MATH377201

This question paper consists of 3
printed pages, each of which is
identified by the reference **MATH3772**.

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

Examination for the Module MATH3772
(May - June 1999)

MULTIVARIATE ANALYSIS

Time allowed: **2 hours**

All four questions may be attempted, but only the best three answers will be taken into account.
Greater credit will be given to complete answers.
All questions carry equal marks.

1. (a) Let $\mathbf{x}$ follow a multivariate normal distribution, $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Describe the contours of constant probability density for $\mathbf{x}$. How can the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ be used to help plot these contours? Give a sketch of these contours for the case

$$\boldsymbol{\mu} = (3,1)^T, \quad \Sigma = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}.$$

(b) For general $\boldsymbol{\mu}$ and Σ, show that $(\mathbf{x}-\boldsymbol{\mu})^T \ \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) \sim \chi^2_p$. How can this result be used in practice to help identify outliers in a dataset?

(c) If $\boldsymbol{\mu} = (3,1)^T$ and $\Sigma = \begin{bmatrix} 3 & -2 \\ -2 & 3 \end{bmatrix}$ as above, explain why $\frac{1}{42}(x_1 - 3x_2)^2 \ \sim \chi^2_1$.

2. (a) Suppose measurements of p variables are made on n_1 individuals from one group $\mathbf{x}_{1,1,} \ldots, \mathbf{x}_{1,n_1}$ and on n_2 individuals from another group, $\mathbf{x}_{2,1,} \ldots, \mathbf{x}_{2,n_2}$. Describe how the union-intersection approach to compare the means of the two groups leads to the statistic

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \ (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^T \ \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2),$$

where $\overline{\mathbf{x}}_1 = n_1^{-1} \displaystyle\sum_{j=1}^{n_1} \mathbf{x}_{1,j}, \ \ \overline{\mathbf{x}}_2 = n_2^{-1} \sum_{k=1}^{n_2} \ \mathbf{x}_{2,k}$

and $S = \frac{1}{n_1+n_2-2} \left\{ \displaystyle\sum_{j=1}^{n} \ \mathbf{x}_{1,j} \ \mathbf{x}_{1,j}^T + \sum_{k=1}^{n_2} \ \mathbf{x}_{2,k} \ \mathbf{x}_{2,k}^T - n_1 \ \overline{\mathbf{x}}_1 \overline{\mathbf{x}}_1^T - n_2 \ \overline{\mathbf{x}}_2 \overline{\mathbf{x}}_2^T \right\}$

Describe the assumptions under which this statistic T^2 will follow Hotelling's T^2 distribution, and give the asymptotic distribution in this case when $n_1 \to \infty$ with n_2 fixed.

(b) In an archaeological study to compare male Egyptian skulls from two epochs, the following two variables were measured

x_1 = maximum breadth (in mm), $\quad x_2$ = nasal height (in mm).

Samples of size 11 were taken from the Early predynastic epoch and the Roman epoch, respectively. The following mean vector and covariance matrix were obtained for each epoch.

$$\overline{\mathbf{x}}_1 = \begin{bmatrix} 130.9 \\ 49.9 \end{bmatrix}, \quad S_1 = \begin{bmatrix} 40 & 16 \\ 16 & 11 \end{bmatrix},$$

$$\overline{\mathbf{x}}_2 = \begin{bmatrix} 132.8 \\ 51.9 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 26 & 2 \\ 2 & 15 \end{bmatrix}.$$

Compare the two epochs on the basis of the information provided here.

[Hints:

1. You may use the fact that the Hotelling T^2 and F distribution are related by $T^2(p, m) = \{mp/(m - p + 1)\} \ \ F(p, m - p + 1)$.
2. Simultaneous $100\alpha\%$ confidence intervals for this problem can be written in the form
$$(\mathbf{a}^T(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) - c, \ \ \mathbf{a}^T(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2) + c)$$
where $c = \{T_\alpha^2(p, \nu) \ \dfrac{n_1 + n_2}{n_1 n_2} \ \mathbf{a}^T S \mathbf{a}\}^{\frac{1}{2}}$ and $T_\alpha^2(p, \nu)$ is the 100α percentage point of the $T^2(p, \nu)$ distribution.]

3. (a) Let $\mathbf{x}(p \times 1)$ be a random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Let $\Sigma = \Gamma\Lambda\Gamma^T$ be the spectral decomposition of Σ, where Γ is an orthogonal matrix and Λ is diagonal. Define the principal components of $\mathbf{x}$ and show that $\text{var}(\mathbf{a}^T\mathbf{x})$, where $\mathbf{a}$ is a vector of coefficients satisfying $\mathbf{a}^T\mathbf{a} = 1$, is maximized when $\mathbf{a}^T\mathbf{x}$ is the first principal component. What is the purpose of principal component analysis in the exploratory analysis of multivariate data?

(b) Suppose an observation $\mathbf{z}(p \times 1)$ can be written as "signal" plus "noise",

$$\mathbf{z} = \mathbf{x} + \mathbf{u}$$

where $\mathbf{x}$ has covariance matrix Σ as above, and $\mathbf{u}$ has covariance matrix $\tau^2 I$, proportional to the identity matrix, with $\mathbf{x}$ and $\mathbf{u}$ uncorrelated. Find the covariance matrix of $\mathbf{z}$ and describe the relationship between the principal components of $\mathbf{z}$ and those of $\mathbf{x}$. Find a vector of coefficients $\mathbf{a}(p \times 1)$ to minimize the ratio of the error variance to total variance, $\text{var}(\mathbf{a}^T\mathbf{u})/\text{var}(\mathbf{a}^T\mathbf{z})$.

4. (a) Suppose an observation $\mathbf{x} \in \mathbb{R}^p$ can come from one of two populations, with probability density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, respectively, and with prior probabilities π_1 and π_2, where $\pi_1 + \pi_2 = 1$. Assuming equal costs of misclassification, derive the classification rule which minimizes the expected cost of misclassification.

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

$$\mathbf{d}^T\Sigma^{-1}\mathbf{x} \geq \mathbf{d}^T\Sigma^{-1}\overline{\boldsymbol{\mu}},$$

where $\mathbf{d} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and $\overline{\boldsymbol{\mu}} = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$, and to the second population otherwise.

(c) Let

$$\Sigma = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix} \text{ and } \boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} 4 \\ 3 \end{bmatrix}.$$

Calculate and sketch the boundary between the two classification regions. How would you classify $\mathbf{x} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$?

(d) For this example, what is the probability of misclassification for individuals from population 2?

END