MTL766: Selected Questions on Multivariate Analysis

Extracted from Past Exams

September 13, 2025

1 Sample Moments, Expectation, and Basic Properties

1. Let $\mathbf{X}$ be an $n \times p$ data matrix. Suppose $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and sample covariance matrix, respectively, computed from $\mathbf{X}$.

   (a) Prove that $\mathbf{S}$ is positive definite if the columns of the mean corrected matrix $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T$ are linearly independent.

   (b) Let $\underline{\mathbf{X}}^T = [X_1, X_2, X_3]$ be a multi-normal random vector having a $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. A random sample of size 10 on $\mathbf{X}$ yielded the sample mean vector $\bar{\mathbf{X}}^T = [0.8, 0.5, 0.4]$ and sample covariance matrix $\mathbf{S} = \begin{pmatrix} 0.85 & 0.63 & 0.17 \\ 0.63 & 0.57 & 0.13 \\ 0.17 & 0.13 & 0.17 \end{pmatrix}$. Use the above information to compute an unbiased estimate of the covariance matrix Cov $\left[\begin{smallmatrix} X_1 - X_2 \\ X_1 - X_3 \end{smallmatrix}\right]$.

2. Let $\underline{\mathbf{X}} = (X_1, X_2, X_3)^T \sim \mathbb{N}_3(\underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 9 \end{pmatrix}$ and $\underline{\mathbf{a}} = (1, -1, 1)^T$. Determine the value of the vector $\mathbf{r} = (r_1, r_2, r_3)^T$, where r_i denotes the correlation between X_i and $\underline{\mathbf{a}}^T \underline{\mathbf{X}}$.

3. To test the "equal correlation" structure in the population, suppose we set the null hypothesis $H_0 : \boldsymbol{\rho} = \boldsymbol{\rho}_0 = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}$ against the alternative hypothesis $H_1 : \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$. The test statistic for a large sample is given by $T = \frac{n-1}{1-\bar{r}^2}((\sum\sum_{i<k} r_{ik} - \bar{r})^2 - \gamma \sum_{k=1}^{p} (\bar{r}_k - \bar{r})^2)$, where $\bar{r}_k$ is the average of off-diagonal elements in the k-th column and $\bar{r}$ is the overall mean of the off-diagonal elements in the sample correlation matrix. It is known that T is χ^2 distributed with d.o.f. $(p+1)(p-2)/2$. Use this to test H_0 at 1% level of significance for 150 samples with correlation matrix $R = \begin{pmatrix} 1 & .7501 & 0.6392 & 0.6363 \\ & 1 & 0.6925 & 0.7386 \\ & & 1 & 0.6625 \\ & & & 1 \end{pmatrix}$.

2 Visualization, Distance, and Geometric Interpretation

1. Justify that $(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$ is a valid statistical distance measure of an observation vector $\mathbf{x}$ from the data mean vector $\bar{\mathbf{x}}$, whose value does not depend on the scales of measurement of components of $\mathbf{x}$.

2. Use the geometric interpretation in n-space of the generalised sample variance computed from an $n \times p$ data matrix $\mathbf{X}$ to justify that it gives a joint measure of variation of p-component variables of the measurement $\mathbf{x}$. What is a major weakness of this measure of joint variation?

3. Prove that the sample generalized variance computed from an $n \times p$ data matrix $\mathbf{X}$ is zero if and only if p deviation vectors defined on $\mathbf{X}$ are linearly dependent.

4. (True/False) For any $\mathbf{X} \in \mathbb{R}^p$, the Mahalanobis distance $M_D(\mathbf{X}, \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = (\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$ is non-negative. Justify your answer.

5. Explain the term Mahalanobis Distance between two points $\mathbf{x}$ and $\mathbf{y}$, taken from a p-dimensional multivariate distribution.

6. Let $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix} \right)$.

   (a) Determine the constant-density ellipse that contains data from this distribution with probability equal to 0.90.

   (b) Determine the direction ratios and lengths of the major and minor axes of the ellipse obtained in part (a). Provide a (rough) sketch of this ellipse.

7. Consider the following data of 10 students' marks (out of 10) in three subjects X, Y and Z. Draw a constellation graph showing the data with equal weights to each subject. Show your calculations. Identify from the graph the best, worst and most average students of the class with justification.

No.	X	Y	Z	No.	X	Y	Z
1	1	2	4	6	2.5	8	6
2	1	6	8	7	3	4	2
3	2	4	5	8	3.5	6	7
4	2	2	3	9	4	10	8
5	2.5	4	3	10	4	8	7

3 Multivariate Normal Distribution and its Properties

1. (True/False) Let X and Y be two univariate random variables. If $aX + bY$ follows a normal distribution for all fixed $a, b \in \mathbb{R}$, then the joint distribution of (X, Y) is bivariate normal. Justify your answer.

2. (True/False) Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $g : \mathbb{R}^p \to \mathbb{R}$ be a continuous function. Then the distribution of $g(\mathbf{X})$ is univariate normal. Justify your answer.

3. (True/False) Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$; $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$. Then $\bar{\mathbf{X}}$ and $\mathbf{X}_1 - \bar{\mathbf{X}}$ are independently distributed. Justify your answer.

4. A company makes a candy bar in three sizes: small (X_1), regular (X_2), and big (X_3). The joint distribution of the weight (in gms) of the candy bars $\underline{\mathbf{X}} = (X_1, X_2, X_3)^T$ follows a multivariate normal distribution with parameters $\underline{\boldsymbol{\mu}} = (3, 5, 7)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & 2 \\ 0 & 2 & 9 \end{pmatrix}$.

   (a) What is the probability that the weight of a regular bar is greater than 8 gm, given that the small size bar weighs 2 gm and the big size bar weighs 10 gm?

   (b) Determine $P(X_1 - 2X_2 + X_3 < 5)$.

5. Suppose $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ is distributed as 3-D normal with parameters: $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 3 \end{pmatrix}$.
   Obtain the conditional distribution of $\begin{pmatrix} X_1 \\ X_3 \end{pmatrix}$ given $X_2 = 2$.

4 Distribution of Sample Statistics

1. (True/False) For any fixed $\mathbf{d} \in \mathbb{R}^p$, the distribution of Mahalanobis distance $M_D(\mathbf{d}, \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = (\mathbf{d} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{d} - \boldsymbol{\mu})$ has a non-central chi-square distribution. Justify your answer.

2. Let $\mathbf{A} \sim W_p(n, \boldsymbol{\Sigma})$. Then using the definition of Wishart distribution or otherwise, prove or disprove $\mathbb{E}(\mathbf{A}) = n\boldsymbol{\Sigma}$.

3. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a random sample from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\bar{\mathbf{X}} = \sum_{i=1}^n \mathbf{X}_i/n$ and $\mathbf{S} = \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T/(n-1)$. Prove or disprove that the distribution of $n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$ converges to a chi-square distribution as $n \to \infty$.

4. Let $\mathbf{X}$ have a $N_p(\underline{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ distribution. Prove that $(\mathbf{X} - \underline{\boldsymbol{\mu}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \underline{\boldsymbol{\mu}})$ has a χ^2-distribution with p-degrees of freedom.

5. Show that an approximate distribution of $n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})$ is a χ^2-distribution with p-degrees of freedom for a large $n - p$, where $\bar{\mathbf{X}}$ is the sample mean vector of a random sample of size n from any population having covariance matrix $\boldsymbol{\Sigma}$.

6. Suppose $\binom{1}{2}, \binom{2}{-2}, \binom{-2}{1}, \binom{1}{-3}, \binom{-2}{3}$ are five observations from 3-dimensional Normal with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Use the above data to explain what is Wishart distribution.

5 Inference on the Mean Vector (Hypothesis Testing & C.I.)

1. Three observations from $\mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are (1,2), (1.2, 1.8), (2,3). Find a 95% minimum volume confidence region for $\boldsymbol{\mu}$.

2. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{\mu}$ are unknown. For testing $H_0 : \boldsymbol{\mu} = (1, 2, 3)^T$ against $H_A : H_0$ is not true, construct a test based on Hotelling's T^2 statistic and write its distribution under H_0.

3. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from $\mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} > 0$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)^T$ are unknown. For testing $H_0 : \mu_1 = \mu_2 - \mu_3$ against $H_A : H_0$ is not true, construct a test based on Hotelling's T^2 statistic and write its distribution under H_0.

4. Consider sample data measuring three attributes X_1, X_2, X_3 of 10 objects from a multivariate normal population. The mean is $\begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}$ and covariance is $\begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 1 & 4 \end{pmatrix}$.

   (a) Construct the 95% simultaneous confidence ellipse for X_1 and X_2.

   (b) Can we assert that the mean of X_1 is the average of the means of X_2 and X_3? Justify with a hypothesis test at a 5% level of significance.

5. Define a 2-sample Hotelling T^2-statistic for testing the equality of mean vectors of two Normal populations $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Find the distribution of this statistic assuming the distribution of a one-sample T^2-statistic.

6. Consider 42 observations on variables X_1 and X_2. The summary statistics are $\bar{\mathbf{x}} = \begin{pmatrix} 0.564 \\ 0.603 \end{pmatrix}$ and $\mathbf{S} = \begin{pmatrix} 0.0144 & 0.0117 \\ 0.0117 & 0.0146 \end{pmatrix}$.

   (a) Stating the assumptions made, test at a 5% level of significance the hypothesis that $[\mu_1, \mu_2] = [0.57, 0.59]$ against any other values.

   (b) Find the simultaneous 95% T^2-confidence intervals for μ_1, μ_2, and $\mu_1 - \mu_2$.

   (c) If the sample size 42 is regarded as large in comparison with variable size p=2, will the assumptions and conclusions of the test in part (a) change? If so, how?

7. Consider an iid sample of size n=5 from an $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution with $\boldsymbol{\Sigma} = \begin{pmatrix} 3 & \eta \\ \eta & 1 \end{pmatrix}$, where η is a known parameter. Suppose the sample mean is $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. For what value of η would the hypothesis $H_0 : \boldsymbol{\mu} = (0, 0)^T$ be rejected in favor of $H_1 : \boldsymbol{\mu} \neq (0, 0)^T$ at the 5% significance level?

8. Consider the following 7 observations from a 3-dimensional Normal distribution:
   $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \\ 4 \end{pmatrix}, \begin{pmatrix} -2 \\ 1 \\ 5 \end{pmatrix}, \begin{pmatrix} 1 \\ -3 \\ 2 \end{pmatrix}, \begin{pmatrix} -2 \\ 3 \\ 4 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \\ 5 \end{pmatrix}$.
   Test if the data is coming from a distribution with mean $\begin{pmatrix} 0 \\ 1 \\ 3 \end{pmatrix}$ at a 95% confidence level. Show all your calculations.

6 Inference on the Covariance Matrix

1. Let a population distribution be $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Based on a random sample of size n from this population, develop a large sample approximate likelihood ratio test for the hypothesis $H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ against $H_1 : \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma}_0$, where $\boldsymbol{\Sigma}_0$ is a given positive definite matrix.