Multivariate Analysis - Homework 1

Please upload your homework on SPOC before 8:00pm, March 9, including all details needed. For R exercises, R markdown is highly encouraged; for other parts, try to use LaTex.

1. Prove that the sample covariance matrix is an unbiased estimator for population covariance matrix.

2. The following are five measurements on the variables Y_1, Y_2 and Y_3:

Y_1	9	2	6	5	8
Y_2	12	8	6	4	10
Y_3	3	4	0	2	1

   Find (by hand) $\bar{\mathbf{y}}$, $\mathbf{S}$ and $\mathbf{R}$.

3. Suppose the random vector $\mathbf{y} = (Y_1, Y_2, Y_3, Y_4)$ with mean vector $\boldsymbol{\mu} = (4, 3, 2, 1)'$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 1 & 1 & 0 \\ 2 & 1 & 9 & -2 \\ 2 & 0 & -2 & 4 \end{bmatrix}$$

   Partition $\mathbf{y}$ as $\mathbf{y}^{(1)} = (Y_1, Y_2)'$ and $\mathbf{y}^{(2)} = (Y_3, Y_4)'$. Furthermore, let matrix

$$\mathbf{A} = [1, 2] \text{ and } \mathbf{B} = \begin{bmatrix} 1 & -2 \\ 2 & -1 \end{bmatrix}$$

   and consider the linear combinations $\mathbf{Ay}^{(1)}$ and $\mathbf{By}^{(2)}$. Find

   (a) $E(\mathbf{y}^{(1)})$

   (b) $E(\mathbf{Ay}^{(1)})$

   (c) $COV(\mathbf{y}^{(1)})$

   (d) $COV(\mathbf{Ay}^{(1)})$

   (e) $E(\mathbf{By}^{(2)})$

   (f) $COV(\mathbf{By}^{(2)})$

   (g) $COV(\mathbf{y}^{(1)}, \mathbf{y}^{(2)})$

   (h) $COV(\mathbf{Ay}^{(1)}, \mathbf{By}^{(2)})$

4. (R exercise.) The following table (data attached) gives partial data from three variables measured in milliequivalents per 100g:

$$y_1 = \text{available soil calcium},$$
$$y_2 = \text{exchangeable soil calcium},$$
$$y_3 = \text{turnip green calcium}.$$

Table. Calcium in Soil and Turnip Greens

Location Number	y_1	y_2	y_3
1	35	3.5	2.80
2	35	4.9	2.70
3	40	30.0	4.38
4	10	2.8	3.21
5	6	2.7	2.73
6	20	2.8	2.81
7	35	4.6	2.88
8	35	10.9	2.90
9	35	8.0	3.28
10	30	1.6	3.20

Define

$$z_1 = y_1 + y_2 + y_3,$$
$$z_2 = 2y_1 - 3y_2 + 2y_3,$$
$$z_3 = -y_1 - 2y_2 - 3y_3.$$

(a) Find the sample mean vector $\bar{\mathbf{z}}$, sample covariance matrix $\mathbf{S}_z$ of $\mathbf{z} = (Z_1, Z_2, Z_3)'$

(b) Find the sample correlation matrix $\mathbf{R}_z$ from $\mathbf{S}_z$.

(c) Find the generalized variance and total variance of $\mathbf{z}$.

(d) Realize the spectral decomposition and Cholesky decomposition of both $\mathbf{S}_z$ and $\mathbf{R}_z$, and get the square root matrix of them.

5. (R exercise.) The attached data are 42 measurements on air-pollution variables recorded at 12:00 noon in the Los Angeles area on different days.

(a) Plot the pairwise scatter plot matrix for all the variables in R. And comment on the output.

(b) Construct the sample mean vector, sample covariance matrix and sample correlation matrix. Interpret the entries in the sample correlation matrix.

(c) Compute the Eucleadian distance matrix and the Mahalanobis/statistical distance matrix among the first five days. Explain the advantage of the Mahalanobis distance.

(d) Describe the overall variability of the data.

(e) Get the Spectral decomposition and Cholesky decomposition of the sample covariance matrix. Observe the difference between the two decompositions.

(f) Obtain a 3-D scatter plot for any three variables that you think make sense. Use any package/command in R except for the one given in the slides.

6. (R exercise.) The attached data “guangdong.xlsx” provide a summary of the high-tech product market in Guangzhou province, China, in 2004. Perform descriptive analysis of it. You can use both graphical and numerical methods.