UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

Friday 8 June 2001 9 to 12

EXPERIMENTAL DESIGN AND APPLIED MULTIVARIATE ANALYSIS

*Attempt any **FOUR** questions. 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 (i) Let $y_1, \ldots, y_n$ be a random sample from $N(\mu, V)$, where μ and V are unknown. Show that if $\ell(\mu, V)$ is the resulting log-likelihood function, then we may write

$$\max_{\mu} \ell(\mu, V) = +\frac{n}{2} \log |V^{-1}S| - \frac{n}{2} \mathrm{tr}(V^{-1}S) + \text{constant},$$

where S is a matrix that you should define.

Hence find the maximum likelihood estimator of V.

(ii) Using (i) above, construct the large-sample likelihood ratio test of

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

2 Whittaker, 'Graphical Methods in Applied Multivariate Statistics' (1990), discusses a dataset for five different mathematics exams for 88 students, for which the sample covariance matrix is

mech	302.29				
vect	125.78	170.88			
alg	100.43	84.19	111.60		
anal	105.07	93.60	110.84	217.88	
stat	116.07	97.89	120.49	153.77	294.37

Explain how you would use this to show that we may estimate

$$\text{var(mech} \mid \text{remaining 4 variables) as 0.62 * var(mech),}$$

and that we may estimate

$$\text{cor(stat, anal|remaining 3 variables) as 0.25.}$$

(This question requires no arithmetical calculations.)

3 Write a brief essay, with appropriate sketch graphs on *two* of the following *four* subjects

(a) principal components analysis

(b) classical multidimensional scaling

(c) hierarchical cluster analysis

(d) multivariate analysis of variance.

4 Explain what is meant by a *balanced incomplete block design* BIBD (t, b, r, k, λ), and a *symmetrical* balanced incomplete block design.

Show that

(i) $bk = \text{tr}$,

(ii) $r(k-1) = \lambda(t-1)$

(iii) $b \geqslant t$.

For a symmetrical balanced incomplete block design with an even number of blocks, show that $r - \lambda$ is a perfect square.

Determine whether or not it is possible to have a balanced incomplete block design with the following parameters. If it is not possible, explain why; if it is possible then give a suitable design

(a) $t = b = 46, \qquad \lambda = 2;$

(b) $t = b = 4, \qquad \lambda = 2.$

5 Explain how k defining contrasts in a 2^n experiment split the treatments into 2^k equal-sized sets (results from lectures should be clearly stated, but need not be proved). Explain the term *aliasing* with reference to fractional replication in 2^n experiments.

In an experiment to investigate what factors affect the compressive strength of concrete cylinders, there are five factors, each at two levels:

S type of sand

C type of cement

W amount of water

M time to mix

T time to mould.

It is assumed that all third and higher order interactions are negligible, and also that all two-factor interactions except ST are negligible. There are only 8 moulds available for the experiment, and all main effects and the ST interaction are of interest. Determine whether the following designs are suitable:

(a) $\{(1), cs, st, mw, ct, cmsw, mstw, cmtw\}$

(b) $\{(1), cs, mt, stw, ctw, msw, cmw, csmt\}$.

For each of (a) and (b), give all the aliases of each contrast of interest.

6 Suppose $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, X is a known $n \times p$ matrix with rank p and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ with $\epsilon_1, \ldots, \epsilon_n$ iid $N(0, \sigma^2)$ random variables. Find the least squares estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Find $\mathbb{E}\hat{\boldsymbol{\beta}}$ and $\mathrm{cov}\hat{\boldsymbol{\beta}}$.

A chemical engineer wishes to find the settings of temperature ξ_1, and amount ξ_2 of catalyst that maximize the yield in a particular industrial process, where $\xi_1 \in [90, 100]$ and $\xi_2 \in [35, 40]$ (in appropriate units). Explain what is meant by coded variables x_1, x_2. The engineer runs the process once each with $(x_1, x_2) = (1, 1)$, $(1, -1), (-1, -1), (-1, 1)$ and n_0 times at $(0, 0)$, and obtain yields $Y_1, Y_2, Y_3, Y_4, Y_5, \ldots, Y_{4+n_0}$ respectively.

Find the least squares estimates of $\beta_0, \beta_1, \beta_2$ for the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

Explain why the n_0 runs at $(0, 0)$ are included. Assuming this model is adequate, how should the engineer proceed using the method of steepest ascent to search for the maximum?

Suppose instead that the above model is inadequate and curvature is indicated. Describe what further runs the engineer should carry out, and indicate what model would then be fitted. Assuming the latter model is adequate, describe briefly how it may be used in investigation of optimal settings of temperature and amount of catalyst.