ST 557: Applied Multivariate Analysis
Fall 2019

Homework 2

Due: Friday October 18
45 Points

1. **(10 points)** Let $\mathbf{X} = [X_1, X_2, X_3]^T$ be a multivariate normal random vector with mean $\boldsymbol{\mu} = [-3, 1, 4]^T$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Which of the following pairs of random variables are independent? Explain.

(a) X_1 and X_2

(b) X_2 and X_3

(c) $\mathbf{Y} = [X_1, X_2]^T$ and X_3

(d) $Y = \frac{X_1+X_2}{2}$ and X_3

(e) X_2 and $Y = X_2 - \frac{5}{2}X_1 - X_3$

2. **(12 points)** The *Mahalanobis distance* (also called *statistical distance* in this textbook) of a point $\mathbf{y} = [y_1, y_2, \ldots, y_p]^T$ from the mean $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots \mu_p]^T$ of a distribution that has covariance matrix $\boldsymbol{\Sigma}$ is given by

$$d_M(\mathbf{y}, \boldsymbol{\mu}) = \sqrt{(\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})}.$$

(Mahalanobis distance may also be computed between $\mathbf{y}$ and another point $\mathbf{x}$ by just substituting $\mathbf{x}$ for $\boldsymbol{\mu}$ in the above expression.) Recall that the Euclidean distance of a point $\mathbf{y}$ from another point $\mathbf{x}$ is given by

$$d_E(\mathbf{y}, \mathbf{x}) = \sqrt{(\mathbf{y} - \mathbf{x})^T (\mathbf{y} - \mathbf{x})} = \sqrt{\sum_{j=1}^{p} (y_j - x_j)^2}$$

(a) Consider the point $\mathbf{y} = [1, 2, -2]^T$, and mean vectors $\boldsymbol{\mu}_1 = [0, 0, 0]^T$ and $\boldsymbol{\mu}_2 = [3, 4, -3.5]^T$. Find the Euclidean distances $d_E(\mathbf{y}, \boldsymbol{\mu}_1)$ and $d_E(\mathbf{y}, \boldsymbol{\mu}_2)$.

(b) For the same two mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ from part (a), find the Mahalanobis distances $d_M(\mathbf{y}, \boldsymbol{\mu}_1)$ and $d_M(\mathbf{y}, \boldsymbol{\mu}_2)$, if the covariance matrix is

$$\Sigma = \begin{bmatrix} 9.0 & 8.1 & -3.6 \\ 8.1 & 9.0 & -4.8 \\ -3.6 & -4.8 & 4.0 \end{bmatrix}$$

(c) If $\bar{\mathbf{X}}$ is the sample mean for an iid sample of n multivariate normal random vectors with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, what is the distribution of $\bar{\mathbf{X}}$? What is the distribution of

$$n\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)=\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)^T\left(\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1}\left(\bar{\mathbf{X}}-\boldsymbol{\mu}\right)?$$

How is this quantity related to the Mahalanobis distance of $\bar{\mathbf{X}}$ from $\boldsymbol{\mu}$?

(d) Given your answers to part (a) - (c), do you think μ_1 or μ_2 is more plausible as the population mean of a distribution with covariance matrix

$$\Sigma=\begin{bmatrix} 9.0 & 8.1 & -3.6 \\ 8.1 & 9.0 & -4.8 \\ -3.6 & -4.8 & 4.0 \end{bmatrix}$$

if the sample mean is $\bar{\mathbf{X}}=[1,2,-2]^T$? Explain your reasoning.

3. **(6 points)** Let $\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3,\boldsymbol{X}_4$ and $\boldsymbol{X}_5$ be independent and identically distributed random vectors with mean vector μ and covariance matrix Σ. Find the mean vector and covariance matrices for each of the following two linear combinations of these random vectors:

(a) $\boldsymbol{Y}_1=\frac{1}{5}\boldsymbol{X}_1+\frac{1}{5}\boldsymbol{X}_2+\frac{1}{5}\boldsymbol{X}_3+\frac{1}{5}\boldsymbol{X}_4+\frac{1}{5}\boldsymbol{X}_5$

(b) $\boldsymbol{Y}_2=\boldsymbol{X}_1-\boldsymbol{X}_2+\boldsymbol{X}_3-\boldsymbol{X}_4+\boldsymbol{X}_5$

4. **(5 points)** Find the maximum likelihood estimates of the 2×1 mean vector μ and the 2×2 covariance matrix Σ based on the random sample

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \boldsymbol{X}_3\\ \boldsymbol{X}_4\end{bmatrix}=\begin{bmatrix}3 & 6\\ 4 & 4\\ 5 & 7\\ 4 & 7\end{bmatrix}$$

from a bivariate normal population.

5. **(12 points)** Exploring the performance of the correlation test for normality: Recall the statistic

$$r_Q=\frac{\sum_{i=1}^n(x_{(j)}-\bar{x})(q_{(j)}-\bar{q})}{\sqrt{\sum_{i=1}^n(x_{(j)}-\bar{x})^2}\sqrt{\sum_{i=1}^n(q_{(j)}-\bar{q})^2}}$$

where x_j is the jth largest value in the sample, and

$$q_{(j)}=\Phi^{-1}\left(\frac{j-\frac{1}{2}}{n}\right)$$

where Φ is the standard normal cdf, and Φ^{-1} is the inverse cdf of the standard normal distribution. The critical values (for a level $\alpha=0.05$ test) for this statistic are given in the following table, for several different sample sizes (a more extensive table may be found on p. 181 of the textbook). If the value of r_Q is below the given value, we reject the hypothesis of normality.

n	Critical Value
5	0.8788
10	0.9198
20	0.9508

(a) Perform 10,000 simulations to see how good this test is at detecting departure from normality when the data are actually Uniform(0, 1), for a sample of size 10:

- Generate 10,000 datasets of 10 observations, and for each dataset compute the test statistic r_Q.
- Report the proportion of the 10,000 datasets for which the resulting test statistic is less than the appropriate critical value.

How often does this test reject the null hypothesis (rejection is the correct decision) that the data are normal at level $\alpha = 0.05$?

(b) Repeat part (a), except this time for data that have a Chi-square distribution with 5 degrees of freedom, and a sample size of 5.

(c) Repeat part (a), except this time for data that have a Chi-square distribution with 2 degrees of freedom, and a sample size of 20.