Lecture 1 Introduction

Shiwei Lan[1]

[1]School of Mathematical and Statistical Sciences
Arizona State University

STP533 Multivariate Analysis
Spring 2025

Table of Contents

Objective of Multivariate Analysis

This course concerns statistical methods to elicit information from data with simultaneous measurements on many variables. The objective of *multivariate analysis* includes

1. Data reduction or structural simplification.
2. Sorting and grouping.
3. Investigation of the dependence among variables.
4. Prediction.
5. Hypothesis construction and testing.

Applications of Multivariate Analysis

1. Data reduction or simplification.
    - Using data on several variables related to cancer patient responses to radio- therapy, a simple measure of patient response to radiotherapy was constructed.
    - Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions.
2. Sorting and grouping.
    - Measurements of several physiological variables were used to develop a screen- ing procedure that discriminates alcoholics from nonalcoholics.
    - The U.S. Internal Revenue Service uses data collected from tax returns to sort taxpayers into two groups: those that will be audited and those that will not.
3. Investigation of the dependence among variables.
    - Data on several variables were used to identify factors that were responsible for client success in hiring external consultants.
    - The associations between measures of risk-taking propensity and measures of socioeconomic characteristics for top-level business executives were used to assess the relation between risk-taking behavior and performance.
4. Prediction.
    - Data on several variables related to the size distribution of sediments were used to develop rules for predicting different depositional environments.
    - Measurements on several accounting and fiancial variables were used to de-velop a method for identifying potentially insolvent property-liability insurers.
5. Hypothesis testing.
    - Several pollution-related variables were measured to determine whether levels for a large metropolitan area were roughly constant throughout the week, or whether there was a noticeable difference between weekdays and weekends.
    - Experimental data on several variables were used to see whether the nature of the instructions makes any difference in perceived risks, as quantified by test scores.

Table of Contents

Arrays

- We focus on analyzing measurements on several variables or characteristics. They can be arranged and displayed in various ways, e.g. graphs, tables, summary statistics, etc.
- Multivariate data consist of $p \geq 1$ *variables or characters* for $n \geq 1$ *items, individuals or experimental units*.
- We use x_{ij} denote the particular value of j-th variable that is observed in i-th item.
- If we arrange the data items in an array, denoted as $\mathbf{X}$, we have

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Example 1.1 (A data array) A selection of four receipts from a university bookstore was obtained in order to investigate the nature of book sales. Each receipt provided, among other things, the number of books sold and the total amount of each sale. Let the first variable be total dollar sales and the second variable be number of books sold. Then we can regard the corresponding numbers on the receipts as four measurements on two variables. Suppose the data, in tabular form, are

Variable 1 (dollar sales):	42	52	48	58
Variable 2 (number of books):	4	5	4	3

Using the notation just introduced, we have

$$\begin{array}{llll} x_{11} = 42 & x_{21} = 52 & x_{31} = 48 & x_{41} = 58 \\ x_{12} = 4 & x_{22} = 5 & x_{32} = 4 & x_{42} = 3 \end{array}$$

and the data array $\mathbf{X}$ is

$$\mathbf{X} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

with four rows and two columns. ■

Descriptive Statistics

- Much of the information contained in the array data can be assessed by calculating certain summary numbers, known as *descriptive statistics*.
- *Sample mean*, $\bar{x}_j$, is defined as

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, \cdots, p.$$

- *Sample variance*, s_j^2, is defined as

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2, \quad j = 1, \cdots, p.$$

- *Sample standard deviation*, s_j, is defined as the root of sample variance $\sqrt{s_j^2}$.
- *Sample covariance*, s_{ij}, is defined as

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad j, k = 1, \cdots, p.$$

Descriptive Statistics

- *Sample correlation (Pearson's correlation coefficient)*, r_{jk}, is defined as

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \quad j, k = 1, \cdots, p.$$

- The value of r must be between -1 and $+1$ inclusive.
- Here r measures the strength of the **linear** association.
- The value of r_{jk} remains unchanged if the measurements x_{ij}'s are subject to affine transformation.
- These sample statistics can be quickly computed in `R` by, e.g. `mean`, `sd`, `cov`, and `cor`.

Descriptive Statistics: Matrix Representation

- *Sample mean*, $\bar{\mathbf{X}}$, can be represented as

- *Sample covariance*, $\mathbf{S}$, can be represented as

- How about *sample variance*, $\mathbf{s}^2$?

Scatter Plots: `plot, points`

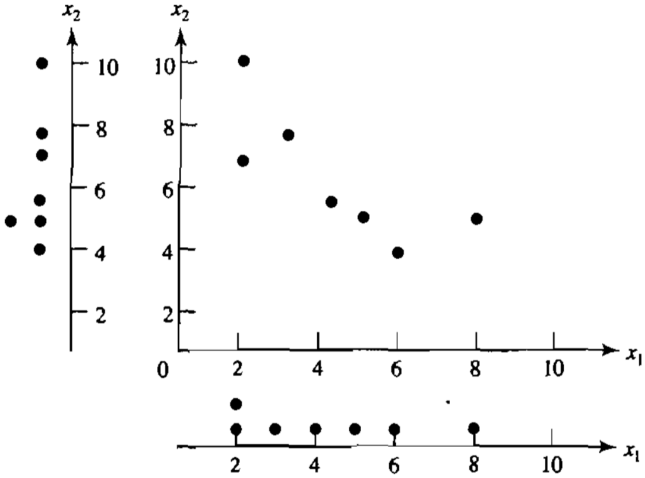

Figure 1.2 Scatter plot and dot diagrams for rearranged data.

3D Plots: `scatterplot3d`

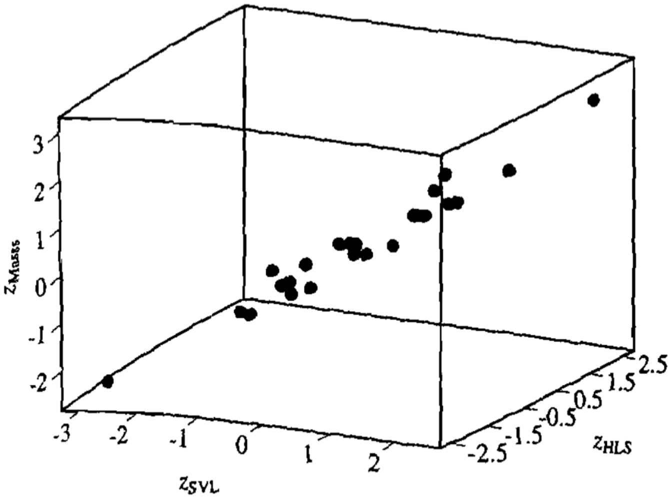

Figure 1.7 3D scatter plot of standardized lizard data.

Pairwise Plots: `pairs`

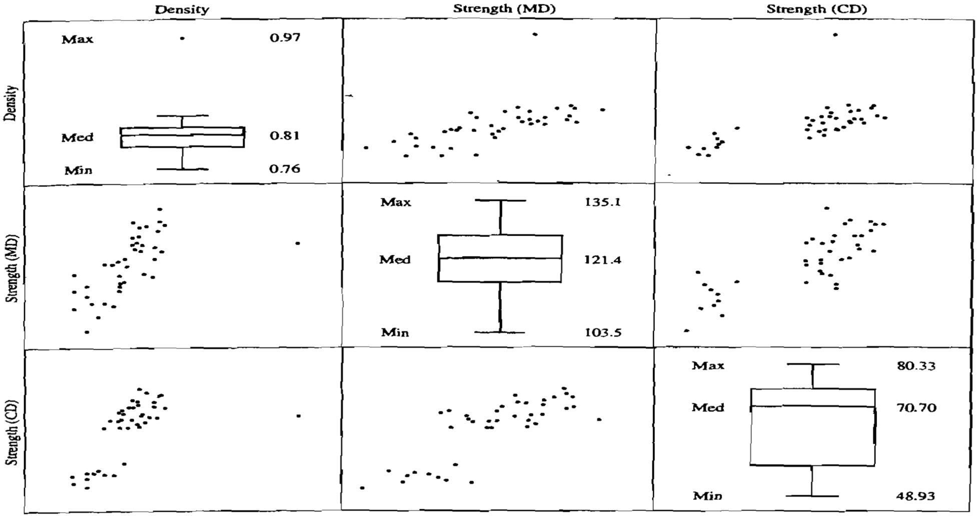

Figure 1.5 Scatter plots and boxplots of paper-quality data from Table 1.2.

Line Plots: `stars`

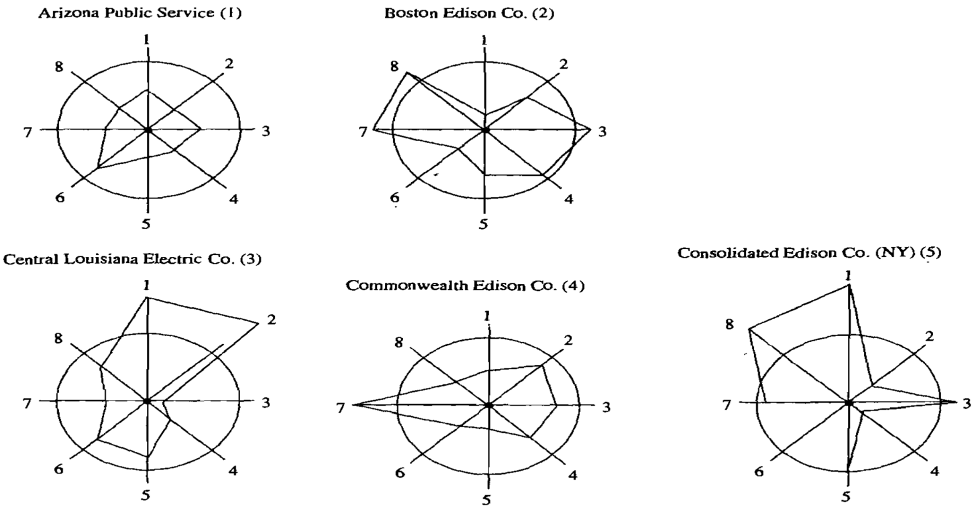

Figure 1.16 Stars for the first five public utilities.

Histogram and Density Plots: `hist`, `density`

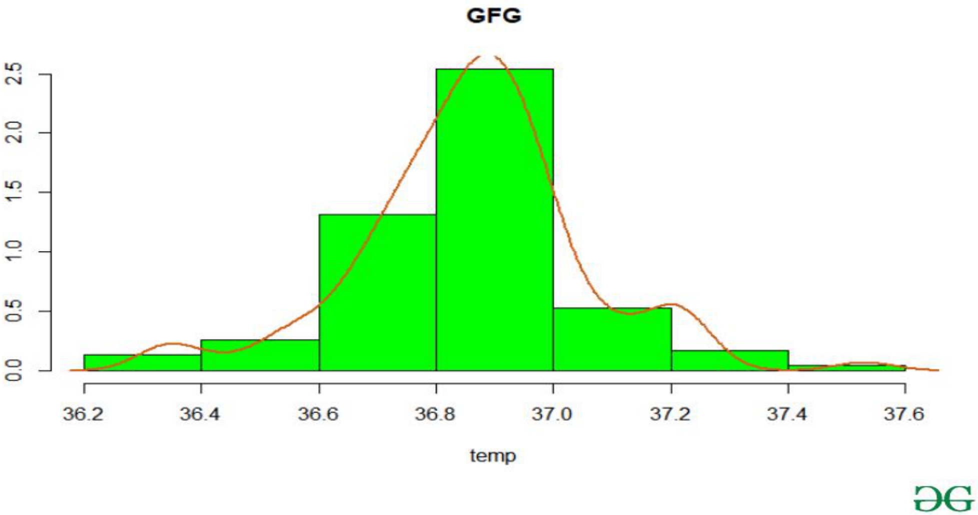

Table of Contents

Distance

- Most multivariate techniques are based upon the simple concept of *distance*, e.g., Euclidean distance.
- The straight-line distance between two points $P = \mathbf{x} = (x_1, \cdots, x_p)$ and $Q = \mathbf{y} = (y_1, \cdots, y_p)$ is given as

$$d(P, Q) = \|\mathbf{x} - \mathbf{y}\|_2 = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$$

- But sometimes the Euclidean distance is insufficient.

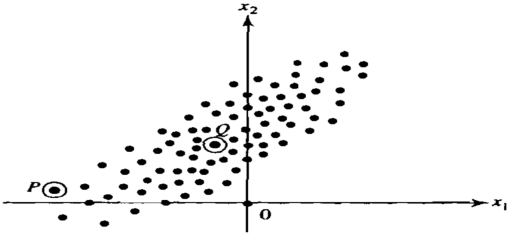

Figure 1.25 A cluster of points relative to a point P and the origin.

Distance

- Mathematically, a distance measures satisfies the following properties
  - $d(P, Q) \geq 0$ and $d(P, Q)$ implies $P = Q$.
  - $d(P, Q) = d(Q, P)$.
  - $d(P, Q) \leq d(P, R) + d(R, Q)$.
- Since we concern more about statistical relationship between variables in multivariate analysis, we need *statistical distance* based on sample variance and covariances

$$d^*(P, Q) = \sqrt{\sum_{j=1}^{p} \frac{(x_j - y_j)^2}{s_j^2}}$$

- This is a weighted distance with *elliptic contours*.

Statistical Distance

- In general, the statistical distance can be written as

$$d^*(P,Q) = \sqrt{\sum_{j,j'=1}^{p} a_{jj'}(x_j - y_j)(x_{j'} - y_{j'})} = \sqrt{(\mathbf{x}-\mathbf{y})^T \mathbf{A} (\mathbf{x}-\mathbf{y})}$$

- What is the matrix **A** in this case??

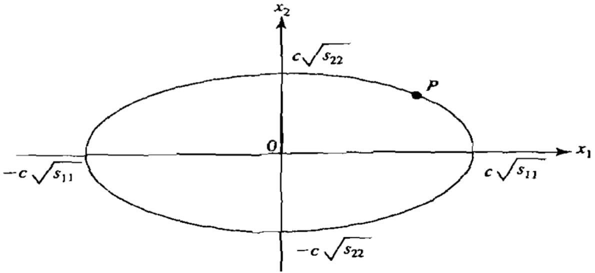

Figure 1.21 The ellipse of constant statistical distance $d^2(O,P) = x_1^2/s_{11} + x_2^2/s_{22} = c^2$.

Statistical Distance

- Consider a point $P = (x_1, x_2)$ and conduct a coordinate rotation as follows:

$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$
$$\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

- How does the Euclidean distance $d(P, O)$ change?
- How does the statistical distance $d^*(P, O)$ change?

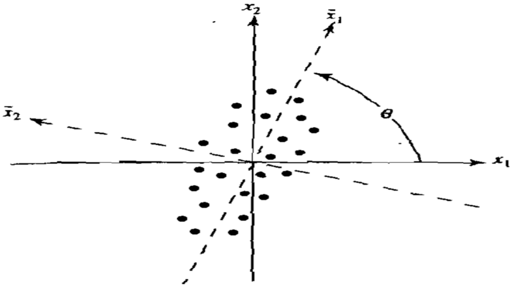

Figure 1.23 A scatter plot for positively correlated measurements and a rotated coordinate system.