MATH477201

This question paper consists of 5 printed pages, each of which is identified by the reference MATH477201.

New Cambridge Elementary Statistical Tables are provided. Only approved basic scientific calculators may be used.

Examination for the Module MATH4772
(January 2003)

MULTIVARIATE AND CLUSTER ANALYSIS

Time allowed: **3 hours**

Attempt not more than THREE questions of the first FOUR.
All candidates must attempt question FIVE.
All questions carry equal marks.

1. The $p \times 1$ random vector $\mathbf{x}$ has a multivariate normal distribution with probability density function

$$f(\mathbf{x}) = |2\pi\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \qquad \mathbf{x} \in \mathbb{R}^p,$$

and moment generating function

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

The matrix Σ is non-singular. We write $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$.

Let $\mathbf{x}$ be partitioned into p_1 and p_2 components, $p_1 + p_2 = p$, with corresponding partitions

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

(a) Consider $\mathbf{y} = \mathbf{x}_1 + M\mathbf{x}_2$, where M is a $p_1 \times p_2$ matrix. Show that

$$\mathrm{Cov}(\mathbf{y}, \mathbf{x}_2) = \Sigma_{12} + M\Sigma_{22}.$$

What value of M results in independence between $\mathbf{y}$ and $\mathbf{x}_2$?

When $\mathbf{y}$ and $\mathbf{x}_2$ are independent the conditional distribution of $\mathbf{y}|\mathbf{x}_2 = \mathbf{x}_2^o$ will be the same as the marginal distribution of $\mathbf{y}$. Use this to show that

$$\mathbf{x}_1|\mathbf{x}_2 = \mathbf{x}_2^o \sim N_{p_1}(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2^o - \boldsymbol{\mu}_2), \Sigma_{11.2}),$$

where $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

(b) Suppose

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 4 \end{pmatrix}\right).$$

Construct a new random vector $\mathbf{z} = (z_1, z_2)^T$ with $z_1 = 2x_1 + x_2$, $z_2 = x_1 - x_2$. Find the variance matrix and correlation matrix of $\mathbf{z}$.

 CONTINUED...

2. (a) Suppose $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is a random sample from $N_p(\boldsymbol{\mu}, \Sigma)$ population (where Σ, assumed positive definite, is unknown). Obtain the log likelihood function $l(\boldsymbol{\mu}, \Sigma)$. Show that the maximum likelihood estimator for $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$.

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from $N_p(\boldsymbol{\mu}_x, \Sigma)$ and $\mathbf{y}_1, \ldots, \mathbf{y}_m$ be a random sample from $N_p(\boldsymbol{\mu}_y, \Sigma)$. The two samples are independent of one another. Show that the union intersection test of the hypothesis $H_0 : \boldsymbol{\mu}_x = \boldsymbol{\mu}_y$ vs. $H_1 : \boldsymbol{\mu}_x \neq \boldsymbol{\mu}_y$, where Σ is unknown, leads to the test statistic

$$T^2 = \frac{nm}{n+m}(\bar{\mathbf{x}} - \bar{\mathbf{y}})^T S_p^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}),$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are sample mean vectors and S_p is the pooled within groups estimate of Σ.

(b) If H_0 is rejected in the overall test, how can simultaneous confidence intervals be used to give insight into the reasons for rejection?

(c) In a study to compare male gorilla skulls with female skulls the following variables were measured:

variable 1 = braincase length, variable 2= braincase height.

A sample of 11 males and 11 females yielded the following summary statistics:

$$\bar{\mathbf{x}} = \begin{pmatrix} 152 \\ 103 \end{pmatrix}, \qquad S_x = \begin{pmatrix} 40 & 10 \\ 10 & 25 \end{pmatrix},$$

$$\bar{\mathbf{y}} = \begin{pmatrix} 142 \\ 101 \end{pmatrix}, \qquad S_y = \begin{pmatrix} 32 & 4 \\ 4 & 17 \end{pmatrix}.$$

The pooled covariance matrix for the two samples and its inverse are given by

$$S_p = \begin{pmatrix} 36 & 7 \\ 7 & 21 \end{pmatrix}, \qquad S_p^{-1} = \begin{pmatrix} 0.03 & -0.01 \\ -0.01 & 0.05 \end{pmatrix}.$$

Explain how S_p is calculated. Compare the sexes on the basis of the information provided.

[Hints:

1. You may use the fact that the Hotelling T^2 and F distribution are related by $T^2(p, \nu) = \{\nu p/(\nu - p + 1)\} F(p, \nu - p + 1)$.
2. Simultaneous 100α percent confidence intervals for this problem can be written in the form

$$(\mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) - c, \mathbf{a}^T(\bar{\mathbf{x}} - \bar{\mathbf{y}}) + c)$$

where $c = \left\{T_\alpha^2(p, \nu)\frac{n+m}{nm}\mathbf{a}^T S_p \mathbf{a}\right\}^{\frac{1}{2}}$ and $T_\alpha^2(p, \nu)$ is the 100α percentage point of the $T^2(p, \nu)$ distribution.]

CONTINUED...

3. (a) Let $\mathbf{x}$ be a p-dimensional random vector with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

(i) Define the principal components $\mathbf{y}$ of $\mathbf{x}$ in terms of the standardized eigenvectors of Σ.

(ii) Obtain the variance-covariance matrix of the principal components $\mathbf{y}$.

(iii) If $\Sigma = \boldsymbol{\alpha}\boldsymbol{\alpha}^T$ for some vector $\boldsymbol{\alpha}$, find the first principal component. What can you say about the other principal components?

(iv) Why do we restrict ourselves to standardized linear combinations when carrying out principal component analysis?

(b) Data were collected on 50 irises from the species *Iris setosa*. The variables are: x_1=sepal length; x_2=sepal width; x_3=petal length; x_4=petal width. The sample mean vector and correlation matrix were

$$\bar{\mathbf{x}} = \begin{pmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{pmatrix}, \qquad R = \begin{pmatrix} 1 & 0.74 & 0.27 & 0.28 \\ 0.74 & 1 & 0.18 & 0.23 \\ 0.27 & 0.18 & 1 & 0.33 \\ 0.28 & 0.23 & 0.33 & 1 \end{pmatrix}.$$

Principal component analysis gave eigenvalues 2.06, 1.02, 0.67, 0.25. The corresponding eigenvectors were the columns of

$$\begin{pmatrix} 0.60 & -0.33 & 0.07 & 0.72 \\ 0.58 & -0.44 & 0.00 & -0.69 \\ 0.38 & 0.63 & 0.68 & -0.09 \\ 0.40 & 0.55 & -0.73 & -0.01 \end{pmatrix}.$$

Interpret these principal components briefly. Assess their relative contribution to total variation. What methods can be used to decide on the number of components we should retain?

CONTINUED....

4. (a) Let $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^p$, denote two probability density functions for populations Π_1 and Π_2 respectively, with prior probabilities π_1 and π_2, where $\pi_1 + \pi_2 = 1$. Consider the allocation rule which assigns $\mathbf{x}$ to Π_1 if

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{\pi_2}{\pi_1}$$

and to Π_2 otherwise. Show that this rule is admissible.

(b) If the two populations have $N_p(\boldsymbol{\mu}_1, \Sigma)$ and $N_p(\boldsymbol{\mu}_2, \Sigma)$ distributions and $\pi_1 = 2\pi_2$, show that this rule classifies $\mathbf{x}$ as coming from the first population if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\Sigma^{-1}\mathbf{x} \geq c$$

and to the second population otherwise, for a suitable constant c. What is the value of c?

(c) Let

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 5 \end{pmatrix}, \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 5 \\ 1 \end{pmatrix}.$$

Calculate and sketch the boundary between the two classification regions for $\pi_1 = 2\pi_2$ and $\pi_1 = \pi_2$ respectively. Compare these boundaries. How would you classify $\mathbf{x} = (3, 2)^T$ under each rule?

(d) If the population parameters have to be replaced by sample estimates, what problem arises in estimating the misclassification probabilities and how might this be overcome?

5. (a) Let a set of centred coordinates of n points in a p dimensional Euclidean space be given by $\mathbf{x}_r = (x_{r1}, \ldots, x_{rp})^T$, so that $\sum_{r=1}^n x_{rj} = 0$, $j = 1, \ldots, p$. The Euclidean distance between the rth and sth points is given by

$$d_{rs}^2 = (\mathbf{x}_r - \mathbf{x}_s)^T(\mathbf{x}_r - \mathbf{x}_s).$$

Let $B = (b_{rs})$ be the positive semi definite inner product matrix

$$B = XX^T$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the $n \times p$ matrix of coordinates.

Show that the elements of B are determined from the Euclidean distances by

$$b_{rs} = a_{rs} - \bar{a}_{r.} - \bar{a}_{.s} + \bar{a}_{..} \qquad (1)$$

where $a_{rs} = -\frac{1}{2}d_{rs}^2$, $\bar{a}_{r.} = \frac{1}{n}\sum_{s=1}^n a_{rs}$, $\bar{a}_{.s} = \frac{1}{n}\sum_{r=1}^n a_{rs}$ and $\bar{a}_{..} = \frac{1}{n^2}\sum_{r=1}^n \sum_{s=1}^n a_{rs}$.

(b) Assuming only B is known, show how to find X from the spectral decomposition of B. Why is X determined only up to arbitrary location, rotation and reflection? (Note: if the rank of B is p, then B has $n - p$ zero eigenvalues).

(c) Estimated car travel times between the 7 cities Hull, Leeds, Liverpool, London, Manchester, Sheffield and York were calculated by the RAC and are given in the matrix D,

$$D = \begin{array}{l} \text{Hull} \\ \text{Leeds} \\ \text{Liverpool} \\ \text{London} \\ \text{Manchester} \\ \text{Sheffield} \\ \text{York} \end{array} \begin{bmatrix} 0 & 73 & 144 & 273 & 112 & 82 & 61 \\ 73 & 0 & 84 & 209 & 52 & 43 & 45 \\ 144 & 84 & 0 & 231 & 43 & 102 & 120 \\ 273 & 209 & 231 & 0 & 218 & 183 & 238 \\ 112 & 52 & 43 & 218 & 0 & 65 & 86 \\ 82 & 43 & 102 & 183 & 65 & 0 & 74 \\ 61 & 45 & 120 & 238 & 86 & 74 & 0 \end{bmatrix}$$

Consider the estimated car travel time to be a measure of distance between cities. We wish to construct a two dimensional map of the 7 cities from D.

A matrix B was calculated from D using equation (1). Since D is not a distance matrix arising from a two dimensional space, B is not positive semi definite. The eigenvalues of B are 45821, 12401, 1983, 407, 0 -56, -1175. The corresponding standardized eigenvectors are the columns of

$$\begin{pmatrix} 0.38 & -0.45 & 0.39 & -0.15 & 0.38 & -0.01 & -0.58 \\ 0.10 & -0.05 & -0.18 & -0.53 & 0.38 & -0.62 & 0.38 \\ 0.10 & 0.72 & 0.03 & -0.37 & 0.38 & 0.41 & -0.14 \\ -0.88 & -0.14 & -0.04 & -0.04 & 0.38 & 0.00 & -0.24 \\ 0.10 & 0.38 & 0.05 & 0.69 & 0.38 & -0.46 & -0.12 \\ -0.02 & -0.17 & 0.50 & 0.19 & 0.38 & 0.33 & 0.66 \\ 0.22 & -0.28 & -0.75 & 0.21 & 0.38 & 0.35 & 0.05 \end{pmatrix}.$$

(i) Describe how to construct a map from B in $p' = 2$ dimensions.

(ii) Using an appropriate measure for the proportion of explained variation compare the choice $p' = 2$ with other choices for p'.

(iii) Obtain the distance between Leeds and London for the map you have described in (i). Compare this value with the original RAC estimate given in D and comment on your findings.

END