NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER II EXAMINATION 2022-2023

MH4501 – MULTIVARIATE ANALYSIS

Apr 2023 Time Allowed: 2 hours

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FIVE (5)** questions and comprises **FOUR (4)** printed pages.
2. Answer **ALL** questions. The marks for each question are indicated at the beginning of each question.
3. Answer each question beginning on a **FRESH** page of the answer book.
4. This is a **RESTRICTED OPEN BOOK** exam. You are only allowed to bring in **ONE DOUBLE-SIDED A4-SIZE REFERENCE SHEET WITH TEXTS HANDWRITTEN OR TYPED ON THE A4 PAPER** (no sticky notes/post-it notes on the reference sheet).
5. Calculators may be used. However, you should write down systematically the steps in the workings.

QUESTION 1. (20 marks)

8 Hong Kong and 8 Singapore investment companies are selected in a comparative study. The data consist of their profits in the first three quarters in 2022. Let $\mu_g = (\mu_{g1}, \mu_{g2}, \mu_{g3})^{\top}$ be the mean vector of the profits in the three quarters for Group $g = 1,2$ (Group 1 is Hong Kong companies; Group 2 is Singapore companies). Suppose that the data of Group g are independent and identically distributed from $N_3(\mu_g, \Sigma)$ for $g = 1,2$ and data from different groups are independent. The sample mean vectors ($\bar{x}_1$ and $\bar{x}_2$) and the pooled sample covariance matrix (S_{pool}) are given as follows:

$$\bar{x}_1 = \begin{pmatrix} \bar{x}_{11} \\ \bar{x}_{12} \\ \bar{x}_{13} \end{pmatrix} = \begin{pmatrix} 70.5 \\ 80.5 \\ 81.0 \end{pmatrix}, \quad \bar{x}_2 = \begin{pmatrix} \bar{x}_{21} \\ \bar{x}_{22} \\ \bar{x}_{23} \end{pmatrix} = \begin{pmatrix} 75.875 \\ 86.125 \\ 86.125 \end{pmatrix}, \quad S_{pool} = \begin{pmatrix} 25.2 & 13.9 & 9.9 \\ 13.9 & 15.6 & 9.6 \\ 9.9 & 9.6 & 15.1 \end{pmatrix}.$$

(a) We consider a parallel profile analysis, i.e., a hypothesis testing for

$$H_0: \begin{pmatrix} \mu_{11} - \mu_{12} \\ \mu_{12} - \mu_{13} \end{pmatrix} = \begin{pmatrix} \mu_{21} - \mu_{22} \\ \mu_{22} - \mu_{23} \end{pmatrix}.$$

Compute the test statistic and draw your conclusion at the 5% significance level.

[*Hint: the null hypothesis can be written as* $H_0: C(\mu_1 - \mu_2) = 0$ *for some* C]

[*The quantiles you may use are:* $F_{0.05}[2,5] = 5.7861, F_{0.05}[2,13] = 3.8056,$ $F_{0.05}[3,4] = 6.5914, F_{0.05}[3,12] = 3.4903.$]

(b) What is the distribution of $\bar{x}_{11} + \bar{x}_{12} + \bar{x}_{13} - \bar{x}_{21} - \bar{x}_{22} - \bar{x}_{23}$?

QUESTION 2. (16 marks)

Suppose that we have five observations, for which we computed a distance matrix as

$$\begin{pmatrix} 0 & 1.5 & 2 & 1.4 & 2.5 \\ 1.5 & 0 & 1.7 & 2.8 & 1.1 \\ 2 & 1.7 & 0 & 2.9 & 2.2 \\ 1.4 & 2.8 & 2.9 & 0 & 3.9 \\ 2.5 & 1.1 & 2.2 & 3.9 & 0 \end{pmatrix}$$

On the basis of this distance matrix, sketch the dendogram that results from hierarchically clustering these 5 observations using **average linkage**. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

QUESTION 3. (20 marks)

Let Π_1 and Π_2 represent two populations that are respectively distributed according to $N_2(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$, where $\mu_1 \neq \mu_2$ and $\Sigma_1 \neq \Sigma_2$. Let $\bar{x}_g$ and S_g be the sample mean vector and sample covariance matrix of n_g observations from $\Pi_g, g = 1,2$. Suppose that

$$n_1 = 50, \quad \bar{x}_1 = \begin{pmatrix} 20 \\ 30 \end{pmatrix}, \quad S_1 = \begin{pmatrix} 16 & 0 \\ 0 & 9 \end{pmatrix};$$
$$n_2 = 60, \quad \bar{x}_2 = \begin{pmatrix} 6 \\ 5 \end{pmatrix}, \quad S_2 = \begin{pmatrix} 6 & 3 \\ 3 & 4 \end{pmatrix};$$

the probability that an individual is coming from Π_1 is twice the probability that an individual is coming from Π_2; and the misclassification costs are the same for both types.

(a) Write down the classification rule for classifying a new data point $x^* = (x_1^*, x_2^*)^\top$.

(b) Classify a new data point $x^* = (18, 20)^\top$ into Π_1 or Π_2.

QUESTION 4. (20 marks)

The eigenvalues and eigenvectors of the sample **correlation** matrix R of six variables from a sample of size $N = 200$ are given as follows:

$\lambda_1 = 3.649,$	$u_1 = (0.361, 0.314, 0.443, 0.452, 0.443, 0.417)^\top,$
$\lambda_2 = 0.954,$	$u_2 = (-0.523, -0.592, -0.153, 0.409, 0.358, 0.239)^\top.$

(a) Using the principal component solution method, calculate the estimates of the factor loading matrix $\hat{L}$ and matrix of specific variances $\hat{\Psi}$ in the following factor analysis model with two common factors: $R \approx \hat{L}\hat{L}^\top + \hat{\Psi}$.

(b) Compute the estimates of the communalities.

(c) Given the standard deviations of these six variables:

$$(0.374, 0.105, 0.493, 1.502, 4.5, 6.439)$$

If a two-factor model is considered for the sample **covariance** matrix S of these six variables, i.e., $S \approx \tilde{L}\tilde{L}^\top + \tilde{\Psi}$, calculate the estimates of the factor loading matrix $\tilde{L}$ and matrix of specific variances $\tilde{\Psi}$.

(d) Briefly explain the differences between principal component analysis and factor analysis.

QUESTION 5. **(24 marks)**

Suppose that a random vector $X = (X_1, X_2, X_3)^\top$ comes from a distribution with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}.$$

(a) Show that the following linear combination form is a principal component for Σ:

$$Y = \frac{1}{\sqrt{3}}(X_1 + X_2 + X_3).$$

(b) What proportion of the total population variation is explained by Y in part (a)?

In the questions below, we also know that Σ has two repeated eigenvalues of $1 - \rho$.

(c) Give the range of ρ such that Σ is positive-definite.

(d) Give the range of ρ such that Y in part (a) is the first principal component for Σ.

(e) In the case that Σ is positive-definite and Y in part (a) is the first principal component for Σ, determine the value of ρ such that the first principal component for Σ explains 80% of total population variation.

(f) In the case that Σ is positive-definite, determine the value of ρ such that $\mathrm{Corr}(Y, X_1) = \sqrt{2/3}$.

END OF PAPER

MH4501 MULTIVARIATE ANALYSIS

Please read the following instructions carefully:

1. **Please do not turn over the question paper until you are told to do so. Disciplinary action may be taken against you if you do so.**

2. You are not allowed to leave the examination hall unless accompanied by an invigilator. You may raise your hand if you need to communicate with the invigilator.

3. Please write your Matriculation Number on the front of the answer book.

4. Please indicate clearly in the answer book (at the appropriate place) if you are continuing the answer to a question elsewhere in the book.