UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday 1 June 2007 1.30 to 3.30

PAPER 47

APPLIED MULTIVARIATE ANALYSIS

*Attempt **THREE** questions.*

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

The questions carry equal weight.

Candidates may use the following without proof.
The value of a that maximises $\dfrac{(a^T b)}{a^T C a}$ *is* $C^{-1}b$ *where* a, b *are* $(p \times 1)$ *and* C *is* $(p \times p)$.

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 (a) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ denote a random sample from a $N_p(\boldsymbol{\mu}, \Sigma)$ distribution. By considering linear compounds of the form $Y = \mathbf{a}^T\mathbf{X}$ for some vector of constants $\mathbf{a}$, derive Hotelling's T^2 statistic for testing the hypothesis that $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ (for Σ unknown) against the alternative $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, showing that it is given by

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)^T S^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0),$$

where $\bar{\mathbf{X}}$ is the mean and S the covariance matrix of the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Given that if $\mathbf{Z} \sim N_p(\mathbf{0}, \Sigma)$ and C has the Wishart distribution $\omega_p(k, \Sigma)$ then $\frac{(k-p+1)}{p} X^T C^{-1} X \sim F_{p,k-p+1}$, derive the distribution of T^2 under the null hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$.

(b) Describe briefly how we can modify this test to test hypotheses of the form

$$H_0 : A^T\boldsymbol{\mu} = \mathbf{0} \quad \text{vs} \quad H_1 : A^T\boldsymbol{\mu} \neq \mathbf{0}$$

where A^T is a specified $(m \times p)$ matrix.

(c) A p-dimensional variate $\mathbf{X}$ comprises measurements on the temperature of an experimental unit at equally-spaced time points $t_1, \ldots, t_p$. A hypothesis states that the mean temperature varies linearly with time, i.e.,

$$H_0 : \mu_i = \alpha + \beta t_i, \quad i = 1, \ldots, p.$$

Show that under this hypothesis

$$(\mu_{i+1} - \mu_i) - (\mu_i - \mu_{i-1}) = 0, \quad i = 2, \ldots, p-1,$$

and hence say how to conduct a significance test of hypothesis H_0. Show how the procedure might be modified to cope with unequal time intervals.

2 (a) Show that the integral $\int_R g(\mathbf{x})\,d\mathbf{x}$ is minimised with respect to R by taking

$$R = \{\mathbf{x} : g(\mathbf{x}) < 0\}.$$

(b) Suppose we have a single population divided into two mutually exclusive and exhaustive subgroups P_1 and P_2. Suppose further that a proportion π_i of individuals belong to group P_i and that if $\mathbf{x}$ is a member of P_i, then $\mathbf{x} \sim f_i(\mathbf{x})$. The range of $\mathbf{x}$ is partitioned into two disjoint sets R_1 and R_2, and we consider the rule where $\mathbf{x}$ is assigned to P_1 if $\mathbf{x} \in R_1$ and to P_2 otherwise.

(i) Derive the total probability of misclassification under this rule in terms of π_1, R_1, f_1 and f_2.

(ii) Using the result proved in (a) above, derive the form of discriminant function that minimises the total probability of misclassification.

(iii) If we associate cost $c(i|j)$ to misclassifying an individual in group j to group i, how would your discriminant function in (ii) change if we chose to minimise the total cost of misclassification?

(c) Suppose we have $\pi_1 = \pi_2 = 1/2$ and two bivariate normal groups with common covariance matrix $\Sigma = \begin{pmatrix} 6 & 2 \\ 2 & 1 \end{pmatrix}$ and respective means

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}.$$

Derive a discriminant function for assigning a new observation $\mathbf{x}$ to one of these two groups.

Find an expression for the probability of misclassifying an individual using this rule when $c(1|2) = c(2|1) = 1$.

If $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, to which group would you assign this individual?

3 (a) What is the purpose of a Principal Components Analysis?

Let $\mathbf{X}$ be a p-variate random variable with covariance matrix Σ. Derive the principal components $\mathbf{Y}$ of $\mathbf{X}$.

Write down the covariance matrix Λ of $\mathbf{Y}$ and show how we can express Λ in terms of Σ. Hence prove that $\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \lambda_i$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p$ are the eigenvalues of Σ.

(b) Let $\mathbf{X} = [X_1, X_2, X_3, X_4]^T$ be a random variable with covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{pmatrix}.$$

Show that the eigenvalues of Σ are $(1-r), (1-r), (1-r)$ and $(1+3r)$. For $0 < r < 1$ find the first principal component, $\mathbf{Y}_1$.

What proportion of the total variation is accounted for by $\mathbf{Y}_1$?

(c) An analysis of protein consumption for 25 European countries published in 1973 gave estimates of the average protein consumption (in grams per person per day) classified into 9 food groups for each country.

The sample variances for the 9 food groups are given by:-

Red meat	11.2
White meat	13.6
Eggs	1.2
Milk	50.5
Fish	11.6
Cereals	120.4
Starch	2.7
Pulses & Nuts	3.9
Fruit & veg	3.3

Total variation = 218.4

Principal component analysis was carried out on both the covariance and correlation matrices. The vectors of coefficients and variances of the first 3 components are given below.

$$
\begin{array}{lcccccc}
 & \multicolumn{3}{c}{\text{Analysis on covariance matrix}} & \multicolumn{3}{c}{\text{Analysis on correlation matrix}} \\
\text{vectors} &
\begin{bmatrix} 0.15 \\ 0.13 \\ 0.07 \\ 0.42 \\ 0.13 \\ -0.86 \\ 0.07 \\ -0.11 \\ -0.02 \end{bmatrix} &
\begin{bmatrix} 0.13 \\ 0.04 \\ 0.02 \\ 0.83 \\ -0.29 \\ 0.40 \\ -0.08 \\ -0.07 \\ -0.17 \end{bmatrix} &
\begin{bmatrix} -0.03 \\ 0.80 \\ 0.10 \\ -0.22 \\ -0.52 \\ -0.04 \\ 0.03 \\ -0.17 \\ -0.02 \end{bmatrix} &
\begin{bmatrix} 0.30 \\ 0.31 \\ 0.43 \\ 0.38 \\ 0.14 \\ -0.44 \\ 0.30 \\ -0.42 \\ -0.01 \end{bmatrix} &
\begin{bmatrix} 0.06 \\ 0.24 \\ 0.04 \\ 0.18 \\ -0.65 \\ 0.23 \\ 0.35 \\ -0.14 \\ -0.54 \end{bmatrix} &
\begin{bmatrix} -0.30 \\ 0.62 \\ 0.18 \\ -0.39 \\ -0.32 \\ 0.10 \\ 0.24 \\ -0.05 \\ 0.41 \end{bmatrix} \\
\text{variances} & 155.2 & 30.7 & 15.6 & 4.01 & 1.64 & 1.13
\end{array}
$$

(i) Compare and contrast the analyses based on the covariance and correlation matrix and use this example to discuss the general principle of whether to standardise the data before a principal components analysis is performed.

(ii) Are we justified in considering only three out of the nine components?

4 (a) Let X be an $(n \times p)$ data matrix in which each row corresponds to a p-variate measurement on one of n individuals.

Assuming the p variates are continuous variables, describe three possible measures of dissimilarity of pairs of individuals. Comment on their relative advantages and disadvantages.

(b) What four properties must be satisfied for a dissimilarity function to be a metric dissimilarity function?

Let $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$, where $x_i, y_i \in \{0, 1\}$ for $i = 1, \ldots, p$. The simple matching coefficient is

$$S_1 = \frac{a+d}{p},$$

where $a = \sum_{i=1}^{p} \mathbb{1}_{\{x_i=1, y_i=1\}}$ and $d = \sum_{i=1}^{p} \mathbb{1}_{\{x_i=0, y_i=0\}}$. State for each of the following whether it is a metric dissimilarity function, and explain your reasoning:

(i) $d_1 = (1 - S_1)$

(ii) $d_2 = \frac{1}{(1+S_1)}$.

(c) The table below shows the values of the simple distance measure $d_{ij} = \max_{1 \leqslant r \leqslant 4} |x_{ir} - x_{jr}|$ for eight individuals, each with measurements on four variables. Use single link and complete link cluster analysis to place the individuals into three groups. Does three seem an appropriate number of groups?

	A	B	C	D	E	F	G	H
A	-	1.2	0.6	1.4	2.4	3.6	2.1	3.3
B		-	0.6	0.4	2.6	3.8	2.3	3.5
C			-	0.8	2.2	3.4	1.9	3.1
D				-	2.3	3.5	2.0	3.2
E					-	1.2	0.6	0.9
F						-	1.5	1.0
G							-	1.2
H								-

END OF PAPER