

## CHAPTER 1

### Introduction to multivariate distributions-1

#### 1.1 Introduction

From the classic textbook of Anderson,

“Multivariate statistical analysis is concerned with data that consists of sets of measurements on a number of individuals or objects. The sample data may be heights and weights of some individuals drawn randomly from a population of schoolchildren in a given city, or the statistical treatment may be made on a collection of measurements, such as lengths and width of petals and lengths and widths of sepals of iris plants taken from three species, or one may study the scores on batteries of mental tests administered to a number of students.”

$p = \#$  of sets of measurements on a given individual,

$n = \#$  of observations = sample size.

**Example 1 :** Length and width of sepal and petal for three northern American species of iris are collected in a famous experiment started by R. A. Fisher. The three species are *iris setosa*, *iris versicolor* and *iris virginica*. We will use this data set later ( Can be obtained at the course webpage). There are 5 variables for each observation, among which the last one is categorical (1,2,3), merely indicating which species the observation comes from. Here,  $n = 150$  (50 for each species), and  $p = 5$ .

**Example 2 :** In an anthropological study, measurements were taken from the

indigenous populations in Africa on three variables, length, weight and volume of skull. The goal was to classify the populations in different tribes according to the observations. There were 200 observations from people across the whole continent.

**Remarks :** In both the above examples, one can assume that  $p \ll n$  since typically many measurements will be taken. Today it is common for  $p \gg 1$ , so  $n/p$  is no longer necessarily large.

**Example 3 :** Vehicle noise are stochastic signals. The power spectrum is discretized to a vector of length  $p = 1200$  with  $n \approx 1200$  samples from the same kind of vehicle.

**Example 4 :** Sloan Digital Sky Survey typically has many observations (say of quasar spectrum) with the spectra of each quasar binned, resulting in a large  $p$ .

**Example 5 :** Stock exchange observes 500 stocks for a specific brand over monthly intervals for twenty years.

### 1.1.1 Preliminaries

The multivariate observations are obtained in form of a vector (we'll use column vectors) with  $p$  dimensions. We write,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$$

for any single observation. The data comes in the form  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$  where each of the  $\mathbf{x}_i$ s are  $p$ -vectors  $(x_{i1}, x_{i2}, \dots, x_{ip})^t$ . The data matrix  $\mathbf{X}$  is stored as,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^t \\ \vdots \\ \mathbf{x}_n^t \end{bmatrix} = \begin{bmatrix} x_{11} \dots x_{1p} \\ \vdots \\ x_{n1} \dots x_{np} \end{bmatrix}$$

which is an  $n \times p$  matrix. the columns correspond to the variables, and the rows correspond to the observations.

The sample mean of a data set is given by

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}^t \mathbf{1}$$

and the sample variance-covariance matrix is

$$S_x = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t = \frac{1}{n-1} \mathbf{X}^t \left( I - \frac{1}{n} \mathbf{1}\mathbf{1}^t \right) \mathbf{X}$$

In this context, we also define the matrix of sums of squares and products (SSP) as

$$A_x = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t$$

i.e.  $S_x = (1/n-1)A_x$ . Sometimes, the matrix  $V_x = (1/n)A_x$  is also defined as the sample variance-covariance matrix.

Observe that  $\bar{\mathbf{x}}$  has components  $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ , the sample mean for the  $j$ th variable. Further, the variance-covariance matrix  $S_x$  has entries

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

the covariance between the  $j$ th and  $k$ th variable. We can also verify that  $S_x$  is a non-negative definite matrix. As in the two-variable case, we define the correlation between the  $j$ th and  $k$ th variables as,

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}$$

The matrix  $R = ((r_{jk}))$  will be known as the correlation matrix. Observe that it has diagonal entries 1, off-diagonal entries between -1 and 1, and is non-negative definite as well. Letting  $D = \text{diag}(s_{11}, \dots, s_{pp})$ , we have,

$$R = D^{-1/2} S D^{-1/2}$$

since  $D^{-1/2} = \text{diag}(\frac{1}{\sqrt{s_{11}}}, \dots, \frac{1}{\sqrt{s_{pp}}})$ . The correlation matrix is the covariance matrix for standardized variables.

*Linear combinations :*

Given any random vector  $\mathbf{x} \in \mathbb{R}^p$ , we define a new random vector  $\mathbf{y}$  obtained by taking linear multiples of the components and adding them, such as

$$\begin{aligned} y_1 &= a_{11}x_1 + \dots + a_{1p}x_p + b_1 \\ &\vdots \\ y_q &= a_{q1}x_1 + \dots + a_{qp}x_p + b_q \end{aligned}$$

In matrix notations, we write  $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ , where  $A = ((a_{ij}))$  and  $\mathbf{b} = (b_1, \dots, b_q)$ . Next, given any set of random vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n$  in  $\mathbb{R}^p$  we can similarly define a new set of random vectors s.t.  $\mathbf{y}_i = A\mathbf{x}_i + \mathbf{b}$ . Then, the mean and variance of the new set is given by

$$\bar{\mathbf{y}} = A\bar{\mathbf{x}} + \mathbf{b}, \text{ and } S_y = AS_xA^t$$

### 1.1.2 Graphical and pictorial techniques

1. Scatter plot in two or three dimensions.
2. Multiple scatter plots : A visual representation for the covariance matrix.  
In a  $p \times p$  grid, the off-diagonal subplots correspond to the scatter plot between those two variables, and in the diagonal grids, we have a box-plot for the particular row variable.
3. Current softwares are able to draw sophisticated plots like multiple-scatter plots, 3-D scatter plots, brush-spin plots (where you can rotate the axes to find out specific patterns in the data), growth curve plots.
4. Stars : When the data set consists of variables with positive measurements, the data can be represented on a multi-axis star with  $p$  equally spaced rays on a circle. The length of each ray represent the value of each variable, and each star will correspond to an observation.
5. Chernoff faces : As suggested by Chernoff,  $p$ -dimensional observations are represented in a 2-dimension face. The characteristics of the faces

(face shape, mouth curvature, nose length, eye size) correspond to the measurements of the p-variables. based on the similarity between faces, observers can visually detect patterns, clusters etc.

### 1.1.3 Distances

The Euclidean distance between two observations on a p-dimensional space is given by

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \dots + (x_p - y_p)^2}$$

However, this distance is unsatisfactory for several statistical purposes. This distance is greatly affected by scales, since the different variables are usually measured in different units. One of the first idea to remove that variability is to use standardized co-ordinates, such as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \dots + \frac{(x_p - y_p)^2}{s_{pp}}}$$

However, even this distance does not take into account the notion of dependence between the variables, which are manifested by the covariance matrix. This is taken into account when we define the Mahalanobis' distance as

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^t S^{-1} (\mathbf{x} - \mathbf{y})}$$

where the matrix  $S$  denote the covariance matrix of the observations.

*Remark :* Here, the distances are computed between two points with reference to a particular data cloud in a p-dimension space. The covariance matrix corresponds to the particular data cloud, and can be completely un-associated to the two points in question.

**Manhattan distance :** The Manhattan distance between two points  $\mathbf{x}$  and  $\mathbf{y}$  is defined as the length of the shortest path from  $\mathbf{x}$  to  $\mathbf{y}$  parallel to the axes. Namely,

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i|$$

This distance is analogous to the least absolute deviation.

**Canberra metric :** This is a distance defined between points with strictly positive co-ordinates. Namely,

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \frac{|x_i - y_i|}{x_i + y_i}$$