MATH577201

This question paper consists of 6
printed pages, each of which is
identified by the reference **MATH5772**.

Statistical tables are not provided.
Where percentage points are needed,
they are given in the question.
Only approved basic scientific
calculators may be used.

©UNIVERSITY OF LEEDS

Examination for the Module MATH5772
(January 2007)

Multivariate and Cluster Analysis

Time allowed: **3 hours**

Attempt not more than FOUR questions.
All questions carry equal marks.

1. (a) Let $\boldsymbol{x} \sim N_2(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{x} = [x_1, x_2]^T$ and $\Sigma = \begin{bmatrix} a & b \\ b & a \end{bmatrix}$ for positive real constants a, b.

Given a particular value of a, for what values of b is Σ a valid variance matrix?

Show that the unit eigenvectors of Σ are $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

Use the eigenvalues and eigenvectors of Σ to sketch the contours of the density of $\boldsymbol{x}$. In particular, give equations defining the major and minor axes of the contours in terms of x_1 and x_2 and indicate the relative lengths of the ellipse along the major and minor axes

(b) Show that the squared Mahalanobis distance of $\boldsymbol{x}$ from the origin is

$$d_M^2(\boldsymbol{x}, \mathbf{0}) = \frac{ax_1^2 - 2bx_1x_2 + ax_2^2}{a^2 - b^2}.$$

Consider points $\boldsymbol{x}_+ = [u, u]^T$ and $\boldsymbol{x}_- = [u, -u]^T$. Show that the squared Mahalanobis distance of $\boldsymbol{x}_+$ from the origin is a fixed multiple of the squared Mahalanobis distance of $\boldsymbol{x}_-$ from the origin, i.e. $d_M^2(\boldsymbol{x}_+, \mathbf{0}) = c d_M^2(\boldsymbol{x}_-, \mathbf{0})$ for all u, and find the constant c.

Also calculate the squared Euclidean distances of $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ from the origin, $d_E^2(\boldsymbol{x}_+, \mathbf{0})$ and $d_E^2(\boldsymbol{x}_-, \mathbf{0})$. Contrast the relationship between the Euclidean distances $d_E^2(\boldsymbol{x}_+, \mathbf{0})$ and $d_E^2(\boldsymbol{x}_-, \mathbf{0})$ to that between the Mahalanobis distances $d_M^2(\boldsymbol{x}_+, \mathbf{0})$ and $d_M^2(\boldsymbol{x}_-, \mathbf{0})$. When would the Mahalanobis and Euclidean distances be equal?

(c) Assume that $\boldsymbol{\mu} = [3, 1]^T$ and $\Sigma = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$. Let $\boldsymbol{y} = [y_1, y_2]^T$ where $y_1 = 2x_1 + x_2$ and $y_2 = x_1 - x_2$. Find a matrix A such that $\boldsymbol{y} = A\boldsymbol{x}$ and hence find the distribution of $\boldsymbol{y}$.

2. (a) Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be i.i.d. $N_p(\boldsymbol{\mu}_x, \Sigma)$ and $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ be i.i.d. $N_p(\boldsymbol{\mu}_y, \Sigma)$. Given that $\bar{\boldsymbol{x}} \sim N_p(\boldsymbol{\mu}_x, \Sigma/n)$ and $\bar{\boldsymbol{y}} \sim N_p(\boldsymbol{\mu}_y, \Sigma/m)$, use the moment generating functions of $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ to show that

$$\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}} \sim N_p\left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y, (n^{-1} + m^{-1})\Sigma\right).$$

Hint: You may use the fact that if $\boldsymbol{z} \sim N_p(\boldsymbol{\mu}, \Sigma)$ then the moment generating function of $\boldsymbol{z}$ is

$$M_{\boldsymbol{z}}(\boldsymbol{t}) = E(\exp\{\boldsymbol{t}^T \boldsymbol{x}\}) = \exp\{\boldsymbol{t}^T \boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T \Sigma \boldsymbol{t}\}.$$

(b) The weight (in grammes) and humerus length (in millimetres) were measured on 49 female sparrows who were injured in a storm. Of these $n = 21$ survived and $m = 28$ died. Denoting measurement on the survivors by $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{21}$ and on those which died by $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{28}$, the summary statistics are

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 24.6 \\ 18.5 \end{bmatrix} \quad \bar{\boldsymbol{y}} = \begin{bmatrix} 25.3 \\ 18.4 \end{bmatrix} \quad 20S_x = \begin{bmatrix} 22.8 & 4.4 \\ 4.4 & 3.4 \end{bmatrix} \quad 27S_y = \begin{bmatrix} 30.8 & 5.9 \\ 5.9 & 4.6 \end{bmatrix}.$$

Use Hotelling's T^2 test to test the null hypothesis that there is no difference between the two groups of sparrows.

(c) Calculate simultaneous 95% confidence intervals for the unit vectors $\boldsymbol{a}_1 = [1, 0]^T$ and $\boldsymbol{a}_2 = [0, 1]^T$. What do your simultaneous confidence intervals tell you about the differences between the two groups of sparrows?

Find a third vector $\boldsymbol{a}_3$ for which the corresponding simultaneous confidence interval will exclude zero and use this vector to help explain the differences between the two groups of sparrows (note that is it not necessary to construct the simultaneous confidence interval for $\boldsymbol{a}_3$).

Hints:

(i) *You may use the fact that the T^2 and F distributions are related by*

$$T^2(p, \nu) = \frac{\nu p}{\nu - p + 1} F(p, \nu - p + 1).$$

(ii) *Simultaneous confidence intervals for this problem are of the form*

$$\boldsymbol{a}^T(\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}}) \pm \sqrt{T^2(p, n+m-2, P\%)(n^{-1} + m^{-1})\boldsymbol{a}^T S \boldsymbol{a}},$$

where $T^2(p, n+m-2, P\%)$ is the upper $P\%$ point of the $T^2(p, n+m-2)$ distribution and S is a pooled estimate of Σ.

(iii) *You may find the percentage points of F distributions in the following R output helpful.*

```
> qf(0.05, df1=2, df2=c(20, 46, 49, 94), lower=F)
[1] 3.493 3.200 3.187 3.093
```

3. (a) Let $\boldsymbol{x}$ be a random vector with mean vector $\boldsymbol{\mu}$ and variance matrix Σ. Let $\boldsymbol{y}$ be the vector of principal components of $\boldsymbol{x}$.

Define $\boldsymbol{y}$ in terms of the eigenvalues and standardised eigenvectors of Σ.

If $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$, find the distribution of $\boldsymbol{y}$.

(b) If $\lambda, \boldsymbol{\gamma}$ are an eigenvalue and corresponding eigenvector of $X^T X$ then show that λ is also an eigenvalue of XX^T and find the corresponding eigenvector. What connection between principal components analysis and classical multi-dimensional scaling follows from this result?

(c) Scientists measured the carapace length (x_1), carapace width (x_2), mean leg length (x_3), and mean antenna length (x_4) in millimetres of 57 adult male stag beetles. The mean vector was $\bar{\boldsymbol{x}} = [25.3, 14.1, 8.9, 7.1]^T$. The sample variance matrix has eigenvalues $7.21, 3.14, 0.75$, and 0.22 with corresponding unit eigenvectors

$$\boldsymbol{\gamma}_{(1)} = \begin{bmatrix} 0.55 \\ 0.60 \\ 0.45 \\ 0.35 \end{bmatrix}, \quad \boldsymbol{\gamma}_{(2)} = \begin{bmatrix} 0.74 \\ -0.66 \\ 0.07 \\ 0.11 \end{bmatrix}, \quad \boldsymbol{\gamma}_{(3)} = \begin{bmatrix} 0.44 \\ 0.57 \\ -0.69 \\ 0.06 \end{bmatrix}, \text{ and } \boldsymbol{\gamma}_{(4)} = \begin{bmatrix} 0.64 \\ 0.57 \\ 0.07 \\ -0.50 \end{bmatrix}.$$

We can reduce the dimension of a data set by retaining only some of the principal components. By considering proportions of total variation and the average of the eigenvalues, suggest how many principal components should be retained for this data set.

Use the eigenvalues and eigenvectors above to interpret the principal components analysis of these data.

(d) Two beetles had observed data vectors

$$\boldsymbol{x}_1 = [28.1, 17.3, 9.7, 8.2]^T \text{ and } \boldsymbol{x}_2 = [27.9, 13.1, 9.1, 6.9]^T.$$

Calculate the values of the first two principal components for these beetles. What do the values tell you about how these beetles differ from a “typical” adult male stag beetle?

4. (a) Consider p-dimensional observations from populations Π_1 and Π_2. Assume that observations from population i have distribution $N_p(\boldsymbol{\mu}_i, \Sigma)$ with density function $f_i(\boldsymbol{x})$ for $i = 1, 2.$. Assume that the prior probability of an observation being from population i is π_i, $i = 1, 2$ with $\pi_1 + \pi_2 = 1$.

Define the Bayesian rule to allocate an observation $\boldsymbol{x}$ to population Π_1 in terms of the densities and prior probabilities of the populations. Show that the allocation rule for the normally distributed populations defined above is to allocate an observation $\boldsymbol{x}$ to Π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right) \geqslant c,$$

giving an explicit expression for c in terms of the prior probabilities. Interpret the probabilities π_1 and π_2 that give rise to $c = 0$.

(b) Thirty two skulls were collected at two sites in Tibet, believed to be from two different ethnic groups (17 skulls of type I and 15 of type II). The length (variable 1) and breadth (variable 2) of each skull were measured in millimetres. The mean vectors of types I and II are $\bar{\boldsymbol{x}}$ and $\bar{\boldsymbol{y}}$ respectively and the pooled variance matrix is S_p where

$$\bar{\boldsymbol{x}} = \begin{bmatrix} 175 \\ 140 \end{bmatrix} \quad \bar{\boldsymbol{y}} = \begin{bmatrix} 186 \\ 139 \end{bmatrix} \quad S_p = \begin{bmatrix} 59 & 9 \\ 9 & 48 \end{bmatrix}.$$

Find the Bayesian allocation rule to determine which population a new skull would be assigned to.

Assuming that $\pi_1 = \pi_2 = 1/2$, Sketch a diagram showing the contours of the densities and the boundary defining the allocation regions. To which type would you allocate a new skull with length 190mm and breadth 152mm?

For what value of π_1 would this skull lie on the line dividing $\mathbb{R}^2$ into regions corresponding to the two populations?

5. (a) Define single and complete linkage and explain how they are used in hierarchical agglomerative cluster analysis.

(b) Consider the points $x_1 = 0, x_2 = 1, x_3 = 9, x_4 = 11$ in $\mathbb{R}$; use Euclidean distance to measure the dissimilarity between points. Show that agglomerative clustering using both single and complete linkage gives the clustering sequence

$$x_1, x_2, x_3, x_4 \longrightarrow (x_1, x_2), x_3, x_4 \longrightarrow (x_1, x_2), (x_3, x_4) \longrightarrow (x_1, x_2, x_3, x_4),$$

where, for example, (x_1, x_2) indicates that points x_1 and x_2 have been combined to form a cluster. How will the dendrograms produced by single and complete linkage differ?

(c) Now add a fifth point y to the set of points. For each of single and complete linkage, find which possible values of y give the following clustering sequence

$$\begin{aligned} x_1, x_2, x_3, x_4, y &\longrightarrow (x_1, x_2), x_3, x_4, y \longrightarrow (x_1, x_2), (x_3, x_4), y \\ &\longrightarrow (x_1, x_2, y), (x_3, x_4) \longrightarrow (x_1, x_2, x_3, x_4, y). \end{aligned}$$

(d) Consider the points $\{x_1, x_2, x_3, x_4, y, z\}$ with $x_1, \ldots, x_4$ as above, $y = -4$, and $z = 14$. Construct a minimum spanning tree for this set of points. Use your MST to draw a dendrogram representing a single-linkage agglomerative clustering of these points.