MULTIVARIATE ANALYSIS

M.Sc., STATISTICS First Year

SEMESTER-II, PAPER-I

LESSON WRITERS

Dr. A. Vasudeva Rao
Honorary Professor,
Department of Statistics,
Acharya Nagarjuna University

Dr. Syed Jilani
Guest Faculty
Department of Statistics
Acharya Nagarjuna University

Dr. S. Bhanu Prakash
Assistant Professor
Freshman Engineering Department
Godavari Global University
Rajamahendravaram - 533296.

Dr. U. Ramkiran
Departnent of Stetistics
Acharya Nagarjuna University

EDITOR

Dr. A. Vasudeva Rao
Honorary Professor,
Department of Statistics,
Acharya Nagarjuna University,

ACADEMIC ADVISOR

Prof. G. V. S. R. Anjaneyulu
Professor of Statistics (Retd.)
Acharya Nagarjuna University

DIRECTOR, I/c.

Prof. V. Venkateswarlu
M.A., M.P.S., M.S.W., M.Phil., Ph.D.

Centre for Distance Education
Acharya Nagarjuna University
Nagarjuna Nagar 522 510

Ph: 0863-2346222, 2346208
0863- 2346259 (Study Material)
Website www.anucde.info
E-mail: anucdedirector@gmail.com

M.Sc., STATISTICS : Multivariate Analysis

First Edition : 2025

No. of Copies :

This book is exclusively prepared for the use of students of M.Sc., STATISTICS Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited circulation only.

Published by:

Prof. V. VENKATESWARLU
Director, I/c
Centre for Distance Education,
Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University.

M.Sc. – Statistics Syllabus
SEMESTER-II
201ST24: Multivariate Analysis

UNIT-I:

The multivariate normal distribution and estimation: The multivariate normal distribution and its properties. Characteristic function of multivariate normal distribution. Sampling from multivariate normal distribution and maximum likelihood estimation, sampling distributions of Sample mean and sample covariance matrix.

UNIT-II:

Inference: Wishart's distribution and its properties. Definition of Hotelling's T^2-distribution (statistic). Invariance property of Hotelling's T^2-statistic. Application of T^2 statistic in tests of mean vector(s) in case of one and two multivariate normal populations. The likelihood ratio principle. Mahalanobis D^2-statistic and its relation with T^2-statistic. Multivariate analysis of variances (MANOVA) for one way classification.

UNIT-III:

Discriminant Analysis: Classification and discrimination procedures for discrimination between two multivariate normal populations, Fisher's discriminant function–separation of two multivariate populations. Classification with several multivariate normal populations. Fisher's method for discrimination among several multivariate populations.

UNIT-IV:

Cluster Analysis: Similarity measures, Euclidian distance and Mahalanobis squared distance-D^2 between two p-dimensional observations (items). Hierarchical Clustering methods - Single Linkage, Complete Linkage, Average Linkage, Ward's method and Centroid Linkage methods. Non-Hierarchical Clustering methods-K-Means method. Multidimensional scaling.

UNIT-V:

Special topics: Principle components analysis - definition, derivation, properties and Computation. Canonical variates and canonical correlations - definition, derivation and computation. Factor Analysis - Orthogonal factor model, Methods of estimating factor loadings - the principal component method and maximum likelihood method of estimation. Factor rotation: orthogonal factor rotation, varimax rotation.

BOOKS FOR STUDY:

1) Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern
2) Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International Mardia, K.V. Multivariate Analysis

BOOKS FOR REFERENCES:

1) Gin. N. C. (1977): Multivariate Statistical Inference. Academic Press
2) Seber, G. A. F. (1984): Multivariate Observations. Wiley
3) Kshirsagar, A. M. (1972): Multivariate Analysis, Marcel Dekker
4) Morrison, D. F. (1976): Multivariate Statistical Methods, 2nd Ed. McGraw Hill
5) Muirhead, R. J. (1982): Aspects of Multivariate Statistical Theory, J. Wiley
6) Rao, C. R. (1973): Linear Statistical Inference and its Applications, 2nd ed. Wiley
7) Sharma S. (1996): Applied Multivariate Techniques, Wiley
8) Srivastava, M. S. and Khatri, C. G. (1979): An Introduction to Multivariate Statistics, North Holland.

SET-I **MODEL QUESTION PAPER** **(201ST24)**

M.Sc. DEGREE EXAMINATION

SECOND SEMESTER

STATISTICS

201ST24 :: MULTIVARIATE ANALYSIS

Time: 3 hours **Maximum: 70 marks**

ANSWER ONE QUESTION FROM EACH UNIT

(Each question carries equal marks)

1. (a) Define the p-variate normal distribution with mean vector μ and dispersion matrix Σ. Derive two important properties of the multivariate normal distribution.

   (b) Prove that the marginal distribution obtained from the multivariate normal distribution is normal.

2 (a) Define the characteristic function of a p-dimensional random variable. Obtain the characteristic function of multivariate normal distribution.

   (b) In the p-variate normal case, show that the sample mean vector and the sample covariance matrix are independently distributed.

UNIT–II

3 (a) Define Hotelling's T^2 statistic. Show that Hotelling's T^2 statistic can be used to test the equality of means of corresponding variables in two MVN populations having the same variance-covariance matrix.

   (b) Explain in detail the likelihood ratio principle.

4 (a) Stating the assumptions clearly, discuss the problem of comparing several multivariate normal population means.

   (b) State and prove the invariance property of Hotelling's T^2 statistic.

UNIT–III

5 (a) Describe the classification between two unknown multivariate normal populations.

   (b) Explain the problem of classification. Distinguish between discrimination and classification.

6 (a) Derive Fisher's linear discriminant function in case of two unknown p-variate populations.

   (b) Describe the method of classification of an individual into one of several p-variate normal populations having a common dispersion matrix ξ, where all the parameters are known.

UNIT–IV

7 (a) Distinguish between cluster analysis and discriminant analysis. Consider the hypothetical distance between pairs of five objects as follows.

$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$

   Cluster the five objects using single linkage method

   (b) Explain the following methods of cluster analysis.

   1) Centeroid Linkage method 2) K-means method.

8 (a) Explain various similarity measures. Explain complete linkage method.

(b) Explain non-hierarchical methods. Describe Ward's method in cluster analysis.

UNIT–V

9 (a) Define principal components. If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then explain how you would compute various principal components.

(b) Define Canonical variables and Canonical correlations. Explain how you estimate canonical correlations.

10 (a) Explain the orthogonal factor model. Explain the ML estimation method of factor loadings.

(b) te and prove two properties of principal components.

CONTENTS

LESSON -1

MULTIVARIATE NORMAL DISTRIBUTION

OBJECTIVES:

- Understand the concepts of the multivariate normal distribution and multivariate analysis and their importance in multivariate statistical analysis.
- Distinguish between different measurement scales**,** with special emphasis on metric and non-metric measurement scales.
- Identify and classify non-metric & metric measurement scales and understand their role in multivariate techniques.
- Understanding the Multivariate Normal (MVN) distribution**.**
- Learning the properties of Multivariate Normal (MVN) distribution**.**

STRUCTURE:

1.1 INTRODUCTION TO MULTIVARIATE ANALYSIS:

Multivariate analysis is not easy to define. Broadly speaking, it refers to all statistical methods that simultaneously analyze multiple measurements on each individual or object under investigation. Any simultaneous analysis of more than two variables can be loosely considered multivariate analysis. As such, many multivariate techniques are extensions of univariate analysis (analysis of single-variable distributions) and bivariate analysis (cross-classification, correlation, analysis of variance, and simple regression used to analyze two variables). For example, simple regression (with one predictor variable) is extended in the multivariate case to include several predictor

variables. Likewise, the single dependent variable found in analysis of variance is extended to include multiple dependent variables in multivariate analysis of variance. In many instances, multivariate techniques are a means of performing in a single analysis what once took multiple analyses using univariate techniques. Other multivariate techniques, however, are uniquely designed to deal with multivariate issues, such as factor analysis, which identifies the structure underlying a set of variables, or discriminant analysis, which differentiates among groups based on a set of variables.

One reason for the difficulty of defining multivariate analysis is that the term multivariate is not used consistently in the literature. Some researchers use multivariate simply to mean examining relationships between or among more than two variables. Others use the term only for problems in which all the multiple variables are assumed to have a multivariate normal distribution. To be considered truly multivariate, however, all the variables must be random and interrelated in such ways that their different effects cannot meaningfully be interpreted separately. Some authors state that the purpose of multivariate analysis is to measure, explain, and predict the degree of relationship among variates (weighted combinations of variables). Thus the multivariate character lies in the multiple variates (multiple combinations of variables), and not only in the number of variables or observations.

The multivariate normal (MVN) distribution plays a central role in multivariate statistical analysis, just as the univariate normal distribution does in classical statistics. Many real-world phenomena-such as measurements in biology, finance, engineering, and social sciences-naturally involve several correlated variables. The MVN distribution provides a powerful framework for modeling such jointly distributed random variables, capturing both their individual behaviors and the dependence structure among them.

1.1.1 SOME BASIC CONCEPTS OF MULTIVARIATE ANALYSIS:

Although multivariate analysis has its roots in univariate and bivariate statistics, the extension to the multivariate domain introduces additional concepts and issues that have particular relevance. These concepts range from the need for a conceptual understanding of the basic building block of multivariate analysis the variate to specific issues dealing with the types of measurement scales used and the statistical issues of significance testing and confidence levels. Each concept plays a significant role in the successful application of any multivariate technique.

The Variate: As previously mentioned, the building block of multivariate analysis is the variate, a linear combination of variables with empirically determined weights. The variables are specified by the researcher, whereas the weights are determined by the multivariate technique to meet a specific objective. A variate of n weighted variables (X_1 to X_n) can be stated mathematically as:

variate value = $w_1X_1 + w_2X_2 + w_3X_3 + ... + w_nX_n$

where X is the observed variable and w is the weight determined by the multivariate technique.

The result is a single value representing a combination of the entire set of variables that best achieves the objective of the specific multivariate analysis. In multiple regression, the variate is determined so as to best correlate with the variable being predicted. In discriminant analysis, the variate is formed so as to create scores for each observation that

maximally differentiates among groups of observations. In factor analysis, variates are formed to best represent the underlying structure or dimensionality of the variables as represented by their inter correlations.

In each instance, the variate captures the multivariate character of the analysis. Thus, in our discussion of each technique, the variate is the focal point of the analysis in many respects. We must understand not only its collective impact in meeting the technique's objective but also each separate variable's contribution to the overall variate effect.

1.1.2 MEASUREMENT SCALES:

Data analysis involves the partitioning, identification, and measurement of variation in a set of variables, either among themselves or between a dependent variable and one or more independent variables. The key word here is measurement because the researcher cannot partition or identify variation unless it can be measured. Measurement is important in accurately representing the concept of interest and is instrumental in the selection of the appropriate multivariate method of analysis. Next we discuss the concept of measurement as it relates to data analysis and particularly to the various multivariate techniques.

There are two basic kinds of data: non-metric (qualitative) and metric (quantitative). Non-metric data are attributes, characteristics, or categorical properties that identify or describe a subject. Non-metric data describe differences in type or kind by indicating the presence or absence of a characteristic or property. Many properties are discrete in that by having a particular feature, all other features are excluded; for example, if one is male, one cannot be female. There is no "amount" of gender, just the state of being male or female. In contrast, metric data measurements are made so that subjects may be identified as differing in amount or degree. Metrically measured variables reflect relative quantity or degree. Metric measurements are appropriate for cases involving amount or magnitude, such as the level of satisfaction or commitment to a job.

1.1.3 NON-METRIC MEASUREMENT SCALES:

Non-metric measurements can be made with either a nominal or an ordinal scale. Measurement with a nominal scale assigns numbers used to label or identify subjects or objects in each category. Nominal scales, also known as categorical scales, provide the number of occurrences or symbols assigned to the objects that have no quantitative meaning beyond indicating the presence or absence of an attribute or characteristic. Therefore, the numbers on the nominally scaled data include no inherent meaning beyond categorization. Examples of nominally scaled data include an individual's sex, religion, or political party. In working with these data, the researcher might assign numbers to each category or class, for example, 2 for females and 1 for males. These numbers only represent categories or classes and do not imply amounts of an attribute or characteristic.

Ordinal scales are the next higher level of measurement precision. Variables can be ordered or ranked with ordinal scales in relation to the amount of the attribute possessed. Every subclass example, different levels of an individual consumer's satisfaction with several new products can be illustrated on an ordinal scale. Numbers utilized in ordinal scales such as these are non-quantitative because they indicate only relative positions in an ordered series. There is no measure of how much satisfaction the consumer receives in absolute terms, nor does the researcher know the exact difference between points on the scale of

satisfaction. Many scales in the behavioral sciences fall into this ordinal category.

1.1.4 METRIC MEASUREMENT SCALES:

Interval scales and ratio scales (both metric) provide the highest level of measurement precision, permitting nearly all mathematical operations to be performed. These two scales have constant units of measurement, so differences between any two adjacent points on any part of the scale are equal. The only real difference between interval and ratio scales is that interval scales have an arbitrary zero point, whereas ratio scales have an absolute zero point. The most familiar interval scales are the Fahrenheit and Celsius temperature scales. Each has a different arbitrary zero point, and neither indicates a zero amount or lack of temperature, because we can register temperatures below the zero point on the scale. Therefore, it is not possible to say that any value on an interval scale is a multiple of some other point on the scale. For example, an 80°F day cannot correctly be said to be twice as hot as a 40°F day, because we know that 80°F, on a different scale, such as Celsius, is 26.7°C. Similarly, 40°F on Celsius is 4.4°C. Although 80°F is indeed twice 40°F, one cannot state that the heat of 80°F is twice the heat of 40°F because, using different scales, the heat is not twice as great; that is, $4.4°C \times 2 \neq 26.7°C$.

Ratio scales represent the highest form of measurement precision because they possess the advantages of all lower scales plus an absolute zero point. All mathematical operations are permissible with ratio-scale measurements. The bathroom scale or other common weighing machines are examples of these scales, for they have an absolute zero point and can be spoken of in terms of multiples when relating one point on the scale to another; for example, 100 pounds is twice as heavy as 50 pounds.

Understanding the different types of measurement scales is important for two reasons. First, the researcher must identify the measurement scale of each variable used, so that non-metric data are not incorrectly used as metric data and vice versa. Second, the measurement scale is critical in determining which multivariate techniques are the most applicable to the data, with considerations made for both independent and dependent variables. In the discussion of the techniques and their classification in later sections of this chapter, the metric or non-metric properties of independent and dependent variables are the determining factors in selecting the appropriate technique.

1.1.5 MEASUREMENT ERROR & MULTIVARIATE MEASUREMENT:

The use of multiple variables and the reliance on their combination (the variate) in multivariate techniques also focuses attention on a complementary issue—measurement error. Measurement error is the degree to which the observed values are not representative of the “true” values. Measurement error has many sources, ranging from data entry errors to the imprecision of the measurement (e.g., imposing seven-point rating scales for attitude measurement when the researcher knows the respondents can accurately respond only to a three-point rating) to the inability of respondents to accurately provide information (e.g., responses as to household income may be reasonably accurate but rarely totally precise). Thus, all variables used in multivariate techniques must be assumed to have some degree of measurement error. The impact of measurement error is to add “noise” to the observed or measured variables. Thus, the observed value obtained represents both the “true” level and the “noise.” When used to compute correlations or means, the “true” effect is partially masked by the measurement error, causing the correlations to weaken and the means to be

less precise. The specific impact of measurement error and its accommodation in dependence relationships.

The researcher's goal of reducing measurement error can follow several paths. In assessing the degree of measurement error present in any measure, the researcher must address both the validity and reliability of the measure. Validity is the degree to which a measure accurately represents what it is supposed to. For example, if we want to measure discretionary income, we should not ask about total household income. Ensuring validity starts with a thorough understanding of what is to be measured and then making the measurement as "correct" and accurate as possible. However, accuracy does not ensure validity. In the above example, the researcher could very precisely define total household income but still have an invalid measure of discretionary income because the "correct" question was not being asked.

If validity is assured, the researcher must still consider the reliability of the measurements. Reliability is the degree to which the observed variable measures the "true" value and is "error free"; thus, it is the opposite of measurement error. If the same measure is asked repeatedly, for example, more reliable measures will show greater consistency than less reliable measures. The researcher should always assess the variables being used and if valid alternative measures are available, choose the variable with the higher reliability.

The researcher may also choose to develop multivariate measurements, also known as summated scales, for which several variables are joined in a composite measure to represent a concept (e.g., multiple-item personality scales or summed ratings of product satisfaction). The objective is to avoid the use of only a single variable to represent a concept, and instead to use several variables as indicators, all representing differing facets of the concept to obtain a more "well-rounded" perspective. The use of multiple indicators allows the researcher to more precisely specify the desired responses. It does not place total reliance on a single response, but instead on the "average" of "typical" response to a set of related responses. For example, in measuring satisfaction, one could ask a single question, "How satisfied are you?" and base the analysis on the single response. or a summated scale could be developed that combined several responses of satisfaction, perhaps in different response formats and in differing areas of interest thought to comprise overall satisfaction. The guiding premise is that multiple responses reflect the "true" response more accurately than does a single response. Assessing reliability and incorporating scales in the analysis are methods the researcher should employ. The impact of measurement error and poor reliability cannot be directly seen because they are embedded in the observed variables. The researcher must therefore always work to increase reliability and validity, which in turn will result in a "truer" portrayal of the variables of interest. Poor results are not always due to measurement error, but the presence of measurement error is guaranteed to distort the observed relationships and make multivariate techniques less powerful. Reducing measurement error, although it takes effort, time, and additional resources, may improve weak or marginal results and strengthen proven results as well.

1.2 APPLICATIONS OF MULTIVARIATE TECHNIQUES:

The published applications of multivariate methods have increased tremendously in recent years. It is now difficult to cover the variety of real-world applications of these methods with brief discussions, as we did in earlier editions of this book. However, in order to give some indication of the usefulness of multivariate techniques, we offer the following

short descriptions of the results of studies from several disciplines. These descriptions are organized according to the categories of objectives given in the previous section. Of course, many of our examples are multifaceted and could be placed in more than one category.
Data reduction or simplification

- Using data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed.
- Track records from many nations were used to develop an index of performance for both male and female athletes.
- 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.
- Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants.
- A matrix of tactic similarities was developed from aggregate data derived from professional mediators. From this matrix the number of dimensions by which professional mediators judge the tactics they use in resolving disputes was determined.

Sorting and grouping

- Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization.
- Measurements of several physiological variables were used to develop a screening procedure that discriminates alcoholics from non alcoholics.
- Data related to responses to visual stimuli were used to develop a rule for separating people suffering from a multiple-sclerosis-caused visual pathology from those not suffering from the disease.
- 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.

Investigation of the dependence among variables

- Data on several variables were used to identify factors that were responsible for success in hiring external consultants.
- Measures of variables related to innovation, on the one hand, and variables related to business environment and business organization, on the other hand, were used to discover why some firms are innovative and some firms are not.
- Data on variables representing the outcomes of the 10 decathlon events in the Olympics were used to determine the physical factors responsible for success in the decathlon.
- 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.

Prediction

- The associations between test scores and several high school performance variables and several college performance variables were used to develop predictors of success in college.
- 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 financial variables were used to develop a method for identifying potentially insolvent property-liability insurers.

- Data on several variables for chickweed plants were used to develop a method for predicting the species of a new plant.

Hypotheses testing

- Several pollution-related variables were measured to determine whether levels for a large metropolitan area were very consistent throughout the period 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.
- Data on many variables were used to investigate the differences in structure of American occupations to determine the support for one of two competing sociological theories.
- Data on several variables were used to determine whether different types of firms in newly industrialized countries exhibited different patterns of innovation.
- The preceding descriptions offer glimpses into the use of multivariate methods in widely diverse fields.

1.3 THE ORGANIZATION OF DATA:

Throughout this lesson, the reader is going to be concerned with analyzing measurements obtained on several variables. As mentioned in the introduction, the data are usually obtained from a sample of some population. That is, we measure or observe the values of p variables for each of n experimental units or individuals. This lesson is intended to introduce the preliminary concepts underlying these first steps of data collection, property measurement (definition), and the organization of the data.

Arrays

Multivariate data arise whenever an investigator is seeking to understand a social or physical phenomenon based on a number of measurements. The principal focus is on understanding the relationships among variables all recorded for each distinct individual or experimental unit in the study.

We will use the notation x_{jk} to indicate the particular value of the k[th] variable that is observed on the j[th] item.

x_{jk} =measurement of the k[th] variable on the j[th] item

Consequently, measurements on variables can be displayed as follows:

	Variable 1	*Variable 2*	...	*Variable k*	⋯	*Variable p*
Item 1:	x_{11}	x_{12}	⋯	x_{1k}	...	x_{1p}
Item 2:	x_{21}	x_{22}	⋯	x_{2k}	...	x_{2p}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
Item j:	x_{j1}	x_{j2}	⋯	x_{jk}	⋯	x_{jp}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
Item n:	x_{n1}	x_{n2}	⋯	x_{nk}	⋯	x_{np}

Or we can display these data as a rectangular array called $\mathbf{X}$ of n rows and p columns:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{k} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{pmatrix}$$

The array $\mathbf{X}$, then, contains the data consisting of all of the observations on all of the variables.

Example 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 the number of books sold. 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:

$x_{11} = 42 \quad x_{21} = 52 \quad x_{31} = 48 \quad x_{41} = 58$

$x_{12} = 4 \quad x_{22} = 5 \quad x_{32} = 4 \quad x_{42} = 3$

and the data array **X** is:

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

with four rows and two columns.

Considering data in the form of arrays facilitates the exposition of the subject matter and allows numerical calculations to be performed in an orderly and efficient manner. The efficiency is twofold, as gains are attained in both describing the numerical calculations as operations on arrays and the implementation of the calculations on computers, which now use many languages and statistical packages to perform array operations. We consider the manipulation of arrays of numbers. At this point, we are concerned only with their value as devices for displaying data.

Example 2 (The arrays $\bar{\mathbf{X}}, \mathbf{S_n}$, and R for bivariate data)

Consider the data introduced in Example 1, Each receipt yields a pair of measurements, total dollar sales, and number of books sold. Find the arrays $\bar{\mathbf{X}}, \mathbf{S_n}$, **and R** .

Since there are four receipts, we have a total of four measurements (observations) on each variable.

The sample means are:

$$\overline{x_1} = \frac{1}{4}\sum_{j=1}^{4} x_{j1} = \frac{1}{4}(42 + 52 + 48 + 58) = 50$$

$$\overline{x_2} = \frac{1}{4}\sum_{j=1}^{4} x_{j2} = \frac{1}{4}(4+5+4+3) = 4$$

$$\overline{X} = \begin{pmatrix} \overline{x_1} \\ \overline{x_2} \end{pmatrix} = \begin{pmatrix} 50 \\ 4 \end{pmatrix}$$

The sample variances and covariances are:

$$S_{11} = \frac{1}{4}\sum_{j=1}^{4}\left(x_{j1} - \overline{x_1}\right)^2 = \frac{1}{4}\left((42-50)^2 + (52-50)^2 + (48-50)^2 + (58-50)^2\right) = 34$$

$$S_{22} = \frac{1}{4}\sum_{j=1}^{4}\left(x_{j2} - \overline{x_2}\right)^2 = \frac{1}{4}\left((4-4)^2 + (5-4)^2 + (4-4)^2 + (3-4)^2\right) = .5$$

$$S_{12} = \frac{1}{4}\sum_{j=1}^{4}\left(x_{j1} - \overline{x_1}\right)\left(x_{j2} - \overline{x_2}\right)$$

$$= \frac{1}{4}\left((42-50)(4-4) + (52-50)(5-4) + (48-50)(4-4) + (58-50)(3-4)\right) = -1.5$$

$$S_{21} = S_{12}$$

and $\mathbf{S_n} = \begin{pmatrix} 34 & -1.5 \\ -1.5 & .5 \end{pmatrix}$

The sample correlation is

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = \frac{-1.5}{\sqrt{34}\sqrt{.5}} = -.36$$

$$r_{21} = r_{12}$$

So, $R = \begin{pmatrix} 1 & -.36 \\ -.36 & 1 \end{pmatrix}$

1.4 MULTIVARIATE NORMAL DISTRIBUTION:

In multivariate analysis, the MVN distribution serves as the foundation for numerous methods, including principal component analysis (PCA), discriminant analysis, regression analysis, confidence region construction, and hypothesis testing. Its mathematical tractability, especially regarding linear transformations and conditional distributions, makes it one of the most widely used models in theory and practice.

A thorough understanding of the MVN distribution includes knowledge of its definition, key properties, and associated matrix algebra. Equally important is the estimation of its parameters-the mean vector and covariance matrix-which forms the basis for inferential procedures in multivariate settings. This lesson introduces the multivariate normal distribution and explores its essential properties., and discusses parameter estimation techniques under this model.

1.5 SYMBOLS AND NOTATIONS:

To describe multivariate quantities, the following symbols and notations are commonly used:

- $\mathbf{X}$: Random vector of order $p \times 1$.
- μ: Mean vector of order $p \times 1$.
- Σ: Covariance matrix of order $p \times p$, symmetric and positive definite.
- Σ^{-1}: Inverse of Σ, also called the precision matrix.
- $|\Sigma|$: Determinant of Σ.
- $N_p(\mu, \Sigma)$: p-variate normal distribution with mean μ and covariance Σ.
- $\mathbf{x}$: Realization of the random vector $\mathbf{X}$.
- $E(\mathbf{X}) = \mu$ and $\mathrm{Cov}(\mathbf{X}) = \Sigma$.
- Superscript T denotes matrix transpose.

1.6 UNDERSTANDING MVN DISTRIBUTION:

Suppose X is a scalar normal variate with mean μ and variance σ^2 then the p.d.f of X can be written as

$$f(x:\mu,\sigma) = ke^{-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)}, \sigma^2 > 0, -\infty < \mu < \infty \quad \rightarrow (1)$$

Where, $k = \frac{1}{\sigma\sqrt{2\pi}}$

Now suppose $\underset{\sim}{X} = \begin{pmatrix} X_1 \\ X_2 \\ : \\ X_p \end{pmatrix}$ is a p-variate random vector and

Its mean vector is given by

$$E(\underset{\sim}{\mathbf{X}}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ : \\ E(X_p) \end{bmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ : \\ \mu_p \end{pmatrix} = \underset{\sim}{\boldsymbol{\mu}} \quad \rightarrow (2)$$

and its variance –covariance matrix is given by

$$V(\underset{\sim}{\mathbf{X}}) = E\left[\left(\underset{\sim}{\mathbf{X}} - E(\underset{\sim}{\mathbf{X}})\right)\left(\underset{\sim}{\mathbf{X}} - E(\underset{\sim}{\mathbf{X}})\right)'\right]$$

$$= E[(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}})']$$

$$V(\underset{\sim}{X})=\begin{bmatrix} E[X_1-\mu_1]^2 & E[(X_1-\mu_1)(X_2-\mu_2)] & \dots & E[(X_1-\mu_p)(X_p-\mu_p)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[X_2-\mu_2]^2 & \dots & E[(X_2-\mu_2)(X_p-\mu_p)] \\ \vdots & \vdots & \vdots & \vdots \\ E[(X_p-\mu_p)(X_1-\mu_1)] & E[(X_p-\mu_p)(X_2-\mu_2)] & \dots & E[X_p-\mu_p]^2 \end{bmatrix}$$

$$=\begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{bmatrix} = \boldsymbol{\Sigma} \text{ (say)} \qquad \rightarrow (3)$$

Where,

$$\sigma_{ij} = E\left[(x_i-\mu_i)(x_j-\mu_j)\right] = \sigma_{ji}$$

clearly, $\boldsymbol{\Sigma}$ is symmetric & positive definite matrix.

Now the multivariate normal density of $\underset{\sim}{\mathbf{X}}$ can be obtained by replacing the positive quantity $(x-\mu)(\sigma^2)^{-1}(x-\mu)$ by the quadratic form

$$(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}) \qquad \rightarrow (4)$$

and is given by

$$f(\underset{\sim}{\mathbf{x}}:\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}) = ke^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})} \qquad \rightarrow (5)$$

Where (k>0) is chosen so that the integral over the entire p-dimensional

Euclidean space of $X_1, X_2, ..., X_p$ is unity. we observe that

$$f(\underset{\sim}{\mathbf{x}}:\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}) \geq 0 \qquad (\because \text{k is chosen as positive})$$

since $\boldsymbol{\Sigma}$ is positive definite

$$(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}) > 0$$

$$\Rightarrow \quad -\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}) < 0$$

$$\Rightarrow \quad e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})} < e^0 = 1$$

i.e. $0 \leq f(\underset{\sim}{\mathbf{x}}:\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}) \leq k$ i.e. $f(\underset{\sim}{\mathbf{x}})$ is bounded.

Now we should find k(>0) such that

$$\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} f(x) = k\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\Sigma^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})}\,\mathbf{d}\underset{\sim}{\mathbf{x}} = 1 \quad \rightarrow (6)$$

$$k^{-1} = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\Sigma^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})}\,\mathbf{d}\underset{\sim}{\mathbf{x}} \quad \rightarrow (7)$$

since $\boldsymbol{\Sigma}^{-1}$ is positive definite $\exists$ a non singular matrix A such that

$$\boldsymbol{\Sigma}^{-1} = A'A \quad \rightarrow (8)$$

then (7) can be written as

$$k^{-1} = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'A'A(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})}\,\mathbf{d}\underset{\sim}{\mathbf{x}} \quad \rightarrow (9)$$

If we use the linear transformation from $\underset{\sim}{\mathbf{X}}$ to a new random vector $\underset{\sim}{\mathbf{Y}}$ such that

$$\underset{\sim}{\mathbf{Y}} = A(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}) \quad \rightarrow (10)$$

then (9) becomes

$$k^{-1} = J(\underset{\sim}{\mathbf{x}})\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}}\,\mathbf{d}\underset{\sim}{\mathbf{y}} \quad \rightarrow (11)$$

where $J(\underset{\sim}{\mathbf{x}})$ is the Jacobian obtained when $\underset{\sim}{\mathbf{X}}$ is transformed into $\underset{\sim}{\mathbf{Y}}$ and is given by

$$J(\underset{\sim}{\mathbf{x}}) = \operatorname{mod}\left|\frac{d\underset{\sim}{\mathbf{y}}}{d\underset{\sim}{\mathbf{x}}}\right|$$

$$= \operatorname{mod}\begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_p} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_p} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \dots & \frac{\partial y_p}{\partial x_p} \end{vmatrix}$$

$$= \operatorname{mod}\left|A^{-1}\right|$$

$$= \frac{1}{\operatorname{mod}|A|} \quad \left(\because \left|A^{-1}\right| = \frac{1}{|A|}\right)$$

Where $|A|$ is determinant of A.

$\therefore$ Equation (11) becomes

$$k^{-1} = \frac{1}{\text{mod}|A|} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}} \mathbf{d}\underset{\sim}{\mathbf{y}}$$

$$= \frac{1}{\text{mod}|A|} \prod_{i=1}^{p} \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}y_i^2} dy_i \right)$$

$$= \frac{1}{\text{mod}|A|} \prod_{i=1}^{p} \sqrt{2\pi} \quad (\because \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_i^2} dy_i = 1)$$

$$= \frac{(2\pi)^{p/2}}{\text{mod}\left|\mathbf{\Sigma}^{-1}\right|^{1/2}} \qquad (\because \left|\mathbf{\Sigma}^{-1}\right| = |A'A| = |A|^2)$$

i.e.,

$$k = \frac{1}{|\mathbf{\Sigma}|^{1/2}(2\pi)^{p/2}}$$

($\because \mathbf{\Sigma}^{-1}$ is positive definite to $\text{mod}\left|\mathbf{\Sigma}^{-1}\right|^{1/2} = \left|\mathbf{\Sigma}^{-1}\right|^{1/2} = \frac{1}{|\mathbf{\Sigma}|^{1/2}}$)

substituting k in (5) we get the p.d.f of the random normal vector $\underset{\sim}{\mathbf{X}}$ and is given by

$$f(\underset{\sim}{\mathbf{x}} : \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{|\mathbf{\Sigma}|^{1/2}(2\pi)^{p/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}} - \boldsymbol{\mu})} \qquad \rightarrow (12)$$

thus (12) is the p.d.f of a multivariate normal vector $\underset{\sim}{\mathbf{X}}$ whose mean vector

and variance-covariance matrix are respectively given by $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$

and is denoted as $n(\underset{\sim}{\mathbf{x}}/\boldsymbol{\mu}, \mathbf{\Sigma})$ and its distribution is denoted as $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$.

NOTE 1:

From (10) ,we may see that

$$E(\underset{\sim}{\mathbf{Y}}) = AE(\underset{\sim}{\mathbf{X}} - \boldsymbol{\mu})$$
$$= A(E(\underset{\sim}{\mathbf{X}}) - \boldsymbol{\mu})$$
$$= A(\boldsymbol{\mu} - \boldsymbol{\mu})$$
$$= \underset{\sim}{\mathbf{0}}$$

i.e. $\underset{\sim}{\mathbf{Y}}$ has zero mean vector.

The variance-covariance matrix $\underset{\sim}{\mathbf{Y}}$ of is given by

$$\text{V}(\underset{\sim}{\mathbf{Y}}) = E[(\underset{\sim}{\mathbf{y}} - E(\underset{\sim}{\mathbf{y}}))(\underset{\sim}{\mathbf{y}} - E(\underset{\sim}{\mathbf{y}}))']$$

$$=E[\underset{\sim}{y}\underset{\sim}{y}'] \quad (\because E(\underset{\sim}{y})=\underset{\sim}{0})$$
$$=E[A(\underset{\sim}{X}-\underset{\sim}{\mu})(\underset{\sim}{X}-\underset{\sim}{\mu})'A']$$
$$=AV(\underset{\sim}{X})A'$$
$$=A\Sigma A'$$

but from (8),

$$\Sigma = (A'A)^{-1}$$
$$= A^{-1}(A')^{-1} \quad (\because A \text{ is a non-singular})$$
$$\therefore V(\underset{\sim}{Y}) = AA^{-1}(A')^{-1}A'$$
$$= I_k I_k$$
$$= I_k$$

Thus if $\underset{\sim}{X}$ is $N_p(\underset{\sim}{\mu}, \Sigma)$,then the random vector $\underset{\sim}{Y}$ defined as $\underset{\sim}{Y} = A(\underset{\sim}{X} - \underset{\sim}{\mu})$ (where A is defined as in (8)) follows $N_p(\underset{\sim}{0}, I_k)$.

In other words , the individual element of $\underset{\sim}{Y}$ are standard normal variates and mutually independent i.e. $Y_i \sim N(0,1)$ with $\text{cov}(Y_i, Y_j) = 0$.

NOTE 2:

In the practical situations 'A' can be computed as follows .since Σ is a symmetric p.d. matrix we may write $\Omega'\Sigma\Omega = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$,when Ω is the normalized latent vector matrix and Λ is the latent root matrix and since Σ is p.d. all $\lambda_1, \lambda_2, ..., \lambda_p$ are positive .Therefore Λ can be written $\Lambda = (\Lambda^{1/2})'(\Lambda^{1/2})$

Where, $\Lambda^{1/2} = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, ..., \sqrt{\lambda_p})$

then

$$\Omega'\Sigma\Omega = (\Lambda^{1/2})'(\Lambda^{1/2})$$
$$\Rightarrow \Sigma = (\Omega^{-1})'(\Lambda^{1/2})'\Lambda^{1/2}\Omega^{-1} = A^{-1}(A^{-1})'$$

where

$$A^{-1} = (\Omega^{-1})'(\Lambda^{1/2})'$$
$$\Rightarrow A = \Lambda^{-1/2}\Omega' \quad (\because (\Lambda^{1/2})' = \Lambda^{1/2})$$

thus, $\underset{\sim}{Y} = \Lambda^{1/2}\Omega'(\underset{\sim}{X} - \underset{\sim}{\mu})$

where $\Lambda^{1/2} = diag(\frac{1}{\sqrt{\lambda_1}}, ..., \frac{1}{\sqrt{\lambda_p}})$

The transformation from $\underset{\sim}{X}$ to $\underset{\sim}{Y}$ follows $N_p(\underset{\sim}{0}, I_k)$ .This transformation is called "whitening".

Eq (7) is the p.d.f . of the multivariate normal variate $\underset{\sim}{X}$ where mean is $\underset{\sim}{\mu}$ and variance – covariance matrix is Σ and is denoted by $n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma)$.

The distribution function of $\underset{\sim}{X}$ is denoted as $N_p(\underset{\sim}{\mu}, \Sigma)$.

Definition: Suppose $\underset{\sim}{\mathbf{X}}$ is a random vector with mean $\underset{\sim}{\boldsymbol{\mu}}$ and the variance-covariance matrix , $\boldsymbol{\Sigma}$. Then $\underset{\sim}{\mathbf{X}}$ is said to follow multivariate normal distribution if its p.d.f. is given by

$$f(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})} \quad \rightarrow (1)$$

It is denoted as $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \Sigma)$.

1.7 PROPERTIES OF MVN DISTRIBUTION:

THEOREM 1:

Let $\underset{\sim}{\mathbf{X}}$ (with p components) be distributed according to $N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ then, $\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\mathbf{X}}$ is distributed according to $N(C\underset{\sim}{\boldsymbol{\mu}}, C\boldsymbol{\Sigma}C')$ for C non-singular.

PROOF:

Since $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ & its p.d.f. is given as

$$f(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})} \quad \rightarrow (1)$$

Now ,consider the linear transformation

$$\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\mathbf{X}} \text{ where C is non-singular}$$

$$\Rightarrow \underset{\sim}{\mathbf{X}} = C^{-1}\underset{\sim}{\mathbf{Y}} \quad \rightarrow (2)$$

Now the p.d.f. 1 becomes in terms of $\underset{\sim}{\mathbf{Y}}$ as

$$g(\underset{\sim}{\mathbf{y}}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(C^{-1}\underset{\sim}{\mathbf{y}}-\underset{\sim}{\mu})'\boldsymbol{\Sigma}^{-1}(C^{-1}\underset{\sim}{\mathbf{y}}-\underset{\sim}{\mu})} J(\underset{\sim}{\mathbf{y}}) \quad \rightarrow (3)$$

where $J(\underset{\sim}{\mathbf{y}})$ is the Jacobian and is given by

$$J(\underset{\sim}{\mathbf{y}}) = \text{mod}\left|\frac{\partial \underset{\sim}{\mathbf{x}}}{\partial \underset{\sim}{\mathbf{y}}}\right| = \text{mod}\left|C^{-1}\right|$$

$$= \frac{1}{\text{mod}|C|}$$

$$= \frac{1}{\sqrt{|C|^2}}$$

$$= \sqrt{\frac{|\boldsymbol{\Sigma}|}{|C||\boldsymbol{\Sigma}||C'|}}$$

$$= \frac{|\boldsymbol{\Sigma}|^{1/2}}{|C\boldsymbol{\Sigma}C'|^{1/2}} \quad \rightarrow (4)$$

Using (4) & (3) ,we get

$$g(\underset{\sim}{y}) = \frac{1}{(2\pi)^{p/2}\left|C\Sigma C'\right|^{1/2}} e^{-\frac{1}{2}(C^{-1}\underset{\sim}{y} - \underset{\sim}{\mu})'\Sigma^{-1}(C^{-1}\underset{\sim}{y} - \underset{\sim}{\mu})}$$

$$= \frac{1}{(2\pi)^{p/2}\left|C\Sigma C'\right|^{1/2}} e^{-\frac{1}{2}[C^{-1}(\underset{\sim}{y} - C\underset{\sim}{\mu})]'\Sigma^{-1}[C^{-1}(\underset{\sim}{y} - C\underset{\sim}{\mu})]}$$

$$= \frac{1}{(2\pi)^{p/2}\left|C\Sigma C'\right|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{y} - C\underset{\sim}{\mu})'(C\Sigma C')^{-1}(\underset{\sim}{y} - C\underset{\sim}{\mu})}$$

$$= n(\underset{\sim}{y} / C\underset{\sim}{\mu}, C\Sigma C') \quad \rightarrow (5)$$

But

$$E(\underset{\sim}{Y}) = CE(\underset{\sim}{X}) = C\underset{\sim}{\mu} \quad \rightarrow (6)$$

$$\& \ \mathrm{V}(\underset{\sim}{Y}) = C\mathrm{V}(\underset{\sim}{X})C' = C\sum C' \quad \rightarrow (7)$$

Now, if we write the multivariate normal p.d.f. of $\underset{\sim}{Y}$ with mean $\underset{\sim}{\mu}$ and the variance-covariance matrix $C\Sigma C'$ that will becomes as (5) and therefore

$$C\underset{\sim}{X} \sim N(C\underset{\sim}{\mu}, C\Sigma C').$$

Hence the proof.

THEOREM 2:

If a multivariate normal vector is divided into two sub vectors and one sub -vector is uncorrelated with other sub-vector ,then those two sub-vectors of variables are independent and each sub-vector is also a multivariate normal vector.

(OR)

Let $\underset{\sim}{X}_{p\times 1} \sim N_p(\underset{\sim}{\mu}, \Sigma) \& \underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{pmatrix}$

Where $\underset{\sim}{X}_1$ **is** $q\times 1$ **and** $\underset{\sim}{X}_2$ **is** $(p-q)\times 1$ **and** $\underset{\sim}{\mu} = \begin{pmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \end{pmatrix}$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where , Σ_{11} **is variance-covariance matrix of** $\underset{\sim}{X}_1$

Σ_{22} **is variance-covariance matrix of** $\underset{\sim}{X}_2$

and Σ_{12} **is covariance matrix of** $\underset{\sim}{X}_1 \& \underset{\sim}{X}_2$

Σ_{21} **is covariance matrix of** $\underset{\sim}{X}_2 \& \underset{\sim}{X}_1$.

Now if $\Sigma_{12} = \Sigma_{21}' = 0_{q\times p-q}$

then , $\underset{\sim}{X}_2$ & $\underset{\sim}{X}_1$ **are independent and** $\underset{\sim}{X}_1 \sim N_q(\underset{\sim}{\mu}_1, \Sigma_{11})$ & $\underset{\sim}{X}_2 \sim N_{p-q}(\underset{\sim}{\mu}_2, \Sigma_{22})$.

PROOF:

We are given $\Sigma_{12} = \mathbf{0}_{pxq} = \Sigma_{21}'$

i.e. the covariance matrix of $\underset{\sim}{X}_{p\times 1}$ is given by

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$$

In order to show that, the random vectors $\underset{\sim}{X}_1$ & $\underset{\sim}{X}_2$ are independently normally distributed, we have to show that

$$n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = n(\underset{\sim}{x}_1/\underset{\sim}{\mu}_1, \Sigma_{11}) n(\underset{\sim}{x}_2/\underset{\sim}{\mu}_2, \Sigma_{22})$$

we have,

$$n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})} \quad \rightarrow (1)$$

consider the Q.F in (1),

i.e., $Q = (\underset{\sim}{x}-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})$

$$= \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}' \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}$$

$$= \begin{bmatrix} (\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1)' & (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)' \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}$$

($\because \Sigma_{11}$ is the variance-covariance matrix of $\underset{\sim}{X}_1$ and hence positive definite)

$$= \begin{bmatrix} (\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1)'\Sigma_{11}^{-1} & (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1 \\ \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2 \end{bmatrix}$$

$$= (\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1)'\Sigma_{11}^{-1}(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1) + (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)$$

$$= Q_1 + Q_2 \quad \rightarrow (2)$$

also we have,

$$|\Sigma| = \begin{vmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{vmatrix} = |\Sigma_{11}||\Sigma_{22}| \quad \rightarrow (3)$$

Using (2) & (3) in (1) we get ,

$$n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{q/2}|\Sigma_{11}|^{1/2}} e^{-\frac{1}{2}Q_1} \frac{1}{(2\pi)^{(p-q)/2}|\Sigma_{22}|^{1/2}} e^{-\frac{1}{2}Q_2}$$

where Q_1 & Q_2 are is as given in (2),

$$\therefore n(\underset{\sim}{x}/\underset{\sim}{\mu},\Sigma) = n(\underset{\sim}{x_1}/\underset{\sim}{\mu}_1,\Sigma_{11}).n(\underset{\sim}{x_2}/\underset{\sim}{\mu_2},\Sigma_{22})$$

Thus ,the joint p.d.f. of the normal variates $X_1, X_2, ..., X_p$ is the product of the marginal p.d.f. of $X_1, X_2, ..., X_q$ and the marginal p.d.f. of $X_{q+1}, ..., X_p$. Thus, the two sets of normal variates are independent.

THEOREM 3:

If $\underset{\sim}{X}_1 \& \underset{\sim}{X}_2$ are independent and are distributed as $N_q(\underset{\sim}{\mu_1},\Sigma_{11}) \& N_{p-q}(\underset{\sim}{\mu}_2,\Sigma_{22})$ respectively then , $\begin{pmatrix}\underset{\sim}{X_1}\\ \underset{\sim}{X_2}\end{pmatrix} \sim N_p\left(\begin{bmatrix}\underset{\sim}{\mu_1}\\ \underset{\sim}{\mu_2}\end{bmatrix},\begin{bmatrix}\Sigma_{11} & 0\\ 0 & \Sigma_{22}\end{bmatrix}\right)$.

PROOF:-

we have given ,

$$\underset{\sim}{X_1} \sim N_q(\underset{\sim}{\mu_1},\Sigma_{11})$$

$$\underset{\sim}{X_2} \sim N_{p-q}(\underset{\sim}{\mu_2},\Sigma_{22})$$

and $\underset{\sim}{X}_1 \& \underset{\sim}{X}_2$ are independent i.e. $\underset{\sim}{X_1}, \underset{\sim}{X_2}$ are uncorrelated.

i.e. $\text{cov}(\underset{\sim}{X_1},\underset{\sim}{X_2}) = \Sigma_{12} = \Sigma_{21} = 0$.

We have to find out the joint p.d.f. of $f(\underset{\sim}{x})$ of $\underset{\sim}{X} = \begin{pmatrix}\underset{\sim}{X_1}\\ \underset{\sim}{X_2}\end{pmatrix}$

we have,

$g(\underset{\sim}{x}) = f(\underset{\sim}{x_1})f(\underset{\sim}{x_2})$ ($\because \underset{\sim}{X_1} \& \underset{\sim}{X_2}$ are independent)

$$= n(\underset{\sim}{x_1}/\underset{\sim}{\mu_1},\Sigma_{11}).n(\underset{\sim}{x_2}/\underset{\sim}{\mu_2},\Sigma_{22})$$

$$= \frac{1}{(2\pi)^{q/2}|\Sigma_{11}|^{1/2}} e^{-\frac{1}{2}Q_1} \frac{1}{(2\pi)^{(p-q)/2}|\Sigma_{22}|^{1/2}} e^{-\frac{1}{2}Q_2}, \text{ (where } Q_i = (\underset{\sim}{x_i} - \underset{\sim}{\mu_i})'\Sigma_{ii}^{-1}(\underset{\sim}{x_i} - \underset{\sim}{\mu_i}), \text{ i=1,2)}$$

$$= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(Q_1+Q_2)} \quad \left(\because |\Sigma| = \begin{vmatrix}\Sigma_{11} & 0\\ 0 & \Sigma_{22}\end{vmatrix} = |\Sigma_{11}||\Sigma_{22}|\right) \quad \rightarrow (1)$$

$$\text{Where } Q_1 + Q_2 = (\underset{\sim}{x_1} - \underset{\sim}{\mu_1})'\Sigma_{11}^{-1}(\underset{\sim}{x_1} - \underset{\sim}{\mu_1}) + (\underset{\sim}{x_2} - \underset{\sim}{\mu_2})'\Sigma_{22}^{-1}(\underset{\sim}{x_2} - \underset{\sim}{\mu_2}) \quad \rightarrow (2)$$

Let us consider

$$Q = (\underset{\sim}{x} - \underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}) \quad \rightarrow (3)$$

where , $\underset{\sim}{x} = \begin{pmatrix}\underset{\sim}{x_1}\\ \underset{\sim}{x_2}\end{pmatrix}, \underset{\sim}{\mu} = \begin{pmatrix}\underset{\sim}{\mu_1}\\ \underset{\sim}{\mu_2}\end{pmatrix}$ is $E(\underset{\sim}{X})$ and the variance-covariance matrix $\underset{\sim}{X}$ is

$$\Sigma = \begin{pmatrix} V(\underset{\sim}{X_1}) & \text{cov}(\underset{\sim}{X_1},\underset{\sim}{X_2}) \\ \text{cov}(\underset{\sim}{X_2},\underset{\sim}{X_1}) & V(\underset{\sim}{X_2}) \end{pmatrix}$$

$$= \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$$

But ,since $\Sigma_{12} = \Sigma'_{21} = \mathbf{0}_{qXp-q}$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \quad \rightarrow (4)$$

Now, $$Q = \begin{pmatrix} \underset{\sim}{x_1} - \underset{\sim}{\mu_1} \\ \underset{\sim}{x_2} - \underset{\sim}{\mu_2} \end{pmatrix}' \begin{pmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{pmatrix} \begin{pmatrix} \underset{\sim}{x_1} - \underset{\sim}{\mu_1} \\ \underset{\sim}{x_2} - \underset{\sim}{\mu_2} \end{pmatrix}$$

$$= (\underset{\sim}{x_1} - \underset{\sim}{\mu_1})'\Sigma_{11}^{-1}(\underset{\sim}{x_1} - \underset{\sim}{\mu_1}) + (\underset{\sim}{x_2} - \underset{\sim}{\mu_2})'\Sigma_{22}^{-1}(\underset{\sim}{x_2} - \underset{\sim}{\mu_2}) \quad \rightarrow (5)$$

from (2) & (5) , $Q_1 + Q_2 = Q$

$$\therefore g(\underset{\sim}{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}Q}$$

Where, Q is given by (3) but $g(\underset{\sim}{X})$ is nothing but $n(\underset{\sim}{x}/\underset{\sim}{\mu}, \Sigma)$.

Thus $\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X_1} \\ \underset{\sim}{X_2} \end{pmatrix} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ where, Σ is as given by (4).

THEOREM 4:

If $X_1, X_2, ..., X_p$ have a joint normal distribution , a necessary & sufficient condition for one subset of some random variables and the subset consisting of the remaining random variables be independent is that each covariance of a variable from one set and a variable from the other set be '0'.

PROOF:-

Necessary condition:

Without loss of generality let us assume that the first q variables form the first subset and the remaining p-q variables form the second subset.

In order to prove the necessary condition, we have given that the variables of $X_1, X_2, ..., X_q$ are independently distributed with the variables $X_{q+1}, X_{q+2}, ..., X_p$ and we have to prove

$$\text{cov}(X_i, X_j) = E[(X_i - E(X_i))(X_j - E(X_j))] = 0$$

where , $1 \le i \le q \;\&\; q+1 \le j \le p$

we have

$$\text{cov}(X_i, X_j) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_i - E(X_i)(x_j - E(X_j)) f(x_1, x_2, ..., x_p) dx_1 dx_2 ... dx_p$$

$$= \left(\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_i - E(X_i)) f_1(x_1 ... x_q) dx_1 ... dx_q \right).$$

$$\left(\int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} (x_j - E(X_j)) f_2(x_{q+1} ... x_p) dx_{q+1} ... dx_p \right)$$

$$\left(\because f(x_1, ..., x_p) = f_1(x_1, ..., x_q) f_2(x_{q+1}, ..., x_p) \right)$$

$$= E[X_i - E(X_i)]E[X_j - E(X_j)]$$
$$= [E(X_i) - E(X_i)][E(X_j) - E(X_j)]$$
$$= 0.0$$
$$= 0$$

Thus if one set of variables is independent of the remaining variables then, the set of variables are uncorrelated with the other set of variables.

Sufficient condition:

Here we have given

$$\underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix} \& \ \underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) \qquad \& \operatorname{cov}(X_i, X_j) = 0$$

where , X_i is from $\underset{\sim}{\mathbf{X_1}}$

X_j is from $\underset{\sim}{\mathbf{X_2}}$

i.e. $\operatorname{cov}(\underset{\sim}{\mathbf{X_1}}, \underset{\sim}{\mathbf{X_2}}) = \sum_{12} = 0_{q \times p-q}$ and we have to prove $\underset{\sim}{\mathbf{X_1}}$ **&** $\underset{\sim}{\mathbf{X}}_2$ are independently distributed.

The proof of this sufficient condition is given in Theorem 2 given above.

Note:

To prove the necessary condition of the above theorem we need not assume $X_1, ..., X_p$ are normally distributed.

1.8 SUMMARY:

In this lesson, the concept of the Multivariate Normal (MVN) distribution was introduced as a fundamental model for describing the joint behaviour of several correlated random variables. Beginning with basic symbols and notations**,** we established a clear mathematical framework for representing vectors, matrices, mean vectors, and covariance structures-elements essential for multivariate analysis.

The probability density function (p.d.f.) of the MVN distribution was presented both in its standard form and through an alternative method of derivation, highlighting the role of linear transformations of normal variables. These derivations illustrated how dependence among variables is incorporated through the covariance matrix, and how geometric features such as ellipsoidal contours arise naturally from the structure of the MVN density.

The section on Estimation in MVN Models discussed the methods used to estimate the mean vector and covariance matrix of a multivariate normal population. Maximum likelihood estimation (MLE) procedures were shown to provide efficient and unbiased estimators, while the sampling distributions of the sample mean vector and sample covariance matrix (Wishart distribution) were described. These results form the backbone of multivariate inference.

Overall, the Multivariate Normal distribution occupies a central position in multivariate statistical analysis. Its mathematical tractability, well-defined inferential properties, and broad applicability make it indispensable for modern statistical modeling. Understanding its density, derivations, estimation procedures, and applications equips students and researchers with a strong foundation for advanced multivariate methods.

1.9 SELF-ASSESSMENT QUESTIONS:

1. What is multivariate analysis? How does it differ from univariate and bivariate analysis?
2. What are measurement scales? How do they influence multivariate analysis?
3. Define the multivariate normal distribution and obtain the bivariate normal density as a particular case of MVN.
4. Let X be a p - variate normal random vector. State and prove a necessary and sufficient condition for one subset of the random variables and the subset consisting of the remaining variables to be independent.
5. Explain and derive any two properties of Multivariate normal distribution.
6. If a multivariate normal vector is divided into two sub vectors and one sub-vector is uncorrelated with other sub-vector, then those two sub-vectors of variables are independent and each sub-vector is also a multivariate normal vector.
7. **If** $\underset{\sim}{\mathbf{X}}_1 \& \underset{\sim}{\mathbf{X}}_2$ **are independent and distributed as** $N_q(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma}_{11}) \& N_{p-q}(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_{22})$ **respectively then ,** $\begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix} \sim N_p\left(\begin{bmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}\right)$.

1.9 SUGGESTED READINGS:

1. Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern.
2. Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International.
3. Mardia, Kent & Bibby. Multivariate Analysis
4. Kshirsagar, A.M. Multivariate Analysis
5. Brenner, D., Bilodeau, M. (1999). Theory of Multivariate Statistics. Germany: Springer.
6. Giri Narayan C. (1995). Multivariate Statistical Analysis.

Prof. A. Vasudeva Rao

LESSON -2

MARGINAL AND CONDITIONAL DISTRIBUTIONS

OBJECTIVES:

- Understand the concept and importance of marginal and conditional distributions in multivariate analysis.
- To derive marginal distribution in the context of the Multivariate Normal Distribution.
- To derive conditional distribution in the context of the Multivariate Normal Distribution.

STRUCTURE:

2.1 INTRODUCTION:

In multivariate analysis, the study of joint distributions of two or more random variables is essential for understanding their combined behavior. However, in many practical situations, interest lies in the behavior of a subset of variables or in the behavior of one variable given the values of others. This leads to the concepts of marginal and conditional distributions**,** which are fundamental tools in multivariate probability theory and statistical inference. These distributions play a crucial role in understanding dependence structures and in simplifying complex multivariate problems.

The Multivariate Normal (MVN) Distribution is a fundamental probability distribution in multivariate statistics. It generalizes the univariate normal distribution to higher dimensions and plays a central role in inference, estimation, classification, regression, and many applied statistical methods.
A random vector

$$X = \left(X_1, X_2, ..., X_p\right)'$$

is said to follow a multivariate normal distribution with mean vector μ and covariance matrix $\sum$, written as

$$X \sim N_p\left(\mu, \Sigma\right).$$

if every linear combination $a'X$ is univariate normally distributed.

2.2 MARGINAL DISTRIBUTION OF MVN DISTRIBUTION:

The marginal distribution of a subset of random variables is obtained from the joint distribution by integrating (or summing) over the remaining variables. It describes the

probability behavior of individual variables or groups of variables without reference to the others.

In multivariate analysis, marginal distributions help in:

- Understanding individual variable behaviour within a multivariate framework
- Examining the distributional properties of subsets of variables
- Establishing results such as the marginal normality of components of an MVN distribution.

Marginal distributions are especially important in simplifying multivariate problems and form the basis for many inferential procedures.

THEOREM:

If $\underset{\sim}{\mathbf{X}}$ has MVN, then any subset of the components of $\underset{\sim}{\mathbf{X}}$ have a (multivariate) normal distribution.

(OR)

Prove that the marginal distribution obtained from the multivariate normal distribution is normal.

(OR)

If $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$, then the marginal distribution of any (sub) set of components of $\underset{\sim}{\mathbf{X}}$ is multivariate normal with means, variances and co-variances obtained by taking the proper components of $\underset{\sim}{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ respectively.

PROOF:

$$\text{Let } \underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix}, \underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}_1} \\ \underset{\sim}{\boldsymbol{\mu}_2} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

where

$$\underset{\sim}{\boldsymbol{\mu}}_1 = E(\underset{\sim}{\mathbf{X}}_1), \underset{\sim}{\boldsymbol{\mu}}_2 = E(\underset{\sim}{\mathbf{X}}_2)$$

$$\boldsymbol{\Sigma}_{11} = \mathrm{V}(\underset{\sim}{\mathbf{X}}_1), \boldsymbol{\Sigma}_{22} = \mathrm{V}(\underset{\sim}{\mathbf{X}}_2)$$

$$\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2) = \left[\mathrm{cov}(\underset{\sim}{\mathbf{X}}_2, \underset{\sim}{\mathbf{X}}_1)\right]' = \boldsymbol{\Sigma}'_{21}$$

Now we shall make a non singular linear transformation to sub vectors

$$\begin{aligned} \underset{\sim}{\mathbf{Y}}_1 &= \underset{\sim}{\mathbf{X}}_1 + \mathbf{M}\underset{\sim}{\mathbf{X}}_2 \\ \underset{\sim}{\mathbf{Y}}_2 &= \underset{\sim}{\mathbf{X}}_2 \end{aligned} \qquad \rightarrow (1)$$

choosing $\mathbf{M}$ so that the components of $\underset{\sim}{\mathbf{Y}}_1$ are uncorrelated with the components of $\underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2$.

The matrix $\mathbf{M}$ must satisfy the equation.

$$\mathrm{cov}(\underset{\sim}{\mathbf{Y}}_1, \underset{\sim}{\mathbf{Y}}_2) = \mathbf{0}_{qXp-q} = \mathrm{E}\left[(\underset{\sim}{\mathbf{Y}}_1 - \mathrm{E}(\underset{\sim}{\mathbf{Y}}_1))(\underset{\sim}{\mathbf{Y}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{Y}}_2))'\right]$$

$$= \mathrm{E}\left[\left\{\underset{\sim}{\mathbf{X}}_1 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_1) + \mathbf{M}(\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2))\right\}\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}'\right]$$

$$= \mathrm{E}\left\{\left\{\underset{\sim}{\mathbf{X}}_1 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_1)\right\}\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}'\right\} + \mathbf{M}\mathrm{E}\left\{\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}\left\{\underset{\sim}{\mathbf{X}}_2 - \mathrm{E}(\underset{\sim}{\mathbf{X}}_2)\right\}'\right\}$$

$$= \text{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2) + \text{M V}(\underset{\sim}{\mathbf{X}}_2)$$

$$= \boldsymbol{\Sigma}_{12} + \text{M}\boldsymbol{\Sigma}_{22} \quad \rightarrow (2)$$

Thus , $\text{M} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ $\rightarrow (3)$

and now, the vector $\underset{\sim}{\mathbf{Y}}_1$ becomes

$$\underset{\sim}{\mathbf{Y}}_1 = \underset{\sim}{\mathbf{X}}_1 - \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \quad \rightarrow (4)$$

and the vector

$$\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix} = \mathbf{C}\begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix} = \mathbf{C}\underset{\sim}{\mathbf{X}},$$

$$\text{where } \mathbf{C} = \begin{pmatrix} I_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0}_{(p-q)xq} & I_{p-q} \end{pmatrix} \quad \rightarrow (5)$$

Since $|\mathbf{C}| = 1$, $\mathbf{C}$ is a non singular matrix.

$\therefore \underset{\sim}{\mathbf{Y}}$ is non-singular transformation of $\underset{\sim}{\mathbf{X}}$.

$\therefore \underset{\sim}{\mathbf{Y}}$ has a normal distribution with mean vector

$$E\left(\underset{\sim}{\mathbf{Y}}\right) = \mathbf{C}\underset{\sim}{\boldsymbol{\mu}} = \mathbf{C}\begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\boldsymbol{\mu}}_2 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\nu}}_1 \\ \underset{\sim}{\boldsymbol{\nu}}_2 \end{pmatrix} \quad \text{(say)} \quad \rightarrow (6)$$

and the variance –covariance matrix

$$V(\underset{\sim}{\mathbf{Y}}) = V\begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} V(\underset{\sim}{\mathbf{Y}}_1) & C\text{ov}(\underset{\sim}{\mathbf{Y}}_1, \underset{\sim}{\mathbf{Y}}_2) \\ C\text{ov}(\underset{\sim}{\mathbf{Y}}_2, \underset{\sim}{\mathbf{Y}}_1) & V(\underset{\sim}{\mathbf{Y}}_2) \end{pmatrix}$$

$$= \begin{pmatrix} V(\underset{\sim}{\mathbf{Y}}_1) & \mathbf{0} \\ \mathbf{0} & V(\underset{\sim}{\mathbf{Y}}_2) \end{pmatrix} \quad \text{(from Eq.(2))}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad \rightarrow (7)$$

which implies $\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix}$ has multivariate normal distribution,where $\underset{\sim}{\mathbf{Y}}_1$ **&** $\underset{\sim}{\mathbf{Y}}_2$ are uncorrelated.

Therefore $\underset{\sim}{\mathbf{Y}}_1$ **&** $\underset{\sim}{\mathbf{Y}}_2$ are independent and have multivariate normal distributions.

In particular, $\underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2$ has a MVN distribution

$\therefore \underset{\sim}{\mathbf{X}}_2 \sim N_{p-q}(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_{22})$. That is the marginal distribution of $\underset{\sim}{\mathbf{X}}_2$ is MVN.

2.3 CONDITIONAL DISTRIBUTION OF MVN DISTRIBUTION :

The conditional distribution describes the probability distribution of one set of random variables given that another set takes specific values. It provides insight into how

variables behave in the presence of information about other variables.

In multivariate analysis, conditional distributions are used to:

- Study dependence and association among variables
- Make predictions and perform regression-type analyses
- Understand conditional normality in multivariate normal distributions.

Conditional distributions are central to modeling relationships and are widely used in multivariate inference and prediction.

CONDITIONAL DISTRIBUTION
THEOREM:

If $\underset{\sim}{\mathbf{X}}$ has multivariate normal distribution, then the conditional distribution of any subset of the components of $\underset{\sim}{\mathbf{X}}$ given the subset of the remaining components of $\underset{\sim}{\mathbf{X}}$ is a (multivariate) normal distribution.

(OR)

Prove that the conditional distribution obtained from the multivariate normal distribution is normal.

PROOF:

$$\text{Let } \underset{\sim}{\mathbf{X}} = \begin{pmatrix} \underset{\sim}{\mathbf{X_1}} \\ \underset{\sim}{\mathbf{X_2}} \end{pmatrix}, \underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}_1} \\ \underset{\sim}{\boldsymbol{\mu}_2} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

where

$$\underset{\sim}{\boldsymbol{\mu}}_1 = E(\underset{\sim}{\mathbf{X}}_1), \underset{\sim}{\boldsymbol{\mu}}_2 = E(\underset{\sim}{\mathbf{X}}_2)$$

$$\boldsymbol{\Sigma}_{11} = \mathrm{V}(\underset{\sim}{\mathbf{X}}_1), \boldsymbol{\Sigma}_{22} = \mathrm{V}(\underset{\sim}{\mathbf{X}}_2)$$

$$\boldsymbol{\Sigma}_{12} = \mathrm{cov}(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2) = \left[\mathrm{cov}(\underset{\sim}{\mathbf{X}}_2, \underset{\sim}{\mathbf{X}}_1)\right]' = \boldsymbol{\Sigma}'_{21}$$

Now we shall make a non singular linear transformation to sub vectors

$$\begin{aligned} \underset{\sim}{\mathbf{Y}}_1 &= \underset{\sim}{\mathbf{X}}_1 + \mathbf{M}\underset{\sim}{\mathbf{X}}_2 \\ \underset{\sim}{\mathbf{Y}}_2 &= \underset{\sim}{\mathbf{X}}_2 \end{aligned} \qquad \rightarrow (1)$$

choosing $\mathbf{M}$ so that the components of $\underset{\sim}{\mathbf{Y}}_1$ are uncorrelated with the components of $\underset{\sim}{\mathbf{Y}}_2 = \underset{\sim}{\mathbf{X}}_2$.

The matrix $\mathbf{M}$ must satisfy the equation.

$$\mathrm{cov}(\underset{\sim}{\mathbf{Y}}_1, \underset{\sim}{\mathbf{Y}}_2) = \mathbf{0}_{qXp-q} = Cov\left(\underset{\sim}{\mathbf{X}}_1 + \mathbf{M}\underset{\sim}{\mathbf{X}}_2, \underset{\sim}{\mathbf{X}}_2\right)$$

$$= = Cov\left(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2\right) + \mathbf{M}V\left(\underset{\sim}{\mathbf{X}}_2\right) = \boldsymbol{\Sigma}_{12} + \mathrm{M}\boldsymbol{\Sigma}_{22} \qquad \rightarrow (2)$$

Thus , $\mathrm{M} = -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ $\rightarrow (3)$

and now, the vector $\underset{\sim}{\mathbf{Y}}_1$ becomes

$$\underset{\sim}{\mathbf{Y}}_1 = \underset{\sim}{\mathbf{X}}_1 - \boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \qquad \rightarrow (4)$$

and the vector

$$\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\mathbf{X}}_2 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix} = \mathbf{C}\begin{pmatrix} \underset{\sim}{\mathbf{X}}_1 \\ \underset{\sim}{\mathbf{X}}_2 \end{pmatrix} = \mathbf{C}\underset{\sim}{\mathbf{X}}, \qquad \rightarrow (5)$$

$$\text{where } \mathbf{C} = \begin{pmatrix} I_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0}_{(p-q)xq} & I_{p-q} \end{pmatrix}$$

Since $|\mathbf{C}| = 1$, $\mathbf{C}$ is a non singular matrix.

$\therefore \underset{\sim}{\mathbf{Y}}$ is non-singular transformation of $\underset{\sim}{\mathbf{X}}$.

$\therefore \underset{\sim}{\mathbf{Y}}$ has a normal distribution with mean vector

$$E\left(\underset{\sim}{\mathbf{Y}}\right) = \mathbf{C}\underset{\sim}{\boldsymbol{\mu}} = \mathbf{C}\begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\mu}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\boldsymbol{\mu}}_2 \\ \underset{\sim}{\boldsymbol{\mu}}_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\boldsymbol{\nu}}_1 \\ \underset{\sim}{\boldsymbol{\nu}}_2 \end{pmatrix} \quad \text{(say)} \qquad \rightarrow (6)$$

and the variance –covariance matrix

$$V(\underset{\sim}{\mathbf{Y}}) = V\begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix} = \begin{pmatrix} V(\underset{\sim}{\mathbf{Y}}_1) & C\text{ov}(\underset{\sim}{\mathbf{Y}}_1, \underset{\sim}{\mathbf{Y}}_2) \\ C\text{ov}(\underset{\sim}{\mathbf{Y}}_2, \underset{\sim}{\mathbf{Y}}_1) & V(\underset{\sim}{\mathbf{Y}}_2) \end{pmatrix}$$

$$= \begin{pmatrix} V(\underset{\sim}{\mathbf{Y}}_1) & \mathbf{0} \\ \mathbf{0} & V(\underset{\sim}{\mathbf{Y}}_2) \end{pmatrix} \qquad \text{(from Eq.(2))}$$

$$= \begin{pmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \qquad \rightarrow (7)$$

which implies $\underset{\sim}{\mathbf{Y}} = \begin{pmatrix} \underset{\sim}{\mathbf{Y}}_1 \\ \underset{\sim}{\mathbf{Y}}_2 \end{pmatrix}$ has multivariate normal distribution, where $\underset{\sim}{\mathbf{Y}}_1 \,\&\, \underset{\sim}{\mathbf{Y}}_2$ are uncorrelated.

Therefore $\underset{\sim}{\mathbf{Y}}_1 \,\&\, \underset{\sim}{\mathbf{Y}}_2$ are independent and are MVN variates. More specifically,

$$\underset{\sim}{\mathbf{Y}}_1 \sim N_q(\underset{\sim}{\boldsymbol{\mu}}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_{11.2})$$
$$\underset{\sim}{\mathbf{Y}}_2 \sim N_{p-q}(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_{22}) \qquad \rightarrow (8)$$

where $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Now, the joint p.d.f. of $\underset{\sim}{\mathbf{Y}}$ is given by

$g(\underset{\sim}{\mathbf{Y}}) = g(\underset{\sim}{\mathbf{Y}}_1)g(\underset{\sim}{\mathbf{Y}}_2)$ ($\because \underset{\sim}{\mathbf{Y}}_1$ *and* $\underset{\sim}{\mathbf{Y}}_2$ are independent)

$$= n\left(\underset{\sim}{y}_1 / \underset{\sim}{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2 , \Sigma_{11.2}\right).n\left(\underset{\sim}{y}_2 / \underset{\sim}{\mu}_2 , \Sigma_{22}\right)$$

$$= \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|}\exp\left\{-\frac{1}{2}(\underset{\sim}{Y}_1 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)'\Sigma_{11.2}^{-1}(\underset{\sim}{Y}_1 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)\right\}$$

$$\text{x}\ \frac{1}{(2\pi)^{(p-q)/2}\left|\Sigma_{22}\right|}\exp\left\{-\frac{1}{2}(\underset{\sim}{Y}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{Y}_2 - \underset{\sim}{\mu}_2)\right\} \qquad \rightarrow \ (9)$$

If we make use of the linear transformation (non singular) as given in (5).

The density function of $\underset{\sim}{X}$ is given by

$$f(\underset{\sim}{X}) = g(\underset{\sim}{Y}(\underset{\sim}{X}))\mathbf{.J}(\underset{\sim}{X})$$

Where, $J(\underset{\sim}{X})$ is the Jacobian and is given by

$$J(\underset{\sim}{X}) = \text{mod}\left|\frac{\partial \underset{\sim}{y}}{\partial \underset{\sim}{x}}\right| = \text{mod}|C| = |I_q|.|I_{p-q}| = 1$$

$$\therefore f(\underset{\sim}{X}_1, \underset{\sim}{X}_2) = f(\underset{\sim}{X}) = g(\underset{\sim}{Y}(\underset{\sim}{X}))$$

$$= \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|}e^{-\frac{1}{2}(\underset{\sim}{x}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{x}_2 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)'\Sigma_{11.2}^{-1}(\underset{\sim}{x}_1 - \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{x}_2 - \underset{\sim}{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\underset{\sim}{\mu}_2)} \qquad \rightarrow \ (10)$$

$$\text{x}\ \frac{1}{(2\pi)^{(p-q)/2}\left|\Sigma_{22}\right|}e^{-\frac{1}{2}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)}$$

Now, By the definition conditional density of $\underset{\sim}{X}$, given that $\underset{\sim}{X}_2 = \underset{\sim}{x}_2$ is that

$$f(\underset{\sim}{x}_1 / \underset{\sim}{x}_2) = \frac{f(\underset{\sim}{x}_1, \underset{\sim}{x}_2)}{f(\underset{\sim}{x}_2)} \qquad \rightarrow \ (11)$$

where $f(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$ is given by (2) and $f(\underset{\sim}{x}_2)$ is the marginal density of $\underset{\sim}{X}_2$ at the point $\underset{\sim}{x}_2$ where is nothing but $n(\underset{\sim}{x}_2 / \underset{\sim}{\mu}_2, \Sigma_{22})$.

i.e. $f(\underset{\sim}{x}_2) = n(\underset{\sim}{x}_2 / \underset{\sim}{\mu}_2, \Sigma_{22})$

$$= \frac{1}{(2\pi)^{(p-q)/2}\left|\Sigma_{22}\right|}\exp\left\{-\frac{1}{2}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)'\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)\right\} \qquad \rightarrow \ (12)$$

Using (10) & (12) in (11) we get,

$$f(\underset{\sim}{x}_1 / \underset{\sim}{x}_2) = \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|}e^{-\frac{1}{2}[(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)]'\Sigma_{11.2}^{-1}[(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1) - \Sigma_{12}\Sigma_{22}^{-1}(\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2)]}$$

→(13)

which is the conditional p.d.f. of $\underset{\sim}{\mathbf{X}}_1$ given that $\underset{\sim}{\mathbf{X}}_2 = \underset{\sim}{\mathbf{x}}_2$.

From (13), it is clear that the density $f(\underset{\sim}{\mathbf{x}}_1/\underset{\sim}{\mathbf{x}}_2)$ is clearly a q-variate normal density with mean,

$$E(\underset{\sim}{\mathbf{X}}_1/\underset{\sim}{\mathbf{X}}_2 = \underset{\sim}{\mathbf{x}}_2) = \underset{\sim}{\boldsymbol{\mu}}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2)$$

$$= \nu(\underset{\sim}{\mathbf{x}}_2) \text{ , say} \qquad \rightarrow (14)$$

and the variances matrix,

$$\text{var}(\underset{\sim}{\mathbf{x}}_1/\underset{\sim}{\mathbf{X}}_2 = \underset{\sim}{\mathbf{x}}_2) = \boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \qquad \rightarrow (15)$$

From (14)&(15) we may observe that the conditional mean of $\underset{\sim}{\mathbf{x}}_1$ is simply a linear function of $\underset{\sim}{\mathbf{x}}_2$ and the conditional co-variance of $\underset{\sim}{\mathbf{x}}_1$ does not depend on $\underset{\sim}{\mathbf{x}}_2$ at all.

Problem:

Let $\underset{\sim}{\mathbf{X}}$ has a trivariate normal distribution with $E(\underset{\sim}{\mathbf{X}}) = \underset{\sim}{\mathbf{0}}$ and variance-covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1/2 & -1/2 & 1/2 \\ -1/2 & 1 & -1/2 \\ 1/2 & -1/2 & 1 \end{bmatrix}.$$

Find the conditional distribution of X_1 given $X_2 = x_2$ and $X_3 = x_3$.

<u>The above result may be put in the following theorem:-</u>

Let the components of $\underset{\sim}{\mathbf{X}}$ be divided in to two groups composing the sub vectors $\underset{\sim}{\mathbf{X}}_1$ & $\underset{\sim}{\mathbf{X}}_2$. Suppose the mean $\underset{\sim}{\boldsymbol{\mu}}$ is similarly divided into $\underset{\sim}{\boldsymbol{\mu}}_1$ & $\underset{\sim}{\boldsymbol{\mu}}_2$ and suppose the co-variance matrix $\boldsymbol{\Sigma}$ of $\underset{\sim}{\mathbf{X}}$ is divided into $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}, \boldsymbol{\Sigma}_{22}$ the co-variance matrices of $\underset{\sim}{\mathbf{X}}_1$ of $\underset{\sim}{\mathbf{X}}_1$ & $\underset{\sim}{\mathbf{X}}_2$, and of $\underset{\sim}{\mathbf{X}}_2$ respectively. Then if the distribution of $\underset{\sim}{\mathbf{X}}$ is normal, the conditional distribution of $\underset{\sim}{\mathbf{X}}_1$ is given $\underset{\sim}{\mathbf{X}}_2 = \underset{\sim}{\mathbf{x}}_2$ is normal with mean $\underset{\sim}{\boldsymbol{\mu}}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\underset{\sim}{\mathbf{x}}_2 - \underset{\sim}{\boldsymbol{\mu}}_2)$ and co-variance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

NOTE:-

The above theorem may simply be asked as follows.If $X_1, X_2, ..., X_P$ have a joint normal distribution, then the conditional distribution of a subset of r.v's given that the remaining r.v's is also having normal distribution.

2.4 SUMMARY:

The Multivariate Normal Distribution possesses elegant and powerful properties regarding marginal and conditional distributions. Marginal distributions of an MVN random vector are themselves multivariate normal, and conditional distributions retain normality with easily interpretable mean and covariance structures. These results greatly simplify multivariate modeling and inference and are central to many multivariate statistical methods.

Marginal and conditional distributions are key concepts in multivariate analysis that allow the study of complex joint distributions in a simplified and meaningful way. Marginal distributions focus on subsets of variables independently of others, while conditional distributions examine variable behavior under given conditions. Together, they provide a comprehensive understanding of dependence, prediction, and inference in multivariate statistical models.

2.5 SELF-ASSESSMENT QUESTIONS:

1. What is meant by the marginal distribution of an MVN vector?
2. Prove that the conditional distribution of a partitioned MVN vector is also normal and derive its mean vector and covariance matrix.
3. Let x be a p - variate normal. Obtain the marginal and conditional distributions of x.
4. Let $X \sim N_1(\mu, \Sigma)$ with $\mu = [2 \ -3 \ 1]$ and $\Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$. Obtain the conditional distribution of X_3 given that $X_1 = x_1$ and $X_2 = x_2$
5. Show that the marginal distribution of any subset of variables from an MVN vector is also normal.
6. What is a conditional distribution? How does it differ from a marginal distribution?

2.6 SUGGESTED READINGS:

1. Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley. (A classic and comprehensive reference on MVN distribution and multivariate inference.)
2. Johnson, R. A. & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Pearson. (Excellent for applied understanding and properties of MVN.)
3. Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). Multivariate Analysis. Academic Press. (Foundational theory, properties, and proofs of MVN results)
4. Seber, G. A. F. (1984). Multivariate Observations. Wiley. (Strong theoretical treatment of multivariate distributions.)
5. Rencher, A. C., Methods of Multivariate Analysis
6. Bilodeau, M. & Brenner, D. (1999). Theory of Multivariate Statistics. Springer. (Accessible theoretical treatment.)

Prof. A. Vasudeva Rao

LESSON -3

CHARACTERISTIC FUNCTION OF MVN DISTRIBUTION

Objectives:

- Define the characteristic function of a random variable and a random vector**.**
- Derive the characteristic function of the MVN distribution**.**
- Use the characteristic function to derive key properties of the MVN distribution.

STRUCTURE:

3.1 INTRODUCTION:

The characteristic function is an important tool in probability theory and statistical inference. It uniquely determines the distribution of a random variable or random vector and is especially useful in multivariate analysis. In this unit, the concept of the characteristic function is introduced and applied to the Multivariate Normal (MVN) distribution**.**

3.2 DEFINITION OF CHARACTERISTIC FUNCTION:

The characteristic function is a mathematical function that completely defines the probability distribution of a random variable (or random vector). It is a unique transformation, similar to a Fourier transform, that always exists for any real-valued random variable, which is a key advantage over the moment-generating function which may not exist for some distributions (e.g., the Cauchy distribution).

For a random vector **X**, its characteristic function, denoted by $\phi_{\mathbf{X}}(\mathbf{t})$, is defined as the expected value of a complex exponential function:

where $(i^2 = -1)$, **t** is a vector of real numbers, and the prime notation denotes the transpose.

The characteristic function of a random vector $\underset{\sim}{\mathbf{X}}$ is

$$\phi_{\mathbf{X}}(\underset{\sim}{\mathbf{t}}) = E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}})$$

defined for every real vector $\underset{\sim}{\mathbf{t}}$.

RESULT:-

If the components of a random vector $\underset{\sim}{\mathbf{X}}$ are independently distributed,

Then,

$$E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}}) = E\left(e^{i\sum_{j=1}^{p} t_j X_j} \right)$$

$$E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}}) = \prod_{j=1}^{p} E(e^{it_j X_j})$$

3.3 CHARACTERISTIC FUNCTION OF THE MVN DISTRIBUTION:

The multivariate normal (MVN) distribution is a generalization of the one-dimensional normal distribution to multiple dimensions. A random vector **X** is said to follow an MVN distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix (denoted as $\mathbf{X}\sim(\boldsymbol{\mu},\boldsymbol{\Sigma})$) if its characteristic function is given by:

$$\phi_{\mathbf{X}}(\underset{\sim}{\mathbf{t}}) = \exp\left[it'\mu - \frac{1}{2} t'\Sigma t \right]$$

This specific functional form is a defining property of the multivariate normal distribution and is frequently used to derive other properties, such as the distributions of linear combinations of MVN variables.

THEOREM 1:-

The characteristic function of $\underset{\sim}{\mathbf{X}}$ which is distributed according to

$N(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ **is** $\phi(\underset{\sim}{\mathbf{t}}) = E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}}) = e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}\underset{\sim}{\mathbf{t}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{t}}}$

for every real vector $\underset{\sim}{\mathbf{t}}$.

PROOF:-

We have given $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

Since, $\boldsymbol{\Sigma}$ and hence $\boldsymbol{\Sigma}^{-1}$ is symmetric and positive definite matrix there exists a non-singular matrix $\mathbf{C}$ such that

$$\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C} = \mathbf{I} \quad \rightarrow (1)$$

$$\Rightarrow \boldsymbol{\Sigma}^{-1} = (\mathbf{C}\mathbf{C}')^{-1} \text{ or } \boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}' \quad \rightarrow (1.a)$$

we have the p.d.f. of $\underset{\sim}{\mathbf{X}}$ is

$$f(\underset{\sim}{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}})} \quad \rightarrow (2)$$

Let us make use of the linear transformation,

$$\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}} = \mathbf{C}\underset{\sim}{\mathbf{Y}} \qquad (\mathbf{C} \text{ is defined as in (1)}) \quad \rightarrow (3)$$

Then the p.d.f. of the new random vector $\underset{\sim}{\mathbf{Y}}$ is

$$g(\underset{\sim}{\mathbf{Y}}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}\underset{\sim}{\mathbf{y}}} \mathrm{J}(\underset{\sim}{\mathbf{y}}) \quad \rightarrow (4)$$

where $\mathrm{J}(\underset{\sim}{\mathbf{y}})$ is the jacobian transformation and is given by

$$\mathrm{J}(\underset{\sim}{\mathbf{y}}) = \mathrm{mod}\left|\frac{\partial \underset{\sim}{\mathbf{x}}}{\partial \underset{\sim}{\mathbf{y}}}\right| = \mathrm{mod}|\mathbf{C}| \quad \left(\because \frac{\partial \underset{\sim}{\mathbf{x}}}{\partial \underset{\sim}{\mathbf{y}}} = \frac{\partial \mathbf{C}\underset{\sim}{\mathbf{y}} + \underset{\sim}{\boldsymbol{\mu}}}{\partial \underset{\sim}{\mathbf{y}}} = \mathbf{C}\right)$$

$$= \mathrm{mod}|\boldsymbol{\Sigma}|^{1/2} \qquad \left(\text{from 1(a)} |\boldsymbol{\Sigma}| = |\mathbf{C}\mathbf{C}'| = |\mathbf{C}|^2\right)$$

$$= |\boldsymbol{\Sigma}|^{1/2} \qquad \left(\because |\boldsymbol{\Sigma}| > 0\right)$$

Therefore (4) becomes from (1),

$$g(\underset{\sim}{\mathbf{Y}}) = \frac{1}{(2\pi)^{p/2}} e^{-\frac{1}{2}\underset{\sim}{\mathbf{y}}'\underset{\sim}{\mathbf{y}}} = n(\underset{\sim}{\mathbf{y}}/\underset{\sim}{\mathbf{0}}, I_p)$$

i.e. $\underset{\sim}{\mathbf{Y}} \sim N_p(\underset{\sim}{\mathbf{0}}, I_p)$.

The characteristic function of $\underset{\sim}{\mathbf{Y}}$ is

$$\phi(\underset{\sim}{\mathbf{u}}) = E(e^{i\underset{\sim}{\mathbf{u}}'\underset{\sim}{\mathbf{Y}}}) = E\left(e^{i\sum_{j=1}^{p} u_j Y_j}\right)$$

$$= \prod_{j=1}^{p} E\left(e^{iu_j Y_j}\right) \qquad (\because Y_j's \text{ are independent})$$

$$= \prod_{j=1}^{p} e^{-\frac{1}{2}u_j^2} \qquad (\because \text{the characteristic function of the standard}$$

$$= e^{-\frac{1}{2}\sum_{j=1}^{p} u_j^2} \qquad \text{normal variate } Y_j \text{ is } e^{-\frac{1}{2}u_j^2})$$

$$= e^{-\frac{1}{2}\underset{\sim}{u}'\underset{\sim}{u}} \qquad \rightarrow (5)$$

Now,

$$\phi(\underset{\sim}{t}) = E(e^{i\underset{\sim}{t}'\underset{\sim}{X}}) = E(e^{i\underset{\sim}{t}'(\mathbf{C}\underset{\sim}{Y}+\underset{\sim}{\mu})}) \qquad \text{(from (3))}$$

i.e.

$$\phi(\underset{\sim}{t}) = E(e^{i\underset{\sim}{t}'\underset{\sim}{\mu}}) = e^{i\underset{\sim}{t}'\underset{\sim}{\mu}} E\left(e^{i\underset{\sim}{t}'\mathbf{C}\underset{\sim}{Y}}\right)$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu}} E\left(e^{i\underset{\sim}{u}'\underset{\sim}{Y}}\right) \qquad (\text{where } \underset{\sim}{u} = \mathbf{C}'\underset{\sim}{t})$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu}} e^{-\frac{1}{2}\underset{\sim}{u}'\underset{\sim}{u}} \qquad \text{(from (5))}$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}'\mathbf{C}\mathbf{C}'\underset{\sim}{t}} \qquad (\because \underset{\sim}{u}' = \underset{\sim}{t}'\mathbf{C})$$

$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}'\Sigma\underset{\sim}{t}} \qquad \text{(from 1.a)}$$

Hence the proof.

3.4 SOME MORE PROPERTIES OF THE MVN DISTRIBUTION BASED ON THE CHARACTERISTIC FUNCTION:

THEOREM 2:-

If every linear combination of the components of a vector $\underset{\sim}{X}$ is normally distributed ,then $\underset{\sim}{X}$ is normally distributed **.**

PROOF:-

Suppose $\underset{\sim}{X}$ is a random vector of p random variables with mean vector $\underset{\sim}{\mu}$ and co-variance matrix Σ .

Let us consider an arbitrary linear combination of $\underset{\sim}{X}$ viz… $\underset{\sim}{c}'\underset{\sim}{X}$, where $\underset{\sim}{c}' = (c_1, c_2, ..., c_p)$.

We have given , $\underset{\sim}{c}'\underset{\sim}{X}$ is normal variate.

We have,

$$E(\underset{\sim}{c}'\underset{\sim}{X}) = \underset{\sim}{c}'E(\underset{\sim}{X})$$

$$= \underset{\sim}{c}'\underset{\sim}{\mu}$$

$$V(\underset{\sim}{c}'\underset{\sim}{X}) = V(\underset{\sim}{c}'\underset{\sim}{X})$$

$$= \underset{\sim}{c}'V(\underset{\sim}{X})\underset{\sim}{c}$$
$$= \underset{\sim}{c}'\Sigma\underset{\sim}{c} \qquad \text{(variance)}$$

If may be noted $\underset{\sim}{c}'\underset{\sim}{\mu}$ & $\underset{\sim}{c}'\Sigma\underset{\sim}{c}$ are scalars and they are respectively the mean & variances of the univariate random variables $\underset{\sim}{c}'\underset{\sim}{X}$. We have given that $\underset{\sim}{c}'\underset{\sim}{X} \sim N(\underset{\sim}{c}'\underset{\sim}{\mu}, \underset{\sim}{c}'\Sigma\underset{\sim}{c})$.

Let $Y = \underset{\sim}{c}'\underset{\sim}{X}$ and

from the univariate normal distribution theory, the characteristic function of Y is given by

$$\psi(t) = E(e^{itY})$$
$$= e^{itE(Y) - \frac{1}{2}t^2V(Y)}$$
$$= e^{it\underset{\sim}{c}'\underset{\sim}{\mu} - \frac{1}{2}t^2\underset{\sim}{c}'\Sigma\underset{\sim}{c}}$$

If we write t=1 then, $\psi(t)$ becomes

$$\phi(\underset{\sim}{c}) = e^{i\underset{\sim}{c}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{c}'\Sigma\underset{\sim}{c}} \qquad \text{where, } \underset{\sim}{X} \sim N(\underset{\sim}{\mu}, \Sigma)$$

which is the characteristic function of a multivariate random vector $\underset{\sim}{X}$ whose mean vector is $\underset{\sim}{\mu}$ & variance-covariance matrix is of Σ.

But the mean & variance-covarince matrix of $\underset{\sim}{X}$ respectively same as $\underset{\sim}{\mu}$ & Σ and therefore, $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$. Hence the proof.

THEOREM 3:-

If $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$ then, $\underset{\sim}{c}'\underset{\sim}{X}$ is uni-normal variate with mean $\underset{\sim}{c}'\underset{\sim}{X}$ and variance $\underset{\sim}{c}'\Sigma\underset{\sim}{c}$.

(OR)

If $X_1, X_2, ..., X_p$ are jointly distributed as p-variate normal then its linear combination follows univariate normal distribution.

PROOF:-

$$\text{Let } \underset{\sim}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \sim N_p(\underset{\sim}{\mu}, \Sigma)$$

Then its characteristic function is given by

$$\phi(\underset{\sim}{t}) = E(e^{i\underset{\sim}{t}'\underset{\sim}{X}}))$$
$$= e^{i\underset{\sim}{t}'\underset{\sim}{\mu} - \frac{1}{2}\underset{\sim}{t}'\Sigma\underset{\sim}{t}} \qquad \rightarrow (1)$$

Let us write $\underset{\sim}{\mathbf{t}} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_p \end{pmatrix} = \begin{pmatrix} t.c_1 \\ t.c_2 \\ \vdots \\ t.c_p \end{pmatrix} = t\underset{\sim}{\mathbf{c}}$

Then (1) becomes ,

$$\phi(\underset{\sim}{\mathbf{t}}) = e^{it\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}t\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}t}$$

$$= e^{\mathrm{it}\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}\mathrm{t}^2\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}}$$

$$= \mathrm{E}\left(e^{\mathrm{itY}}\right)$$

$$= \psi(t) \quad \text{,say} \qquad \rightarrow (2)$$

where Y is normal variate with mean $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}}$ and variance $\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}$.

In other words (2) is the characteristic function $\psi(t)$ of a uni-normal variate whose mean is $\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}}$ and variance is $\underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}$.

If we consider the linear combination of the components of the normal random vector $\underset{\sim}{\mathbf{X}}$ viz.,

$$\mathrm{Y} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}$$

$$= c_1\mathrm{X}_1 + c_2\mathrm{X}_2 + \ldots + c_p\mathrm{X}_p$$

its mean and variance are given by

$$E(\mathrm{Y}) = E(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}) = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}} \quad \& \quad \mathrm{V}(\mathrm{Y}) = \mathrm{V}(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}) = \underset{\sim}{\mathbf{c}}'\mathrm{V}(\underset{\sim}{\mathbf{X}})\underset{\sim}{\mathbf{c}} = \underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}}$$

Thus, from the above explanation it follows that $\mathrm{Y} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}}$ follows uvi-variate normal distribution,

i.e. , . $\mathrm{Y} = \underset{\sim}{\mathbf{c}}'\underset{\sim}{\mathbf{X}} \sim \mathrm{N}(\underset{\sim}{\mathbf{c}}'\underset{\sim}{\boldsymbol{\mu}}, \underset{\sim}{\mathbf{c}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{c}})$

Hence the proof.

3.5 SUMMARY:

- The multivariate normal distribution is one of the most important and widely used distributions in statistics and data science due to its analytical tractability and rich geometric and probabilistic properties.
- Its behavior is completely determined by its mean vector and covariance matrix**,** making it mathematically elegant and practically useful.
- The characteristic function offers a powerful tool for deriving and proving many results about MVN distributions, including linear transformations and independence properties.
- The structural properties such as marginals**,** conditionals**,** and affine transformations make the MVN family closed under many operations commonly used in statistical inference.

- Because of these features, the MVN distribution forms the foundation of many classical multivariate methods and modern machine-learning techniques.

3.6 SELF-ASSESSMENT QUESTIONS:

1. Derive the characteristic function of an MVN vector $X \sim N_p(\mu, \Sigma)$.
2. Define the characteristic function of p-variate normal distribution.Hence find the $Cov(X_i, X_j)$
3. State and prove characteristic function of $\underset{\sim}{\mathbf{X}}$ which is distributed according to $N(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ is $\phi(\underset{\sim}{\mathbf{t}}) = E(e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\mathbf{X}}}) = e^{i\underset{\sim}{\mathbf{t}}'\underset{\sim}{\boldsymbol{\mu}} - \frac{1}{2}\underset{\sim}{\mathbf{t}}'\boldsymbol{\Sigma}\underset{\sim}{\mathbf{t}}}$ for every real vector $\underset{\sim}{\mathbf{t}}$.
4. State two important properties of characteristic functions.
5. Explain the significance of the characteristic function in multivariate analysis.

3.7 SUGGESTED READINGS:

1. Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis**.** Wiley. (A classic and comprehensive reference on MVN distribution and multivariate inference.)
2. Johnson, R. A. & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis**.** Pearson. (Excellent for applied understanding and properties of MVN.)
3. Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis***.** Academic Press. (Foundational theory, properties, and proofs of MVN results.)
4. Bilodeau, M. & Brenner, D. (1999). Theory of Multivariate Statistics**.**Springer.(Accessible theoretical treatment.).

Prof. A. Vasudeva Rao

LESSON -4

ML ESTIMATION AND SAMPLING DISTRIBUTIONS

OBJECTIVES:

- Understand the concept of random sampling from a Multivariate Normal (MVN) distribution.
- Derive and explain the sampling distribution of the sample mean vector.
- Derive and interpret the sampling distribution of the sample covariance matrix.
- Apply Maximum Likelihood Estimation (MLE) to estimate the mean vector and covariance matrix of an MVN distribution.
- Analyze the independence properties of the sample mean and sample covariance matrix.

STRUCTURE:

4.1 INTRODUCTION:

Maximum likelihood estimation is a method for estimating the parameters of a probability distribution by finding the values that make the observed data most likely, given the model. In multivariate analysis, it is used to estimate the parameters of a model, such as covariance matrix and mean vector.

Here is a step-by-step explanation:

1. Specify the model: Define the multivariate model, such as a multivariate normal distribution.
2. Define the likelihood function: The likelihood function is the probability of observing the data given the model parameters.
3. Define the log-likelihood function: The log-likelihood function is the logarithm of the likelihood function, which is used for computational convenience.
4. Find the maximum likelihood estimates: Find the values of the model parameters that

maximize the log-likelihood function. This is typically done using numerical optimization methods, such as the Newton-Raphson method or gradient-based method.

5. Estimate the model parameters: The maximum likelihood estimates are the values of the model parameters that maximize the log-likelihood function. These estimates are used to summarize the data and make inferences about the population.

In multivariate statistical analysis, sample data are used to make inferences about unknown population parameters. Two of the most important sample statistics are the sample mean vector and the sample covariance matrix**,** as they provide information about the central tendency and variability of multivariate data. To use these statistics effectively in estimation and hypothesis testing, it is necessary to understand their sampling distributions**.**

The sampling distribution describes the probability distribution of a statistic obtained from repeated random samples of the same size drawn from a population. When the underlying population follows a multivariate normal distribution**,** the sampling distributions of the sample mean vector and the sample covariance matrix have well-defined and tractable forms. The sample mean vector follows a multivariate normal distribution, while the sample covariance matrix follows a Wishart distribution**.**

Knowledge of these sampling distributions forms the theoretical basis for many multivariate inference techniques such as confidence regions for the mean vector, hypothesis testing using Hotelling's T^2 statistic, multivariate analysis of variance (MANOVA), and likelihood-based estimation methods. Hence, the study of sampling distributions of the sample mean vector and covariance matrix is fundamental to multivariate statistical theory and applications.

4.1.1 OVERVIEW OF MULTIVARIATE NORMAL DISTRIBUTION:

The multivariate normal distribution is a generalization of the one-dimensional (univariate) normal distribution to a higher number of dimensions. A random vector $\mathbf{X}$ of p dimensions is considered to be multivariate normal, denoted as $\underset{\sim}{\mathbf{X}} \sim N_p(\underset{\sim}{\boldsymbol{\mu}}, \Sigma)$, if every linear combination of its components is normally distributed.

It is completely characterized by two parameters: the **mean vector** ($\boldsymbol{\mu}$), which is a p-dimensional vector of the expected values for each variable, and the **covariance matrix**($\boldsymbol{\Sigma}$), a $p \times p$ symmetric, positive semi-definite matrix that contains the variances of each variable on the diagonal and the co-variances between variable pairs in the off-diagonal elements. The contours of constant density for the MVN distribution are ellipsoids centered at $\boldsymbol{\mu}$.
Key properties include:

- Any subset of variables from a MVN vector also has a MVN distribution (marginal distributions are normal).
- Any linear combination of the components is also normally distributed.
- Zero covariance between components implies statistical independence, a property unique to the normal distribution family.
- Conditional distributions of any subset of variables, given values of other variables, are also multivariate normal.

4.1.2 IMPORTANCE IN STATISTICAL MODELLING AND DATA ANALYSIS:

The MVN distribution plays a central and fundamental role in multivariate statistical analysis, similar to the univariate normal in standard statistics. Its theoretical tractability and desirable mathematical properties, such as being closed under affine transformations, make it the default assumption for many classical multivariate techniques like Principal Component Analysis, Factor Analysis, and Discriminant Analysis. It is widely used in:

- Regression modeling, especially in econometrics and psychometrics.
- Machine learning, where it is used to approximate feature distributions in classification and Bayesian inference.
- Finance, particularly in portfolio modeling and risk assessment (though its lack of tail dependence can be a limitation).
- Biological and social sciences, for analyzing relationships between multiple correlated variables, such as the classic example of father's and son's heights.

Its importance also stems from the multivariate central limit theorem, which states that the distribution of sample means from a large variety of underlying distributions approaches a multivariate normal distribution.

4.2 SAMPLING FROM MVN DISTRIBUTION:

Let us assume that the p×*1* **vectors** $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n$ represent a random sample from a multivariate normal population with mean vector $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix Σ. Since $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n$ are mutually independent (by virtue of randomization) and each has distributed as the joint p.d.f. of all the observations is the product of the marginal normal densities.

i.e. $f(\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n) = f(\underset{\sim}{\mathbf{X}}_1) f(\underset{\sim}{\mathbf{X}}_2) \cdots f(\underset{\sim}{\mathbf{X}}_n)$

$$= \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})} \right\}$$

$$\left[\because f(\underset{\sim}{\mathbf{x}}_j) = n(\underset{\sim}{\mathbf{x}}_j / \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right]$$

$$= \frac{1}{(2\pi)^{np/2} |\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n} (\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}})} \quad \text{--- (1)}$$

When the numerical values of the observations become available, there may be substituted for $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n$ in equation (1). The resulting expression, now considered as a function of $\underset{\sim}{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ and for a fixed set of observations $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, \ldots, \underset{\sim}{\mathbf{X}}_n$, it is called as "**the likelihood function**" and is denoted as $L(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$.

4.3 ML ESTIMATION OF MEAN VECTOR ($\underset{\sim}{\mu}$) AND DISPERSION MATRIX (Σ):

Consider the likelihood function $L(\underset{\sim}{\mu}, \Sigma)$ given in Eq. (1). i.e.

$$L(\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{j=1}^{n}(\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu})} \quad \text{---- (2)}$$

Now the maximum likelihood estimates of $\underset{\sim}{\mu}$ and Σ can be obtained by maximizing $L(\underset{\sim}{\mu}, \Sigma)$.

In order to obtain the MLE's of $\underset{\sim}{\mu}$ and Σ ,let us consider logarithms of (2) and is given by

$$\log L(\underset{\sim}{\mu}, \Sigma) = \frac{-np}{2} \log(2\pi) - \frac{n}{2} \log|\Sigma| - \frac{1}{2} \sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}) \quad \text{---- (3)}$$

Consider the last term of (3) and as if is a scalar we may write if as

$$\sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu})$$

$$= tr\left[\sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}) \right]$$

$$= \sum_{j=1}^{n} tr\left[(\underset{\sim}{x}_j - \underset{\sim}{\mu})' \Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu}) \right]$$

$$= \sum_{j=1}^{n} tr\left[\Sigma^{-1} (\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})' \right]$$

$$(\because \ tr(AB) = tr(BA))$$

$$= tr\left[\Sigma^{-1} \left\{ \sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})' \right\} \right] \quad \text{---- (4)}$$

Now consider

$$\sum_{j=1}^{n} (\underset{\sim}{x}_j - \underset{\sim}{\mu})(\underset{\sim}{x}_j - \underset{\sim}{\mu})'$$

$$= \sum_{j=1}^{n} [\underset{\sim}{x}_j - \underset{\sim}{\bar{x}} + \underset{\sim}{\bar{x}} - \underset{\sim}{\mu}][\underset{\sim}{x}_j - \underset{\sim}{\bar{x}} + \underset{\sim}{\bar{x}} - \underset{\sim}{\mu}]'$$

Where $\underset{\sim}{\bar{x}} = \frac{1}{n}(\underset{\sim}{x}_1 + \underset{\sim}{x}_2 + \cdots + \underset{\sim}{x}_n)$

$$= \sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' + \sum_{j=1}^{n} \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)'$$

(Since the cross product terms)

$$\sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)' = \left(n\overline{\underset{\sim}{\mathbf{x}}} - n\overline{\underset{\sim}{\mathbf{x}}} \right) \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)' = 0$$

$$\left[\because \sum_{j=1}^{n} \underset{\sim}{\mathbf{x}}_j = n\overline{\underset{\sim}{\mathbf{x}}} \right]$$

and similarly

$$\sum_{j=1}^{n} \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' = 0$$

Thus

$$\sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}} \right)'$$

$$= \sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' + n \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)'$$

Substituting this in (4) we get ,

$$\sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}} \right)' \boldsymbol{\Sigma}^{-1} \left(\underset{\sim}{\mathbf{x}}_j - \underset{\sim}{\boldsymbol{\mu}} \right)$$

$$= tr \left[\boldsymbol{\Sigma}^{-1} \left\{ \sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' + n \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)' \right\} \right]$$

$$= tr \left[\boldsymbol{\Sigma}^{-1} \left\{ \sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' \right\} \right] + n \left[tr \left\{ \boldsymbol{\Sigma}^{-1} \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)' \right\} \right]$$

$$= tr \left[\boldsymbol{\Sigma}^{-1} \left\{ \sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' \right\} \right] + n \left[tr \left\{ \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)' \boldsymbol{\Sigma}^{-1} \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \right\} \right]$$

[Since tr (AB) =tr (BA)]

$$= tr \left[\boldsymbol{\Sigma}^{-1} \left\{ \sum_{j=1}^{n} \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right) \left(\underset{\sim}{\mathbf{x}}_j - \overline{\underset{\sim}{\mathbf{x}}} \right)' \right\} \right] + n \left\{ \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right)' \boldsymbol{\Sigma}^{-1} \left(\overline{\underset{\sim}{\mathbf{x}}} - \underset{\sim}{\boldsymbol{\mu}} \right) \right\} \quad \text{--- (5)}$$

(Since trace of scalar is scalar)

Substituting (5) in (3),we get

$$\log L\left(\underset{\sim}{\mu},\Sigma\right)=\frac{-np}{2}\log\left(2\pi\right)+\frac{n}{2}\log\left|\boldsymbol{\Sigma}^{-1}\right|$$
$$-\frac{1}{2}\left[tr\left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)'\right)\right\}\right]$$
$$-\frac{n}{2}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right) \quad \text{---- (6)}$$

Since $\boldsymbol{\Sigma}^{-1}$ is positive definite

$$\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\overline{\underset{\sim}{\mathbf{x}}}-\underset{\sim}{\boldsymbol{\mu}}\right)>0 \quad \forall\underset{\sim}{\boldsymbol{\mu}}\neq\overline{\underset{\sim}{\mathbf{x}}}$$
$$=0 \quad \text{if } \underset{\sim}{\boldsymbol{\mu}}=\overline{\underset{\sim}{\mathbf{x}}}$$

From (6), we can observe that if the last term is zero then (6) becomes maximum that is $\log L\left(\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)$ can be maximized with respect to $\underset{\sim}{\boldsymbol{\mu}}$ at $\underset{\sim}{\boldsymbol{\mu}}=\overline{\underset{\sim}{\mathbf{x}}}$

$\therefore$ The MLE of $\underset{\sim}{\boldsymbol{\mu}}$ is $\overline{\underset{\sim}{\mathbf{x}}}$ and substituting the MLE of $\underset{\sim}{\boldsymbol{\mu}}$ $\left(\overline{\underset{\sim}{\mathbf{x}}}\right)$ in (6)

We get

$$\log L\left(\underset{\sim}{\boldsymbol{\mu}},\boldsymbol{\Sigma}\right)=\frac{-np}{2}\log\left(2\pi\right)+\frac{n}{2}\log\left|\boldsymbol{\Sigma}^{-1}\right|$$
$$-\frac{1}{2}\left[tr\left\{\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)'\right)\right\}\right] \quad \text{--- (7)}$$

Now we have to maximize (7) w.r.t. $\boldsymbol{\Sigma}$ as the equation is free of $\underset{\sim}{\boldsymbol{\mu}}$

We can prove that (7) attains its maximum value at $\boldsymbol{\Sigma}=\hat{\boldsymbol{\Sigma}}$,

Where, $$\hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)' \quad \text{--- (8)}$$

Thus $\hat{\boldsymbol{\Sigma}}$ (given by (8)) is the MLE of $\boldsymbol{\Sigma}$.

The maximum value of the likelihood can be obtained by substituting the MLEs of $\underset{\sim}{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}$ respectively given by

$$\hat{\underset{\sim}{\boldsymbol{\mu}}}=\overline{\underset{\sim}{\mathbf{x}}} \quad \text{and } \hat{\boldsymbol{\Sigma}}=\frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)\left(\underset{\sim}{\mathbf{x}}_{\mathbf{j}}-\overline{\underset{\sim}{\mathbf{x}}}\right)'$$

in (2) and it is given by

$$= (2\pi)^{-np/2} \left|\hat{\boldsymbol{\Sigma}}\right|^{-n/2} e^{\frac{-n}{2} tr(\mathrm{I_p})}$$

$$= (2\pi)^{-np/2} \left|\hat{\boldsymbol{\Sigma}}\right|^{-n/2} e^{-np/2}$$

$$= \text{const. x } \left|\hat{\boldsymbol{\Sigma}}\right|^{-\frac{n}{2}}$$

$$= \text{const. x } (\text{generilised variance})^{\frac{-n}{2}}$$

since generalized variance is defined as $\left|\hat{\boldsymbol{\Sigma}}\right|$.The generalized variance determines the *peakedness* of the likelihood function and consequently is a natural measure of variability when the parent population is multivariate normal.

NOTE:-

1. MLEs posses an invariance property which means if $\hat{\theta}$ is the MLE of θ then $h(\hat{\theta})$ is the MLE of $h(\theta)$,where $h(\theta)$ is a function of θ .

For Example :-

- If $\hat{\underset{\sim}{\mu}}$ MLE of $\underset{\sim}{\mu}$ and $\hat{\boldsymbol{\Sigma}}$ is the MLE of $\boldsymbol{\Sigma}$, then $\hat{\underset{\sim}{\mu}}' \hat{\boldsymbol{\Sigma}}^{-1} \hat{\underset{\sim}{\mu}}$ is the MLE of $\underset{\sim}{\mu}' \boldsymbol{\Sigma}^{-1} \underset{\sim}{\mu}$.
- If σ_{ij} is the ij^{th} element of $\boldsymbol{\Sigma}$ and $\hat{\sigma}_{ij}$ is the ij^{th} element of $\hat{\boldsymbol{\Sigma}}$ where $\hat{\boldsymbol{\Sigma}}$ is the MLE of σ_{ij} .

Where $\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left(X_{ik} - \bar{X}_i\right)\left(X_{jk} - \bar{X}_j\right)$

$$= COV\left(X_i, X_j\right).$$

2. From equation (6) the log-likelihood and hence the joint p.d.f depends on the whole set of observations $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$ only through the sample mean $\bar{\underset{\sim}{X}}$ and the sum of squares and cross product matrix,

$$\sum_{j=1}^{n} \left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)' = n\hat{\boldsymbol{\Sigma}}$$

We may express this fact by saying that $\hat{\underset{\sim}{\mu}}\left(\text{or } \bar{\underset{\sim}{x}}\right)$ and $\hat{\boldsymbol{\Sigma}}$ are sufficient statistics. Thus the MLEs $\hat{\underset{\sim}{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are sufficient statistics of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

3. The MLE of $\boldsymbol{\Sigma}$ is

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)'$$

Thus formula is not convenient to compute $\hat{\Sigma}$ and the following is the convenient formula for computation

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \bar{\underset{\sim}{x}}\ \bar{\underset{\sim}{x}}'$$

Explanation:-

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}}\right)'$$

$$= \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\bar{\underset{\sim}{x}}' - \frac{1}{n}\sum_{j=1}^{n}\bar{\underset{\sim}{x}}\underset{\sim}{x}_j' + \frac{1}{n}\sum_{j=1}^{n}\bar{\underset{\sim}{x}}\ \bar{\underset{\sim}{x}}'$$

$$= \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \bar{\underset{\sim}{x}}\ \bar{\underset{\sim}{x}}' - \bar{\underset{\sim}{x}}\bar{\underset{\sim}{x}}' + \bar{\underset{\sim}{x}}\bar{\underset{\sim}{x}}'$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{j=1}^{n}\underset{\sim}{x}_j\underset{\sim}{x}_j' - \bar{\underset{\sim}{x}}\ \bar{\underset{\sim}{x}}'$$

4.4 SAMPLING DISTRIBUTIONS OF THE MLEs $\hat{\underset{\sim}{\mu}}$ & $\hat{\Sigma}$ AND THEIR INDEPENDENCE:

Before going to obtain the sampling distribution of $\hat{\underset{\sim}{\mu}}$ *and* $\hat{\Sigma}$, let us prove the following result which is useful in obtaining the sampling distributions of $\hat{\underset{\sim}{\mu}}$ *and* $\hat{\Sigma}$.

Result: Suppose $\underset{\sim}{X}_1,...,\underset{\sim}{X}_n$ are independent where $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$. Let

$C = (c_{\alpha j})_{n\times n}$ be an orthogonal matrix then

$$\underset{\sim}{Y}_\alpha = \sum_{j=1}^{n} c_{\alpha j}\underset{\sim}{X}_j \sim Np(\underset{\sim}{\upsilon}_\alpha, \Sigma)$$

Where $\underset{\sim}{\upsilon}_\alpha = \sum_{j=1}^{n} c_{\alpha j}\underset{\sim}{\mu}_j$ and $\underset{\sim}{Y}_1, \ldots, \underset{\sim}{Y}_n$ are independent.

Theorem:- Let $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ be an independent random sample from $N_p(\underset{\sim}{\mu}, \Sigma)$. Then the MLE of μ say $\hat{\mu}$ (also the sample mean) is distributed according to

$N_p\left(\underset{\sim}{\mu}, \Sigma / n\right)$ and is independent of the MLE of Σ given by

$$\hat{\Sigma} = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\underset{\sim}{X}_{\alpha} - \underset{\sim}{\bar{X}}\right)\left(\underset{\sim}{X}_{\alpha} - \underset{\sim}{\bar{X}}\right)'$$

and $n\hat{\Sigma}$ is distributed as $\sum_{\alpha=1}^{n}\underset{\sim}{Z}_{\alpha}\underset{\sim}{Z}'_{\alpha}$ where $\underset{\sim}{Z}_{\alpha} \sim N_p\left(\underset{\sim}{0}, \Sigma\right)$ and is $\underset{\sim}{Z}_{1,\ldots\ldots}\underset{\sim}{Z}_n$ are independent.

Proof:- We have given a random sample $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ where $\underset{\sim}{X}_{\alpha} \sim N_p\left(\underset{\sim}{\mu}, \Sigma\right)$ and is independent of $\underset{\sim}{X}_{\beta}$ for $\alpha \neq \beta$ .We have the MLE's of μ and Σ are respectively given by

$$\hat{\mu} = 1/n\sum_{\alpha=1}^{n}\underset{\sim}{X}_{\alpha} = \underset{\sim}{\bar{X}} \quad \text{and} \quad \hat{\Sigma} = \frac{1}{n}\sum_{\alpha=1}^{n}\left(\underset{\sim}{X}_{\alpha} - \underset{\sim}{\bar{X}}\right)\left(\underset{\sim}{X}_{\alpha} - \underset{\sim}{\bar{X}}\right)' \quad (1)$$

Now there exists an $n \times n$ orthogonal matrix $B = \left(b_{\alpha\beta}\right)$ with the last row i.e.

$$b_{n\beta} = 1/\sqrt{n} \quad \forall \beta \qquad \textbf{(1.a)}$$

Let us define a new random sample $\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_n$ from the given random sample $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ using the orthogonal transformation from the orthogonal matrix B .

Thus

$$\underset{\sim}{Z}_{\alpha} = \sum_{\beta=1}^{n} b_{\alpha\beta}\underset{\sim}{X}_{\beta} \quad \text{for } \alpha = 1, 2, \cdots, n \qquad \textbf{---(2)}$$

In particular,

$$\underset{\sim}{Z}_n = \sum_{\beta} b_{n\beta}\underset{\sim}{X}_{\beta}$$

$$= \sum_{\beta}\frac{1}{\sqrt{n}}\underset{\sim}{X}_{\beta} \qquad \text{[The last row of } B \text{ is as given in (1.a)]}$$

$$\underset{\sim}{Z}_n = \sqrt{n}\underset{\sim}{\bar{X}} \qquad \text{[From (1)]} \qquad \textbf{--- (3)}$$

Let us consider

$$\sum_{\alpha=1}^{n}\underset{\sim}{Z}_{\alpha}\underset{\sim}{Z}'_{\alpha} = \sum_{\alpha=1}^{n}\left(\sum_{i=1}^{n} b_{\alpha i}\underset{\sim}{X}_i\right)\left(\sum_{j=1}^{n} b_{\alpha j}\underset{\sim}{X}_j\right)' \qquad \text{[Using (2)]}$$

i.e.

$$= \sum_{\alpha=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{\alpha i} b_{\alpha j} \underset{\sim}{X}_i \underset{\sim}{X}_j'$$

$$= \sum_{i=j=1}^{n}\sum_{\alpha=1}^{n} b_{\alpha i}^2 \underset{\sim}{\mathbf{X}}_\mathbf{i} \underset{\sim}{\mathbf{X}}_\mathbf{i}' + \sum_{i\neq j=1}^{n}\sum_{\alpha=1}^{n} b_{\alpha i} b_{\alpha j} \underset{\sim}{\mathbf{X}}_\mathbf{i} \underset{\sim}{\mathbf{X}}_\mathbf{j}'$$

$$= \sum_{i=1}^{n} \underset{\sim}{\mathbf{X}}_\mathbf{i} \underset{\sim}{\mathbf{X}}_\mathbf{i}' \left(\sum_{\alpha=1}^{n} b_{\alpha i}^2 \right) + \sum_{i\neq j=1}^{n} \underset{\sim}{\mathbf{X}}_\mathbf{i} \underset{\sim}{\mathbf{X}}_j' \left(\sum_{\alpha=1}^{n} b_{\alpha i} b_{\alpha j} \right)$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha} \underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha}' \qquad \text{--- (4)}$$

[$\because$ B is the orthogonal matrix and as a consequence

$$\sum_{\alpha=1}^{n} b_{\alpha i} b_{\beta j} = 0 \quad \text{and} \quad \sum_{\alpha=1}^{n} b_{\alpha i}^2 = 1$$]

Now consider $n\hat{\boldsymbol{\Sigma}}$ from (1) i.e.

$$n\hat{\boldsymbol{\Sigma}} = \sum_{\alpha=1}^{n} \left(\underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha} - \bar{\underset{\sim}{\mathbf{X}}} \right) \left(\underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha} - \bar{\underset{\sim}{\mathbf{X}}} \right)'$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha} \underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha}' - \sum_{\alpha=1}^{n} \bar{\underset{\sim}{\mathbf{X}}} \bar{\underset{\sim}{\mathbf{X}}}'$$

$$\left(\because \bar{\underset{\sim}{\mathbf{X}}} \sum_{\alpha=1}^{n} \left(\underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha} - \bar{\underset{\sim}{\mathbf{X}}} \right)' = 0 \right)$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha} \underset{\sim}{\mathbf{X}}_\boldsymbol{\alpha}' - n \bar{\underset{\sim}{\mathbf{X}}} \bar{\underset{\sim}{\mathbf{X}}}'$$

$$= \sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} - \underset{\sim}{\mathbf{Z}}_{\mathbf{n}} \underset{\sim}{\mathbf{Z}}'_{\mathbf{n}} \qquad \text{[Using (3) and (4)]}$$

$$= \sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} \qquad \text{--- (5)}$$

From (3) and (5) we observe that $\bar{\underset{\sim}{\mathbf{X}}}$ ($\hat{\underset{\sim}{\boldsymbol{\mu}}}$) is distributed according to the distribution of $\underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$ and $n\hat{\boldsymbol{\Sigma}}$ (and hence $\hat{\boldsymbol{\Sigma}}$)is distributed according to the distribution of $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n-1}}$.

Also, since $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$ are obtained from $\underset{\sim}{\mathbf{X}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{X}}_{\mathbf{n}}$ using the orthogonal linear transformation (using orthogonal matrix $\mathbf{B}$) $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$ are independently distributed as Multivariate normal distribution with common covariance matrix ' $\boldsymbol{\Sigma}$ '. Therefore $\hat{\underset{\sim}{\boldsymbol{\mu}}}$ and $\hat{\boldsymbol{\Sigma}}$ are independently distributed.

Now let us obtain the mean vector of $\underset{\sim}{\mathbf{Z}}_{\mathbf{1}}, \cdots, \underset{\sim}{\mathbf{Z}}_{\mathbf{n}}$

From (3)

$$E\left(\underset{\sim}{\mathbf{Z}}_{\mathbf{n}}\right) = \sqrt{n} E\left(\bar{\underset{\sim}{\mathbf{X}}}\right)$$

$$= \sqrt{n} \frac{1}{n} \left(E\left(\underset{\sim}{\mathbf{X}}_{\mathbf{1}} + \underset{\sim}{\mathbf{X}}_{\mathbf{2}} + \cdots + \underset{\sim}{\mathbf{X}}_{\mathbf{n}} \right) \right)$$

[$\because \underset{\sim}{\mathbf{X}}_{\mathbf{i}}{}'s$ are independent]

$$= \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} E\left(X_i\right) \quad = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{\mu} = \sqrt{n} \underset{\sim}{\mu}$$

$$\left(\because \underset{\sim}{\mathbf{X}}_{\mathbf{i}} \sim Np\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right) \right)$$

Thus $\underset{\sim}{\mathbf{Z}}_{\mathbf{n}} \sim Np\left(\sqrt{n}\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$

i.e. $\sqrt{n}\bar{\underset{\sim}{\mathbf{X}}} \sim Np\left(\sqrt{n}\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$

i.e. $\bar{\underset{\sim}{\mathbf{X}}} \sim Np\left(\underset{\sim}{\boldsymbol{\mu}}, \frac{\boldsymbol{\Sigma}}{n}\right)$

From (2), we have

$$E\left(\underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}}\right) = \sum_{\beta=1}^{n} b_{\alpha\beta} E\left(\underset{\sim}{\mathbf{X}}_{\boldsymbol{\beta}}\right) \quad [\because \underset{\sim}{\mathbf{X}}_{\boldsymbol{\beta}}\text{'}s \text{ are independent }]$$

$$= \sum_{\beta=1}^{n} b_{\alpha\beta} \underset{\sim}{\boldsymbol{\mu}}$$

$$= \sum_{\beta=1}^{n} b_{\alpha\beta} \frac{1}{\sqrt{n}} \sqrt{n} \underset{\sim}{\boldsymbol{\mu}}$$

$$= \sqrt{n} \underset{\sim}{\boldsymbol{\mu}} \sum_{\beta=1}^{n} b_{\alpha\beta} b_{n\beta} \qquad [\because b_{n\beta} = \frac{1}{\sqrt{n}}]$$

$$= \underset{\sim}{\mathbf{0}} \quad \forall \alpha \neq n$$

Thus each of $\underset{\sim}{\mathbf{z}}_{\mathbf{1}}, \ldots, \underset{\sim}{\mathbf{z}}_{\mathbf{n-1}}$ are distributed as $Np\left(\underset{\sim}{0}, \boldsymbol{\Sigma}\right)$ .Therefore from (5)

$n\hat{\boldsymbol{\Sigma}}$ is distributed as $\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}} \underset{\sim}{\mathbf{z}}'_{\boldsymbol{\alpha}}$, where $\underset{\sim}{\mathbf{z}}_{\boldsymbol{\alpha}} \sim N\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$

and is independent of $\underset{\sim}{\mathbf{z}}_{\boldsymbol{\beta}} \left(\beta \neq \alpha\right)$

Thus the MLE's of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are independently distributed.

Hence the proof.

4.5 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN VECTOR $\left(\bar{\mathbf{X}}\right)$:

In multivariate normal (MVN) models, the sampling distributions for the sample mean vector and the sample covariance matrix are the Multivariate Normal distribution and the Wishart distribution, respectively. A key property is that these two sample statistics are statistically independent of each other.

Suppose a random sample of n observation vectors $\mathbf{X_1}, \mathbf{X_2}, \ldots., \mathbf{X_n}$ is drawn independently from a $\mathbf{p}$ - dimensional multivariate normal population with mean vector μ and covariance matrix Σ, denoted as $\mathbf{X}_i \sim N_p(\boldsymbol{\mu}, \Sigma)$,

The sample mean vector is defined as $\bar{\mathbf{X}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i}$

We have

$$E\left(\underset{\sim}{\bar{\mathbf{X}}}\right)=\frac{1}{n}\left(E\left(\underset{\sim}{\mathbf{X}}_1\right)+E\left(\underset{\sim}{\mathbf{X}}_2\right)+\cdots+E\left(\underset{\sim}{\mathbf{X}}_n\right)\right)$$

$$=\frac{n\underset{\sim}{\boldsymbol{\mu}}}{n}=\underset{\sim}{\boldsymbol{\mu}}$$

Thus $\underset{\sim}{\bar{\mathbf{X}}}$ is an unbiased estimator of $\underset{\sim}{\boldsymbol{\mu}}$. Thus sample mean is an unbiased estimator of the population mean vector $\underset{\sim}{\boldsymbol{\mu}}$.

The sampling distribution of the sample mean vector $\left(\underset{\sim}{\bar{\mathbf{X}}}\right)$ is also a multivariate normal distribution with:

- **Mean Vector:** The same mean vector as the population, $E\left(\underset{\sim}{\bar{\mathbf{X}}}\right)=\underset{\sim}{\boldsymbol{\mu}}$
- **Variance-covariance Matrix:** The population covariance matrix scaled by the inverse of the sample size, $Var\left(\underset{\sim}{\bar{\mathbf{X}}}\right)=\frac{1}{n}\boldsymbol{\Sigma}$.

So, the distribution is:

$$\underset{\sim}{\bar{\mathbf{X}}}\sim N_p\left(\underset{\sim}{\boldsymbol{\mu}},\frac{1}{n}\boldsymbol{\Sigma}\right)$$

This result holds exactly for any sample size when the population is MVN.

4.6 SAMPLING DISTRIBUTION OF THE SAMPLE COVARIANCE MATRIX (S):

We have from Eq. (5),

$$E\left(\hat{\boldsymbol{\Sigma}}\right)=\frac{1}{n}E\left(\sum_{\alpha=1}^{n-1}\underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha}\right)$$

$$=\frac{1}{n}\sum_{\alpha=1}^{n-1}E\left(\underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha}\right) \quad [\because z_{\alpha}{}'s \text{ are independent}]$$

$$=\frac{1}{n}\sum_{\alpha=1}^{n-1}V\left(\underset{\sim}{\mathbf{Z}}_{\alpha}\right) \quad [\because E\left(\underset{\sim}{\mathbf{Z}}_{\alpha}=\underset{\sim}{\mathbf{0}}\right)]$$

$$=\frac{1}{n}\sum_{\alpha=1}^{n-1}\boldsymbol{\Sigma}=\frac{n-1}{n}\boldsymbol{\Sigma}$$

Thus $\hat{\boldsymbol{\Sigma}}$ is not an unbiased estimator of $\boldsymbol{\Sigma}$.

But $\frac{n}{n-1}\hat{\boldsymbol{\Sigma}}=\frac{1}{n-1}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\alpha}-\underset{\sim}{\bar{\mathbf{X}}}\right)\left(\underset{\sim}{\mathbf{X}}_{\alpha}-\underset{\sim}{\bar{\mathbf{X}}}\right)'=\mathbf{S}$

(say) is an unbiased estimator of $\boldsymbol{\Sigma}$ [$\because E\left(\frac{n}{n-1}\hat{\boldsymbol{\Sigma}}\right)=\boldsymbol{\Sigma}$]

Hence $\mathbf{S}=\frac{1}{n-1}\sum_{\alpha=1}^{n}(\underset{\sim}{\mathbf{X}}_{\alpha}-\bar{\underset{\sim}{\mathbf{X}}})(\underset{\sim}{\mathbf{X}}_{\alpha}-\bar{\underset{\sim}{\mathbf{X}}})'$ is called the sample covariance matrix and is an unbiased estimator of $\boldsymbol{\Sigma}$. The sampling distribution of the scaled sample covariance matrix is the Wishart distribution**,** which is the multivariate generalization of the chi-squared distribution.

Specifically, the matrix $(n-1)S$ follows a Wishart distribution with parameters:

- **Degrees of Freedom:** $v=n-1$.
- **Scale Matrix (or parameter):** $\sum$, the population covariance matrix..

This is denoted as:

$$(n-1)S\sim W_p(v,\Sigma) \text{ or } \mathrm{W}_p(\Sigma,n-1)$$

where p is the dimension of the vectors. The Wishart distribution is a distribution over symmetric, positive-definite matrices.

4.7 SAMPLE MEAN VECTOR AND SAMPLE DISPERSION MATRIX ARE INDEPENDENT:

Theorem:- Let $\underset{\sim}{\mathbf{X}}_1,\underset{\sim}{\mathbf{X}}_2,\ldots,\underset{\sim}{\mathbf{X}}_n$ be a random sample of size n from $N_p(\mu,\Sigma)$. Then $\bar{\underset{\sim}{\mathbf{X}}}$ is distributed according to $N_p\left(\underset{\sim}{\mu},\Sigma/n\right)$ and is independent of sample covariance matrix S is given by $S=\frac{1}{n-1}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\alpha}-\bar{\underset{\sim}{\mathbf{X}}}\right)\left(\mathbf{X}_{\alpha}-\bar{\underset{\sim}{\mathbf{X}}}\right)'$ and $(n-1)S$ is distributed as $\sum_{\alpha=1}^{n}\underset{\sim}{Z}_{\alpha}Z_{\alpha}^{1}$ where $\underset{\sim}{Z}_{\alpha}\sim \mathrm{N_p}\left(\underset{\sim}{0},\Sigma\right)$ and is $Z_1,Z_2,....,Z_n$ are independent.

Proof: We have given a random sample $\underset{\sim}{\mathbf{X}}_1,\underset{\sim}{\mathbf{X}}_2,\ldots,\underset{\sim}{\mathbf{X}}_n$ where $\underset{\sim}{X}_{\alpha}\sim \mathrm{N_p}\left(\underset{\sim}{\mu},\Sigma\right)$ and is independent of $\underset{\sim}{X}_{\beta}$ for $\alpha\neq\beta$ where the sample mean of $\bar{\underset{\sim}{\mathbf{X}}}$ and sample variance-covariance matrix $\mathbf{S}$ respectively given by $\bar{\underset{\sim}{X}}=\frac{1}{n}\sum_{\alpha=1}^{n}\underset{\sim}{X}_{\alpha}$

$$S=\frac{1}{n-1}\sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathbf{X}}_{\alpha}-\bar{\underset{\sim}{\mathbf{X}}}\right)\left(\mathbf{X}_{\alpha}-\bar{\underset{\sim}{\mathbf{X}}}\right)' \qquad \ldots\ldots(1)$$

Now $\exists$ a $n\times n$ orthogonal matrix $B=\left(b_{X\beta}\right)$ with the last row

i.e., $b_{n\beta}=\frac{1}{\sqrt{n}}\forall\beta$

$$\left(\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}}\right) \qquad \ldots\ldots(1.\ \mathrm{A})$$

Let us define a new random sample $\underset{\sim}{\mathbf{Z}}_1,\underset{\sim}{\mathbf{Z}}_2,\ldots,\underset{\sim}{\mathbf{Z}}_n$ from the given random sample $\underset{\sim}{\mathbf{X}}_1,\underset{\sim}{\mathbf{X}}_2,\ldots,\underset{\sim}{\mathbf{X}}_n$ using the orthogonal transformation from the orthogonal matrix $\mathbf{B}$.

Thus,

$$\underset{\sim}{\mathbf{Z}}_{\alpha} = \sum_{\beta=1}^{n} b_{\alpha\beta}\underset{\sim}{\mathbf{X}}_{\beta} \quad \text{for } \alpha = 1, 2, \cdots, n \qquad \text{.....(2)}$$

In particular, $\underset{\sim}{\mathbf{Z}}_{\mathbf{n}} = \sum_{\beta} b_{n\beta}\underset{\sim}{\mathbf{X}}_{\beta}$

$$= \sum_{\beta} \frac{1}{\sqrt{n}}\underset{\sim}{\mathbf{X}}_{\beta}$$ [The last row of $\mathbf{B}$ is as given in (1.A)]

$$\underset{\sim}{\mathbf{Z}}_{\mathbf{n}} = \frac{1}{\sqrt{n}}\ \mathrm{n}\underset{\sim}{\bar{\mathrm{X}}}$$

$$\underset{\sim}{\mathbf{Z}}_{\mathbf{n}} = \sqrt{n}\underset{\sim}{\bar{\mathrm{X}}} \qquad \text{..... (3)}$$

Let us consider

$$\sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha} = \sum_{\alpha=1}^{n}\left(\sum_{i=1}^{n} b_{\alpha i}\underset{\sim}{\mathbf{X}}_{\mathbf{i}}\right)\left(\sum_{j=1}^{n} b_{\alpha j}\underset{\sim}{\mathbf{X}}_{\mathbf{j}}\right)'$$

$$\sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha} = \sum_{\alpha=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{\alpha i}b_{\alpha j}\underset{\sim}{X}_{i}\underset{\sim}{X}_{j}'$$

$$= \sum_{i=j=1}^{n}\sum_{\alpha=1}^{n} b_{\alpha i}^{2}\underset{\sim}{\mathbf{X}}_{\mathbf{i}}\underset{\sim}{\mathbf{X}}'_{\mathbf{i}} + \sum_{i\neq j=1}^{n}\sum_{\alpha=1}^{n} b_{\alpha i}b_{\alpha j}\underset{\sim}{\mathbf{X}}_{\mathbf{i}}\underset{\sim}{\mathbf{X}}'_{\mathbf{j}}$$

$$\sum_{\alpha=1}^{n} \underset{\sim}{\mathbf{Z}}_{\alpha}\underset{\sim}{\mathbf{Z}}'_{\alpha} = \sum_{i=1}^{n} \underset{\sim}{\mathbf{X}}_{\mathbf{i}}\underset{\sim}{\mathbf{X}}'_{\mathbf{i}}\left(\sum_{\alpha=1}^{n} b_{\alpha i}^{2}\right) + \sum_{i\neq j=1}^{n} \underset{\sim}{\mathbf{X}}_{\mathbf{i}}\underset{\sim}{\mathbf{X}}'_{j}\left(\sum_{\alpha=1}^{n} b_{\alpha i}b_{\alpha j}\right)$$

$\{$*In orthogonal matrix sum of squares is one*$\}$

$$= \sum_{\mathrm{i}=1}^{\mathrm{n}} \underset{\sim}{\mathrm{X}}_{\mathrm{i}}\underset{\sim}{\mathrm{X}}'_{\mathrm{i}}(1) + \sum_{\mathrm{i}\neq\mathrm{j}=1}^{\mathrm{n}} \underset{\sim}{\mathrm{X}}_{\mathrm{i}}\underset{\sim}{\mathrm{X}}'_{\mathrm{j}}(0)$$

$$= \sum_{\alpha=1}^{\mathrm{n}} \underset{\sim}{\mathrm{Z}}_{\alpha}\underset{\sim}{\mathrm{Z}}'_{\alpha} \qquad \text{……(3A)}$$

Now consider $(n-1)S$ from (1) i.e.,

$$(n-1)S = \sum_{\alpha=1}^{n}\left(\underset{\sim}{\mathrm{X}}_{\alpha}-\underset{\sim}{\bar{\mathrm{X}}}\right)\left(\underset{\sim}{\mathrm{X}}_{\alpha}-\underset{\sim}{\bar{\mathrm{X}}}\right)' \qquad \text{..... (4)}$$

$$(n-1)S = \sum_{\alpha=1}^{n}\left(\underset{\sim}{X}_{\alpha}-\bar{\underset{\sim}{X}}\right)\left(\underset{\sim}{X}_{\alpha}-\bar{\underset{\sim}{X}}\right)'$$

$$(n-1)S = \sum_{\alpha=1}^{n}\underset{\sim}{X}_{\alpha}\underset{\sim}{X}'_{\alpha} - \sum_{\alpha=1}^{n}\bar{\underset{\sim}{X}}\bar{\underset{\sim}{X}}'$$

$$= \sum_{\alpha=1}^{n}\underset{\sim}{X}_{\alpha}\underset{\sim}{X}'_{\alpha} - n\bar{\underset{\sim}{X}}\bar{\underset{\sim}{X}}'$$

$$(n-1)S = \sum_{\alpha=1}^{n}\underset{\sim}{Z}_{\alpha}\underset{\sim}{Z}'_{\alpha} - \underset{\sim}{Z}_{n}\underset{\sim}{Z}'_{n}$$

$$= \sum_{\alpha=1}^{n-1}\mathbf{\underset{\sim}{Z}_{\alpha}\underset{\sim}{Z}'_{\alpha}} \qquad \text{…….(5)}$$

From (3) and (5) we observe that $\bar{\underset{\sim}{X}}$ is distributed according to the distribution of $\mathbf{\underset{\sim}{Z}_{n}}$ and $(n-1)S$ is distributed according to the distribution of .

Also, since $\mathbf{\underset{\sim}{Z}_1,\underset{\sim}{Z}_2,\ldots,\underset{\sim}{Z}_n}$ are obtained from $\mathbf{\underset{\sim}{X}_1,\underset{\sim}{X}_2,\ldots,\underset{\sim}{X}_n}$ using the orthogonal linear transformation $\mathbf{\underset{\sim}{Z}_1,\underset{\sim}{Z}_2,\ldots,\underset{\sim}{Z}_n}$ are independently distributed as Multivariate normal distribution with common covariance matrix are independently distributed.

Now, let us obtain the mean vector of $\mathbf{\underset{\sim}{Z}_1,\underset{\sim}{Z}_2,\ldots,\underset{\sim}{Z}_n}$

From (3) $E\left(\mathbf{\underset{\sim}{Z}_n}\right) = \sqrt{n}E\left(\bar{\underset{\sim}{X}}\right)$

$$= \sqrt{n}\,\frac{1}{n}\left(E\left(\mathbf{\underset{\sim}{X}_1}+\mathbf{\underset{\sim}{X}_2}+\cdots+\mathbf{\underset{\sim}{X}_n}\right)\right)$$

$[\because \mathbf{\underset{\sim}{X}_i}'s$ are independent]

$$= \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}E\left(X_i\right) = \sqrt{n}\,\frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{\mu} = \sqrt{n}\underset{\sim}{\mu}$$

$\left(\because \mathbf{\underset{\sim}{X}_i} \sim Np\left(\mathbf{\underset{\sim}{\mu}, \Sigma}\right)\right)$

Thus $\mathbf{\underset{\sim}{Z}_n} \sim Np\left(\sqrt{n}\mathbf{\underset{\sim}{\mu}, \Sigma}\right)$

i.e. $\sqrt{n}\bar{\underset{\sim}{X}} \sim Np\left(\sqrt{n}\mathbf{\underset{\sim}{\mu}, \Sigma}\right)$

i.e. $\bar{\underset{\sim}{X}} \sim Np\left(\mathbf{\underset{\sim}{\mu}}, \frac{\mathbf{\Sigma}}{n}\right)$

From (2), we have

$$E\left(\mathbf{\underset{\sim}{Z}_{\alpha}}\right) = \sum_{\beta=1}^{n}b_{\alpha\beta}E\left(\mathbf{\underset{\sim}{X}_{\beta}}\right) \quad [\because \mathbf{\underset{\sim}{X}_{\beta}}'s \text{ are independent}]$$

$$E\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right)=\sum_{\beta=1}^{n} b_{\alpha\beta}E\left(\underset{\sim}{\mathbf{X}}_{\beta}\right)$$

$$E\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right)=\sum_{\beta=1}^{n} b_{\alpha\beta}\ \underset{\sim}{\boldsymbol{\mu}}$$

$$E\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right)=\sum_{\beta=1}^{n} b_{\alpha\beta}\frac{1}{\sqrt{n}}\sqrt{n}\ \underset{\sim}{\boldsymbol{\mu}}$$

$$E\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right)=\sqrt{n}\ \underset{\sim}{\boldsymbol{\mu}}\sum_{\beta=1}^{n} b_{\alpha\beta}b_{n\beta}$$

$$E\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right)=\underset{\sim}{\mathbf{0}}\quad\forall\alpha\neq n$$

Thus each of $\underset{\sim}{\mathbf{z}}_{1},\underset{\sim}{\mathbf{z}}_{2},\ldots,\underset{\sim}{\mathbf{z}}_{n-1}$ are distributed as $\mathbf{N}_{p}\left(\underset{\sim}{\mathbf{0}},\boldsymbol{\Sigma}\right)$.

Therefore from (5) $(\mathbf{n-1})\mathbf{S}$ is distributed as $\sum_{\alpha=1}^{n-1}\underset{\sim}{\mathbf{z}}_{\alpha}\underset{\sim}{\mathbf{z}}_{\alpha}'$, where $\underset{\sim}{\mathbf{z}}_{\alpha}\sim N\left(\underset{\sim}{\mathbf{0}},\boldsymbol{\Sigma}\right)$

and is independent of $\underset{\sim}{\mathbf{z}}_{\beta}\left(\beta\neq\alpha\right)$

∴ Sample mean vector $\underset{\sim}{\bar{\mathbf{X}}}$ and sample covariance matrix $\mathbf{S}$ respectively are independently distributed.

4.8 SUMMARY:

This lesson focused on sampling from the multivariate normal (MVN) distribution and the associated maximum likelihood estimation (MLE) and sampling distributions of key sample statistics**.** The multivariate normal distribution was introduced as a fundamental model in multivariate analysis, characterized completely by its mean vector and variance–covariance matrix**.**

A random sample drawn from an MVN population was considered, and the behaviour of important sample statistics was examined. The sample mean vector was shown to follow a multivariate normal distribution with the same mean vector and a scaled covariance matrix. The sample covariance matrix was shown to follow the Wishart distribution**,** which serves as the multivariate analogue of the chi-square distribution.

These sampling distributions form the theoretical basis for multivariate inference. Maximum Likelihood Estimation was applied to estimate the unknown parameters of the MVN distribution. The MLE of the mean vector was obtained as the sample mean vector**,** while the MLE of the variance-covariance matrix was derived as a scaled version of the sample dispersion matrix. The distinction between the MLE and the unbiased estimator of the covariance matrix was highlighted.

A key and unique result of multivariate normal theory-that the sample mean vector and the sample covariance matrix are statistically independent-was also

established. This property greatly simplifies the development of multivariate test statistics and confidence regions.

Overall, understanding sampling and estimation under the multivariate normal framework is essential for effective statistical modelling, inference, and data analysis involving multiple correlated variables**.**

4.9 SELF-ASSESSMENT QUESTIONS:

1. Derive the MLEs of the mean vector, μ and the variance-covariance matrix , Σ based on a random sample of size n draw from the normal population $N_p(\mu,\Sigma)$.
2. Obtain the maximum likelihood estimates of the mean vector and the covariance matrix in a p - variate normal.
3. Describe the method of sampling from a multivariate normal distribution.
4. Derive the MLEs of the mean vector and covariance matrix for an MVN distribution.
5. What are the sampling distributions of the sample mean and sample covariance matrix?
6. In the p -variate normal case, show that mean vector and the sample variance- covariance matrix are independently distributed.
7. Derive Sample mean vector for Multivariate normal distribution.
8. Find the covariance matrix of the multivariate normal distribution which has the quadratic form $2x_1^2 + x_2^2 + 4x_2^2 - x_1x_2 - 2x_1x_2$.
9. Derive the Samping distribution of Sample Variance-Covariance Matrix.
10. Let $X_1, X_2, ..., X_n$ be a random sample of size n from $N_p(\mu,\Sigma)$. If $\bar{X}$ denotes the sample and S denote the sample covariance matrix. Then determine the distribution of $\bar{X}$ and $(n-1)S$.

4.10 SUGGESTED READINGS:

1. Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern.
2. Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International.
3. Mardia, Kent & Bibby. Multivariate Analysis
4. Kshirsagar, A.M. Multivariate Analysis
5. Brenner, D., Bilodeau, M. (1999). Theory of Multivariate Statistics. Germany: Springer.
6. Giri Narayan C. (1995). Multivariate Statistical Analysis.
7. Tong, Y. L. (1990). The Multivariate Normal Distribution.

Dr. U. Ramkiran

LESSON-5

WISHART's DISTRIBUTION

OBJECTIVES:

- **Understanding the role of Wishart distribution**
  Explain the importance of the Wishart distribution as the sampling distribution of the sample covariance matrix in multivariate normal populations.
- **Define the Wishart distribution formally**
  State the definition, parameters, and notation of the Wishart distribution and relate it to the chi-square distribution.
- **Understanding connections with multivariate normal theory**
  Explain how the Wishart distribution arises from multivariate normal random vectors and its role in multivariate inference.
- **Understanding the properties of Wishart distribution**
  Build the necessary theoretical background for further study in multivariate hypothesis testing and estimation.

STRUCTURE

5.1 INTRODUCTION:

The Wishart distribution is a probability distribution used in statistics and probability theory to describe the behaviour of a sample covariance matrix or a sample correlation matrix. It is named after John Wishart, who first introduced it in 1928.

Given a set of p-dimensional multivariate normal random vectors, the Wishart distribution describes the probability distribution of the sample covariance matrix, which is a $p \times p$ matrix. The distribution is characterized by two parameters: the degrees of freedom (n) and the scale matrix (Σ).

Multivariate inference deals with statistical procedures for drawing conclusions about population parameters when observations are in the form of vectors rather than single measurements. Unlike univariate inference, multivariate inference accounts for the interrelationships among variables.

The Wishart distribution has several important applications in statistics and data analysis, including:

(i) Covariance Matrix Estimation

(ii) Multivariate Analysis of Variance (MANOVA)

(iii) Principal Component Analysis (PCA)

(iv) Factor Analysis

(v) Bayesian Analysis

The Wishart distribution is a generalization of the Chi-Squared distribution and is closely related to other distributions, such as the multivariate gamma distribution and the inverse Wishart distribution.

5.2 DEFINITION OF WISHART DISTRIBUTION:

From Lesson 4, we know that the sample mean ($\bar{\underset{\sim}{\mathbf{X}}}$) and the sample covariance matrix $\left(\mathbf{S} = \frac{1}{n-1} \sum_{\alpha=1}^{n} (\underset{\sim}{\mathbf{X}}_{\alpha} - \bar{\underset{\sim}{\mathbf{X}}})(\underset{\sim}{\mathbf{X}}_{\alpha} - \bar{\underset{\sim}{\mathbf{X}}})' \right)$ are independently distributed. Also it may be seen that

$$\bar{\underset{\sim}{\mathbf{X}}} \sim N\left(\underset{\sim}{\boldsymbol{\mu}}, \frac{\boldsymbol{\Sigma}}{n} \right)$$

$$S = \frac{1}{n-1} \sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} \text{ , where } \underset{\sim}{\mathbf{Z}}_{\alpha} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$$

i.e. **(n-1)S** is distributed according to the distribution of $\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha}$, where $\underset{\sim}{\mathbf{Z}}_{1}, \ldots, \underset{\sim}{\mathbf{Z}}_{n}$ are independently distributed as $N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$. The matrix $\sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha}$ is called " Wishart random matrix" and it is distributed according to "wishart distribution" with (n-1) degrees of freedom.

And is denoted as $W_{n-1}\left(\boldsymbol{\Sigma}\right)$, where $\boldsymbol{\Sigma}$ is the covariance matrix of Wishart random matrix . Hence it may be noted that

$$\boldsymbol{\Sigma} = E\left(\frac{1}{n-1} \sum_{\alpha=1}^{n-1} \underset{\sim}{\mathbf{Z}}_{\alpha} \underset{\sim}{\mathbf{Z}}'_{\alpha} \right) = E\left(\frac{\text{Wishart random matrix}}{\text{degrees of freedom}} \right)$$

Thus **(n-1)S** (and hence **S**) provides independent information about $\boldsymbol{\Sigma}$ and the distribution os S does not depend on $\underset{\sim}{\boldsymbol{\mu}}$. This allows us to construct a statistics for making inferences about $\underset{\sim}{\boldsymbol{\mu}}$ as we shall see in the later lessons.

Definition of Wishart distribution:

If $\underset{\sim}{Z}_1, \underset{\sim}{Z}_2, \ldots, \underset{\sim}{Z}_n$ are independently distributed as $N_P(\underset{\sim}{0}, \Sigma)$ Then the matrix A written as $A = \sum_{i=1}^{n} \underset{\sim}{Z}_i \underset{\sim}{Z}_i'$ is called as Wishart random matrix and its distribution is called as Wishart distribution with 'n' degrees of freedom and if may be denoted as $A \sim W_n(\Sigma)$, where Σ is the parametric matrix of wishart distribution. The P.d.f. of wishart distribution is given by

P.D.F. Of Wishart Distribution :

The p.d.f. of $A \sim W_n(\Sigma)$ is given by

$$W_n\left(\frac{A}{\Sigma}\right) = \frac{|A|^{(n-p-1)/2} \, e^{tr(A\Sigma^{-1})/2}}{2^{np/2} \, \pi^{p(p-1)/4} \, |\Sigma|^{n/2} \prod_{i=1}^{p} \left|\left(\frac{1}{2}(n+1-i)\right)\right|}$$

A is positive definite and $\Gamma(.)$ is gamma function.

5.3 SOME PROPERTIES OF THE WISHART DISTRIBUTION:

Property – 1:

Sum Property: The sum of independent Wishart matrices with the same covariance matrix is also Wishart.

- If A_1 is distributed as $W_{m_1}(\Sigma)$ independently of A_2, which is distributed as $W_{m_2}(\Sigma)$, then $A_1 + A_2$ is distributed as $W_{m_1+m_2}(\Sigma)$. That is the degrees of freedom are added.

Proof: - Since $A_1 \sim W_{m_1}(\Sigma)$.

A_1 may be written as

$$A_1 = \sum_{\alpha=1}^{m_1} \underset{\sim}{Z}_\alpha \underset{\sim}{Z}_\alpha' \quad \text{, where} \quad \underset{\sim}{Z}_\alpha \sim Np(\underset{\sim}{0}, \Sigma)$$

Also since A_2 is independently distributed as $W_{m_2}(\Sigma)$,

We may write

$$\mathbf{A}_2 = \sum_{\alpha=m_1+1}^{m_1+m_2} \underset{\sim}{\mathbf{z}}_{\alpha}\underset{\sim}{\mathbf{z}}'_{\alpha} \text{ , where } \underset{\sim}{\mathbf{z}}_{\alpha} \sim Np\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$$

Since $\mathbf{A}_1$ and $\mathbf{A}_2$ are independent, $\underset{\sim}{\mathbf{z}}_{\mathbf{1}}, \ldots, \underset{\sim}{\mathbf{z}}_{\mathbf{m_1+m_2}}$ are independent and as a consequence .

$$\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2 = \sum_{\alpha=1}^{m_1+m_2} \underset{\sim}{\mathbf{z}}_{\alpha}\underset{\sim}{\mathbf{z}}'_{\alpha} \sim \mathrm{W}_{m_1+m_2}\left(\boldsymbol{\Sigma}\right)$$

Hence the proof.

Property – 2:

- If $\mathbf{A} \sim \mathrm{W}_m\left(\boldsymbol{\Sigma}\right)$,then $\mathbf{CAC}' \sim \mathrm{W}_m\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$

Proof:-Given $\mathbf{A} \sim \mathrm{W}_m\left(\boldsymbol{\Sigma}\right)$

$$\therefore \mathbf{A} = \sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{z}}_{\alpha}\underset{\sim}{\mathbf{z}}'_{\alpha} \text{ , where } \underset{\sim}{\mathbf{z}}_{\alpha} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\right)$$

$$\mathbf{CAC}' = \mathrm{W}_m \sum_{\alpha=1}^{m} \mathbf{C}\underset{\sim}{\mathbf{z}}_{\alpha}\underset{\sim}{\mathbf{z}}'_{\alpha}\mathbf{C}' = \sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Y}}_{\alpha}\underset{\sim}{\mathbf{Y}}'_{\alpha}$$

where $\underset{\sim}{\mathbf{Y}}_{\alpha} = \mathbf{C}\underset{\sim}{\mathbf{z}}_{\alpha} \sim N_p\left(\underset{\sim}{\mathbf{0}}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$

[$\because E\left(\underset{\sim}{\mathbf{Y}}_{\alpha}\right) = CE\left(\underset{\sim}{\mathbf{z}}_{\alpha}\right) = \underset{\sim}{\mathbf{0}}$

$V\left(\underset{\sim}{\mathbf{Y}}_{\alpha}\right) = V\left(\mathbf{C}\underset{\sim}{\mathbf{z}}_{\alpha}\right) = \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'$ and $\underset{\sim}{\mathbf{Y}}_{\alpha}$ is normal random vector]

$$\mathbf{CAC}' \sim \mathrm{W}_m\left(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}'\right)$$

Hence the proof .

5.4 IMPORTANCE OF WISHART DISTRIBUTION

- **Sampling Distribution:** Models the distribution of sample covariance matrices from multivariate normal data, essential for understanding data variability.
- **Bayesian Statistics:** Acts as a conjugate prior for the inverse of the covariance matrix (precision matrix), simplifying Bayesian computations.
- **Multivariate Inference:** Underpins likelihood-ratio tests and statistics like Hotelling's T^2, used for multivariate mean comparisons.
- **Eigenvalue Analysis:** Its eigenvalues reveal insights into data dimensionality and structure, used in random matrix theory and analyzing functional brain networks.
- **Applications:** Found in wireless communication (MIMO channels), finance, and medical imaging (Diffusion Tensor Imaging).

5.5 SUMMARY:

The Wishart distribution provides the theoretical foundation for statistical inference involving covariance matrices, particularly when dealing with data sampled from a multivariate normal population. Its properties, such as additivity and the ability to model positive definite matrices, make it an efficient and mathematically convenient tool for the analysis of variance-covariance structures in high-dimensional data.

Its primary use is in modeling sample covariance matrices and serving as a key component in more complex Bayesian models, such as the Normal-Wishart conjugate prior for vector autoregression models. The development of related models, such as the Wishart Autoregressive (WAR) processes, demonstrates its continued relevance in modern fields like quantitative finance and signal processing for modeling time-varying volatility.

5.6 SELF-ASSESSMENT QUESTIONS:

1. Define the Wishart distribution. State and prove the additive property of Wishart distribution.
2. Let A follows Wishers distribution $W_p(n,\Sigma)$ what is the distribution of $W_1(n,\Sigma)$.
3. State the conditions under which a random matrix follows a Wishart distribution. What is the relationship between the Wishart distribution and the multivariate normal distribution?
4. Explain the importance of the Wishart distribution in multivariate analysis.
5. Show that the Wishart distribution generalizes the chi-square distribution.
6. State and explain the reproductive property of the Wishart distribution.
7. Derive the marginal distribution of a principal submatrix of a Wishart matrix.
8. Explain the conditional distribution of partitioned Wishart matrices.
9. Discuss the role of Wishart distribution in multivariate hypothesis testing.

5.7 SUGGESTED READINGS:

1. Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern.
2. Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International.

3. Morrison, D.F. (2004): Multivariate Statistical Methods (Fourth Edition). Duxbury Press, New York.
4. Rao, C.R. (2001): Linear Statistical Inference and its Applications (Second Edition), Wiley Inter Science, New York.
5. Mardia, K.V., Kent, J. T and Bibby, J. M. (1979): Multivariate Analysis. Academic Press, New York.
6. Brenner, D., Bilodeau, M. (1999). Theory of Multivariate Statistics. Germany: Springer.
7. Giri Narayan C. (1995). Multivariate Statistical Analysis.
8. Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
9. Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

Dr. U. Ramkiran

LESSON-6

HOTELLING'S T^2 STATISTIC AND ITS APPLICATIONS

OBJECTIVES:

After completing this lesson, students will be able to:

- Understand the inferences about mean vector(s) of a MVN distribution(s) and the need for multivariate testing when multiple correlated variables are involved.
- Define and derive Hotelling's T^2 statistic and explain its role as the multivariate analogue of Student's *t*-test.
- State and interpret the assumptions underlying Hotelling's T^2 test and assess their importance in practical applications.
- Perform statistical inference on population mean vectors**,** including one-sample and two-sample Hotelling's T^2 problems.
- Develop appropriate test statistics for testing hypotheses about mean vectors when the population covariance matrix is known and unknown.
- Understand the distribution of Hotelling's T^2 and its transformation to the F-distribution for hypothesis testing.
- Derive Hotelling's T^2 as a Likelihood Ratio Test (LRT) under multivariate normality assumptions.
- Understand and apply the invariance property of Hotelling's T^2 statistic under linear transformations.

STRUCTURE:

6.1 Introduction

6.2 Inferences About Mean Vector(s)

6.3 Developing Test Statistics when Σ is known

6.4 Hotelling's T^2 Statistic

6.4.1 Assumptions of Hotelling's T^2

6.4.2 Definition of Hotelling's T^2 - statistic (distribution)

6.5 Deriving Hotelling's T^2-Statistic as the Likelihood Ratio Test of $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$

6.6 Invariance Property of Hotelling's T^2

6.7 Applications of Hotelling's T^2 - statistic

6.7.1 For testing the significance of one sample mean vector.

6.7.2 A Two Sample Problem when the covariance matrices are equal but unknown

6.7.3 The Two Sample Problem when the covariance matrices are not equal

6.8 Distribution of Hotelling's T^2

6.1 INTRODUCTION:

In many practical situations, researchers are interested in comparing groups or testing hypotheses involving several related response variables simultaneously**.** Performing separate univariate t-tests for each variable ignores the correlations among the variables and leads to an increased risk of Type I error**.** To address this limitation, Hotelling introduced the T^2 statistic, which serves as the multivariate extension of Student's t-test**.**

Hotelling's T^2 statistic is designed to test hypotheses about mean vectors of multivariate normal populations. It provides a single test that jointly considers all variables and incorporates their covariance structure**,** thereby offering a more powerful and informative assessment of group differences than separate univariate tests.

The statistic plays a central role in multivariate inference and forms the theoretical foundation for several important techniques, including Multivariate Analysis of Variance (MANOVA)**.** In particular, the two-sample Hotelling's T^2 test is equivalent to a one-way MANOVA with two groups.

Hotelling's T^2 has wide applicability in fields such as medicine, engineering, psychology, education, economics, and quality control**,** where outcomes are inherently multivariate. By accounting for inter-relationships among variables, it enables researchers to draw valid and meaningful conclusions from complex multivariate data.

Thus, Hotelling's T^2 statistic is a fundamental and powerful tool in multivariate analysis, providing a coherent framework for hypothesis testing involving multiple correlated variables.

6.2 INFERENCES ABOUT MEAN VECTOR (S): ONE SAMPLE PROBLEM:

Suppose $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, ..., \underset{\sim}{\mathbf{X}}_n$ is a random sample from a multivariate normal population. Now, our statistical problem is whether the given sample has come from the multivariate normal population, whose mean vector is given by $\underset{\sim}{\mu} = \underset{\sim}{\mu}_0$. In other words, we have to test

$$H_0 : \underset{\sim}{\boldsymbol{\mu}} = \underset{\sim}{\boldsymbol{\mu}}_0 \quad \text{vs} \quad H_1 : \underset{\sim}{\boldsymbol{\mu}} \neq \underset{\sim}{\boldsymbol{\mu}}_0$$

based on the given random sample $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, ..., \underset{\sim}{\mathbf{X}}_n$.

TWO SAMPLE PROBLEM:

Suppose we have two different samples from two different multivariate normal populations $N_p(\underset{\sim}{\boldsymbol{\mu}}^{(1)}, \boldsymbol{\Sigma})$ and $N_p(\underset{\sim}{\boldsymbol{\mu}}^{(2)}, \boldsymbol{\Sigma})$ with common variance-covariance matrix $\boldsymbol{\Sigma}$. Now, our statistical problem is whether the two normal populations have the same mean vector or not. In other words, our problem is equivalent to test the hypothesis

$$H_0 : \underset{\sim}{\boldsymbol{\mu}}^{(1)} = \underset{\sim}{\boldsymbol{\mu}}^{(2)} \quad \text{vs} \quad H_1 : \underset{\sim}{\boldsymbol{\mu}}^{(1)} \neq \underset{\sim}{\boldsymbol{\mu}}^{(2)}$$

based on the given two samples.

For developing the test statistics in the above two problems, we have to consider whether the common covariance matrix $\boldsymbol{\Sigma}$ is known or not. First, let us develop the test statistics for the above one-sample case as well as two-sample case assuming the population variance-covariance matrix $\boldsymbol{\Sigma}$ is known.

6.3 TEST STATISTICS WHEN Σ IS KNOWN :

One-Sample problem:

Before discussing one-sample problem, **let us prove the following important result.**

Result 1: If a *p*-component vector $\underset{\sim}{Y} \sim N_p(\underset{\sim}{0}, \Sigma)$, where Σ is non- singular (positive definite), then

$$\underset{\sim}{Y}'\Sigma^{-1}\underset{\sim}{Y} \sim \chi^2_p \qquad \rightarrow (1)$$

where χ^2_p is Chi-square distribution with *p* d.f .

Solution: We have given $\underset{\sim}{\mathbf{Y}} \sim N_p(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma})$.

Since, Σ is p.d.f $\exists$ a non-singular matrix **C** such that,

$$\mathbf{C}\,\boldsymbol{\Sigma}\,\mathbf{C}' = \mathbf{I}$$

$$\Rightarrow \Sigma = \mathbf{C}^{-1}\,\mathbf{I}\,(\mathbf{C}')^{-1} = (\mathbf{C}'\,\mathbf{C})^{-1} \qquad \rightarrow (2)$$

Let us define the linear transformation,

$$\underset{\sim}{\mathbf{Z}} = \mathbf{C}\underset{\sim}{\mathbf{Y}} \qquad \rightarrow (3)$$

Then, $E(\underset{\sim}{Z}) = C\,E(\underset{\sim}{Y}) = \underset{\sim}{0}$

$$V(\underset{\sim}{\mathbf{Z}}) = V(\mathbf{C}\underset{\sim}{\mathbf{Y}}) = C\,V(\underset{\sim}{\mathbf{Y}})\,C' = \mathbf{C}\,\boldsymbol{\Sigma}\,\mathbf{C}' = I \quad (\text{from (2)})$$

Since the transformation is linear,

$\underset{\sim}{Z} \sim N_p(\underset{\sim}{0}, I)$ i.e ., $Z_1, Z_2, ..., Z_p$, the individual components of $\underset{\sim}{Z}$ are distributed as N(0,1).

Further, since the covariances are zeros, $Z_1, Z_2, ..., Z_p$ are independent which follows from the normality of the components.

$$\therefore \quad \underset{\sim}{Z}'\underset{\sim}{Z} = Z_1^2 + Z_2^2 + \ldots\ldots\ldots\ldots + Z_p^2 \sim \chi^2_p$$

$$\Rightarrow \underset{\sim}{Y}'C'C\underset{\sim}{Y} \sim \chi^2_p \qquad (\text{from (3)})$$

$$\Rightarrow \underset{\sim}{Y}'\Sigma^{-1}\underset{\sim}{Y} \sim \chi^2_p \qquad (\text{from (2)})$$

Hence the result (1) .

Result 2:

If $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is a random sample of size *n* drawn from a multivariate normal population with known variance-covariance matrix Σ, then obtain the test statistic for testing

$$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0 \quad \text{vs} \quad H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$$

and the critical region of size 'α' as well as the confidence region for $\underset{\sim}{\mu}$ of confidence 1- α.

Solution: We have given the random sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ from $N_p(\underset{\sim}{\mu}, \Sigma)$, where Σ is known .Now, we know that, the sample mean, $\bar{\underset{\sim}{X}} \sim N_p(\underset{\sim}{\mu}, \Sigma/n)$.

Define the random vector, $\underset{\sim}{Y} = \sqrt{n}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})$ $\rightarrow (4)$

With $E(\underset{\sim}{Y}) = \sqrt{n}\ E(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) = \sqrt{n}\ (\underset{\sim}{\mu} - \underset{\sim}{\mu}) = \underset{\sim}{0}$.

$$V(\underset{\sim}{Y}) = V(\sqrt{n}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})) = n\ V(\bar{\underset{\sim}{X}}) = n\ \Sigma/n = \Sigma .$$

Thus, the mean vector of $\underset{\sim}{Y}$ is $\underset{\sim}{0}$ and covariance matrix is Σ.

Further, since the transformation in (4) is linear,

$$\underset{\sim}{Y} \sim N_p(\underset{\sim}{0}, \Sigma)$$

Now, from the above **Result 1**, it immediately follows

$$\underset{\sim}{Y}' \Sigma^{-1} \underset{\sim}{Y} \sim \chi^2_p$$

$$\Rightarrow \quad \sqrt{n}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})' \ \Sigma^{-1}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})\ \sqrt{n} \sim \chi^2_p \qquad \text{(from (4))}$$

$$\Rightarrow \quad n\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu})' \ \Sigma^{-1}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) \sim \chi^2_p$$

Thus, the test statistic for $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ is given by

$$n\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0)' \ \Sigma^{-1}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0) \qquad \rightarrow (5)$$

which follows χ^2 distribution with *p* d.f.

Let $\chi^2_p(\alpha)$ be the number such that $\Pr\{\chi^2_p \geq \chi^2_p(\alpha)\} = \alpha$.

Thus, $$\Pr\{\, n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0)' \ \Sigma^{-1}\ (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0) \geq \chi^2_p(\alpha)\ \} = \alpha$$

and to test $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ (given), we use

$$n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0)' \, \Sigma^{-1} (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0) \geq \chi^2_p(\alpha) \qquad \rightarrow (6)$$

as critical region.

Similarly, we use the inequality,

$$(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}^*)' \, \Sigma^{-1} (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}^*) \leq \chi^2_p(\alpha) \qquad \rightarrow (7)$$

for obtaining the confidence region for $\underset{\sim}{\mu}$ (the set of all $\underset{\sim}{\mu}^*$ satisfying (7)) with confidence $1 - \alpha$.

Hence the result.

Two Sample Problem :-

Result 3:

Suppose we have a sample $\underset{\sim}{X}^{(1)}_1, \underset{\sim}{X}^{(1)}_2, ..., \underset{\sim}{X}^{(1)}_{n_1}$ from $N_p(\underset{\sim}{\mu}^{(1)}, \Sigma)$ and another sample $\underset{\sim}{X}^{(2)}_1, \underset{\sim}{X}^{(2)}_2, ..., \underset{\sim}{X}^{(2)}_{n_2}$ from $N_p(\underset{\sim}{\mu}^{(2)}, \Sigma)$, where Σ is known. Now, under the null hypothesis

$$H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$$

$$\frac{n_1 n_2}{n_1 + n_2} \left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \right)' \Sigma^{-1} \left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \right) \right] \sim \chi^2_p,$$

where, $\bar{\underset{\sim}{X}}^{(1)}$ = mean of the random sample $\underset{\sim}{X}^{(1)}_1, \underset{\sim}{X}^{(1)}_2, ..., \underset{\sim}{X}^{(1)}_{n_1}$

and $\bar{\underset{\sim}{X}}^{(2)}$ = mean of the random sample $\underset{\sim}{X}^{(2)}_1, \underset{\sim}{X}^{(2)}_2, ..., \underset{\sim}{X}^{(2)}_{n_2}$.

Solution:

From the given hypothesis, we have

$$\bar{\underset{\sim}{X}}^{(1)} \sim N_p\left(\underset{\sim}{\mu}^{(1)}, \Sigma/n_1\right) \;\&\; \bar{\underset{\sim}{X}}^{(2)} \sim N_p\left(\underset{\sim}{\mu}^{(2)}, \Sigma/n_2\right) \qquad \rightarrow (1)$$

Now define, $\underset{\sim}{Y} = \bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)} \qquad \rightarrow (2)$

with mean vector, $E(\underset{\sim}{Y}) = \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} \qquad \rightarrow (3)$

and variance- covariance matrix,

$$V(\underset{\sim}{Y}) = V(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)})$$

$$= V(\bar{\underset{\sim}{X}}^{(1)}) + V(\bar{\underset{\sim}{X}}^{(2)}) - Cov(\bar{\underset{\sim}{X}}^{(1)}, \bar{\underset{\sim}{X}}^{(2)}) - Cov(\bar{\underset{\sim}{X}}^{(2)}, \bar{\underset{\sim}{X}}^{(1)})$$

$$= \frac{\Sigma}{n_1} + \frac{\Sigma}{n_2} = \Sigma\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \qquad \rightarrow (4)$$

(since the two samples are independent and as a consequence the covariance

matrices $Cov(\underset{\sim}{\bar{\mathbf{X}}}^{(1)}, \underset{\sim}{\bar{\mathbf{X}}}^{(2)}) = \mathbf{0}$ & $Cov(\underset{\sim}{\bar{\mathbf{X}}}^{(2)}, \underset{\sim}{\bar{\mathbf{X}}}^{(1)}) = \mathbf{0}$

Since the transformation used in (2) is linear, we have

$$\underset{\sim}{\mathbf{Y}} \sim N_p\left(\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)}, \boldsymbol{\Sigma}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$$

$$\Rightarrow \quad \underset{\sim}{\mathbf{Y}} - (\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)}) \sim N_p\left(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right)$$

Now, from the above theorem (1), it immediately follows

$$\left[\underset{\sim}{\mathbf{Y}} - (\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)})\right]' \left[\boldsymbol{\Sigma}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]^{-1} \left[\underset{\sim}{\mathbf{Y}} - (\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)})\right] \sim \chi^2_p$$

But, under the null hypothesis, $H_0 : \underset{\sim}{\boldsymbol{\mu}}^{(1)} = \underset{\sim}{\boldsymbol{\mu}}^{(2)}$

$$\left(\underset{\sim}{\bar{\mathbf{X}}}^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)}\right)' \left[\Sigma\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\right]^{-1} \left(\underset{\sim}{\bar{\mathbf{X}}}^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)}\right) \sim \chi^2_p \qquad \text{(from(2))}$$

$$\Rightarrow \frac{n_1 n_2}{n_1 + n_2}\left(\underset{\sim}{\bar{\mathbf{X}}}^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)}\right)' \Sigma^{-1} \left(\underset{\sim}{\bar{\mathbf{X}}}^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)}\right) \sim \chi^2_p$$

Hence the proof.

6.4 HOTELLING'S T^2 STATISTIC:

The Hotelling's T^2 was develop by Harold Hotelling (1895−1973) to extend the univariate t-test with one dependent variable to a multivariate t-test with two or more dependent variables (Hotelling, 1931).

Hotelling's T^2 test is indeed an extension of the univariate t-test to analyze data with multiple response variables. It is commonly used in multivariate analysis to compare means across groups or to test hypotheses about the mean vector of multivariate data. The power of Hotelling's T^2 tests for one-group and two-group designs can be calculated based on sample sizes, alpha level, effect size, and the variance-covariance structure of the data. Options are provided to specify these parameters and solve for required sample sizes.

6.4.1 ASSUMPTIONS OF HOTELLING'S T^2:

The following assumptions are made when using Hotelling's T^2 to analyze one or two samples of data:

(i) Multivariate Normality: The data should follow a multivariate normal distribution within each group.

(ii) Homogeneity of Covariance Matrices: The covariance matrices of the groups should be equal (homoscedasticity).

(iii) Independence: Observations within and between groups should be independent.

6.4.2 DEFINITION OF HOTELLING'S T^2 - STATISTIC (DISTRIBUTION):

Suppose $\underset{\sim}{\mathbf{y}}$ is a *p*-variate random vector distributed according to $N_p(\underset{\sim}{\mathbf{0}}, \mathbf{\Sigma})$ and let $\mathbf{B} = \sum_{i=1}^{n} \underset{\sim}{\mathbf{z}}_i \underset{\sim}{\mathbf{z}}_i'$ (where each $\underset{\sim}{\mathbf{z}}_i \sim N_p(\underset{\sim}{0}, \Sigma)$ and are independent) is a Wishart random matrix and is distributed as Wishart distribution with n degrees of freedom i.e. $\mathbf{B} \sim \mathbf{W_n(\Sigma)}$. Now, if $\underset{\sim}{\mathbf{y}}$ and $\mathbf{B}$ are independent then the quantity

$$T^2 = \underset{\sim}{\mathbf{y}}' \left(\frac{\mathbf{B}}{n} \right)^{-1} \underset{\sim}{\mathbf{y}}$$

is called as Hotelling's T^2 statistic and it distribution is called as Hotelling's T^2-distribution with n d.f. and is denoted as $T^2 \sim T_n^2$.

Nature Of T^2 - statistic (Distribution) :-

We can write,

$$T^2 = \sqrt{n}\,(\underset{\sim}{\bar{X}} - \mu_0)' \left(\frac{\sum_{i=1}^{n} (\underset{\sim}{X}_i - \underset{\sim}{\bar{X}})(\underset{\sim}{X}_i - \underset{\sim}{\bar{X}}_i)'}{n - 1} \right)^{-1} \sqrt{n}\,(\underset{\sim}{\bar{X}} - \mu_0)$$

which is of the form,

$$\begin{pmatrix} \text{multivariate} \\ \text{normal r.v} \end{pmatrix}' \left(\frac{\text{Wishart random matrix}}{\text{d.f}} \right)^{-1} \begin{pmatrix} \text{multivariate} \\ \text{normal r.v} \end{pmatrix}.$$

Since the multivariate normal random vector and the Wishart random matrix, given in T^2 are independently distributed ($\because \underset{\sim}{\bar{X}}$ & S are independently distributed). Their joint distribution is the product of the marginal normal and Wishart distributions and therefore T^2-distribution can be obtained from this.

6.5 DERIVING HOTELLING'S T^2 STATISTIC AS THE LIKELIHOOD RATIO TEST OF $H_0: \underset{\sim}{\boldsymbol{\mu}} = \underset{\sim}{\boldsymbol{\mu}}_0$:

There is a general principle for constructing test procedures called the Likelihood Ratio (LR) principle method and the T^2-statistic can be derived as the LR test of $H_0: \underset{\sim}{\boldsymbol{\mu}} = \underset{\sim}{\boldsymbol{\mu}}_0$ as explained below.

Suppose $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, ..., \underset{\sim}{\mathbf{X}}_n$ (n > p) is given random sample from $N_p(\underset{\sim}{\mu}, \Sigma)$, the likelihood function is

$$L(\underset{\sim}{\mu},\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu})'\Sigma^{-1}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu})} \rightarrow (1)$$

Under the hypothesis, $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$, the likelihood becomes,

$$L(\underset{\sim}{\mu}_0,\Sigma) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu}_0)'\Sigma^{-1}(\underset{\sim}{x}_\alpha-\underset{\sim}{\mu}_0)} \rightarrow (2)$$

The likelihood ratio criterion is

$$\lambda = \frac{\max_{\Sigma} L(\underset{\sim}{\mu}_0,\Sigma)}{\max_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma)} \rightarrow (3)$$

i.e., the numerator is the maximum of the likelihood function for $\underset{\sim}{\mu}, \Sigma$ is the parameter space restricted by the null hypothesis ($\underset{\sim}{\mu} = \underset{\sim}{\mu}_0$) and Σ is positive definite and the denominator is the maximum over the entire parameter space (Σ is positive definite). When the parameters are unrestricted the MLE's of $\underset{\sim}{\mu}$ *and* Σ from (1) are given by

$$\hat{\underset{\sim}{\mu}}_\Omega = \bar{\underset{\sim}{x}}$$

$$\hat{\Sigma}_\Omega = \frac{1}{n}\sum_\alpha (\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})' \rightarrow (4)$$

When $\underset{\sim}{\mu} = \underset{\sim}{\mu}_0$, the likelihood function given by (2), minimizes at

$$\hat{\Sigma}_0 = \frac{1}{n}\sum_\alpha (\underset{\sim}{x}_\alpha - \underset{\sim}{\mu}_0)(\underset{\sim}{x}_\alpha - \underset{\sim}{\mu}_0)' \rightarrow (5)$$

Substituting (4) in (1), we get (after simplification),

$$\max_{\underset{\sim}{\mu},\Sigma} L(\underset{\sim}{\mu},\Sigma) = \frac{1}{(2\pi)^{np/2}\left|\hat{\Sigma}_\Omega\right|^{n/2}} e^{-np/2} \rightarrow (6)$$

Similarly, substituting (5) in (2), we get

$$\max_{\Sigma} L(\underset{\sim}{\mu}_0,\Sigma) = \frac{1}{(2\pi)^{np/2}\left|\hat{\Sigma}_0\right|^{n/2}} e^{-np/2} \rightarrow (7)$$

Substituting (6) & (7) in (3),we get,

$$\lambda = \left(\frac{\left|\hat{\Sigma}_{\Omega}\right|}{\left|\hat{\Sigma}_{0}\right|} \right)^{n/2}$$

$$\Rightarrow \lambda^{2/n} = \frac{\left|\hat{\Sigma}_{\Omega}\right|}{\left|\hat{\Sigma}_{0}\right|} = \frac{\left|\sum_{\alpha=1}^{n} (\underset{\sim}{x}_{\alpha} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{\alpha} - \bar{\underset{\sim}{x}})'\right|}{\left|\sum_{\alpha=1}^{n} (\underset{\sim}{x}_{\alpha} - \underset{\sim}{\mu}_{0})(\underset{\sim}{x}_{\alpha} - \underset{\sim}{\mu}_{0})'\right|} \rightarrow (7a)$$

$$\Rightarrow \lambda^{2/n} = \frac{|A|}{\left|A + n\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})'\right|}$$

Where $A = \sum_{\alpha=1}^{n} (\underset{\sim}{x}_{\alpha} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{\alpha} - \bar{\underset{\sim}{x}})' \rightarrow (8)$

Consider the matrix, $\underset{(p+1)\times(p+1)}{B} = \begin{bmatrix} A & \sqrt{n}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0}) \\ \sqrt{n}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})' & -1 \end{bmatrix}$

$$= \begin{bmatrix} B_{11} & \vdots & B_{12} \\ \cdots & \vdots & \cdots \\ B_{21} & \vdots & B_{22} \end{bmatrix}.$$

We have, $|B| = |B_{11}|\,|B_{22} - B_{21}B_{11}^{-1}B_{12}|$

$$= |B_{22}|\,|B_{11} - B_{12}B_{22}^{-1}B_{21}|$$

$$\therefore |B| = |-1|\,\left|A - \sqrt{n}\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})\,(-1)^{-1}\sqrt{n}\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})'\right|$$

$$= \left|A + n\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})'\right|.$$

$$\therefore \lambda^{2/n} = \frac{|A|}{-\begin{vmatrix} A & \sqrt{n}\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0}) \\ \sqrt{n}\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})' & -1 \end{vmatrix}}$$

$$= \frac{|A|}{-|A|\,\left|-1 - \sqrt{n}\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})'\,A^{-1}\sqrt{n}\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_{0})\right|}$$

$$= \frac{|\mathbf{A}|}{|\mathbf{A}|\,|1 + n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})' \mathbf{A}^{-1} (\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})|}$$

$$= \frac{1}{1 + n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})' \mathbf{A}^{-1} (\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})}.$$

Where $\mathbf{A}$ is as given in (8) . But we have,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A} = \frac{1}{n-1}\sum_{\alpha}(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})'$$

$$\Rightarrow \quad \mathbf{A} = (n-1)\,\mathbf{S}$$

$$\therefore \ \lambda^{2/n} = \frac{1}{1+n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})'[(n-1)S]^{-1}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})}$$

$$= \frac{1}{1+\frac{n}{(n-1)}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})' S^{-1}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})} = \frac{1}{1+T^2/(n-1)} \qquad \rightarrow (9)$$

where, $T^2 = n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})' S^{-1}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})$ is Hotelling's T^2 - statistic.

Now, from (7a) & (9), we can see

$$1 + \frac{T^2}{(n-1)} = \frac{\left|\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha - \underset{\sim}{\mu_0})(\underset{\sim}{x}_\alpha - \underset{\sim}{\mu_0})'\right|}{\left|\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})'\right|}$$

$$\Rightarrow \quad T^2 = (n-1)\left[\frac{\left|\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha - \underset{\sim}{\mu_0})(\underset{\sim}{x}_\alpha - \underset{\sim}{\mu_0})'\right|}{\left|\sum_{\alpha=1}^{n}(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_\alpha - \bar{\underset{\sim}{x}})'\right|} - 1\right]$$

$$= (n-1)\left[\frac{|\hat{\Sigma}_0|}{|\hat{\Sigma}_\Omega|} - 1\right] \qquad \rightarrow (10)$$

In this formula, we need not find the inverse of a matrix, where as in the original formula we have to evaluate S^{-1}.

6.6 INVARIANCE PROPERTY OF HOTELLING'S T^2 STATISTIC:

Result : T^2 - statistic is invariant (unchanged) under changes in the units of measurements for $\underset{\sim}{X}$ of the form,

$$\underset{\sim}{Y} = C\underset{\sim}{X} + \underset{\sim}{d}, \text{ where C is non-singular} \qquad \rightarrow (1)$$

Proof :- We have, $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$,

$$\text{i.e., } E(\underset{\sim}{X}) = \underset{\sim}{\mu} \Rightarrow E(\underset{\sim}{Y}) = C\underset{\sim}{\mu} + \underset{\sim}{d} \quad (\because \text{from (1)}) \qquad \rightarrow (2)$$

Now, we have the T^2 - statistic for testing, $H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ vs $H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu}_0$ based on the given sample $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is

$$T_x^2 = n\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_0)'\, S_x^{-1}\, (\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_0) \qquad \rightarrow (3)$$

where $$S_x = \frac{1}{n-1} \sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})' \qquad \rightarrow (3a)$$

From (1) we can see $\underset{\sim}{Y} \sim N_p(C\underset{\sim}{\mu} + d, C\Sigma C')$.

Now, the T^2 - statistic for testing,

$$H_0 : \underset{\sim}{\mu}_Y = \underset{\sim}{\mu}_{Y_0} \text{ vs } H_1 : \underset{\sim}{\mu}_Y \neq \underset{\sim}{\mu}_{Y_0}$$

where, $\underset{\sim}{\mu}_Y = C\underset{\sim}{\mu} + \underset{\sim}{d}$ & $\underset{\sim}{\mu}_{Y_0} = C\underset{\sim}{\mu}_0 + \underset{\sim}{d}$ $\qquad \rightarrow (4)$

based on the sample $\underset{\sim}{Y}_1, \underset{\sim}{Y}_2, ..., \underset{\sim}{Y}_n$ is given by

$$T_y^2 = n\,(\bar{\underset{\sim}{Y}} - \underset{\sim}{\mu}_{Y_0})'\, S_Y^{-1} (\underset{\sim}{Y} - \underset{\sim}{\mu}_{Y_0}) \qquad \rightarrow (5)$$

where, $\bar{\underset{\sim}{Y}} = C\,\bar{\underset{\sim}{x}} + \underset{\sim}{d}$ (from (1))

$$S_y = \frac{1}{n-1} \sum_{i=1}^{n} (\underset{\sim}{y}_i - \bar{\underset{\sim}{y}})(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}})' \qquad \rightarrow (6)$$

and $\underset{\sim}{\mu}_{Y_0}$ is given by (4).

In order to show that the Hotelling's T^2 is invariant under the changes in the units of measurements, we have to show, $T_y^2 = T_x^2$.

For that, consider T_y^2 given from (5),

$$T_y^2 = n\,(\bar{\underset{\sim}{y}} - \underset{\sim}{\mu}_{y_0})'\, S_y^{-1}\, (\bar{\underset{\sim}{y}} - \underset{\sim}{\mu}_{y_0})$$

$$= n\,(C\bar{\underset{\sim}{x}} - C\underset{\sim}{\mu}_0)'\, S_y^{-1}\, (C\bar{\underset{\sim}{x}} - C\underset{\sim}{\mu}_0) \qquad \text{(using (4))}$$

$$= n\,(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_0)'\, C'\, S_y^{-1}\, C(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu}_0) \qquad \rightarrow (7)$$

But, $\mathbf{S_y} = \frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}})(\underset{\sim}{y}_i - \bar{\underset{\sim}{y}})'$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(C\underset{\sim}{x}_i - C\bar{\underset{\sim}{x}})(C\underset{\sim}{x}_i - C\bar{\underset{\sim}{x}})' \qquad \text{(using (1))}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}C(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'C'$$

$$= CS_xC' \qquad \text{(from (3a))}$$

$$\Rightarrow \mathbf{S_y^{-1}} = (C')^{-1}S_x^{-1}C^{-1}$$

$$\Rightarrow C'S_y^{-1}C = S_x^{-1}$$

Using this in (7), we get

$$T_y^2 = n(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0})'S_x^{-1}(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu_0}) = T_x^2 \qquad \text{(from (3))}$$

Thus, T^2 is invariant under the changes in the units of measurements.

NOTE : The above theorem may be stated as " The Hotellings T^2 is invariant under linear transformation (or under changes in the location and scale) of the sample .

6.7 APPLICATIONS OF HOTELLING'S T^2 STATISTIC:

6.7.1 For testing the significance of one sample mean vector $\bar{\underset{\sim}{x}}$:

Suppose $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ is a random sample from a *p*-variate normal population $N_p(\underset{\sim}{\mu}, \Sigma)$, where both $\underset{\sim}{\mu}$ and Σ are assumed as unknown. Now, our statistical problem is whether the given sample has come from the multivariate normal population, whose mean vector is $\underset{\sim}{\mu_0}$. In other words, we want to test the hypothesis

$$H_0 : \underset{\sim}{\mu} = \underset{\sim}{\mu_0} \text{ vs } H_1 : \underset{\sim}{\mu} \neq \underset{\sim}{\mu_0} \qquad \rightarrow (1),$$

where $\underset{\sim}{\mu_0}$ is the given mean vector.

For testing the above hypothesis, derive the test statistic.

Solution:

We have given a random sample of size n viz., $\underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_n$ from $N_p(\underset{\sim}{\mu}, \Sigma)$, where both $\underset{\sim}{\mu}$ and Σ are unknown.

Now, we know that the mean vector

$$\bar{\underset{\sim}{X}} \sim N_p(\underset{\sim}{\mu}, \Sigma/n) \text{ (since } \bar{\underset{\sim}{X}} \text{ is a linear function of the sample)}$$

Define the random vector, $\underset{\sim}{Y} = \sqrt{n}(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}) \qquad \rightarrow (2)$

Whose population mean vector and population variance-covariance matrix are

respectively given by

$$E(\underset{\sim}{\mathbf{Y}})=\sqrt{n}\,E(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}})=\sqrt{n}(\underset{\sim}{\boldsymbol{\mu}}-\underset{\sim}{\boldsymbol{\mu}})=\underset{\sim}{\mathbf{0}}.$$

$$V(\underset{\sim}{\mathbf{Y}})=V(\sqrt{n}\,(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}}))=n\,V(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}})=n\,V(\bar{\underset{\sim}{\mathbf{X}}})=n\,\Sigma/n=\Sigma.$$

Thus, the mean vector of $\underset{\sim}{Y}$ is $\underset{\sim}{0}$ and covariance matrix is Σ.

Further, since the transformation in (2) is linear, we have

$$\underset{\sim}{\mathbf{Y}} \sim N_p(\underset{\sim}{0},\Sigma) \qquad \rightarrow (3)$$

We have the sample variance-covariance matrix

$$\mathbf{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{\mathbf{X}}_i-\bar{\underset{\sim}{\mathbf{X}}})(\underset{\sim}{\mathbf{X}}_i-\bar{\underset{\sim}{\mathbf{X}}})' \qquad \rightarrow (4)$$

Now, we know that (n-1)**S** follows Wishart distribution with n-1 degrees freedom and parameter $\boldsymbol{\Sigma}$ that is

$$(n-1)\mathbf{S} \sim W_{n-1}(\boldsymbol{\Sigma}) \qquad \rightarrow (5)$$

Further, we know that the sample mean vector $\bar{\underset{\sim}{X}}$ and the sample variance-covariance matrix **S** are independently distributed.

From (2), it immediately follows that the random vector $\underset{\sim}{\mathbf{Y}}$ and the random matrix(n-1)**S** distribute independently.

Now, by the definition of Hotelling's T^2 distribution, the statistic

$$T^2=\underset{\sim}{\mathbf{Y}}'\left(\frac{(n-1)S}{(n-1)}\right)^{-1}\underset{\sim}{\mathbf{Y}} \qquad \rightarrow (6)$$

follows Hotelling's T^2 distribution with n-1d.f. i.e.

$$T^2 \sim T^2_{n-1}$$

Substituting (2) in (6), we can see that

$$T^2=n(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}})'\,\mathbf{S}^{-1}(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}}) \sim T^2_{n-1} \qquad \rightarrow (7)$$

Now, under $H_0: \underset{\sim}{\boldsymbol{\mu}}=\underset{\sim}{\boldsymbol{\mu}}_0$ (7) becomes

$$T^2=n(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}}_0)'\,\mathbf{S}^{-1}(\bar{\underset{\sim}{\mathbf{X}}}-\underset{\sim}{\boldsymbol{\mu}}_0) \sim T^2_{n-1} \qquad \rightarrow (8)$$

where, $\bar{\underset{\sim}{\mathbf{x}}}=\frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{\mathbf{x}}_i$ and $\mathbf{S}=\frac{1}{n-1}\sum_{i=1}^{n}(\underset{\sim}{\mathbf{X}}_i-\bar{\underset{\sim}{\mathbf{X}}})(\underset{\sim}{\mathbf{X}}_i-\bar{\underset{\sim}{\mathbf{X}}})'$

Thus, the formula (8) gives us the Hotelling's T^2 statistic which can be used to test (1) and follows T^2_{n-1}

At the given α level of significance, H_0 may be rejected infavour of H_1 if

$$\frac{T^2}{n-1}\left(\frac{n-p}{p}\right) > F_{p,n-p}(\alpha) \quad \text{(or)} \quad T^2 > T_0^2 . \qquad \rightarrow (5)$$

Where, $T_0^2=\frac{(n-1)\,p}{n-p}\,F_{p,n-p}(\alpha)$ and $F_{p,n-p}(\alpha)$ is the upper $100\,\alpha^{th}$ percentile

of the F-distribution and can be obtained from the F-tables.

6.7.2 A Two Sample Problem when the covariance matrices are equal but unknown (A Use Of T^2-statistic):

Another situation in which the T^2-statistic is used is that in which the null hypothesis is that the mean of one normal population is equal to the mean of the other, where the covariance matrices are assumed equal but unknown.

Suppose $\underset{\sim}{X}_1^{(1)}, \underset{\sim}{X}_2^{(1)}, ..., \underset{\sim}{X}_n^{(1)}$ is a sample from $N_p(\underset{\sim}{\mu}^{(1)}, \Sigma)$ and $\underset{\sim}{X}_1^{(2)}, \underset{\sim}{X}_2^{(2)}, ..., \underset{\sim}{X}_n^{(2)}$ is a another sample (independent of the first sample) from $N_p(\underset{\sim}{\mu}^{(2)}, \Sigma)$.

Now, we wish to test the null hypothesis,

$$H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)} \text{ or } \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} = 0 \text{ ,against } H_1 : \underset{\sim}{\mu}^{(1)} \neq \underset{\sim}{\mu}^{(2)} \quad \rightarrow (1)$$

The sample means from the hypothesis ,

$$\underset{\sim}{\bar{X}}^{(1)} = \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{X}_i^{(1)} \sim N_p\left(\underset{\sim}{\mu}^{(1)}, \Sigma / n_1\right)$$

and $$\underset{\sim}{\bar{X}}^{(2)} = \frac{1}{n_2}\sum_{i=1}^{n_2} \underset{\sim}{X}_i^{(2)} \sim N_p\left(\underset{\sim}{\mu}^{(2)}, \Sigma / n_2\right)$$

Now define, $\underset{\sim}{Y} = \underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)} \quad \rightarrow (2)$

with mean, $E(\underset{\sim}{Y}) = E(\underset{\sim}{\bar{X}}^{(1)}) - E(\underset{\sim}{\bar{X}}^{(2)}) = \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} \quad \rightarrow (3)$

and the variance- covariance matrix,

$V(\underset{\sim}{Y}) = V(\underset{\sim}{\bar{X}}^{(1)}) + V(\underset{\sim}{\bar{X}}^{(2)})$ ($\because$ The two samples are independent)

$$= \frac{1}{n_1}\Sigma + \frac{1}{n_2}\Sigma$$

$$= \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma \quad \rightarrow (4)$$

Since the transformation in (2) is linear, from (3) and (4) it follows

$$\underset{\sim}{\mathbf{Y}} \sim N_p\left(\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)}, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{\Sigma}\right)$$

i.e., $$\underset{\sim}{\mathbf{Y}} - (\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)}) \sim N_p\left(\underset{\sim}{\mathbf{0}}, \left(\frac{1}{n_1}+\frac{1}{n_2}\right)\boldsymbol{\Sigma}\right)$$

i.e., $$\frac{\left(\underset{\sim}{\bar{\mathbf{X}}}^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)}\right) - \left(\underset{\sim}{\boldsymbol{\mu}}^{(1)} - \underset{\sim}{\boldsymbol{\mu}}^{(2)}\right)}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim N_p(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}) \quad \text{(using (2))} \quad \rightarrow (5)$$

The sample covariance matrix from sample 1, which is denoted by $\mathbf{S}_1$ and is given by

$$\mathbf{S}_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(\underset{\sim}{\mathbf{X}}_i^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(1)})(\underset{\sim}{\mathbf{X}}_i^{(1)} - \underset{\sim}{\bar{\mathbf{X}}}^{(1)})' .$$

Similarly, the sample covariance matrix from sample 2, denoted by S_2 and is given by

$$\mathbf{S}_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(\underset{\sim}{\mathbf{X}}_i^{(2)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)})(\underset{\sim}{\mathbf{X}}_i^{(2)} - \underset{\sim}{\bar{\mathbf{X}}}^{(2)})' .$$

Let us denote, $\mathbf{S} = \frac{(n_1-1)\mathbf{S}_1 + (n_2-1)\mathbf{S}_2}{n_1 + n_2 - 2} \quad \rightarrow (6)$

We know that $(n_1-1)\mathbf{S}_1$ and $(n_2-1)\mathbf{S}_2$ are Wishart random matrices and are distributed as $W_{n_1-1}(\boldsymbol{\Sigma})$ & $W_{n_2-1}(\boldsymbol{\Sigma})$ respectively, where $W_{n_1-1}(\boldsymbol{\Sigma})$ is Wishart distribution with $(n_1 - 1)$ d.f and $W_{n_2-1}(\boldsymbol{\Sigma})$ is Wishart distribution with $(n_2 - 1)$ d.f. both have the parametric matrix Σ.

By assumption, the samples are independent, so $(n_1-1)\,S_1$ and $(n_2-1)\,S_2$ are also independent . Therefore from (6), $(n_1+n_2-2)S$ is distributed as Wishart distribution with $n_1 + n_2 - 2$ d.f and with the parametric matrix Σ,

i.e. $(n_1+n_2-2)S \sim W_{n_1+n_2-2}(\Sigma) \quad \rightarrow (7)$

Since, the sample variance-covariance matrix is independently distributed with the sample mean vector, S_1 is independently distributed with $\underset{\sim}{\bar{\mathbf{X}}}^{(1)}$ and since the two

samples are independent, S_1 is independently distributed with $\bar{\underset{\sim}{X}}^{(2)}$ and therefore S_1 is independently distributed with $\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}$.

Similarly, S_2 is independently distributed with $\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}$.

Therefore, $S = \dfrac{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2}{n_1 + n_2 - 2}$ is independently distributed with $\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}$.

Thus, from the above explanation and from (5) & (6) and by the definition of T^2 -distribution, we have

$$T^2 = \left[\frac{\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right]' S^{-1} \left[\frac{\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right]$$

$$= \left(\frac{n_1 n_2}{n_1 + n_2}\right)\left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right]' S^{-1}\left[\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right] \rightarrow (8)$$

is distributed as T^2-distribution with $n_1 + n_2 - 2$ d.f .

Now, by virtue of the relation between T^2 and F – distribution, we have

$$\frac{T^2}{n_1 + n_2 - 2} \sim \frac{p}{n_1 + n_2 - 2 - (p-1)} F_{p,\, n_1 + n_2 - 2 - (p-1)}$$

i.e., $$\frac{T^2}{n_1 + n_2 - 2} \sim \frac{p}{n_1 + n_2 - p - 1} F_{p,\, n_1 + n_2 - p - 1}$$

under $H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$ i.e., $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)} = \underset{\sim}{0}$, (8) becomes

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) S^{-1}\left(\bar{\underset{\sim}{X}}^{(1)} - \bar{\underset{\sim}{X}}^{(2)}\right) \qquad \rightarrow (9)$$

if $T^2 > \dfrac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p - 1} F_{p,\, n_1 + n_2 - p - 1}(\alpha)$, where $F_{p,\, n_1 + n_2 - p - 1}(\alpha)$ is table F-

value at α level of significance with (p, $n_1 + n_2 - p - 1$) d.f., then $H_0 : \underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$ may be rejected .

6.7.3 The Two Sample Problem when the Covariance matrices are unequal :

In the above problem, we have assumed that the covariance matrices of both the populations are assumed as equal i.e., $\Sigma_1 = \Sigma_2 = \Sigma$.

Now, let us suppose that $\Sigma_1 \neq \Sigma_2$ i.e., the population covariance matrices are not equal .

In this case, no tests are available for making inferences about $\underset{\sim}{\mu}^{(1)} = \underset{\sim}{\mu}^{(2)}$,when the sizes of the samples are small . However, if $n_1 \& n_2$ are large i.e., in case of large samples, we have the following result.

Result : Let the sample sizes be such that n_1-p and n_2-p are large . An approximation $100(1-\alpha)\%$ confidence region for $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ is given by all $\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$ satisfying ,

$$\left[\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right]' \left[\frac{1}{n_1} S_1 + \frac{1}{n_2} S_2\right]^{-1} \left[\left(\underset{\sim}{\bar{x}}^{(1)} - \underset{\sim}{\bar{x}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right] \leq \chi^2_p(\alpha)$$

where, $\chi^2_p(\alpha)$ is χ^2 - table values with p.d.f at $100\alpha\%$ level of significance.

Proof:- $E(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}) = \underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}$

$$\& \; V(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}) = V(\underset{\sim}{\bar{X}}^{(1)}) + V(\underset{\sim}{\bar{X}}^{(2)}) = \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2 .$$

By the central limit theorem,

$$\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)} \sim N_p\left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}, \frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right) .$$

If $\Sigma_1 \& \Sigma_2$ are known,

$$\left[\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right]' \left[\frac{1}{n_1}\Sigma_1 + \frac{1}{n_2}\Sigma_2\right]^{-1}$$

$$\left[\left(\underset{\sim}{\bar{X}}^{(1)} - \underset{\sim}{\bar{X}}^{(2)}\right) - \left(\underset{\sim}{\mu}^{(1)} - \underset{\sim}{\mu}^{(2)}\right)\right] \sim \chi^2_p(\alpha),$$

approximately, when $n_1 \& n_2$ are large, with high probability $S_1 \to \Sigma_1$ and $S_2 \to \Sigma_2$.

Consequently, the approximation holds with $S_1 \& S_2$, in place of Σ_1 and Σ_2 respectively. Hence the theorem.

6.8 DISTRIBUTION OF HOTELLING'S T^2 STATISTIC:

Theorem :- Let $\underset{\sim}{Y} \sim N_p(\underset{\sim}{v}, \Sigma)$ and let **A** be a Wishart random matrix independently distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{Z}_\alpha \underset{\sim}{Z}'_\alpha$, where $\underset{\sim}{Z}_\alpha$'s are i.i.d $\sim N_p(\underset{\sim}{0}, \Sigma)$. Also let

$$T^2 = m\underset{\sim}{Y}'\mathbf{A}^{-1}\underset{\sim}{Y} \qquad \rightarrow (1)$$

then, $\frac{T^2}{m}\left(\frac{m - p + 1}{p}\right)$ is distributed as a non-central ***F*** with

p and *m-p+1* d.f. and non-centrality parameter $\underset{\sim}{v}'\Sigma^{-1}\underset{\sim}{v}$. Further, if $\underset{\sim}{v} = \underset{\sim}{0}$, then

$$\frac{T^2}{m}\left(\frac{m - p + 1}{p}\right) \sim F_{p, m-p+1} \qquad \rightarrow (2)$$

and the distribution of T^2 is called T^2-distribution.

Proof :- Since Σ is positive definite, there exists a non-singular **C** such that

$$\mathbf{C}\Sigma\mathbf{C}' = \mathbf{I_p} \text{ so that, } \Sigma = (\mathbf{C}'\mathbf{C})^{-1} \qquad \rightarrow (3)$$

Define, $\underset{\sim}{Y}^* = \mathbf{C}\underset{\sim}{Y}$ and $\mathbf{A}^* = \mathbf{CAC}'$ $\qquad \rightarrow (4)$

We can see that, $\mathbf{E}(\underset{\sim}{Y}^*) = \mathbf{C}\underset{\sim}{v} = \underset{\sim}{v}^*$ (say)

$$V(\underset{\sim}{Y}^*) = V(\mathbf{C}\underset{\sim}{Y}) = \mathbf{C}\Sigma\mathbf{C}' = \mathbf{I} \qquad \text{(using (3))} \qquad \rightarrow (5)$$

Thus, $\underset{\sim}{Y}^* \sim N_p(\underset{\sim}{v}^*, \mathbf{I})$.

Since **A** is distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{Z}_\alpha \underset{\sim}{Z}'_\alpha$, $\mathbf{A}^* = \mathbf{CAC}'$ is distributed as

$$\mathbf{C}\sum_{\alpha=1}^{m} \underset{\sim}{Z}_\alpha \underset{\sim}{Z}'_\alpha \mathbf{C}' = \sum_{\alpha=1}^{m} \underset{\sim}{Z}^*_\alpha \underset{\sim}{Z}^{*\prime}_\alpha \qquad \rightarrow (6)$$

where, $\underset{\sim}{Z}^*_\alpha = \mathbf{C}\underset{\sim}{Z}_\alpha \sim N_p(\mathbf{C}\underset{\sim}{0}, \mathbf{C}\Sigma\mathbf{C}') = N_p(\underset{\sim}{0}, \mathbf{I})$.

Eq (1) can be written as

$$\begin{aligned} T^2 &= m\underset{\sim}{Y}'\mathbf{A}^{-1}\underset{\sim}{Y} \\ &= m\underset{\sim}{Y}'\mathbf{C}'(\mathbf{C}')^{-1}\mathbf{A}^{-1}(\mathbf{C})^{-1}\mathbf{C}\underset{\sim}{Y} \\ &= m(\mathbf{C}\underset{\sim}{Y})'(\mathbf{CAC}')^{-1}(\mathbf{C}\underset{\sim}{Y}) \\ &= m\underset{\sim}{Y}^{*\prime}\mathbf{A}^{*\prime}\underset{\sim}{Y}^* \qquad \rightarrow (7) \end{aligned}$$

where, $\underset{\sim}{\mathbf{Y}}^* \sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$ and $\mathbf{A}^*$ is independently distributed as $\sum_{\alpha=1}^{m} \underset{\sim}{\mathbf{Z}}_\alpha^* \underset{\sim}{\mathbf{Z}}_\alpha^{*\prime}$ in which $\underset{\sim}{\mathbf{Z}}_\alpha^*$'s are i.i.d $\sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$.

Also, since $\underset{\sim}{\mathbf{Y}}$ and $\mathbf{A}$ are independently distributed, from eq. (4) , $\underset{\sim}{\mathbf{Y}}^*$ and $\mathbf{A}^*$ are also independently distributed .

Let $\mathbf{\Omega} = (\omega_{ij})_{p\times p}$ is an orthogonal matrix in which first row is defined by

$$\omega_{1j} = \frac{Y_j^*}{\sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^*}} \quad , \text{ j=1,2,.....,p} \qquad \rightarrow (8)$$

where , Y_j^* is j^{th} component of $\underset{\sim}{\mathbf{Y}}^*$.

Now define , $\underset{\sim}{\mathbf{U}} = \mathbf{\Omega}\underset{\sim}{\mathbf{Y}}^*$

$$\mathbf{B} = \mathbf{\Omega}\mathbf{A}^*\mathbf{\Omega}' \qquad \rightarrow (9)$$

The i^{th} component of $\underset{\sim}{\mathbf{U}}$ is given by

$$U_i = \sum_{j=1}^{p} \omega_{ij} Y_j^*$$

$$= \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^*} \sum_{j=1}^{p} \omega_{ij}\omega_{1j} \qquad \text{[using (8)]}$$

$$= \begin{cases} \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^*} & if \;\; i = 1 \\ 0 & \text{otherwise} \end{cases}$$

(Since , $\mathbf{\Omega}$ is orthogonal matrix)

Thus , $$\underset{\sim}{\mathbf{U}} = \begin{bmatrix} \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^*} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \rightarrow (10)$$

From equation (7) ,

$$\frac{T^2}{m} = \underset{\sim}{\mathbf{Y}}^{*\prime}\mathbf{I}\mathbf{A}^{*\prime}\mathbf{I}\underset{\sim}{\mathbf{Y}}^{*}$$

$$= \underset{\sim}{\mathbf{Y}}^{*\prime}\mathbf{\Omega}'\mathbf{\Omega}\mathbf{A}^{*\prime}\mathbf{\Omega}'\mathbf{\Omega}\underset{\sim}{\mathbf{Y}}^{*} \qquad (\because\ \mathbf{\Omega} \text{ is orthogonal})$$

$$= \left(\mathbf{\Omega}\underset{\sim}{\mathbf{Y}}^{*}\right)'\left(\mathbf{\Omega}\mathbf{A}^{*}\mathbf{\Omega}'\right)^{-1}\left(\mathbf{\Omega}\underset{\sim}{\mathbf{Y}}^{*}\right) \qquad (\because\ \mathbf{\Omega^{-1}} = \mathbf{\Omega}')$$

$$= \underset{\sim}{\mathbf{U}}'\mathbf{B^{-1}}\underset{\sim}{\mathbf{U}} \qquad [\text{ using (9)}]$$

$$= \left(\sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} \quad 0 \quad \cdots \quad 0\right)\begin{bmatrix} b^{11} & b^{12} & \cdots & b^{1p} \\ b^{21} & b^{22} & \cdots & b^{2p} \\ \vdots & \vdots & & \vdots \\ b^{p1} & b^{p2} & \cdots & b^{pp} \end{bmatrix}\begin{bmatrix} \sqrt{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$= b^{11}\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*} \qquad \rightarrow (11)$$

where , b^{11} is first diagonal element of $\mathbf{B^{-1}}$.
But we know that ,

$$b^{11} = \frac{1}{b_{11} - \underset{\sim}{\mathbf{b}}_{\mathbf{1}}'\mathbf{B}_{\mathbf{22}}^{-1}\underset{\sim}{\mathbf{b}}_{\mathbf{1}}} \quad , \quad \text{where} \quad \mathbf{B} = \begin{bmatrix} b_{11} & \underset{\sim}{\mathbf{b}}_{\mathbf{1}}' \\ \underset{\sim}{\mathbf{b}}_{\mathbf{1}} & b_{22} \end{bmatrix}$$

Thus , from eq (11) ,

$$\frac{T^2}{m} = \frac{\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^{*}}{b_{11.2.....p}} \qquad \rightarrow (12)$$

where , $b_{11.2.....p} = b_{11} - \underset{\sim}{\mathbf{b}}_{\mathbf{1}}'\mathbf{B}_{\mathbf{22}}^{-1}\underset{\sim}{\mathbf{b}}_{\mathbf{1}}$.

Let us suppose that $\mathbf{\Omega}$ is fixed (given). Then, just as we show $\mathbf{A}^{*}$ is distributed as $\sum_{\alpha=1}^{m}\underset{\sim}{\mathbf{Z}}_{\alpha}^{*}\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}}^{*\prime}$, we can show that $\mathbf{\Omega}\mathbf{A}^{*}\mathbf{\Omega}'$ is distributed as $\sum_{\alpha=1}^{m}\underset{\sim}{\mathbf{V}}_{\alpha}\underset{\sim}{\mathbf{V}}_{\alpha}'$, when $\underset{\sim}{\mathbf{V}}_{\boldsymbol{\alpha}} = \mathbf{\Omega}\underset{\sim}{\mathbf{Z}}_{\boldsymbol{\alpha}}^{*}$ and $\underset{\sim}{\mathbf{V}}_{\boldsymbol{\alpha}}$'s are $\text{i.i.d} \sim N_p(\underset{\sim}{\mathbf{0}}, \mathbf{I})$.

Now, with little difficult , we may show that $b_{11.2.....p} = b_{11} - \underset{\sim}{\mathbf{b}}_{\mathbf{1}}'\mathbf{B}_{\mathbf{22}}^{-1}\underset{\sim}{\mathbf{b}}_{\mathbf{1}}$ is conditionally distributed as $\sum_{\alpha=1}^{m-(p-1)}\omega_{\alpha}^{2}$,

where $each\ \omega_{\alpha}$ is $i.i.d \sim N(0,1)$.

Therefore, $\sum_{\alpha=1}^{m-(p-1)}\omega_{\alpha}^{2} \sim \chi_{m-(p-1)}^{2}$.

More over, the conditional distribution of $b_{11.2.....p}$ does not dependent on $\boldsymbol{\Omega}$, we have $b_{11.2.....p}$ is unconditionally distributed as $\chi^2_{m-(p-1)}$.

Also, since $\underset{\sim}{\mathbf{Y}}^* \sim N_p(\underset{\sim}{\nu}^*, \mathbf{I})$, $\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^* = \sum_{i=1}^{p} Y_i^{*2}$

where , $Y_i^* \sim N(\nu_i^*, 1)$ and Y_i^*'s are independent .

Thus , $\underset{\sim}{\mathbf{Y}}^{*\prime}\underset{\sim}{\mathbf{Y}}^*$ has non-central χ^2-distribution with non-centrality

$$= \sum_{i=1}^{p} \nu_i^{*2} = \underset{\sim}{\nu}^{*\prime}\underset{\sim}{\nu}^* = \underset{\sim}{\nu}'\mathbf{C}'\mathbf{C}\underset{\sim}{\nu} \qquad \text{[from(5)]}$$

$$= \underset{\sim}{\nu}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\nu} \qquad \text{[from (3)]}$$

Thus , $\frac{T^2}{m}$ is distributed as the ratio of non-central χ_p^2 and an independent central $\chi^2_{m-(p-1)}$.

Thus , $\frac{T^2}{m}\left[\frac{m-(p-1)}{p}\right] \sim F_{(p,m-p+1)}$ (non-central) and non-centrality parameter $\underset{\sim}{\nu}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\nu}$.

If $\underset{\sim}{\nu} = \mathbf{0}$ then, $\underset{\sim}{\nu}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\nu} = \mathbf{0}$ and therefore in this case, the distribution is central . $F_{(p,m-p+1)}$,the distribution of T^2 is called T^2-distribution with 'm' degrees of freedom .

6.9 SUMMARY:

Hotelling's T^2 statistic is a fundamental tool in multivariate statistical analysis**,** developed as a natural extension of Student's *t*-test to situations involving multiple correlated response variables**.** It provides a unified framework for testing hypotheses about population mean vectors while accounting for the covariance structure among variables.

The statistic is constructed using the sample mean vector and sample covariance matrix and, under the assumption of multivariate normality, follows a distribution that can be transformed into an F-distribution for hypothesis testing. Hotelling's T^2 can be applied in both one-sample and two-sample problems, enabling simultaneous comparison of several characteristics across populations.

A key theoretical feature of Hotelling's T^2 is its close relationship with the Mahalanobis distance, showing that it measures the standardized multivariate distance between mean vectors. The statistic also possesses the important invariance property, ensuring that results remain unchanged under linear transformations and changes of measurement units. These properties make Hotelling's T^2 a robust and reliable method for multivariate inference.

Hotelling's T^2 - statistic plays a central role in multivariate analysis by:

- Providing an effective method for simultaneous testing of multiple means
- Preserving the overall significance level**,** unlike multiple univariate tests
- Serving as the theoretical foundation for MANOVA and multivariate control charts
- Being widely applicable in medicine, engineering, economics, psychology, and social sciences.
-

Despite its strengths, the method requires adherence to assumptions such as multivariate normality and adequate sample size relative to the number of variables. When these conditions are satisfied, Hotelling's T^2 offers a powerful, elegant, and statistically sound approach for multivariate hypothesis testing.

In conclusion, Hotelling's T^2 statistic remains an indispensable tool in modern statistical methodology, bridging theory and application in the analysis of multivariate data.

6.10 SELF ASSESSMENT QUESTIONS:

1. Define Hotelling's T^2 statistic. Show that Hotelling's T^2 statistic can be used to test the equality of means of corresponding variables in two MVN populations having the same variance-covariance matrix.
2. Explain in detail the likelihood ratio principle in multivariate testing.
3. Derive the invariance property of Hotelling's T^2 statistic.
4. Discuss the applications of Hotelling's T^2 statistic in fields such as medicine, quality control, economics, and social sciences with suitable examples**.**
5. Explain the construction of confidence regions for the population mean vector using Hotelling's T^2 statistic.
6. Derive the test statistic and explain how it is used to compare two multivariate population mean vectors.
7. Given a multivariate sample from a normal population, apply the one-sample Hotelling's T^2 test to test whether the mean vector equals a specified value.
8. What is meant by the pooled covariance matrix in the two-sample T^2 test? List any four practical applications of Hotelling's T^2 statistic.

6.11 SUGGESTED READINGS:

1. Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern.
2. Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International.
3. Mardia, K.V., Kent, J. T and Bibby, J. M. (1979): Multivariate Analysis. Academic Press, New York.
4. Brenner, D., Bilodeau, M. (1999). Theory of Multivariate Statistics. Germany: Springer.
5. Giri Narayan C. (1995). Multivariate Statistical Analysis.
6. Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
7. Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

Dr. U. Ramkiran

LESSON-7

MAHALANOBIS D^2 STATISTIC AND ITS APPLICATIONS

OBJECTIVES:

After successful completion of this unit, the students will be able to:

- Understand the concept of statistical distance in a multivariate framework and the need for Mahalanobis D^2 over Euclidean distance.
- Define and derive Mahalanobis D^2 statistic and to know its properties.
- Know the relationship between Mahalanobis D^2 and Hotelling's T^2 statistic.
- Apply Mahalanobis D^2 in hypothesis testing.

STRUCTURE:

7.1 INTRODUCTION:

7.2

In multivariate statistical analysis, it is often necessary to measure the distance or dissimilarity between observations or populations described by several correlated variables. Traditional distance measures such as the Euclidean distance treat all variables as independent and equally scaled, making them inappropriate when variables are correlated or measured in different units.

To overcome these limitations, Professor P. C. Mahalanobis introduced the **Mahalanobis** D^2 statistic**,** a covariance-adjusted measure of distance that incorporates both the variances and covariances of the variables. Unlike Euclidean distance, Mahalanobis D^2 standardizes the data using the variance–covariance matrix, providing a meaningful measure of separation in a multivariate setting.

The **Mahalanobis D^2** statistic plays a central role in multivariate inference**,** serving as the basis for important techniques such as Hotelling's T^2 test**,** discriminant analysis**,** cluster analysis**,** and multivariate outlier detection**.** Under the assumption of multivariate normality, the statistic follows a chi-square distribution**,** which allows it to be used for hypothesis testing and statistical decision-making.

Because of its ability to account for correlation structure and scale differences among variables, **Mahalanobis D^2** has wide applications in biology, medicine, quality control, economics, psychology, and social sciences**.** It remains one of the most fundamental and powerful tools in multivariate statistical analysis.

7.3 DEFINITION OF MAHALANOBIS D^2 STATISTIC (MAHALANOBIS SQUARED DISTANCE):

The Mahalanobis D^2 statistic**,** introduced by P. C. Mahalanobis**,** is a measure of distance between a multivariate observation and a population (or between two populations) that takes into account the variances and covariances among the variables.

For a p -variate random vector $\underset{\sim}{\mathbf{X}}$ with mean vector $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix $\boldsymbol{\Sigma}$, the Mahalanobis D^2 statistic is defined as

$$D^2 = \left(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}}\right)' \boldsymbol{\Sigma}^{-1} \left(\underset{\sim}{\mathbf{X}} - \underset{\sim}{\boldsymbol{\mu}}\right)$$

Where $\underset{\sim}{\mathbf{X}}$ = observation vector, $\underset{\sim}{\boldsymbol{\mu}}$ = population mean vector, and

$\boldsymbol{\Sigma}$ = variance–covariance (dispersion) matrix.

For measuring the distance between two multivariate populations with mean vectors $\mu_1$and μ_2, the Mahalanobis distance is:

$$D^2 = \left(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2\right)' \boldsymbol{\Sigma}^{-1} \left(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2\right)$$

The Mahalanobis D^2 statistic is scale-invariant**,** accounts for correlation among variables**,** and reduces to the squared Euclidean distance when the variables are uncorrelated with equal variances. It is widely used in multivariate hypothesis testing, classification, outlier detection, and discriminant analysis**.**

7.4 PROPERTIES OF MAHALANOBIS D^2 STATISTIC:

- **Scale Invariance:** The value of the value of D^2 does not change if the units of measurement of the variables are changed.
- **Accounts for Covariance:** It considers the correlation structure of the variables, which is a significant advantage over Euclidean distance.
- **Dimensionless:** It is a unitless measure of distance.
- **Zero Minimum:** The minimum possible value is zero, occurring when the observation.
- **Robust to Linear Transformations:**The distance remains unchanged under non-singular linear transformations of the data.

7.5 DERIVATION OF MAHALNOBIS D^2 TEST STATISTIC FOR TWO-SAMPLE PROBLEM AND IT'S RELATIONSHIP WITH HOTELLING'S T^2 STATISTIC:

Suppose $\boldsymbol{\pi}_1 : N_p(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma})$ and $\boldsymbol{\pi}_2 : N_p(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma})$ are two p-variate normal populations with mean vectors $\underset{\sim}{\boldsymbol{\mu}}_1$ and $\underset{\sim}{\boldsymbol{\mu}}_2$ respectively. Both the populations have the common dispersion matrix $\boldsymbol{\Sigma}$

Suppose $\underset{\sim}{\mathbf{x}}_{11}, \underset{\sim}{\mathbf{x}}_{12}, \ldots \underset{\sim}{\mathbf{x}}_{1n_1}$ be a random sample of size 'n_1' from population $\boldsymbol{\pi}_1 : N_p(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma})$ and let $\underset{\sim}{\mathbf{x}}_{21}, \underset{\sim}{\mathbf{x}}_{22}, \ldots, \underset{\sim}{\mathbf{x}}_{2n_2}$ be a random sample of size 'n_2' from population $\boldsymbol{\pi}_2 : N_p(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma})$. $H_0 : \underset{\sim}{\boldsymbol{\mu}}_1 = \underset{\sim}{\boldsymbol{\mu}}_2$ vs $\mathrm{H}_1 : \underset{\sim}{\boldsymbol{\mu}}_1 \neq \underset{\sim}{\boldsymbol{\mu}}_2$.

Now our problem is to test $H_0 : \underset{\sim}{\boldsymbol{\mu}}_1 = \underset{\sim}{\boldsymbol{\mu}}_2$ vs $\mathrm{H}_1 : \underset{\sim}{\boldsymbol{\mu}}_1 \neq \underset{\sim}{\boldsymbol{\mu}}_2$ or to test the significance of

the difference $\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2$ or to test the separation between the two populations $\boldsymbol{\pi}_1$ *and* $\boldsymbol{\pi}_2$ is significant based on the above given samples based on Mahalanobis D^2 statistic, which is as explained below.

The Mahalanobis D^2 test statistic (which is nothing but Mahalanobis squared distance between $\underset{\sim}{\bar{\mathbf{x}}}_1$ and $\underset{\sim}{\bar{\mathbf{x}}}_2$) is given by

$$D^2 = (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1} (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (1)$$

where

$$\underset{\sim}{\bar{\mathbf{x}}}_1 = \frac{1}{n_1} \sum_{\alpha=1}^{n_1} \underset{\sim}{\mathbf{x}}_{1\alpha} \text{ (unbiased estimator of } \underset{\sim}{\boldsymbol{\mu}}_1 \text{)}$$

$$\underset{\sim}{\bar{\mathbf{x}}}_2 = \frac{1}{n_2} \sum_{\alpha=1}^{n_2} \underset{\sim}{\mathbf{x}}_{2\alpha} \text{ (unbiased estimator of } \underset{\sim}{\boldsymbol{\mu}}_2 \text{)}$$

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2} \qquad (2)$$

is an unbiased estimator of $\boldsymbol{\Sigma}$ based on the pooled samples.

$$\mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{\alpha=1}^{n_1} (\underset{\sim}{\mathbf{x}}_{1\alpha} - \underset{\sim}{\bar{\mathbf{x}}}_1)(\underset{\sim}{\mathbf{x}}_{1\alpha} - \underset{\sim}{\bar{\mathbf{x}}}_1)'$$

$$\mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{\alpha=1}^{n_2} (\underset{\sim}{\mathbf{x}}_{2\alpha} - \underset{\sim}{\bar{\mathbf{x}}}_2)(\underset{\sim}{\mathbf{x}}_{2\alpha} - \underset{\sim}{\bar{\mathbf{x}}}_2)'$$

Now define, $\underset{\sim}{\mathbf{Y}} = \underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2 \qquad (3)$

with mean, $\mathrm{E}(\underset{\sim}{\mathbf{Y}}) = \mathrm{E}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) = \underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2$ and the variance- covariance matrix,

$V(\underset{\sim}{\mathbf{Y}}) = V(\underset{\sim}{\bar{\mathbf{x}}}_1) + V(\underset{\sim}{\bar{\mathbf{x}}}_2)$ ($\because$ The two samples are independent)

$$= \frac{1}{n_1}\boldsymbol{\Sigma} + \frac{1}{n_2}\boldsymbol{\Sigma} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma} \qquad (4)$$

Thus, we have

$$\underset{\sim}{\mathbf{Y}} \sim \mathrm{N}_p\left(\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}\right)$$

$$\Rightarrow \underset{\sim}{\mathbf{Y}} - (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2) \sim \mathrm{N}_p\left(\underset{\sim}{\mathbf{0}}, \left(\frac{n_1 + n_2}{n_1 n_2}\right)\boldsymbol{\Sigma}\right)$$

$$\Rightarrow \frac{\underset{\sim}{\mathbf{Y}} - (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)}{\sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \sim \mathrm{N}_p(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma}) \qquad (5)$$

We know that

$$(n_1\text{-}1)\mathbf{S}_1 \sim w_{n_1\text{-}1}(\mathbf{\Sigma}) \ \& \ (n_2\text{-}1)\mathbf{S}_2 \sim w_{n_2\text{-}1}(\mathbf{\Sigma})$$

Since, the samples are independent, we have $(n_1\text{-}1)\,S_1$ and $(n_2\text{-}1)\,S_2$ are also independent . Therefore from (2), $(n_1+n_2-2)S$ is distributed as Wishart distribution with n_1+n_2-2 d.f and with the parametric matrix Σ,

i.e. $(n_1+n_2-2)S \sim w_{n_1+n_2\text{-}2}(\Sigma)$ (6)

Since, S and $\underset{\sim}{Y} = \underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2$ are independently distributed, by applying the definition of T^2-distribution to the Eqs. (5) & (6), we have

$$T^2 = \left[\frac{\underset{\sim}{Y} - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}}\right]' S^{-1} \left[\frac{\underset{\sim}{Y} - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}}\right]$$

$$= \left(\frac{n_1 n_2}{n_1+n_2}\right)\left[\underset{\sim}{Y} - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right]' \mathbf{S}^{-1}\left[\underset{\sim}{Y} - (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\right] \quad (7)$$

and is distributed as T^2-distribution with n_1+n_2-2 d.f.

Now, under $H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ i.e. $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2 = 0$, (7) becomes

$$T^2 = \left(\frac{n_1 n_2}{n_1+n_2}\right)\underset{\sim}{Y}' \mathbf{S}^{-1}\underset{\sim}{Y} \sim T^2_{n_1+n_2-2}$$

$$\Rightarrow \left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right)' \mathbf{S}^{-1}\left(\underset{\sim}{\bar{x}}_1 - \underset{\sim}{\bar{x}}_2\right) \sim T^2_{n_1+n_2-2} \quad \text{(From Eq. (3))}$$

$$\Rightarrow \left(\frac{n_1 n_2}{n_1+n_2}\right) D^2 \sim T^2_{n_1+n_2-2} \quad \text{(From Eq. (1))} \quad (8)$$

Now, by virtue of the relation between T^2 and F – distribution, we have

$$\frac{T^2}{n_1+n_2\text{-}2} = \frac{1}{n_1+n_2\text{-}2}\left(\frac{n_1 n_2}{n_1+n_2}\right) D^2 \sim \frac{p}{n_1+n_2\text{-}2\text{-}(p\text{-}1)} F_{p,\, n_1+n_2\text{-}2\text{-}(p\text{-}1)}$$

$$\Rightarrow D^2 \sim \left(\frac{n_1+n_2}{n_1 n_2}\right)\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,\, n_1+n_2-p-1}$$

Thus, if $D^2 > \left(\frac{n_1+n_2}{n_1 n_2}\right)\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,\, n_1+n_2-p-1}(\alpha)$ (9)

where $F_{p,\, n_1+n_2-p-1}(\alpha)$ is table F- value at α level of significance with (p, n_1+n_2-p-1) d.f., then $\mathbf{H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2}$ may be rejected .

Thus, the D^2 test statistic can be used for testing $H_0: \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2$ vs $H_1 : \underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$ or equally for testing the significance of the difference $\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2$.

If H_0 is rejected , we can conclude that the separation between the two populations π_1 *and* π_2 is significant.

Thus, from Eq. (8), we may notice that the Hotelling's T^2 and Mahalanobis D^2 are closely associated as with the following relationship between them

$$T^2 = \left(\frac{n_1 n_2}{n_1+n_2}\right) D^2 \quad (10)$$

7.6 SUMMARY:

Mahalanobis D^2 statistic is a fundamental measure in multivariate statistical analysis used to quantify the distance between observations or populations when multiple correlated variables are involved. Unlike Euclidean distance, Mahalanobis D^2 incorporates the variance–covariance structure of the data, thereby adjusting for differences in scale and correlation among variables.

The statistic is defined as a quadratic form involving the inverse of the covariance matrix and follows a chi-square distribution under the assumption of multivariate normality. Mahalanobis D^2 serves as the theoretical foundation for several important multivariate techniques, including Hotelling's T^2 test**,** discriminant analysis**,** and multivariate outlier detection**.** Its invariance under linear transformations makes it a robust and reliable distance measure in multivariate space.

Mahalanobis D^2 statistic provides a powerful and meaningful way to assess similarity or dissimilarity in multivariate data by accounting for correlation and variability among variables. Its applications extend across diverse fields such as biology, medicine, quality control, economics, psychology, and social sciences**,** where simultaneous consideration of multiple characteristics is essential.

Despite its reliance on assumptions such as multivariate normality and a non-singular covariance matrix, Mahalanobis D^2 remains an indispensable tool in modern statistical

practice. When these assumptions are reasonably satisfied, the statistic offers accurate inference, effective classification, and insightful interpretation of complex multivariate datasets. Overall, Mahalanobis D^2 continues to play a crucial role in both theoretical development and practical applications of multivariate statistics.

7.7 SELF-ASSESSMENT QUESTIONS:

1. Define Mahalanobis D^2 and explain its importance in multivariate analysis.
2. Explain in detail the properties of Mahalanobis D^2 statistic.
3. Establish the relationship between Mahalanobis D^2 and Hotelling's T^2 statistic.
4. Explain the use of Mahalanobis D^2 in hypothesis testing.
5. Given a multivariate observation and the corresponding covariance matrix, compute the Mahalanobis D^2 statistic.
6. Using Mahalanobis D^2, identify whether an observation is an outlier at a given level of significance.

7.7 SUGGESTED READINGS:

1. Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern.
2. Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International.
3. Morrison, D.F. (2004): Multivariate Statistical Methods (Fourth Edition). Duxbury Press, New York.
4. Mardia, K.V., Kent, J. T and Bibby, J. M. (1979): Multivariate Analysis. Academic Press, New York.
5. Brenner, D., Bilodeau, M. (1999). Theory of Multivariate Statistics. Germany: Springer.
6. Giri Narayan C. (1995). Multivariate Statistical Analysis.

Dr. U. Ramkiran

LESSON-8

MANOVA FOR ONE - WAY CLASSIFICATION

OBJECTIVES:

After completing this lesson, students will be able to:

- **Understand the need for MANOVA**
  Explain why Multivariate Analysis of Variance (MANOVA) is required when multiple correlated response variables are analyzed simultaneously.
- **Formulate the one-way MANOVA model**
  Express the one-way classification MANOVA model using matrix notation and identify treatment and error components.
- **Understand distributional assumptions**
  State and verify assumptions such as multivariate normality, homogeneity of covariance matrices, and independence of observations.

STRUCTURE:

8.1 INTRODUCTION TO MANOVA:

Multivariate Analysis of Variance (MANOVA) is a generalization of the univariate Analysis of Variance (ANOVA) employed when two or more correlated response variables are observed for each experimental unit. Unlike ANOVA, which tests for differences among group means for a single dependent variable, MANOVA simultaneously examines differences among the mean vectors of multiple groups.

The principal aim of MANOVA is to assess whether variations in the levels of one or more independent (classification) variables produce statistically significant effects on a set of dependent variables considered jointly. By incorporating all response variables into a single analysis, MANOVA effectively accounts for the interrelationships among the variables and provides a more comprehensive evaluation of group effects.

MANOVA is especially appropriate for experimental and observational studies involving multidimensional outcomes. It overcomes the limitation of inflated Type I error rates that arise when multiple univariate ANOVA tests are conducted independently for each response variable, by offering a single global test of significance.

In many agricultural experiments, generally the data on more than one character is observed. One common example is grain yield and straw yield. The other characters on which the data is generally observed are the plant height, number of green leaves, germination count, etc. The analysis is normally done only on the grain yield and the best treatment is identified on the basis of this character alone. The straw yield is generally not taken into account. If we see the system as a whole, the straw yield is also important either for the cattle feed or for mulching or manuring, etc. Therefore, while analyzing the data, the straw yield should also be taken into consideration. Similarly, in varietal trials also the data is collected on several plant characteristics and quality parameters. In these experimental situations also the data is generally analyzed separately for each of the characters. The best treatment or genotype is identified separately for each of the characters. In these situations, Multivariate Analysis of Variance (MANOVA) can be helpful.

In the case of one-way classification, MANOVA tests the hypothesis that the mean vectors corresponding to different levels of a single factor are equal. Owing to its ability to handle multiple correlated responses simultaneously, MANOVA is extensively applied in disciplines such as medicine, psychology, education, agriculture, economics, and the social sciences.

Consequently, MANOVA constitutes a powerful and efficient statistical technique for investigating group differences in multivariate data, yielding more meaningful and reliable inferences than those obtained from separate univariate analyses.

8.1.1 LIMITATIONS OF UNIVARIATE ANOVA:

Univariate Analysis of Variance (ANOVA) is designed to compare group means **for** a single response variable**.** When multiple response variables are present, applying separate ANOVA tests to each variable leads to several problems:

- Inflated Type I error rate due to multiple testing
- Ignoring correlations among response variables
- Loss of overall group effect interpretation
- Reduced statistical power in detecting joint differences

Thus, univariate ANOVA is inadequate when responses are correlated and must be analyzed simultaneously.

8.1.2 MOTIVATION FOR MULTIVARIATE TESTING:

Multivariate Analysis of Variance (MANOVA) extends ANOVA to situations involving two or more dependent variables**.** The main motivations are:

- To test equality of mean vectors across groups
- To account for correlations among responses
- To provide a single overall test for group differences
- To increase efficiency and interpretability of results.

8.1.3 EXAMPLES OF ONE-WAY CLASSIFICATION WITH MULTIPLE RESPONSES:

- Agriculture**:** Effect of fertilizer type on yield, plant height, and leaf area
- Medicine**:** Effect of treatment on blood pressure, cholesterol, and heart rate
- Education**:** Teaching methods compared using math score, reading score, and reasoning ability

In all cases, there is one classification factor (grouping variable) and multiple response variables**.**

8.2 COMPARISION OF SEVERAL MULTIVARIATE POPULATION MEANS:

In multivariate analysis, interest often lies in comparing several population mean vectors corresponding to different groups. MANOVA provides a formal framework for testing whether these mean vectors are equal.

The null hypothesis is:

$$\mathbf{H_0 : \mu_1 = \mu_2 = ... = \mu_g}$$

against the alternative that at least one mean vector differs.

This comparison is carried out using SSCP matrices and appropriate multivariate test statistics.

8.2.1 Definition of MANOVA:

MANOVA is a statistical method for comparing means of multiple dependent variables across different levels of one or more independent variables. Instead of comparing univariate means, MANOVA compares vectors of group means. The fundamental idea is to create a linear combination of the dependent variables that maximizes the differences between the groups.

8.2.2 ONE-WAY MANOVA MODEL:

(ONE–WAY MANOVA) MULTIVARIATE ANALYSIS OF VARIANCE :-

Suppose we have ‘g’ populations, each is distributed multivariate normal with mean vectors $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, ..., \underset{\sim}{\mu}_g$ respectively. Let us suppose that all populations have the same covariance matrix Σ. Thus, we have the ‘g’ populations.

$$\Pi_1 \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$$

$$\Pi_2 \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$$

$$\vdots$$

$$\Pi_g \sim N_p(\underset{\sim}{\mu}_g, \Sigma)$$

Now , we have a sample of size ‘ n_i ’ from i^{th} population Π_i . Thus, we have ‘g’ samples from the ‘g’ populations as follows :

$$\text{Population } \Pi_1 \ : \ \underset{\sim}{\mathbf{X}}_{11}, \underset{\sim}{\mathbf{X}}_{12}, ..., \underset{\sim}{\mathbf{X}}_{1n_1}$$

$$\text{Population } \Pi_2 \ : \ \underset{\sim}{\mathbf{X}}_{21}, \underset{\sim}{\mathbf{X}}_{22}, ..., \underset{\sim}{\mathbf{X}}_{2n_2}$$

$$\vdots$$

$$\text{Population } \Pi_g \ : \underset{\sim}{\mathbf{X}}_{g1}, \underset{\sim}{\mathbf{X}}_{g2}, ..., \underset{\sim}{\mathbf{X}}_{gng}$$

Using the above random samples , MANOVA is used to investigate whether the population mean vectors are same and if not , which mean components differ significantly. Thus , the null hypothesis is

$$H_0 : \underset{\sim}{\boldsymbol{\mu}}_1 = \underset{\sim}{\boldsymbol{\mu}}_2 = ... = \underset{\sim}{\boldsymbol{\mu}}_g \qquad \rightarrow (1)$$

ASSUMPTIONS CONCERNING THE STRUCTURE OF THE DATA :

- Observation Independence: Each observation should be independent of one another. For example, one student's performance should not influence another's. The random samples from different populations are independent.
- Multivariate Normality: The combined dependent variables should be approximately normally distributed for each group of the independent variable.
- All populations have a common covariance matrix $\boldsymbol{\Sigma}$. That is Homogeneity of Variance-Covariance Matrices: The variance-covariance matrix of the dependent variables should be similar for all groups. This means that the spread and relationship between variables should be consistent across groups.
- Absence of Multicollinearity: The dependent variables should not be too highly correlated. If two variables are very similar, it doesn't add value to have both.

Suppose, the mean vector of i^{th} population is written as

$$\underset{\sim}{\boldsymbol{\mu}}_i = \underset{\sim}{\boldsymbol{\mu}} + \underset{\sim}{\boldsymbol{\tau}}_i \qquad \rightarrow (2)$$

Here, $\underset{\sim}{\boldsymbol{\mu}}$ is the overall mean vector of all population and $\underset{\sim}{\boldsymbol{\tau}}_i$ is a component due to the specific population, then the null hypothesis (1) can be written as

$$H_0 : \underset{\sim}{\boldsymbol{\tau}}_1 = \underset{\sim}{\boldsymbol{\tau}}_2 = ... = \underset{\sim}{\boldsymbol{\tau}}_g = \underset{\sim}{\mathbf{0}} \qquad \rightarrow (3)$$

The response $\underset{\sim}{\mathbf{X}}_{ij}$, distributed as $N_p(\underset{\sim}{\boldsymbol{\mu}} + \underset{\sim}{\boldsymbol{\tau}}_i, \boldsymbol{\Sigma})$, can be expressed in the suggestive form,

$$\underset{\sim}{\mathbf{X}}_{ij} = \underset{\sim}{\boldsymbol{\mu}} + \underset{\sim}{\boldsymbol{\tau}}_i + \underset{\sim}{\boldsymbol{\varepsilon}}_{ij} \qquad \rightarrow (4)$$

$$\begin{pmatrix} \text{overall} \\ \text{mean} \end{pmatrix} \begin{pmatrix} \text{treatment} \\ \text{effect} \end{pmatrix} \begin{pmatrix} \text{random} \\ \text{error} \end{pmatrix}$$

$i = 1,2,.....,g \ \ \& \ \ j = 1,2,....., n_i$.

where , $\underset{\sim}{\boldsymbol{\varepsilon}}_{ij} \sim N_p(\underset{\sim}{\mathbf{0}}, \boldsymbol{\Sigma})$ are independent random variables. (4) is called as MANOVA model for comparing of population mean vectors. Here $\underset{\sim}{\boldsymbol{\mu}}$ is overall mean vector and $\underset{\sim}{\boldsymbol{\tau}}_i$ represents the i^{th} treatment effect with

$$\sum_{i=1}^{g} n_i \underset{\sim}{\tau}_i = \underset{\sim}{0} \qquad \rightarrow (5)$$

A vector of observations may be decomposed as suggested by model (4) . Thus,

$$\underset{\sim}{x}_{ij} \quad = \quad \underset{\sim}{\bar{x}} \quad + \quad (\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}) \quad + \quad (\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i) \qquad \rightarrow (6)$$

$$\begin{pmatrix}\text{observation}\end{pmatrix} \quad \begin{pmatrix}\text{overall}\\ \text{sample}\\ \text{mean } \underset{\sim}{\mu}\end{pmatrix} \quad \begin{pmatrix}\text{estimated}\\ \text{treatment}\\ \text{effect } \underset{\sim}{\tau}_i\end{pmatrix} \quad \begin{pmatrix}\text{residual}\\ \hat{\underset{\sim}{\varepsilon}}_{ij}\end{pmatrix}$$

When $\underset{\sim}{\bar{x}}_i$ = mean of i^{th} sample $\underset{\sim}{x}_{i1}, \underset{\sim}{x}_{i2}, ..., \underset{\sim}{x}_{in_i}$

$$\underset{\sim}{\bar{x}} = \frac{1}{g}\left(\underset{\sim}{\bar{x}}_1 + \underset{\sim}{\bar{x}}_2 + ... + \underset{\sim}{\bar{x}}_g\right) \qquad \textit{(general mean)}$$

From (6), we may write the cross product,

$$\left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}\right)\left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}\right)' = \left(\left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i\right) + \left(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}\right)\right)\left(\left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i\right) + \left(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}\right)\right)'$$

$$= \left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i\right)\left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i\right)' + \left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i\right)\left(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}\right)'$$

$$+ \left(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}\right)\left(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}\right)' + \left(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}\right)\left(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i\right)'$$

Summing the cross product over i and j , we get

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})' = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)'$$

$$+ \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})'$$

$$+ \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_{ij} - \underset{\sim}{\bar{x}}_i)'$$

$$+ \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})' \qquad \rightarrow (7)$$

But, since $\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i) = \underset{\sim}{0}$ *., Eq (7) becomes ,*

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})' = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_i)'$$

$$+\sum_{i=1}^{g} n_i(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}})(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}})' \qquad \rightarrow (8)$$

$$\Rightarrow \begin{pmatrix} \text{total(corrected)} \\ \text{sum of square} \\ \text{\& cross products} \end{pmatrix} = \begin{pmatrix} \text{residual(within)} \\ \text{sum of squares} \\ \text{\& cross products} \end{pmatrix} + \begin{pmatrix} \text{treatment(between)} \\ \text{sum of squares} \\ \text{\& cross products} \end{pmatrix}$$

That is (8) may be written as

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}})' = W + B \qquad \rightarrow (9)$$

where , $W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}})'$

$$= (n_1 - 1)S_1 + (n_2 - 1)S_2 + ... + (n_g - 1)S_g$$

where , S_i *is sample covariance matrix of* i^{th} *sample .*

and $B = \sum_{i=1}^{g} n_i(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}})(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}})'$

Now, we summarise the calculations leading to the test statistic in a MANOVA table . MANOVA table for comparing population mean vectors :-

Source of variation	Matrix of sum of squares & cross product	Degrees of freedom
Treatments	$B = \sum_{i=1}^{g} n_i(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}})(\underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}} - \underset{\sim}{\bar{\mathbf{x}}})'$	g-1
Residual(error)	$W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}}_{\mathbf{i}})'$	n-g
Total (correlated for the mean)	$B + W = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}})(\underset{\sim}{\mathbf{x}}_{\mathbf{ij}} - \underset{\sim}{\bar{\mathbf{x}}})'$	n-1

Now one of the test statistic for testing (3) involves generalized variances and is given by

$$\Lambda^* = \frac{|W|}{|B + W|} \qquad \rightarrow (10)$$

The quantity Λ^* is called Wilk's lamda and related to likelihood ratio criterion. The exact distribution of Λ^* can be derived for the special cases listed in the following table .

Distribution of Wilk's lamda , Λ^* **:-**

No . of variables	No .of groups	Sampling distribution for multivariate normal data

p - 1	g ≥ 2	$\left[\frac{\sum_{i=1}^{g} n_i - g}{g-1}\right]\left[\frac{1-\Lambda^*}{\Lambda^*}\right] \sim F_{\left(g-1,\sum_{i=1}^{g} n_i - g\right)}$
p = 2	g ≥ 2	$\left[\frac{\sum_{i=1}^{g} n_i - g - 1}{g-1}\right]\left[\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right] \sim F_{\left(2(g-1),2\sum_{i=1}^{g} n_i - g - 1\right)}$
p ≥ 1	g = 2	$\left[\frac{\sum_{i=1}^{g} n_i - p - 1}{p}\right]\left[\frac{1-\Lambda^*}{\Lambda^*}\right] \sim F_{\left(p,\sum_{i=1}^{g} n_i - p - 1\right)}$
p ≥ 1	g = 3	$\left[\frac{\sum_{i=1}^{g} n_i - p - 2}{p}\right]\left[\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right] \sim F_{\left(2p,2(\sum_{i=1}^{g} n_i - p - 2)\right)}$

Bartlett has shown that if H_o is true and $\sum_{i=1}^{g} n_i = n$ is large,

$$-\left[n-1-\frac{(p+g)}{2}\right]\ln \Lambda^* = -\left[n-1-\frac{(p+g)}{2}\right]\ln \frac{|W|}{|B+W|}$$

has approximately a χ^2 - distribution with p(g-1) d.f. consequently.

8.3 SUMMARY:

MANOVA is an option for statistical testing of multivariate experiments. The dependent variables are random normal. The test is more senstive than other parametrics to violations of normality and homogeneity of variance. MANOVA tests whether independent variables affect an abstract combination of dependent variables. For most, use MANOVA as an omnibus test followed by post hoc comparisons of interest to control FWER. Care should be taken in selecting the dependent variables of interest.

Multivariate Analysis of Variance (MANOVA) for one-way classification is a natural extension of univariate ANOVA to situations where multiple correlated response variables

are observed for each experimental unit. In this framework, observations are classified according to a single factor, and the primary objective is to test whether the population mean vectors corresponding to different groups are equal.

MANOVA overcomes the limitations of conducting separate univariate ANOVA tests by jointly analyzing all response variables. It accounts for the correlation structure among the variables and provides a single overall test of group differences, thereby controlling the inflation of Type I error. The method is based on the partitioning of total variation into between-groups (hypothesis) and within-groups (error) components using Sum of Squares and Cross-Products (SSCP) matrices.

The comparison of several multivariate population means is carried out using standard MANOVA test criteria such as Wilks' Lambda, Pillai's Trace, Hotelling–Lawley Trace, and Roy's Largest Root. Each statistic offers a different perspective on group separation, with Pillai's Trace being the most robust under departures from assumptions.

The validity of one-way MANOVA depends on key assumptions, including multivariate normality, homogeneity of covariance matrices, and independence of observations. When these assumptions are reasonably satisfied, MANOVA provides reliable and efficient inference. In cases of assumption violations, careful interpretation and the choice of robust test statistics are essential.

In conclusion, MANOVA for one-way classification is a powerful and comprehensive statistical technique for comparing groups when multiple responses are involved. By integrating information across correlated variables, it yields more meaningful and interpretable results than separate univariate analyses and is widely applicable in disciplines such as medicine, education, psychology, agriculture, economics, and the social sciences

8.4 SELF-ASSESSMENT QUESTIONS:

1. Explain in detail the procedure of carrying out MANOVA of one way classification.
2. Discuss the multivariate analysis of variance for one-way classified data. How can we test the equality of means of several groups using MANOVA?
3. An experiment was conducted to evaluate the effects of various training programs (Program A, Program B, and Program C) on employee productivity and job satisfaction over a three month period. The data collected for each training program, with five replications, are shown in the Table. Perform a one- way MANOVA at 5% significance level and draw an inference using Wilks' lambda.

Replication	**Program A**		**Program B**		**Program C**	
1	50	7	32	6	35	5
2	46	6	45	7	40	4
3	48	7	60	8	47	5
4	53	9	48	5	50	5
5	48	6	50	4	38	6

4. Explain the key difference between comparing means in ANOVA versus comparing mean vectors in MANOVA.
5. Name at least three assumptions required for the proper application of MANOVA.

8.5 SUGGESTED READINGS:

1. Anderson, T.W.(2000). An Introduction to Multivariate Statistical Analysis, 3rd Edition, Wiley Eastern.
2. Johnson, A. and Wichern, D.W.(2001). Applied Multivariate Statistical Analysis, Prentice Hall and International.
3. Morrison, D.F. (2004): Multivariate Statistical Methods (Fourth Edition). Duxbury Press, New York.
4. Rao, C.R. (2001): Linear Statistical Inference and its Applications (Second Edition), Wiley Inter Science, New York.
5. Mardia, K.V., Kent, J. T and Bibby, J. M. (1979): Multivariate Analysis. Academic Press, New York.
6. Brenner, D., Bilodeau, M. (1999). Theory of Multivariate Statistics. Germany: Springer.
7. Giri Narayan C. (1995). Multivariate Statistical Analysis.

Dr. U. Ramkiran

LESSON -9
DISCRIMINANT ANALYSIS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of Expected (or average) cost of misclassification and Total Probability of Misclassification
- To know the concept of Discriminant analysis
- To acquire knowledge about significance of Discriminant analysis
- To understand the purpose and objectives of pivotal provisions of the ECM and TPM regions

STRUCTURE

9.1. INTRODUCTION

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects (or observations) and with allocating new objects (observations) to previously defined groups. Discriminant analysis is rather exploratory in nature. As a seperatory procedure, it is often employed on a onetime basis in order to investigate observed differences when causal relationships are not well understood. Classification procedure are less exploratory in the sense that they lead to well defined rules, which can be used for assigning new objects. Classification ordinarily requires more problem structure than discrimination. Thus, the immediate goals of discrimination and classification, respectively, are as follows:

Goal 1: To describe either graphically (in three or fewer dimensions) or algebraically, the differential features of objects (observations) from several known collections (populations). We try to find “discriminants” whose numerical values are such that the collections are separated as much as possible.

Goal 2: To sort objects (observations) into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign a new object to the labeled classes.

We shall follow convention and use the term discrimination to refer to 'Goal 1'. This terminology was introduced by R. A. Fisher in the first modern treatment of separatory problems. A more descriptive term for this goal, however, is separation. We shall refer to the second goal as classification, or allocation. A function that separates may sometimes serve as an allocator, and, conversely, an allocator rule may suggest a discriminatory procedure. Thus, in practice Goal 1& Goal2 frequently overlap and distinction between separation and allocation is not clear.

The problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of several categories on the basis of these measurements. The investigator cannot identify the individual with a category directly but must use these measurements. In many cases it can be assumed that these are a finite number of categories or populations from which the individual may have come and each population is characterized by a probability distribution of the measurements. Thus, an individual is considered as a random observation from this population. The question is: Given an individual with certain measurements, from which population did it arise?

In some, instances, the categories are specified before hand in the sense that the probability distributions of the measurements are completely known. In other cases, the form of each distribution may be known, but the parameters of the distribution must be estimated from a sample from that population .In some other cases, the form of the distribution of the populations may not be known.

Let us give an example of a problem of discrimination and classification. Prospective students applying for admission into college are given a battery of tests; the vector of scores is a set of measurements $\underset{\sim}{\mathbf{x}}$. The prospective students may be a member of one population consisting of these students who will successfully complete college training or, rather, have potentialities for successfully completing training, or he/she may be member of the other population, those who will not complete the course successfully. The problem is to classify a student applying for admission on the basis of these scores on the entrance examination. Before that we have to describe or explore the differential scores between the two categories of the students from the past information. Also, we have to prepare a discriminant function that separates the two categories of students clearly as much as possible. This problem is called discrimination.

9.2 DISCRIMINATION AND CLASSIFICATION:

To fix ideas, we list below situations where one may be interested in

(1). Separating or discriminating two classes of objects.

Or (2). Assigning a new object to one of the two classes .

Or both (1)&(2).

It is convenient to label the classes $\boldsymbol{\pi}_1 \,\&\, \boldsymbol{\pi}_2$. The objects are ordinarily separated or classified on the basis of measurements on, for instance, P associated random variables. $\underset{\sim}{\mathbf{X}}' = (X_1, X_2, ..., X_p)$. The observed values of $\underset{\sim}{\mathbf{X}}$ differ to some extend from one class to the other (of the values of $\underset{\sim}{\mathbf{X}}$ were not very different for objects in $\boldsymbol{\pi}_1 \,\&\, \boldsymbol{\pi}_2$, there would be no problem; i.e., it would be indistinguishable and new objects could be assigned to either class

indiscriminately). We can think of the totality of values from the first class as being the population of $\underset{\sim}{\mathbf{x}}$ values for $\boldsymbol{\pi}_1$ and those from the second class as the population of $\underset{\sim}{\mathbf{x}}$ values for $\boldsymbol{\pi}_2$. These two populations can then be described by probability density functions $f_1(\underset{\sim}{\mathbf{x}}) \& f_2(\underset{\sim}{\mathbf{x}})$, and consequently, we can talk of assigning observations to populations (or objects to classes).

The following are some more examples:

(1). Separation of two species of chickweed based on the measurements sepal and petal lengths, petal left depth, bract length, scarious tip length and pollen diameter.

(2). Discrimination of successful and unsuccessful college students based on the entrance examination scores, high school grade point average and number of high school activities.

(3). Classification of purchasers of a new product and laggards (those slow to purchase) based on particulars of education, income, family size and amount of previous brand switching.

(4). Discriminating male-skulls and female-skulls based on the anthropological measurements like circumference and volume on ancient skulls.

(5). Separating good and poor credit risks based on the particulars of income, age, member of credit cards and family size.

From the above examples, it is clear that allocation or classification rules are usually developed from learning samples. Measured characteristics of randomly selected objects known to come from each of the two populations are examined for differences. Essentially, the set of possible sample outcomes is divided into two regions $R_1 \& R_2$, such that is a new observation falls in R_1, it is allocated to population $\boldsymbol{\pi}_1$ and is it falls in R_2, we allocate it to population $\boldsymbol{\pi}_2$. Thus one set of observed values favours $\boldsymbol{\pi}_1$, the other set of values favours $\boldsymbol{\pi}_2$. Here, it may be noted that classification rules cannot usually provide and error-free method of assignment. This is because there may not be a clear distinction between the measured characteristics of the populations ; i.e. the groups may overlap. It is then possible, for example, to incorrectly classify a $\boldsymbol{\pi}_2$ object as belonging to $\boldsymbol{\pi}_1$ or a $\boldsymbol{\pi}_1$ object as belonging to $\boldsymbol{\pi}_2$.

A good classification procedure should result in a few misclassifications. In other words, the chances or probabilities of misclassification should be small. As we shall see, there are additional features that an "optimal" classification rule should be possessed.

9.3 STANDARDS OF GOOD CLASSIFICATION:

In constructing a procedure of classification, it is desired to minimize the probability of misclassification or more specifically it is desired to minimize on the average the bad effects of misclassification.

Suppose an individual is an observation from either population $\boldsymbol{\pi}_1$ or population $\boldsymbol{\pi}_2$. The classification of an observation depends on the vector of measurements

$$\underset{\sim}{\mathbf{x}} = \left(x_1, x_2, \ldots . x_p\right)'_{p\times 1}$$

on that individual. We set up a rule that if an individual is characterized by certain sets of values of $x_1, x_2, \ldots, x_p$ it will be classified as from $\boldsymbol{\pi_1}$; if it has other values it is classified as from $\boldsymbol{\pi_2}$.

We can think of an observation $\underset{\sim}{\mathbf{x}}$ as a point in a P-dimensional space. We divide this space into two regions $R_1 \& R_2$ if the observation falls in R_1, we classify it as coming from $\boldsymbol{\pi_1}$ and if it falls in R_2 we classify it as coming from $\boldsymbol{\pi_2}$.

Usually, the statistician can make two kinds of errors in classification. If the individual is actually from $\boldsymbol{\pi_1}$ and is misclassified into $\boldsymbol{\pi_2}$; or if it is actually from $\boldsymbol{\pi_2}$ and is misclassified into $\boldsymbol{\pi_1}$. We need to know the relative undesirability of these two kinds of misclassification.

Let $f_1(\underset{\sim}{\mathbf{x}}) \& f_2(\underset{\sim}{\mathbf{x}})$ be the p.d.f.'s associated with the $p \times 1$ random vector $\underset{\sim}{\mathbf{X}}$ for populations $\boldsymbol{\pi_1} \& \boldsymbol{\pi_2}$ respectively. An object, with associated measurements $\underset{\sim}{\mathbf{x}}$, must be assigned to either $\boldsymbol{\pi_1}$**(or)**$\boldsymbol{\pi_2}$. Let $\boldsymbol{\Omega}$ be the sample space that is the collection of all possible observations $\underset{\sim}{\mathbf{x}}$. Let R_1 be that set of $\underset{\sim}{\mathbf{x}}$ values for which we classify objects as $\boldsymbol{\pi_1}$ and $R_2 = \Omega - R_1$ be the remaining $\underset{\sim}{\mathbf{x}}$ values for which we classify objects as $\boldsymbol{\pi_2}$. Since every object must be assigned to one and only one of the two populations, the sets $R_1 \& R_2$ be mutually exclusive and exhaustive.

9.4 EXPECTED (OR AVERAGE) COST OF MISCLASSIFICATION (ECM):

In order to obtain ECM we consider the following conditional probabilities:

P (correctly classifying an observation (object) that actually is drawn from $\boldsymbol{\pi_1}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi_1}) = \int_{R_1} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \text{P(1/1) (say)} \quad (1)$$

P(correctly classifying an observation that actually is drawn from $\boldsymbol{\pi_2}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi_2}) = \int_{R_2 = \Omega - R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \text{P(2/2)} \quad \text{(say)} \quad (2)$$

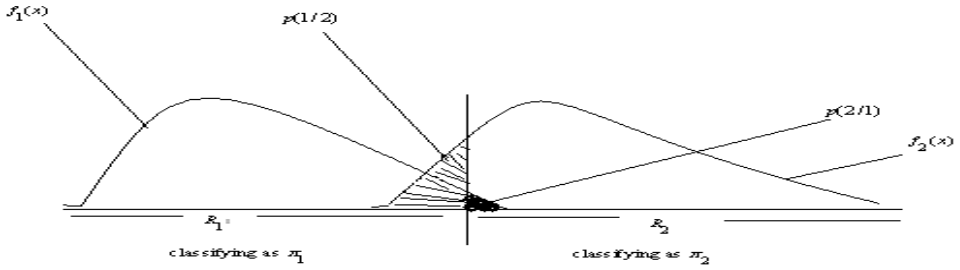
P(misclassifying an observation that is drawn from $\boldsymbol{\pi_1}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi_1}) = \int_{R_2} f_1(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \text{P(2/1)} \quad \text{(say)} \quad (3)$$

P(misclassifying an observation that is drawn from $\boldsymbol{\pi_2}$)

$$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi_2}) = \int_{R_1} f_2(\underset{\sim}{\mathbf{x}}) d\underset{\sim}{\mathbf{x}} = \text{P(1/2)} \quad \text{(say)} \quad (4)$$

Misclassification probabilities when p=1:



Let

p_1 = prior probability of $\boldsymbol{\pi_1}$

= P(drawing an observation from $\boldsymbol{\pi_1}$) = P($\boldsymbol{\pi_1}$) (5)

and p_2 = prior probability of $\boldsymbol{\pi_2}$

= P(drawing an observation from $\boldsymbol{\pi_2}$) = P($\boldsymbol{\pi_2}$) (6)

Now the overall probabilities of correctly or incorrectly classifying objects can be derived as the product of the prior and conditional classification probabilities. Thus we get

P(correctly classified as $\boldsymbol{\pi_1}$)=P(observations comes from $\boldsymbol{\pi_1}$ and is correctly Classified as $\boldsymbol{\pi_1}$)

$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi}_1).P(\boldsymbol{\pi}_1)$ = P(1/1). p_1 (from (1)&(5)) (7)

similarly

P(correctly classified as $\boldsymbol{\pi_2}$)=P(2/2). p_2 (from (2)&(6)) (8)

P(misclassified as $\boldsymbol{\pi_1}$) = P(observations comes from $\boldsymbol{\pi_2}$ and is misclassified as $\boldsymbol{\pi_1}$)

$= P(\underset{\sim}{\mathbf{X}} \in R_1 / \boldsymbol{\pi}_2).P(\boldsymbol{\pi}_2)$ = P(1/2). p_2 (from (4)&(6)) (9)

P(misclassified as $\boldsymbol{\pi_2}$) = P(observation comes from $\boldsymbol{\pi_1}$ and is misclassified as $\boldsymbol{\pi_2}$)

$= P(\underset{\sim}{\mathbf{X}} \in R_2 / \boldsymbol{\pi}_1).P(\boldsymbol{\pi}_1)$ = P(2/1). p_1 (from (3)&(5)) (10)

A good classification rule must take into account the misclassification costs. Although the statistician may not know these costs in each case, he will often have at least a rough idea of them. The costs of misclassification can be defined by a cost matrix C:

True population| Classified as

	π_1	π_2
π_1	0	$C(2/1)$
π_2	$C(1/2)$	0

The costs are

(1). Zero for correct classification .

(2). C(1/2) is cost involved when an observation drawn from $\boldsymbol{\pi}_2$ is incorrectly classified into $\boldsymbol{\pi}_1$.

(3). C(2/1) is cost involved when an observation actually drawn from $\boldsymbol{\pi}_1$ is incorrectly classified as $\boldsymbol{\pi}_2$.

Clearly, a good classification procedure is one which minimize in some sense or the cost of misclassification. Now , the expected cost of misclassification(ECM) is obtained by multiplying the off-diagonal entries in (11) by their probabilities of occurrence. Consequently a reasonable classification rule should has an ECM as small as possible. From the above the ECM may be defined as follows :

ECM= C(1/2) .P(misclassification into $\boldsymbol{\pi}_1$) + C(2/1) .P(misclassification into $\boldsymbol{\pi}_2$)

=C(1/2) .P(1/2). p_2 +C(2/1) . P(2/1) . p_1 (11)

Definition:

Expected (or average) cost of misclassification (ECM) is the sum of the products of costs of each misclassification multiplied by the probability of its occurrence. Its formula is given by Eq. (11).

Result (Optional ECM regions or Bayes regions):

The regions R_1 & R_2 that minimize ECM are defined by the values

of $\underset{\sim}{\mathbf{x}}$ for which the following inequalities hold.

$$R_1 = \left\{ \underset{\sim}{\mathbf{x}} \Big/ \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \left(\frac{C(1/2)}{C(2/1)} \right) \Big/ \left(\frac{p_1}{p_2} \right) \right\} \quad (1)$$

(density ratio) $\geq$ (cost ratio)/(prior probability ratio)

$$R_2 = \left\{ \underset{\sim}{\mathbf{x}} \Big/ \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \left(\frac{C(1/2)}{C(2/1)} \right) \Big/ \left(\frac{p_1}{p_2} \right) \right\} \quad (2)$$

Proof:

From Eq. (11), we have the expected cost of misclassification (ECM) as

ECM=C(1/2) .P(1/2). p_2 +C(2/1) . P(2/1) . p_1 (3)

But , we have

$$P(1/2) = \int_{R_1} f_2(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} \quad \text{and} \quad P(2/1) = \int_{R_2} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} \tag{4}$$

using Eqs. (4) in (3) we get

$$\text{ECM} = \text{C}(1/2)\ p_2 \int_{R_1} f_2(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} + \text{C}(2/1)\ p_1 \int_{R_2} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} \tag{5}$$

Noting that $\Omega = R_1 \cup R_2$ so that the total probability

$$1 = \int_{\Omega} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} = \int_{R_1} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} + \int_{R_2} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} \quad (\because R_1 \,\&\, R_2 \text{ are disjoint}) \tag{6}$$

using (6) in (5),we get

$$\text{ECM} = \text{C}(1/2)\ p_2 \int_{R_1} f_2(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}} + \text{C}(2/1)\ p_1 \left[1 - \int_{R_1} f_1(\underset{\sim}{\mathbf{x}})d\underset{\sim}{\mathbf{x}}\right]$$

$$= \int_{R_1} [C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}})]d\underset{\sim}{\mathbf{x}} + C(2/1)p_1 \tag{7}$$

Now p_1, p_2, C(1/2) and C(2/1) are non-negative. In addition $f_1(\underset{\sim}{\mathbf{x}}) \,\&\, f_2(\underset{\sim}{\mathbf{x}})$ are Non-negative for all $\underset{\sim}{\mathbf{x}}$ and are the only qualities in ECM that depend on $\underset{\sim}{\mathbf{x}}$. Therefore, minimization of ECM is equivalent to minimize the function

$$\int_{R_1} [C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}})]d\underset{\sim}{\mathbf{x}} \tag{8}$$

But, from the theory of integration (8) will be minimized is R_1 includes there values of $\underset{\sim}{\mathbf{x}}$ for which the integrand

$$\text{C}(1/2)\ p_2\ f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) \le 0 \tag{9}$$

and for all $\underset{\sim}{\mathbf{x}}$ those not included in R_1 or equivalently for all $\underset{\sim}{\mathbf{x}}$ those included in R_2

$$C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) > 0 \tag{10}$$

Thus from (9),

$$R_1 = \{\underset{\sim}{\mathbf{x}} / C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}}) - C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) \le 0\}$$

$$= \{\underset{\sim}{\mathbf{x}} / C(2/1)p_1 f_1(\underset{\sim}{\mathbf{x}}) \ge C(1/2)p_2 f_2(\underset{\sim}{\mathbf{x}})\}$$

$$= \left\{\underset{\sim}{\mathbf{x}} \Big/ \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \ge \left(\frac{C(1/2)}{C(2/1)}\right) \Big/ \left(\frac{p_1}{p_2}\right)\right\} \tag{11}$$

$$(\because all f_1, f_2, p_1, p_2, C(1/2) \,\&\, C(2/1) \text{ are all positive})$$

Similarly from Eq. (10),

$$R_2 = \left\{ \underset{\sim}{\mathbf{x}} \Big/ \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \left(\frac{C(1/2)}{C(2/1)} \right) / \left(\frac{p_1}{p_2} \right) \right\} \qquad (12)$$

where (12) gives (2).

REMARK:

It is clear from Eqs. (1) & (2) that the implementation of the minimum ECM rules requires

(1). The ratio of p.d.f.'s is f_1 / f_2 is to be evaluated at a new observation $\underset{\sim}{\mathbf{x}}_0$.

(2). The cost ratio $\frac{C(1/2)}{C(2/1)}$

(3). The prior probability ratio $\frac{p_1}{p_2}$

The appearance of ratios in the definition of the optimal classification regions has significance as often it is much easier to specify the ratios than their component parts.

Special cases of ECM regions:

Case(1): (Equal prior probabilities i.e. $p_1 = p_2$ or $\frac{p_1}{p_2} = 1$)

In this case (1) & (2) become

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \frac{C(1/2)}{C(2/1)}; \qquad R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \frac{C(1/2)}{C(2/1)}$$

Case (2): (Equal misclassification costs that is C(1/2)=C(2/1))

In this case $\frac{C(1/2)}{C(2/1)} = 1$ and therefore (1) & (2) become

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq \frac{p_2}{p_1} \quad \& \quad R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < \frac{p_2}{p_1}$$

Case (3): $p_1 = p_2 \,\&\, C(1/2) = C(2/1)$

In this case $\frac{p_1}{p_2} = 1 = \frac{C(1/2)}{C(2/1)}$ and therefore (1) & (2) become

$$R_1 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} \geq 1; \qquad R_2 : \frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})} < 1$$

NOTE:

(1). When the prior probabilities are not known, they are often taken to be equal.

(2). Similarly when the misclassification costs are unknown, they are often taken to be equal.

(3). If $\frac{C(1/2)}{C(2/1)} = \frac{p_1}{p_2}$ then $C(1/2)p_2 = C(2/1)p_1$ and hence

$$R_1 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq 1; \qquad R_2 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < 1$$

9.5 OPTIMAL TOTAL PROBABILITY OF MISCLASSIFICATION (TPM) REGIONS :

Criteria other than the ECM can be used to derive "optimal" classification procedures. For example, one might ignore the costs of misclassification and Choose R_1 & R_2 to minimize the total probability of misclassification (TPM).

TPM=P(misclassifying as $\boldsymbol{\pi_1}$ observation or misclassifying a $\boldsymbol{\pi_2}$ observation)

= P($\underset{\sim}{x}$ comes from $\boldsymbol{\pi_1}$ and is misclassified)+

P($\underset{\sim}{x}$ comes from $\boldsymbol{\pi_2}$ and is misclassified)

$$\Rightarrow TPM = P(\underset{\sim}{X} \in R_2 / \boldsymbol{\pi_1}).P(\boldsymbol{\pi}_1) + P(\underset{\sim}{X} \in R_1 / \boldsymbol{\pi}_2).P(\boldsymbol{\pi}_2)$$

$$= P_1 \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x} + P_2 \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}$$

$$= p_1 P(2/1) + p_2 P(1/2) \qquad (1)$$

But, when C(1/2)=C(2/1) (i.e. when misclassification costs are equal) we get from equation (12) of page 14,

$$ECM = C(1/2)[p_1 P(2/1) + p_2 P(1/2)] \qquad (2)$$

Now, from (1) & (2), it can be easily seen that minimizing (1) is equivalent to minimizing (2). In other words, minimizing TPM is equivalent to minimizing ECM with equal misclassification costs. Thus the optional TPM regions R_1 & R_2 are same as those given in case(2) of page 20. Thus

$$R_1 = \left\{ \underset{\sim}{x} / \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \frac{p_2}{p_1} \right\}$$
$$R_2 = \left\{ \underset{\sim}{x} / \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \frac{p_2}{p_1} \right\} \qquad (3)$$

ALOCATING A NEW OBSERVATION $\underset{\sim}{x}_0$ BASED ON BAYE'S POSTERIOR PROBABILITIES

We can also allocate a new observation $\underset{\sim}{\mathbf{x}}_0$ to the population with the largest posterior probability $P(\boldsymbol{\pi}_i / \underset{\sim}{\mathbf{x}}_0)$. By Baye's rule, the "posterior" probabilities are

$$P(\boldsymbol{\pi}_i / \underset{\sim}{\mathbf{x}}_0) = \text{P}(\boldsymbol{\pi}_1 \text{ occurs and observe } \underset{\sim}{\mathbf{x}}_0) / \text{P(observe } \underset{\sim}{\mathbf{x}}_0)$$

$$= \text{P(observe } \underset{\sim}{\mathbf{x}}_0 / \boldsymbol{\pi}_1).\text{P}(\boldsymbol{\pi}_1) \;/\; \{ \text{P(observe } \underset{\sim}{\mathbf{x}}_0 / \boldsymbol{\pi}_1).\text{P}(\boldsymbol{\pi}_1) + \text{P(observe } \underset{\sim}{\mathbf{x}}_0 / \boldsymbol{\pi}_2).\text{P}(\boldsymbol{\pi}_2) \}$$

$$= \frac{f_1(\underset{\sim}{\mathbf{x}}_0).p_1}{f_1(\underset{\sim}{\mathbf{x}}_0).p_1 + f_2(\underset{\sim}{\mathbf{x}}_0).p_2}$$

$$= \frac{p_1 f_1(\underset{\sim}{\mathbf{x}}_0)}{p_1 f_1(\underset{\sim}{\mathbf{x}}_0) + p_2 f_2(\underset{\sim}{\mathbf{x}}_0)} \tag{1}$$

$$P(\boldsymbol{\pi}_2 / \underset{\sim}{\mathbf{x}}_0) = 1 - P(\boldsymbol{\pi}_1 / \underset{\sim}{\mathbf{x}}_0)$$

$$= 1 - \frac{p_1 f_1(\underset{\sim}{\mathbf{x}}_0)}{p_1 f_1(\underset{\sim}{\mathbf{x}}_0) + p_2 f_2(\underset{\sim}{\mathbf{x}}_0)}$$

$$= \frac{p_2 f_2(\underset{\sim}{\mathbf{x}}_0)}{p_1 f_1(\underset{\sim}{\mathbf{x}}_0) + p_2 f_2(\underset{\sim}{\mathbf{x}}_0)} \tag{2}$$

Now classify an observation $\underset{\sim}{\mathbf{x}}_0$ into $\boldsymbol{\pi}_1$ when

$$P(\boldsymbol{\pi}_1 / \underset{\sim}{\mathbf{x}}_0) > P(\boldsymbol{\pi}_2 / \underset{\sim}{\mathbf{x}}_0)$$

$$\Rightarrow p_1 f_1(\underset{\sim}{\mathbf{x}}_0) > p_2 f_2(\underset{\sim}{\mathbf{x}}_0)$$

($\because$ Numerators of (1) & (2) are equal)

$$\Rightarrow \frac{f_1(\underset{\sim}{\mathbf{x}}_0)}{f_2(\underset{\sim}{\mathbf{x}}_0)} > \frac{p_2}{p_1} \tag{3}$$

Now from(3), it can be seen that allocating a new observation to a population based on Baye's posterior probabilities is same as optional TPM rule.

NOTE:

The above method is also equivalent to classify a new observation using optional ECM (Baye's method) rule when misclassification costs are equal.

9.6 CONCLUSION

Discriminant Analysis and Classification provide systematic statistical tools to separate known groups and to allocate new observations into appropriate populations. This lesson emphasized two major goals:

(i) Describing the differences between populations (separation) and
(ii) Assigning new observations to one of the populations (classification).
(iii) Using probability density functions, prior probabilities, and misclassification costs, we derived rules that minimize the chance or cost of wrong decisions.

The Expected Cost of Misclassification (ECM) serves as a fundamental criterion. The Bayes Rule provides the regions (decision boundaries) that minimize ECM or, in special cases, minimize the Total Probability of Misclassification (TPM)**.**
Special cases such as equal prior probabilities or equal misclassification costs simplify the classification rule. When misclassification costs and priors are unknown, they are commonly assumed equal. Bayes posterior probabilities offer another intuitive approach for assigning new observations.

9.7 SELF ASSESSMENT QUESTIONS

1. Explain discriminant analysis. Distinguish between discrimination and classification.
2. Explain in detail the standards of good classification.
3. Obtain the minimum expected or average cost of misclassification regions.
4. Discuss different special cases of minimum ECM regions and show how each case leads to simplified classification rules.
5. Derive the total probability of misclassification (TPM) regions and show that minimizing TPM is equivalent to minimizing ECM under equal misclassification costs.
6. Explain Bayes posterior probability classification and prove that it leads to the same rule as the minimum-TPM classifier.

9.8 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard *A. Johnson and*
2. *D. W. Wichern.*
3. **An Introduction to Multivariate Statistical Analysis** by *T.W. Anderson*
4. **Multivariate Statistical Methods: A Primer** by *Bryan F.J. Manly*
5. **Psychometric Theory** by *Jum C. Nunnally & Ira H. Bernstein*

Dr. A. Vasudeva Rao

LESSON -10

CLASSIFICATION BETWEEN TWO MULTIVARIATE NORMAL (MVN) POPULATIONS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of classification into one of two known multivariate normal populations and classification into one of two multivariate normal populations when the parameters are unknown
- To know the concept of classification analysis
- To acquire knowledge about significance of classification analysis
- To understand the purpose and objectives of pivotal provisions of the classification into one of two known multivariate normal populations and classification into one of two multivariate normal populations when the parameters are unknown

STRUCTURE

10.1. INTRODUCTION

10.1.1 classification into one of two mvn populations when the parametrers are known

One of the most fundamental problems in multivariate statistics is to classify an observation vector **x** into one of two populations, say $\pi 1$ and $\pi 2$. When both populations are assumed to follow multivariate normal distributions with completely specified parameters (mean vectors and covariance matrices are known), the classification rule is obtained by comparing their likelihood functions.

If $\pi 1 \sim Np(\mu 1, \Sigma)$ and $\pi 2 \sim Np(\mu 2, \Sigma)$ with common covariance matrix, then the optimal rule is based on a linear discriminant function.

10.1.2 classification into one of two mvn populations when the parametrers are unknown

- In practice, the true mean vectors μ1, μ2 and covariance matrices Σ (or Σ1, Σ2) are rarely known. Instead, they must be estimated from sample data obtained from each population.
- The sample mean vectors are used to estimate μ1 and μ2.
- If equal covariance matrices are assumed, the pooled sample covariance matrix is used to estimate Σ.
- The resulting empirical discriminant function resembles Fisher's Linear Discriminant, but with estimated parameters.
- This approach makes the method applicable in real-world classification problems (medicine, finance, biology, etc.). It also introduces new issues, such as the impact of estimation error on misclassification probabilities and the need for large-sample approximations.

10.1.3 classification into one of two mvn populations with unequal dispersion mataries

- When the two multivariate normal populations have unequal (Dispersion) covariance matrices, the problem becomes more complex. Unlike the equal covariance case, where the decision boundary is linear, here the likelihood ratio test leads to a quadratic classification rule.
- For π1 ~ Np(μ1, Σ1) and π2 ~ Np(μ2, Σ2), the log-likelihood ratio contains quadratic terms in x.
- The resulting Quadratic Discriminant Function (QDF) is used for classification.
- Geometrically, the separating surface between populations is no longer a hyperplane, but a quadratic surface (ellipsoidal, hyperbolic, or parabolic).
- This case is the most general form of the normal classification problem and is particularly important when populations have markedly different variances and correlations. However, it requires large sample sizes for stable estimation of separate covariance matrices and is more computationally demanding.

10.2 CLASSIFICATION INTO ONE OF TWO MVN POPULATIONS (with common covariance matrix $\sum$) WHEN THE PARAMETRERS ARE KNOWN

Classification procedures based on normal populations predominate in statistical practice because of their simplicity and reasonably high efficiency across a wide variety of population models. We assume $f_1(\underset{\sim}{\mathbf{x}}) \& f_2(\underset{\sim}{\mathbf{x}})$ are multivariate normal densities; the first with mean vector $\underset{\sim}{\boldsymbol{\mu}}_1$,and the second with mean vector $\underset{\sim}{\boldsymbol{\mu}}_2$ and both with common matrix $\boldsymbol{\Sigma}$. Now the p.d.f. of the

two populations $\boldsymbol{\pi_1} \boldsymbol{\&} \boldsymbol{\pi_2}$ are given by

$$f_i(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)} \quad \text{for} \quad i=1,2,\ldots.$$

(1)

The ratio of densities after simplification is

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)+\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)}$$

$$= e^{-\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x}+\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_1+\frac{1}{2}\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1+\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_2-\frac{1}{2}\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{x}+\frac{1}{2}\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2}$$

$$\Rightarrow \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x}-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2)}$$

$$(\because \underset{\sim}{\mu}_i'\Sigma^{-1}\underset{\sim}{x} = \underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_i \ \ for \ \ i=1,2)$$

$$= e^{[(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)]}$$

(2)

$$\text{for} \quad (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) = \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_2-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2$$

$$= \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2 \quad (\because \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1)$$

By minimum ECM classification rule, we have

$$R_1 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

where p_1 and p_2 are prior probabilities of π_1 and π_2.

C(1/2) = cost involved when an observation drawn from π_2 is incorrectly classified into π_1.

C(2/1) = cost involved when an observation drawn from π_1 is incorrectly classified into π_2.

From (2) we have, after taking logarithms on both sides

$$R_1 : (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) \geq \log K$$

$$R_2 : (\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1} + \underset{\sim}{\mu_2}) < \log K \qquad (3)$$

$$where \;\; K = \frac{C(1/2).p_2}{C(2/1)p_1}$$

The regions $R_1 \& R_2$ given by (3) are called as minimum ECM regions for two normal populations.

NOTES:

1. The first term of (3) viz

$$(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x} = \underset{\sim}{l}'\underset{\sim}{x}, \quad \text{where } \underset{\sim}{l} = \Sigma^{-1}(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2}) \qquad (4)$$

is the well known Fisher (linear) discriminant function, which is actually obtained by Fisher with entirely different argument which we will discuss later.

10.3 CLASSIFICATION INTO ONE OF TWO MVN POPULATIONS (with common covariance matrix $\sum$) WHEN THE PARAMETRERS ARE UNKNOWN

Now the p.d.f. of the two populations $\pi_1 \& \pi_2$ are given by

$$f_i(\underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_i)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu_i})} \qquad \text{for} \qquad i=1,2,….$$

(1)

The ratio of densities after simplification is

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{-\frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_1) + \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x} - \underset{\sim}{\mu}_2)}$$

$$= e^{-\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x} + \frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_1 + \frac{1}{2}\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1 + \frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu_2} - \frac{1}{2}\underset{\sim}{\mu_2}'\Sigma^{-1}\underset{\sim}{x} + \frac{1}{2}\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu_2}}$$

$$\Rightarrow \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x} - \underset{\sim}{\mu_2}'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu_1}'\Sigma^{-1}\underset{\sim}{\mu_1} - \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu_2})}$$

$$(\because \underset{\sim}{\mu_i}'\Sigma^{-1}\underset{\sim}{x} = \underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_i \;\; for \;\; i = 1,2)$$

$$= e^{[(\underset{\sim}{\mu_1}-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1}+\underset{\sim}{\mu_2})]}$$

(2)

$$\text{for } (\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1}+\underset{\sim}{\mu}_2)=\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu_1}+\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_2-\underset{\sim}{\mu_2}'\Sigma^{-1}\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2}'\Sigma^{-1}\underset{\sim}{\mu_2}$$

$$=\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu_1}-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2 \quad (\because \underset{\sim}{\mu_1}'\Sigma^{-1}\underset{\sim}{\mu_2}=\underset{\sim}{\mu_2}'\Sigma^{-1}\underset{\sim}{\mu_1})$$

By minimum ECM classification rule, we have

$$R_1: \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2: \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

where p_1 and p_2 are prior probabilities of π_1 and π_2.

C(1/2) = cost involved when an observation drawn from π_2 is incorrectly classified into π_1.

C(2/1) = cost involved when an observation drawn from π_1 is incorrectly classified into π_2.

From (2) we have, after taking logarithms on both sides

$$R_1: (\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1}+\underset{\sim}{\mu_2}) \geq \log K$$

$$R_2: (\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu_1}-\underset{\sim}{\mu_2})'\Sigma^{-1}(\underset{\sim}{\mu_1}+\underset{\sim}{\mu_2}) < \log K$$

$$where \;\; K=\frac{C(1/2).p_2}{C(2/1)p_1}$$

(3)

The regions R_1 & R_2 given by (3) are called as minimum ECM regions for two normal populations.

Suppose $\underset{\sim}{X}_{11}, \underset{\sim}{X}_{12}, ... \underset{\sim}{X}_{1n_1}$, be a random sample of size 'n_1', from population $\pi_1 : N(\underset{\sim}{\mu}_1, \Sigma)$ and let $\underset{\sim}{X}_{21}, \underset{\sim}{X}_{22}, ..., \underset{\sim}{X}_{2n_2}$ be a random sample of size 'n_2' from population $\pi_2 : N(\underset{\sim}{\mu}_2, \Sigma)$. Since $\underset{\sim}{\mu_1}, \underset{\sim}{\mu_2}$ & Σ are unknown we replace them with their unbiased estimators viz.,

$$\underset{\sim}{\bar{X}}_1=\frac{1}{n_1}\sum_{\alpha=1}^{n_1}\underset{\sim}{X}_{1\alpha}, \underset{\sim}{\bar{X}}_2=\frac{1}{n_2}\sum_{\alpha=1}^{n_2}\underset{\sim}{X}_{2\alpha}$$

(4)

and

$$S=\frac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$$

(5)

where

$$S_1 = \frac{1}{n_1 - 1}\sum_{\alpha=1}^{n_1}(\underset{\sim}{\mathbf{X}}_{1\alpha} - \underset{\sim}{\bar{\mathbf{X}}}_1)(\underset{\sim}{\mathbf{X}}_{1\alpha} - \underset{\sim}{\bar{\mathbf{X}}}_1)'$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{\alpha=1}^{n_2}(\underset{\sim}{\mathbf{X}}_{2\alpha} - \underset{\sim}{\bar{\mathbf{X}}}_2)(\underset{\sim}{\mathbf{X}}_{2\alpha} - \underset{\sim}{\bar{\mathbf{X}}}_2)'$$

(6)

Now, the estimated (or sample) minimum ECM regions can be obtained from the above method replacing $\underset{\sim}{\boldsymbol{\mu}}_1, \underset{\sim}{\boldsymbol{\mu}}_2 \,\&\, \boldsymbol{\Sigma}$ with their unbiased estimators $\underset{\sim}{\bar{\mathbf{X}}}_1, \underset{\sim}{\bar{\mathbf{X}}}_2 \,\&\, S$ (given by (1) & (2)) respectively. They are form equations as follows :

$$\hat{R}_1 : (\underset{\sim}{\bar{\mathbf{X}}}_1 - \underset{\sim}{\bar{\mathbf{X}}}_2)' S^{-1}\underset{\sim}{\mathbf{X}} - \frac{1}{2}(\underset{\sim}{\bar{\mathbf{X}}}_1 - \underset{\sim}{\bar{\mathbf{X}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{X}}}_1 + \underset{\sim}{\bar{\mathbf{X}}}_2) \geq \log\left[c(1/2)p_2 / c(2/1)p_1\right]$$

(7)

$$\hat{R}_2 : (\underset{\sim}{\bar{\mathbf{X}}}_1 - \underset{\sim}{\bar{\mathbf{X}}}_2)' S^{-1}\underset{\sim}{\mathbf{X}} - \frac{1}{2}(\underset{\sim}{\bar{\mathbf{X}}}_1 - \underset{\sim}{\bar{\mathbf{X}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{X}}}_1 + \underset{\sim}{\bar{\mathbf{X}}}_2) < \log \qquad \left[c(1/2)p_2 / c(2/1)p_1\right]$$

(8)

from (4)&(5), the estimated sample minimum classification ECM rule for two normal populations is given by

Allocate $\underset{\sim}{\mathbf{X}}_0$ to $\boldsymbol{\pi}_1$ if

$$(\underset{\sim}{\bar{\mathbf{X}}}_1 - \underset{\sim}{\bar{\mathbf{X}}}_2) S^{-1}\underset{\sim}{\mathbf{X}} - \frac{1}{2}(\underset{\sim}{\bar{\mathbf{X}}}_1 - \underset{\sim}{\bar{\mathbf{X}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{X}}}_1 + \underset{\sim}{\bar{\mathbf{X}}}_2) \geq \log K$$

(9)

where K= $\left[c(1/2)p_2 / c(2/1)p_1\right]$

Allocate $\underset{\sim}{\mathbf{X}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

NOTE:

(1) The estimated or sample minimum TPM rule for two normal populations with unknown parameters can be obtained from (6) replacing K with $\left(p_2 / p_1\right)$.

(2) When $p_1 = p_2 \,\&\, \quad c\left(1/2\right) = c\left(2/1\right)$,the estimated or sample minimum ECM rule is equivalent to sample ML rule and is given by

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}\underset{\sim}{\mathbf{x}}_0 \geq \frac{1}{2}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 + \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (10)$$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

(3) The estimated minimum ECM rule or sample ML rule amounts to comparing the scalar variable (univariate normal variable)

$$y = \hat{\underset{\sim}{l}}'\underset{\sim}{\mathbf{x}} \qquad \text{,where } \hat{\underset{\sim}{l}} = S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \tag{11}$$

evaluated at $\underset{\sim}{\mathbf{x}}_0$ is

$$y_0 = \hat{\underset{\sim}{l}}'\underset{\sim}{\mathbf{x}}_0$$

with the number

$$\begin{aligned} \hat{m} &= \frac{1}{2}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)' S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 + \underset{\sim}{\bar{\mathbf{x}}}_2) \\ &= \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \end{aligned} \tag{12}$$

where

$$\bar{y}_1 = (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)^{-1} S^{-1}\underset{\sim}{\bar{\mathbf{x}}}_1 = \hat{\underset{\sim}{l}}'\underset{\sim}{\bar{\mathbf{x}}}_1$$
$$\bar{y}_2 = \hat{\underset{\sim}{l}}'\underset{\sim}{\bar{\mathbf{x}}}_2$$

Thus allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$y_0 \geq \hat{m} \tag{13}$$

otherwise allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$

That is, the estimated minimum ECM rule for two normal populations is to creating two univariate normal populations for the y values by taking an appropriate linear combination of the observations from populations

$\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ and then assigning a new new observation $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ or $\boldsymbol{\pi}_2$ depending upon whether $y_0 = \hat{\underset{\sim}{l}}'\underset{\sim}{\mathbf{x}}_0$ falls to the right or left of the midpoint $\hat{m}$, between the two normal means $\bar{y}_1$ and $\bar{y}_2$.

(4) The linear function (7) is known as Fisher linear discriminant function, which is obtained by Fisher with a different argument for separating two populations.

10.4 CLASSIFICATION INTO ONE OF TWO MVN POPULATIONS WHEN $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Now the p.d.f. of the two populations $\boldsymbol{\pi}_1 \,\&\, \boldsymbol{\pi}_2$ are given by

$$f_i(\underset{\sim}{\mathbf{x}}) = \frac{1}{(2\boldsymbol{\pi})^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)} \qquad \text{for i=1,2} \tag{1}$$

The ratio of densities after simplification is

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)+\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)}$$

$$= e^{-\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x}+\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_1+\frac{1}{2}\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1+\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_2-\frac{1}{2}\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{x}+\frac{1}{2}\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2}$$

$$\Rightarrow \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = e^{\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{x}-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2)}$$

$$(\because \underset{\sim}{\mu}_i'\Sigma^{-1}\underset{\sim}{x} = \underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{\mu}_i \ \ for \ \ i=1,2)$$

$$= e^{[(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)]}$$

(2)

$$\text{for } (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) = \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_2-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2$$

$$= \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_2 \quad (\because \underset{\sim}{\mu}_1'\Sigma^{-1}\underset{\sim}{\mu}_2 = \underset{\sim}{\mu}_2'\Sigma^{-1}\underset{\sim}{\mu}_1)$$

By minimum ECM classification rule, we have

$$R_1 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2 : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \left(\frac{C(1/2)}{C(2/1)}\right)\left(\frac{p_2}{p_1}\right)$$

Where p_1 and p_2 are prior probabilities of π_1 and π_2.

C(1/2) .P(misclassification into π_1), C(2/1) .P(misclassification into π_2)

From (2) we have, after taking logarithms on both sides

$$R_1 : (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \geq \log K$$

$$R_2 : (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x}-\frac{1}{2}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) < \log K$$

$$where \ \ K = \frac{C(1/2).p_2}{C(2/1)p_1}$$

(3)

The regions R_1 & R_2 given by (3) are called as minimum ECM regions for two normal populations.

Here we have $\boldsymbol{\pi}_1 : N(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{\pi}_2 : N(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_2)$ when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

Let $f_1(\underset{\sim}{\mathbf{x}})$ be the p.d.f. of $\boldsymbol{\pi}_1$ and $f_2(\underset{\sim}{\mathbf{x}})$ be the p.d.f. of $\boldsymbol{\pi}_2$. Then on simplification,

$$\log\left[\frac{f_1(\underset{\sim}{\mathbf{x}})}{f_2(\underset{\sim}{\mathbf{x}})}\right] = \log f_1(\underset{\sim}{\mathbf{x}}) - \log f_2(\underset{\sim}{\mathbf{x}})$$

$$= 1/2\,\underset{\sim}{\mathbf{x}}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}} + (\underset{\sim}{\boldsymbol{\mu}}_1'\boldsymbol{\Sigma}_1^{-1} - \underset{\sim}{\boldsymbol{\mu}}_2'\boldsymbol{\Sigma}_2^{-1}) - \lambda \quad (4)$$

where $\lambda = 1/2\log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + 1/2(\underset{\sim}{\boldsymbol{\mu}}_1'\boldsymbol{\Sigma}_1^{-1}\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2'\boldsymbol{\Sigma}_2^{-1}\underset{\sim}{\boldsymbol{\mu}}_2)$

(5)

we have general formula for minimum ECM region and is given by

$$R_1 : \log\left[f_1(\underset{\sim}{\mathbf{x}}) / f_2(\underset{\sim}{\mathbf{x}})\right] \geq \log\ \mathrm{k} \quad \text{,where } \mathrm{K=c(1/2)p_2 / c(2/1)}\,p_1$$

$$R_2 : \log\left[f_1(\underset{\sim}{\mathbf{x}}) / f_2(\underset{\sim}{\mathbf{x}})\right] < \log\ \mathrm{k}$$

where $\lambda = 1/2\log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + 1/2(\underset{\sim}{\boldsymbol{\mu}}_1'\boldsymbol{\Sigma}_1^{-1}\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2'\boldsymbol{\Sigma}_2^{-1}\underset{\sim}{\boldsymbol{\mu}}_2)$

(6)

Now, the minimum ECM regions for classification of two normal populations when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ is given by:

$$R_1 : -1/2\,\underset{\sim}{\mathbf{x}}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}} + (\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \underset{\sim}{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}} - \lambda \geq \log\ \mathrm{k}$$

$$\mathrm{R}_2 : -1/2\,\underset{\sim}{\mathbf{x}}'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}} + (\underset{\sim}{\mu}_1'\boldsymbol{\Sigma}_1^{-1} - \underset{\sim}{\mu}_2'\boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}} - \lambda < \log\ \mathrm{k}$$

where λ & k are given as (2) & (3) (7)

The allocation rule that minimizes the ECM is given by :

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$-1/2\,\underset{\sim}{\mathbf{x}}_0'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}}_0 + (\underset{\sim}{\boldsymbol{\mu}}_1'\boldsymbol{\Sigma}_1^{-1} - \underset{\sim}{\boldsymbol{\mu}}_2'\boldsymbol{\Sigma}_2^{-1})\underset{\sim}{\mathbf{x}}_0 - \lambda \geq \log\ \mathrm{k} \quad (8)$$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

In practice, the classification rule in (5) is implemented by substituting the sample quantities $\bar{\underset{\sim}{\mathbf{x}}}_1, \bar{\underset{\sim}{\mathbf{x}}}_2, S_1$ and S_2 for $\underset{\sim}{\boldsymbol{\mu}}_1, \underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ respectively.

QUADRATIC CLASSIFICATION RULE (NORMAL POPULATIONS WITH $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$)

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$ if

$$1/2\underset{\sim}{\mathbf{x}}_0'(S_1^{-1}-S_2^{-1})\underset{\sim}{\mathbf{x}}_0+(\underset{\sim}{\bar{\mathbf{x}}}_1'S_1^{-1}-\underset{\sim}{\bar{\mathbf{x}}}_2'S_2^{-1})\underset{\sim}{\mathbf{x}}_0-\hat{\lambda}\geq \log\ k \qquad (9)$$

allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$ otherwise.

$$\text{Where}\quad \hat{\lambda}=1/2\log\left(\frac{|S_1|}{|S_2|}\right)+1/2(\underset{\sim}{\bar{\mathbf{x}}}_1'S_1^{-1}\underset{\sim}{\bar{\mathbf{x}}}_1-\underset{\sim}{\bar{\mathbf{x}}}_2'S_2^{-1}\underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (10)$$

NOTE:

(1). Minimum TPM rule or quadratic classification rule when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$ is a special case of (6) when $K = p_2 / p_1$.

(2). If the misclassification costs are equal and prior probabilities are equal (i.e. $C(1/2) = C(2/1)$ & $p_1 = p_2$). Then the MC rule or QCR is obtained by taking K=1 or log K=0 in the rule (6).

10.5 CONCLUSION

Discriminate analysis is a powerful multivariate statistical tool used for classification and separation of groups based on several quantitative variables. Fisher's discriminate function provides an optimal linear combination of variables that maximizes the separation between populations. Using methods such as Mahalanobis distance, prior probabilities, and classification rules**,** it enables researchers to classify new observations with high accuracy. The technique is widely applicable in medical diagnosis, finance, quality control**,** biological studies, and social sciences**.** Overall, discriminant analysis gives a systematic and mathematically sound procedure for discriminating and classifying individuals into predefined groups.

10.6 SELF ASSESSMENT QUESTIONS:

1. Explain the procedure for classification into one of two multivariate normal (MVN) populations when the parameters (mean vectors and common dispersion matrix) are known.
2. Describe the method of classification into one of two multivariate normal (MVN) populations when the parameters are unknown and must be estimated from samples.
3. Discuss the classification procedure for two multivariate normal (MVN) populations when the dispersion matrices of the two populations are unequal.

10.7 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair & William C. Black
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein

Dr. A. Vasudeva Rao

LESSON -11

CLASSIFICATION WITH SEVERAL MVN POPULATIONS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of classification with several populations
- To know the concept of classification with several populations
- To acquire knowledge about the importance of classification with several populations
- To understand the purpose and objectives of classification with several populations.

STRUCTURE

11.1. INTRODUCTION

In multivariate statistical analysis, one of the major objectives is classifying an individual (observation) into one of several known populations. These populations may represent different groups such as disease categories, customer segments, manufacturing quality levels, or species classifications.

When each population is described by a p-variate distribution**,** usually the Multivariate Normal (MVN) distribution, classification rules are constructed to minimize misclassification.

For more than two populations, the classification problem becomes more complex because we must compute posterior probabilities, compare them across all populations, and assign the observation to the group with the minimum expected classification cost (ECM) or maximum posterior probability (TPM)**.**

If the parameters (means, covariance matrices, and prior probabilities) are known**,** classification rules are direct.

If parameters are unknown**,** they must be estimated from samples, leading to estimated versions of the Bayes allocation rules.

11.2 CLASSIFICATION WITH SEVERAL MVN POPULATIONS WITH COMMON DISPERSION MATRIX

We have 'g' multivariate normal populations

$$\boldsymbol{\pi}_1 : N_p(\underset{\sim}{\mu}_1, \boldsymbol{\Sigma})$$
$$\boldsymbol{\pi}_2 : N_p(\underset{\sim}{\mu}_2, \boldsymbol{\Sigma})$$
$$\vdots$$
$$\boldsymbol{\pi}_g : N_p(\underset{\sim}{\mu}_g, \boldsymbol{\Sigma})$$

Let $f_i(\underset{\sim}{\mathbf{x}})$ be the density associated with population $\boldsymbol{\pi}_i$,*i*=l,2..….g.

Let $p_i = the$ prior probability of population $\boldsymbol{\pi}_i, i = 1,2,...g$

Since, $\boldsymbol{\pi}_i, i = 1,2,...g$ are MVN populations, we have

$$f_i(\underset{\sim}{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}\right|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)} \qquad \text{for i=1,2,....,g}$$
$$\Rightarrow \log f_i(\underset{\sim}{\mathbf{x}}) = -\left(\frac{p}{2}\right)\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}((\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)) \tag{1}$$

Now, the linear discriminant scores are given by

$$d_i(\underset{\sim}{\mathbf{x}}) = \log f_i(\underset{\sim}{\mathbf{x}}) + \log p_i$$
$$= -\left(\frac{p}{2}\right)\log(2\pi) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}\right| - \frac{1}{2}\underset{\sim}{\mathbf{x}}'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} + \underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} - \frac{1}{2}\underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\boldsymbol{\mu}}_i + \log p_i \tag{2}$$

(after simplification)

The first three terms are same for $d_1(\underset{\sim}{\mathbf{x}}), d_2(\underset{\sim}{\mathbf{x}}), ...d_g(\underset{\sim}{\mathbf{x}})$ and consequently, they can be ignored for allocatory purposes. Now, the linear discriminant scores become

$$d_i(\underset{\sim}{\mathbf{x}}) = \underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\mathbf{x}} - \frac{1}{2}\underset{\sim}{\boldsymbol{\mu}}_i'\boldsymbol{\Sigma}^{-1}\underset{\sim}{\boldsymbol{\mu}}_i + \log p_i \tag{3}$$

The relevant sample quantities for population $\boldsymbol{\pi}_i$ are

$\underset{\sim}{\bar{\mathbf{x}}}_i$ = sample mean vector
S_i = sample covariance matrix and
n_i = sample size

and the pooled estimate of $\boldsymbol{\Sigma}$,

$$S = \frac{(n_1-1)S_1 + (n_2-1)S_2 + ... + (n_g-1)S_g}{n_1 + n_2 + ... + n_g - g} \tag{4}$$

Now, an estimate of $d_i(\underset{\sim}{\mathbf{x}})$ viz, $\hat{d}_i(\underset{\sim}{\mathbf{x}})$ is given by

$$\hat{d}_i(\underset{\sim}{\mathbf{x}}) = \underset{\sim}{\bar{\mathbf{x}}}_i' S^{-1}\underset{\sim}{\mathbf{x}} - \frac{1}{2}\underset{\sim}{\bar{\mathbf{x}}}_i' S^{-1}\underset{\sim}{\mathbf{x}} + \log p_i \tag{5}$$

consequently, the estimated minimum TPM rule for equal covariance normal populations is as follows:

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if the linear discriminant score $\hat{d}_k(\underset{\sim}{\mathbf{x}}) = \max_i \{\hat{d}_i(\underset{\sim}{\mathbf{x}})\}$ (6)

Where $\hat{d}_i(\underset{\sim}{\mathbf{x}})$ is given by (5)

NOTE:-

1. In the above minimum TPM rules , for any case , if $p_1 = p_2 = p_3 = ... = p_g = 1/g$, we may ignore those term $\log p_i$ is discriminant scores , as it is same for all discriminant scores. In this case the minimum TPM rule is reduced to ML rule in which case the allocation rules are same as above except ignoring $\log p_i$.
2. An equivalent classifier for common covariance matrix case can be obtained from (1) by ignoring the term

$$-\frac{1}{2}\log|\boldsymbol{\Sigma}| \text{ and is given by}$$

$$-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)+\log p_i$$

The classification rule with sample estimates instead for unknown populations quantities is given by Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$,if

$$-\frac{1}{2}D_k^2(\underset{\sim}{\mathbf{x}}) + \log p_k \quad \text{is} \quad \text{largest} \quad \text{for} \quad k=1,2,\ldots,g. \tag{7}$$

where $D_k^2(\underset{\sim}{\mathbf{x}}) = (\underset{\sim}{\mathbf{x}} - \bar{\underset{\sim}{\mathbf{x}}}_k)'S^{-1}(\underset{\sim}{\mathbf{x}} - \bar{\underset{\sim}{\mathbf{x}}}_k)$ is Mahalnobis squared distance between $\underset{\sim}{\mathbf{x}}$ and the sample mean $\bar{\underset{\sim}{\mathbf{x}}}_k$.

Thus , we see the rule (7) or equivalently rule (6) assigns $\underset{\sim}{\mathbf{x}}$ to the closest population (the distance in penalized by $\log p_i$).

3. In note(2) , if we assume $p_1, p_2, p_3, ..., p_g$ are equal and hence allocation rule may be significant as follows:

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$, if $-\frac{1}{2}D_k^2(\underset{\sim}{\mathbf{x}})$ is largest

Or equivalently $D_k^2(\underset{\sim}{\mathbf{x}})$ smallest (8)

In other words, we are allocating $\underset{\sim}{\mathbf{x}}$ to that population whose sample mean vector is closest to $\underset{\sim}{\mathbf{x}}$. This rule is also called as ML classification rule.

11.3 CLASSIFICATION AMONG SEVERAL MVN POPULATIONS WITH UNEQUAL DISPERSION MATRICES

We have 'g' multivariate normal populations

$$\begin{aligned}
&\boldsymbol{\pi}_1 : N_p(\underset{\sim}{\mu}_1, \boldsymbol{\Sigma}_1)\\
&\boldsymbol{\pi}_2 : N_p(\underset{\sim}{\mu}_2, \boldsymbol{\Sigma}_2)\\
&\vdots\\
&\boldsymbol{\pi}_g : N_p(\underset{\sim}{\mu}_g, \boldsymbol{\Sigma}_g)
\end{aligned}$$

Let $f_i(\underset{\sim}{\mathbf{x}})$ be the density associated with population $\boldsymbol{\pi}_i$, $i = 1, 2...g$

Let $p_i = the$ prior probability of population $\boldsymbol{\pi}_i, i = 1, 2, ...g$

Since, $\boldsymbol{\pi}_i, i = 1, 2, ...g$ are MVN populations, we have

$$f_i(\underset{\sim}{\mathbf{x}}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}}-\underset{\sim}{\boldsymbol{\mu}}_i)} \qquad \text{for i=1,2,....,g} \tag{1}$$

From Eq. (1) we have

$$\log(p_i f_i(\underset{\sim}{\mathbf{x}})) = \log p_i - \frac{p}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i) \tag{2}$$

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if

$$\begin{aligned}\log(p_k f_k(\underset{\sim}{\mathbf{x}})) &= \max_i \log(p_i f_i(\underset{\sim}{\mathbf{x}})) \\ &= \log p_k - \frac{p}{2}\log(2\boldsymbol{\pi}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_k\right| - \frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)\end{aligned} \tag{3}$$

The constant $p/2 \log(2\boldsymbol{\pi})$ can be ignored in (2) since it is same for all populations. We therefore define the quadratic discrimination score for i^{th} population is

$$d_i^Q(\underset{\sim}{\mathbf{x}}) = \log p_i - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_i\right| - \frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i)'\boldsymbol{\Sigma}_i^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}}_i) \quad \text{for i=1,2,...,g} \tag{4}$$

The quadratic score $d_i^Q(\underset{\sim}{\mathbf{x}})$ is composed contributions from the generalized variance $\left|\boldsymbol{\Sigma}_i\right|$, the prior probability p_i, and Mahalnobis (or statistical) squared distance between $\underset{\sim}{\mathbf{x}}$ and population mean μ_i.

Using discriminant scores the classification rule (4) becomes

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$

$$\text{The quadratic score } d_k^Q(\underset{\sim}{\mathbf{x}}) = \max_i \{d_i^Q(\underset{\sim}{\mathbf{x}})\} \tag{5}$$

where $d_i^Q(\underset{\sim}{\mathbf{x}})$ is given by (4).

In practice, the $\underset{\sim}{\boldsymbol{\mu}}_i$ and $\boldsymbol{\Sigma}_i$ are unknown and hence a training set of correctly classified observations is often available for the construction of estimates. The relevant sample quantities for population $\boldsymbol{\pi}_i$ are

$\underset{\sim}{\bar{\mathbf{x}}}_i$ = sample mean vector
S_i = sample covariance matrix and
n_i = sample size

Using the above estimation in (4) , we get the estimate of the quadratic discriminant score $\hat{d}_i^Q(\underset{\sim}{\mathbf{x}})$ as

$$\hat{d}_i^Q(\underset{\sim}{\mathbf{x}}) = \log p_i - \frac{1}{2}\log|S_i| - \frac{1}{2}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\bar{\mathbf{x}}}_i)' S_i^{-1}(\underset{\sim}{\mathbf{x}} - \underset{\sim}{\bar{\mathbf{x}}}_i) \tag{6}$$

and the classification rule based on the sample is as follows:

Allocate $\underset{\sim}{\mathbf{x}}$ to $\boldsymbol{\pi}_k$ if the quadratic score

$$\hat{d}_k^Q(\underset{\sim}{\mathbf{x}}) = \max_i \{\hat{d}_i^Q(\underset{\sim}{\mathbf{x}})\} \tag{7}$$

where $\hat{d}_i^Q(\underset{\sim}{\mathbf{x}})$ is given by (6) for $i = 1, 2, ..., g$.

NOTE:-

(1). In the above minimum TPM rules , for any case , if $p_1 = p_2 = p_3 = ... = p_g = 1/g$, we may ignore those term $\log p_i$ is discriminant scores , as it is same for all discriminant scores. In this case the minimum TPM rule is reduced to ML rule in which case the allocation rules are same as above except ignoring $\log p_i$.

11.4 CONCLUSION

In this unit, we examined several important methods for classifying multivariate observations into populations. Fisher's Linear Discriminant Function provides a powerful approach for separating two populations by transforming multivariate data into a single discriminating variable. This method does not require normality, but it implicitly assumes equal covariance matrices**.**

For more than two populations, two general decision-theoretic approaches were discussed: the Minimum Total Probability of Misclassification (TPM) rule and the Minimum Expected Cost of Misclassification (ECM) rule. The TPM rule focuses on minimizing overall misclassification probability, whereas ECM incorporates prior probabilities and misclassification costs, making it more flexible and realistic for practical applications.
When the populations follow multivariate normal distributions, classification rules become more explicit through quadratic **or** linear discriminant scores.

- With unequal covariance matrices, we derive the quadratic discriminant function**.**
- With equal covariance matrices, the rule simplifies to a linear discriminant function**,** which corresponds to the Bayes rule**,** ML rule**,** and in special cases **to** Fisher's discriminant**.**

11.5 SELF ASSESSMENT QUESTIONS:

1. Explain the problem of classification into one of the two known multivariate normal populations
2. Describe the method of classification of an individual into one of several p-variate normal populations having a common dispersion matrix. ξ where all the parameters are known.

11.6 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein

Dr. Syed Jilani

LESSON -12

FISHERS LINEAR DISCRIMINANT ANALYSIS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of Fishers linear discriminate analysis
- To know the concept of Fishers linear discriminate analysis
- To acquire knowledge about significance of Fishers linear discriminate analysis

STRUCTURE

12.1. INTRODUCTION

In multivariate statistical analysis, one of the central problems is the classification (discrimination) of an observation into one of several known populations. This problem arises frequently in practice—for example, assigning patients to diagnostic groups based on medical measurements, classifying credit applicants as low-risk or high-risk based on financial indicators, or determining the origin of agricultural products using chemical characteristics.

When the probability distributions of the populations are fully specified—that is, the functional form of the distribution and all of its parameters (means, variances, and covariances) are known—statistical theory provides optimal decision rules for classification. This setting represents the "ideal" or theoretical case**,** as in practice parameters are usually estimated from data. Nonetheless, studying this case is fundamental because it provides the benchmark for performance and forms the basis for practical extensions.

One of the earliest and most influential approaches to this problem was proposed by Sir Ronald A. Fisher (1936) in his seminal work on linear discriminate analysis (LDA)**.** Fisher's method aims to construct a linear discriminate function**,** i.e., a linear combination of the observed variables, such that the separation among populations is maximized.

The method works by finding the projection (linear function) of the multivariate data that best separates the groups relative to the within-group variability. For two populations, this reduces to Fisher's linear discriminate function; for more than two populations, it leads to a set of discriminate functions that can be used for classification.

12.2 FISHERS DISCRIMINANT FUNCTION-SEPARATION OF TWO POPULATIONS (NOT NECESSARY MULTIVARIATE NORMAL)

Fishers idea was to transform the multivariate observations $\underset{\sim}{\mathbf{x}}$'s to univariate observation y's such that the y's derived from population $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ were separated as much as possible . Fisher suggested taking linear combination of $\underset{\sim}{\mathbf{x}}$'s to create y's because they are simple function of $\underset{\sim}{\mathbf{x}}$ and are easily handled mathematically .

Fisher's approach does not assume that the populations are normal.

If does, however , implicitly assume the population covariance matrices are equal because a pooled estimate of the common covariance matrix is used.

Let $\underset{\sim}{\mathbf{x}}_{11},\underset{\sim}{\mathbf{x}}_{12},...,\underset{\sim}{\mathbf{x}}_{1n_1}$ be a random sample of size n_1 from population $\boldsymbol{\pi}_1$ and let $\underset{\sim}{\mathbf{x}}_{21},\underset{\sim}{\mathbf{x}}_{22},...,\underset{\sim}{\mathbf{x}}_{2n_2}$ be a random sample of size n_2 from population $\boldsymbol{\pi}_2$ .Now

$\overline{\underset{\sim}{\mathbf{x}}}_1$ be the mean of 1st sample

S_1 be the sample covariance matrix of 1st sample

$\overline{\underset{\sim}{\mathbf{x}}}_2$ be the mean of 2nd sample

S_2 be the sample covariance matrix of 2nd sample

Denote $S = \dfrac{(n-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$ (1)

Which is a pooled sampled covariance matrix.

Now, Fisher's idea is as follows

Consider the linear combination

$y = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}$, when $\underset{\sim}{\mathbf{w}}$ is $|\mathbf{x}|$ vector of real number (2)

using the linear transformation, the multivariate observation of 1st sample will be transformed into univariate observations given by

$y_{11},y_{12},...,y_{1n_1}$

when $y_{1i} = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}_{1i}, i = 1,2,...,n_1$

similarly the second sample $\underset{\sim}{\mathbf{x}}_{21},\underset{\sim}{\mathbf{x}}_{22},...,\underset{\sim}{\mathbf{x}}_{2n_2}$ will be transformed into

$y_{21},y_{22},...,y_{2n_2}$

when $y_{2i} = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}_{2i}, i = 1,2,...,n_2$

Now $\overline{y}_1 = \underset{\sim}{\mathbf{w}}'\overline{\underset{\sim}{\mathbf{x}}}_1$

$\overline{y}_2 = \underset{\sim}{\mathbf{w}}'\overline{\underset{\sim}{\mathbf{x}}}_2$

and $s_y^2 = \dfrac{\sum_{i=1}^{n_1}(y_{1i}-\bar{y}_1)^2 + \sum_{i=1}^{n_2}(y_{2i}-\bar{y}_2)^2}{n_1+n_2-2}$ (3)

consider

$$\text{separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y} \quad (4)$$

Now, Fisher's idea is to select the linear combination $\underset{\sim}{\mathbf{w}}$ such that the separation given in (4) is maximum. In other words, the objective is to select the linear combination of $\underset{\sim}{\mathbf{x}}$ (i.e. $\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}$) to achieve maximum separation between the sample means $\bar{y}_1$ & $\bar{y}_2$. Equation (4) may be written as

$$separartion^2 = \frac{(\text{squared distance between sample mean } \bar{y}_1 \text{ and } \bar{y}_2)}{\text{pooled sample variance of y}}$$

$$= \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{\left[\underset{\sim}{\mathbf{w}}'(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)\right]^2}{s_y^2} \qquad \text{(from (3))}$$

but $s_y^2 = \underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}$ (from (3) & (1))

$$\therefore \frac{\text{squared distance between } \bar{y}_1 \text{ \& } \bar{y}_2}{\text{pooled variance of y}} = \frac{(\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}})^2}{\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}} = \phi \text{ say} \quad (5)$$

where $\underset{\sim}{\mathbf{d}} = (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)$

Now, as per Fisher's idea, (5) has to be maximized w.r.t. $\underset{\sim}{\mathbf{w}}$.

Which implies

$$\frac{\partial \phi}{\partial \underset{\sim}{\mathbf{w}}} = \frac{(\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}) \frac{\partial}{\partial \underset{\sim}{\mathbf{w}}} (\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}})^2 - (\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}})^2 \frac{\partial}{\partial \underset{\sim}{\mathbf{w}}} (\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}})}{\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}} = 0$$

$$\Rightarrow (\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}) 2 (\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}}).\underset{\sim}{\mathbf{d}} - 2(\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}})^2 S \underset{\sim}{\mathbf{w}} = 0$$

$$\Rightarrow (\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}) \underset{\sim}{\mathbf{d}} - (\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}}) S \underset{\sim}{\mathbf{w}} = 0$$

$$\Rightarrow \underset{\sim}{\mathbf{w}} = \frac{(\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}})}{(\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}})} S^{-1} \underset{\sim}{\mathbf{d}} \quad (\because \text{ S is positive defined matrix}) \quad (6)$$

$$= CS^{-1}\underset{\sim}{\mathbf{d}}$$

where $C = \dfrac{\underset{\sim}{\mathbf{w}}' S \underset{\sim}{\mathbf{w}}}{\underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{d}}}$ and C is ratio of two scalars thus $\underset{\sim}{\mathbf{w}}$ is a scalar multiplier of the vector $S^{-1}\underset{\sim}{\mathbf{d}}$.

Using

$\underset{\sim}{\mathbf{w}} = CS^{-1}\underset{\sim}{\mathbf{d}}$ in (5) we get

$$\phi = \frac{C^2(\underset{\sim}{\mathbf{d}}'S^{-1}\underset{\sim}{\mathbf{d}})^2}{C^2\underset{\sim}{\mathbf{d}}'S^{-1}SS^{-1}\underset{\sim}{\mathbf{d}}} \qquad (7)$$

$$= \underset{\sim}{\mathbf{d}}'S^{-1}\underset{\sim}{\mathbf{d}}$$

Now using $\underset{\sim}{\mathbf{w}} = CS^{-1}\underset{\sim}{\mathbf{d}}$ in (5) we get

$$\phi = \underset{\sim}{\mathbf{d}}'S^{-1}\underset{\sim}{\mathbf{d}} \qquad (8)$$

Thus, from (7) & (8), we can see that for either

$$\underset{\sim}{\mathbf{w}} = CS^{-1}\underset{\sim}{\mathbf{d}} \text{ or } \underset{\sim}{\mathbf{w}} = S^{-1}\underset{\sim}{\mathbf{d}}$$

the same ratio ϕ, we are setting . Thus ϕ will be maximized if we take

$$\underset{\sim}{\mathbf{w}} = S^{-1}\underset{\sim}{\mathbf{d}} = S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (9)$$

and the maximum value of ϕ is

$$\phi_m = \underset{\sim}{\mathbf{d}}'S^{-1}\underset{\sim}{\mathbf{d}}$$
$$= (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)'S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (10)$$
$$= D^2 \quad \text{(say)}$$

Now, the linear function

$$Y = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{X}}$$
$$= (\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)'S^{-1}\underset{\sim}{\mathbf{X}} \qquad \text{(from (2) \& (9))} \qquad (11)$$

is called as Fisher's linear discriminant function . and the maximum ratio D^2,where D^2 given by (10),is called the sample squared distance or squared Mahalnobis distance between sample means $\underset{\sim}{\bar{\mathbf{x}}}_1$ and $\underset{\sim}{\bar{\mathbf{x}}}_2$.

The linear discriminant function given by (11) converts the two multivariate samples into two univariate samples such that the corresponding univariate sample means are sepearted as much as possible to the relative to pooled sample variance .
We can employ (11) as a classification device as given below.

12.2.1 an allocation rule based on fisher's discriminant function:

We have the Fisher's linear discriminant function

$$y = \underset{\sim}{\mathbf{w}}'\underset{\sim}{\mathbf{x}}, \qquad \text{where } \underset{\sim}{\mathbf{w}} = S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (1)$$

Let 'm' be the midpoint between $\bar{y}_1$ and $\bar{y}_2$ and is given by

$$m = (\bar{y}_1 + \bar{y}_2)/2$$
$$= 1/2\,(\underset{\sim}{\bar{\mathbf{x}}}_1 - \underset{\sim}{\bar{\mathbf{x}}}_2)'S^{-1}(\underset{\sim}{\bar{\mathbf{x}}}_1 + \underset{\sim}{\bar{\mathbf{x}}}_2) \qquad (2)$$

Now , the allocation rule or classification rule based on Fisher's discriminant function is as follows:

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_1$,if

$$y_0 = (\bar{\underset{\sim}{\mathbf{x}}}_1 - \bar{\underset{\sim}{\mathbf{x}}}_2)' S^{-1} \underset{\sim}{\mathbf{x}}_0 \geq m \text{ or } y_0 - m \geq 0$$

Allocate $\underset{\sim}{\mathbf{x}}_0$ to $\boldsymbol{\pi}_2$,if

$$y_0 < m \text{ or } y_0 - m < 0 \qquad (3)$$

NOTE:

(1). If $\boldsymbol{\pi}_1 \sim \underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma}$ and $\boldsymbol{\pi}_2 \sim \underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma}$ then the Mahalnobis distance between $\underset{\sim}{\boldsymbol{\mu}}_1$ and $\underset{\sim}{\boldsymbol{\mu}}_2$ is denoted by $\Delta_{\underset{\sim}{\boldsymbol{\mu}}_1, \underset{\sim}{\boldsymbol{\mu}}_2}$ and is given by

$$\Delta^2_{\underset{\sim}{\boldsymbol{\mu}}_1, \underset{\sim}{\boldsymbol{\mu}}_2} = (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2)$$

(2). $\Delta^2_{\underset{\sim}{\mathbf{x}}, \underset{\sim}{\boldsymbol{\mu}}} = (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})' \boldsymbol{\Sigma}^{-1} (\underset{\sim}{\mathbf{x}} - \underset{\sim}{\boldsymbol{\mu}})$

(3). Mahalnobis D^2 test statistic to test separation between $\boldsymbol{\pi}_1$ *and* $\boldsymbol{\pi}_2$ (or)

$H_0 : \underset{\sim}{\boldsymbol{\mu}}_1 = \underset{\sim}{\boldsymbol{\mu}}_2$ vs $H_1 : \underset{\sim}{\boldsymbol{\mu}}_1 \neq \underset{\sim}{\boldsymbol{\mu}}_2$

suppose $\boldsymbol{\pi}_1 : N_p(\underset{\sim}{\boldsymbol{\mu}}_1, \boldsymbol{\Sigma})$ and $\boldsymbol{\pi}_2 : N_p(\underset{\sim}{\boldsymbol{\mu}}_2, \boldsymbol{\Sigma})$ $\bar{\underset{\sim}{\mathbf{x}}}_1, S_1$ are the sample mean and sample covariance matrix of a sample drawn from $\boldsymbol{\pi}_1$ and $\bar{\underset{\sim}{\mathbf{x}}}_2, S_2$ are …$\boldsymbol{\pi}_2$.

Now Mahalnobis D^2 test statistic is given by

$$D^2 = (\bar{\underset{\sim}{\mathbf{x}}}_1 - \bar{\underset{\sim}{\mathbf{x}}}_2)' S^{-1} (\bar{\underset{\sim}{\mathbf{x}}}_1 - \bar{\underset{\sim}{\mathbf{x}}}_2)$$

where $S = \dfrac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$

under $H_0 : \left(\dfrac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p}\right)\left(\dfrac{n_1 n_2}{n_1 + n_2}\right) D^2 \sim F_{p, n_1 + n_2 - p - 1}$

which can be used as for testing the significant difference $\underset{\sim}{\boldsymbol{\mu}}_1 - \underset{\sim}{\boldsymbol{\mu}}_2$. If H_0 is rejected , we can conclude that the separation between the two populations $\boldsymbol{\pi}_1$ *and* $\boldsymbol{\pi}_2$ is significant.

(4). Two sample T^2 and Mahalnobis D^2 are closely associated as

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right) D^2$$

(5). In case of two normal populations with common covariance matrix, Fisher's

method is corresponds to a particular case of minimum ECM rule with equal prior probabilities and equal costs of TPM rule with equal prior probabilities . Further, it is same as ML rule .

(6). The expression in minimum ECM rule for two multivariate normal populations $w = (\bar{\underset{\sim}{x}}_1 - \bar{\underset{\sim}{x}}_2)' S^{-1} (\underset{\sim}{x} - 1/2(\bar{\underset{\sim}{x}}_1 + \bar{\underset{\sim}{x}}_2))$ is frequently called Anderson's classification.

(7). Fisher's method is also a special case of allocation rule based on Bayeson posterior probabilities when the prior probabilities p_1 and p_2 are same for the case of two multivariate normal populations.

12.3 FISHER'S METHOD FOR DISCRIMINATING AMONG SEVERAL POPULATIONS WHEN PARAMETERS ARE SPECIFIED

Fisher also proposed a several population extension of his discriminant method, which was discussed for the case of two populations. The motivation behind the Fisher discriminant analysis is the need to obtain a reasonable representation of the population that involves only a few linear combinations of the observations, such as $\underset{\sim}{l}_1' \underset{\sim}{x}, \underset{\sim}{l}_2' \underset{\sim}{x}$ and so on. His approach has several advantages and one is interested in separating several populations for

1) Visual inspection or
2) Graphical descriptive purposes.

It allows for the follows:-

1. Convenient representation of the g populations that reduce the dimension from a very large number of characteristics to a relatively few linear combinations. Of course, some information – needed for optimal classification- may be lost unless the population means lie completely in the lower dimensional space selected.

2. Plotting of the means of the first two or three linear combinations (discriminates). This helps display the relationship and possible groupings of the populations.

3. Scatter plots of the sample values of the first two discriminates, which can indicate outliers or other abnormalities in the data.

The primary purpose of Fishers Discriminant analysis is to separate populations. However, it can also be used to classify a new observation into one of the populations. It is not necessary to assume that the g populations are multivariate normal. However we assume the population covariance matrices are equal and of full rank. That is $\Sigma_1 = \Sigma_2 = - - - - = \Sigma_g = \Sigma$. Thus, we have g populations with mean vectors $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots\ldots \underset{\sim}{\mu}_g$ and common covariance matrix $\sum$.

Let $\bar{\underset{\sim}{\mu}} = \frac{1}{g} \sum_{i=1}^{g} \underset{\sim}{\mu}_i$

and $B = \sum_{i=1}^{g} (\underset{\sim}{\mu_i} - \underset{\sim}{\bar{\mu}})(\underset{\sim}{\mu_i} - \underset{\sim}{\bar{\mu}})'$ (1)

we consider the linear combination $y = \underset{\sim}{l}\underset{\sim}{x}$

which has expected value

$$E(y) = \underset{\sim}{l}' E(\underset{\sim}{X}/\pi_i) = \underset{\sim}{l}' \underset{\sim}{\mu}_i \text{ (for population } \pi_i)$$

$$= \mu_{iy} \text{ (say)}$$

and variance $V(y) = \underset{\sim}{l}' cov(\underset{\sim}{X}, \underset{\sim}{X}') \underset{\sim}{l}$

$$= \underset{\sim}{l}' \Sigma \underset{\sim}{l} = \sigma_y^2 \text{ for all populations.} \quad (2)$$

we defuse the overall mean,

$$\bar{\mu}_y = \frac{1}{g}\sum_{i=1}^{g} \mu_{iy} = \frac{1}{g}\sum_{i=1}^{g} \underset{\sim}{l}' \underset{\sim}{\mu_i}$$

$$= \underset{\sim}{l}' \left(\frac{1}{g}\sum_{i=1}^{g} \underset{\sim}{\mu_i}\right)$$

$$= \underset{\sim}{l}' \underset{\sim}{\bar{\mu}} \quad \text{(From (1))} \quad (3)$$

and form the ratio

$$\frac{\text{sum of squared distances from populations to over all mean of Y}}{\text{common population variance of Y}}$$

$$= \frac{\sum_{i=1}^{g} (\mu_{iy} - \bar{\mu}_y)^2}{\sigma_y^2}$$

$$= \frac{\sum_{i=1}^{g} (\underset{\sim}{l}' \underset{\sim}{\mu_i} - \underset{\sim}{l}' \underset{\sim}{\bar{\mu}})^2}{\underset{\sim}{l}' \sum \underset{\sim}{l}}$$

$$= \frac{\underset{\sim}{l}' \sum_{i=1}^{g} (\underset{\sim}{\mu_i} - \underset{\sim}{\bar{\mu}})(\underset{\sim}{\mu_i} - \underset{\sim}{\bar{\mu}})' \underset{\sim}{l}}{\underset{\sim}{l}' \sum \underset{\sim}{l}}$$

$$= \frac{\underset{\sim}{l}' B \underset{\sim}{l}}{\underset{\sim}{l}' \sum \underset{\sim}{l}} \quad \text{(from (1))}$$

Thus $$\frac{\sum_{i=1}^{g} (\mu_{iy} - \bar{\mu}_y)^2}{\sigma_y^2} = \frac{\underset{\sim}{l}' B \underset{\sim}{l}}{\underset{\sim}{l}' \sum \underset{\sim}{l}} \quad (4)$$

The ratio (4) measures the variability between the groups of Y- values relative to the common variability within the groups. We can then choose $\underset{\sim}{l}$ to maximize the ratio (4) Thus if we write

$$\lambda = \frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\Sigma\underset{\sim}{l}} \tag{5}$$

Then we have to maximize (5) with respect to $\underset{\sim}{l}$ when implies

$$\frac{\partial\lambda}{\partial\underset{\sim}{l}} = 0 \Rightarrow (\underset{\sim}{l}'\Sigma\underset{\sim}{l})\frac{\partial\underset{\sim}{l}'}{\partial\underset{\sim}{l}}B\underset{\sim}{l} - (\underset{\sim}{l}'B\underset{\sim}{l})\frac{\partial\underset{\sim}{l}'}{\partial\underset{\sim}{l}}\Sigma\underset{\sim}{l} = 0$$

$$\Rightarrow (\underset{\sim}{l}'\Sigma\underset{\sim}{l})B\underset{\sim}{l} - (\underset{\sim}{l}'B\underset{\sim}{l})\Sigma\underset{\sim}{l} = 0$$

$$\Rightarrow \quad B\underset{\sim}{l} - (\frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\Sigma\underset{\sim}{l}})\,\Sigma\underset{\sim}{l} = 0$$

$$\Rightarrow \Sigma^{-1}B\underset{\sim}{l} - (\frac{\underset{\sim}{l}'B\underset{\sim}{l}}{\underset{\sim}{l}'\Sigma\underset{\sim}{l}})\underset{\sim}{l} = 0$$

$$\Rightarrow (\Sigma^{-1}B - \lambda I)\underset{\sim}{l} = 0 \quad \text{(using (5))} \tag{6}$$

Thus $\underset{\sim}{l}$ is the latent vector corresponding to a latent root λ of $\Sigma^{-1}B$. As, we are seeking for a $\underset{\sim}{l}$ which maximizes λ, let λ_1 be the non zero largest latent root of $\Sigma^{-1}B$ and $\underset{\sim}{l}_1$ be the corresponding latent vector. Now, the linear combination, $Y_1 = \underset{\sim}{l}_1'\underset{\sim}{X}$ is called Fisher's first linear discriminant .

Similarly if λ_2 is the next non Zero largest latent root of $\Sigma^{-1}B$ and $\underset{\sim}{l}_2$ correspondent latent vector then , $Y_2 = \underset{\sim}{l}_2'\underset{\sim}{X}$ is Fisher's second linear discriminant.

Let $\lambda_1 > \lambda_2 > ..... > \lambda_s > 0$ denote the $s \le \min(g-1,p)$ non zero eigen values of $\Sigma^{-1}B$ and let $\underset{\sim}{l}_1, \underset{\sim}{l}_2, ... \underset{\sim}{l}_s$ be the corresponding latent vectors. Now, the linear combinations

$$Y_k = \underset{\sim}{l}_k'\underset{\sim}{X}(k \le s) \tag{7}$$

is Fisher's k^{th} linear discriminant.

12.4 FISHERS METHOD FOR DISCRIMINATING SEVERAL POPULATIONS WHEN PARAMETERS ARE UNKNOWN

Fisher's sample linear discriminants:

In general, $\sum$ and the $\underset{\sim}{\mu_i'}s$ are unknown, but we have a training set consisting of correctly classified observations. Suppose the training set consist of a random sample of size n_i from population $\pi_i, i = 1,2,3.....g$

Let $\underset{\sim}{x_i'}$ be the mean vector and S_i be the covariance matrix of ith sample. Now denote the sample between groups matrix.

$$B_0 = \sum_{i=1}^{g} (\underset{\sim}{\bar{x}_i} - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}_i} - \underset{\sim}{\bar{x}})'$$

Where, $\underset{\sim}{\bar{x}} = \frac{1}{g}\sum_{i=1}^{g} \underset{\sim}{\bar{x}_i}$. (8)

B_0 is an estimate of B

Also, an estimate of $\sum$ is based on the sample within groups matrix is

$$W = \sum_{i=1}^{g} (n_i - 1)S_i \quad (9)$$

Consequently, $S_p = \frac{W}{(n-g)}, n = \sum_{i=1}^{g} n_i$. (10)

is an estimate of $\sum$.

We consider the linear transformation,

$$y = \underset{\sim}{l'x} \quad (11)$$

Under the linear transformation, (11) the given multi variate samples can be transformed into univariate samples whose means and variances are given by

Means : $\bar{y}_1, \bar{y}_{2,}.....\bar{y}_g$

Variances: $s_{y_1}^2, s_{y_2}^2, ...s_{y_g}^2$

We denote the overall sample as

$$\bar{y} = \frac{1}{g}\sum_{i=1}^{g} \bar{y}_i \quad (12)$$

Now form the ratio,

$\lambda =$ sum of squared distances fro sample means to overall mean

Total within samples variation

$$= \frac{\sum_{i=1}^{g}(\bar{y}_i - \bar{y})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2}$$

$$= \frac{\sum_{i=1}^{g}(\underset{\sim}{l}'\underset{\sim}{\bar{x}}_i - \underset{\sim}{l}'\underset{\sim}{\bar{x}})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{l}'\underset{\sim}{x}_{ij} - \underset{\sim}{l}'\underset{\sim}{\bar{x}}_i)^2} \qquad \text{(from (11))}$$

$$= \frac{\underset{\sim}{l}'\sum_{i=1}^{g}(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})'\underset{\sim}{l}}{\underset{\sim}{l}'\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}}_j)(\underset{\sim}{x}_{ij} - \underset{\sim}{x}_i)'\underset{\sim}{l}}$$

$$= \frac{\underset{\sim}{l}'\sum_{i=1}^{g}(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}})(\underset{\sim}{\bar{x}}_i - \underset{\sim}{\bar{x}}_i)'\underset{\sim}{l}}{\underset{\sim}{l}'\sum_{i=1}^{g}(n_i - 1)S_i\underset{\sim}{l}}$$

(By using the definition of sample covariance matrix)

$$= \frac{\underset{\sim}{l}'B_0\underset{\sim}{l}}{\underset{\sim}{l}'W\underset{\sim}{l}} \qquad \text{(from (8) \& (9))} \qquad (13)$$

The ratio (13) measures the variability between the groups of g values relative to the total variability within the groups.

Now, Fisher suggested to choose $\underset{\sim}{l}$ such that λ given by (13) is maximum, Maximization of λ with respect to $\underset{\sim}{l}$ implies.

$$\frac{\partial\lambda}{\partial\underset{\sim}{l}} = 0 \Rightarrow (W^{-1}B_0 - \lambda I)\underset{\sim}{l} = 0 \qquad (14)$$

(See (6) of page 29 for derivation particulars)

Now, if we denote $\lambda_1 \geq \lambda_2 \geq ..... \geq \lambda_s > 0$ (where $s = \min(g-1, p)$) are s eigen values of (14) and let $\underset{\sim}{l}_1, \underset{\sim}{l}_2, ... \underset{\sim}{l}_s$ the corresponding eigen vectors, then

Fisher's K-th sample linear discriminate is given by

$$y_k = \underset{\sim}{l}_k'\underset{\sim}{x}(k \leq s)$$

Thus, Fisher's sample linear discriminates are eigen vectors of $W^{-1}B_0$,

Where B_0 and W are as assigned in (8) & (9).

Note:

1. If sample means $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots\ldots\ldots \underset{\sim}{x}_g$ and sample covariance matrix $S_1, S_2, \ldots\ldots\ldots S_g$ are given , then B_0 and W can be completed using (8) and (9) respectively.
2. If raw samples from g populations are given, then B_0 and W can be computed as follows:

First compute the individual sample covariance matrices $S_1, S_2, \ldots\ldots\ldots S_g$ from the given samples and then use (9) to compute W. Now, compute the sample covariance matrix S from the combined samples of g samples given by

$$S = \frac{1}{n-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})(\underset{\sim}{x}_{ij} - \underset{\sim}{\bar{x}})' \quad \text{,when } n = \sum_{i=1}^{g} n_i \, .$$

Now, B_0 can be computed from the following relationship.

$$(n-1)S = W + B_0 \quad .$$

3. It may be noted that the pooled sample covariance matrix S_p and combined sample covariance matrix S are connected by the

$$(n-1)S = W + B_0$$

Thus, if the individual sample covariance matrix $S_1, S_2, \ldots\ldots\ldots S_g$ and the combined sample covariance matrix S are given, then one can obtain W and B_0 can be obtained as follows

$$W = (n_1 - 1)S_1 + (n_2 - 1)S_2 + \ldots\ldots\ldots + (n_g - 1)Sg$$
$$B_0 = (n-1)S - W$$

Now, Fisher's discriminates can be constructed using the eigen vectors of $W^{-1}B_0$.

4. We know that $W^{-1}B_0$ is not a symmetric matrix. Many computer

Packages can compute eigen values and eigen vectors only for the symmetric matrices. However, the eigen vectors of $W^{-1}B_0$ can be computed as follows:

Suppose, λ is a ch root and $\underset{\sim}{l}$ is Ch. Vector of $W^{-1}B_0$,then we have

$$(W^{-1}B_0 - \lambda I)\underset{\sim}{l} = 0$$

The above equations may be rewritten as

$$(W^{-1}B_0 - \lambda W^{-\frac{1}{2}}W^{-\frac{1}{2}})\underset{\sim}{l} = 0$$
$$\Rightarrow (W^{-\frac{1}{2}}B_0 - \lambda W^{\frac{1}{2}})\underset{\sim}{l} = 0$$
$$\Rightarrow (W^{-\frac{1}{2}}B_0W^{-\frac{1}{2}} - \lambda)W^{\frac{1}{2}}\underset{\sim}{l} = 0$$
$$\Rightarrow (W^{-\frac{1}{2}}B_0W^{-\frac{1}{2}} - \lambda)\underset{\sim}{w} = 0$$

($W^{\frac{1}{2}}$ is square root of W and $W^{-\frac{1}{2}}$ is a inverse of $W^{\frac{1}{2}}$).

Where, $\underset{\sim}{w} = W^{\frac{1}{2}}\underset{\sim}{l}$ or $\underset{\sim}{l} = W^{-\frac{1}{2}}\underset{\sim}{w}$.

Thus if $\underset{\sim}{w}$ is latent vector of the matrix, $W^{-\frac{1}{2}}B_0W^{-\frac{1}{2}}$. Corresponding to the latent root of λ, then latest vector $\underset{\sim}{l}$ of $W^{-\frac{1}{2}}B_0$ corresponding root λ may be obtained as $\underset{\sim}{l} = W^{-\frac{1}{2}}\underset{\sim}{w}$

For all practical purposes, for the construction of Fisher's discriminant functions we use the above method.

Classification of a new observation among several populations using Fisher's discriminants

Mainly, Fisher's discriminates were derived for the purpose of obtaining a low dimensional representation of the data that separate the populations as much as possible. Although they were derived from separatory considerations, the discriminates also provide the basis for a classification rule.

12.5 CONCLUSION

We extended Fisher's discriminate analysis from two populations to the more general case of several (g) populations**,** both when the population parameters are known and when they are unknown and must be estimated from sample data.

Fisher's central idea remains the same:
to find linear combinations of the original variables that maximize the separation among population means relative to the within-population variability**.**

When the parameters are fully specified, the discriminate functions are obtained through the Eigen value–eigenvector decomposition of the matrix $\Sigma^{-1}B$, where

- B represents between-group variability, and
- Σ represents the common within-group covariance matrix**.**

The eigenvectors corresponding to the largest eigenvalues provide the Fisher's discriminant functions**.** These linear combinations reduce dimensionality and allow clear visual separation of the groups.

When parameters are unknown, they are replaced by sample-based estimates:

- B is estimated using sample means (between-group matrix), and
- W is estimated using pooled within-group sample covariance matrices**.**

The discriminate functions obtained in this case are called Fisher's sample discriminants**,** and they are the eigenvectors of $W^{-1}B$. These functions not only help visualize differences between populations but also serve as the basis for classification rules, allowing new observations to be assigned to the population whose discriminant scores they most closely match.

12.6 SELF ASSESSMENT QUESTIONS:

1. Explain Fisher's method for discriminating among several populations when parameters are specified
2. Explain Fishers method for discriminating several populations when parameters are Unknown

12.7 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein

Dr. Syed Jilani

LESSON -13
CLUSTER ANALYSIS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of Cluster analysis
- To know the concept of Cluster analysis
- To acquire knowledge about significance of Cluster analysis

STRUCTURE

13.1. INTRODUCTION

Cluster Analysis is an unsupervised multivariate statistical technique used to group a set of objects (observations, variables, or cases) into clusters such that:

- Objects within the same cluster are highly similar
- Objects from different clusters are dissimilar

The goal is to uncover the natural structure **or** pattern present in multivariate data without using any prior group labels**.**

Cluster analysis is widely used in:

- Data mining
- Marketing segmentation
- Bioinformatics
- Image recognition
- Pattern classification
- Social sciences and medical research

Cluster analysis relies on similarity (or dissimilarity) measures, which quantify how close or far apart two observations are. The most commonly used measures are based on distance**.**

In multivariate analysis, the concept of similarity or dissimilarity plays a central role in understanding relationships among objects. When observations are described by several

variables, the distance between them indicates how *close* or *far* they are in multidimensional space. These distance measures form the foundation of several multivariate techniques such as cluster analysis**,** multidimensional scaling (MDS)**,** discriminant analysis**,** and nearest-neighbour classification**.**

A similarity measure quantifies how alike two objects are. In most applications, this is expressed in terms of a distance, where:

- Small distance → high similarity
- Large distance → low similarity

Different measures capture different aspects of variation, and choosing the appropriate distance metric is crucial for accurate data analysis.

The most commonly used distance-based similarity measures include the Squared Euclidean Distance**,** Chebyshev Distance**,** and Minkowski Distance, each with its mathematical form and geometric interpretation.

13.2 SIMILARITY MEASURES

Cluster analysis groups objects so that objects within the same cluster are similar and objects in different clusters are dissimilar. To do this, we use similarity or distance measures.

- Squared Euclidean Distance
- Chebyshev Distance
- Minkowski Distance

Distance, such as the Euclidean distance, is a dissimilarity measure and has some well known properties:

1. $d(p, q) \geq 0$ for all p and q, and $d(p, q) = 0$ if and only if $p = q$,

2. $d(p, q) = d(q,p)$ for all p and q,

3. $d(p, r) \leq d(p, q) + d(q, r)$ for all p, q, and r, where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

A distance that satisfies these properties is called a metric**.** Following is a list of several common distance measures to compare multivariate data. We will assume that the attributes are all continuous.

13.2.1 Squared Euclidean Distance:

The Squared Euclidean Distance (SED) between two points (objects) in a p-dimensional space is defined as the sum of the squared differences between the corresponding coordinates of the two points.

For two observations

$$X_i = \left(x_{i1}, x_{i2}, ..., x_{ip}\right)$$

$$X_j = \left(x_{j1}, x_{j2}, ..., x_{jp}\right)$$

the Squared Euclidean Distance is:

$$d^2\left(\mathrm{i}, \mathrm{j}\right) = \sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2$$

Interpretation

- It measures how far apart two objects are in multidimensional space.

- It gives more weight to large differences because the differences are squared.
- It is widely used in clustering algorithms like k-means**,** Ward's method**,** and MDS when distances are required.

Squared Euclidean Distance: Advantages

1. Computationally simpler and faster

- The square root is not computed**,** unlike Euclidean distance.
- This reduces computation time, especially in large datasets and high-dimensional clustering (e.g., k-means).

2. Emphasizes larger differences

- Squaring magnifies large deviations.
- Hence, objects that differ strongly on some variables are placed much farther apart.
- Useful when large deviations are important for clustering or classification.

3. Consistent with many clustering criteria

- Methods like k-means**,** Ward's method**,** and minimum-variance clustering are based on minimizing sum of squared distances**.**
- SED directly matches the Within-Cluster Sum of Squares (WCSS) objective.

4. Geometrically interpretable

- Although the square root is removed, the interpretation of distance is still consistent with Euclidean geometry.
- Preserves relative ordering of distances (monotonic with Euclidean distance).

Disadvantages

1. Sensitive to outliers

- Squaring increases the effect of extreme values disproportionately.
- A single large deviation can dominate the distance and distort clustering results.

2. Requires variables to be on the same scale

- If variables have different units (e.g., height in cm, weight in kg), the larger-scale variable dominates the squared distance.
- Standardization (z-scores) is necessary before computing SED.

3. Ignores correlation between variables

- Assumes variables are independent.
- In multivariate data with correlated variables, SED may misrepresent true dissimilarity.
- (Mahalanobis distance handles this better.)

4. Provides distances in squared units

- Distances are not in original units, so interpretation (e.g., "actual distance") is less intuitive compared to Euclidean distance.

5. Can lead to over-separation of clusters

- Because large differences are heavily penalized, clusters may appear artificially separated in high-dimensional spaces.

Example:

Let two observations be:

$X_1=(2,4,5)$, $X_2=(3,7,1)$

$$d_{12}^2 = (2-3)^2 + (4-7)^2 + (5-1)^2$$
$$= (-1)^2 + (-3)^2 + (4)^2$$
$$= 1 + 9 + 16 = 26$$

So, **Squared Euclidean Distance = 26**.

13.2.2 Minkowski Distance

The Minkowski distance is a generalization of the Euclidean distance.

With the measurement, x_{ik}, $i = 1, \ldots, N$, $k = 1, \ldots, p$, the Minkowski distance is

$$d_M(i, j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}}$$

where $\lambda \geq 1$. It is also called the L_λ metric.

- $\lambda = 1$: L_1 metric, Manhattan or City-block distance.
- $\lambda = 2$: L_2 metric, Euclidean distance.
- $\lambda \to \infty$: L_∞ metric, Supremum distance.

$$\lim_{\lambda \to \infty} \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}} = \max\left(|x_{i1} - x_{j1}|, \ldots, |x_{ip} - x_{jp}| \right)$$

Note that λ and p are two different parameters. Dimension of the data matrix remains finite.

Advantages of Minkowski Distance

1. Highly flexible
- By varying q, it can behave like several popular distance metrics.
- Users can tune the distance measure depending on the structure of the data.

2. Includes many useful metrics as special cases
- Useful in clustering, pattern recognition, and machine learning because one formula covers:
  - Manhattan
  - Euclidean
  - Chebyshev

3. Can control contribution of variable differences
- Lower q reduces the influence of large differences.
- Higher q highlights large deviations.
- Helps adapt to different data patterns.

4. Useful in machine learning and pattern recognition
- Many algorithms allow selecting the order qqq to improve classification or clustering performance.

Disadvantages of Minkowski Distance

1. Sensitive to variable scale
- Like Euclidean and Manhattan distances, variables with larger numeric values dominate the distance.
- Requires **standardization** (z-scores).

2. Sensitive to outliers (for large q)

- For q>2, large deviations are magnified.
- Makes clustering unstable if dataset contains extreme values.

3. Computational difficulty for non-integer q

- For fractional orders (e.g., q=1.5), computation becomes slower and more complex.

4. Choosing the right q is not straightforward

- No universal rule for selecting qqq.
- Often requires trial-and-error or cross-validation.

5. Not suitable for categorical variables

- Only applicable to continuous or numeric variables.

13.2.3 Chebyshev Distance

The Chebyshev Distance (also called L_∞ norm or maximum metric**)** between two points in a *p*-dimensional space is defined as the maximum absolute difference among their corresponding coordinates.

For two observations

$$X_i = \left(x_{i1}, x_{i2},.....,x_{ip}\right), X_j = \left(x_{j1}, x_{j2},.....,x_{jp}\right)$$

the Chebyshev distance is:

$$d_{ij} = \max_{k=1,.....,\mathrm{p}} \left|x_{ik} - x_{jk}\right|$$

Interpretation

- It measures the greatest deviation between two points along any coordinate.
- Only the largest coordinate difference contributes to the distance; smaller differences are ignored.
- Geometrically, it forms square-shaped (in 2D) **or** cube-shaped (in 3D) contours, unlike the circular Euclidean distance.

Example

Let

$X_1=(4,9,2)$, $X_2=(7,3,5)$

$$\left|4-7\right| = 3$$

$$\left|9-3\right| = 3$$

$$\left|2-5\right| = 3$$

Chebyshev distance:

$$d_{12} = \max\left(3,6,3\right) = 6$$

Advantages of Chebyshev Distance

1. Simple and fast to compute

- Requires only absolute differences and a max operation.

- Useful in large datasets and real-time systems.

2. Captures the dominant difference

- Good when the **largest coordinate difference** decides similarity.
- Useful in quality control, chess (king's moves), and bottleneck problems.

3. Robust when small variations are unimportant

- If similarity should depend only on the **worst-case difference**, Chebyshev is appropriate.

4. Works well in grid-based or discrete spaces

- Frequently used in:
  - Image processing
  - Pattern recognition
  - Chessboard distances
  - Robotics path planning

Disadvantages

1. Ignores all but the largest difference

- Smaller but meaningful differences across several variables are **completely neglected**.
- Poor for datasets where *overall* variation matters.

2. Very sensitive to noise / outliers

- A single noisy measurement (large deviation) dominates the distance.

3. Requires variables to be on the same scale

- Same issue as Euclidean, Manhattan, and Minkowski distances.
- Standardization is necessary in multivariate applications.

4. Geometry may not match natural clustering

- Square/cube contours may not reflect the **natural shapes** of clusters in real data.
- Often gives unnatural groupings compared to Euclidean distance.

13.3 EUCLIDEAN DISTANCE

The Euclidean Distance between two points in a *p*-dimensional space is the straight-line distance between them. It is derived from the Pythagorean theorem.
For two observations:

$$X_i = \left(x_{i1}, x_{i2}, ..., x_{ip} \right)$$

$$X_j = \left(x_{j1}, x_{j2}, ..., x_{jp} \right)$$

the Euclidean distance is defined as:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}$$

Interpretation

- It measures the actual geometric distance between two objects.
- Most commonly used in cluster analysis**,** MDS**,** nearest-neighbour classification**,** and multivariate space.
- Produces circular (2D) **or** spherical (3D) distance contours.

Example

Let

X1=(2,5,6), X2=(5,1,3)

$$d_{12} = \sqrt{(2-5)^2 + (5-1)^2 + (6-3)^2}$$
$$= \sqrt{(-3)^2 + (4)^2 + (3)^2}$$
$$= \sqrt{(-3)^2 + (4)^2 + (3)^2}$$
$$= 5.83$$

Advantages of Euclidean Distance

1. Intuitive and easy to understand
- Direct extension of the Pythagorean theorem.
- Matches our natural perception of distance.

2. Geometrically meaningful
- Represents actual spatial distance.
- Good for visualization, clustering, and MDS.

3. Most commonly used in clustering
- Works well when clusters are spherical or compact.
- Basis for many algorithms like k-means, hierarchical clustering (single, complete, average linkage).

4. Sensitive to overall differences
- Takes all variable differences into account, not just maximum (Chebyshev) or sum of absolute differences (Manhattan).

Disadvantages

1. Sensitive to scale of measurement
- Variables with large numeric range dominate the distance.
  Solution: Standardize (z-scores) before computing distance.

2. Sensitive to outliers
- Squaring magnifies extreme deviations (similar to squared Euclidean).

3. Assumes variables are uncorrelated
- Does not consider variable relationships.
  Solution: Use **Mahalanobis distance** when variables are correlated.

4. Less effective in high dimensions
- Suffers from **curse of dimensionality**:
  - Distances converge
  - Loss of discriminating power
  - Clustering performance deteriorates

5. Not suitable for categorical variables
- Only applicable to continuous or numeric variables.

13.4 Mahalanobis Distance

Let **X** be a N × p matrix. Then the ith row of **X** is

$$x_i^T = (x_{i1}, ..., x_{ip})$$

The Mahalanobis distance is

$$d_{M\,\mathrm{H}}(\mathrm{i},\mathrm{j}) = \left(\left(x_i - x_j\right)^T \Sigma^{-1}\left(x_i - x_j\right)\right)^{\frac{1}{2}}$$

where $\sum$ is the p×p sample covariance matrix.

13.5 CONCLUSION

Similarity and distance measures play a foundational role in multivariate analysis, as they provide the quantitative basis for comparing observations in multidimensional space. The choice of an appropriate distance metric directly influences the outcomes of clustering, classification, and dimensionality reduction techniques.

The Squared Euclidean Distance**,** Chebyshev Distance**,** and Minkowski Distance represent flexible and widely used dissimilarity measures, each capturing different patterns of variation depending on whether overall differences, maximum deviations, or generalized norms are emphasized. The Euclidean Distance**,** being the most intuitive and geometrically interpretable metric, remains central to many multivariate techniques, especially when variables are on similar scales and uncorrelated.

However, real-world datasets often contain variables that differ in scale and exhibit interdependence. In such cases, the Mahalanobis Squared Distance (D^2) provides a more robust and statistically sound measure by incorporating variance–covariance structure, allowing for meaningful comparisons even when variables are correlated.

Overall, understanding and selecting the appropriate distance measure is essential for accurate data interpretation, effective clustering, and reliable multivariate modeling. A clear knowledge of these measures enhances analytical decisions and leads to more insightful conclusions in multivariate statistical studies.

13.6 SELF ASSESSMENT QUESTIONS:

1. Explain the concept of similarity and dissimilarity measures in multivariate analysis. Why are they important for clustering and multidimensional scaling?
2. Derive the formula for Squared Euclidean Distance and discuss its advantages and disadvantages. In what situations is it preferred over Euclidean Distance?
3. Define Chebyshev Distance. Provide a numerical example
4. What is Minkowski Distance? Discuss how Euclidean Distance and Manhattan Distance arise as special cases.
5. Derivation of Mahalnobis D^2 test statistic to test $H_0: \underset{\sim}{\boldsymbol{\mu}}_1 = \underset{\sim}{\boldsymbol{\mu}}_2$ vs $\mathrm{H}_1: \underset{\sim}{\boldsymbol{\mu}}_1 \neq \underset{\sim}{\boldsymbol{\mu}}_2$ and the relationship between Hotelling's T^2 and Mahalnobis D^2

13.7 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein.

Dr. Syed Jilani

LESSON -14

HIERARCHICAL CLUSTERING METHODS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of Hierarchical Clustering methods
- To know the concept of Hierarchical Clustering methods
- To acquire knowledge about significance of is Hierarchical Clustering methods

STRUCTURE

14.1. INTRODUCTION

Multivariate methods deal with the analysis of data of more than two variables recorded from n sample objects selected from a specified population. Since the sample objects are selected from a specified population, the units are assumed to be homogeneous in respect of some characteristics. However, the values of different variables recorded from sample objects are not strictly uniform, though there should not be any systematic difference in the objects. In general, we expect some variations in the values of the variables, even if the sample objects are uniform in respect of some characters. For example, the income or the expenditure of middle class of people in a country are not exactly uniform, though they belong to the same class.

Again, the people of a country can be classified as rich, upper middle class, lower middle class and poor. For each class of people there may be common variable which influences the economic condition. For example, the income of a person depends on his education. This is true for every class of people. But their income or expenditure are not uniform. Therefore, there may be some systematic difference in values of the variables recorded from sample objects, there may be some similarities in the recorded observations of sample objects. Those sample objects which are similar in their recorded information may form a group. Dissimilar objects fall in different groups. In general, the objects that share similar characteristics are

found together. In statistics, the search for relatively homogeneous objects is called cluster analysis.

The cluster analysis has wide application in biology, medicine, agriculture, marketing, etc. The numerical taxonomy in the field of biology is used to classify the animals into class, order and families. Different species of plants have different characteristics. Therefore, plant specimens can be classify into homogeneous groups. In agriculture, the land fertility of a particular region may not be homogeneous for any type of crop. Then the pieces of land sharing similar fertility for a particular may be grouped together. The milk production of cows, even of the same type, may vary due to lactation period. Then the cows of the same lactation period may be grouped together. In economics, the people of a city center may be grouped according to their socio-economic condition. In marketing, people can be grouped according to the similar buying habits. In medicine, the patients having similar disease may be clustered together.

Since similar objects form a cluster, all the sample points in any cluster will provide similar information about the population characteristics. Thus, for further analysis one may include one object from each cluster analysis is a data reduction technique in rows of the data matrix.

What is Clustering?

Cluster analysis is a technique used in data mining and machine learning to group similar objects into clusters. K-means clustering is a widely used method for cluster analysis where the aim is to partition a set of objects into K clusters in such a way that the sum of the squared distances between the objects and their assigned cluster mean is minimized.
Hierarchical clustering and k-means clustering are two popular techniques in the field of unsupervised learning used for clustering data points into distinct groups. While k-means clustering divides data into a predefined number of clusters, hierarchical clustering creates a hierarchical tree-like structure to represent the relationships between the clusters.

Example:

Let's try understanding this with a simple example. A bank wants to give credit card offers to its customers. Currently, they look at the details of each customer and, based on this information, decide which offer should be given to which customer.
Now, the bank can potentially have millions of customers. Does it make sense to look at the details of each customer separately and then make a decision? Certainly not! It is a manual process and will take a huge amount of time.
So what can the bank do? One option is to segment its customers into different groups. For instance, the bank can group the customers based on their income:

Can you see where I'm going with this? The bank can now make three different strategies or offers, one for each group. Here, instead of creating different strategies for individual customers, they only have to make 3 strategies. This will reduce the effort as well as the time.

The groups I have shown above are known as clusters, and the process of creating these groups is known as clustering. Formally, we can say that:
In clustering, we do not have a target to predict. We look at the data, try to club similar observations, and form different groups. Hence it is an unsupervised learning problem.
We now know what clusters are and the concept of clustering. Next, let's look at the properties of these clusters, which we must consider while forming the clusters.

Let $X(n \times p)$ be a data matrix from a specified population. Let the values of the p variables observed from n sample objects be denoted by $X_1, X_2, ..., X_n$. The objective of the cluster analysis is to group these n vector of values into $n_1 (n_1 < n)$ vectors so that the elements in a group are homogeneous. Here the method of clustering is on the basis of one-sample observations. Let $X_{ij}[i = 1, 2, ..., n_j; j = 1, 2, ..., m]$ be the vector of values of p variables of i-th object in j-th sample. Here the objective of clustering is to form m_1 groups (m_1<m) of sample observations in different groups are heterogeneous.

From above discursion is is clear that the CA reduces the sample observations in size. It has similar property of other data reduction technique. Namely, PCA . this analysis has a similarity with DA in respect of classification of observations. But DA derives a rule for an allocating an object to its proper properties based on some prior information of the group membership of the object. Whereas, the CA identifies homogeneous groups or clusters.

There is no unified approach on what actually constitute a cluster. As per the definition what we have discussed above, a cluster constitutes with a similar object. Then, we need to decide on a measure of inter-object similarities. Also, a decision is needed to specify a procedure for forming the clusters, based on the chosen measure of similarity. The criterion of similarity in observations varies from researcher to researcher. However, the basic criterion is that the objects in a cluster should be closer to each other than to objects in other clusters. As a preliminary technique to identify the similarity of objects, one can use the diagram of sample objects. Let us consider that from each of 'n' sample object values of p variables are recorded. These values can be represented a p-dimensional diagram. The values of each variable are plotted in each separate axis. If n sets of values are plotted in p-axes, a diagram will be formed. The cluster can be formed with those objects, which lie nearer in an area of the diagram but are dispersed from another area. The cluster can also be formed mathematically calculating distances among sample objects.

14.2 BASIC STEPS OF CLUSTER ANALYSIS:

In CA, the sample objects are clustered on the basis of some characteristics. Therefore, to start with the analysis, a number of decisions must be made regarding the characteristics to be considered, the variables to be included in the analysis, the measurement of distance between objects and the criterion to group the objects.

The selection of variables for any CA is important, since the exclusion of important variables will be poor or misleading findings. For eg., if any marketing research the consumers are needed to be clustered, their tastes and habits and their economic capacities must be considered. Otherwise, the clustering of consumers will not be fruitful. The initial choice of variables determines the characteristics that can be used to identify subgroups.

After the selection of variables, the next important point to be considered is to measure the distance and similarity become objects. Two objects will be included in two separate groups, if their distance is maximum and they will be included in one group if they are close to each other. Therefore, one of the important steps in cluster analysis is to measure the distance among objects.

The measurement of similarity of their distance is divided into two main parts. One of this is (a) distance – type measure, and another is (b) matching – type measure.

14.3 CLUSTER LINKAGE METHODS:

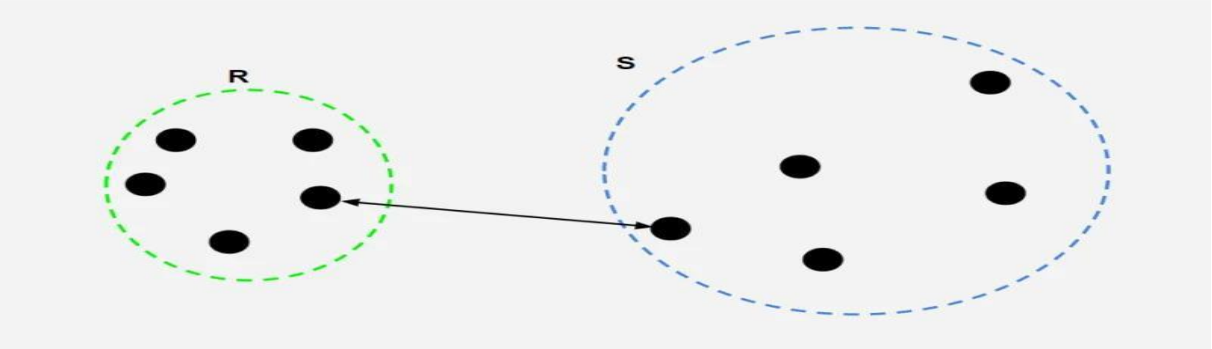
The linkage method that you choose determines how the distance between two clusters is defined. At each amalgamation stage, the two closest clusters are joined. At the beginning, when each observation constitutes a cluster, the distance between clusters is just the inter-observation distance. Subsequently, after observations are joined together, a linkage rule is necessary for calculating inter-cluster distances when there are multiple observations in a cluster. You might want to try several linkage methods and compare results. Depending on the characteristics of your data, some methods may provide "better" results than others.

14.3.1 Single Linkage Method:

Single linkage agglomerative clustering is a hierarchical clustering algorithm that works by iteratively merging the two closest clusters based on the minimum distance between their closest members. The steps involved in it are:

1. Start with assigning each observation to its own cluster.
2. Compute the distance between all pairs of clusters using a chosen distance metric (e.g., Euclidean distance).
3. Merge the two closest clusters into a single cluster.
4. Recompute the distance between the new cluster and all remaining clusters.
5. Repeat steps 3 and 4 until all observations belong to a single cluster, or until a pre-defined number of clusters has been reached.

In single linkage agglomerative clustering, the distance between two clusters is defined as the minimum distance between any two points in the clusters. This is why it's also called the "nearest neighbor" or "single linkage" clustering.



One disadvantage of single linkage agglomerative clustering is that it can produce long, trailing clusters that do not represent well-defined groups, also known as chaining phenomenon. This can be overcome by using other linkage criteria such as complete linkage, average linkage, or Ward's linkage.

Example 1: Numerical Example

	A	**B**	**C**	**D**
A	0	2	6	10
B	2	0	5	9
C	6	5	0	4
D	10	9	4	0

Distance Matrix:

A-B = 2

A-C = 6

A-D = 10

B-C = 5

B-D = 9

C-D = 4

Step 1: Smallest distance = 2 → Merge A & B → Cluster {A,B}

Step 2: Update distances:

(A,B)-C = min(6,5) = 5

(A,B)-D = min(10,9) = 9

Step 3: Next smallest = 4 → Merge C & D → Cluster {C,D}

Step 4: Distance between clusters:

D({A,B},{C,D}) = min(6,10,5,9) = 5

Final:

A-B merge at 2

C-D merge at 4

(AB)-(CD) merge at 5

14.3.2 Complete linkage agglomerative clustering

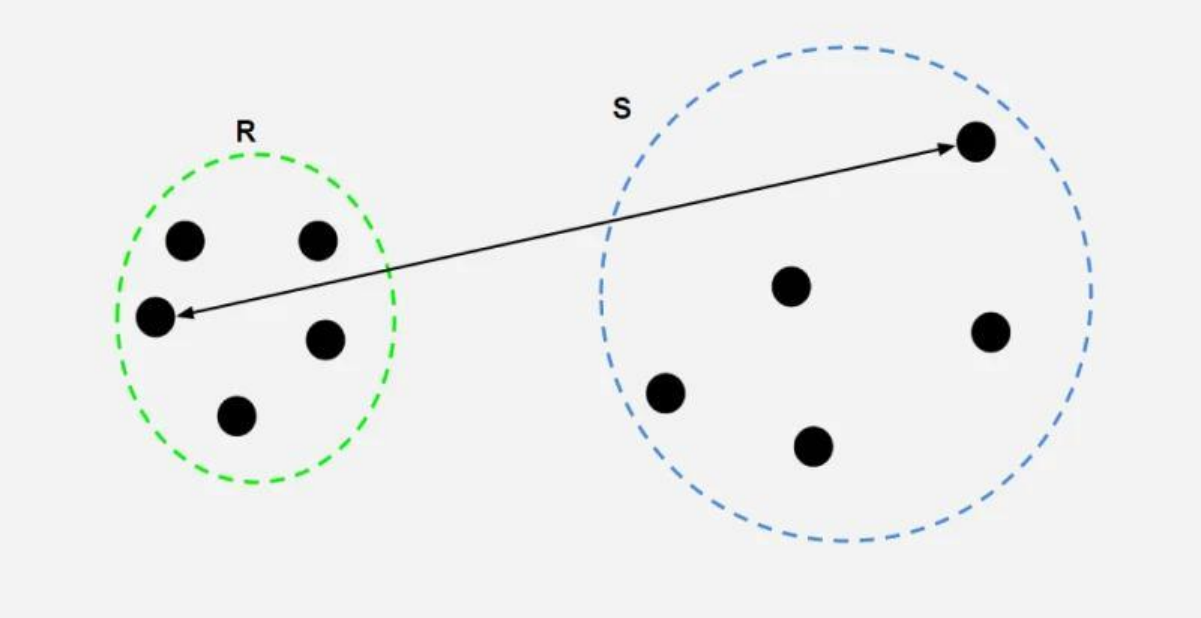
Complete linkage agglomerative clustering is another hierarchical clustering algorithm that works by iteratively merging the two closest clusters based on the maximum distance between their furthest members.

The steps involved in the complete linkage agglomerative clustering algorithm are:

1. Start with assigning each observation to its own cluster.

2. Compute the distance between all pairs of clusters using a chosen distance metric (e.g., Euclidean distance).
3. Merge the two closest clusters into a single cluster.
4. Recomputed the distance between the new cluster and all remaining clusters.
5. Repeat steps 3 and 4 until all observations belong to a single cluster, or until a pre-defined number of clusters has been reached.

1) In complete linkage agglomerative clustering, the distance between two clusters is defined as the maximum distance between any two points in the clusters. This is why it's also called the "furthest neighbor" or "complete linkage" clustering.
2)
Compared to single linkage agglomerative clustering, complete linkage tends to produce more compact, spherical clusters that are less prone to the chaining phenomenon. However, it's more sensitive to outliers and can produce unbalanced clusters if there are extreme values or noise in the data.



Example:
This document provides a detailed, step-by-step explanation of Complete Linkage Agglomerative Hierarchical Clustering using a sample dataset of five objects: A, B, C, D, and E. Complete Linkage considers the **maximum distance** between elements of two clusters when merging.

1. Distance Matrix

	A	B	C	D	E
A	0	4	6	7	10
B	4	0	5	9	11
C	6	5	0	4	8
D	7	9	4	0	6
E	10	11	8	6	0

The above table contains the pairwise Euclidean distances between all objects.

2. Step-by-Step Clustering Process

Step 1: Find the Closest Pair

We examine all distances and identify the smallest value. The minimum distance is **4**, which occurs for the pairs (A, B) and (C, D). We merge one pair first—here, we merge (A, B).

Step 2: Update Distances Using Complete Linkage

Complete linkage defines the distance between two clusters as the maximum pairwise distance between elements of the clusters. Thus, distances from cluster (AB) to other objects are computed as:

- d(AB, C) = max(d(A, C)=6, d(B, C)=5) = 6
- d(AB, D) = max(d(A, D)=7, d(B, D)=9) = 9
- d(AB, E) = max(d(A, E)=10, d(B, E)=11) = 11

The updated distance matrix becomes:

	AB	C	D
AB	0	6	9
C	6	0	4
D	9	4	0

Distances to E are handled separately in continuation tables to avoid clutter.

Step 3: Merge the Next Closest Pair

The smallest remaining distance is 4 for the pair (C, D). Thus, we merge C and D to form cluster (CD).

Step 4: Recompute Distances Between Clusters (AB), (CD), and E

Compute complete linkage distances:

• d(AB, CD) = max(6, 7, 5, 9) = 9

• d(CD, E) = max(8, 6) = 8

Thus, the updated distance matrix (clusters AB, CD, and E) becomes:

	AB	CD	E
AB	0	9	11
CD	9	0	8
E	11	8	0

Step 5: Merge Clusters (CD) and E

The smallest distance is 8, so we merge (CD, E) to form cluster (CDE).

Step 6: Final Merge

Compute complete linkage distance between clusters (AB) and (CDE):

Distances involved: A–C=6, A–D=7, A–E=10, B–C=5, B–D=9, B–E=11.

Complete linkage distance = max(all above) = 11.

Thus, clusters (AB) and (CDE) are merged to form the final single cluster.

3. Final Dendrogram Structure

1. Merge A and B at height 4.

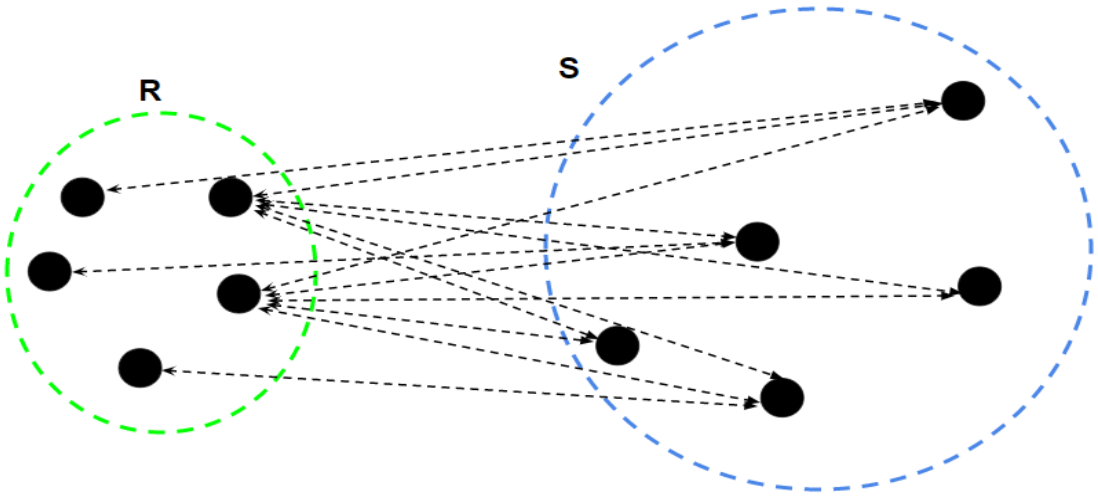
2. Merge C and D at height 4.
3. Merge (CD) with E at height 8.
4. Merge (AB) with (CDE) at height 11.

14.3.3 AVERAGE LINKAGE METHOD

Another hierarchical clustering algorithm that is commonly used in bioinformatics and evolutionary biology is the Unweighted Pair Group Method with Arithmetic Mean (UPGMA).

The steps involved in the UPGMA algorithm:

1. Begin by assigning each data point to its own cluster.
2. Compute the pairwise distances between all clusters based on the distance metric of choice, such as Euclidean distance, Manhattan distance, or Pearson correlation.
3. Find the two closest clusters based on the pairwise distances and merge them into a single cluster. The distance between the two clusters is calculated as the average of the pairwise distances between their members.
4. Update the pairwise distances between the new cluster and all remaining clusters. The distance between the new cluster and any other cluster is calculated as the average of the pairwise distances between the members of the new cluster and the members of the other cluster.
5. Repeat steps 3 and 4 until all data points belong to a single cluster.



Example:
We have five observations in $\mathbb{R}^2$:

$X_1 = (1, 1)$
$X_2 = (2, 1)$
$X_3 = (4, 3)$
$X_4 = (5, 4)$
$X_5 = (5, 5)$

We use Euclidean distance:

$d(X_i, X_j) = \text{sqrt}[\ (x_i1 - x_j1)^2 + (x_i2 - x_j2)^2\ ]$.

Step 1: Compute the Distance Matrix

Compute distances between all pairs (X_i, X_j).

1. $d(X_1, X_2)$:

$d(X_1, X_2) = \text{sqrt}[\ (1 - 2)^2 + (1 - 1)^2\ ]$

$= \text{sqrt}[\ 1^2 + 0^2\ ]$

$= 1.000$ (approx)

2. $d(X_1, X_3)$:

$d(X_1, X_3) = \text{sqrt}[\ (1 - 4)^2 + (1 - 3)^2\ ]$

$= \text{sqrt}[\ (-3)^2 + (-2)^2\ ]$

$= \text{sqrt}[\ 9 + 4\ ]$

$= \text{sqrt}[13] \approx 3.606$

3. $d(X_1, X_4)$:

$d(X_1, X_4) = \text{sqrt}[\ (1 - 5)^2 + (1 - 4)^2\ ]$

$= \text{sqrt}[\ (-4)^2 + (-3)^2\ ]$

$= \text{sqrt}[\ 16 + 9\ ]$

$= \text{sqrt}[25] = 5.000$

4. $d(X_1, X_5)$:

$d(X_1, X_5) = \text{sqrt}[\ (1 - 5)^2 + (1 - 5)^2\ ]$

$= \text{sqrt}[\ (-4)^2 + (-4)^2\ ]$

$= \text{sqrt}[\ 16 + 16\ ]$

$= \text{sqrt}[32] \approx 5.657$

5. $d(X_2, X_3)$:

$d(X_2, X_3) = \text{sqrt}[\ (2 - 4)^2 + (1 - 3)^2\ ]$

$= \text{sqrt}[\ (-2)^2 + (-2)^2\ ]$

= sqrt[4 + 4]

= sqrt[8] ≈ 2.828

6. $d(X_2, X_4)$:

$d(X_2, X_4) = sqrt[(2 - 5)^2 + (1 - 4)^2]$

= sqrt[$(-3)^2$ + $(-3)^2$]

= sqrt[9 + 9]

= sqrt[18] ≈ 4.243

7. $d(X_2, X_5)$:

$d(X_2, X_5) = sqrt[(2 - 5)^2 + (1 - 5)^2]$

= sqrt[$(-3)^2$ + $(-4)^2$]

= sqrt[9 + 16]

= sqrt[25] = 5.000

8. $d(X_3, X_4)$:

$d(X_3, X_4) = sqrt[(4 - 5)^2 + (3 - 4)^2]$

= sqrt[$(-1)^2$ + $(-1)^2$]

= sqrt[1 + 1]

= sqrt[2] ≈ 1.414

9. $d(X_3, X_5)$:

$d(X_3, X_5) = sqrt[(4 - 5)^2 + (3 - 5)^2]$

= sqrt[$(-1)^2$ + $(-2)^2$]

= sqrt[1 + 4]

= sqrt[5] ≈ 2.236

10. $d(X_4, X_5)$:

$d(X_4, X_5) = sqrt[(5 - 5)^2 + (4 - 5)^2]$

= sqrt[0^2 + $(-1)^2$]

= 1.000

Distance matrix (0 on the diagonal, rounded to 3 decimals):

	X1	**X2**	**X3**	**X4**	**X5**
X1	0.000	1.000	3.606	5.000	5.657
X2	1.000	0.000	2.828	4.243	5.000
X3	3.606	2.828	0.000	1.414	2.236
X4	5.000	4.243	1.414	0.000	1.000

	X1	X2	X3	X4	X5
X5	5.657	5.000	2.236	1.000	0.000

Step 2: Hierarchical Agglomerative Clustering (Average Linkage)

Initially, each object is its own cluster:

$C_1 = \{X_1\}$

$C_2 = \{X_2\}$

$C_3 = \{X_3\}$

$C_4 = \{X_4\}$

$C_5 = \{X_5\}$

We repeatedly merge the two clusters with the smallest inter-cluster distance, using average linkage.

Stage 1: First Merge

From the distance matrix, the smallest non-zero distances are:

$d(X_1, X_2) = 1.000$

$d(X_4, X_5) = 1.000$

We have a tie; we may merge either pair first.

Assume we merge X_1 and X_2 first.

Merge clusters: $C_1 = \{X_1\}$ and $C_2 = \{X_2\}$

New cluster: $C_{12} = \{X_1, X_2\}$

Height (distance level): 1.000

Current clusters:

$C_{12} = \{X_1, X_2\}$

$C_3 = \{X_3\}$

$C_4 = \{X_4\}$

$C_5 = \{X_5\}$

Now compute distances from C_{12} to the remaining singletons using average linkage:

$d(C_{12}, C_3) = (1 / (2 \cdot 1)) \cdot [\, d(X_1, X_3) + d(X_2, X_3) \,]$

$= (3.606 + 2.828) / 2$

$\approx 6.434 / 2$

≈ 3.217

$d(C_{12}, C_4) = (1 / (2 \cdot 1)) \cdot [\, d(X_1, X_4) + d(X_2, X_4) \,]$

$= (5.000 + 4.243) / 2$

$\approx 9.243 / 2$

≈ 4.622

$d(C_{12}, C_5) = (1 / (2 \cdot 1)) \cdot [\, d(X_1, X_5) + d(X_2, X_5) \,]$

$= (5.657 + 5.000) / 2$

$\approx 10.657 / 2$

≈ 5.329

Distances among $\{C_3, C_4, C_5\}$ remain as in the original matrix:

$d(C_3, C_4) = d(X_3, X_4) \approx 1.414$

$d(C_3, C_5) = d(X_3, X_5) \approx 2.236$

$d(C_4, C_5) = d(X_4, X_5) = 1.000$

New inter-cluster distances (rounded):

	C12	C3	C4	C5
C12	–	3.217	4.622	5.329
C3	3.217	–	1.414	2.236
C4	4.622	1.414	–	1.000
C5	5.329	2.236	1.000	–

Stage 2: Second Merge

The smallest distance is now:

$d(C_4, C_5) = 1.000$

So we merge clusters C_4 and C_5:

Merge clusters: $C_4 = \{X_4\}$, $C_5 = \{X_5\}$

New cluster: $C_{45} = \{X_4, X_5\}$

Height (distance level): 1.000

Current clusters:

$C_{12} = \{X_1, X_2\}$

$C_3 = \{X_3\}$

$C_{45} = \{X_4, X_5\}$

Compute distances involving C_{45}:

$d(C_{12}, C_{45}) = (1 / (2 \cdot 2)) \cdot \Sigma\_\{i \in \{X_1, X_2\}\} \Sigma\_\{j \in \{X_4, X_5\}\} d(i, j)$

$= (1 / 4) \cdot [\, d(X_1, X_4) + d(X_1, X_5) + d(X_2, X_4) + d(X_2, X_5) \,]$

$= (1 / 4) \cdot [\, 5.000 + 5.657 + 4.243 + 5.000 \,]$

$= (1 / 4) \cdot 19.900 \approx 4.975$

$d(C_3, C_{45}) = (1 / (1 \cdot 2)) \cdot [\, d(X_3, X_4) + d(X_3, X_5) \,]$

$= (1 / 2) \cdot [\, 1.414 + 2.236 \,]$

$\approx 3.650 / 2$

≈ 1.825

Existing distance $d(C_{12}, C_3) \approx 3.217$ remains.

Updated distances:

	C12	**C3**	**C45**
C12	–	3.217	4.975
C3	3.217	–	1.825
C45	4.975	1.825	–

Stage 3: Third Merge

The smallest distance now is:

$d(C_3, C_{45}) \approx 1.825$

So we merge C_3 and C_{45}:

Merge clusters: $C_3 = \{X_3\}$, $C_{45} = \{X_4, X_5\}$

New cluster: $C_{345} = \{X_3, X_4, X_5\}$

Height (distance level): ≈ 1.825

Current clusters:

$C_{12} = \{X_1, X_2\}$

$C_{345} = \{X_3, X_4, X_5\}$

Now compute the distance between these two clusters using average linkage:

$d(C_{12}, C_{345}) = (1 / (2 \cdot 3)) \cdot \Sigma\_\{i \in \{X_1, X_2\}\} \Sigma\_\{j \in \{X_3, X_4, X_5\}\} d(i, j).$

We need the 6 pairwise distances:

$d(X_1, X_3) \approx 3.606$

$d(X_1, X_4) = 5.000$

$d(X_1, X_5) \approx 5.657$

$d(X_2, X_3) \approx 2.828$

$d(X_2, X_4) \approx 4.243$

$d(X_2, X_5) = 5.000$

Sum $= 3.606 + 5.000 + 5.657 + 2.828 + 4.243 + 5.000 \approx 26.334$

So:

$d(C_{12}, C_{345}) = 26.334 / 6 \approx 4.389$

There are now only two clusters, so the final merge is at height ≈ 4.389.

Stage 4: Final Merge

Merge clusters: $C_{12} = \{X_1, X_2\}$, $C_{345} = \{X_3, X_4, X_5\}$

New cluster: $C_{12345} = \{X_1, X_2, X_3, X_4, X_5\}$

Height (distance level): ≈ 4.389

Summary of merges:

Step	Merged Clusters	New Cluster	Distance (Height)
1	$\{X_1\}$, $\{X_2\}$	C_{12}	1.000
2	$\{X_4\}$, $\{X_5\}$	C_{45}	1.000
3	$\{X_3\}$, C_{45}	C_{345}	≈ 1.825
4	C_{12}, C_{345}	C_{12345}	≈ 4.389

14.3.4 Ward Linkage Method

Ward's method (also called the minimum variance method) is a hierarchical agglomerative clustering technique in which, at each stage, the pair of clusters merged is the one that causes the smallest increase in the total within-cluster sum of squares (WSS).

Let Ca and Cb be two clusters with sizes

na = |Ca| and nb = |Cb|, and let

- $\bar{a}$ = mean vector of cluster Ca
- $\bar{b}$ = mean vector of cluster Cb

Then the increase in **WSS** when Ca and Cb are merged is

$$\Delta(Ca, Cb) = \left(na * \frac{nb}{na + nb}\right) * \left|\left|\bar{a} - \bar{b}\right|\right|^2$$

where $|| \bar{a} - \bar{b} ||^2$ is the squared Euclidean distance between the two cluster means.

Algorithm idea (Ward's method):

1. Start with each observation as its own cluster.
2. At each step, compute Δ(Ca, Cb) for all pairs of current clusters.
3. Merge the pair with the **smallest** Δ(Ca, Cb).
4. Continue until all observations are in a single cluster.

(b) Worked Example with the 6 Observations

We have six points in R^2:

- X1 = (1, 1)

- X2 = (2, 1)
- X3 = (3, 2)
- X4 = (8, 8)
- X5 = (9, 8)
- X6 = (9, 9)

We use **squared Euclidean distance**:

$|| x - y ||^2 = (x1 - y1)^2 + (x2 - y2)^2$.

Step 1: Squared Euclidean Distance Matrix

Compute pairwise squared distances:

- $d^2(X1, X2) = (1 - 2)^2 + (1 - 1)^2 = 1$
- $d^2(X1, X3) = (1 - 3)^2 + (1 - 2)^2 = 4 + 1 = 5$
- $d^2(X1, X4) = (1 - 8)^2 + (1 - 8)^2 = 49 + 49 = 98$
- $d^2(X1, X5) = (1 - 9)^2 + (1 - 8)^2 = 64 + 49 = 113$
- $d^2(X1, X6) = (1 - 9)^2 + (1 - 9)^2 = 64 + 64 = 128$
- $d^2(X2, X3) = (2 - 3)^2 + (1 - 2)^2 = 1 + 1 = 2$
- $d^2(X2, X4) = (2 - 8)^2 + (1 - 8)^2 = 36 + 49 = 85$
- $d^2(X2, X5) = (2 - 9)^2 + (1 - 8)^2 = 49 + 49 = 98$
- $d^2(X2, X6) = (2 - 9)^2 + (1 - 9)^2 = 49 + 64 = 113$
- $d^2(X3, X4) = (3 - 8)^2 + (2 - 8)^2 = 25 + 36 = 61$
- $d^2(X3, X5) = (3 - 9)^2 + (2 - 8)^2 = 36 + 36 = 72$
- $d^2(X3, X6) = (3 - 9)^2 + (2 - 9)^2 = 36 + 49 = 85$
- $d^2(X4, X5) = (8 - 9)^2 + (8 - 8)^2 = 1 + 0 = 1$
- $d^2(X4, X6) = (8 - 9)^2 + (8 - 9)^2 = 1 + 1 = 2$
- $d^2(X5, X6) = (9 - 9)^2 + (8 - 9)^2 = 0 + 1 = 1$

Now form the **squared distance matrix**:

	X1	**X2**	**X3**	**X4**	**X5**	**X6**
X1	0	1	5	98	113	128
X2	1	0	2	85	98	113
X3	5	2	0	61	72	85
X4	98	85	61	0	1	2

	X1	X2	X3	X4	X5	X6
X5	113	98	72	1	0	1
X6	128	113	85	2	1	0

Step 2: Hierarchical Clustering Using Ward's Method

Initial clusters (each point is its own cluster):

- C1 = {X1}
- C2 = {X2}
- C3 = {X3}
- C4 = {X4}
- C5 = {X5}
- C6 = {X6}

For singletons (na = nb = 1), we have

$$\Delta(Ca, Cb) = \left(1 * \frac{1}{1+1}\right) * \left|\left|\bar{a} - \bar{b}\right|\right|^2 = \left(\frac{1}{2}\right) * d^2,$$

So the Ward distance between two singletons is **half** of their squared Euclidean distance.

We proceed stage by stage.

Stage 1: First Merge

Look at the smallest squared distances in the matrix:

- $d^2(X1, X2) = 1$
- $d^2(X4, X5) = 1$
- $d^2(X5, X6) = 1$

All give Δ = (1/2)*1 = 0.5. There is a tie; we may choose any.

For definiteness, merge X1 and X2 first.

Merge: C1 = {X1}, C2 = {X2} → C12 = {X1, X2}

Cluster size: n12 = 2

Cluster mean:

$\bar{x}12 = ((1 + 2)/2 , (1 + 1)/2) = (1.5, 1)$

Increase in WSS:

$\Delta(C1, C2) = (1*1/(1+1)) * d^2(X1, X2) = (1/2)*1 = 0.5$

Current clusters:

- C12 = {X1, X2}
- C3 = {X3}

- C4 = {X4}
- C5 = {X5}
- C6 = {X6}

Stage 2: Second Merge

Among the remaining singletons {X3, X4, X5, X6}, the smallest squared distances are

- $d^2(X4, X5) = 1$
- $d^2(X5, X6) = 1$

So again $\Delta = 0.5$ for these possible merges.

Choose to merge X4 and X5 next.

Merge: C4 = {X4}, C5 = {X5} → C45 = {X4, X5}

Cluster size: n45 = 2

Cluster mean:

$\bar{x}45 = ((8 + 9)/2 , (8 + 8)/2) = (8.5, 8)$

Increase in WSS:

$\Delta(C4, C5) = (1*1/(1+1)) * d^2(X4, X5) = (1/2)*1 = 0.5$

Current clusters:

- C12 = {X1, X2}
- C3 = {X3}
- C45 = {X4, X5}
- C6 = {X6}

Stage 3: Third Merge

We now have four clusters: C12, C3, C45, C6.

We must compute $\Delta(Ca, Cb)$ for all pairs using

$\Delta(Ca, Cb) = (na * nb / (na + nb)) * || \bar{a} - \bar{b} ||^2$.

First compute means:

- C12: $\bar{x}12 = (1.5, 1)$, n12 = 2
- C3: $\bar{x}3 = (3, 2)$, n3 = 1
- C45: $\bar{x}45 = (8.5, 8)$, n45 = 2
- C6: $\bar{x}6 = (9, 9)$, n6 = 1

Now compute Δ for the pairs:

1. $\Delta(C12, C3)$

$|| \bar{x}12 - \bar{x}3 ||^2$

$= (1.5 - 3)^2 + (1 - 2)^2$

$= (-1.5)^2 + (-1)^2$

$= 2.25 + 1 = 3.25$

$\Delta(C12, C3) = (2*1/(2+1)) * 3.25 = (2/3)*3.25 \approx 2.167$

2. $\Delta(C45, C6)$

$|| \bar{x}45 - \bar{x}6 ||^2$

$= (8.5 - 9)^2 + (8 - 9)^2$

$= (-0.5)^2 + (-1)^2$

$= 0.25 + 1 = 1.25$

$\Delta(C45, C6) = (2*1/(2+1)) * 1.25 = (2/3)*1.25 \approx 0.833$

3. $\Delta(C12, C45)$

$|| \bar{x}12 - \bar{x}45 ||^2$

$= (1.5 - 8.5)^2 + (1 - 8)^2$

$= (-7)^2 + (-7)^2$

$= 49 + 49 = 98$

$\Delta(C12, C45) = (2*2/(2+2)) * 98 = (4/4)*98 = 98.0$

4. $\Delta(C12, C6)$

$|| \bar{x}12 - \bar{x}6 ||^2$

$= (1.5 - 9)^2 + (1 - 9)^2$

$= (-7.5)^2 + (-8)^2$

$= 56.25 + 64 = 120.25$

$\Delta(C12, C6) = (2*1/(2+1)) * 120.25 = (2/3)*120.25 \approx 80.17$

5. $\Delta(C3, C45)$

$|| \bar{x}3 - \bar{x}45 ||^2$

$= (3 - 8.5)^2 + (2 - 8)^2$

$= (-5.5)^2 + (-6)^2$

$= 30.25 + 36 = 66.25$

$\Delta(C3, C45) = (1*2/(1+2)) * 66.25 = (2/3)*66.25 \approx 44.17$

6. $\Delta(C3, C6)$

$|| \bar{x}3 - \bar{x}6 ||^2$

$= (3 - 9)^2 + (2 - 9)^2$

$= (-6)^2 + (-7)^2$

$= 36 + 49 = 85$

$\Delta(C3, C6) = (1*1/(1+1)) * 85 = (1/2)*85 = 42.5$

The **smallest** Δ is:

$\Delta(C45, C6) \approx 0.833$

So we merge C45 and C6.

Merge: C45 and C6 → C456 = {X4, X5, X6}

Cluster size: n456 = 3

Cluster mean:

$\bar{x}456 = ((8 + 9 + 9)/3 , (8 + 8 + 9)/3) = (26/3, 25/3) \approx (8.67, 8.33)$

Increase in WSS:

$\Delta(C45, C6) \approx 0.833$

Current clusters:

- C12 = {X1, X2}
- C3 = {X3}
- C456 = {X4, X5, X6}

Stage 4: Fourth Merge

Now we have three clusters: C12, C3, C456.

Means and sizes:

- C12: $\bar{x}12 = (1.5, 1)$, n12 = 2
- C3: $\bar{x}3 = (3, 2)$, n3 = 1
- C456: $\bar{x}456 \approx (8.67, 8.33)$, n456 = 3

Compute Δ again:

1. $\Delta(C12, C3)$ (already calculated)

$|| \bar{x}12 - \bar{x}3 ||^2 = 3.25$

$\Delta(C12, C3) = (2*1/(2+1)) * 3.25 = (2/3)*3.25 \approx 2.167$

2. $\Delta(C12, C456)$

$|| \bar{x}12 - \bar{x}456 ||^2$

$\approx (1.5 - 8.67)^2 + (1 - 8.33)^2$

$\approx (-7.17)^2 + (-7.33)^2$

$\approx 51.39 + 53.73 \approx 105.12$

$\Delta(C12, C456) \approx (2*3/(2+3)) * 105.12$

$= (6/5)*105.12$

≈ 126.17

3. $\Delta(C3, C456)$

$|| \bar{x}3 - \bar{x}456 ||^2$

$\approx (3 - 8.67)^2 + (2 - 8.33)^2$

$\approx (-5.67)^2 + (-6.33)^2$

$\approx 32.15 + 40.07 \approx 72.22$

$\Delta(C3, C456) \approx (1*3/(1+3)) * 72.22$

$= (3/4)*72.22$

≈ 54.17

Smallest Δ is:

$\Delta(C12, C3) \approx 2.167$

So we merge C12 and C3.

Merge: C12 and C3 → C123 = {X1, X2, X3}

Cluster size: n123 = 3

Cluster mean:

$\bar{x}123 = ((1 + 2 + 3)/3 , (1 + 1 + 2)/3) = (6/3, 4/3) = (2, 1.33)$

Increase in WSS:

$\Delta(C12, C3) \approx 2.167$

Current clusters:

- C123 = {X1, X2, X3}
- C456 = {X4, X5, X6}

Stage 5: Final Merge

Only two clusters remain: C123 and C456.

Sizes and means:

- C123: n123 = 3, $\bar{x}123 = (2, 1.33)$
- C456: n456 = 3, $\bar{x}456 \approx (8.67, 8.33)$

Compute $\Delta(C123, C456)$:

$|| \bar{x}123 - \bar{x}456 ||^2$

$\approx (2 - 8.67)^2 + (1.33 - 8.33)^2$

$\approx (-6.67)^2 + (-7.00)^2$

$\approx 44.49 + 49.00 \approx 93.49$

$\Delta(C123, C456) = (3*3/(3+3)) * 93.49$

$= (9/6)*93.49$

$= (3/2)*93.49$

≈ 140.17

This is a **very large** increase in WSS compared to previous merges.

Merge: C123 and C456 → C123456 (all points in one cluster).

This completes the clustering.

Step	Merged Clusters	New Cluster	Δ (Increase in WSS)
1	{X1}, {X2}	C12	0.50
2	{X4}, {X5}	C45	0.50
3	C45, {X6}	C456	≈ 0.83
4	C12, {X3}	C123	≈ 2.17
5	C123, C456	C123456	≈ 140.17

(c) Suggested 2-Cluster Solution and Interpretation

From the merge sequence and the jump in Δ:

- Up to Step 4, increases in WSS are small (0.5, 0.5, 0.83, 2.17).
- The final merge (Step 5) has a **massive increase** in WSS (≈ 140.17).

So, a natural **2-cluster solution** is obtained **before** the last merge:

- **Cluster 1:** C123 = {X1, X2, X3}
- **Cluster 2:** C456 = {X4, X5, X6}

Interpretation:

- Cluster {X1, X2, X3} lies in the **lower-left** region of the plane, with small internal variation.
- Cluster {X4, X5, X6} lies in the **upper-right** region, again with small internal variation.
- Merging these two clusters causes a very large increase in within-cluster sum of squares, so Ward's method strongly supports **two compact, well-separated clusters**.

14.3.5 Centroid Linkage Method:

In hierarchical agglomerative clustering, the centroid linkage method defines the distance between two clusters as the Euclidean distance between their centroids (mean vectors).

Let Ca and Cb be two clusters with sizes

- na = |Ca|,
- nb = |Cb|,

and let

- m_a = centroid (mean vector) of Ca,
- m_b = centroid (mean vector) of Cb.

Then

$$m_a = \frac{1}{n_a}\sum_{i \in C_a} x_i, m_b = \frac{1}{n_b}\sum_{i \in C_b} x_j,$$

and the centroid linkage distance is

$$d_{centroid}(C_a, C_b) = ||m_a - m_b|| = \sqrt{\sum_k (m_{a,k} - m_{b,k})^2}$$

Algorithm (Centroid Linkage):

1. Start with each observation as its own cluster.
2. At each step, compute d_centroid(Ca, Cb) for all pairs of clusters.
3. Merge the pair of clusters whose **centroids are closest** (smallest d_centroid).
4. Recompute centroids for the new clusters and repeat.

This method is described in multivariate analysis texts such as Anderson (2000) and Johnson & Wichern (2001).

Worked Example with the 5 Observations

Observations in $\mathbb{R}^2$:

- X1 = (1, 2)
- X2 = (2, 1)
- X3 = (3, 2)
- X4 = (7, 8)
- X5 = (8, 9)

We use **Euclidean distance**:

$$d(X_i, X_j) = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2}$$

Step 1: Distance Matrix Between All Pairs

Compute d(Xi, Xj) for all i< j.

1. d(X1, X2)

$d(X1, X2) = \sqrt{[(1-2)^2 + (2-1)^2]}$

$= \sqrt{[(-1)^2 + (1)^2]}$

$= \sqrt{(1+1)} = \sqrt{2} \approx 1.414$

2. d(X1, X3)

$d(X1, X3) = \sqrt{[(1-3)^2 + (2-2)^2]}$

$= \sqrt{[(-2)^2 + 0^2]}$

$= \sqrt{(4)} = 2.000$

3. d(X1, X4)

$d(X1, X4) = \sqrt{[(1-7)^2 + (2-8)^2]}$

$= \sqrt{[(-6)^2 + (-6)^2]}$

$= \sqrt{(36+36)} = \sqrt{72} \approx 8.485$

4. d(X1, X5)

$d(X1, X5) = \sqrt{[(1-8)^2 + (2-9)^2]}$

$= \sqrt{[(-7)^2 + (-7)^2]}$

$= \sqrt{(49+49)} = \sqrt{98} \approx 9.899$

5. d(X2, X3)

$d(X2, X3) = \sqrt{[(2-3)^2 + (1-2)^2]}$

$= \sqrt{[(-1)^2 + (-1)^2]}$

$= \sqrt{(1+1)} = \sqrt{2} \approx 1.414$

6. d(X2, X4)

$d(X2, X4) = \sqrt{[(2-7)^2 + (1-8)^2]}$

$= \sqrt{[(-5)^2 + (-7)^2]}$

$= \sqrt{(25+49)} = \sqrt{74} \approx 8.602$

7. d(X2, X5)

$d(X2, X5) = \sqrt{[(2-8)^2 + (1-9)^2]}$

$= \sqrt{[(-6)^2 + (-8)^2]}$

$= \sqrt{(36+64)} = \sqrt{100} = 10.000$

8. d(X3, X4)

$d(X3, X4) = \sqrt{[(3-7)^2 + (2-8)^2]}$

$= \sqrt{[(-4)^2 + (-6)^2]}$

$= \sqrt{(16+36)} = \sqrt{52} \approx 7.211$

9. d(X3, X5)

$d(X3, X5) = \sqrt{[(3 - 8)^2 + (2 - 9)^2]}$

$= \sqrt{[(-5)^2 + (-7)^2]}$

$= \sqrt{(25 + 49)} = \sqrt{74} \approx 8.602$

10. d(X4, X5)

$d(X4, X5) = \sqrt{[(7 - 8)^2 + (8 - 9)^2]}$

$= \sqrt{[(-1)^2 + (-1)^2]}$

$= \sqrt{(1 + 1)} = \sqrt{2} \approx 1.414$

Distance matrix (rounded to 3 decimals):

	X1	**X2**	**X3**	**X4**	**X5**
X1	0.000	1.414	2.000	8.485	9.899
X2	1.414	0.000	1.414	8.602	10.000
X3	2.000	1.414	0.000	7.211	8.602
X4	8.485	8.602	7.211	0.000	1.414
X5	9.899	10.000	8.602	1.414	0.000

Step 2: Hierarchical Clustering with Centroid Linkage

Initially, each observation is its own cluster:

- C1 = {X1}
- C2 = {X2}
- C3 = {X3}
- C4 = {X4}
- C5 = {X5}

For a cluster C with points Xi, the **centroid** is

$m(C) = (1 / |C|) \Sigma_{i \in C} Xi.$

The **cluster distance** is:

$d\_centroid(Ca, Cb) = || m(Ca) - m(Cb) ||.$

For singletons, $m(\{Xi\}) = Xi$, so initially $d\_centroid(\{Xi\}, \{Xj\}) = d(Xi, Xj)$.

Stage 1: First Merge

From the distance matrix, the smallest non-zero distances are:

- d(X1, X2) = 1.414
- d(X2, X3) = 1.414
- d(X4, X5) = 1.414

We have a tie. We can choose one pair.

Assume we first merge X1 and X2.

Merge: C1 = {X1}, C2 = {X2} → C12 = {X1, X2}

Cluster size: |C12| = 2

Centroid of C12:

m(C12) = ((1 + 2)/2 , (2 + 1)/2) = (1.5, 1.5)

Current clusters:

- C12 = {X1, X2}
- C3 = {X3}
- C4 = {X4}
- C5 = {X5}

Now compute **centroid distances** from C12 to the remaining singletons:

1. d_centroid(C12, C3)

m(C3) = X3 = (3, 2)

Difference: (1.5 − 3, 1.5 − 2) = (−1.5, −0.5)

|| m(C12) − m(C3) ||

= √[(−1.5)² + (−0.5)²]

= √(2.25 + 0.25) = √2.5 ≈ 1.581

2. d_centroid(C12, C4)

m(C4) = X4 = (7, 8)

Difference: (1.5 − 7, 1.5 − 8) = (−5.5, −6.5)

|| m(C12) − m(C4) ||

= √[(−5.5)² + (−6.5)²]

= √(30.25 + 42.25)

= √72.5 ≈ 8.515

3. d_centroid(C12, C5)

m(C5) = X5 = (8, 9)

Difference: (1.5 − 8, 1.5 − 9) = (−6.5, −7.5)

|| m(C12) − m(C5) ||

= √[(−6.5)² + (−7.5)²]

= √(42.25 + 56.25)

= √98.5 ≈ 9.925

Distances among C3, C4, C5 are still the original point distances:

- d(C3, C4) = d(X3, X4) ≈ 7.211
- d(C3, C5) = d(X3, X5) ≈ 8.602
- d(C4, C5) = d(X4, X5) ≈ 1.414

Updated inter-cluster distances (Stage 1):

	C12	**C3**	**C4**	**C5**
C12	–	1.581	8.515	9.925
C3	1.581	–	7.211	8.602
C4	8.515	7.211	–	1.414
C5	9.925	8.602	1.414	–

Stage 2: Second Merge

The smallest distance now is:

- d(C4, C5) = 1.414

So we merge C4 and C5.

Merge: C4 = {X4}, C5 = {X5} → C45 = {X4, X5}

Cluster size: |C45| = 2

Centroid of C45:

m(C45) = ((7 + 8)/2 , (8 + 9)/2) = (7.5, 8.5)

Current clusters:

- C12 = {X1, X2}
- C3 = {X3}
- C45 = {X4, X5}

Now compute distances:

1. d_centroid(C12, C3) (unchanged from before)

d(C12, C3) ≈ 1.581

2. d_centroid(C3, C45)

m(C3) = (3, 2), m(C45) = (7.5, 8.5)

Difference: (3 − 7.5, 2 − 8.5) = (−4.5, −6.5)

|| m(C3) − m(C45) ||

= √[(−4.5)² + (−6.5)²]

= √(20.25 + 42.25)

= √62.5 ≈ 7.906

3. d_centroid(C12, C45)

m(C12) = (1.5, 1.5), m(C45) = (7.5, 8.5)

Difference: (1.5 − 7.5, 1.5 − 8.5) = (−6, −7)

|| m(C12) − m(C45) ||

= √[(−6)² + (−7)²]

= √(36 + 49)

= √85 ≈ 9.220

Updated inter-cluster distances (Stage 2):

	C12	**C3**	**C45**
C12	–	1.581	9.220
C3	1.581	–	7.906
C45	9.220	7.906	–

Stage 3: Third Merge

The smallest distance is:

- d(C12, C3) ≈ 1.581

So we merge C12 and C3.

Merge: C12 and C3 → C123 = {X1, X2, X3}

Cluster size: |C123| = 3

Centroid of C123:

$$m(C123) = \left(\frac{1+2+3}{3}, \frac{2+1+2}{3}\right) = (2, \frac{5}{3} \approx (2.1.667)$$

C45 remains as before with centroid (7.5, 8.5).

Now compute

d_centroid(C123, C45):

Difference: (2 − 7.5, 1.667 − 8.5) = (−5.5, −6.833)

|| m(C123) − m(C45) ||

≈ √[(−5.5)² + (−6.833)²]

≈ √(30.25 + 46.71)

≈ √76.96 ≈ 8.772

Only two clusters remain:

- C123 = {X1, X2, X3}

- C45 = {X4, X5}

Final distance between them ≈ 8.772.

Summary of merges (Centroid linkage):

Step	Merged Clusters	New Cluster	Centroid of New Cluster	Distance at Merge
1	{X1}, {X2}	C12	(1.5, 1.5)	1.414
2	{X4}, {X5}	C45	(7.5, 8.5)	1.414
3	C12, {X3}	C123	(2, 1.667)	1.581
4	C123, C45	C12345	(all points)	8.772

Note the large jump from about 1.581 to 8.772 in the last merge.

(c) 2-Cluster Solution and Interpretation

From the merge sequence:

- Up to **Step 3**, the distances between merging clusters are small (1.414, 1.414, 1.581).
- At **Step 4**, merging C123 and C45 requires a much larger distance (~8.772).

A natural **2-cluster solution** is to **cut the dendrogram** before the last big jump:

- **Cluster 1:** C123 = {X1, X2, X3}
- **Cluster 2:** C45 = {X4, X5}

Interpretation:

- Cluster {X1, X2, X3} forms a compact group in the **lower-left** region of the (x1, x2) plane.
- Cluster {X4, X5} forms a compact group in the **upper-right** region.
- The centroids of these two clusters are far apart, so centroid linkage clearly separates the data into **two well-separated clusters**, consistent with the geometry of the points.

14.4 CONCLUSION

Cluster analysis is a powerful and versatile multivariate statistical technique that helps researchers group a set of objects into clusters based on their similarity across multiple characteristics. The primary goal is to ensure that objects within the same cluster are highly similar, while objects in different clusters are significantly different. To achieve this, several clustering methods are available, each offering a distinct strategy for measuring similarity and forming clusters. Methods such as single linkage**,** complete linkage, and average linkage differ in the way they compute inter-cluster distances—whether based on the nearest neighbour, farthest neighbour, or average pairwise distance**.** Ward's method, on the other hand, focuses on minimizing within-cluster variance and tends to create compact, homogeneous groups, while the centroid method relies on the geometric center of clusters to guide the merging process.

The selection of an appropriate clustering method depends on multiple factors: the nature of the dataset, the scale and measurement of the variables, the expected cluster shapes, and the presence of noise or outliers. For example, single linkage works well for elongated clusters but may produce chaining effects, whereas complete and average linkage generate more compact and stable clusters. Ward's method is particularly effective when the goal is to minimize variability within clusters and form clusters of roughly equal size.

The process of cluster analysis involves several well-defined steps, beginning with careful selection and standardization of variables, computation of similarity or distance measures, selection of a suitable clustering algorithm, and finally, determining the optimal number of clusters. Once clusters are formed, their validity must be assessed using visual tools such as dendrograms or quantitative indices such as silhouette scores. Proper interpretation of clusters is crucial to ensure they reflect meaningful patterns rather than random groupings.
Overall, cluster analysis serves as an essential exploratory tool in many fields—including market segmentation, biology, psychology, finance, and machine learning—by revealing hidden structures and relationships within complex datasets. By applying appropriate methodological choices and thorough validation, researchers can derive insightful, data-driven classifications that support strong decision-making and deeper understanding of underlying phenomena.

14.5 SELF ASSESSMENT QUESTIONS:

- Explain the different types of clustering with suitable examples.
- Describe the basic steps involved in performing cluster analysis.
- Compare and contrast single linkage, complete linkage, and average linkage methods.
- Discuss the advantages and disadvantages of Ward's method.
- Explain how the centroid method works. What are reversals in dendrograms?
- Elaborate on the role of distance measures in cluster analysis. Give examples.

14.6 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein

Dr. Syed Jilani

LESSON -15

NON-HIERARCHICAL CLUSTERING METHODS

OBJECTIVES:

After studying this unit, you should be able to:

- To understand the concept and purpose of Non-Hierarchical Clustering methods
- To know the concept of Non-Hierarchical Clustering methods
- To acquire knowledge about significance of is Non-Hierarchical Clustering methods

STRUCTURE

15.1. INTRODUCTION

K–Means is one of the most widely used partitioning clustering techniques in multivariate data analysis. Its objective is to divide a set of *n* homogeneous observations into *k* distinct, non-overlapping groups (clusters) such that observations within a cluster are as similar as possible**,** while observations between clusters are as different as possible**.** The method is based on minimizing the within-cluster sum of squares (WCSS) and uses the Euclidean distance as the primary measure of similarity. Because of its simplicity, computational efficiency, and ability to handle large datasets, K–Means is frequently applied in data mining, pattern recognition, market segmentation, and bioinformatics.

Multidimensional Scaling (MDS) is a powerful exploratory technique used to convert a matrix of similarities or dissimilarities among a set of objects into a geometric representation in a low-dimensional space, usually 2D or 3D. The central idea of MDS is to position objects in such a way that distances on the map reflect their original dissimilarities**:** similar items appear close together, while dissimilar items appear far apart.

MDS is widely used in behavioural sciences, psychometrics, marketing (perceptual mapping), ecology, and machine learning. It accommodates both metric (interval/ratio distances) and non-metric (ordinal) data and provides an intuitive visual understanding of complex multivariate relationships.

15.2. NON-HIERARCHICAL CLUSTERING METHODS – DEFINITION

Non-hierarchical clustering methods, also called partitioning methods**,** are clustering techniques in which the dataset is divided directly into a pre-specified number of clusters (k) without forming a hierarchical structure.

Unlike hierarchical clustering, these methods do not produce a dendrogram. Instead, they assign objects to clusters based on distance or similarity measures, and iteratively update the cluster centers or cluster memberships until an optimal partition is obtained.
These methods aim to minimize within-cluster variation and maximize between-cluster separation**.**

Examples of Non-Hierarchical Methods

1. K-Means Clustering
2. K-Medoids (PAM – Partitioning Around Medoids)
3. CLARA (Clustering Large Applications)
4. CLARANS (Clustering Large Applications based on Randomized Search)

15.2.1: K-Means Clustering Method:

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into a pre-defined number of clusters. The goal is to group similar data points together and discover underlying patterns or structures within the data.

Recall the first property of clusters – it states that the points within a cluster should be similar to each other. So, our aim here is to minimize the distance between the points within a cluster.

There is an algorithm that tries to minimize the distance of the points in a cluster with their centroid – the k-means clustering technique.

K-means is a centroid-based algorithm or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid.

The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

Optimization plays a crucial role in the k-means clustering algorithm. The goal of the optimization process is to find the best set of centroids that minimizes the sum of squared distances between each data point and its closest centroid.

Here's how it works:

1. **Initialization**: Start by randomly selecting K points from the dataset. These points will act as the initial cluster centroids.
2. **Assignment**: For each data point in the dataset, calculate the distance between that point and each of the K centroids. Assign the data point to the cluster whose centroid is closest to it. This step effectively forms K clusters.
3. **Update centroids**: Once all data points have been assigned to clusters, recalculate the centroids of the clusters by taking the mean of all data points assigned to each cluster.

4. **Repeat**: Repeat steps 2 and 3 until convergence. Convergence occurs when the centroids no longer change significantly or when a specified number of iterations is reached.
5. **Final Result**: Once convergence is achieved, the algorithm outputs the final cluster centroids and the assignment of each data point to a cluster.

Objective of k means Clustering

The main objective of k-means clustering is to partition your data into a specific number (k) of groups, where data points within each group are similar and dissimilar to points in other groups. It achieves this by minimizing the distance between data points and their assigned cluster's center, called the centroid.

Here's an objective:

- **Grouping similar data points:** K-means aims to identify patterns in your data by grouping data points that share similar characteristics together. This allows you to discover underlying structures within the data.
- **Minimizing within-cluster distance:** The algorithm strives to make sure data points within a cluster are as close as possible to each other, as measured by a distance metric (usually Euclidean distance). This ensures tight-knit clusters with high cohesiveness.
- **Maximizing between-cluster distance:** Conversely, k-means also tries to maximize the separation between clusters. Ideally, data points from different clusters should be far apart, making the clusters distinct from each other.

How to Apply K-Means Clustering Algorithm?

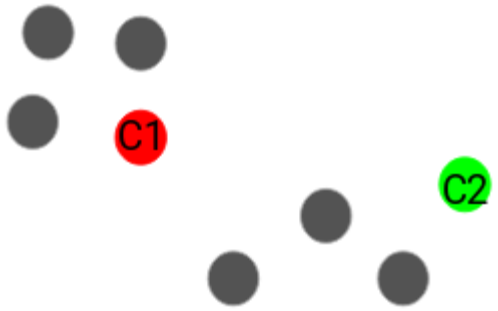
Let's now take an example to understand how K-Means actually **works**:

Time needed: 10 minutes

We have these 8 points, and we want to apply k-means to create clusters for these points.

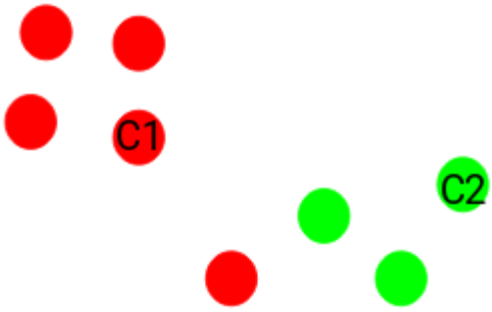
Here's how we can do it.

1. **Choose the number of clusters *k***
   The first step in k-means is to pick the number of clusters, k.
2. **Select k random points from the data as centroids**
   Next, we randomly select the centroid for each cluster. Let's say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:



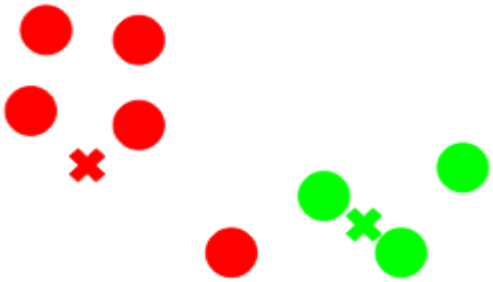
Here, the red and green circles represent the centroid for these clusters.

3. **Assign all the points to the closest cluster centroid**
   Once we have initialized the centroids, we assign each point to the closest cluster



centroid:
Here you can see that the points closer to the red point are assigned to the red cluster, whereas the points closer to the green point are assigned to the green cluster.

4. **Recompute the centroids of newly formed clusters**
   Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:



Here, the red and green crosses are the new centroids.

5. **Repeat steps 3 and 4**

We then repeat steps 3 and 4:

The step of computing the centroid and assigning all the points to the cluster based on their distance from the centroid is a single iteration.

Stopping Criteria for K-Means Clustering
There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

6. Centroids of newly formed clusters do not change
7. Points remain in the same cluster
8. Maximum number of iterations is reached

We can stop the algorithm if the centroids of newly formed clusters are not changing. Even after multiple iterations, if we are getting the same centroids for all the clusters, we can say that the algorithm is not learning any new pattern, and it is a sign to stop the training.

Another clear sign that we should stop the training process is if the points remain in the same cluster even after training the algorithm for multiple iterations.
Finally, we can stop the training if the maximum number of iterations is reached. Suppose we have set the number of iterations as 100. The process will repeat for 100 iterations before stopping.

15.3 MULTIDIMENSIONAL SCALING (MDS)

Multidimensional Scaling (MDS) is a multivariate technique used to visually represent the similarity or dissimilarity among a set of objects.

It maps high-dimensional data into a low-dimensional space (usually 2D or 3D) such that:

- Similar objects are placed close to each other.
- Dissimilar objects are placed far apart**.**

MDS is commonly used in:

- Psychology and behavioural sciences
- Marketing (perceptual mapping)
- Ecology and genetics
- Classification and clustering diagnostics

1. Basic Idea of MDS

- Suppose we have n objects and a matrix of pairwise distances (dissimilarities):
  $D = \left[d_{ij} \right],\ i, j = 1 \& \ n$
  MDS attempts to find points $x_1, x_2 \ldots\ldots\ldots x_n$ in a lower-dimensional space such that:
  $\left\| x_i - x_j \right\| \approx d_{ij}$

Algorithms:

- Classical MDS (Torgerson-Gower)
- Metric Scaling (Kruskal)

2. **Non-metric MDS**

- Only the rank order of dissimilarities is preserved.
- Perfect for ordinal or non-metric distances (Likert ratings, preferences).

Goal:

$\left\| x_i - x_j \right\| \approx f\left(d_{ij} \right)$

where f is a monotonic transformation.

Used extensively in perceptual mapping and psychological research.

3. Classical (Torgerson-Gower) MDS

Classical MDS works directly with the distance matrix**.**

Step 1: Start with distance matrix D^2

Compute squared distances:

$$D^2 = \left[d_{ij}^2 \right]$$

Step 2: Double-centering

Convert the distance matrix into a scalar product matrix B:

$$B = \frac{1}{2} J D^2 J$$

Where

$$J = I - \frac{1}{n} ee^1$$

is the centering matrix.

Step 3: Obtain eigenvalues and eigenvectors

If

$$B = V \nabla V^1$$

then:

- Λ= Lambda = diagonal matrix of eigenvalues
- V = matrix of eigenvectors

Step 4: Form the configuration

Coordinates in a k-dimensional space:

$$X_k = V_k \nabla_k^{1/2}$$

This gives the best-fitting low-dimensional representation**.**

4. Stress and Goodness-of-Fit

Kruskal's Stress Formula

$$Stress = \sqrt{\sum_{i<j} \left(\hat{d}_{ij} - d_{ij} \right)^2} \Big/ \sum_{i<j} d_{ij}^2$$

Where:

- d_{ij} observed dissimilarity
- $\hat{d}_{ij}$ reproduced distance

Rules of thumb

Stress Value	Interpretation
< 0.05	Excellent fit
0.05–0.10	Good
0.10–0.20	Fair
> 0.20	Poor fit

Applications of MDS:

- **Marketing**: perceptual maps of brands (taste similarity, quality)
- **Psychology**: similarity of stimuli, personality traits
- **Sociology**: social distance, attitude analysis
- **Bioinformatics**: genetic distance visualization
- **Machine Learning**: visualizing high-dimensional clusters

Advantages:

- Works with a distance or dissimilarity matrix directly.
- Enables visualization of high-dimensional relationships.
- Non-metric MDS handles ordinal data.
- Flexible and widely applicable.

Limitations:

- Sensitive to local minima (for non-metric MDS).
- Computation can be heavy for very large n.
- Interpretation of axes is often subjective.
- Requires a good metric of dissimilarity.

15.4 CONCLUSION

K–Means is an efficient and conceptually simple clustering tool for partitioning a dataset into *k* homogeneous groups. By iteratively updating cluster centroids and minimizing within-cluster variance, it produces compact and well-separated clusters. However, it is sensitive to initial seed selection and assumes spherical cluster shapes, which may limit performance for complex or non-linear structures. Despite these limitations, it remains a fundamental and widely applied clustering technique due to its speed, scalability, and interpretability.

MDS offers a flexible framework for visualizing the hidden structure of multivariate data by mapping objects onto a low-dimensional coordinate system that preserves their pairwise distances as faithfully as possible. It simplifies complex similarity relationships into an interpretable spatial form, making patterns, groupings, and underlying dimensions readily apparent. Although computationally intensive for large datasets and somewhat subjective in interpreting dimensions, MDS remains a valuable tool for exploratory data analysis, perceptual mapping, and evaluating clustering and classification results.

15.5 SELF ASSESSMENT QUESTIONS:

1. Explain the steps involved in the K–Means clustering algorithm.
2. What are the main differences between metric and non-metric Multidimensional Scaling (MDS)?
3. Discuss the advantages and limitations of the K–Means clustering method.

15.6 SUGGESTED READING BOOKS:

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein

Dr. Syed Jilani

LESSON -16
PRINCIPLE COMPONENT ANALYSIS

Learning Objectives

To understand the concept and purpose of Principle Component Analysis (PCA).
To learn the Mathematical Derivation and Computation of Principle Components.
To study the properties and computation of Principle Components.

STRUCTURE

16.1. INTRODUCTION

Suppose $X_1, X_2, ..., X_p$ are the given random variables. Then, principle component analysis (P.C.A) is concerned with explaining the variance-covariance structure of the variables through a few standardized linear combinations (SLC) of the original variables (we call a linear combination $l_1X_1 + l_2X_2 + ... + l_pX_p$ as an SLC if $\sum_i l_i^2 = 1$).

Algebraically, principal components (PCs) are particular standard linear combinations (SLCs) of the components of the original pattern and geometrically, these LCs represent the selection of new coordinate system obtained by rotating the original system with X_1, X_2,…, X_p as the coordinate axes. The new axes represent the directions with maximum variability an provide a simpler and more parsimonious (avoiding of excess) description of the covariance structure. As we shall see, principle components depend solely on the covariance matrix (or the correlation matrix) of the random variables $X_1, X_2, ..., X_p$. Their development does not require a multivariate normal assumption.

The general objections of P.C.A are
(i) Data-reduction and
(ii) Interpretation.

Although p components required to reproduce the total system variability, often much of this variability can be accounted by a small number ‘k’(<p) of the principal components. If there is almost as much information in the k components as there is in the original ‘p’ variables, then the ‘k’ principal components replace the original ‘p’ components of the pattern. And the

original data set consisting of n measurements on p-component pattern is reduced to one consisting of n measurements on k-principal component pattern. In other words, PCA reduces the dimensionality of the given data, losing as little information as possible. This technique was developed by Hotelling(1933).

An analysis of principle components often reveals relationships that were not previously suspected and there by allows interpretations that would not ordinarily result. In other words, the key problem is the interpretation of the principle components.

PCs may be inputs to a multiple regression analysis or cluster analysis. Moreover, (scaled) principle components are one factoring of the covariance matrix for the factor analysis model. Suppose we consider a sample of ***n*** students and they are asked to write five papers mechanics (X_1), vectors (X_2), algebra (X_3), analysis(X_4) and statistics (X_5). The examination in the first two papers is conducted in the closed book system, where as in the remaining three papers in the open book system.

Thus, we have totally '5n' observations so that ***n*** observations on each paper. One question which can be asked concerning this data is how the results on the five different papers should be combined to produce an overall scare various answers are possible. One obvious answer would be to use the overall mean that is the linear combination $(X_1 + X_2 + X_3 + X_4 + X_5)/5$. But, can one do better than this? This is one of the questions that principle component analysis seeks to answer.

If $\underset{\sim}{X}$ is a random vector with mean $\underset{\sim}{\mu}$ and variance – covariance matrix Σ, then the principle component transformation is the transformation.

$$\underset{\sim}{X} \rightarrow \underset{\sim}{Y} = \Omega' (\underset{\sim}{X} - \underset{\sim}{\mu}) \rightarrow (1)$$

where, Ω is orthogonal matrix, such that

$$\Omega'\Sigma\Omega = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots\ldots\ldots\ldots, \lambda_p), \quad \lambda_1 \geq \lambda_2 \geq \ldots\ldots\ldots\ldots\ldots \geq \lambda_p > 0$$

The strict positivity of the eigen values λ_i is guaranteed if, Σ is positive definite. The i^{th} principle component of $\underset{\sim}{X}$ may be defined as the i^{th} element of the vector $\underset{\sim}{Y}$, namely as $Y_i = \underset{\sim}{\omega}_i' (\underset{\sim}{x} - \underset{\sim}{\mu})$, where $\underset{\sim}{\omega}_i$ is the i^{th} column of Ω and may be called the i^{th} vector of principle components leadings.

16. 2 DEFINITION OF PRINCIPLE COMPONENT

If **X** is a pattern (random vector) with covariance matrix **Σ,** then the **first PC** is defined as the SLC of **X** given by

$$Y_1 = \boldsymbol{\omega}_1' \mathbf{X} = \omega_{11} X_1 + \omega_{12} X_2 + \ldots + \omega_{1p} X_p$$

where $\boldsymbol{\omega}_1 = \begin{pmatrix} \omega_{11} \\ \omega_{12} \\ . \\ . \\ \omega_{1p} \end{pmatrix}$

such that v(Y_1) is larger than the variance of any other SLC Y= $\boldsymbol{\alpha}'\mathbf{X}$ that is $V(Y_1) \geq V(Y)$.
In other words, Y_1 has the largest variance among all SLCs of **X**.
The second PC Y_2 of **X** is defined as the SLC of **X** given by

$Y_2 = \boldsymbol{\omega}_2'\mathbf{X}$

which is uncorrelated with Y_1 (the first PC) and $V(Y_2) \leq V(Y_1)$.
In general the k^{th} PC Y_k of **X** is defined as the SLC of **X** given by

$$Y_k = \boldsymbol{\omega}_k'\mathbf{X}$$

which is uncorrelated with first k-1 PCs and $V(Y_k) \leq V(Y_i)$ for i = 1,2,…,k-1.

16.3 DERIVATION OF THE PRINCIPLE COMPONENTS

Suppose $\underset{\sim}{X}$ is px1 random vector with mean vector $\underset{\sim}{\mu}$ and covariance-matrix Σ i.e., $\underset{\sim}{X} \sim (\underset{\sim}{\mu},\Sigma)$, then by definition, the first principle component is the SLC of $\underset{\sim}{X}$ which has largest variance among all SLC's of $\underset{\sim}{X}$. Thus, we should seek a LC of $\underset{\sim}{X}$ viz.,

$Y = \underset{\sim}{\omega}'\underset{\sim}{X} \rightarrow (1)$

with largest variance,

$V(Y) = \underset{\sim}{\omega}'\, \mathbf{V}(\underset{\sim}{X})\, \underset{\sim}{\omega} = \underset{\sim}{\omega}'\, \boldsymbol{\Sigma}\, \underset{\sim}{\omega} \rightarrow (2)$

such that $\underset{\sim}{\omega}'\underset{\sim}{\omega} = 1$.

Thus, we have to maximize (2) subject to the condition

$\underset{\sim}{\omega}'\underset{\sim}{\omega} = 1 \rightarrow (3)$

which is equivalent to maximizing the function,

$\phi(\underset{\sim}{\omega},\lambda) = \underset{\sim}{\omega}'\boldsymbol{\Sigma}\underset{\sim}{\omega} - \lambda(\underset{\sim}{\omega}'\underset{\sim}{\omega}-1) \rightarrow (4)$

w.r.t $\underset{\sim}{\omega}$ and λ, where 'λ' is a Lagranges multiplier. This implies to solve the equations,

$$\frac{\partial \phi}{\partial \underset{\sim}{\omega}} = \underset{\sim}{0} \Rightarrow \Sigma\underset{\sim}{\omega} = \lambda\underset{\sim}{\omega}$$

i.e., $(\Sigma-\lambda I)\underset{\sim}{\omega} = \underset{\sim}{0} \rightarrow (5)$

$$\frac{\partial \phi}{\partial \lambda} = 0 \Rightarrow \underset{\sim}{\omega}'\underset{\sim}{\omega} = 1 \rightarrow (6)$$

Using (5) & (6), from (2), we get, $V(Y) = \lambda\underset{\sim}{\omega}'\underset{\sim}{\omega} = \lambda \rightarrow (7)$

From (5), to have a non-zero solution for $\underset{\sim}{\omega}$, we must have,

$|\Sigma-\lambda I| = 0 \rightarrow (8)$

We know that (8) is a characteristic equation, and ' λ ' is a latent root andfrom (5), $\underset{\sim}{\omega}$ is the corresponding latent vector of the equation. But, weknow that, solving (8) for ' λ ' gives p-latent roots (positive),

$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ..... \geq \lambda_p \geq 0 \rightarrow (9)$

With the corresponding latent vectors, $\omega_1, \omega_2, ..., \omega_p$ respectively,

i.e., we have from (5), $\Sigma \underset{\sim}{\omega}_i = \lambda_i \underset{\sim}{\omega}_i$, i= 1,2,……p $\rightarrow (9.a)$

Since λ_1 is the largest latent root among all latent roots and $\underset{\sim}{\omega}_1$ is the corresponding latent vector.

From (1) and (7), $Y_1 = \underset{\sim}{\omega}_1' \underset{\sim}{X}$, is the first principle component with variance,

$V(Y_1) = \lambda_1$.

Let us denote the first principle component by Y_1. Now, $Y_1 = \underset{\sim}{\omega}_1' \underset{\sim}{X} \rightarrow (10)$

$V(Y_1) = \lambda_1 \rightarrow (11)$

Now, let us show that for $2 \leq k \leq p$, $Y_K = \underset{\sim}{\omega}_K' \underset{\sim}{X} \rightarrow (12)$

is the K^{th} principle component with variance, $V(Y_K) = \lambda_K \rightarrow (13)$

By definition, Y_K should uncorrelated with $Y_1, Y_2, ...Y_{K-1}$, which can be easily verified as follows (for J = 1,2,……..k-1).

$Cov(Y_k, Y_i) = Cov(\boldsymbol{\omega_k' X, \omega_i' X})$

$= \boldsymbol{\omega_k'}$ Cov $(\mathbf{X, X'})$ $\boldsymbol{\omega}_i$

$= \boldsymbol{\omega_k' \Sigma \omega_i}$

$= \lambda_i \boldsymbol{\omega_k' \omega_i}$ [From(5)]

$= 0$ ($\because \underset{\sim}{\omega}_K$ & $\underset{\sim}{\omega}_J$ are orthogonal vectors)

Also by definition, the k^{th} PC Y_k has largest variance than $Y_{k+1}, \ldots, Y_p$ which can also be verifying from (13).

$\lambda_K \geq \lambda_{K+1} \geq ..... \geq \lambda_P \geq 0$

$\Rightarrow V(Y_K) \geq V(Y_{K+1}) \geq ..... \geq V(Y_P) \geq 0$.

Hence the proof.

Remark :-

The above result may be asked as no standard linear combination (SLC) of $\underset{\sim}{X}$ has a variance larger than λ_1 ,the variance of first principle combination.

From the above result, we may say that construction (derivation) of principle components of a given random vector $\underset{\sim}{X}$ is equivalent to theproblem of the construction (derivation) of the latent roots and latent vectorsof the variance-covariance matrix Σ of $\underset{\sim}{X}$ in case of known Σ.

Note:-

(1) If Σ is not known, we may construct the principle component of the random vector $\underset{\sim}{X}$ based on the sample variance – covariance matrix or sample correlation matrix.

(2) If the population correlation matrix ' ρ ' is given, we may use it in place of Σ to construct the principle components.

(3) The principle components of the random vector $\underset{\sim}{X}$ derived frompopulation (sample) correlation matrix are different from the principle components derived from population (sample) covariance matrix.

16.4 PROPERTIES OF PRINCIPLE COMPONENTS

Property 1. Sum of the variances of all p.c's equal to the trace of Σ.

OR

Sum of the variances of all PCs is equal to the sum of the variances of the components of original pattern (or equal to the trace of covariance matrix of the pattern).

Proof :- Let $Y_1, Y_2, \ldots\ldots\ldots\ldots, Y_p$ are the p.c's obtained from random variable $\underset{\sim}{X}$.

Let us denote, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots\ldots\ldots\ldots, \lambda_p)_{p\times p}$

$$\Omega = (\underset{\sim}{\omega}_1, \underset{\sim}{\omega}_2, \ldots\ldots\ldots\ldots, \underset{\sim}{\omega}_p)_{p\times p}$$

where, λ_i's are the latent roots and $\underset{\sim}{\omega}_i$'s are the latent vectors of

the covariance matrix Σ of the random variable $\underset{\sim}{X}$. Then, we have,

$\Omega'\Sigma\Omega = \Lambda$.

$\Rightarrow \text{Tr}(\Lambda) = \text{Tr}(\Omega'\Sigma\Omega)$

$\Rightarrow \lambda_1+\lambda_2+\ldots\ldots\ldots\ldots\ldots+\lambda_p = \text{Tr}(\Sigma\Omega\Omega') = \text{Tr}(\Sigma I)$ ($\because \Omega$ is orthogonal matrix)

$\Rightarrow V(Y_1)+V(Y_2)+\ldots\ldots\ldots\ldots\ldots+V(Y_p) = \text{Tr}(\Sigma)$ $\quad (\because \lambda_i = V(Y_i))$

$\Rightarrow V(Y_1)+V(Y_2)+\ldots\ldots\ldots\ldots\ldots+V(Y_p) = V(X_1)+V(X_2)+\ldots\ldots\ldots\ldots\ldots+V(X_p)$

Hence Proved.

Property 2. Product of the variances of PC's is equal to the determinant of Σ i.e., $|\Sigma|$(or generalized variance).

OR

The generalized variance of **Y** is equal to the generalized variance of **X.** That is $|\Lambda| = |\boldsymbol{\Sigma}|$.

Proof :- Let $Y_1, Y_2, \ldots\ldots\ldots\ldots, Y_p$ are the PC's obtained from random vector $\underset{\sim}{X}$.

Let us denote $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots\ldots\ldots\ldots, \lambda_p)_{p\times p}$

$$\Omega = (\underset{\sim}{\omega}_1, \underset{\sim}{\omega}_2, \ldots\ldots\ldots\ldots, \underset{\sim}{\omega}_p)_{p\times p}$$

where, λ_i's are the latent roots and $\underset{\sim}{\omega}_i$'s are the latent vectors of the covariance matrix Σ of the random vector $\underset{\sim}{X}$. Then, we have,

We have $\Omega'\Sigma\Omega = \Lambda$

$\Rightarrow |\Lambda| = |\Omega'\Sigma\Omega|$

$\Rightarrow \lambda_1\lambda_2\ldots\ldots\ldots\ldots\lambda_p = |\Sigma\Omega\Omega'|$

$\Rightarrow V(Y_1)V(Y_2)\ldots\ldots\ldots\ldots V(Y_p) = |\Sigma I| = |\Sigma|$

Hence the proof.

Property 3. The sum of the first k eigen values divided by the sum of all eigen values

$$\frac{\lambda_1+\lambda_2+\ldots\ldots\ldots\ldots\ldots+\lambda_k}{\lambda_1+\lambda_2+\ldots\ldots\ldots\ldots\ldots+\lambda_p} = \frac{\lambda_1+\lambda_2+\ldots\ldots\ldots\ldots\ldots+\lambda_k}{\text{Tr}(\Sigma)}$$

represents the 'Proportion of total variation' explained by the first K principle components.

Property 4. The principle components of a random vector are not scale invariant. It is one disadvantage of principle component analysis.

Theorem :-An orthogonal transformation $\underset{\sim}{Y} = C\underset{\sim}{X}$ of a random vector $\underset{\sim}{X}$ leaves invariant the generalized variance and the sum of the variance of the components.

Proof :- We have given $\underset{\sim}{X}$ is the original random vector and $\underset{\sim}{Y}$ is the transformed random variable using the orthogonal matrix C.

Now, we have to show that, $|\text{cov}(\underset{\sim}{X}, \underset{\sim}{X}')| = |\text{cov}(\underset{\sim}{Y}, \underset{\sim}{Y}')|$,

where $|\ |$ is determinant and $\sum_{i=1}^{p} V(X_i) = \sum_{i=1}^{p} V(Y_i)$

since, 'C' is orthogonal we have, $C'C = CC' = I \quad \rightarrow (1)$

Now, $\text{cov}(Y, Y') = \text{cov}(C\underset{\sim}{X}, (C\underset{\sim}{X})')$

$= C\,\text{cov}(\underset{\sim}{X}, \underset{\sim}{X}')C'$

$= C\Sigma C'$

$\Rightarrow |\text{cov}(\underset{\sim}{Y}, \underset{\sim}{Y}')| = |C\Sigma C'|$

$= |C||\Sigma||C'|$

$= |\Sigma||CC'|$

$= |\Sigma||I|$ ($\because$ from (1))

$= |\Sigma|$

$= |\mathrm{cov}(\underset{\sim}{X},\underset{\sim}{X}')|$

$\Rightarrow$ generalized variance of $\underset{\sim}{Y}$ = generalized variance of $\underset{\sim}{X}$.

We have, $\sum_{i=1}^{p} V(X_i) = Tr(\Sigma)$

$= Tr(\Sigma I)$

$= Tr(\Sigma CC')$ ($\because$ from (1))

$= Tr(C\Sigma C')$

$= \sum_{i=1}^{p} V(Y_i)$ ($\because C\Sigma C'$ is covariance matrix of $\underset{\sim}{Y}$)

$\Rightarrow$ Sum of the variances of original variables (total population variance)

= Sum of variances of principle components.

$= \lambda_1+\lambda_2+.....+\lambda_p$.

Note :-The above theorem may be stated as follows. The generalized variance of the vector of principle components is the generalized variance of the original vector and the sum of the variances of the principle components is the sum of the variances of the original variates.

Results :-If $\underset{\sim}{X}$ is a random vector with covariance matrix Σ and $Y_i = \underset{\sim}{\omega}_i'\underset{\sim}{X}$ is the i^{th} principle component of the random vector $\underset{\sim}{X} = (X_1,X_2,.....,X_p)'$,then the correlation coefficient between i^{th} principle component and J^{th} originalvariable (that correlation coefficient between Y_i and X_J) is given by

$\rho_{Y_i,X_J} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{JJ}}}\omega_{iJ}$ i,J = 1,2,…….,p, where $\sigma_{JJ} = V(X_J)$

λ_i is i^{th} largest root of Σ and ω_{iJ} is J^{th} component of $\underset{\sim}{\omega}_i$, when $\underset{\sim}{\omega}_i$ is thelatent vector of Σ corresponding to λ_i.

Proof :- Denote $\underset{\sim}{l}_J = \begin{pmatrix} 0 \\ 0 \\ : \\ J \\ 0 \\ 0 \end{pmatrix} \rightarrow J^{th}$ position

Now, $X_J = \underset{\sim}{l}_J' \underset{\sim}{X}$. Also we have given, $Y_i = \underset{\sim}{\omega}_i' \underset{\sim}{X} \rightarrow (1)$

$$\rho_{Y_i, X_J} = \frac{\text{cov}\,(\underset{\sim}{\omega}_i' \underset{\sim}{X}\,,\, \underset{\sim}{l}_J' \underset{\sim}{X})}{\sqrt{\lambda_i \sigma_{JJ}}} \frac{\text{cov}\,(Y_i, X_J)}{\sqrt{V(Y_i) V(X_J)}}$$

$= \frac{\text{cov}\,(\underset{\sim}{\omega}_i' \underset{\sim}{X}\,,\, \underset{\sim}{l}_J' \underset{\sim}{X})}{\sqrt{\lambda_i \sigma_{JJ}}}$ ($\because V(Y_i) = \lambda_i$, the i^{th} larger latent rootof Σ, using (1) &

σ_{JJ} is J^{th} diagonalelement of Σ).

$$= \frac{\underset{\sim}{\omega}_i' \,\text{cov}\,(\underset{\sim}{X}\,,\, \underset{\sim}{X}')\, \underset{\sim}{l}_J}{\sqrt{\lambda_i}\sqrt{\sigma_{JJ}}}$$

$$= \frac{\underset{\sim}{\omega}_i' \,\Sigma\, \underset{\sim}{l}_J}{\sqrt{\lambda_i}\sqrt{\sigma_{JJ}}} \rightarrow (2)$$

($\because \Sigma$ is covariance matrix of $\underset{\sim}{X}$)

Since $\underset{\sim}{\omega}_i$ is the latent vector of Σ corresponding to latent root λ_i, we have

$\Sigma \underset{\sim}{\omega}_i = \lambda_i \underset{\sim}{\omega}_i$

$\Rightarrow \underset{\sim}{\omega}_i' \Sigma = \lambda_i \underset{\sim}{\omega}_i'$ (Taking transpose & $\Sigma = \Sigma'$) $\rightarrow (3)$

Using (3) in (2), we get,

$$\rho_{Y_i, X_J} = \frac{\lambda_i \,\underset{\sim}{\omega}_i'\, \underset{\sim}{l}_J}{\sqrt{\lambda_i}\sqrt{\sigma_{JJ}}}$$

$= \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{JJ}}} (\omega_{i1}0 + \omega_{i2}0 + \ldots\ldots\ldots\ldots + \omega_{iJ}1 + \omega_{iJ+1}0 + \ldots\ldots\ldots + 0)$

$= \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{JJ}}} \omega_{iJ}$ for i,J = 1,2,……,p

Hence the proof.

16.5 COMPUTATION OF PRINCIPLE COMPONENTS

From the given data, we have to calculate the sample dispersion matrix $\mathbf{S}$. Now, we can compute the first principal component Y_1 and its variance

$Y_1 = \omega_{11}X_1 + \omega_{12}X_2 + ... + \omega_{1p}X_p = \boldsymbol{\omega}_1'\underset{\sim}{\mathbf{x}}$, where $\boldsymbol{\omega}_1'\boldsymbol{\omega}_1 = 1$ and $Var(Y_1) = \lambda_1$ (1)

from the following iterative equation.

$\mathbf{S}_1\underset{\sim}{\boldsymbol{\omega}}_1 = \lambda_1\underset{\sim}{\boldsymbol{\omega}}_1$, where $\mathbf{S}_1$=$\mathbf{S}$ (2)

Equation (2) can be written as an iterative equation given by

$\lambda_1^{(i+1)}\underset{\sim}{\boldsymbol{\omega}}_1^{(i+1)} = \underset{\sim}{\boldsymbol{\beta}} = \mathbf{S}_1\underset{\sim}{\boldsymbol{\omega}}_1^{(i)}$, i=0,1,... (3)

From Eq.(3), we can compute

$\lambda_1^{(i+1)} = \sqrt{\underset{\sim}{\boldsymbol{\beta}}'\underset{\sim}{\boldsymbol{\beta}}}$ and $\underset{\sim}{\boldsymbol{\omega}}_1^{(i+1)} = \underset{\sim}{\boldsymbol{\beta}} / \lambda_1^{(i+1)}$ (4)

Now, the above iterative equation (3) will be initiated with $\underset{\sim}{\boldsymbol{\omega}}_1^{(0)} = \begin{pmatrix} 1 \\ 0 \\ . \\ . \\ . \\ 0 \end{pmatrix}_{px1}$

Eq. (3) will be solved iteratively until two successive values of λ_1 (computed using Eq. (4)) do agree upto 4 decimal places. The corresponding $\underset{\sim}{\boldsymbol{\omega}}_1$ is the first principal component and it's variance is λ_1.

Computing second principal component:

We have to replace the sample dispersion matrix $\mathbf{S}_1$ with the adjusted dispersion matrix $\mathbf{S}_2$, given by,

$\mathbf{S}_2$=$\mathbf{S}_1 - \lambda_1\boldsymbol{\omega}_1\boldsymbol{\omega}_1'$ (5)

Now, the second PC can be computed in the same way as computed the first PC by solving the following equation iteratively.

$\mathbf{S}_2\underset{\sim}{\boldsymbol{\omega}}_2 = \lambda_2\underset{\sim}{\boldsymbol{\omega}}_2$ (6)

Thus, the second PC and it's variance are given by

$Y_2 = \boldsymbol{\omega}_2'\underset{\sim}{\mathbf{x}}$, and $Var(Y_2) = \lambda_2$ (7)

Computing third principal component:

We have to replace the matrix $\mathbf{S}_2$ with the adjusted dispersion matrix $\mathbf{S}_3$, given by,

$$\mathbf{S}_3 = \mathbf{S}_2 - \lambda_2 \boldsymbol{\omega}_2 \boldsymbol{\omega}_2' \quad (8)$$

Now, the third PC can be computed in the same way as computed the second PC by solving the following equation iteratively.

$$\mathbf{S}_3 \underset{\sim}{\boldsymbol{\omega}}_3 = \lambda_3 \underset{\sim}{\boldsymbol{\omega}}_3 \quad (9)$$

Thus, the third PC and its variance are given by

$$Y_3 = \boldsymbol{\omega}_3' \underset{\sim}{\mathbf{x}}, \quad \text{and} \quad Var(Y_3) = \lambda_3 \quad (10)$$

Similarly, one can compute the remaining PCs iteratively.

16.6 SUMMARY

This lesson introduces the concept and methodology of Principal Component Analysis (PCA), a powerful statistical tool used for dimensionality reduction and data interpretation by transforming a set of correlated variables into a smaller set of uncorrelated principal components (PCs) that preserve as much variability as possible. PCA achieves this by identifying standardized linear combinations (SLCs) of the original variables, which are determined by the eigen values and eigenvectors of the covariance or correlation matrix. The derivation of principal components involves maximizing variance under orthonormal constraints using Lagrange multipliers, resulting in mutually uncorrelated PCs. The lesson also explores important properties of PCA, such as the sum and product of the variances of PCs, their lack of scale invariance, and the invariance of generalized variance under orthogonal transformations. Applications of PCA are highlighted in fields like regression, clustering, and factor analysis, and the mathematical derivation of the correlation between original variables and principal components is provided to aid interpretation.

16.7 SELF ASSESSMENT QUESTIONS

1. What is the primary objective of Principal Component Analysis (PCA)?
2. Define a standardized linear combination (SLC) and explain its role in PCA.
3. How is the first principal component of a random vector derived?
4. What condition ensures that the eigen values of a covariance matrix are strictly positive?
5. How does PCA achieve dimensionality reduction while preserving information?
6. Explain why principal components are uncorrelated.
7. What is meant by the “proportion of total variation explained” in PCA?
8. Why are principal components not scale-invariant, and what are the implications of this?
9. What is the relationship between the variance of PCs and the trace of the covariance matrix?
10. How can the correlation between an original variable and a principal component be computed?
11. Describe the iterative method for computing principal components from sample data.

16.8 SUGGESTED READINGS

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Principal Component Analysis** by I.T. Jolliffe
4. **Modern Multivariate Statistical Techniques** by Alan J. Izenman
5. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
6. **The Elements of Statistical Learning** by Trevor Hastie, Robert Tibshirani, and Jerome Friedman

Dr. S. BHANU PRAKASH

LESSON -17

CANONICAL CORRELATION ANALYSIS

Learning Objectives

To understand the concept and purpose of Canonical Correlation Analysis (CCA).
To define and interpret canonical variates.
To learn the Mathematical Derivation
To compute canonical correlation using covariance matrices.

STRUCTURE

17.1. INTRODUCTION

Canonical correlations analysis seeks to identify and quantity the associations between two sets of variables. Holding (1936), who initially developed the technique, provided the example of relating arithmetic seed and arithmetic power to reading speed and reading power. Other example includes relating governmental policy variables with economic goal variables and relating college" performance "variables with pre college "Achievement" variables.

A statistical method for examining the connections between two sets of variables is canonical correlation analysis, or CCA. CCA explores the underlying structure of two multi-variable datasets and looks into how they relate to one another overall, in contrast to simple correlation, which measures the relationship between two individual variables. When examining complicated data, where variables within each set may be interrelated and straightforward pairwise correlations may not provide the whole picture, this is especially helpful.

CCA achieves this by creating canonical variates – new, composite variables formed by taking weighted sums (linear combinations) of the original variables within each of the two sets. The primary goal is to find these weights in a way that maximizes the correlation between the resulting canonical variates from the two different sets. The strength of these relationships between the paired canonical variates is then quantified by canonical correlations, which are essentially the correlation coefficients between these newly formed variables.

Consider the following scenario: a researcher wishes to investigate the relationship between a collection of personality qualities and a collection of academic performance metrics. The

degree of correlation between "overall personality" and "academic aptitude" could be determined by using CCA to find latent dimensions or canonical variates that represent these concepts.

17. 2 DEFINITION OF CANONICAL VARIATE

Canonical variates are new composite variables formed by taking linear combinations (weighted sums) of the original variables within each of two distinct sets. Canonical Correlation Analysis (CCA) seeks to determine the weights for each variable that maximize the correlation between the canonical variates derived from these two sets. For instance, given one set of physiological variables (such as weight and waist circumference) and another set of exercise variables (like the number of chin-ups and sit-ups), CCA might generate a "body size" canonical variate from the physiological measures and an "exercise capacity" canonical variate from the exercise measures.

Suppose $X_1, X_2, ..., X_p$ and $Y_1, Y_2, ..., Y_q$ are two sets of p and q variables then the variates

$U = a_1X_1 + a_2X_2 + ... + a_pX_p$ and $V = b_1Y_1 + b_2Y_2 + ... + b_qY_q$

are said to be canonical variates if the coefficients *a*'s and *b*'s are selected such that the correlation between U and V is maximum.

17. 3 DEFINITION OF CANONICAL CORRELATIONS

Canonical correlations are the correlation coefficients that quantify the strength of the linear relationship between corresponding pairs of canonical variates derived from the two sets of original variables. The objective of Canonical Correlation Analysis (CCA) is to identify the linear combinations that maximize the correlation between these pairs of canonical variates.

For example, for two sets of variables, $X = (X_1, X_2, ..., X_p)$ and $Y = (Y_1, Y_2, ..., Y_q)$, the first canonical correlation ρ_1 is defined as:

$$\rho_1 = \max_{a,b} corr(a^T X, b^T Y) = \max_{a,b} \frac{\text{cov}(a^T X, b^T Y)}{\sqrt{V(a^T X)V(b^T Y)}}$$

where '*a*' and '*b*' are weight vectors (coefficients) for linear combinations.

Subsequent canonical correlations are found similarly, with the constraint that the new canonical variates are uncorrelated with the previously found canonical variates.

17.4 DERIVATION AND COMPUTATION OF CANONICAL CORRELATION

Suppose the random vector $\underset{\sim}{X}$ of p components has the covariance matrix $\sum_\lambda$ (which is assumed to be positive definite). We partition $\underset{\sim}{X}$ into two sub vectors of p_1 and p_2 components respectively, that is

$$\underset{\sim}{X} = \begin{pmatrix} \underset{\sim}{X}^{(1)} \\ \underset{\sim}{X}^{(2)} \end{pmatrix}_{p\times 1} \quad \underset{\sim}{X}^{(1)} \text{ is } p_1 \times 1 \text{ and } \underset{\sim}{X}^{(2)} \text{ is } p_2 \times 1 \; (p = p_1 + p_2) \qquad (1)$$

For convenience we shall assume $p_1 \le p_2$. The covariance matrix $\sum$ is partitioned similarly p_1 and p_2 rows and columns.

$$\Sigma = \begin{pmatrix} \sum_{(p_1 \times p_2)} 11 & \sum_{(p_1 \times p_2)} 12 \\ \sum_{(p_1 \times p_2)} 21 & \sum_{(p_1 \times p_2)} 22 \end{pmatrix} \quad (2)$$

Now we are interested in measures of association between first group of p_1 variables $\underset{\sim}{X}^{(1)}$ and the second group of p_2 variables $\underset{\sim}{X}^{(2)}$.The $p_1 p_2$ elements of $\sum_{12}$ measure the association between two groups. When p_1 and p_2 are relatively large, interpreting the elements of $\sum_{12}$ collectively is ordinarily hopeless. Moreover, it is often linear combinations of variables that are interesting and useful for predictive purposes.The main task of canonical correlation analysis is to summarize the associations between the $\underset{\sim}{X}^{(1)}$ and $\underset{\sim}{X}^{(2)}$ sets in terms of a few carefully chosen conversances (or correlations) rather than the $p_1 p_2$ covariance in $\sum_{12}$.

Consider an arbitrary linear combination

$$U = \underset{\sim}{\alpha}' \underset{\sim}{X}^{(1)} \quad (3)$$

Of the components of $\underset{\sim}{X}^{(1)}$ and an arbitrary linear combination.

$$V = \underset{\sim}{\beta}' \underset{\sim}{X}^{(2)} \quad (4)$$

Of the components of $\underset{\sim}{X}^{(2)}$. Since the correlation of multiple of U and a multiple of V is the same as the correlations of U and V, we can make an arbitrary normalizations of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$.

We therefore require $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ to be such that

$$V(u) = \text{cov}(\underset{\sim}{\alpha}' \underset{\sim}{X}^{(1)}, \underset{\sim}{X}^{(1)'} \underset{\sim}{\alpha}) = \underset{\sim}{\alpha}' \text{cov}(\underset{\sim}{X}^{(1)}, \underset{\sim}{X}^{(2)}) \underset{\sim}{\alpha} = \underset{\sim}{\alpha}' \sum_{11} \underset{\sim}{\alpha} = 1 \quad (5)$$

and $V(V) = \underset{\sim}{\beta}' \sum_{22} \underset{\sim}{\beta} = 1 \quad (6)$

Then the correlation between U and V is

$$\text{cov}(U,V) = \frac{\text{cov}(U,V)}{\sqrt{V(U)V(U)}}$$

$$= \text{cov}(U,V) \quad \text{(using (5) \& (6))}$$

$$= \text{cov}(\underset{\sim}{\alpha}' \underset{\sim}{X}^{(1)}, \underset{\sim}{X}^{(2)'} \underset{\sim}{\beta})$$

$$= \underset{\sim}{\alpha}' \sum_{12} \underset{\sim}{\beta} \quad (7)$$

Now, we shall $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ such that $\text{cov}(U,V)$ is as large as possible. Thus, the algebraic problem is to find $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ such that $\underset{\sim}{\alpha}' \sum_{12} \underset{\sim}{\beta}$ is maximum subject to (5) &(6) consider.

$$\phi = \underset{\sim}{\alpha}' \sum_{12} \underset{\sim}{\beta} - \frac{1}{2}\rho(\underset{\sim}{\alpha}' \sum_{11} \underset{\sim}{\alpha} - 1) - \frac{1}{2}\lambda(\underset{\sim}{\beta}' \sum_{22} \underset{\sim}{\beta} - 1) \quad (8)$$

When ρ and λ are Lagrange multiplies. Now, our problem is to solve the following equations simultaneously or jointly.

$$\frac{\partial \phi}{\partial \underset{\sim}{\alpha}} = \underset{\sim}{0} \Rightarrow \Sigma_{12}\,\underset{\sim}{\beta} - \rho \Sigma_{11}\,\underset{\sim}{\alpha} = \underset{\sim}{0} \quad (9)$$

$$\frac{\partial \phi}{\partial \underset{\sim}{\beta}} = \underset{\sim}{0} \Rightarrow \Sigma'_{12}\,\underset{\sim}{\alpha} - \lambda \Sigma_{22}\,\underset{\sim}{\beta} = \underset{\sim}{0} \quad (10)$$

remultiplying (9) by $\underset{\sim}{\alpha}'$ and (10) by $\underset{\sim}{\beta}'$ and using (5) & (6),we get

$$\underset{\sim}{\alpha}' \Sigma_{12}\,\underset{\sim}{\beta} = \rho$$

$$\underset{\sim}{\beta}' \Sigma'_{12}\,\underset{\sim}{\alpha} = \lambda \Rightarrow \underset{\sim}{\alpha}' \Sigma_{12}\,\underset{\sim}{\beta} = \lambda' = \lambda$$

That $\rho = \underset{\sim}{\alpha}' \Sigma_{12}\,\underset{\sim}{\beta} = \lambda$ (11)

Thus, the equations (9) and (10) because

$$\Sigma_{12}\,\underset{\sim}{\beta} - \rho \Sigma_{11}\,\underset{\sim}{\alpha} = \underset{\sim}{0}_{p_1 \times 1} \quad (12)$$

And $\Sigma_{21}\,\underset{\sim}{\alpha} - \rho \Sigma_{22}\,\underset{\sim}{\beta} = \underset{\sim}{0}_{(p_2 \times 1)}$ (13)

Since Σ is positive definite, being principle diagonal sub matrices Σ_{11} and Σ_{22} are also positive definite and hence Σ_{11}^{-1} and Σ_{22}^{-1} are also exist.

On the above equation can be solved simultaneously to get solutions for ρ, $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ as follows:

re-multiplying (12) by ρ and (13) by Σ_{22}^{-1}we get

$$\Sigma_{12}(\rho \underset{\sim}{\beta}) = \rho^2 \Sigma_{11}\,\underset{\sim}{\alpha} \quad (14)$$

$$\Sigma_{22}^{-1} \Sigma_{21}\,\underset{\sim}{\alpha} = \rho \underset{\sim}{\beta} \quad (15)$$

Using (15) and (14) we get

$$\Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}\,\underset{\sim}{\alpha} = \rho^2 \Sigma_{11}\,\underset{\sim}{\alpha} \quad (16)$$

Since Σ_{11} is positive finite, we may write

$\Sigma_{11} = \Sigma_{11}^{\frac{1}{2}} \Sigma_{11}^{\frac{1}{2}}$, when $\Sigma_{11}^{\frac{1}{2}}$ is square root matrix. (17)

using (17) in (16), we get

$$\Sigma_{22}^{-1} \Sigma_{21}\,\underset{\sim}{\alpha} = \rho \underset{\sim}{\beta} \quad (18)$$

$$\left|\Sigma_{11}^{\frac{1}{2}}\right|\left|\Sigma_{11}^{\frac{1}{2}}\right| = \left|\Sigma_{11}\right| \neq 0$$

$\left|\sum_{11}^{\frac{1}{2}}\right| \neq 0 \Rightarrow \sum_{11}^{-\frac{1}{2}}$ inverse of $\left|\sum_{11}^{\frac{1}{2}}\right|$ exists

re-multiplying (18) with $\sum_{11}^{-\frac{1}{2}}$ we get

$$\sum_{11}^{\frac{1}{2}} \sum_{12} \sum_{22}^{-1} \sum_{21} \sum_{11}^{-\frac{1}{2}} \underset{\sim}{a} = \rho^2 \underset{\sim}{a} \quad (19)$$

Where $\underset{\sim}{a} = \sum_{11}^{\frac{1}{2}} \underset{\sim}{\alpha}$ (20)

$$\Rightarrow \underset{\sim}{\alpha} = \sum_{11}^{-\frac{1}{2}} \underset{\sim}{a} \quad (21)$$

(19)causes written as

$$(\sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-1} \sum_{21} \sum_{11}^{-\frac{1}{2}} - \rho^2 I)\underset{\sim}{a} = 0 \quad (22)$$

Thus ρ^2 is a latent root and $\underset{\sim}{a}$ is the corresponding latent vector of

$$\sum_{11}^{-\frac{1}{2}} \sum_{12} \sum_{22}^{-\frac{1}{2}} \sum_{21} \sum_{11}^{-\frac{1}{2}} \quad (23)$$

Once after getting $\underset{\sim}{a}, \underset{\sim}{\alpha}$ can be obtained using (21).

Now, from equations (15),

$$\underset{\sim}{\beta} = \rho^{-1} \sum_{22}^{-1} \sum_{21} \underset{\sim}{\alpha} \quad (24)$$

Where $\underset{\sim}{\alpha}$ is given by (21)

Let $\rho_1^2 \geq \rho_2^2 \geq ..... \geq \rho_{\rho_1}^2$ are the eigen values of (23) and $\underset{\sim}{a}_1, \underset{\sim}{a}_2, ... \underset{\sim}{a}_{p_1}$ are the corresponding eigen vectors of (23), then the i^{th} pair of canonical variables are given by

$U_i = \underset{\sim}{\alpha}_i' \underset{\sim}{X}^{(1)}$ and $V_i = \underset{\sim}{\beta}_i' \underset{\sim}{X}^{(2)}$ (25)

When $\underset{\sim}{\alpha}_i = \sum_{11}^{-\frac{1}{2}} \underset{\sim}{a}_i$ (from (21))

And $\underset{\sim}{\beta}_i = \rho_i^{-1} \sum_{22}^{-1} \sum_{21} \underset{\sim}{\alpha}_i$ (from (24))

And the canonical correlation of the ith pair of canonical variables is given by ρ_i

Since $\underset{\sim}{a}_1, \underset{\sim}{a}_2, ... \underset{\sim}{a}_{p_1}$ are orthogonal vectors, we can see easily $\underset{\sim}{\alpha}_1, \underset{\sim}{\alpha}_2, ... \underset{\sim}{\alpha}_{p_1}$ and $\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, ... \underset{\sim}{\beta}_{p_1}$ are also orthogonal vectors. As a consequence,

$\text{cov}(U_i, U_j) = \underset{\sim}{\alpha}_i' \sum_{11} \underset{\sim}{\alpha}_i$ (from (25))

$= \underset{\sim}{a}_i' \underset{\sim}{a}_j = 0$ for $i \neq j$ (26)

(Since $\underset{\sim}{a}_i, \underset{\sim}{a}_j'$ are orthogonal)

Similarly,

$\text{cov}(V_i, V_j) = \underset{\sim}{\beta}_i' \Sigma_{22} \underset{\sim}{\beta}_j$

$= \underset{\sim}{\beta}_i' \rho_j^{-1} \Sigma_{21} \underset{\sim}{\alpha}_j$ (from (13))

$= \rho_j^{-1} \rho_i \underset{\sim}{\alpha}_i' \Sigma_{11} \underset{\sim}{\alpha}_j$ (from (12))

$= 0$ (from (26))

And $\text{cov}(U_i, V_j) = \underset{\sim}{\alpha}_i' \Sigma_{12} \underset{\sim}{\beta}_j = \rho_j \underset{\sim}{\alpha}_i' \Sigma_{11} \underset{\sim}{\alpha}_j$ (from (12))

$= 0$ (from (26))

Thus, a canonical variable is uncorrelated with any other canonical variable except its paired canonical variable. More clearly, the canonical variable U_i is highly correlated with V_i and uncorrelated with all other canonical variables $U_j = (j \neq i)$ and $V_j = (j \neq i)$

For example, from **Johnson, R. A., & Wichern, D. W. (2007).** *Applied Multivariate Statistical Analysis* (6th ed., **pp. 545–555**) a small dataset with 3 observations and two sets of variables are considered:

Set *X: X₁, X₂* Set *Y: Y₁, Y₂*

Data Table:

Obs	X_1	X_2	Y_1	Y_2
1	2	3	4	6
2	4	5	6	8
3	6	7	8	10

Compute the means:

$\bar{X}_1 = \frac{2+4+6}{3} = 4, \quad \bar{X}_2 = \frac{3+5+7}{3} = 5 \; \bar{Y}_1 = \frac{4+6+8}{3} = 6, \quad \bar{Y}_2 = \frac{6+8+10}{3} = 8$

Subtract the means from each value to get the centered matrices:

$X_c = \begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 2 & 2 \end{bmatrix}, \quad Y_c = \begin{bmatrix} -2 & -2 \\ 0 & 0 \\ 2 & 2 \end{bmatrix}$

sample covariance matrix

$S = \frac{1}{n-1} X^T X$ (for centered data)

With n = 3, we use 1/2 as the scaling factor.

Now compute each covariance matrix:

$S_{XX} = \frac{1}{2} X_c^T X_c = \frac{1}{2} \begin{bmatrix} (-2)^2 + 0^2 + 2^2 & (-2)(-2) + 0(0) + 2(2) \\ (-2)(-2) + 0(0) + 2(2) & (-2)^2 + 0^2 + 2^2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 8 & 8 \\ 8 & 8 \end{bmatrix} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$

Similarly,

$$S_{YY} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}, \quad S_{XY} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$$

We compute the canonical correlation matrix:

$$M = S_{XX}^{-1} S_{XY} S_{YY}^{-1} S_{YX}$$

But since: $S_{XX} = S_{YY} = S_{XY} = \begin{bmatrix} 4 & 4 \\ 4 & 4 \end{bmatrix}$

This matrix is singular (non-invertible), as:

$$\det(S_{XX}) = 4 \cdot 4 - 4 \cdot 4 = 0$$

Hence, only one canonical correlation can be found.

Therefore,

- First Canonical Correlation $\rho_1 = 1.0$ (perfect correlation)
- Second Canonical Correlation $\rho_2 = 0$ (undefined due to rank deficiency)

17.5 SUMMARY

This lesson introduces Canonical Correlation Analysis (CCA), a technique developed by Hotelling (1936) to examine relationships between two sets of variables. CCA creates canonical variates—linear combinations of variables in each set—that are maximally correlated. The method identifies these combinations by solving an eigenvalue problem derived from the partitioned covariance matrix. The resulting canonical correlations measure the strength of association between the variates. The lesson also includes a numerical example and highlights that when covariance matrices are singular, only one valid canonical correlation may exist.

17.6 SELF ASSESSMENT QUESTIONS

1. Who originally developed Canonical Correlation Analysis (CCA), and what was the initial example provided?
2. Define canonical variates.
3. Define canonical correlation.
4. What are canonical variates, and how are they formed?
5. What are canonical correlations, and how do they relate to canonical variates?
6. Derive the canonical correlations step by step.
7. In the numerical example provided, why was only one canonical correlation found?

17.7 SUGGESTED READINGS

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Methods of Multivariate Analysis(2nd ed., Section 11.1–11.5)** by Rencher, A. C.
4. **Multivariate Analysis** by Mardia, K. V., Kent, J. T., & Bibby, J. M.

Dr. S. BHANU PRAKASH

LESSON -18
FACTOR ANALYSIS

Learning Objectives

To understand the concept and purpose of Factor Analysis.
To learn the main estimation methods for factor loading and commonalties.
To recognize the properties and challenges of factor analysis.
To understand how to evaluate the adequacy of factor models.

STRUCTURE

18.1. INTRODUCTION

Factor analysis is a mathematical model which attempts to explain the correlation between a large set of variables in terms of a small number of underlying unobservable factors. In other words, the essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable, random quantities called factors. Basically, the factor model is motivated by the following argument. Suppose variables can be grouped by their correlations. That is all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. Factor analysis was originally developed by psychologists interested in psychometric measurement.

Arguments over the psychological interpretations of several early studies and the lack of powerful computing facilities impelled its developments a statistical method. The advent of high-speed computers has generated a renewed interest in the theoretical and computational aspects of factor analysis. Most of the original techniques have been abandoned and early

controversies resolved in the make of recent developments.It is still true that each application of the technique must be examined on its own merits to determine its success.

Factor analysis can be considered as an extension of principal component analysis. Both can be viewed as attempts to approximate the covariance matrix $\mathbf{\Sigma}$.However, the approximation based on the factor analysis model is more elaborate. The primary question is factor analysis is whether the data are consistent with a prescribed structure.

In order to get a feel for the subject we first describe a simple example.
Example 1 (Spearmon,1904): In children examinations performance in classics (x_1),French (x_2) and English (x_3).It is found that the correlation matrix is given by

$$\begin{pmatrix} 1 & 0.83 & 0.78 \\ & 1 & 0.67 \\ & & 1 \end{pmatrix}$$

Although this matrix has full rank, its dimensionality can be effectively reduced from *p*=3 to *p*=1by expressing the three variables as follows

$$\left.\begin{aligned} x_1 &= \lambda_1 f + u_1 \\ x_2 &= \lambda_2 f + u_2 \\ x_3 &= \lambda_3 f + u_3 \end{aligned}\right\} \qquad (1)$$

In these equations f is an underlying 'common factor' and λ_1, λ_2 and λ_3 are known as factor loadings.The terms u_1, u_2 and u_3 represent random disturbance terms. The common factor may be interpreted as 'general ability' (or 'intelligence') and u_i will have small variance x_i is closely related to general ability. The variation is u_i consist of two parts which we shall not try to disentangle in practice.First,this variance represents the extent to which an individual's ability at classics, say, differs from his general ability and second it represents the fact that the examination is only an approximate measure of his ability in the subject.The model defined in (1) can be generalized to include k >1 common factors.

18.2 ORTHOGONAL FACTOR MODEL

The observable random vector $\underset{\sim}{\mathbf{x}}$ with p component has mean $\underset{\sim}{\boldsymbol{\mu}}$ and covariance matrix $\mathbf{\Sigma}$. The factor model postulates that $\underset{\sim}{\mathbf{x}}$ is linearly dependent upon a few unobservable random variables $F_1,\ldots,F_k$ called common factors and *p* additional sources of variations $u_1, u_2, \cdots, u_p$ called random disturbances or error or specific factors. In particular, the factor analysis model is

$$\left.\begin{aligned} X_1 &= \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1k}F_k + u_1 \\ X_2 &= \mu_2 + \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2k}F_k + u_2 \\ &\vdots \\ X_p &= \mu_p + \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pk}F_k + u_p \end{aligned}\right\} \quad (1)$$

(or) in matrix notation.

$$\underset{\sim}{\mathbf{x}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \boldsymbol{\Lambda}_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)} \quad (2)$$

Where $\underset{\sim}{\mathbf{X}} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, \underset{\sim}{\boldsymbol{\mu}} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \underset{\sim}{\mathbf{F}} = \begin{pmatrix} F_1 \\ F_2 \\ \vdots \\ F_k \end{pmatrix}, \underset{\sim}{\mathbf{u}} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix}, \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1k} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2k} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pk} \end{pmatrix}$

The matrix $\boldsymbol{\Lambda}$ is called the matrix of factor loadings, where λ_{ij} is the loading of i^{th} variable (X_i) on j^{th} factor (F_j).Note that the i^{th} specific factor u_i is associated only with the i^{th} response X_i.

The p deviations $X_1-\mu_1, X_2-\mu_2, \cdots, X_p-\mu_p$ are expressed in terms of k+p random variables $F_1, F_2, \cdots, F_k, \mu_1, \mu_2, \cdots, \mu_p$ are unobservable.

From (1),it may be noted that each equation looks like a multiple regression equation but for one exception. The common factor in (1) $F_1, F_2, \cdots, F_k$ are unobservable where as in multiple regression equation the independent variables can be observed.This distinguishes the factor model from the multivariate regression model. With so many unobservable quantities (k+p) a direct verification of the factor model (1) from observations on $X_1, \ldots, X_p$ is hopeless. However, with some additional assumptions about the random vectors $\underset{\sim}{F}$ and $\underset{\sim}{\boldsymbol{\mu}}$, the model in(2) implies certain covariance relationships,which can be checked.

We assume that

$$\left.\begin{aligned} &E(\underset{\sim}{\mathbf{F}}) = \underset{\sim}{0},\ V(\underset{\sim}{\mathbf{F}}) = E(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}') = I_{k\times k} \\ &E(\underset{\sim}{\mathbf{u}}) = \underset{\sim}{0},\ V(\underset{\sim}{\mathbf{u}}) = E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}') \\ &\qquad = \boldsymbol{\psi} = \text{doag}(\psi_1, \psi_2, \cdots, \psi_p) \\ &\text{and } \text{cov}(\underset{\sim}{\mathbf{u}}, \underset{\sim}{\mathbf{F}}) = E(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{F}}') = 0_{p\times k} \end{aligned}\right\} \quad (3)$$

The model (2) with the assumptions (3) is called the 'Orthogonal Factor model'

The assumption (3) implies the following implicit assumptions.

- All common factors are standardized to have variance 1 and uncorrelated with one another $\left(V(\underset{\sim}{\mathbf{F}})=I\right)$
- All specific factors (random disturbances) are have zero means and uncorrelated $\left(V(\underset{\sim}{\mathbf{u}})=\boldsymbol{\psi}=\text{doag}(\psi_1,\psi_2,\cdots,\psi_p)\right)$
- Common factor and specific factor are uncorrelated $\left(\text{cov}(\underset{\sim}{\mathbf{u}},\underset{\sim}{\mathbf{F}})=0\right)$.

The Orthogonal model with k common factors

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)}=\underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)}+\Lambda_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)}+\underset{\sim}{\mathbf{u}}_{(p\times 1)} \qquad (4)$$

Where $\mathbf{X}_i = i^{th}$ response variable

$\boldsymbol{\mu}_i$ = mean of $\mathbf{X}_i$

λ_{ij} = loading of $\mathbf{X}_i$ on $\mathbf{F}_j$

$\mathbf{F}_j$ = j^{th} common factor

$\mathbf{u}_i$ = i^{th} specific factor.

The unobservable random vectors $\underset{\sim}{\mathbf{F}}$ and $\underset{\sim}{\mathbf{u}}$ satisfy $\underset{\sim}{\mathbf{F}}$ and $\underset{\sim}{\mathbf{u}}$ are independent

$E(\underset{\sim}{\mathbf{u}})=0$, $V(\underset{\sim}{\mathbf{u}})=\psi=\text{doag}(\psi_1,\psi_2,\cdots,\psi_p)$

The orthogonal factor model implies a covariance structure for $\underset{\sim}{\mathbf{X}}$.From the model in (4), we have

$$\begin{aligned}\left(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}\right)\left(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}\right)' &= \left(\Lambda\underset{\sim}{\mathbf{F}}+\underset{\sim}{\mathbf{u}}\right)\left(\Lambda\underset{\sim}{\mathbf{F}}+\underset{\sim}{\mathbf{u}}\right)' \\ &= \Lambda\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}'\Lambda'+\Lambda\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{u}}'+\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{F}}'\Lambda'+\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\end{aligned}$$

so that

$$\begin{aligned}\boldsymbol{\Sigma}=V\left(\underset{\sim}{\mathbf{X}}\right) &= E\left(\left(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}\right)\left(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}\right)'\right) \\ &= \Lambda E\left(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}'\right)\Lambda'+\Lambda E\left(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{u}}'\right)+E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{F}}'\right)\Lambda'+E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{u}}'\right) \\ &= \Lambda\Lambda'+\psi \qquad \text{(from(3))} \qquad (5)\end{aligned}$$

Also from the model (4),we have

$$\left(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}\right)\underset{\sim}{\mathbf{F}}'=\left(\Lambda\underset{\sim}{\mathbf{F}}+\underset{\sim}{\mathbf{u}}\right)\underset{\sim}{\mathbf{F}} \quad =\Lambda\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}'+\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{F}}'$$

$$\text{cov}\left(\underset{\sim}{\mathbf{X}},\underset{\sim}{\mathbf{F}}\right)=E\left(\left(\underset{\sim}{\mathbf{X}}-\underset{\sim}{\boldsymbol{\mu}}\right)\underset{\sim}{\mathbf{F}}'\right)=\Lambda E\left(\underset{\sim}{\mathbf{F}}\underset{\sim}{\mathbf{F}}'\right)+E\left(\underset{\sim}{\mathbf{u}}\underset{\sim}{\mathbf{F}}'\right)=\Lambda \quad \text{(From(3))} \qquad (6)$$

From the model (1), we have

$$X_i = \mu_i + \sum_{j=1}^{k} \lambda_{ij} F_j + u_i \ , \qquad i=1,2,\cdots,p$$

Covariance-structure for the orthogonal factor model

$$\left.\begin{array}{l} 1.\ V\left(\underset{\sim}{\mathbf{X}}\right) = \mathbf{\Lambda\Lambda}' + \psi \\ \quad or\ X_i = \mu_i = \sum_{j=1}^{k} \lambda_{ij} F_j + u_i \\ \quad V\left(X_i\right) = \sum_{j=1}^{k} \lambda_{ij}^2 + \psi_i \ and\ \ \mathrm{cov}\left(X_i, X_l\right) = \sum_{j=1}^{k} \lambda_{ij}\lambda_{lj} \\ 2.\ \mathrm{cov}\left(\underset{\sim}{\mathbf{X}}, \underset{\sim}{\mathbf{F}}\right) = \mathbf{\Lambda}\ \ or\ \ \mathrm{cov}\left(X_i, F_j\right) = \lambda_{ij} \end{array}\right\} \qquad (7)$$

From the above,thus $V\left(X_i\right)$ can be split into two parts.

First $h_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2$ is called the *communality* and represents the variance of X_i which is shared with the other variables via the common factors.

In particular $\lambda_{ij}^2 = \left[\mathrm{cov}\left(X_i, F_j\right)\right]^2$ represents the extent to which X_i depends on the j^{th} common factor. On the other hand ψ_i is called specific or unique variance and is due to the specific factor u_i it explains the variability in X_i not shared with other variables.

Thus from (7)

$$\underset{V\left(X_i\right)}{\sigma_{ii}} = \underbrace{\lambda_{i1}^2 + \lambda_{i2}^2 + \cdots \lambda_{ik}^2}_{\text{communality}} + \underset{\text{specific variance}}{\psi_i} \qquad (8)$$

$$\underset{V\left(X_i\right)}{\sigma_{ii}} = \underset{\text{communality}}{h_i^2} + \underset{\text{specific variance}}{\psi_i} \qquad (9)$$

so that the i^{th} communality is the sum of squares of the loadings of the i^{th} variable on k common factors.

Note**:**The validity of the k-factor model can be expressed in terms of a simple condition on $\mathbf{\Sigma}$

From (5) we have

$$\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \psi \qquad (10)$$

The converse also holds. If $\mathbf{\Sigma}$ can be decomposed into the form (10), then the k-factor model holds For $\underset{\sim}{\mathbf{X}}$. However, $\underset{\sim}{\mathbf{F}}$ and $\underset{\sim}{\mathbf{u}}$ are not uniquely determined by $\underset{\sim}{\mathbf{X}}$.

18.3 SCALE INVARIANCE PROPERTY

Statement: Factor analysis is invariant of scaling of variables

Proof: Suppose $\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda_x \underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}}$ (1)

is the factor model.

Now rescaling the variables of $\underset{\sim}{\mathbf{X}}$ is equivalent to set

$$\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\mathbf{X}}, \textit{where } \mathrm{C}=\mathrm{diag}(\mathrm{c}_1,\mathrm{c}_2,\cdots,\mathrm{c}_\mathrm{p})$$

$$= \begin{pmatrix} \mathrm{c}_1 & 0 & 0 & \cdots & 0 \\ 0 & \mathrm{c}_1 & 0 & \cdots & 0 \\ 0 & 0 & \mathrm{c}_1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathrm{c}_1 \end{pmatrix} \begin{pmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \\ \vdots \\ \vdots \\ \mathrm{X}_\mathrm{p} \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} \mathrm{c}_1\mathrm{X}_1 \\ \mathrm{c}_2\mathrm{X}_2 \\ \vdots \\ \vdots \\ \mathrm{c}_\mathrm{p}\mathrm{X}_\mathrm{p} \end{pmatrix} (2)$$

Premultiplying (1) with C we get

$$\underset{\sim}{\mathbf{Y}} = C\underset{\sim}{\boldsymbol{\mu}} + C\Lambda\underset{\sim}{\mathbf{F}} + C\underset{\sim}{\mathbf{u}}$$

$$\textit{and } \mathrm{V}(\underset{\sim}{\mathbf{Y}}) = C\Lambda_x\Lambda_x'C' + C\psi_x C'$$

$$\textit{i.e., } \Sigma_\mathrm{y} = \Lambda_y\Lambda_y' + \psi_y \, (3)$$

when $\Lambda_y = C\Lambda_x$

$$\psi_y = C\psi_x C \qquad (\because C = C')$$

$$= \mathrm{diag}(\mathrm{c}_1^2\psi_1, \mathrm{c}_2^2\psi_2, \cdots, \mathrm{c}_p^2\psi_p)$$

From (1)

$$\mathrm{V}(\underset{\sim}{\mathbf{X}}) = \Lambda_x\Lambda_x' + \psi_x$$

$$\Rightarrow \Sigma_\mathrm{x} = \Lambda_x\Lambda_x' + \psi_x \, (4)$$

But we have

$$\Sigma_\mathrm{y} = C\Sigma_x C \qquad (\because C = C')$$

$$= C\Lambda_x\Lambda_x'C' + C\psi_x\mathrm{C}$$

$$= \Lambda_y\Lambda_y' + \psi_y$$

Which is nothing but (3).

Thus the factor loading matrix Λ_y for the scaled random vector $\underset{\sim}{\mathbf{Y}}$ is obtained by scaling the factor loading matrix Λ_x of the original random vector $\underset{\sim}{\mathbf{X}}$. Similarly the specific variance matrix ψ_y for the scaled random vector $\underset{\sim}{\mathbf{Y}}$ is obtained by premultiplying and postmuliplying the specific variance matrix ψ_x of the original r.v. $\underset{\sim}{\mathbf{X}}$ by C. In other words,factor analysis(unlike principal component analysis) is unaffected by a rescaling of the variables.

18.4 NON-UNIQUENESS OF FACTOR LOADINGS PROPERTY

Statement: Non-uniqueness of factor loadings(Rotated Factors)

Proof: Let T is any $k \times k$ orthogonal matrix. So that, $TT' = T'T = I$ .Then the factor model

$$\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda \underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}} \qquad (1)$$

Can be written as

$$\begin{aligned} \underset{\sim}{\mathbf{X}} &= \underset{\sim}{\boldsymbol{\mu}} + \Lambda TT' \underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}} \\ &= \underset{\sim}{\boldsymbol{\mu}} + (\Lambda T)(T' \underset{\sim}{\mathbf{F}}) + \underset{\sim}{\mathbf{u}} \\ &= \underset{\sim}{\boldsymbol{\mu}} + \Lambda^* \underset{\sim}{\mathbf{F}}^* + \underset{\sim}{\mathbf{u}} \qquad (2) \end{aligned}$$

where , $\Lambda^* = \Lambda T$ and $\underset{\sim}{\mathbf{F}}^* = T' \underset{\sim}{\mathbf{F}}$

Since, $E(\underset{\sim}{\mathbf{F}}^*) = T'E(\underset{\sim}{\mathbf{F}}) = \underset{\sim}{\mathbf{0}}$ and $\mathrm{V}(\underset{\sim}{\mathbf{F}}^*) = T'V(\underset{\sim}{\mathbf{F}})T = T'IT = I$.

It is impossible, on the basis of observations on $\underset{\sim}{\mathbf{X}}$ to distinguish the loadings Λ from those of Λ^*. That is the factor $\underset{\sim}{\mathbf{F}}$ and $\underset{\sim}{\mathbf{F}}^* = T' \underset{\sim}{\mathbf{F}}$ have the same statistical properties and even though the loadings Λ^* are in general different from the loadings Λ, they both generate the same covariance matrix. That is

$$\Sigma = \Lambda\Lambda' + \psi \qquad (3)$$

$$= \Lambda TT' \Lambda' + \psi$$

$$= \Lambda^* \Lambda^{*\prime} + \psi \qquad (4)$$

Thus the variance-covariance matrix $\boldsymbol{\Sigma}$ can be decomposed as either (3) or (4).And if Λ is the factor loadings, then $\Lambda^* = \Lambda T$ (for any orthogonal matrix T), is also the factor loadings. However, the communalities given by the diagonal elements of $\Lambda\Lambda' = \Lambda^* \Lambda^{*\prime}$ are uneffected by the choice of T.

This determinacy in the definition of factor loadings is usually resolved by rotating (multiplying by an orthogonal matrix). The factor loadings Λ to satisfy an arbitrary constant such as $\Lambda' \psi^1 \Lambda$ is diagonal or $\Lambda' D^{-1} \Lambda$ is diagonal, $D = diag(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp})$. Where in either case the diagonal elements are written in decreasing order.. Once the loadings and

specific variances are obtained, factors are identified and estimated values for the factors themselves (called factor scores) are frequently constructed.

18.5 METHODS OF ESTIMATION

Given observations $\underset{\sim}{\mathbf{X}}_1, \underset{\sim}{\mathbf{X}}_2, ..., \underset{\sim}{\mathbf{X}}_n$ on p generally correlated variables, factor analysis seeks to the question,'Does the factor model

$$\underset{\sim}{\mathbf{X}} = \underset{\sim}{\boldsymbol{\mu}} + \Lambda \underset{\sim}{\mathbf{F}} + \underset{\sim}{\mathbf{u}} \qquad (1)$$

With a small number of factors, adequately represent the data?

In essence, we tackle this statistical model building problem by trying to verify the covariance relationship

$$\Sigma = \Lambda\Lambda' + \psi \qquad (2)$$

The sample covariance matrix S is an estimator of the unknown $\boldsymbol{\Sigma}$.If the off-diagonal elements are small or those of the sample correlation matrix R are essentially zero,the variables are not related and factor analysis will not prove useful.In these circumstances,the specific factors play the dominant role,whereas the major aim of the factor analysis is to determine a few important common factors.

If $\boldsymbol{\Sigma}$ appears to deviate significantly from diagonal matrix then a factor model can be entertained and the initial problem is one of the estimating the factor loadings $\lambda_{ij}'s$ and specific variances $\psi_i's$.We shall consider two of the most popular methods of parameter estimation.

1.Principal factor method (Analysis).

2. Maximum likelihood method (factor analysis)

The solution from either method can be rotated in order to simplify the interpretation of factors

Principal factor analysis:

We have the factor model (k-factor)

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \Lambda_{(p\times k)} \underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)}$$

(1)

Where $\underset{\sim}{\mathbf{X}}$ = p-component random vector

$\underset{\sim}{\boldsymbol{\mu}}$ = mean of $\underset{\sim}{\mathbf{X}}$

$\boldsymbol{\Lambda}$ = matrix of factor loadings

$\underset{\sim}{\mathbf{F}}$ = vector of common factors

$\underset{\sim}{\mathbf{u}}$ = p-component random vector

with covariance matrix of $\underset{\sim}{\mathbf{X}}$

$$\mathbf{\Sigma} = \mathbf{\Lambda\Lambda}' + \boldsymbol{\psi} \qquad (2)$$

Where $\boldsymbol{\psi} = V(\underset{\sim}{\mathbf{u}}) = diag\left(\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \cdots, \boldsymbol{\psi}_k\right)$.

In practical situation,since $\mathbf{\Sigma}$ is not known, $\mathbf{\Sigma}$ is replaced by its estimate the sample covariance matrix $\mathbf{S}$ which is obtained from the observations $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$.Since,factor analysis is invariant of the scaling of the variables the correlation matrix $\mathbf{R}$,computed from the observations $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ on p-variable random vector $\underset{\sim}{\mathbf{X}}$,may also be used in place of $\mathbf{S}$.

Let us suppose the data is summarized by the correlation matrix $\mathbf{R}$ so that an estimate of $\mathbf{\Lambda}$ and $\boldsymbol{\psi}$ is rough for the standardized variables.

Now our problem is to obtain the estimates of $\mathbf{\Lambda}$ and $\boldsymbol{\psi}$ from equation (2),replacing the unknown $\mathbf{\Sigma}$ with known $\mathbf{R}$ (when the variables standardized $\mathbf{\Sigma}$ is equivalent to the population correlation matrix $\boldsymbol{\rho}$).Then we have

$$\mathbf{R} = \hat{\Lambda}\hat{\Lambda}' + \hat{\boldsymbol{\psi}} \qquad (3)$$

Comparing the diagonal elements on both sides,we get

$1 = \hat{h}_i^2 + \hat{\psi}_i$ for i=1,2,---,p

$where \quad \hat{h}_i^2 = \sum_{j=1}^{k} \hat{\boldsymbol{\lambda}}_{ij}^2$

Is the preliminaryestimate of the i^{th} communality h_i^2 and may be obtained either of the following two ways:

1) The square of the multiple correlation coefficient of the i^{th} variable X_i on the remaining p-1 variables.

2) The largest absolute correlation coefficient between X_i and one of the remaining p-1 variables. i.e., $\max_{j \neq i} \left| r_{ij} \right|$

Note that the estimated communality h_i^2 is higher when X_i is highly correlated with the other as we would expect. Now $\hat{\boldsymbol{\psi}}$=diag$\left(\hat{\boldsymbol{\psi}}_i\right) = diag\left(1 - \hat{h}_i^2\right)$ has to be subtracted from $\mathbf{R}$ to obtained the matrix

$$\mathbf{R}-\hat{\boldsymbol{\psi}}=\begin{bmatrix}\hat{h}_1^2 & r_{12} & \cdots & r_{1p}\\ r_{12} & \hat{h}_2^2 & \cdots & r_{2p}\\ \vdots & \vdots & & \vdots\\ r_{1p} & r_{2p} & \cdots & \hat{h}_p^2\end{bmatrix} \tag{4}$$

Which is called the reduced correlation matrix because the 1's on the diagonal have been replaced by the estimated*communalities* $\hat{h}_i^2$.

Suppose $a_1 \geq a_2 \geq \cdots \geq a_p$ are eigen values of $\mathbf{R}-\hat{\boldsymbol{\psi}}$ and $\underset{\sim}{\boldsymbol{\omega}}_1, \underset{\sim}{\boldsymbol{\omega}}_2, \cdots, \underset{\sim}{\boldsymbol{\omega}}_p$ are the corresponding eigen vectors, then we may decompose $\mathbf{R}-\hat{\boldsymbol{\psi}}$ as

$$\mathbf{R}-\hat{\boldsymbol{\psi}}=\sum_{i=1}^{p} a_i \underset{\sim}{\boldsymbol{\omega}}_i \underset{\sim}{\boldsymbol{\omega}}_i' \tag{5}$$

Suppose the first K eigen values $a_1, a_2, \cdots, a_k$ are positive then

$$\mathbf{R}-\hat{\boldsymbol{\psi}}=\sum_{i=1}^{p} a_i \underset{\sim}{\boldsymbol{\omega}}_i \underset{\sim}{\boldsymbol{\omega}}_i' = \hat{\Lambda}\hat{\Lambda}' \tag{6}$$

Where, $\hat{\Lambda}=\left[\sqrt{a_1}\,\underset{\sim}{\boldsymbol{\omega}}_1 \quad \sqrt{a_2}\,\underset{\sim}{\boldsymbol{\omega}}_2 \quad \cdots \quad \sqrt{a_k}\,\underset{\sim}{\boldsymbol{\omega}}_k\right]_{p\times k} = \boldsymbol{\Omega}\mathbf{A}^{\frac{1}{2}}$ (7)

$\Omega=\left(\underset{\sim}{\boldsymbol{\omega}}_1 \quad \underset{\sim}{\boldsymbol{\omega}}_2 \quad \cdots \quad \underset{\sim}{\boldsymbol{\omega}}_k\right)$ and $\mathbf{A}=diag\left(a_1, a_2, \cdots, a_k\right)$ is the estimate of the factor loading matrix $\hat{\boldsymbol{\Lambda}}$ .Since, $\boldsymbol{\Omega}$ is orthogonal matrix, we may see that

$$\hat{\boldsymbol{\Lambda}}'\hat{\boldsymbol{\Lambda}}=\mathbf{A}^{1/2}\boldsymbol{\Omega}'\boldsymbol{\Omega}\mathbf{A}^{1/2}=\mathbf{A}^{1/2}\mathbf{I}\mathbf{A}^{1/2}=\mathbf{A} \tag{8}$$

Finally, the revised estimates of the specific variances are given in terms of $\hat{\boldsymbol{\Lambda}}$ by

$$\hat{\boldsymbol{\psi}}_i = 1-\sum_{j=1}^{k}\hat{\boldsymbol{\lambda}}_{ij}^2, i=1,2,\cdots,p \tag{9}$$

Where $\hat{\boldsymbol{\lambda}}_{ij}$ is the $(i,j)^{th}$ element of the estimated factor loading s matrix $\hat{\boldsymbol{\Lambda}}$ given by (7).Then the principal factor solution is permissible is all the $\hat{\boldsymbol{\psi}}_i$ are non-negative.

Thus for the k factor model (1) the principal factor estimates of the factors loading matrix $\boldsymbol{\Lambda}$ is given by (7) and the estimates of communalities h_i^2 are given by the diagonal elements of $\hat{\Lambda}\hat{\Lambda}'$.

i.e. $h_i^2=\sum_{j=1}^{k}\hat{\boldsymbol{\lambda}}_{ij}^2$ (10)

The estimates of the specific variables $\hat{\boldsymbol{\psi}}_i'$s are given by (9)

Note:

- The principal factor analysis can be performed iteratively with the communality estimates given by (10)
- becoming the initial estimates for the next stage.
- If we are given the sample covariance matrix **S**,it may be converted into **R** and then above analysis can be performed.

For example, consider the open/closed book data of the following table with correlation matrix.

$$\begin{bmatrix} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 1 & 0.610 & 0.485 & 0.437 \\ & & 1 & 0.711 & 0.665 \\ & & & 1 & 0.607 \\ & & & & 1 \end{bmatrix}$$

If k>2 then **S**<0 and the factor model is not well defined. The principal factor solutions for k=1 and k=2, where we estimate the ith communality $\hat{h}_i^2$ by $\max_j |r_{ij}|$,are given in the table.

The eigen values of the reduced correlation matrix are 2.84,0.38, 0.08,0.02 and -0.05, suggesting that the two-factor solution fits the data well.

In the above table principal factor solutions for the open/closed book data with k=1 and k=2 factors.

variable	k=1		k=2		
	$\hat{h}_i^2$	$\lambda(1)$	$\hat{h}_i^2$	$\lambda(1)$	$\lambda(2)$
1	0.417	0.646	0.543	0.646	0.354
2	0.506	0.711	0.597	0.711	0.303
3	0.746	0.864	0.749	0.864	-0.051
4	0.618	0.786	0.680	0.786	-0.249
5	0.551	0.742	0.627	0.742	-0.276

The first factor represents overall performance and for k=2, the second factor, which is much less important $(a_2 = 0.38 << 2.84 = a_1)$, represents a contrast acrossthe range $h_i^2 << 1$ for all i, and therefore a fair proportion of the variance of each variable is left unexplained by the common factor.

18.6 PRINCIPAL COMPONENT METHOD (PRINCIPAL COMPONENT SOLUTION OF THE FACTOR MODEL)

Suppose $\underset{\sim}{\mathbf{x}}_1,\ldots,\underset{\sim}{\mathbf{x}}_n$ are observations on **p** generally correlated variables and the data is summarized either into the sample correlation matrix **R**.

Let the orthogonal factor model with k common factors

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)} = \underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)} + \Lambda_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)} + \underset{\sim}{\mathbf{u}}_{(p\times 1)} \quad (1)$$

Where $\underset{\sim}{\mathbf{X}}$ = p-component random vector

$\underset{\sim}{\boldsymbol{\mu}}$ = mean of $\underset{\sim}{\mathbf{X}}$

$\boldsymbol{\Lambda}$ = matrix of factor loadings

$\underset{\sim}{\mathbf{F}}$ = vector of common factors

$\underset{\sim}{\mathbf{u}}$ = vector of random disturbances

with $\quad V(\underset{\sim}{\mathbf{X}}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi}$ (1.a)

$\boldsymbol{\psi} = V(\underset{\sim}{\mathbf{u}})$

Now the principal component method is to obtain the estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$ using the sample covariance matrix **S** or sample correlated matrix **R**.

Suppose $a_1 \geq a_2 \geq \cdots \geq a_p$ are the latent roots of **S** (or **R**) and let us consider the first 'k' roots

i.e. $a_1, a_2, \cdots, a_p$

Let $\underset{\sim}{\boldsymbol{\omega}}_1, \underset{\sim}{\boldsymbol{\omega}}_2, \cdots, \underset{\sim}{\boldsymbol{\omega}}_k$ be the corresponding latent vectors. Then the estimated matrix of factor loadings is given by

$$\hat{\boldsymbol{\Lambda}} = \left[\sqrt{a_1}\,\underset{\sim}{\boldsymbol{\omega}}_1 \quad \sqrt{a_2}\,\underset{\sim}{\boldsymbol{\omega}}_2 \quad \cdots \quad \sqrt{a_k}\,\underset{\sim}{\boldsymbol{\omega}}_k\right]_{p\times k} \quad (2)$$

and the estimated specific variances are provided by the diagonal elements of the matrix

$$\mathbf{S} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\left(\mathbf{R} - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'\right) \quad (3)$$

so that $\hat{\boldsymbol{\psi}} = \text{diag}\left(\hat{\boldsymbol{\psi}}_1, \hat{\boldsymbol{\psi}}_2, \cdots, \hat{\boldsymbol{\psi}}_k\right)$ with $\hat{\boldsymbol{\psi}}_i = s_{ii} - \sum_{j=1}^{k}\hat{\boldsymbol{\lambda}}_{ij}^2$ $\left(\hat{\boldsymbol{\psi}}_i = r_{ii} - \sum_{j=1}^{k}\hat{\boldsymbol{\lambda}}_{ij}^2\right)$ estimates of communalities are given by the diagonal elements of $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$

i.e. $h_i^2 = \sum_{j=1}^{k}\hat{\lambda}_{ij}^2$

Note:

- Consider the residual matrix $\mathbf{S} - \left(\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\psi}}\right)$ resulting from the approximation of **S** by the

principal componentsolution. The diagonal elements are zero and if the other elements are also small,we may subjectively take the 'k' factor model to be appropriate.

•The contribution to the total sample variance = $\mathrm{Tr}(\mathbf{S})$ from the j^{th} common factor is given by

$$\sum_{i=1}^{p}\hat{\lambda}_{ij}^{2}=\left(\sqrt{a_j}\ \underset{\sim}{\omega}_{j}\right)'\left(\sqrt{a_j}\ \underset{\sim}{\omega}_{j}\right)$$

$$=a_j \qquad \left(\because\ \underset{\sim}{\omega}_{j}'\underset{\sim}{\omega}_{j}=1\right)$$

= j^{th} latent root of $\mathbf{S}$ (Where is a_j the j^{th} latent root of $\mathbf{S}$)

Thus, proportion of total sample variance due to j^{th} factor

$$=\begin{cases}\dfrac{a_j}{\mathrm{Tr(S)}} & \text{for factor analysis of 'S'}\\ \dfrac{a_j}{p} & \text{for factor analysis of 'R'}\end{cases}$$

•(For worked out examples see page no's: 388-391 of Applied Multivariate Analysis by Johnson&Wichern)

18.7 MAXIMUM LIKELIHOOD FACTOR ANALYSIS

Suppose $\underset{\sim}{x}_1,\ldots,\underset{\sim}{x}_n$ are 'n' observations drawn on $\underset{\sim}{X}$ which follows population $Np\left(\underset{\sim}{\mu},\mathbf{\Sigma}\right)$ and $\underset{\sim}{X}$ is having the following k factor model

$$\underset{\sim}{\mathbf{X}}_{(p\times 1)}=\underset{\sim}{\boldsymbol{\mu}}_{(p\times 1)}+\Lambda_{(p\times k)}\underset{\sim}{\mathbf{F}}_{(k\times 1)}+\underset{\sim}{\mathbf{u}}_{(p\times 1)}$$

(1)

Where,

$\underset{\sim}{\mu}$ = mean of $\underset{\sim}{X}$

$\mathbf{\Lambda}$ = matrix of factor loadings

$\underset{\sim}{F}$ = vector of common factors

$\underset{\sim}{u}$ = vector of random disturbances

with the assumptions

$E(\underset{\sim}{F})=0=E(\underset{\sim}{u})$

$V(\underset{\sim}{F})=I$, $V(\underset{\sim}{u})=\psi=\mathrm{doag}(\psi_1,\psi_2,\cdots,\psi_p)$

$\mathrm{cov}(\underset{\sim}{F},\underset{\sim}{u})=0$

These assumptions implicitly imposes the restriction on $\mathbf{\Sigma}$ as follows

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \psi \qquad (2)$$

Since $\underset{\sim}{\mathbf{X}} \sim Np\left(\underset{\sim}{\boldsymbol{\mu}}, \boldsymbol{\Sigma}\right)$, its log -likelihood is given by

$$\log L = \frac{-n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)$$

if we with its MLE $\bar{\mathbf{X}}$, then $\log L$ becomes

$$\begin{aligned} l = \log L &= \frac{-n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{n}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)'\boldsymbol{\Sigma}^{-1}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right) \\ &= -\left(\frac{n}{2}\log\left|2\pi\boldsymbol{\Sigma}\right| + \frac{n}{2}\ \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\mathrm{S}_n)\right) \end{aligned} \qquad (3)$$

Where

$$\mathrm{S}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)'\left(\underset{\sim}{\mathbf{x}}_i - \underset{\sim}{\boldsymbol{\mu}}\right)$$

$\boldsymbol{\Sigma}$ is as given by **(2)**

Maximizing (3) is equivalent to minimizing the following function w.r.t. $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$

$$\mathrm{F}(\Lambda,\bar{\boldsymbol{\psi}}) = \log\left|\boldsymbol{\Sigma}\right| + \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\mathrm{S}_n\right) - \log\left|\mathrm{S}_n\right| - \mathrm{p} \qquad (4)$$

$\left(\because \left|\mathrm{S}_n\right| \text{and p are constants}\right)$

Since from (2), $\boldsymbol{\Lambda}$ is not uniquely determined. We have minimize (4) subject to the following uniqueness condition

$$\boldsymbol{\Lambda}'\psi^{-1}\boldsymbol{\Lambda} = \Delta, \text{a diagonal matrix} \qquad (5)$$

The MLEs $\hat{\boldsymbol{\Lambda}}$ and $\hat{\psi}$ obtained by minimizing (4) subject to (5) satisfy

$$\left(\hat{\psi}^{\frac{-1}{2}}\mathrm{S}_n\hat{\psi}^{\frac{-1}{2}}\right)\left(\hat{\psi}^{\frac{-1}{2}}\hat{\boldsymbol{\Lambda}}\right) = \left(\hat{\psi}^{\frac{-1}{2}}\hat{\boldsymbol{\Lambda}}\right)\left(1+\hat{\boldsymbol{\Delta}}\right) \qquad (6)$$

so that the j^{th} column of $\hat{\psi}^{\frac{-1}{2}}\hat{\boldsymbol{\Lambda}}$ is the (non-normalised) eigen vector of $\hat{\psi}^{\frac{-1}{2}}\mathrm{S}_n\hat{\psi}^{\frac{-1}{2}}$ corresponding to eigen value $1+\hat{\boldsymbol{\Delta}}_i$

where $\hat{\boldsymbol{\Delta}}_1 \geq \hat{\boldsymbol{\Delta}}_2 \geq \cdots \geq \hat{\boldsymbol{\Delta}}_k$

clearly,for the above, the MLE of $\hat{\boldsymbol{\Lambda}}$ can be obtained only for a given $\hat{\psi}$,whose initial value can be taken as

$$\hat{\boldsymbol{\psi}}^{(0)} = diag\left(\hat{\boldsymbol{\psi}}_1^{(0)}, \hat{\boldsymbol{\psi}}_2^{(0)}, \cdots, \hat{\boldsymbol{\psi}}_k^{(0)}\right)$$

where $\hat{\boldsymbol{\psi}}_i = \left(1 - \frac{1}{2}\frac{k}{p}\right)\left(\frac{1}{s^{ii}}\right)$

where s^{ii} is the i^{th} diagonal element of $\mathbf{S} = \dfrac{n\,\mathbf{S}_n}{(n-1)}$

The next modified value of $\hat{\psi}$ is given by

$$\hat{\boldsymbol{\psi}}^{(1)} = diag\left(\hat{\boldsymbol{\psi}}_1^{(1)}, \hat{\boldsymbol{\psi}}_2^{(1)}, \cdots, \hat{\boldsymbol{\psi}}_k^{(1)}\right)$$

Where $\hat{\boldsymbol{\psi}}^{(1)}$ is the i^{th} diagonal element of the computed matrix $\mathbf{S}_n - \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}'$

Using this $\hat{\boldsymbol{\psi}}^{(1)}$,we can obtained the revised value of $\hat{\boldsymbol{\Lambda}}$ using (6).

This procedure is to be continued until the latest estimates $\hat{\boldsymbol{\Lambda}}$ and $\hat{\boldsymbol{\psi}}$ satisfy the relation (5).

Note:Ordinarily the observations are standardized and a sample correlation matrix is factor analyzed.Of the data is summarized into a sample correlation matrix $\mathbf{R}$,then the above method of maximum likelihood factor analysis may becarried out replacing $\mathbf{S}_n$ or $\mathbf{S}$ by $\mathbf{R}$ to get the some estimates of $\boldsymbol{\Lambda}$ and $\boldsymbol{\psi}$ .This is due to the fact that the MLEs are scale invariant.

An workedout example is given in page no:394 of Applied Multivariate Statistical analysis by Richard A.Jhon and Wichern.

18.8 FACTOR ROTATION

We have

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\psi} \\ &= \boldsymbol{\Lambda}\mathrm{T}\mathrm{T}'\boldsymbol{\Lambda}' + \boldsymbol{\psi} \\ &= \left(\boldsymbol{\Lambda}\mathrm{T}\right)\left(\boldsymbol{\Lambda}\mathrm{T}\right)' + \boldsymbol{\psi} \\ &= \boldsymbol{\Lambda}^*\left(\boldsymbol{\Lambda}^*\right)' + \boldsymbol{\psi}\end{aligned}$$

where $\boldsymbol{\Lambda}^* = \boldsymbol{\Lambda}\mathrm{T}$, T is an orthogonal matrix

Thus if $\boldsymbol{\Lambda}$ is a factor loadings matrix which reproduce $\boldsymbol{\Sigma}$,then any other factor loadings matrix $\boldsymbol{\Lambda}^*$ obtained from $\boldsymbol{\Lambda}$ by an orthogonal transformation (T) have the same ability to reproduce the covariance matrix (or correlation matrix).From matrix algebra,we know that an orthogonal transformation corresponds to a rigid rotation of the coordinate axes.For this reason, an orthogonal transformation of the factor loadings and the implied orthogonal transformation of the factor is called "factor rotation".

Let $\hat{\boldsymbol{\Lambda}}$ be the $p \times k$ matrix of estimated factor loadings obtained by any method, then

$\hat{\boldsymbol{\Lambda}}^* = \hat{\boldsymbol{\Lambda}}\mathrm{T}$ where TT′=T′T=I (1)

is a $p \times k$ matrix of rotated loadings. Moreover, the estimated covariance (or correlation) matrix remains unchanged,

since $\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\psi}} = \hat{\boldsymbol{\Lambda}}^*\left(\hat{\boldsymbol{\Lambda}}^*\right)' + \hat{\boldsymbol{\psi}}$ (2)

since, the original loadings may not be readily interpretable, it is usual practice to rotate them until a " sample structure" is achieved. Ideally, we should like to see a pattern of loadings such that each variable loadings highly on a single factor and has small-to-moderate loadings on the remaining factors. Of course, it is not always possible to get this simple structure.

A convenient analytical choice of rotation is given by the "varimax method" described below:
The varimax method of orthogonal rotation was provided by **kaiser**(1958).Its rationale is to provide axes with a few large loadings and as many near zero loadings as possible. This is accomplished by an iterative maximization of a quadratic function of the loadings.
Devote the matrix of rotated loadings as

$\hat{\Delta} = \hat{\mathbf{\Lambda}}\mathrm{T}$

Now the $(i,j)^{th}$ element of $\mathbf{\Lambda}$ viz; δ_{ij} represents the loadings of the i^{th} variable on the j^{th} factor.
The function ϕ that the variance criterion maximizes is the sum of the variances of the squared loadings within each column of the loadings is normalized by its communality, that is

$$\phi = \sum_{i=1}^{k}\sum_{j=1}^{p}\left(d_{ij}^2 - \bar{d}_i\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{p} d_{ij}^4 - p\sum_{i=1}^{k}\bar{d}_i^2$$

Where $d_{ij} = \dfrac{\delta_{ij}}{h_i}$ *and* $\bar{d} = \dfrac{1}{p}\sum_{j=1}^{p} d_{ij}^2$

h_i^2 is the i^{th} communality is the i^{th} diagonal element of $\hat{\Delta}\,\hat{\Delta}'$
The varimax criterion ϕ is a function of T, and the iterative algorithm proposed by Kariser finds the orthogonal matrix $\mathbf{G}$ which maximizes ϕ.
In the case where k=2,the calculations simplify. For then T is given by

$$\mathrm{T}=\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

and represents a rotation of the coordinate axis clockwise by an angle θ. The value of θ can be determined by the relation $\mathrm{T}'\mathrm{T}=\mathrm{I}$.
In the case where k>2,an iterative solution for the rotation is used.
See example 9.8,9.9,9.10,9.11 in the pages 401-408 of AMVA by Richard Johanson & Wichern.

18.9 SUMMARY

This lesson explains Factor Analysis, a method that models the correlations among many observed variables using a smaller number of unobservable common factors. It introduces the orthogonal factor model, where observed variables are expressed as linear combinations of uncorrelated common factors and specific unique factors, with assumptions on their variances and independence. The covariance matrix is decomposed into communalities and specific variances, and factor analysis is shown to be scale invariant but allows non-unique factor loadings that can be rotated—commonly by the varimax method—to improve interpretability. The lesson covers estimation techniques such as Principal Factor Analysis, Principal

Component Method, and Maximum Likelihood, focusing on how to estimate factor loadings and specific variances from sample data, and emphasizes the importance of factor rotation for clearer, meaningful factor structures.

18.10 SELF ASSESSMENT QUESTIONS

1. What is the primary purpose of factor analysis and how does it explain correlations among variables?
2. How is the covariance matrix decomposed in the orthogonal factor model, and what do communalities and specific variances represent?
3. Why is factor analysis invariant to the scaling of variables, and how does this differ from principal component analysis?
4. What does the non-uniqueness of factor loadings mean, and how is this ambiguity resolved in factor analysis?
5. Describe the principal factor method of estimating factor loadings and communalities. How are communalities initially estimated?
6. Outline the principal component method for estimating factor loadings and specific variances. How does this method relate to the sample covariance or correlation matrix?
7. What is the maximum likelihood approach in factor analysis, and what are its main estimation steps?
8. Explain the purpose of factor rotation and how an orthogonal rotation preserves the covariance structure.
9. What is the varimax rotation method, and why is it commonly used in factor analysis?

18.11 SUGGESTED READINGS

1. **Applied Multivariate Statistical Analysis** by Richard A. Johnson and Dean W. Wichern
2. **An Introduction to Multivariate Statistical Analysis** by T.W. Anderson
3. **Multivariate Statistical Methods: A Primer** by Bryan F.J. Manly
4. **Multivariate Data Analysis** by Joseph F. Hair, William C. Black, et al.
5. **Psychometric Theory** by Jum C. Nunnally & Ira H. Bernstein

Dr. S. BHANU PRAKASH