UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

Tuesday 7 June, 2005 1:30 to 3:30

APPLIED MULTIVARIATE ANALYSIS

*Attempt **THREE** questions.*

*There are **FOUR** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Suppose the p-dimensional vector X is distributed as $N_p(\mu, V)$. Show that if we partition X into components X_1, X_2, so that $X^T = (X_1^T, X_2^T)$, then the covariance matrix of X_1 conditional on $X_2 = x_2$ is $V_{11} - V_{12}V_{22}^{-1}V_{21}$, where

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

In a classic heredity study, Frets (1921) measured the head lengths and head breadths on the first and second adult sons in a sample of families. Let (X_1, X_2) be the head length and breadth of the first son and (Y_1, Y_2) the corresponding quantities for the second son.

Considering these four measurements as the 4-vector $Z^T = (X^T, Y^T) = (X_1, X_2, Y_1, Y_2)$, a reasonable model for the population from which the sample has come is a Normal population with mean vector

$$\mu = (\mu_1, \mu_2, \mu_1, \mu_2)^T$$

and dispersion matrix

$$V = \begin{pmatrix} a & b & c & c \\ b & a & c & c \\ c & c & a & b \\ c & c & b & a \end{pmatrix}$$

for some positive a, b, c such that V is positive definite, $a > c$ and $b > c$.

Obtain

(i) the joint distribution of $X_1 - Y_1$ and $X_2 - Y_2$;

(ii) the marginal distribution of $X_1 - Y_1$;

and

(iii) the conditional distribution of $X_1 - Y_1$ given that $X_2 - Y_2 = 0$.

Comment on the differences between (ii) and (iii)

2 Let $X_1, X_2, \ldots X_p$ be p variables observed on a random sample of n individuals from a population with covariance matrix V, and let $(x_{i1}, x_{i2}, \ldots, x_{ip})$ be the values of these variables for the ith individual of the sample.

(i) Show how to choose ℓ_1 such that $\ell_1^T \ell_1 = 1$, and such that

$$\text{var } (\ell_1^T X) \quad \text{is maximum}\,.$$

(ii) Now show how to choose ℓ_2 such that $\ell_2^T \ell_2 = 1$, $\ell_2^T \ell_1 = 0$ and var $(\ell_2^T X)$ is maximum.

Hence define the principal components of the sample described above.

Discuss briefly the desirability or otherwise of standardising the data before extracting principal components.

An experiment was conducted to determine the semantic attributes of 292 words, in which each of a number of children rated each word for the following characteristics

(1) friendly/unfriendly
(2) good/bad
(3) nice/awful
(4) brave/not brave
(5) big/little
(6) strong/weak
(7) moving/still
(8) fast/slow.

Each child gave each word a score from 0 to 10 for each characteristic with the given adjectives attached to the end of the scale, and the resulting scores were averaged over children to give a 292×8 matrix. The dataset was standardised, and eigenvalues and eigenvectors extracted from the correlation matrix. The first three eigenvalues were 4.77, 1.53 and 0.81 respectively and their corresponding eigenvectors had the following coefficients

Characteristic

	1	2	3	4	5	6	7	8
Vector								
1	0.87	0.88	0.87	0.89	0.58	0.85	0.49	0.61
2	-0.42	-0.42	-0.44	0.10	0.26	0.19	0.70	0.61
3	-0.13	-0.10	-0.15	0.06	0.72	0.22	-0.29	-0.32

Interpret these eigenvalues and coefficients of the associated eigenvectors.

3 Fisher's classic "iris" data consists of a table 150×5, of which the first 3 rows are given in the Splus6 output below. There are 3 distinct species, denoted here by "setosa", "versicolor" and "virginica", and we wish to construct a classification tree to sort the 150 iris specimens into species according to the values of Sepal Length, Sepal Width, Petal Length and Petal Width. Explain carefully the construction of the Splus object "iris.tr" in the output below, and sketch the resulting classification tree.

```
> iris[1:3,]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
> iris.tr <- tree(Species~.,iris);summary(iris.tr)

Classification tree:
tree(formula = Species ~ ., data = iris)
Variables actually used in tree construction:
[1] "Petal.Length" "Petal.Width"  "Sepal.Length"
Number of terminal nodes:  6
Residual mean deviance:  0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
> iris.tr
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

1) root 150 329.600 setosa ( 0.3333 0.33330 0.33330 )
   2) Petal.Length<2.45 50   0.000 setosa ( 1.0000 0.00000 0.00000 ) *
   3) Petal.Length>2.45 100 138.600 versicolor ( 0.0000 0.50000 0.50000 )
     6) Petal.Width<1.75 54  33.320 versicolor ( 0.0000 0.90740 0.09259 )
      12) Petal.Length<4.95 48   9.721 versicolor ( 0.0000 0.97920 0.02083 )
        24) Sepal.Length<5.15 5   5.004 versicolor ( 0.0000 0.800000.20000 ) *
        25) Sepal.Length>5.15 43   0.000 versicolor ( 0.0000 1.000000.00000 ) *
      13) Petal.Length>4.95 6   7.638 virginica ( 0.0000 0.33330 0.66670 ) *
     7) Petal.Width>1.75 46   9.635 virginica ( 0.0000 0.02174 0.97830 )
      14) Petal.Length<4.95 6   5.407 virginica ( 0.0000 0.16670 0.83330 )*
      15) Petal.Length>4.95 40   0.000 virginica ( 0.0000 0.00000 1.00000 ) *
```

4 Write brief essays, which should include appropriate sketch graphs, on two of the following topics.

(i) Multivariate Analysis of Variance

(ii) Factor Analysis

(iii) Clustering Algorithms.

END OF PAPER