UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday 6 June, 2003 1:30 to 4:30

PAPER 42

Experimental Design and Multivariate Analysis

Attempt **FOUR** *questions.*

There are **six** *questions in total.*

The questions carry equal weight.

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

1 Applied Multivariate Analysis

Suppose that $x_1, \ldots, x_n$ is a random sample from the p-variate normal distribution $N(\mu, V)$.

(i) Show that if $\ell(\mu, V)$ is the corresponding log-likelihood function, then

$$-2\ell(\mu, V) = n \log|V| + ntr(V^{-1}S) + n(\bar{x} - \mu)^T V^{-1}(\bar{x} - \mu),$$

where you should define $\bar{x}, S$.

(ii) State without proof the formulae for the maximum likelihood estimators of μ, V, and (also without proof) the joint distribution of these two quantities.

(iii) Show that the generalised likelihood ratio test of

$$H_0 : V \quad \text{is a diagonal matrix}$$

may be written as

$$\text{reject } H_0 \text{ if } \log|R| < \text{ constant},$$

where R is the sample correlation matrix derived from $x_1, \ldots, x_n$. What is the form of this test in the case $p = 2$?

2 Applied Multivariate Analysis

Each of a class of 13 students is given a set of 10 "Yes/No" questions to answer, and the corresponding data matrix is read into S-Plus, with the response "Yes" corresponding to a 1, and "No" to a 0.

(i) Explain carefully how the distance matrix is computed. (You may assume that the function

$$\text{dist2full } (d)$$

converts $(d_{ij}\,,\, 1 \leqslant i < j \leqslant 13)$ to the "full" distance matrix $(d_{ij}\,,\, 1 \leqslant i, j \leqslant 13)$.)

(ii) Explain briefly the results of the hierarchical clustering algorithm, and sketch the graph that you would expect it to give.

```
> a _ read.table("students", header=T)
> a _ as.matrix(a); a
         eggs meat coffee beer UKres Cantab Fem sports driver Left.h
  Philip    1    1      1    0     1      1   0      0      1      1
    Chad    1    1      1    0     0      0   0      1      1      0
  Graham    1    1      1    1     1      1   0      1      1      0
     Tim    1    1      1    1     1      1   0      1      0      0
    Mark    1    1      0    1     1      1   0      0      0      1
  Juliet    0    1      1    0     1      0   1      0      0      0
Garfield    0    1      1    1     0      0   0      1      0      0
 Nicolas    1    1      1    1     0      0   0      1      1      0
Frederic    1    1      0    1     0      0   0      1      1      0
    John    1    1      1    1     0      0   0      0      1      0
   Sauli    1    1      0    0     1      0   0      1      1      0
    Fred    1    1      1    0     0      0   0      1      0      0
  Gbenga    1    1      1    0     0      0   0      1      0      0
```

```
> d _ dist(a, metric="binary") ; round(dist2full(d), 2)

      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
 [1,] 0.00 0.50 0.33 0.44 0.38 0.62 0.78 0.56 0.67  0.50  0.50  0.62  0.62
 [2,] 0.50 0.00 0.38 0.50 0.78 0.71 0.50 0.17 0.33  0.33  0.33  0.20  0.20
 [3,] 0.33 0.38 0.00 0.12 0.44 0.67 0.50 0.25 0.38  0.38  0.38  0.50  0.50
 [4,] 0.44 0.50 0.12 0.00 0.38 0.62 0.43 0.38 0.50  0.50  0.50  0.43  0.43
 [5,] 0.38 0.78 0.44 0.38 0.00 0.75 0.75 0.67 0.62  0.62  0.62  0.75  0.75
 [6,] 0.62 0.71 0.67 0.62 0.75 0.00 0.67 0.75 0.88  0.71  0.71  0.67  0.67
 [7,] 0.78 0.50 0.50 0.43 0.75 0.67 0.00 0.33 0.50  0.50  0.71  0.40  0.40
 [8,] 0.56 0.17 0.25 0.38 0.67 0.75 0.33 0.00 0.17  0.17  0.43  0.33  0.33
 [9,] 0.67 0.33 0.38 0.50 0.62 0.88 0.50 0.17 0.00  0.33  0.33  0.50  0.50
[10,] 0.50 0.33 0.38 0.50 0.62 0.71 0.50 0.17 0.33  0.00  0.57  0.50  0.50
[11,] 0.50 0.33 0.38 0.50 0.62 0.71 0.71 0.43 0.33  0.57  0.00  0.50  0.50
[12,] 0.62 0.20 0.50 0.43 0.75 0.67 0.40 0.33 0.50  0.50  0.50  0.00  0.00
[13,] 0.62 0.20 0.50 0.43 0.75 0.67 0.40 0.33 0.50  0.50  0.50  0.00  0.00
> h _  hclust(d, method="compact"); h

$merge:
      [,1] [,2]
 [1,]  -12  -13
 [2,]   -3   -4
 [3,]   -2   -8
 [4,]   -9    3
 [5,]  -10    4
 [6,]   -1   -5
 [7,]   -7    1
 [8,]    2    6
 [9,]    5    7
[10,]  -11    8
[11,]   -6   10
[12,]    9   11
$height:
 [1] 0.0000000 0.1250000 0.1666667 0.3333333 0.3333333 0.3750000 0.4000000
 [8] 0.4444444 0.5000000 0.6250000 0.7500000 0.8750000
> plclust(h)
```

3 Applied Multivariate Analysis

The S-Plus output given below shows the dataset "trees" (slightly edited), in which the 3 columns correspond respectively to Girth (in inches) at a height of 4.5 feet from the ground, Height (in feet) of the tree, and Volume (in cubic feet) of useable wood from the tree, for a sample of 31 cherry trees.

(i) How is the sample covariance matrix of the log of these 3 columns computed?

(ii) Discuss carefully the rationale behind the command princomp( ) and explain (with a suitable sketch graph) what the output is telling you. What other use might you make of the princomp( ) command?

```
 >trees
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
3    8.8     63   10.2
4   10.5     72   16.4
5   10.7     81   18.8
6   10.8     83   19.7
7   11.0     66   15.6
8   11.0     75   18.2
 ..............
 ..............
28  17.9     80   58.3
29  18.0     80   51.5
30  18.0     80   51.0
31  20.6     87   77.0
> b _ as.matrix(trees); var(log(b))
            Girth     Height     Volume
Girth  0.05458445 0.01061861 0.12008415
Height 0.01061861 0.00734844 0.02926209
Volume 0.12008415 0.02926209 0.27695632

> first.pc _  princomp(log(b), cor=F)
> summary(first.pc)
Importance of components:
                         Comp.1     Comp.2      Comp.3
  Standard deviation     0.5671603 0.07312407 0.030648264
 Proportion of Variance  0.9808315 0.01630435 0.002864141
 Cumulative Proportion   0.9808315 0.99713586 1.000000000
```

4 Design of Experiments

Let n be the incidence matrix for a general block design with t treatments to be compared in b blocks, each containing $k(\leqslant t)$ experimental units, so that $n_{ij} = 1$ if the ith treatment occurs in the jth block, and $n_{ij} = 0$ otherwise. Interpret $(nn^T)_{ii}$ and $(nn^T)_{il}, i \neq l$, in words.

Consider a design where $nn^T = (r - \lambda)I_t + \lambda \mathbf{1}_t \mathbf{1}_t^T$, where I_t is the $t \times t$ identity matrix and $\mathbf{1}_t$ is a $t \times 1$ vector of ones. Determine what kind of design this is

$$\text{(i)} \quad \text{if } r = \lambda \quad \text{and} \quad \text{(ii)} \quad \text{if } \lambda < r\,.$$

In case (ii), show that $k < t$ and $b \geqslant t$.

The main effects of five different types of hardwood A, B, C, D, E on paper strength are being investigated. Four observations are obtained on each type, giving the following analysis of variance

	df	ss
Hardwood types	*	42.4
Residual	*	46.7
Total (corrected)	*	89.1

Complete the missing degrees of freedom, and test whether there is a significant difference between the hardwood types.

In fact, further information reveals that the observations were obtained over a period of five days as shown below.

Day	1	2	3	4	5
Hardwood types:	$ABDE$	$BCDE$	$ACDE$	$ABCD$	$ABCE$

What type of design is this?

The sum of squares for days is 28.4 and the sum of squares for hardwood types (adjusted for days) is 18.3. Test whether there is a significant difference between the hardwood types.

[**Hint:** If $\mathbb{P}(Y > F_{m,n}(\alpha)) = \alpha$ where $Y \sim F_{m,n}$, then

$$F_{5,14}(0.10) = 2.31 \quad F_{5,10}(0.10) = 2.52 \quad F_{4,15}(0.10) = 2.36 \quad F_{4,11}(0.10) = 2.54$$
$$F_{5,14}(0.05) = 2.96 \quad F_{5,10}(0.05) = 3.33 \quad F_{4,15}(0.05) = 3.06 \quad F_{4,11}(0.05) = 3.36 \quad]$$

5 Design of Experiments

Explain how to construct a $\frac{1}{2^k}$ replicate of a 2^m experiment (you may quote results from lectures without proof). If $k = 1$, show that each contrast is aliased with one other contrast.

An experimenter wishes to investigate all main effects and all two-factor interactions except AF of six factors A, B, C, D, E, F (each at two levels) affecting the operation of an industrial process. In order to carry out the experiment, normal production must be stopped, and it is decided that normal production can only be interrupted for four days. Treating days as blocks, explain how to construct a suitable design if eight treatment combinations can be tested in a single day, and only 4 days are available for the experiment. You may assume that third and higher order interactions are negligible. Give the partition of the degrees of freedom in the resulting analysis of variance table.

6 Design of Experiments

A scientist wishes to maximize the yield, y, during crystal growth as a function of three coded variables x_1, x_2, x_3. As part of a search procedure for the maximum, trials are run at the 14 points given below, with associated yields $y_1, \ldots, y_{14}$ as shown.

x_1	x_2	x_3	Yield
-1	-1	-1	y_1
-1	-1	1	y_2
-1	1	-1	y_3
-1	1	1	y_4
1	-1	-1	y_5
1	-1	1	y_6
1	1	-1	y_7
1	1	1	y_8
0	0	0	y_9
0	0	0	y_{10}
0	0	0	y_{11}
0	0	0	y_{12}
0	0	0	y_{13}
0	0	0	y_{14}

Consider the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$, where errors for different runs are independent $N(0, \sigma^2)$ random variables. Find the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ in terms of $y_1, \ldots, y_{14}$, and write down the distribution of $\hat{\boldsymbol{\beta}}$.

After examining the fit of this model, six more trials are run as shown below.

x_1	x_2	x_3	Yield
-1.682	0	0	y_{15}
1.682	0	0	y_{16}
0	-1.682	0	y_{17}
0	1.682	0	y_{18}
0	0	-1.682	y_{19}
0	0	1.682	y_{20}

What design is formed by all 20 points?

A full second order model is fitted to these 20 points. Write down this model. The residual sum of squares is found to be 1860.98, and the sum of squares from the six centre points is

$$\sum_{i=9}^{14} \left(y_i - \frac{1}{6} \sum_{k=9}^{14} y_k \right)^2 = 859.33 \ .$$

Describe how to test for lack of fit of this model.

[**Hint:** If $\mathbb{P}(Y > F_{m,n}(\alpha)) = \alpha$ where $Y \sim F_{m,n}$, then

$$F_{9,10}(0.10) = 2.35 \quad F_{5,5}(0.10) = 3.45 \quad F_{4,6}(0.10) = 3.18$$
$$F_{9,10}(0.05) = 3.02 \quad F_{5,5}(0.05) = 5.05 \quad F_{4,6}(0.05) = 4.53 \qquad]$$