the variance matrices for both groups appear to be very similar, so that they may be assumed to be the same. Using your function from part (i) to classify observations, estimate the probability that a randomly selected noncarrier is misclassified as a carrier. (6 marks)

(iii) The cost of measuring the variable $X_2$ is high, and it is hoped to develop a test using only the value of $X_1$. What value should be used as a lower limit to ensure that the probability of missing a carrier is the same as that using the rule determined in part (i)? (7 marks)

(iv) What proportion of non-carriers will be falsely diagnosed as carriers by the rule in part (iii)? (4 marks)

End of Question Paper