STRATHMORE UNIVERSITY
INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN STATISTICS

STA 8103 MULTIVARIATE DATA ANALYSIS TIME: 3 HOURS

DATE: 20th August 2018 Time: 3 Hrs

INSTRUCTIONS: Answer Question 1 and any other TWO.

Question 1 (20 Marks)

(a) (i) Explain two properties of covariance matrix for a multivariate distribution (2 Marks)

(ii) Differentiate between profile analysis and factor Analysis (2 Marks)

(b) 50 packets of match boxes are manufactured each of two ways. Two characteristics, X_1 and X_2 are measured. The summary statistics produced by method 1 and 2 are summarised as:

$$\bar{X}_1 = \begin{bmatrix} 8.3 \\ 4.1 \end{bmatrix}, S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}, \bar{X}_2 = \begin{bmatrix} 10.2 \\ 3.9 \end{bmatrix}, S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

Test at 0.05 level whether there is any significant difference in the characteristics measured. (6 Marks)

(c) Consider the following <u>five</u> observations of two variables:

X_1	95	100	105	95	105
X_2	210	200	190	195	205

(i) Find the sample mean vector for this data (2 Marks)

(ii) Find the sample covariance matrix (3 Marks)

(d) Consider the covariance matrix $\begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$ and the derived correlation matrix $R = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$. Determine

(i) Construct two principal components from R (4 Marks)

(ii) Determine the proportion of total variance in the data accounted for by the first principal component. (1 Mark)

Question 2 (20 Marks)

(a) Let $\underline{X} = (x_1, x_2, .., x_p)'$ be a p-dimensional column vector of random variables each which is distributed by $N_p \sim (\mu, \Sigma)$, with $\Sigma > 0$, as shown below

$$f(\underline{X}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2} Q\right] \quad -\infty < \underline{X} < \infty$$

where, $Q = (\underline{X} - \underline{\mu})' \Sigma^{-1} (\underline{X} - \underline{\mu})$.

(i) Derive the Maximum Likelihood Estimator of $\underline{\mu}$ (5 Marks)

(ii) Derive the Maximum Likelihood Estimator of Σ (5 Marks)

(b) Suppose $\underline{X} = x_1, x_2, ..., x_n$ are n independent observations and that $X \sim N(\mu, \sigma^2)$,

(i) Show that the joint distribution of $x_1, x_2, ..., x_n$ is the n-variate normal distribution. (3 Marks)

(ii) Determine the moment generating function of X (7 Marks)

Question 3 (20 Marks)

(a) Let the number of variates p=3 and the number of common factors m=1 and suppose the random variables X_1, X_2 and X_3 have the positive definite covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{bmatrix}$$

and that $\Psi = \text{cov}(\varepsilon) = \begin{bmatrix} 0.19 & 0 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0 & 0.75 \end{bmatrix}$

(i) Show that standardized random variables are given by; $Z_1 = 0.9F_1 + \varepsilon_1$, $Z_2 = 0.7F_1 + \varepsilon_2$, $Z_3 = 0.5F_1 + \varepsilon_3$ where $\text{var}(F_1)$ = cov $(\varepsilon_1, F_1) = 0$ (4 Marks)

(ii) Calculate the communicalities, hence interpret your results (3 Marks)

- b) Two groups of students were studied on their performance on three different subject out of a possible score of 15 marks. The data collected on students who attend classes during the day, n_1 and those who attend evening classes n_2 is as shown below.

Unit	Group	
	Day class, $n_1 = 37$	Evening class, $n_2 = 12$
Mathematics	13	9
English	10	5
Science	12	9

(i) Develop the variance –covariance matrix S (3 Marks)
(ii) Determine whether the two profiles are parallel (4 Marks)
(iii) Assuming the profiles are parallel determine whether the profiles are coincident (6 Marks)

Question 4 (20 Marks)

Suppose we have the following population of four observations made on three random variables X_1, X_2, and X_3:

X_1	X_2	X_3
1.0	6.0	9.0
4.0	12.0	10.0
3.0	12.0	15.0
4.0	10.0	12.0

Suppose the derived correlation matrix is given by;

$$\rho = \begin{bmatrix} 1 & .83 & .35 \\ .83 & 1 & .62 \\ .35 & .62 & 1 \end{bmatrix}$$

and the corresponding eigen value-eigen vector pairs:

$$\lambda_1 = 2.2 \; e_1 = [0.58 \quad 0.63 \quad 0.51]' \quad \lambda_2 = 0.62 \; e_2 = [0.54 \quad -0.16 \quad 0.82]'$$
$$\lambda_3 = 0.16, e_3 = [0.60 \quad -0.75 \quad 0.26]'$$

(a) Find the three population principal components variables Y_1, Y_2, and Y_3 for the standardized random variables Z_1, Z_2, and Z_3 (3 Marks)

(b) Determine the proportion of total population variance due to each principal component (8 Marks)

(c) Obtain the correlations between the original random variables X_i and the principal components Y_i. Hence interpret the correlation matrix. (9 Marks)

Question 5 (20 Marks)

(a) Outline Three assumptions about the structure of the data for one way Multivariate Analysis of Variance (3 Marks)

(b) Explain the objective of canonical correlation analysis (2 Marks)

(c) Let $Z^{(1)} = [Z_1^{(1)}, Z_2^{(2)}]'$ and $Z^{(2)} = [Z_1^{(1)}, Z_2^{(2)}]'$ be standardized variables $Z = [Z^{(1)}, Z^{(2)}]'$ and

$$\text{Covariance } Z = \begin{bmatrix} \Sigma_{11} & | & \Sigma_{12} \\ - & | & - \\ \Sigma_{21} & | & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{bmatrix}$$

Determine the first and second canonical correlations. Hence comment on your results

(15 Marks)