Data provided: Graph Paper

MAS465

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2015–2016**

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 Parts (i)–(vi) of this question refer to a study reported by Sokal and Rohlf (1981) of air pollution in 41 US cities, between 1969 and 1971. The variables are as follows:

`SO2` Sulphur dioxide content of air in μg per cubic metre

`Temp` Average annual temperature in degrees Fahrenheit

`Firms` Number of manufacturers employing 20 or more workers

`Pop` Population in thousands in 1970

`Wind` Average annual wind speed in miles per hour

`Rain` Average annual rainfall in inches

`Raindays` Average number of rainy days per year

A principal components analysis was carried out on the data, and an R transcript is given below.

(i) Explain why `cor=TRUE` is used in the `princomp` command. ***(1 mark)***

(ii) In the principal components analysis, it is decided to use only the first few components. Using an informal graphical technique, how many components would you choose? ***(3 marks)***

(iii) Phoenix is in the top left of the plot of the first two principal components. Explain why Phoenix gets such a low score on the first principal component.

There is an outlier on PC1, with a very high score. Can you suggest some characteristics of this city? ***(7 marks)***

(iv) Explain why Phoenix seems to get such a high score for the second principal component. ***(3 marks)***

(v) Interpret the third principal component, justifying your answer briefly. ***(2 marks)***

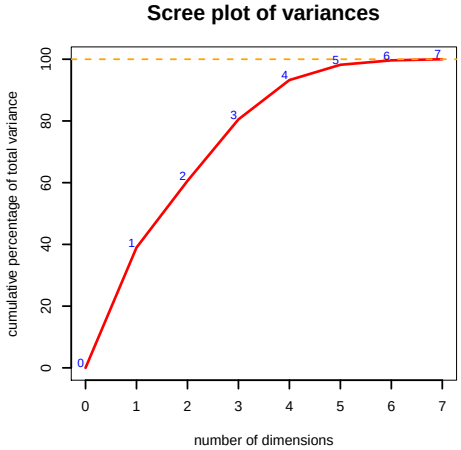
(vi) Comment on the fourth and fifth principal components. Is there anything in the plots of these principal components that strikes you? ***(3 marks)***

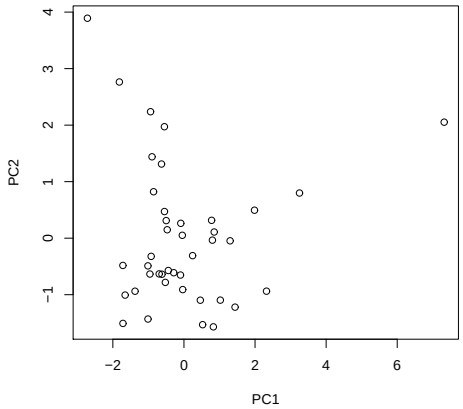
(vii) Let M be a $p \times p$ real matrix. Quoting any results you wish from the course, sketch how you would find the maximum value of $\dfrac{a'Ma}{a'a}$ as a varies over all vectors.

Verify your answer if $M = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. ***(6 marks)***

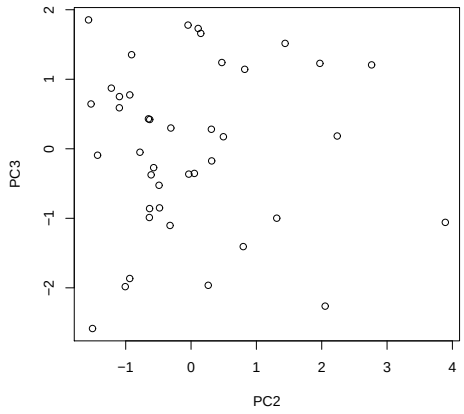
1 (continued)

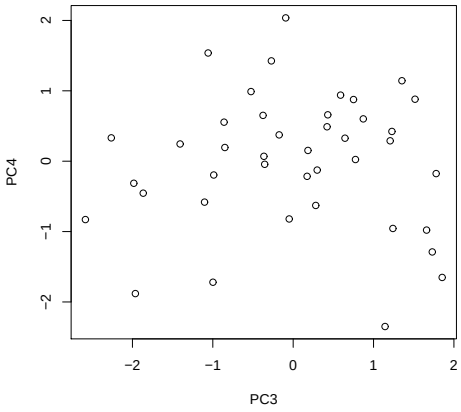
```
> attach(airpoll)
> airpoll[1:5,]
              SO2 Temp Firms Pop Wind  Rain Raindays
Phoenix        10 70.3   213 582  6.0  7.05       36
Little Rock    13 61.0    91 132  8.2 48.52      100
San Fransisco  12 56.7   453 716  8.7 20.66       67
Denver         17 51.9   454 515  9.0 12.95       86
Harford        56 49.1   412 158  9.0 43.37      127

> apply(airpoll,2,mean)
     SO2     Temp    Firms      Pop     Wind     Rain Raindays
   30.05    55.76   463.10   608.61     9.44    36.77   113.90
> airpoll.pca<-princomp(airpoll,cor=TRUE)
> summary(airpoll.pca)
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6  Comp.7
Standard deviation       1.65  1.230  1.181  0.944 0.5889 0.3167 0.15973
Proportion of Variance   0.39  0.216  0.199  0.127 0.0495 0.0143 0.00364
Cumulative Proportion    0.39  0.606  0.805  0.932 0.9820 0.9964 1.00000
> print(airpoll.pca$loadings,digits=2)

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
SO2       0.49                 0.40  -0.73  -0.18   0.15
Temp     -0.32         -0.68  -0.19  -0.16  -0.61
Firms     0.54   0.23  -0.27          0.16         -0.75
Pop       0.49   0.28  -0.34  -0.11   0.35          0.65
Wind      0.25          0.31  -0.86  -0.27  -0.15
Rain            -0.63  -0.49  -0.18  -0.16   0.55
Raindays  0.26  -0.68   0.11   0.11   0.44  -0.50
> screeplot(airpoll,cor=TRUE)
> airpoll.pc<-predict(airpoll.pca)
> airpoll.pc[1:5,]
               Comp.1  Comp.2  Comp.3  Comp.4  Comp.5    Comp.6   Comp.7
Phoenix       -2.7160  3.8914 -1.0584  1.5374 -0.0331 -0.654428  0.09061
Little Rock   -1.7177 -0.4824 -0.8500  0.1945 -0.1421  0.544992 -0.19867
San Fransisco -0.9390  2.2372  0.1837  0.1528  0.1856  0.295074  0.00714
Denver        -0.5499  1.9719  1.2286  0.4227  0.3837 -0.130106 -0.15799
Harford        0.4604 -1.0973  0.5897  0.9387 -0.7427  0.400483 -0.26019
```

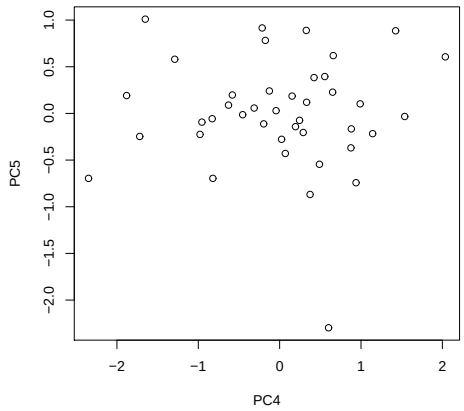




1 (continued)


PC3
PC2
−2
−1
0
1
2
−1
0
1
2
3
4


PC4
PC3
−2
−1
0
1
2
−2
−1
0
1
2


PC5
PC4
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
−2
−1
0
1
2

2 Corbet et al (1970) recorded the presence or absence of 13 characteristics in about 300 water vole skulls divided into samples from 14 populations from Britain and the rest of Europe. A similarity matrix was constructed from this data.

```
       1     2     3     4     5     6     7     8     9    10    11    12    13
2  0.099
3  0.033 0.022
4  0.183 0.114 0.042
5  0.148 0.224 0.059 0.068
6  0.198 0.039 0.053 0.085 0.051
7  0.462 0.266 0.322 0.435 0.268 0.025
8  0.628 0.442 0.444 0.406 0.240 0.129 0.014
9  0.113 0.070 0.046 0.047 0.034 0.002 0.106 0.129
10 0.173 0.119 0.162 0.331 0.177 0.039 0.089 0.237 0.071
11 0.434 0.419 0.339 0.505 0.469 0.390 0.315 0.349 0.151 0.430
12 0.762 0.633 0.781 0.700 0.758 0.625 0.469 0.618 0.440 0.538 0.607
13 0.530 0.389 0.482 0.579 0.597 0.498 0.374 0.562 0.247 0.383 0.387 0.084
14 0.586 0.435 0.550 0.530 0.552 0.509 0.369 0.471 0.234 0.346 0.456 0.090 0.038
```

The British populations are numbered 1–6, and the remaining populations are taken from 8 European sites. The non-British populations are known to come from two species, *Arvicola terrestris*, and *Arvicola sapidus*; the wide geographical separation of the European sites suggest that each of these consists of either one species or the other, but not both. The authors' main aim was to test the hypothesis that both species were present in Britain.

An R analysis using the similarity matrix was performed with a view to producing a graphical representation of the 14 populations. Some of the results are given below, followed plots of some of the principal coordinates against each other. The fourth plot superimposes a minimum spanning tree onto the first; the fifth gives a plot based on ordinal scaling, while the final plot gives a dendrogram coming from a cluster analysis.

(i) Comment on the list of eigenvalues, and what it implies about the plots below. With the aid of an informal graphical technique, how many dimensions would you recommend to provide an adequate representation of the data? ***(7 marks)***

(ii) The first plot gives the default output from classical multidimensional scaling. Without referring to the original data matrix, give two ways in which the remaining plots indicate that there is some distortion in the first plot, justifying your answers. ***(6 marks)***

(iii) Is there any reason to conclude that the British populations may indeed be comprised of both European species? ***(3 marks)***

(iv) Is there any reason to suspect that the British populations might, in fact, be comprised of species different from the two European species? ***(3 marks)***

(v) If there were no distortion in the Kruskal plot, which would be closest pair of populations be? ***(2 marks)***

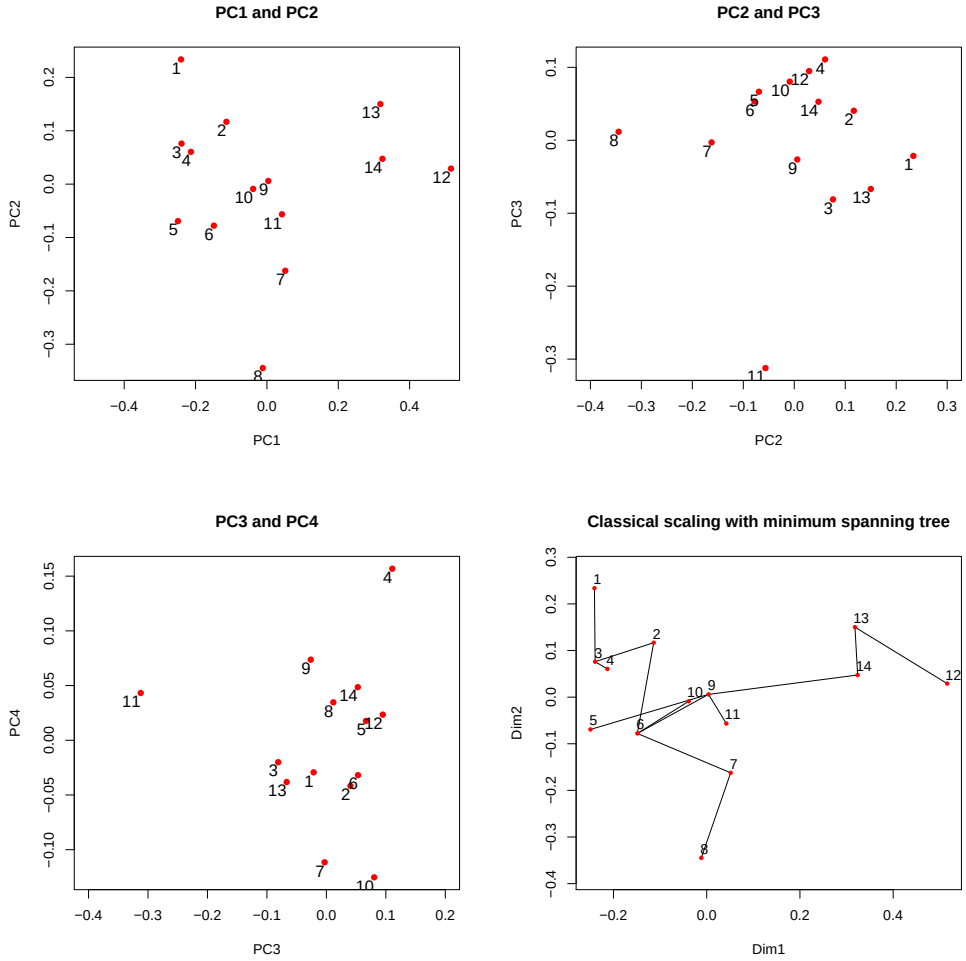
2 (continued)

(vi) A k-means clustering analysis was also performed, with 3 clusters. The output consisted of one cluster with 8 populations, and two with 3 populations each. Referring to the multidimensional scaling plots as well as to the hierarchical cluster plot, identify the clusters.

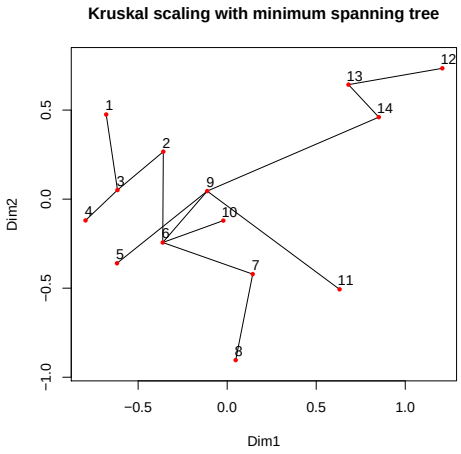
The `kmeans(voles,3)` command reports that these three clusters account for 76.9% of the total sum of squares. What does this mean?

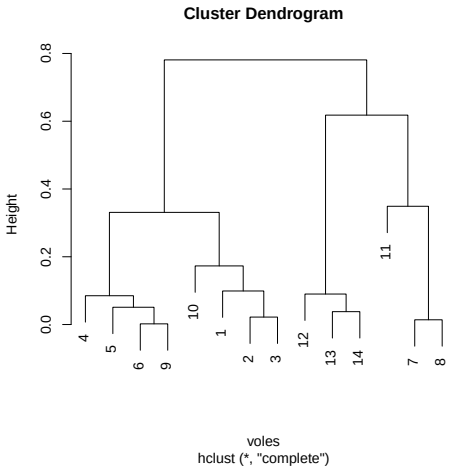
(4 marks)

```
> voles.cmds<-cmdscale(voles,eig=TRUE)
> round(100*voles.cmds$eig,2)
 [1]  73.60  26.26  14.93   6.99   2.96   1.93   0.00  -1.14  -1.28  -2.85
[11]  -4.25  -5.26  -7.41 -10.98
```



2 (continued)


Kruskal scaling with minimum spanning tree
Dim2
Dim1
−1.0
−0.5
0.0
0.5
−0.5
0.0
0.5
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14


Cluster Dendrogram
Height
0.0
0.2
0.4
0.6
0.8
4
5
6
9
10
1
2
3
12
13
14
11
7
8
voles
hclust (*, "complete")

3 (i) Measurements of x_1 (stiffness) and x_2 (bending strength) for a new sample of $n_A = 20$ pieces of the highest grade of timber from a site A were taken. The means were given by $\overline{x}_A = (\overline{x}_1, \overline{x}_2) = (17.6, 81.8)$, while the variance matrix of the sample was $S = \begin{pmatrix} 2.1 & 2.2 \\ 2.2 & 7.4 \end{pmatrix}$, with inverse $S^{-1} = \begin{pmatrix} 0.7 & -0.2 \\ -0.2 & 0.2 \end{pmatrix}$.

The company wishes to compare the new sample with its established population, where the mean stiffness is $\overline{x} = (17.1, 83.0)$.

For the first two parts, R gives `qt(0.975,19)=2.093`.

(a) Perform a t-test to test the hypothesis that the mean stiffness in the new sample is equal to 17.1. ***(2 marks)***

(b) Perform a t-test to test the hypothesis that the mean bending strength in the new sample is equal to 83.0. ***(2 marks)***

(c) Test the hypothesis that $\overline{x}_A = \overline{x}$. You may use the R output `qf(0.95,2,18)=3.555`. ***(6 marks)***

(d) Discuss briefly the results of parts (a)–(c), bearing in mind that $\overline{x}_A$ seems to be not too far from $\overline{x}$, illustrating your answer with a sketch of the multivariate confidence region. (You may like to compute the correlation between the two variables to help with your answer.) ***(3 marks)***

(ii) Suppose that $x_1, \ldots, x_n$ are independent observations of a bivariate normal distribution $x \sim N_2(\mu, \Sigma)$, where neither μ nor Σ are known. Write $\overline{x}$ for the sample mean, and $S = \begin{pmatrix} a & c \\ c & b \end{pmatrix}$ for the sample variance.

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) Suppose that $\Sigma = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$ is diagonal, and $\mu = \overline{x}$. By explicitly computing $\Sigma^{-1}S$, show that

$$\ell(\mu, \Sigma) = -\tfrac{1}{2}(n-1)(\alpha^{-1}a + \beta^{-1}b) - n\log(2\pi) - \tfrac{1}{2}n\log(\alpha\beta).$$

By computing $\dfrac{\partial \ell}{\partial \alpha}$, deduce that under the constraint that Σ is diagonal, its MLE is

$$\hat{\Sigma} = \frac{n-1}{n}\mathrm{diag}(S) = \frac{n-1}{n}\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}. \qquad \textit{(5 marks)}$$

(b) Develop a likelihood-ratio test to test the null hypothesis that Σ is diagonal, using Wilks's Theorem. ***(7 marks)***

4 A study was performed in Malaysia (1985) relating to the rate of settling of red blood cells out of suspension in blood plasma. This rate increases if the level of certain proteins in the blood plasma rise, and two plasma proteins, whose levels are denoted `F` and `G` in what follows, were considered. Subjects were classed as healthy or unhealthy depending on the value of this rate of settling.

The mean values of `F` and `G` of 26 healthy subjects were 4.66 and 61.81 respectively. The corresponding values of 6 subjects classified as not healthy were 5.97 and 66.88. The pooled within group variances of the two measurements were 1.022 and 63.067, with covariance -0.316.

A key aim of the study was to determine the strength of any relationship between the levels of the two plasma proteins and the rate of settling.

(i) Estimate Fisher's linear discriminant function for classifying a subject as healthy or not on the basis of measurements of the levels of plasma proteins `F` and `G`. ***(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 answer to part (i) to classify subjects as healthy or unhealthy, estimate the probability of misclassifying a randomly selected healthy subject as unhealthy. You may leave your answer in the form $\Phi(z)$ for some number z which you should determine. ***(5 marks)***

(iii) Suppose that for a particular subject, the `G` sample is contaminated, and an assessment is made of the subject purely based on the level of the plasma protein `F`. What value of `F` 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 (i)? ***(6 marks)***

(iv) What proportion of healthy subjects will be falsely diagnosed as unhealthy by the rule in (iii)? (Again, you may give your answer in a form involving $\Phi(z)$ for some z.) ***(3 marks)***

(v) The classification of the patient as unhealthy or healthy is purely based on whether or not the rate of settling lies above or below a certain threshold. What would be a better tool to see how this is related to the levels of `F` and `G`? Give the R command for this. ***(3 marks)***

End of Question Paper