Data provided: Graph Paper

MAS465

The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2014–2015**

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 The heptathlon is an athletics event for women consisting of seven disciplines: two throwing events (the shot put and the javelin), two jumping events (the high jump and the long jump), all measured in metres, and three running events (the 200m, the 800m and the 100m hurdles), measured in seconds. Points are awarded for every discipline, with more points awarded for faster times and longer distances, and summed to give an overall score.

Data were collected from the Olympic Games in Seoul in 1988, where the competitors are listed in order of how they finished, with the winner being given first.

Below is part of an R transcript on the results; parts (i)–(vii) of this question refer to this data.

(i) In the command `princomp(heptathlon[,-8],cor=TRUE)`, discuss why

(a) `heptathlon[,-8]` is used, and not `heptathlon` ***(2 marks)***

(b) `cor=TRUE` is necessary. ***(1 mark)***

(ii) The data set gives the results of the events prior to converting them to points. If the data for each discipline had been presented after converting to points, how might your answers to (i)(b) change? What would you do to decide whether to include `cor=TRUE`? ***(2 marks)***

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

(iv) Interpret the first principal component, justifying your answer briefly. What does the competition winner score on this component? ***(3 marks)***

(v) Interpret the second principal component. The data set omits the competitors who did not complete the event, but, of those who finished, the competitor who came last overall did very well at one event. Which event was this? ***(3 marks)***

(vi) Interpret the third principal component. A Swiss athlete appears to be an outlier on PC3. How might her performance have been? ***(3 marks)***

(vii) A Chinese athlete did rather well at the high jump, but was a slow 800m runner. Where might one expect to see her on the last graph? ***(2 marks)***

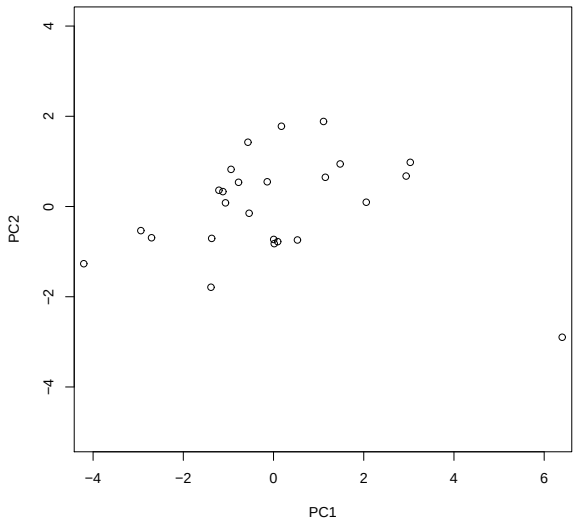
(viii) Suppose that two correlation matrices S_1 and S_2 satisfy $S_1 = tS_2 + (1-t)I_p$. Explain that the *ratios* between the correlations of pairs of distinct variables are the same for S_1 and S_2. Show that Principal Components Analysis (using the correlation matrix) only depends on these ratios in the sense that the principal components for the matrix S_1 are the same as those for S_2. ***(3 marks)***

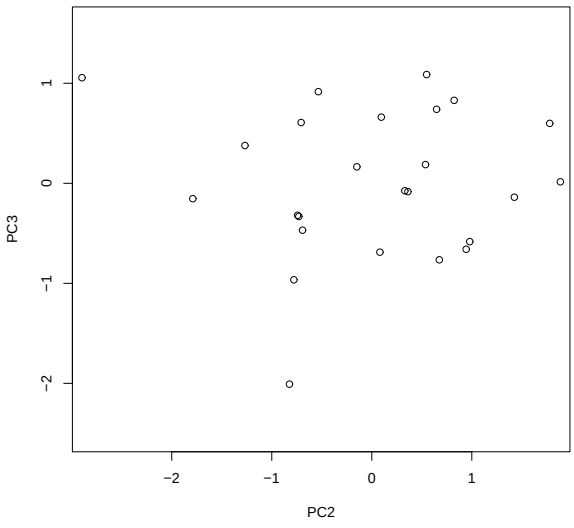
(ix) If PCA is applied to a $p \times p$ correlation matrix R with $R_{ij} = \rho$ for all $i \neq j$ (as is often the case in biological situations), then the first principal component is (proportional to) $(1, \ldots, 1)'$. What proportion of the total variance does this explain? ***(3 marks)***

```
> attach(heptathlon)
```

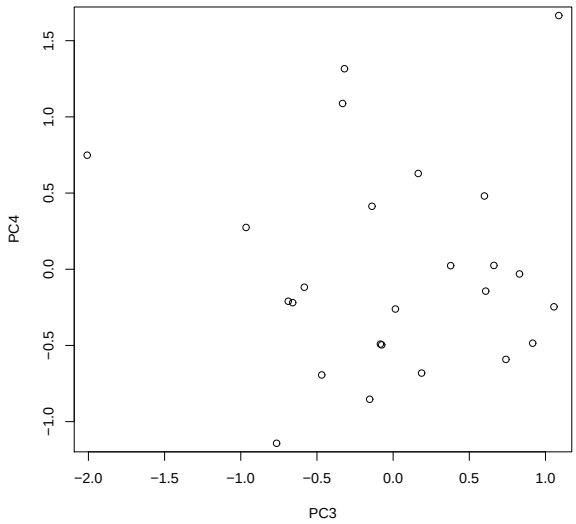
1 (continued)

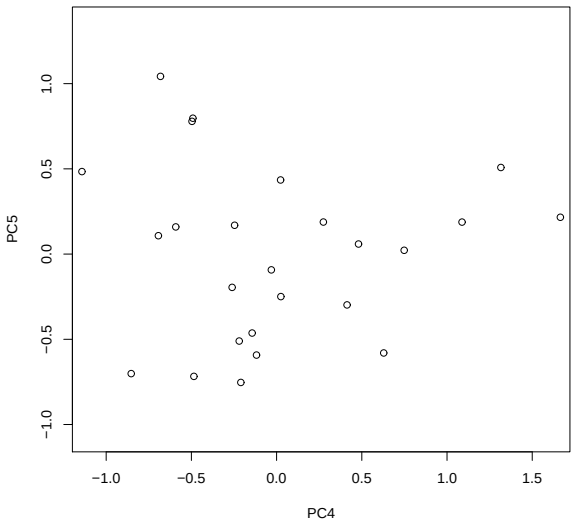
```
> heptathlon[1:3,]
             hurdles highjump  shot run200m longjump javelin run800m score
Joyner-Kersee   12.69     1.86 15.80   22.56     7.27   45.66  128.51  7291
John            12.85     1.80 16.23   23.65     6.71   42.56  126.12  6897
Behmer          13.20     1.83 14.20   23.10     6.68   44.54  124.20  6858

> hep.pca<-princomp(heptathlon[,-8],cor=TRUE)
> summary(hep.pca)
Importance of components:
                        Comp.1  Comp.2   Comp.3  Comp.4   Comp.5   Comp.6    Comp.7
Standard deviation     2.11194 1.09285 0.721813 0.67614 0.495244 0.270103 0.2213617
Proportion of Variance 0.63718 0.17062 0.074431 0.06531 0.035038 0.010422 0.0070001
Cumulative Proportion  0.63718 0.80780 0.882230 0.94754 0.982578 0.993000 1.0000000
> loadings(hep.pca)
Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
hurdles   0.453 -0.158                 0.783 -0.380
highjump -0.377  0.248 -0.368  0.680               -0.434
shot     -0.363 -0.289  0.676  0.124 -0.512        -0.218
run200m   0.408  0.260         0.361 -0.650         0.453
longjump -0.456         0.139  0.111  0.184  0.590  0.612
javelin         -0.842 -0.472  0.121 -0.135         0.173
run800m   0.375 -0.224  0.396  0.603  0.504 -0.156
>
> hep.pc<-predict(hep.pca)
> hep.pc[1:3,]
              Comp.1   Comp.2   Comp.3    Comp.4   Comp.5   Comp.6   Comp.7
Joyner-Kersee -4.2064 -1.26802  0.37754  0.023476  0.43479  0.34633  0.35510
John          -2.9416 -0.53453  0.91592 -0.485256 -0.71756 -0.24300  0.14699
Behmer        -2.7043 -0.69276 -0.46865 -0.693643  0.10770  0.24412 -0.13232

> eqscplot(hep.pc[,1],hep.pc[,2])
> eqscplot(hep.pc[,2],hep.pc[,3])
> eqscplot(hep.pc[,3],hep.pc[,4])
> eqscplot(hep.pc[,4],hep.pc[,5])
```





1 (continued)





2 Ten different wine tasters independently sampled seven wines:

`J`	Jurtschitsch Chardonnay
`Z`	Ziniel Chardonnay
`M`	Markowitsch Chardonnay
`R1`	Ritinitis Noble Retsina
`R2`	Retsina
`K`	Krems Chardonnay
`CN`	Castel Nova Chardonnay

Each taster gave the wines various scores for colour, smell, taste, fun and overall impression; based on these ratings, distances were computed. An edited R session occurs at the end of the question.

(i) Why might multidimensional analysis be suitable for this sort of analysis? ***(2 marks)***

(ii) With the aid of an informal graphical technique, how many dimensions would you recommend to provide an adequate representation of the data? ***(5 marks)***

(iii) Is the plot on the first two dimensions a good representation of the data? Justify your answer. ***(4 marks)***

(iv) Is it possible to plot all the points in any Euclidean space? Bearing in mind how the data was constructed, should this be surprising? Why might this have happened? ***(3 marks)***

(v) In the plot of the first two dimensions, two wines, `K` and `CN`, are plotted in almost exactly the same position. Are they joined by the minimum spanning tree? From the data, which wines would you expect to be their closest neighbours? Is this reflected in the spanning tree? ***(7 marks)***

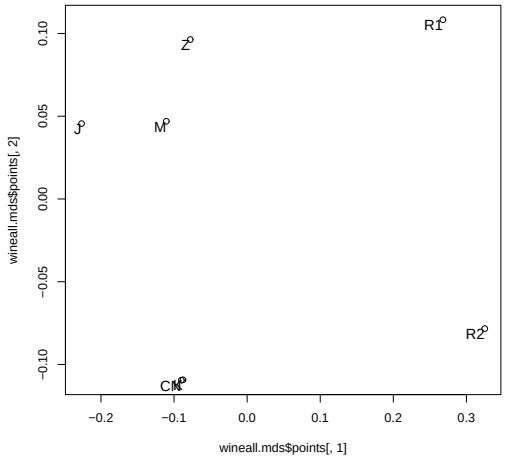
(vi) The plot of Kruskal scaling looks rather different to the plot of classical scaling. Indeed, it seems as though several points are all plotted at one location, and several at another. Which wines might you expect to be in the two locations? ***(2 marks)***

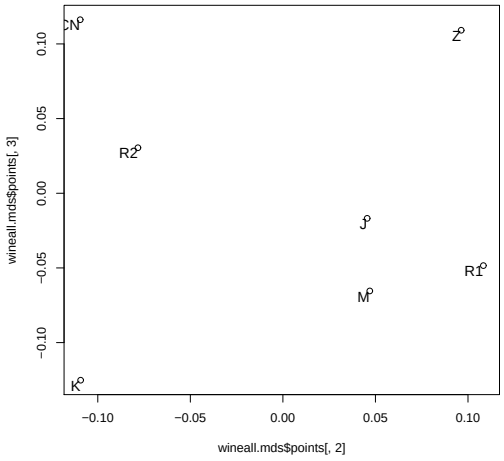
(vii) Comment on the differences between the Sammon and Kruskal plots. ***(2 marks)***

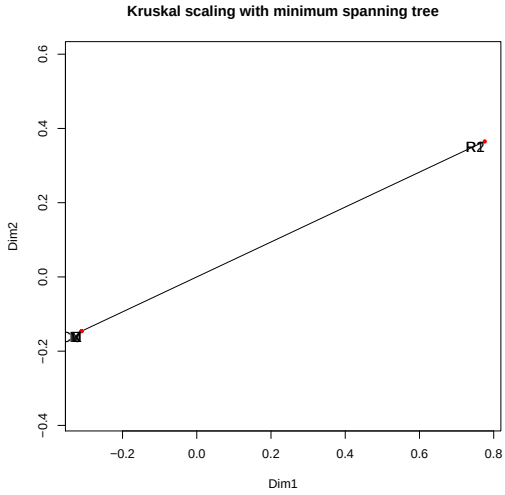
```
> wineset
       J     Z     M    R1    R2     K
Z  0.222
M  0.213 0.242
R1 0.501 0.396 0.407
R2 0.573 0.448 0.478 0.223
K  0.242 0.314 0.234 0.432 0.444
CN 0.262 0.245 0.260 0.452 0.434 0.268
> wine.tr<-spantree(wineset)
> wine.mds<-cmdscale(wineset)
```

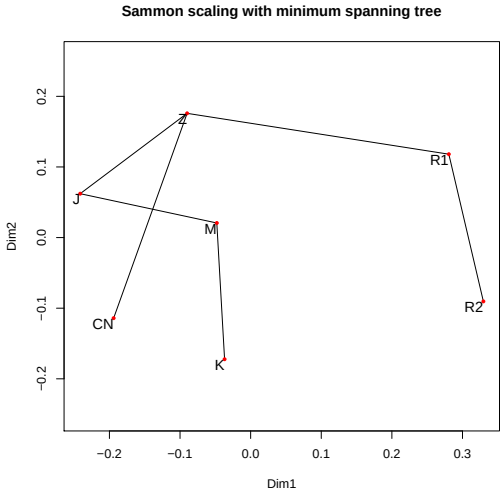
2 (continued)

```
> wine.mds
      [,1]    [,2]
J  -0.2266  0.0455
Z  -0.0779  0.0963
M  -0.1106  0.0469
R1  0.2678  0.1083
R2  0.3251 -0.0783
K  -0.0878 -0.1093
CN -0.0900 -0.1095
> wineall.mds<-cmdscale(wineset,k=5,eig=TRUE)
> wineall.mds$eig
[1]  2.63e-01  5.54e-02  4.89e-02  2.11e-02  1.08e-02 -3.78e-17 -3.22e-04
```









3 Measurements are taken on 6 approximately two-year old boys of height, chest circumference and mid-upper-arm circumference, all measured in centimetres (`Height`, `Circ` and `MUAC` in the R session below) in a highland area of a certain country.

```
> apply(boys,2,mean)
Height   Circ   MUAC
  82.0   60.2   14.5
> S<-var(boys)
> S
       Height  Circ MUAC
Height  31.6   8.0  0.5
Circ     8.0   3.2  1.3
MUAC     0.5   1.3  1.9
> solve(S)
        Height Circ   MUAC
Height  0.2    -0.6    0.4
Circ   -0.6     2.6   -1.6
MUAC    0.4    -1.6    1.5
```

(i) The corresponding means for children in lowland areas are considered to be 88cm, 58.4cm and 16cm respectively. Compute the statistic which you should use to test the hypothesis that the means for the highland children are the same, stating the distribution of the test statistic under this hypothesis. ***(6 marks)***

(ii) Given the R output

```
qf(.99,2,2)=99,     qf(.99,2,3)=30.82,  qf(.99,2,4)=18
qf(.99,3,2)=99.17,  qf(.99,3,3)=29.46,  qf(.99,3,4)=16.69
qf(.99,4,2)=99.25,  qf(.99,4,3)=28.71,  qf(.99,4,4)=15.98,
```

explain that we can reject the null hypothesis at a 1% level of significance. ***(1 mark)***

(iii) By computing an appropriate expression involving the F-statistic, and taking projections of the multivariate confidence ellipsoid onto each variable, compute the simultaneous 99% confidence intervals for the actual mean of each variable around the observed mean.

Verify that these confidence intervals contain $\mu = (88.0, 58.4, 16.0)'$. ***(6 marks)***

(iv) Given that $t_5(0.005) = 4.032$, give the 99% univariate confidence intervals for each variable mean individually.

Verify that these confidence intervals contain μ. ***(4 marks)***

(v) Comment on the apparent contradiction between the simultaneous multivariate test of part (ii), and the individual confidence intervals in parts (iii) and (iv).

To illustrate your argument, draw a sketch of a confidence region in the bivariate case (i.e., make a 2-dimensional sketch), its projections onto the co-ordinate axes, and suggesting how the apparent contraction can be seen on your sketch.

Conversely, again using your sketch, explain how a different set of observations might have individual means all significant at a given level, whereas the joint multivariate test would not allow us to reject the null hypothesis at the same level. ***(8 marks)***

4 Thomson and Randall-Maciver (1905) gave a data set of 150 Egyptian skulls from predynastic, 12th-13th dynasties and postdynastic periods (coded in the analyses below as periods 1, 2 and 3 respectively). The measurements available (all converted to cm) are maximum breadth (`mb`), basibregmatic height (`bh`), basialiveolar length (`bl`) and nasal height (`nh`). Skulls 1–60 come from period 1, skulls 61–90 from period 2, and skulls 91–150 from period 3.

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

(i) The first question is whether the measurements change over time; nonconstant measurements would indicate interbreeding with immigrant populations.

How would you test the hypothesis that the three groups of skulls are the same sizes? Discuss the R command you would use, and the number of degrees of freedom involved. ***(4 marks)***

(ii) Justifying your answer, what linear combination of the four measurements shows the greatest differences between the three periods? ***(4 marks)***

(iii) From which period are skulls most difficult to classify correctly with this technique? ***(2 marks)***

(iv) What proportion of the skulls used in the analysis does the linear discriminant analysis classify correctly? ***(2 marks)***

(v) It is proposed to try to understand whether or not the skulls from 12th-13th dynasties were more like the predynastic or postdynastic ones. With this in mind, it is proposed to repeat the Linear Discriminant Analysis simply with the skulls from periods 1 and 3, and then using Fisher's linear discriminant function to decide whether each skull from period 2 is closer to period 1 or to period 3.

(a) Compute Fisher's linear discriminant function for the LDA on the skulls from periods 1 and 3, and hence give the condition that a new observation is classified as period 1. ***(6 marks)***

(b) To assess the reliability of this method, the investigators wonder how it would work for a skull from period 1 or 3. Give a command that you would use in R to calculate the probability of misclassification.

What assumptions does this calculation make, and does this seem reasonable? (You are not expected to give any analysis.) ***(5 marks)***

(c) If a skull from period 3 were to be misclassified as from period 1, what features might be the most likely cause? ***(2 marks)***

```
     ***  Summary Statistics for data in:  skulls ***

period:1
           mb      bh      bl      nh
    Mean:  13.1867 13.3150  9.9117  5.0383
Std Dev.:   0.4956  0.4543  0.5129  0.2841
------------------------------------------------
period:2
           mb      bh      bl      nh
    Mean:  13.4467 13.3800  9.6033  5.0567
Std Dev.:   0.3481  0.4979  0.4552  0.3549
------------------------------------------------
```

4 (continued)

```
period:3
          mb      bh      bl      nh
    Mean: 13.5833 13.1317  9.4017  5.1667
Std Dev.:  0.4662  0.5107  0.4817  0.3287
>
> skull.lda<-lda(period~mb+bh+bl+nh,prior=c(1,1,1)/3)
> skull.lda
Coefficients of linear discriminants:
          LD1        LD2
mb -0.1256624 -0.0615404
bh  0.0277626 -0.1913655
bl  0.1485194  0.0592604
nh -0.0890593  0.1761597

Proportion of trace:
  LD1   LD2
0.882 0.118
>
> skull.pred<-predict(skull.lda,data.frame(cbind(mb,bh,bl,nh)))
>
> table(skull.pred$class,period)
   period
     1  2  3
  1 36  7 11
  2 12 16 14
  3 12  7 35
>
> skulls1<-skulls[1:60,]
> skulls3<-skulls[91:150,]
> m1<-apply(skulls1,2,mean)
> m3<-apply(skulls3,2,mean)
> pooled<-(var(skulls1)+var(skulls3))/2
> pooled
           mb          bh         bl        nh
mb  0.23146328 -0.00429096 0.00187288 0.0215876
bh -0.00429096  0.23361299 0.05776554 0.0370240
bl  0.00187288  0.05776554 0.24755650 0.0106483
nh  0.02158757  0.03702401 0.01064831 0.0943658
> sinv<-solve(pooled)
> a<-t(m1-m3)%*%sinv
> a
          mb     bh     bl      nh
[1,] -1.5894 0.4797 2.0210 -1.4126
> a%*%m1
         [,1]
[1,] -1.6578
> a%*%m3
         [,1]
[1,] -3.5883
```

End of Question Paper