The University Of Sheffield.

SCHOOL OF MATHEMATICS AND STATISTICS **Autumn Semester 2012–13**

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 p dimensional observations are available on items which may come from one of k distinct Normal populations $N_p(\mu_i, \Sigma)$, $i = 1, \ldots, k$, where the μ_i and Σ are known and Σ is a positive definite variance matrix.

(i) Show that the maximum likelihood rule classifies a new observation x into that population whose Mahalanobis distance from x is the smallest.

(2 marks)

(ii) Consider the case of three known Normal populations with common variance matrix Σ and means μ_1, μ_2, μ_3, where $\mu_3 = \dfrac{\mu_1 + \mu_2}{2}$. Let Δ_{ij} be the Mahalanobis distance between populations i and j. Show that $\Delta_{13} = \dfrac{\Delta_{12}}{2}$.

(5 marks)

(iii) If

$$P_{ij} = P(\text{classify a type } j \text{ as type } i),$$

then it is a standard result that, in the case of $k = 2$, the two misclassification probabilities P_{12} and P_{21} are each equal to $\Phi(-\Delta/2)$, where Δ is the Mahalanobis distance between the two populations and Φ denotes the cumulative distribution function of the standard Normal distribution $N(0, 1)$. Deduce that in the case of part (ii) above

$$P_{13} = P_{23} = \Phi(-\Delta_{12}/4).$$

(3 marks)

(iv) Using μ_1, μ_2, μ_3 defined in part (ii), show that the other four probabilities of misclassification are given by

$$P_{21} = P_{12} = \Phi(-3\Delta_{12}/4) \quad \text{and} \quad P_{31} = P_{32} = \Phi(-\Delta_{12}/4) - \Phi(-3\Delta_{12}/4).$$

[HINT: you may find it useful to sketch μ_1, μ_2, μ_3.] ***(6 marks)***

(v) Bivariate measurements are made on the setting times and bonding strengths from two samples of adhesive, both of size 47, those in the first sample having been previously identified as acceptable and those in the second as being of unacceptable standard. The sample means and pooled variance matrix of the two samples were

$$\mu_1 = \begin{pmatrix} 11 \\ 16 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 18 \\ 12 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} 10 & 2 \\ 2 & 4 \end{pmatrix}.$$

The data will be used to construct a sample discriminant rule for identifying substandard batches on the basis of such bivariate measurements. The first procedure proposed is to have just two categories 'acceptable' and 'unacceptable'. What proportion of unacceptable batches is likely to be erroneously passed as acceptable when using this rule? ***(9 marks)***

2 Measurements (in Newtons per square centimetre) were made on each of 11 samples of timber of x_1 (stiffness) and x_2 (bending strength) both before and after a new resin treatment. The sample mean vectors and the variance matrices of the sample means were as follows

$$\text{Before treatment:} \quad \bar{x}_B = \begin{pmatrix} \bar{x}_{1B} \\ \bar{x}_{2B} \end{pmatrix} = \begin{pmatrix} 1420 \\ 9175 \end{pmatrix}, \quad S_B = \begin{pmatrix} 265 & 350 \\ 350 & 2525 \end{pmatrix}$$

$$\text{After treatment:} \quad \bar{x}_A = \begin{pmatrix} \bar{x}_{1A} \\ \bar{x}_{2A} \end{pmatrix} = \begin{pmatrix} 1440 \\ 9295 \end{pmatrix}, \quad S_A = \begin{pmatrix} 240 & 375 \\ 375 & 2610 \end{pmatrix}.$$

The covariance between before and after treatment measurements of x_1 was 105 and the covariance between before and after treatment measurements of x_2 was 375. The covariances between non-corresponding measurements before and after treatment were negligible.

(i) Calculate the variance matrix of the sample mean change in strength measurements as a result of the new resin treatment. ***(7 marks)***

(ii) Do these data provide evidence that the new resin treatment changes the overall strength of the timber on average? ***(11 marks)***

(iii) What linear combination of the two measurements of strength shows the greatest mean change as a result of the treatment? ***(7 marks)***

3 An investigation on housing values was conducted in the suburbs of Boston. A data set of 506 sampled houses measured 12 variables of interest as follows:

- `crim`: per capita crime rate by town.
- `zn`: proportion of residential land zoned for lots over 25,000 sq.ft.
- `indus`: proportion of non-retail business acres per town.
- `nox`: nitrogen oxides concentration (parts per 10 million).
- `rm`: average number of rooms per dwelling.
- `age`: proportion of owner-occupied units built prior to 1940.
- `dis`: weighted mean of distances to five Boston employment centres.
- `tax`: full-value property-tax rate per \$10,000.
- `ptratio`: pupil-teacher ratio by town.
- `black`: $1000(Bk - 0.63)^2$ where Bk is the proportion of black people by town.
- `lstat`: lower status of the population (percent).
- `medv`: median value of owner-occupied homes in \$1000s.

A principal components analysis was performed in R and part of the output is given below.

(i) Explain why it is considered a better option to use the correlation matrix instead of the variance matrix in order to compute the principal components. ***(2 marks)***

(ii) What do you expect the sample correlation coefficient between PC2 and PC3 to be? Justify your answer. ***(2 marks)***

(iii) Given that the correlation matrix has been used for the calculation of the principal components and using the R output, make a scree plot and decide which principal components you would keep. ***(9 marks)***

(iv) Give the interpretation of the first 3 principal components. ***(10 marks)***

(v) By looking at the output (see Figure 1 on page 7) a practitioner claims that the above principal component analysis is wrong, because the data do not seem to be Normally distributed. Is she correct? ***(2 marks)***

3 (continued)

R output

```
First 5 rows of the data
  crim  zn   indus nox  rm   age    dis tax ptratio  black lstat medv
1  0.01 18.0  2.31 0.53 6.57  65.2  4.09 296 15.3 396.90  4.98 24.0
2  0.03  0.0  7.07 0.46 6.42  78.9  4.96 242 17.8 396.90  9.14 21.6
3  0.03  0.0  7.07 0.46 7.18  61.1  4.96 242 17.8 392.83  4.03 34.7
4  0.03  0.0  2.18 0.45 6.99  45.8  6.06 222 18.7 394.63  2.94 33.4
5  0.07  0.0  2.18 0.45 7.15  54.2  6.06 222 18.7 396.90  5.33 36.2


Variance sub-matrix (for variables crim, zn, indus, nox)

           crim      zn      indus    nox
crim      73.986   -40.216   23.992  0.419
zn       -40.216   543.937  -85.413 -1.396
indus     23.992   -85.413   47.064  0.607
nox        0.419    -1.396    0.607  0.013

Eigenvalues of the correlation matrix
$values
 [1] 5.99 1.52 1.14 0.81 0.64 0.52 0.40 0.25 0.21 0.19 0.18 0.15

Principal components (1-9)
Loadings:
        Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
crim     0.242         0.492 -0.113 -0.420  0.654  0.181
zn      -0.267  0.163  0.449  0.256 -0.344 -0.375 -0.329 -0.347
indus    0.349 -0.131                       -0.418  0.241        -0.444
nox      0.341 -0.275         0.209        -0.203         0.234  0.434
rm      -0.224 -0.497  0.276 -0.315               -0.427  0.498 -0.264
age      0.317 -0.285 -0.165                0.121 -0.602 -0.228  0.325
dis     -0.316  0.399  0.151                -0.112 -0.128  0.324  0.110
tax      0.321         0.328 -0.205 -0.257 -0.405  0.104  0.109
ptratio  0.213  0.337        -0.772        -0.107 -0.236 -0.237
black   -0.200        -0.557 -0.111 -0.778                0.164
lstat    0.332  0.224         0.313         0.105 -0.378        -0.616
medv    -0.287 -0.475        -0.143                0.159 -0.567 -0.176
```

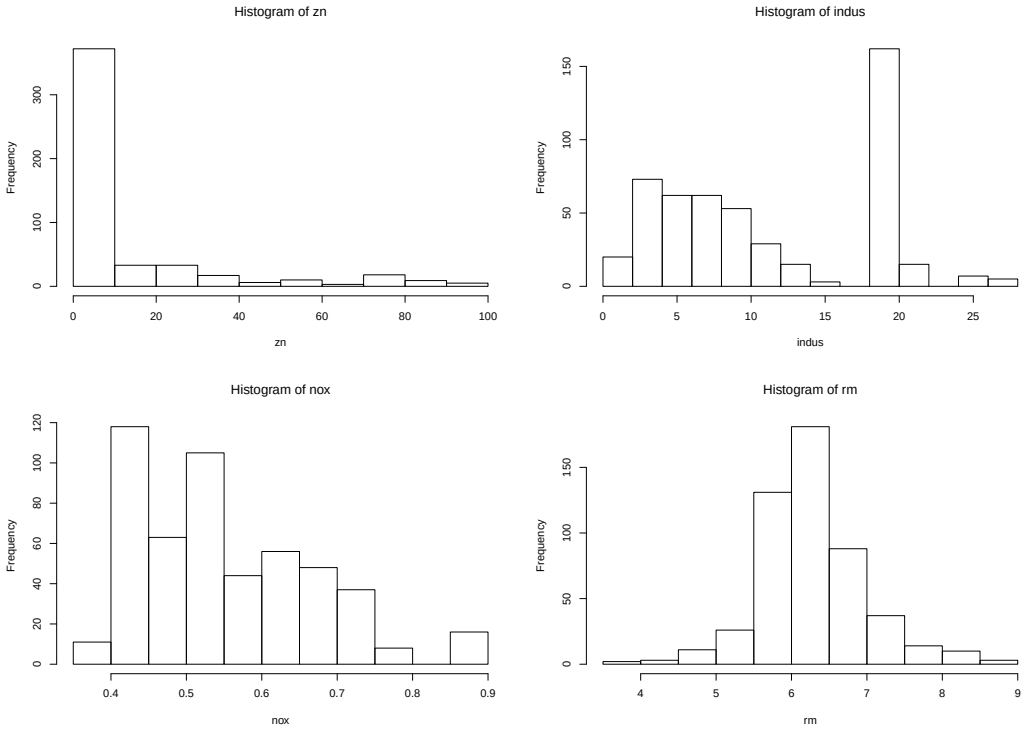

Figure 1: Histogram of four variables: `zn`, `indus`, `nox` and `rm`

4 A study was made of samples of ellipsoidal pebbles taken from two tributaries of a river, one tributary flowing into the upper part of the river and the other into the lower. On each pebble, the maximum and minimum diameter was measured. The 21 pebbles from the upper tributary gave mean maximum and minimum diameters of 10.3mm and 8.5mm with sample variances 1.69mm^2 and 1.30mm^2 and covariance 0.1mm^2. The measurements of the 31 pebbles from the lower tributary gave mean results of 9.1mm and 7.9mm with sample variances 1.33mm^2 and 1.21mm^2 and sample covariance 0.14mm^2.

(i) Calculate Fisher's linear discriminant function for classifying a pebble with maximum and minimum diameters (x_1, x_2) as deriving from the upper tributary (x_1) or lower tributary (x_2). ***(12 marks)***

(ii) Assuming that these measurements are adequately modelled by bivariate Normal distributions with a common variance matrix and that the classification of pebbles uses Fisher's discriminant function, estimate the probability of misclassifying a randomly selected lower tributary pebble as deriving from the upper tributary. ***(8 marks)***

(iii) Two further pebbles whose labels were lost during the study had diameters $(9.7, 8.2)$ and $(9.9, 8.0)$. What are the best assessments of the sources of these two pebbles? ***(5 marks)***

End of Question Paper