MATH377209

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.

©**UNIVERSITY OF LEEDS**

Examination for the Module MATH3772
(May/June 2000)

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. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ is a random sample from a $N_p(\boldsymbol{\mu}, \Sigma)$ population. Show that the union-intersection test of the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ vs $H_A : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ leads to the test statistic

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

where $\bar{\mathbf{x}}$ is the sample mean, and S is the sample covariance matrix.

When H_0 is true, $\frac{(N-p)}{p(N-1)} T^2$ has an F distribution with p and $(N-p)$ degrees of freedom.

How can this testing approach be adapted to the case $p = 4$ with $\boldsymbol{\mu}^T = (\mu_1, \mu_2, \mu_3, \mu_4)$ to test $H_0 : \mu_1 = \mu_3, \ \ \mu_2 = \mu_4$?

In a study of hearing among 82 boxers, four variables were recorded all at intensity 110 decibel sound pressure level:

$$\begin{aligned} X_1 &= \text{Right Ear Wave 1 time,} \\ X_2 &= \text{Right Ear Wave 5 time,} \\ X_3 &= \text{Left Ear Wave 1 time,} \\ X_4 &= \text{Left Ear Wave 5 time.} \end{aligned}$$

Q.1 Continued

The sample mean is given by $\bar{\mathbf{x}}^T = (1.648, 5.634, 1.625, 5.672)$ and the sample covariance matrix is given by $S = \begin{pmatrix} 0.023 & 0.011 & 0.009 & 0.008 \\ 0.011 & 0.072 & 0.008 & 0.003 \\ 0.009 & 0.008 & 0.010 & 0.007 \\ 0.008 & 0.003 & 0.007 & 0.050 \end{pmatrix}$.

What can you say about the difference between the means of the right and left ears?

2. (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 Σ be used to help plot these contours? Give a sketch of these contours for the case

$$\boldsymbol{\mu} = (4, 1)^T, \quad \Sigma = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}.$$

(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 $\boldsymbol{\mu} = (4, 1)^T$ and $\Sigma = \begin{bmatrix} 5 & -2 \\ -2 & 5 \end{bmatrix}$ as above, explain why $\frac{21}{69}(x_1 - 4x_2)^2 \ \sim \chi^2_1$.

3. (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

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

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

(c) Let

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

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?

4. Define the principal components of p random variables $x_1, x_2, \ldots, x_p$.

Show that the first two principal components are obtained from the standardized eigenvectors corresponding to the largest two eigenvalues of the covariance matrix Σ of $x_1, x_2, \ldots, x_p$.

Show that principal components are not scale invariant and discuss the implications of this as regards using the covariance or correlation matrix in a principal component analysis.

Obtain the three principal components when the covariance matrix takes the special form $\Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

END