UNIVERSITY OF CAMBRIDGE

M. P*HIL. IN* STATISTICAL SCIENCE

Monday 12 June 2006 1.30 to 3.30

APPLIED MULTIVARIATE ANALYSIS

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

Marks for each question are indicated on the paper in square brackets. Each question is worth a total of 20 marks.

You may use the following results without proof.

Given $X \sim N_p(\boldsymbol{\mu}, \Sigma)$ *and* $(p \times q)$ *matrix* A *with* $q < p$, $A^T X \sim N_q(A^T \boldsymbol{\mu}, A^T \Sigma A)$*;*

and

$\int_R g(x)dx$ *is minimised with respect to* R *by* $R^* = \{x : g(x) < 0\}$*, for any function* g*.*

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) Given *iid* observations $Z_1, \ldots, Z_p \sim N(0,1)$ state how these can be used to obtain a single observation $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ for a given $(p \times 1)$ vector $\boldsymbol{\mu}$ and for positive definite non-singular $(p \times p)$ matrix Σ. Hence, or otherwise, derive the density function for the p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. [8]

(b) Suppose $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ is partitioned as $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ where $\mathbf{X}_1$ is $(q \times 1)$ with $q < p$ and the corresponding partitions of $\boldsymbol{\mu}$ and Σ are $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ where Σ_{11} and Σ_{22} are positive semi-definite symmetric matrices. Show that the marginal distribution of $\mathbf{X}_1$ is $N_q(\boldsymbol{\mu}_1, \Sigma_{11})$. [3]

(c) Suppose we have data $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N_p(\boldsymbol{\mu}_x, \Sigma)$ with Σ unknown. State the form of the best test statistic for testing $H_0 : \boldsymbol{\mu}_x = \boldsymbol{\mu}_0$ vs $H_1 : \boldsymbol{\mu}_x \neq \boldsymbol{\mu}_0$. What is the null distribution of this test statistic? [3]

Suppose $\mathbf{Y} = A\mathbf{X} + b$ where A is a $(p \times p)$ non-singular matrix and b a $(p \times 1)$-vector. Show that your test statistic above for testing $H_0 : \boldsymbol{\mu}_x = \boldsymbol{\mu}_0$ is identical to that used in testing $H_0 : \boldsymbol{\mu}_y = A\boldsymbol{\mu}_0 + \mathbf{b}$. What does this tell you about the properties of your test statistic with regards non-singular linear transformations of the variables? [6]

2 (a) A population is divided into two groups and the prior probability that an individual belongs to group i is π_i, $i = 1, 2$. A measurement, $\mathbf{x} \in \mathcal{X}$, is taken on each individual. The probability distribution of $\mathbf{x}$ for individuals from group i is p-variate normal with mean $\boldsymbol{\mu}_i$ and common covariance matrix Σ, for $i = 1, 2$.

Assuming that the cost of assigning a member of group 1 to group 2 is the same as the cost of assigning a member of group 2 to group 1, show that an appropriate decision rule for assigning an individual with measurement $\mathbf{x}$ to one of the two groups is given by: if $\mathbf{L}^T\mathbf{x} - \ 1/2\ \mathbf{L}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \leqslant \log \pi_2/\pi_1$ allocate to group 1 otherwise allocate to group 2, where $\mathbf{L} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. [8]

(b) An experiment was carried out to investigate the amount of time that beetles spend on different activities. Define $\mathbf{X}$ to be a bivariate vector measurement, X_1 being the amount of time spent eating and X_2 the amount of time spent at rest.

The experiment considered two equally-common separate species *fattus* and *apathus* and yielded sample means $\bar{\mathbf{x}}_f = \begin{pmatrix} 6 \\ 3 \end{pmatrix}$ and $\bar{\mathbf{x}}_a = \begin{pmatrix} 2 \\ 9 \end{pmatrix}$ for the *fattus* and *apathus* species respectively, with pooled covariance matrix

$$S = \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}.$$

(i) Use these data to derive a linear function that will discriminate between the two species of beetle. State any assumptions you make. [4]

(ii) By sketching a graph, allocate each of the following five beetles to one of the two species

$$\begin{pmatrix} 0 \\ 5 \end{pmatrix}, \begin{pmatrix} 0 \\ 3 \end{pmatrix}, \begin{pmatrix} 4 \\ 7 \end{pmatrix}, \begin{pmatrix} 5 \\ 6 \end{pmatrix}, \begin{pmatrix} 9 \\ 8 \end{pmatrix}. \quad [5]$$

(iii) If it is known that the individuals in (ii) come from a location where *fattus* are twice as common as *apathus*, how should your graph in (ii) change and how would you now classify the fifth individual in (ii) above? [3]

3 (a) What is the purpose of principal components analysis? [1]

Let $\mathbf{X}$ be a p-variate random variable with covariance matrix Σ. Derive the first principal component of $\mathbf{X}$. [5]

Define the remaining principal components of $\mathbf{X}$. [1]

(b) One hundred 13 year old children were assessed in Physics, Biology, Mathematics and Physical Education. Each child was given a mark out of 300 for each of the four disciplines.

Let $\mathbf{X}$ be the vector of marks obtained by an individual child, where X_1 corresponds to the mark in Physics, X_2 the mark in Biology and X_3, X_4 the marks in Mathematics and Physical Education respectively. The sample covariance is given by

$$\Sigma = \begin{bmatrix} 904 & 650 & 696 & 20 \\ 650 & 950 & 650 & 0 \\ 696 & 650 & 904 & -20 \\ 20 & 0 & -20 & 404 \end{bmatrix}$$

(i) *Verify* that one of the principal components has coefficients proportional to $(1, 1, 1, 0)$. [3]

(ii) Given that the remaining principal components are proportional to $(1, 0, -1, 10)$, $(1, -2, 1, 0)$ and $(10, 0, -10, -2)$ calculate the proportion of total variation attributable to each of the four components. [4]

(iii) Interpret the components where possible and say what conclusions you would draw from this analysis. Explain your answer. [4]

(iv) Suppose now that the Physics teacher decides to assess his students by giving them marks out of 1000 rather than marks out of 300, like his colleagues. Explain how this might affect the sample covariance matrix and hence, how you might consider modifying your method of analysing the data. [2]

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 that the p variates are continuous variables describe three possible measures of dissimilarity of pairs of individuals. Comment on their relative advantages and disadvantages. [3]

(b) What four properties must be satisfied for a dissimilarity function to be a metric dissimilarity coefficient? [2]

The values of four binary variables are measured for each of four individuals as follows:

Individual	Variable			
	1	2	3	4
1	1	1	1	0
2	0	0	1	1
3	1	1	1	1
4	0	1	0	1

Construct a dissimilarity matrix for the four individuals using (i) the simple matching coefficient and (ii) Jaccard's coefficient. [4]

If S_{rt} denotes the simple matching coefficient show that $d_{rt} = 1 - S_{rt}$ is a metric dissimilarity coefficient. [4]

(c) Five subjects were each given three psychological tests. The scores for each subject on each test were recorded and the Euclidean distances between each pair of subjects were calculated as follows:

	Subject				
	A	B	C	D	E
A	0	-	-	-	-
B	4.2	0	-	-	-
C	5.9	7.6	0	-	-
D	1.2	7.0	10.3	0	-
E	6.1	2.6	5.4	7.8	0

Using single-link clustering, cluster the five subjects. Sketch the dendrogram and interpret the results. [4]

How would your dendrogram change if you used a complete-link clustering algorithm? [3]

END OF PAPER