Chapter 3: Applications of Matrices in Statistics

Learning Objectives:

1. *Recognize multivariate data can be organized into a data matrix.*
2. *Compute and interpret population and sample mean vectors, covariance matrices and correlation matrices.*
3. *Describe bivariate normal distribution.*
4. *Compute generalized sample variance and total sample variance.*
5. *Determine sample mean vector, sample covariance matrix and sample correlation matrix for a linear combination of random variables.*

3.1 Multivariate Data

A single multivariate observation is a collection of measurements on p different random variables X_1, $X_2 \ldots$, X_p taken on the same item or trial.

If n observations have been obtained, the entire data set can be written as an $n \times p$ matrix:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

We say that the data are a sample of size n from a p-variate population.
The sample then consists of n measurements, each of which has p components.
Data matrices are denoted by large uppercase bold letters, such as $\mathbf{X}$ and $\mathbf{Y}$.

Recall from the module MS0140 Statistics for Data Science that a random variable is a variable whose numeric value is based on the outcome of a random event. A random variable is represented by an uppercase letter, such as X and Y. For example, X could be a random variable representing the height of a randomly selected student. The value of X is unknown until a student is selected and his or her height is measured. A realization or a particular value of a random variable is denoted by its corresponding lowercase letter, such as x and y.

By partitioning into its columns, $\mathbf{X}$ can be written in this way:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{pmatrix}$$

where the column vector $\mathbf{x}_j$ represents all n measurements on the j^{th} variable.

Alternatively, we can partition $\mathbf{X}$ into its rows:

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{\chi}_1^{'} \\ \boldsymbol{\chi}_2^{'} \\ \vdots \\ \boldsymbol{\chi}_n^{'} \end{pmatrix}$$

where the row vector $\boldsymbol{\chi}_i^{'}$ represents all p measurements on the i^{th} observation.

Example 1.

The table below shows the quiz scores in Mathematics and Science for 4 students.

	Mathematics	**Science**
Ali	12	16
Bob	18	20
Cindy	14	16
Dawn	20	18

(a) Write the quiz scores above as a data matrix $\mathbf{X}$.

(b) State the number of variables p and the number of observations n.

(c) Write the individual vector for each variable.

(d) Write the individual vector for each observation.

3.1.1 Mean Vector

Population mean vector is denoted as: $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$

where μ_j is a scalar that represents the population mean of the j^{th} variable.

Hence, sample mean vector is denoted as: $\overline{\mathbf{x}} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix}$

where $\overline{x}_j$ is a scalar that represents the sample mean of data collected on the j^{th} variable.

The sample mean vector can be computed using matrix operations:

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}_n$$

Example 2.

Using matrix operations, find the sample mean vector for the data matrix $\mathbf{X}$ in Example 1. Interpret the elements in the vector obtained.

$$\mathbf{X} = \begin{bmatrix} 12 & 16 \\ 18 & 20 \\ 14 & 16 \\ 20 & 18 \end{bmatrix}$$

3.1.2 Covariance Matrix

Population covariance matrix is denoted as: $\Sigma = \mathrm{Cov}(\mathbf{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{pmatrix}$

where σ_{ij} is a scalar that represents the population **covariance** between i^{th} and j^{th} variables. Specifically, diagonal elements σ_{jj} is a scalar that represents the population **variance** of the j^{th} variable.

Observe that the covariance matrix is always a $p \times p$ symmetric matrix. (Why?)

Covariance between two variables describes the relationship between the variables. The sign of the covariance indicates the direction of the relationship – either positive or negative. However, the numerical value gives no valuable information about the strength of the relationship, because $-\infty < \mathrm{cov}(X, Y) < \infty$.

Nevertheless, if two variables are independent, then the covariance between the two variables will be zero. Note that the converse is not necessarily true.

The sample covariance matrix is given by:

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ c_{12} & c_{22} & \cdots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{1p} & c_{2p} & \cdots & c_{pp} \end{pmatrix}$$

where c_{ij} is a scalar that represents the sample covariance between data collected on the i^{th} and j^{th} variables. Specifically, c_{jj} is a scalar that represents the sample variance of the data collected on the j^{th} variable.

The sample covariance matrix can be computed using matrix operations:

$$\mathbf{C} = \frac{1}{n-1}\left(\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}'\right)'\left(\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}}'\right)$$

Sometimes, it is desirable to assign a single numerical value for the variation encapsulated in the sample covariance matrix **C**. Two options for this are:

Generalized sample variance = $|\mathbf{C}|$

Total sample variance = $\text{tr}(\mathbf{C})$

Example 3.

Continue from Examples 1 and 2.

(a) Compute the sample covariance matrix **C**. Which subject has higher variability in the quiz scores? What is the relationship between quiz scores of both subjects? Explain.

$$\mathbf{X} = \begin{bmatrix} 12 & 16 \\ 18 & 20 \\ 14 & 16 \\ 20 & 18 \end{bmatrix}$$

$$\bar{\mathbf{x}} = \begin{bmatrix} 16 \\ 17.5 \end{bmatrix}$$

(b) Hence, compute the generalized sample variance and total sample variance.

3.1.3 Correlation Matrix

Population correlation matrix is denoted as: $\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$

where ρ_{ij}, $i \neq j$, is a scalar that represents the population **correlation** between i^{th} and j^{th} variables.

Observe that the correlation matrix is always a $p \times p$ symmetric matrix, with diagonal elements always 1. (Why?)

Similar to covariance, correlation between two variables describes the relationship between the variables. In fact:

$$\boxed{\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{SD}(X) \cdot \operatorname{SD}(Y)}}$$

The sign of the correlation indicates the direction of the relationship – either positive or negative. The numerical value indicates the strength of the relationship, because $-1 \leq \operatorname{corr}(X, Y) \leq 1$. In particular, correlation of 1 indicates perfect positive relationship.

If two variables are independent, then the correlation between the two variables will be zero. Note that the converse is not necessarily true.

Then, the sample correlation matrix is denoted as: $\mathbf{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{pmatrix}$

where $\boxed{r_{ik} = \dfrac{c_{ik}}{\sqrt{c_{ii}}\sqrt{c_{kk}}}}$ is a scalar that represents the sample correlation between data collected on the i^{th} and j^{th} variables.

The sample standard deviation matrix is defined as: $\mathbf{S} = \begin{pmatrix} \sqrt{c_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{c_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{c_{pp}} \end{pmatrix}$

Thus, its inverse is: $\mathbf{S}^{-1} = \begin{pmatrix} \frac{1}{\sqrt{c_{11}}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{c_{22}}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sqrt{c_{pp}}} \end{pmatrix}$

Then, sample correlation matrix can be computed using matrix operations:

$$\mathbf{R} = \mathbf{S}^{-1}\mathbf{C}\mathbf{S}^{-1}$$

Example 4.

Continue from Example 3.

(a) Compute the sample correlation matrix **R** by computing the sample correlation between Mathematics and Science quiz scores.

$$\mathbf{C} = \begin{bmatrix} 13.3 & 5.3 \\ 5.3 & 3.7 \end{bmatrix}$$

(b) Compute the sample correlation matrix **R** by using matrix operations.

(c) Comment on the relationship between Mathematics and Science quiz scores.

3.2 Bivariate Data

In particular, let us write out the data matrix, sample mean vector, sample covariance matrix and sample correlation matrix for bivariate data.

Data matrix

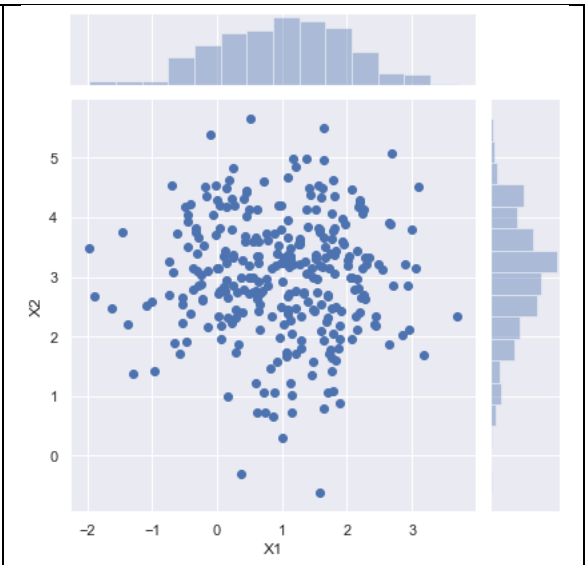
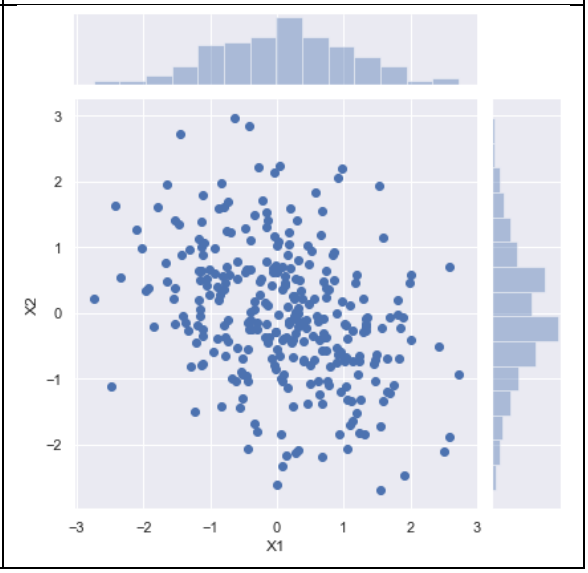
Sample mean vector

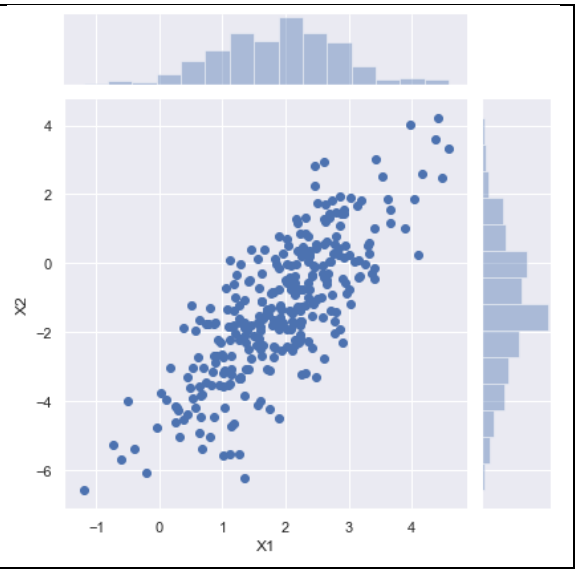
Sample covariance matrix

Sample correlation matrix

Example 5.

In this example, we will understand **bivariate normal distribution** by visualization.

(a)	$\bar{\mathbf{x}} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	
(b)	$\bar{\mathbf{x}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}$	

(c)	$\overline{\mathbf{x}} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$, $\mathbf{C} = \begin{bmatrix} 1 & 1.6 \\ 1.6 & 4 \end{bmatrix}$	

Example 6.

Given the following data matrix $\mathbf{X} = \begin{pmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{pmatrix}$. Assume normal distribution.

(a) State the number of variables and observations.

(b) Write out the individual vectors of each variable.

(c) Compute the sample mean vector, sample covariance matrix and sample correlation matrix by matrix method.

(d) Make sense of the data by plotting.

(e) From part (c), state the following:

(i) SD(X_1)

(ii) SD(X_2)

(iii) relationship between X_1 and X_2

(f) What is the generalized sample variance?

(g) What is the total sample variance?

3.3 Sample Values of Linear Combinations of Random Variables

The linear combination of p random variables $X_1, X_2, ..., X_p$ can be expressed as:

$$\mathbf{a}'\mathbf{X} = \begin{pmatrix} a_1 & a_2 & \dots & a_p \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = a_1X_1 + a_2X_2 + ... + a_pX_p$$

We use bold uppercase letters, such as **X**, to represent a vector of random variables, such as $\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$.

$\mathbf{a}'\mathbf{X}$ forms a new random variable.

In general, consider the q linear combinations of the p random variables $X_1, X_2, ..., X_p$:

$$\begin{aligned} Y_1 &= a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p \\ Y_2 &= a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p \\ &\vdots \\ Y_q &= a_{q1}X_1 + a_{q2}X_2 + ... + a_{qp}X_p \end{aligned}$$

This can be rewritten in matrix form: $\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \dots & a_{qp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$

Or simply, $\mathbf{Y} = \mathbf{AX}$, where $\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix}$, $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \dots & a_{qp} \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$.

AX is a **linear transformation** of **X**.
Y forms a new random variable.

The transformed data matrix $\mathbf{Y}$ can be obtained by direct application of the linear combinations. Alternatively, the data matrix $\mathbf{X}$ can be transformed one observation vector at a time, to obtain the transformed data matrix $\mathbf{Y}$. For example, the i^{th} observation vector $\boldsymbol{\chi}_1'$ can be transformed by

$$\mathbf{y}_i = \mathbf{A}\boldsymbol{\chi}_i .$$

Example 7.

Given data $\mathbf{X}$ in Example 6, and the following linear combinations:

$$Y_1 = X_1 + 2X_2$$
$$Y_2 = 3X_1 - X_2$$

(a) Obtain the transformed data $\mathbf{Y}$ by:

(i) direct application of the linear combinations.

(ii) transforming the data in $\mathbf{X}$ one observation vector at a time using matrix algebra method.

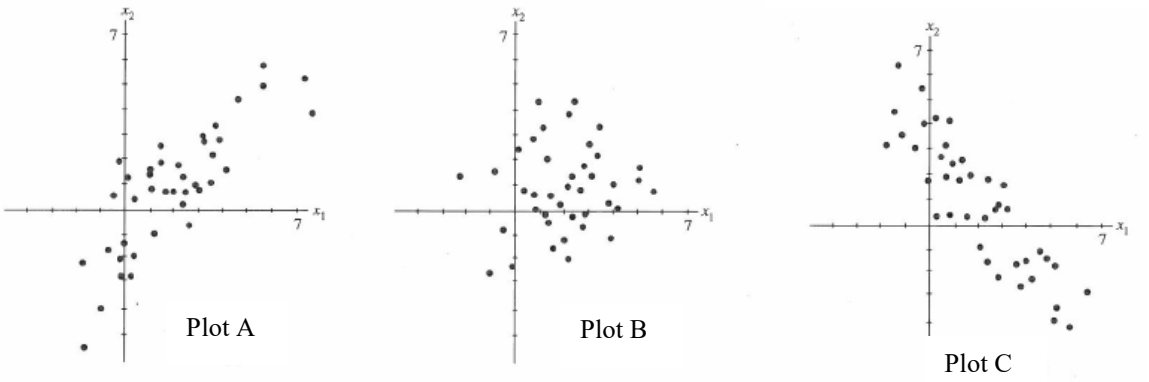
(b) Compute the sample mean vector and sample covariance matrix of data $\mathbf{Y}$.

(c) What is the correlation between Y_1 and Y_2?

Tutorial 3

1. Scatterplots below show three sets of bivariate data; all having but different variability structures.

(I) $\mathbf{C} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$ (II) $\mathbf{C} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}$ (III) $\mathbf{C} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}$



(a) Find the correlation in each sample covariance matrix.

(b) Match the covariance matrix to the correct scatterplot.

(c) Show that the generalized sample variances are all the same.

2. For each of the following data matrices, compute the sample mean vector, sample covariance matrix and sample correlation matrix. Interpret the elements in the vector/matrix.

(a) $\mathbf{X} = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix}$ (b) $\mathbf{X} = \begin{bmatrix} 3 & 4 \\ 6 & -2 \\ 3 & 1 \end{bmatrix}$

3. Given the data matrix $\mathbf{X} = \begin{pmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{pmatrix}$ and the linear combinations:

$$Y_1 = 2X_1 + 2X_2 - X_3$$
$$Y_2 = X_1 - X_2 + 3X_3$$

(a) Compute the sample mean vector and sample covariance matrix of data $\mathbf{X}$.

(b) Find the sample mean vector and sample covariance matrix of data $\mathbf{Y}$.

4. The following is a simple scenario of data taken from a telemarketing company. The number of successful phone solicitations in 5 days of a particular week by a part-time (PT) and a full-time (FT) employee is recorded in the table below:

X_1: Number by PT employee	1	4	2	5	3
X_2: Number by FT employee	9	12	10	8	11

(a) Find the sample mean vector and sample covariance matrix of data $\mathbf{X}$. Interpret the matrices, and explain the correlation between X_1 and X_2.

Abby, an administration executive of the company, decided to create three extra variables for her records for the 5 days:

Y_1: Total number of successful phone solicitations					
Y_2: Average number of successful phone solicitations					
Y_3: Difference in number of successful phone solicitations between FT and PT (FT – PT)					

(b) Express Y_1, Y_2 and Y_3 as linear combinations of X_1 and X_2.

(c) Find the sample mean vector and sample covariance matrix of the transformed data $\mathbf{Y}$.

Answers

1. (a) 0, 0.8, –0.8 (b) (I) → Plot B, (II) → Plot A, (III) → Plot C
   (c) 9

2. (a) $\begin{bmatrix}5\\2\end{bmatrix}, \begin{bmatrix}16 & -2\\-2 & 1\end{bmatrix}, \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}$
   (b) $\begin{bmatrix}4\\1\end{bmatrix}, \begin{bmatrix}3 & -4.5\\-4.5 & 9\end{bmatrix}, \begin{bmatrix}1 & -0.866\\-0.866 & 1\end{bmatrix}$

3. (a) $\begin{bmatrix}3\\1\\5\end{bmatrix}, \begin{bmatrix}3 & -1.5 & 0\\-1.5 & 1 & 0.5\\0 & 0.5 & 1\end{bmatrix}$ (b) $\begin{bmatrix}3\\17\end{bmatrix}, \begin{bmatrix}3 & 4.5\\4.5 & 13\end{bmatrix}$

4. (a) $\bar{\mathbf{x}} = \begin{bmatrix}3\\10\end{bmatrix}, \begin{bmatrix}2.5 & 0\\0 & 2.5\end{bmatrix}$
   (b) $Y_1 = X_1 + X_2$, $Y_2 = 0.5X_1 + 0.5X_2$, $Y_3 = -X_1 + X_2$
   (c) $\begin{bmatrix}13\\6.5\\7\end{bmatrix}, \begin{bmatrix}5 & 2.5 & 0\\2.5 & 1.25 & 0\\0 & 0 & 5\end{bmatrix}$

Practical 3

Data can be downloaded from Blackboard.

In a study on the cost of transporting milk from farms to dairy plants, a survey was taken of firms engaged in milk transportation. Cost data on X_1 = fuel, X_2 = repair and X_3 = capital, all measured on a per-mile basis from 34 firms were taken. The table below shows part of the data collected.

Fuel (X_1)	**Repair (X_2)**	**Capital (X_3)**
16.44	12.43	11.23
7.19	2.7	3.92
⋮	⋮	⋮
17.32	6.86	4.44

(a) Plot a scatter matrix to visualize the relationship between X_1, X_2 and X_3. Are the data approximately multivariate normal?

(b) Find the sample mean vector, sample covariance matrix and sample correlation matrix of data $\mathbf{X}$.

(c) Find the generalized sample variance and total sample variance.

(d) Compute the sample mean and sample variance of a firm's operational cost of transporting milk per mile, which comprises of fuel cost and repair cost.

(e) Compute the sample mean and sample variance of a firm's total cost of ownership of the trucks per mile for transporting milk, which comprises of fuel cost, repair cost and capital cost.

(f) Determine the sample covariance matrix for the variables in parts (d) and (e).

Answers

(e) Sample mean = \$28.51, sample variance = \$60.62

(f) $\begin{bmatrix} 33.595 & 41.116 \\ 41.116 & 60.624 \end{bmatrix}$