NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER II EXAMINATION 2024–2025

MH4501 – MULTIVARIATE ANALYSIS

April/May 2025 Time Allowed: 2 hours

INSTRUCTIONS TO CANDIDATES

1. This examination paper contains **FIVE (5)** questions and comprises **SIX (6)** printed pages.
2. Answer each question beginning on a **FRESH** page of the answer book.
3. 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 WITHOUT ANY ATTACHMENTS** (e.g. sticky notes, post-it notes, gluing or stapling of additional papers).
4. Calculators may be used. However, you should write down systematically the steps in the workings.

QUESTION 1 **(20 marks)**

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

$$\begin{aligned}
\lambda_1 &= 1.676 \quad u_1 = \begin{pmatrix} 0.710 & 0.703 & 0.024 & 0.031 \end{pmatrix}^\top, \\
\lambda_2 &= 1.145 \quad u_2 = \begin{pmatrix} 0.186 & -0.237 & 0.837 & 0.457 \end{pmatrix}^\top, \\
\lambda_3 &= ? \quad u_3 = \begin{pmatrix} -0.140 & 0.117 & -0.421 & 0.889 \end{pmatrix}^\top, \\
\lambda_4 &= 0.217 \quad u_4 = \begin{pmatrix} 0.664 & -0.661 & -0.349 & 0.026 \end{pmatrix}^\top,
\end{aligned}$$

where the information about λ_3 is lost.

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

(b) Compute the estimates of the communalities.

(c) In order to obtain at least 80% for the cumulative proportion of total standardized sample variance explained by the common factors, how many common factors should we consider?

(d) Suppose that we rotate the factor loadings using varimax and obtain the following rotated factor loadings with missing numbers:

- On the first rotated common factor F_1^*: $\begin{pmatrix} 0.911 & ? & -0.002 & ? \end{pmatrix}^\top$
- On the second rotated common factor F_2^*: $\begin{pmatrix} ? & -0.220 & ? & 0.490 \end{pmatrix}^\top$

We only know that all the missing numbers are positive. What are the values of these four missing numbers?

See next page.

QUESTION 2 **(20 marks)**

Suppose that $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^\top$ is a random vector with the covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix},$$

i.e., diagonal elements are ones; off-diagonal elements are equal to $\rho > 0$. Denote by I the p-dimensional identity matrix and by $J = 1_p 1_p^\top$, where $1_p = (1, \ldots, 1)^\top$ is the p-dimensional vector of ones. It is known that J has $p - 1$ eigenvalues, which are equal to zero and one non-zero eigenvalue, which is equal to p.

(a) We start with a general result: Let A be a square $p \times p$ matrix and b be a scalar. Prove that λ is an eigenvalue of A if and only if $\lambda + b$ is an eigenvalue of $A + bI$.

(b) Find the scalars c and d such that the covariance matrix Σ can be written as $\Sigma = cI + dJ$.

(c) What proportion of the total population variation is explained by the first principal component for the covariance matrix Σ?

(d) Show that the first principal component for the covariance matrix Σ is

$$\boldsymbol{y}_1 = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} \boldsymbol{x}_i.$$

(e) Find the correlation coefficient between the first principal component $\boldsymbol{y}_1$ and the variable $\boldsymbol{x}_1$.

QUESTION 3 **(20 marks)**

We consider multivariate analysis of variance (MANOVA) throughout this question.

(a) Suppose there are 2 variables and 3 groups in the dataset. Below are the sample statistics table as well as the MANOVA table. However, there are some missing values in the tables.

	Group 1	Group 2	Group 3	Overall
Sample size	$n_1 = 8$	$n_2 = 8$	n_3	n
Mean	$\bar{x}_1$	$\bar{x}_2 = \begin{pmatrix} 1.1 \\ 5.0 \end{pmatrix}$	$\bar{x}_3 = \begin{pmatrix} 1.3 \\ 0.3 \end{pmatrix}$	$\bar{x} = \begin{pmatrix} 1.1 \\ 2.4 \end{pmatrix}$
Covariance matrix	$S_1 = \begin{pmatrix} 3.8 & 0.9 \\ 0.9 & 0.5 \end{pmatrix}$	S_2	$S_3 = \begin{pmatrix} 3.5 & 1.7 \\ 1.7 & 1.1 \end{pmatrix}$	S_{pool}

	Sum of squares and cross products (SSCP)	Degrees of freedom
Treatment	B	df_B
Residual	$W = \begin{pmatrix} 79.1 & 31.1 \\ 31.1 & 24.2 \end{pmatrix}$	$df_W = 22$
Total	$B + W$	df_{B+W}

i. Provide the values of n_3, n, $\bar{x}_1$, S_2, S_{pool}, B, df_B, $B + W$, and df_{B+W} which are missing in the two tables above (with 1 decimal place).

ii. Find the observed value of the Wilk's lambda statistic λ^* (with 2 decimal places).

(b) Assume another MANOVA study, with the same number of variables, the same number of groups, and the same sample size for each group, as in Question 3.(a). Suppose this second study results in an observed value of Wilk's lambda statistic $\tilde{\lambda}^* = 0.29$. Which is more likely to result in the rejection of H_0: $\mu_1 = \mu_2 = \mu_3$, λ^* or $\tilde{\lambda}^*$? Justify your answer.

(c) Consider general SSCP matrices $\boldsymbol{B}$ and $\boldsymbol{W}$ such that $\boldsymbol{W}$ and $\boldsymbol{B} + \boldsymbol{W}$ are invertible but $\boldsymbol{B}$ is not necessarily invertible, and denote by $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p$ the eigenvalues of $\boldsymbol{W}^{-1}\boldsymbol{B}$. Express the corresponding Wilks' lambda statistic $\boldsymbol{\lambda}^*$ as a function the eigenvalues $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p$ only.

See next page.

QUESTION 4 **(20 marks)**

Assume two bivariate normal populations π_1 and π_2, with different population mean vectors μ_1 and μ_2 but the same population covariance matrix Σ. A set of observations is collected from each population, thus μ_1 and μ_2 are estimated by

$$\bar{x}_1 = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \quad \text{and} \quad \bar{x}_2 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

respectively. Meanwhile, Σ is estimated by

$$S_{pool} = \begin{pmatrix} 7 & -1 \\ -1 & 5 \end{pmatrix}.$$

(a) Find S_{pool}^{-1}.

(b) Find the linear discriminant function $d_{12}(x)$ (for convenience, denote $x = (x_1, x_2)^T$).

(c) i. Assume equal prior probabilities $p_1 = p_2$, and equal misclassification costs $c(1|2) = c(2|1)$. Use the minimum ECM rule to classify $x_0 = (0, 1)^\top$.

ii. Suppose assumptions on prior probabilities and misclassification costs are

$$p_1 = \frac{1}{7} p_2 \quad \text{and} \quad c(1|2) = 10c(2|1).$$

Once again, use the minimum ECM rule to classify $x_0 = (0, 1)^\top$.

See next page.

QUESTION 5 **(20 marks)**

The table below provides a data set containing 8 observations with only 1 feature:

Obs.	#1	#2	#3	#4	#5	#6	#7	#8
X	1.5	2.5	2.0	3.5	6.5	5.5	9.0	3.5

We wish to identify three clusters of this data set using K-means clustering with $K = 3$. We use the Euclidean distance measure. Suppose that we initially assign the observations #1, #2 as Cluster 1, the observations #3, #4, #5 as Cluster 2, and #6, #7, #8 as Cluster 3.

(a) What are cluster centroids and cluster assignments after the first iteration of K-means clustering?

(b) Continue the algorithm of K-means clustering until it converges. Report the cluster centroids and cluster assignments after each iteration.

END OF PAPER

CONFIDENTIAL

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.